Data Mining

What is Data Mining

- The process of discovering useful patterns, models, or algorithms from large datasets.
- Goals:
    - Modeling the data.
    - Extracting knowledge in the form of algorithms, statistical structures, or summaries.
    - Supporting decision-making across various domains.

Approaches to Data Mining

1. Modeling

- Data mining often involves creating a model from data (usually through machine learning).
- Example:
    - Detecting phishing emails → Build a *word-weight model* (positive for phishing words, negative for safe words).
    - Once model is ready, algorithm application is straightforward.
- Challenge: Building the *best model* is the most difficult step.

2. Statistical Modeling

- Origin of the term *data mining* comes from statistics.
- Initially, it had a negative meaning (*data dredging* – extracting patterns not supported by data).
- Today, it means constructing a statistical model of data.
- Example:
    - Data drawn from a Gaussian distribution.
    - The mean and standard deviation fully describe the model.

Data Mining

3. Machine Learning

- Some equate data mining with **machine learning (ML)**.
- ML uses data as **training sets** to build predictive models.
- Algorithms:
    - Bayesian networks
    - Support Vector Machines (SVM)
    - Decision Trees
    - Hidden Markov Models
    - Deep Learning
- **Usefulness**: Best when the data's relationship to the problem is unclear (e.g., *Netflix Challenge* – predicting movie ratings).
- **Limitations**:
    - Sometimes direct algorithms perform better (e.g., resume detection).
    - Many ML models (especially deep learning) lack **explainability**, which may be unacceptable in sensitive areas (e.g., insurance premiums).

4. Computational/Algorithmic Modeling

Computational modeling is the computer science approach to data mining, where the focus is on designing algorithms to analyze and represent data, rather than only building statistical or machine-learning models.

- **Idea:** Treat data mining as solving a complex query or problem using computation.
- **Example:** Instead of fitting data to a Gaussian distribution (statistics), we may just compute the **mean** and **standard deviation** directly as answers.
- **Approaches:**
    1. **Summarization** → Condense data into compact forms (e.g., PageRank, clustering).
    2. **Feature Extraction** → Focus on the most important aspects of data (e.g., frequent itemsets, collaborative filtering).

Data Mining

Core Techniques in Data Mining

1. Summarization

- Compactly representing large data using concise statistics or measures.
- Examples:
    - **PageRank (Google):** Assigns a single number to each webpage reflecting its importance.
    - **Clustering:** Grouping data points close in a multidimensional space.
        - Example: *John Snow (1854)* used clustering of Cholera cases on a London map to trace contaminated wells.

2. Feature Extraction

- Identifies the **most significant features** or relationships in the data.
- Examples:

1. Frequent Itemsets

- A model used when data comes in the form of baskets (small sets of items).
- Goal : Find groups of items that frequently appear together across many baskets.
- Classic Example (Market Basket Problem):
    - Each basket = items a customer buys in one shopping trip.
    - If "hamburger" and "ketchup" often appear together in baskets, this pair becomes a *frequent itemset*.
- Usefulness:
    - Helps in association rule mining (e.g., “If a customer buys hamburger, they are likely to also buy ketchup”).
    - Applied in supermarkets, recommendation systems, cross-selling strategies.

2. Similar Items

- A model where data is treated as a collection of sets, and the task is to find pairs of sets with high overlap.

- Example (Amazon Customers):
    - Each customer = set of items they bought.
    - To recommend new products:
        - Find customers who have similar purchase histories.
        - Recommend items that similar customers bought.
- This process is called *Collaborative Filtering*.
    - Instead of clustering all customers into fixed groups, it works better to find a few "nearest neighbors" (customers with similar tastes).
- Why not clustering?
    - Customers usually have multiple diverse interests.
    - Example: One customer may buy both *books* and *electronics*.
    - Clustering into only one group may miss this variety.
    - Collaborative filtering allows multiple similarity-based recommendations.

Statistical Limits on Data Mining

- Data mining often aims to discover unusual events hidden in massive data.
- Challenge: Over-mining leads to false positives (patterns that appear significant but are random artifacts).
- Bonferroni's Principle warns against this overzealous interpretation.

1. Total Information Awareness (TIA)

- **Background:**
    - After the 9/11 terrorist attacks (2001), it was noted that four people were enrolled in flight schools, learning to fly commercial aircraft without airline affiliations.
    - This raised the idea that early signs of terrorism might already exist in available data but were not detected.

Data Mining

- **The Program:**
    - TIA (Total Information Awareness) was created to integrate and mine multiple data sources:
        - Credit card records
        - Hotel bookings
        - Travel data
        - Other personal information
    - Goal: Detect and track suspicious/terrorist activities.
- **Outcome:**
    - TIA faced **major privacy concerns**.
    - Eventually shut down by **U.S. Congress** due to fears of surveillance misuse.
    - Technical Problem**:** Looking for too many patterns simultaneously → many innocent or irrelevant activities will look suspicious.

2 Bonferroni's Principle

- **Concept:** Even if data is random, some unusual events will appear significant by chance.
- **Bonferroni Correction (formal):** A statistical adjustment to reduce false positives when testing many hypotheses.
- **Informal Rule (Principle):**
    - Calculate the **expected number of occurrences** of the event under the assumption of randomness.
    - If this expected number is much larger than the actual number of real cases you hope to find → most findings will be **bogus**.
- **Implication:** To detect rare events (like terrorists), you must search for patterns so rare that they are *unlikely to occur in random data*.

The Core Idea

When you search through **a lot of data** for "unusual" events,
you will almost always find **something unusual** —
but most of the time, it's just **random coincidence**, not something meaningful.

Data Mining

So, **Bonferroni's Principle** warns us:
Before believing an event is meaningful, check whether it could simply happen by chance, given the amount of data you're looking at.

Example

Imagine you flip a coin:

- Probability of getting 10 heads in a row is very small (1 in 1024).
- If **one person** flips coins 10 times, and they get 10 heads in a row → that looks special.

But now imagine:

- **A million people** each flip a coin 10 times.
- Even though the probability is small, among so many people, you can **expect** that someone will get 10 heads in a row *just by chance*.

So, if you see one person with 10 heads in a row, you should NOT immediately think they have a "magic coin." It's probably just coincidence.

Example of Bonferroni's Principle

- **Scenario Assumptions:**
    1. 1 billion possible suspects.
    2. Each person goes to a hotel 1 day out of 100.
    3. Hotels hold 100 people → about 100,000 hotels.
    4. Data collected over 1000 days.
- **Detection Rule:** Look for pairs of people who are in the same hotel on two different days.
- **Calculation (assuming all behavior is random):**
    - Probability that two people meet in the same hotel on one day = 10^{-9}
    - Probability they meet on two different days = 10^{-18}
    - Number of pairs of people ≈ 5×10^{17}
    - Number of pairs of days ≈ 5×10^{5}
    - Expected "suspicious" events = $5\times10^{17} \times 5\times10^{5} \times 10^{-18}$ =250,000.

- **Interpretation:**
    - Even with **no evil-doers**, 250,000 innocent pairs will look suspicious.
    - If there are only 10 true terrorist pairs, police must sift through a huge number of false alarms.
    - **Conclusion:** Approach is statistically infeasible.

Importance of Words in Documents

- In data mining, document classification often requires identifying important words.
- Common words (e.g., *the, and*) are not useful → called **stop words** (usually removed).
- Rare words can be more useful indicators, but not all rare words are informative.
    - Example: *albeit* (rare but not topic-specific).
    - Example: *chukker* (rare and strongly linked to polo).

TF.IDF (Term Frequency × Inverse Document Frequency)

A statistical measure that evaluates how important a word is to a document relative to a collection of documents.

(a) Term Frequency (TF)

$$TF = \frac{\text{fij}}{max\,(f)}$$

- f_{ij}: frequency of term *i* in document *j*.
- Denominator: highest frequency of any word in document *j*.
- Normalizes word frequency → most frequent word in a document has TF = 1.

(b) Inverse Document Frequency (IDF)

$$IDF = log_2(N / n_i)$$

- N: total number of documents.
- n_i: number of documents containing term *i*.
- Rare terms across the collection get higher IDF values.

(c) TF.IDF Score

$$TF.IDF_{ij} = TF_{ij} \times IDF_i$$

- High score → word is frequent in a document but rare in the collection.
- Useful to identify topic-specific words.

---

3. Example

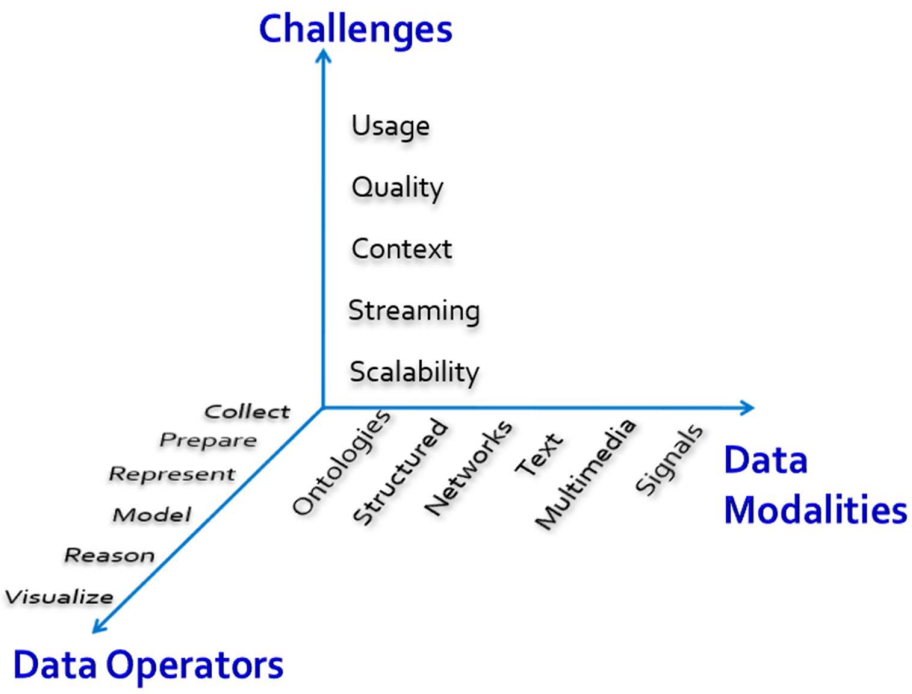
3. Example

- Collection size: $N = 2^{20} = 1,048,576$ docs.
- Word *w* appears in $2^{10} = 1024$ docs →

$$IDF_w = \log_2\left(\frac{2^{20}}{2^{10}}\right) = \log_2(2^{10}) = 10$$

- **Case 1 (Document j):**
  - *w* appears 20 times (max frequency in doc).
  - $TF_{wj} = 20/20 = 1.$
  - $TF.IDF = 1 \times 10 = 10.$
- **Case 2 (Document k):**
  - *w* appears once, max word frequency = 20.
  - $TF_{wk} = 1/20 = 0.05.$
  - $TF.IDF = 0.05 \times 10 = 0.5.$

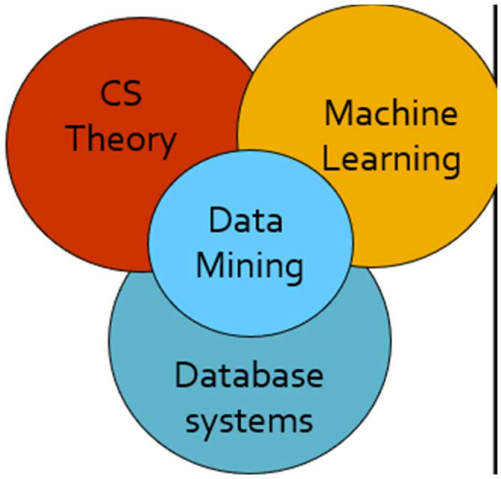
Data Mining



Data Mining: Cultures

Overlapping Fields

- Databases (DB): Large-scale data, simple queries.
- Machine Learning (ML): Small data, complex models.
- CS Theory: (Randomized) algorithms.



Data Mining

Perspectives in Data Mining

Culture	View of Data Mining	Output
Database person	Extreme form of analytic processing – queries over large datasets	Query answers
Machine learning person	Inference of models from data	Model parameters

Focus of This Course

- **Overlap with:** ML, statistics, AI, databases.
- **Emphasis on:**
    - Scalability → handling **big data** efficiently.
    - Algorithms → designing efficient, correct methods.
    - Computing architectures → hardware/software considerations.
    - Automation → reducing human effort in processing large datasets.