
# Introduction to Artificial Intelligence

## Subject 6: Introduction to Data Mining

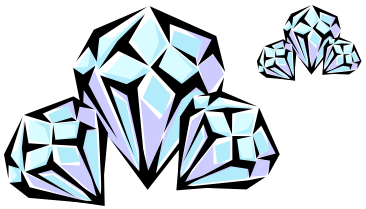
# Contents..

- What is data mining
- Why it is needed
- Data types & attributes
- KDD Framework
- Evaluation and Interpretation of Results

# WHAT IS DATA MINING

# What is data mining?

Data mining is the process of finding **useful** and often **hidden patterns** from very large amounts of data by using advanced techniques.



# Why do we need data mining?

- Really, really huge amounts of raw data!!
  - In the digital age, TB of data is generated by the second
    - Mobile devices, digital photographs, web documents.
    - Facebook updates, Tweets, Blogs, User-generated content
    - Transactions, sensor data, surveillance data
    - Queries, clicks, browsing
  - Cheap storage has made possible to maintain this data
- Need to analyze the raw data to extract knowledge

# WHY IT IS NEEDED

# Why do we need data mining?

- Large amounts of data can be more powerful than complex algorithms and models
  - Google has solved many Natural Language Processing problems, **simply by looking at the data**
  - Example: misspellings, synonyms

- Data is power!
  - Today, the collected data is one of the biggest assets of an online company
    - Query logs of Google
    - The friendship and updates of Facebook
    - Tweets and follows of Twitter
    - Amazon transactions

# So, what is Data?

- Collection of data objects and their attributes
- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | No | Married | 60K | No |
| 7 | Yes | Divorced | 220K | No |
| 8 | No | Single | 85K | Yes |
| 9 | No | Married | 75K | No |
| 10 | No | Single | 90K | Yes |

Size: Number of objects
Dimensionality: Number of attributes

# DATA TYPES & ATTRIBUTES

# Types of data

- Numeric data: Data that represents quantifiable values and can be measured on a scale.
- Categorical data: Data used to label or categorize items into distinct groups.
- Transaction Data: A special form of record data where each record represents a single event or exchange and involves a set of items.
- Document Data: Data that is typically unstructured text from sources like reports, emails, or web pages.
- Ordered sequences: Data where the order of the elements is significant and carries crucial information.
- Graph data: Data representing entities (nodes) and the relationships (edges) between them.

# Types of data: Categorical Data

- Data that represents qualitative characteristics or labels that describe a group or category.

**Types of Categorical Data:**

1. **Nominal:** Categories with **no inherent order** (e.g., Colors, Countries, Brands).
2. **Ordinal:** Categories with a **meaningful order or rank** (e.g., Size: Small, Medium, Large / Satisfaction: Low, Medium, High).

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | High | No |
| 2 | No | Married | Medium | No |
| 3 | No | Single | Low | No |
| 4 | Yes | Married | High | No |
| 5 | No | Divorced | Medium | Yes |
| 6 | No | Married | Low | No |
| 7 | Yes | Divorced | High | No |
| 8 | No | Single | Medium | Yes |
| 9 | No | Married | Medium | No |
| 10 | No | Single | Medium | Yes |

# Types of data: Numerical Data

- Data that represents quantitative, measurable quantities as numbers.

**Types of Numerical Data:**

1. **Discrete:** Countable numbers, usually whole numbers (e.g., Number of cars, Number of children).
2. **Continuous:** Measurable numbers that can take any value within a range, often with decimals (e.g., Height, Weight, Time).

| *Tid* | Refund | Marital Status | Taxable Income | Cheat |
|---|---|---|---|---|
| 1 | Yes | Single | High | No |
| 2 | No | Married | Medium | No |
| 3 | No | Single | Low | No |
| 4 | Yes | Married | High | No |
| 5 | No | Divorced | Medium | Yes |
| 6 | No | Married | Low | No |
| 7 | Yes | Divorced | High | No |
| 8 | No | Single | Medium | Yes |
| 9 | No | Married | Medium | No |
| 10 | No | Single | Medium | Yes |

# Types of data: Document Data

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.
  - Bag-of-words representation – no ordering

| | team | coach | play | ball | score | game | win | lost | timeout | season |
|---|---|---|---|---|---|---|---|---|---|---|
| Document 1 | 3 | 0 | 5 | 0 | 2 | 6 | 0 | 2 | 0 | 2 |
| Document 2 | 0 | 7 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| Document 3 | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 3 | 0 |

# Types of data: Transaction Data

- Each record (transaction) is a set of items.

| *TID* | *Items* |
|---|---|
| 1 | Bread, Coke, Milk |
| 2 | Chips, Bread |
| 3 | Chips, Coke, Diaper, Milk |
| 4 | Beer, Bread, Diaper, Milk |
| 5 | Coke, Diaper, Milk |

- A set of items can also be represented as a binary vector, where each attribute is an item.
- A document can also be represented as a set of words (no counts)

Sparsity: average number of products bought by a customer

# Types of data: Ordered Data

- Genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

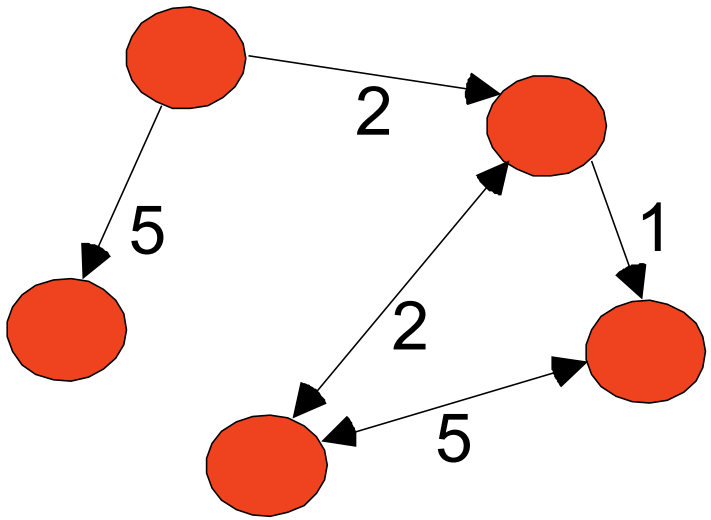
- Data is a long ordered string

# Types of data: Ordered Data

- Time series
  - Sequence of ordered (over “time”) numeric values.

# Types of data: Graph Data

- Examples: Web graph and HTML Links




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

# EXAMPLES OF REAL-WORLD BIG DATA

# Examples of Real-World Big Data: transaction data

- Billions of real-life customers:
  - WALMART: **20 M** transactions per day
  - AT&T **300 M** calls per day
  - Credit card companies: **billions** of transactions per day.

- **The point cards** allow companies to collect information about specific users

# Examples of Real-World Big Data: document data

- Web as a document repository: estimated **50 billions** of web pages
- Wikipedia: **4 million** articles (and counting)
- Online news portals: steady stream of 100's of new articles every day
- X (Twitter): ~300 million tweets every day

# Examples of Real-World Big Data: network data

- Web: 50 billion pages linked via hyperlinks
- Facebook: 500 million users
- Twitter: 300 million users
- Instant messenger: ~1billion users
- Blogs: 250 million blogs worldwide, presidential candidates run blogs

# Examples of Real-World Big Data: environmental data

- Climate data (just an example)
- "a database of temperature, precipitation and pressure records managed by the National Climatic Data Center, Arizona State University and the Carbon Dioxide Information Analysis Center"
- "6000 **temperature stations**, 7500 **precipitation stations**, 2000 **pressure stations**"
  - Spatiotemporal data

# Examples of Real-World Big Data: Behavioral data

- Mobile phones today record a large amount of information about the user behavior
  - GPS records position
  - Camera produces images
  - Communication via phone and SMS
  - Text via facebook updates
- Amazon collects all the items that you browsed, placed into your basket, read reviews about, purchased.
- Google and Bing record all your browsing activity via toolbar plugins. They also record the queries you asked, the pages you saw and the clicks you did.
- Data collected for millions of users on a daily basis

# KDD FRAMEWORK

# Knowledge Discovery in Databases (KDD)

- The KDD is a comprehensive, multi-step process designed to extract **valid**, **novel**, **useful**, and **understandable** knowledge from data.

# The Steps of Knowledge Discovery (KDD) process

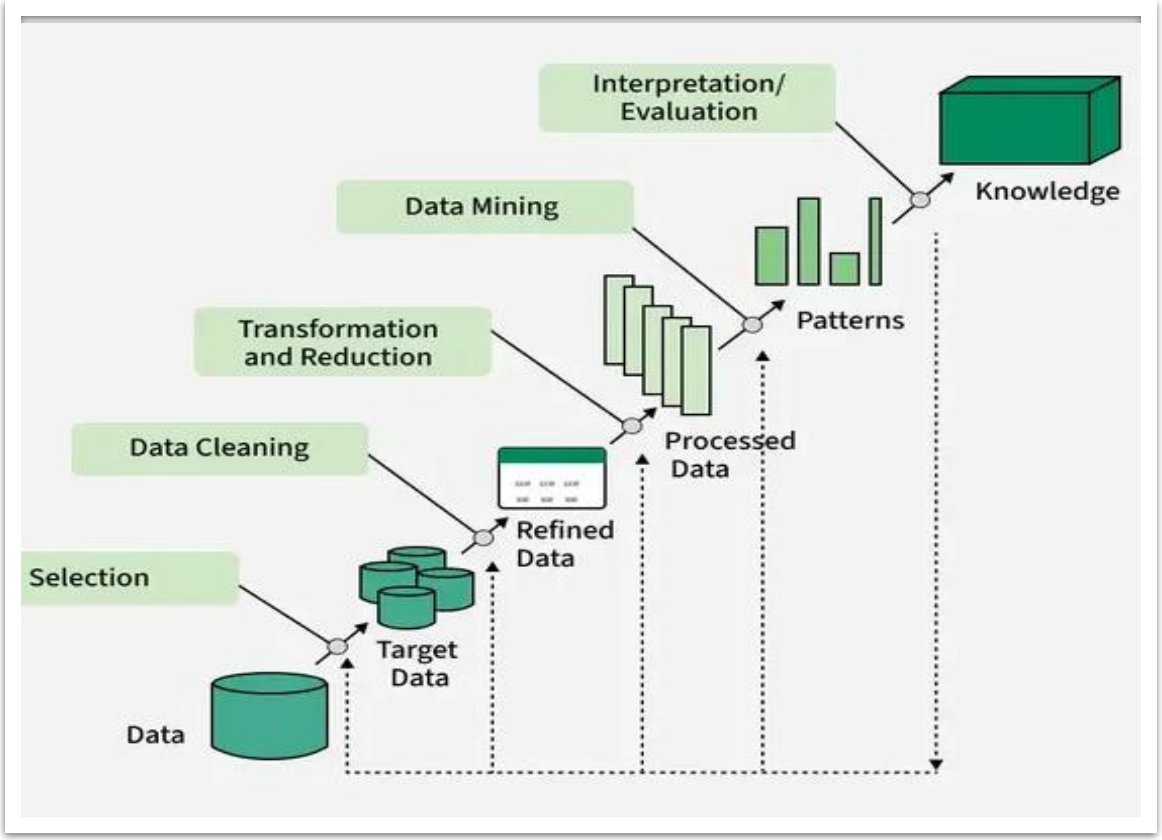
**Step 1: Selection**
We start by gathering the raw data we are interested in, and then selecting only the parts that are relevant to our problem.

**Step 2: Data Cleaning**
We "clean" the selected data by fixing errors and problems.

**Step 3: Transformation and Reduction**
We reshape and simplify the data to make it ready for analysis.

**Step 4: Data Mining**
This is the core step where we use algorithms to automatically search for patterns in the prepared data.

**Step 5: Interpretation/Evaluation**
We look at the discovered patterns, check if they are useful and meaningful, and turn them into knowledge.




https://www.geeksforgeeks.org/dbms/kdd-process-in-data-mining/

# KDD FRAMEWORK:

## STEP 1. DATA SELECTION

# Sampling

- Sampling is the main technique employed for data selection.
  - It is often used for both the preliminary investigation of the data and the final data analysis.
- Statisticians take samples because **obtaining** the entire set of data of interest is too expensive or time consuming.
- Sampling is used in data mining because **processing** the entire set of data of interest is too expensive or time consuming.
- Example: From a company's full database, we choose only the customer purchase records from the last year.

# Sampling …

- The **key principle** for effective sampling is the following:
  - Using a sample will work almost **as efficiently as** using **entire datasets,** if the sample is representative.
  - A sample is representative if it has approximately the same (significant) property as the original dataset.

# Types of Sampling

1. **Simple Random Sampling**
   - There is an equal probability of selecting any particular item.
2. **Sampling without replacement**
   - As each item is selected, it is removed from the population.
3. **Sampling with replacement**
   - Objects are not removed from the population as they are selected for the sample.
     - In sampling with replacement, the same object can be picked up more than once.
4. **Stratified sampling**
   - Split the data into several partitions; then draw random samples from each partition.

# KDD FRAMEWORK:

## STEP 2. DATA CLEANING AND PREPROCESSING

# Data Cleaning

- Data Cleaning is a foundational step in the KDD process, ensuring that the dataset is accurate, consistent, and suitable for meaningful analysis.
- Poor-quality data can distort patterns, compromise model performance, and lead to incorrect decisions.
- This stage focuses on detecting and resolving errors, inconsistencies, and imperfections in the raw data.
- Example: We fill in missing customer ages, correct typos in product names, and remove duplicate entries.

# Key Components of Data Cleaning

**1. Handling Missing Values**

- Missing data can arise from system errors, user mistakes, corrupted inputs, etc. **To maintain completeness and avoid bias:**
  - **Imputation Techniques:** Replace missing values using the mean, median, or mode for numerical or categorical variables.
  - **Predictive Imputation:** Use models (e.g., k-NN) to estimate the most probable value.
  - **Deletion:** Remove records only when the missingness is extensive or irreparable.

**2. Reducing Noisy Data**

- Noisy data contains **random errors and outliers**. To smooth or correct noise:

**3. Removing Duplicates**

- Duplicate records can skew statistics and mislead analysis.

**4. Exploratory Data Analysis (EDA)**

- EDA is performed to understand data distributions, and trends before running algorithms.
- It helps decide how to transform the data and which mining techniques to use.

# KDD FRAMEWORK:

## STEP 3. DATA TRANSFORMATION AND REDUCTION

# Data Transformation and Reduction

**Data transformation** reshapes data into formats that algorithms can process using various techniques, such as:

**1. Normalization**

- Standardizes data by scaling features into a common range (e.g., 0–1 or −1–1).
- Prevents attributes with large numeric ranges from dominating others.
- Improves performance for distance-based algorithms (e.g., k-means, k-NN).

**2. Discretization**

- Converts continuous variables into discrete categories.
- Example: We convert salaries into categories (e.g., "Low," "Medium," "High")

**3. Data Aggregation**

- Combines multiple data points to create summary-level information.
- Examples: daily averages, monthly totals, or summarized transactions.

# Data Transformation and Reduction

**Data reduction** reduces data size without losing essential information using various techniques, such as:

**1. Principal Component Analysis (PCA)**

- Reduces the number of variables by creating new, combined features that capture the most important patterns in the data.
- Example: Instead of using 10 smartphone sensors to classify physical activity, PCA can combine these 10 sensor streams into 3-4 new, independent "meta-sensors" that capture the essential patterns of movement.

**2. Numerosity Reduction**

- Reduces the number of records or objects while maintaining the overall data distribution and key patterns.
- Example: Replacing a million daily sales transactions with 365 daily summary totals.

**3. Data Compression**

- Encodes data more compactly without losing critical information. Facilitates faster storage, retrieval, and processing.
- Example: Using an algorithm for JPEG images or MP3 for audio to significantly reduce file size while preserving perceptual quality.

# KDD FRAMEWORK:

## STEP 4. DATA MINING

# What is Data Mining again?

- Data Mining is the process of discovering valuable, previously unknown patterns from large datasets through automatic or semi-automatic means.

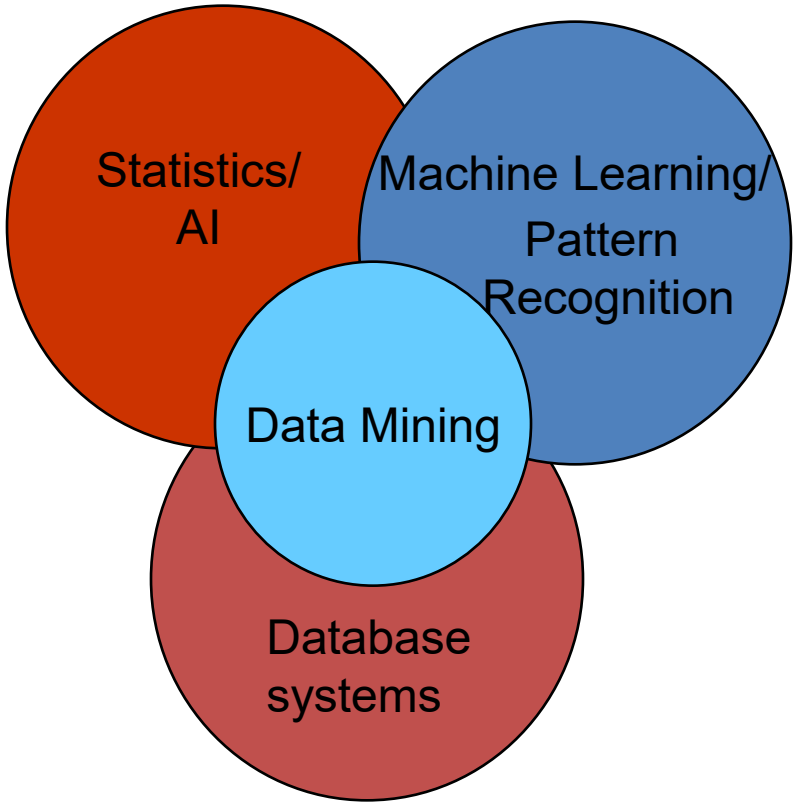
- It involves exploring vast amounts of data to extract useful information that can drive decision-making.

# Why data mining?

- Commercial point of view
  - Data has become the key competitive advantage of companies
    - Examples: Facebook, Google, Amazon
  - Being able to extract useful information out of the data is key for exploiting them commercially.
- Scientific point of view
  - Scientists are at a unique position as they can gather enormous amounts of information.
    - Examples: Sensor data, social network data, gene data
  - We need the tools to analyze such data to get a better understanding of the world and advance science
- Scale (in data size and feature dimension)
  - Why not use traditional analytic methods?
  - The amount and the complexity of data does not allow for manual processing of the data. We need automated techniques.

# Connections of Data Mining with other areas

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be **unsuitable** due to:
  - Massiveness of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



# Connections of Data Mining with other areas

- Databases: Concentrate on large-scale data storage and efficient retrieval (querying).
- AI (machine-learning): Focuses on learning patterns from data to enable intelligent decision-making and automation.
- Statistics: Centers on inference, measurement of uncertainty, and precise analytical methods to generalize findings from samples to populations.

# Meaningfulness of Answers

- A big data-mining **risk** is that you will "discover" patterns that are **meaningless**.
- Statisticians call it Bonferroni's principle: (roughly) if you look in more places for interesting patterns than your amount of data will support, you are bound to find inaccurate results.

Tan, M. Steinbach and V. Kumar,
Introduction to Data Mining

# What can we do with data mining?

- Some examples:
  1. Frequent itemsets and Association Rules extraction
  2. Clustering
  3. Coverage
  4. Classification
  5. Ranking

# 1. Frequent Itemsets and Association Rules

- Given a set of records each of which contain some number of items from a given collection;
  - Identify sets of items (itemsets) occurring frequently together
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

| *TID* | *Items* |
|---|---|
| 1 | Bread, Coke, Milk |
| 2 | Chips, Bread |
| 3 | Chips, Coke, Gum, Milk |
| 4 | Chips, Bread, Gum, Milk |
| 5 | Coke, Gum, Milk |

Itemsets Discovered:
{Milk,Coke}
{Gum, Milk}

Rules Discovered:
{Milk} --> {Coke}
{Gum, Milk} --> {Chips}

# 1. Frequent Itemsets: Applications

- Text mining: finding associated phrases in text
  - There are lots of documents that contain the phrases "association rules", "data mining" and "efficient algorithm"

- Recommendations:
  - Users who buy item 'X' often buy item 'Y' as well.
  - Users who watched James Bond movies, also watched Jackie Chan movies.
  - Recommendations make use of item and user similarity.

# Association Rule Discovery: Application

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys gum and milk, then he is very likely to buy chips.
    - So, don’t be surprised if you find six-packs stacked next to gums!

# 2. Clustering Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.

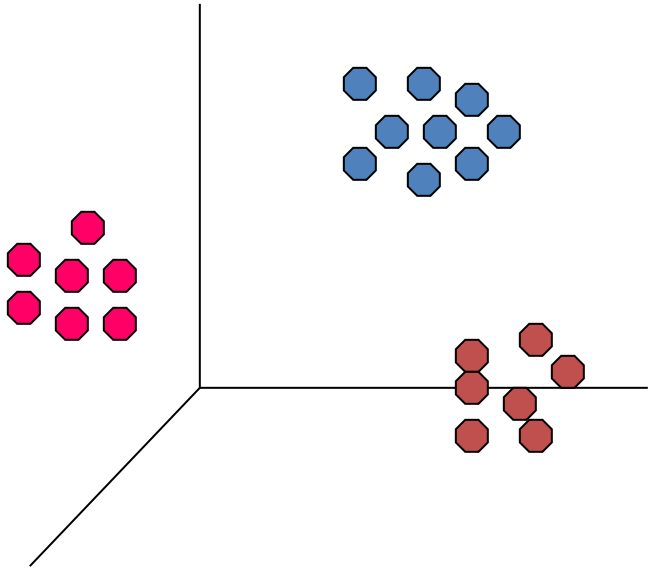
- Similarity Measures?
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

# Illustrating Clustering

Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



# Clustering: Application of Document Clustering

- Document Clustering:
  - Goal: To find groups of documents that are **similar** to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to link a new document or search term to the collected documents.

# 3. Coverage

- **Data coverage** refers to the extent to which a dataset covers the full range of scenarios, conditions, or variations that a system might encounter in the real world.

- Given a set of customers, a set of items, and a **record** of which customer purchased which items, the goal is to identify the smallest possible subset of items so that every customer is "covered".

- A customer is considered **covered** if the selected set contains **at least one item that the customer has purchased**.

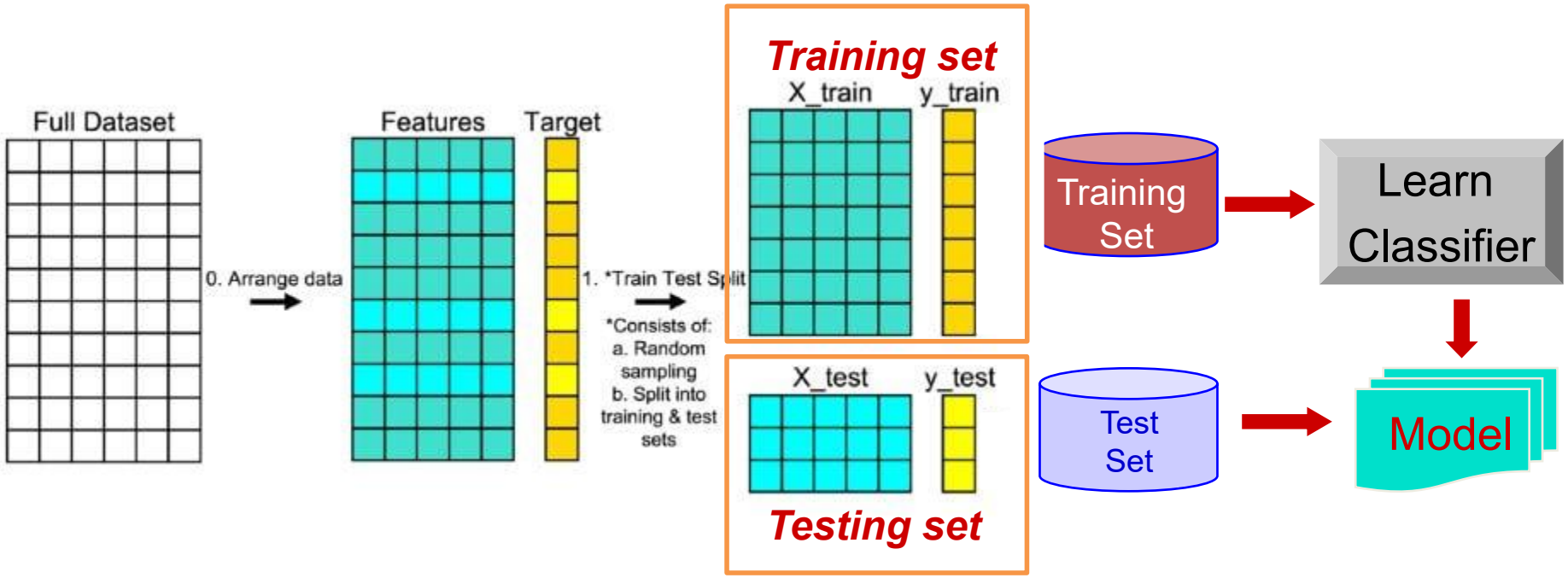
- Application:
  - A company wants to design a promotional catalog to mail to customers.

# 4. Classification: Definition

- Given a collection of records (*training set* )
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes. *F(class)=?*

- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model.
  - Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

# Classification Example



# Classification: Email Spam Filtering (Application1)

- **Goal:** To determine if an incoming email message is legal (legitimate) or unwanted (**Spam**).
- **Input Features (X):** Text content of the email (word frequency, punctuation, capitalization), sender's address, and presence of suspicious keywords (e.g., "free," "winner," “money").
- **Class Attribute (Y):** A **binary label** that is either **[Spam]** or **[Not Spam]**.
- **Application:** Automatically moving unwanted emails into the junk folder to improve user experience and security.

# Classification: Medical Diagnosis (Application2)

- **Goal:** To predict if a patient has a specific medical condition (e.g., diabetes, heart disease) based on their clinical data.
- **Input Features (X):** Patient measurements (e.g., blood pressure, blood sugar level, BMI, age, family history, ECG results).
- **Class Attribute (Y):** A **binary or multi-class label** indicating the diagnosis, such as **[Positive for Condition]** or **[Negative for Condition]**.
- **Application:** Assisting doctors in making faster, more accurate preliminary diagnoses and identifying high-risk patients.

# 5. Ranking: Link Analysis Ranking Example

- The concept behind Link Analysis Ranking (like PageRank) is simple: **A page is important if important pages link to it.**
  - Simple Counting: If Page A has 10 links pointing to it, and Page B has 5 links pointing to it, **Page A** is more **important**.

- If Page A is linked to by The New York Times (a highly authoritative source), and Page B is linked to by 10 unknown blogs, Page A is still more important.

# KDD FRAMEWORK:

## STEP 5. EVALUATION AND INTERPRETATION OF RESULTS

# Pattern Evaluation

- Pattern Evaluation is the stage in the KDD process where the **discovered** patterns from data mining are **assessed** for their relevance, validity, novelty, and usefulness.

- **Not all** extracted patterns are meaningful, so this step helps filter out noise and highlight truly valuable insights.

- A key component of this stage is computing an interestingness scores, which may include **statistical measures** (e.g., confidence, accuracy).

- **To assist understanding, visualization and summarization techniques**—such as charts, heatmaps, association rule graphs, or cluster profiles—**are applied**. These tools make complex patterns easier for analysts to interpret and compare.

# Interpretation and Knowledge Presentation

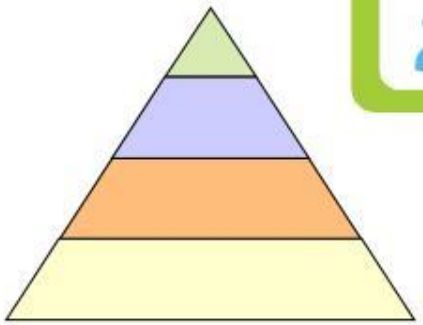
- Interpretation involves transforming the evaluated patterns into **clear, actionable knowledge.**
- This stage **focuses on explaining the results** in a way that **aligns** with **business goals or research objectives**.
- **Effective communication**—through dashboards, reports, data stories, or visual summaries—enables decision-makers to quickly grasp findings and translate them into informed actions and strategies.

# Data Interpretation

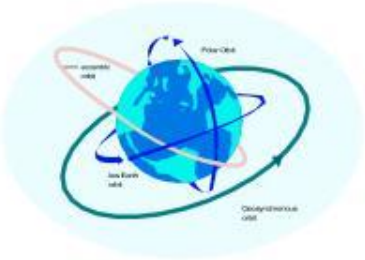
**Includes……**

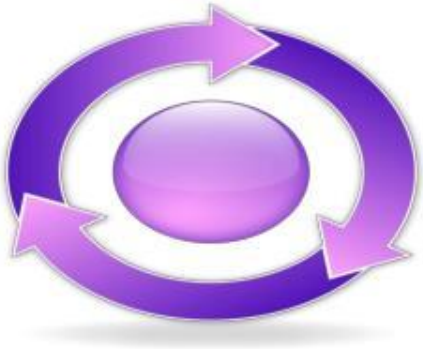
## Graphs


## Charts



## Diagrams & Figures





## Tables

| Name | Thread pitch (mm) | Minor diameter tolerance | Nominal diameter (mm) | Head shape | Price for 50 screws | Available at factory outlet? | Number in stock | Flat or Phillips head? |
|---|---|---|---|---|---|---|---|---|
| M4 | 0.7 | 4g | 4 | Pan | $10.08 | Yes | 276 | Flat |
| M5 | 0.8 | 4g | 5 | Round | $13.89 | Yes | 183 | Both |
| M6 | 1 | 5g | 6 | Button | $10.42 | Yes | 1043 | Flat |
| M8 | 1.25 | 5g | 8 | Pan | $11.98 | No | 298 | Phillips |
| M10 | 1.5 | 6g | 10 | Round | $16.74 | Yes | 488 | Phillips |
| M12 | 1.75 | 7g | 12 | Pan | $18.26 | No | 998 | Flat |
| M14 | 2 | 7g | 14 | Round | $21.19 | No | 235 | Phillips |
| M16 | 2 | 8g | 16 | Button | $23.57 | Yes | 292 | Both |
| M18 | 2.1 | 8g | 18 | Button | $25.87 | No | 664 | Both |
| M20 | 2.4 | 8g | 20 | Pan | $29.09 | Yes | 486 | Both |
| M24 | 2.55 | 9g | 24 | Round | $33.01 | Yes | 982 | Phillips |
| M28 | 2.7 | 10g | 28 | Button | $35.66 | No | 1067 | Phillips |
| M36 | 3.2 | 12g | 36 | Pan | $41.32 | No | 434 | Both |

# Difference Between KDD and Data Mining

| Parameter | KDD (Knowledge Discovery in Databases) | Data Mining |
|---|---|---|
| **Definition** | A complete end-to-end process of discovering valid, useful, and understandable knowledge from data. | A step within KDD that focuses specifically on extracting patterns and insights from data. |
| **Objective** | To produce meaningful knowledge that supports interpretation and decision-making. | To find patterns, trends, or relationships in data. |
| **Scope/Steps** | Includes data selection, cleaning, transformation, mining, evaluation, and knowledge representation. | Primarily uses algorithms to analyze data and extract patterns. |
| **Techniques Used** | Uses a combination of preprocessing, transformation, mining, and evaluation methods. | Uses algorithms such as classification, clustering, regression, association rules, etc. |
| **Output** | Generates structured knowledge and insights that can aid in decision-making or predictions. | Raw patterns or relationships detected in the data. |
| **Focus** | Focuses on the discovery of useful knowledge. | Focuses on the analytical algorithms used for pattern extraction. |

# References


- Tan, M. Steinbach and V. Kumar, Introduction to Data Mining.
- https://www.cs.uoi.gr/~tsap/teaching/2012f-cs059/slides-en.html
- KDD Process in Databases, geeksforgeeks, https://www.geeksforgeeks.org/dbms/kdd-process-in-data-mining/

That's all
for Today