Unit 1
Introduction

introduction, KDD, data pre-processing, similarity measurement, data visualization

Rupak Raj Ghimire

Objective

- Introduction
- KDD
- Data pre-processing,
- Similarity measurement
- Summary Statistics
- Data visualization

Data

- Data is a collection of facts, figures, statistics, or any other type of information that can be recorded and analyzed.
  - **Forms**: text, numbers, images, audio, video
  - **Source of Data**:
    - created by people, machines, or sensors, and
    - collected from a wide range of sources, such as websites, social media, databases, and sensors.

Information

- Information is a collection of data that has been processed, organized, or structured in a way that makes it meaningful, useful, and relevant to a particular context or purpose
- Information is created when data is analyzed, interpreted, and presented in a way that can be easily understood and used by humans or machines
- Information provides knowledge, insight, and understanding about a particular topic, situation, or phenomenon
- It can be used to support decision-making, problem-solving, and communication
- It conveyed in the form of reports, charts, graphs, tables, or other visual representations

Example of Data and Information

- Set of Marks = { 2, 5, 7, 9, 11 }
    - It is a data
    - This dataset is considered to be data because it is a collection of raw, unorganized numbers that don't necessarily convey any meaning or context on their own.
- Average of these numbers = 6.8
    - It is now information
    - Interpreted, presented it in a report or a chart, along with some context and explanation, such as "the average score on a test for a group of students", then we have turned the data into information.

Database

- A database is an organized collection of data that is stored and managed using specialized software
- A database allows users to store, retrieve, update, and delete data in an efficient manner (Operations)
- Example
  - The data in a relational database is organized into tables, which are composed of rows and columns.

    Each row represents a record, and each column represents a specific piece of information about that record.
  - For example, a table in a customer database might include columns for the customer's name, address, phone number, and email address.

Database Management System (DBMS)

- A DBMS is a software system that is designed to manage and manipulate databases
- It provides a set of tools and services that allow users to store, access, modify, and maintain data in an organized and secure way
- A DBMS provides a way to create and manage databases, define the data structures, and enforce data integrity
  - Example: PostgreSQL, Oracle, MySQL etc.

Type of Databases

- Relational Database
- Object Oriented Database
- NoSQL
- Graph Database
- Network Database etc.

Retrieving Data from Database

- SQL is a programming language used to communicate with and manipulate databases.
- Type of SQL
  - Data Query Language (DQL) – select
  - Data Manipulation Language (DML) – insert, update,delete
  - Data Definition Language (DDL) – create
  - Data Control Language (DCL) – grant, revoke
- Reference
  - https://learnsql.com/blog/sql-basics-cheat-sheet/

Transactional Database

- An operational database system (also known as a transnational database system) is a type of database that is designed to support the day-to-day operations of an organization.
  - It is optimized for transnational processing, which involves capturing, storing, and updating data in real-time as business transactions occur
  - Typically used to support online transaction processing (OLTP), which involves frequent and rapid database access and updates by multiple users simultaneously

Transactional Database

- Operational database systems are designed to ensure data consistency, reliability, and availability
- ACID
  - Atomicity - Transaction
  - Consistency - Data Quality
  - Isolation - Concurrency
  - Durability - Recovery

Data warehouse

- A data warehouse refers to a data repository that is maintained separately from an organization’s operational databases
- A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management’s decision making process - William H. Inmon

Data warehouse

- Subject-oriented
  - A data warehouse is organized around major subjects such as customer, supplier, product, and sales.
  - Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers.
  - Data warehouses typically provide a simple and concise view of particular subject issues by excluding data that are not useful in the decision support process.

Data warehouse

- Integrated:
  - A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and online transaction records.
  - Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on
- Time-variant:
  - Data are stored to provide information from an historic perspective (e.g., the past 5–10 years). Every key structure in the data warehouse contains,either implicitly or explicitly, a time element.

Data warehouse

- Nonvolatile:
  - A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: initial loading of data and access of data.

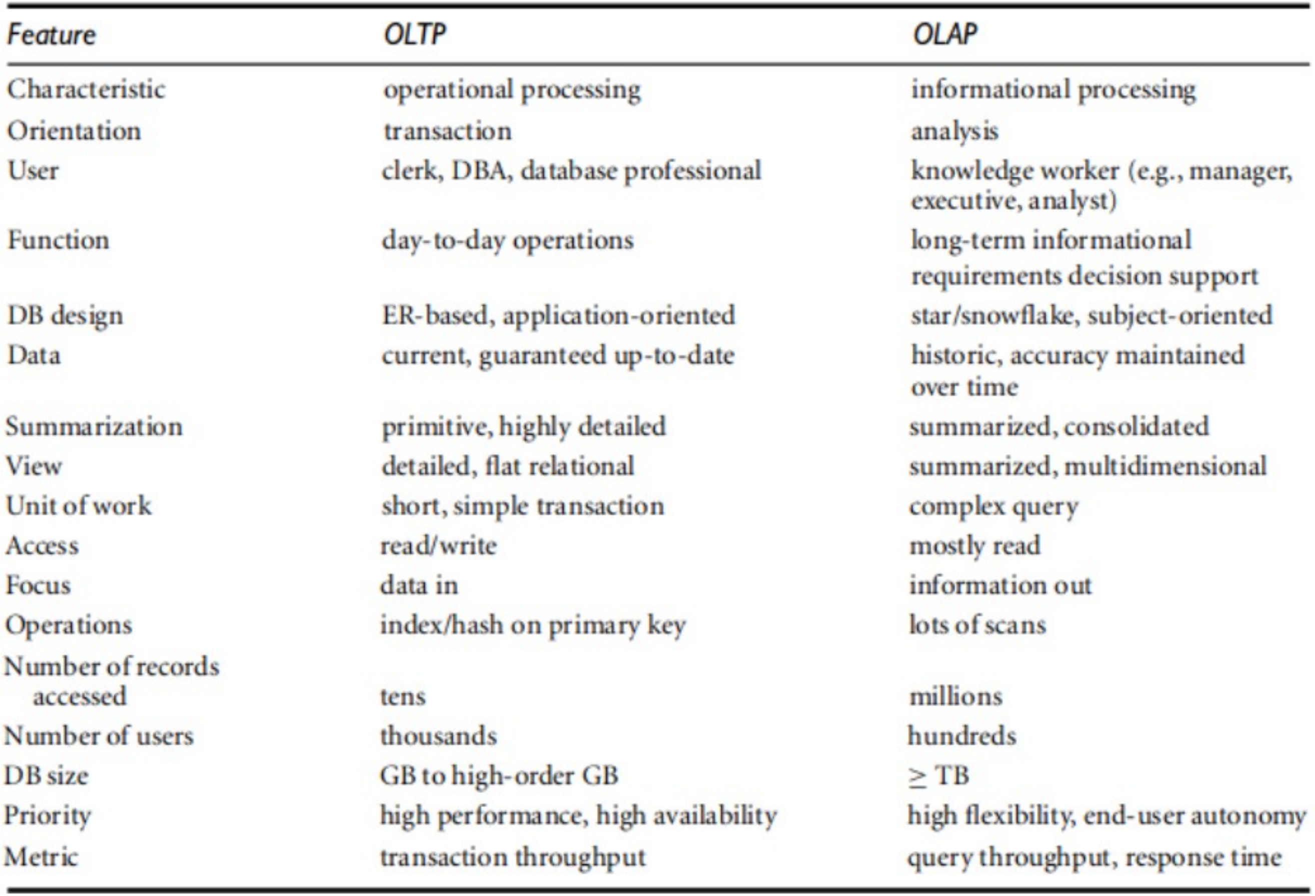
OLTP

- The major task of online operational database systems is to perform online transaction and query processing. These systems are called **online transaction processing** (OLTP) systems.
- They cover most of the day-to-day operations of an organization such as purchasing, inventory, manufacturing, banking, payroll, registration, and accounting.

OLAP

- Data warehouse systems, on the other hand, serve users or knowledge workers in the role of data analysis and decision making. Such systems can organize and present data in various formats in order to accommodate the diverse needs of different users. These systems are known as **online analytical processing** (OLAP) systems.

OLTP vs OLAP



Feature	*OLTP*	*OLAP*
Characteristic	operational processing	informational processing
Orientation	transaction	analysis
User	clerk, DBA, database professional	knowledge worker (e.g., manager, executive, analyst)
Function	day-to-day operations	long-term informational requirements decision support
DB design	ER-based, application-oriented	star/snowflake, subject-oriented
Data	current, guaranteed up-to-date	historic, accuracy maintained over time
Summarization	primitive, highly detailed	summarized, consolidated
View	detailed, flat relational	summarized, multidimensional
Unit of work	short, simple transaction	complex query
Access	read/write	mostly read
Focus	data in	information out
Operations	index/hash on primary key	lots of scans
Number of records accessed	tens	millions
Number of users	thousands	hundreds
DB size	GB to high-order GB	≥ TB
Priority	high performance, high availability	high flexibility, end-user autonomy
Metric	transaction throughput	query throughput, response time

Data Warehouse Models

- Enterprise Warehouse
- Data Marts
- Virtual Warehouse

Enterprise Warehouse

- Collects all of the information about subjects spanning the entire organization
- It provides corporate-wide data integration, usually from one or more operational systems or external information providers, and is cross-functional in scope
- It typically contains detailed data as well as summarized data
- Data range in size from a few gigabytes to hundreds of gigabytes, terabytes, or beyond
- An enterprise data warehouse may be implemented on traditional mainframes, computer super servers, or parallel architecture platforms
- It requires extensive business modeling and may take years to design and build

Data Marts

- A data mart contains a subset of corporate-wide data that is of value to a specific **group of users**.
- For example, a marketing data mart may confine its subjects to customer, item, and sales. The data contained in data marts tend to be summarized
- Depending on the source of data, data marts can be categorized as independent or dependent
- Independent data marts are sourced from data captured from one or more operational systems or external information providers, or from data generated locally within a particular department or geographic area
- Dependent data marts are sourced directly from enterprise data warehouses

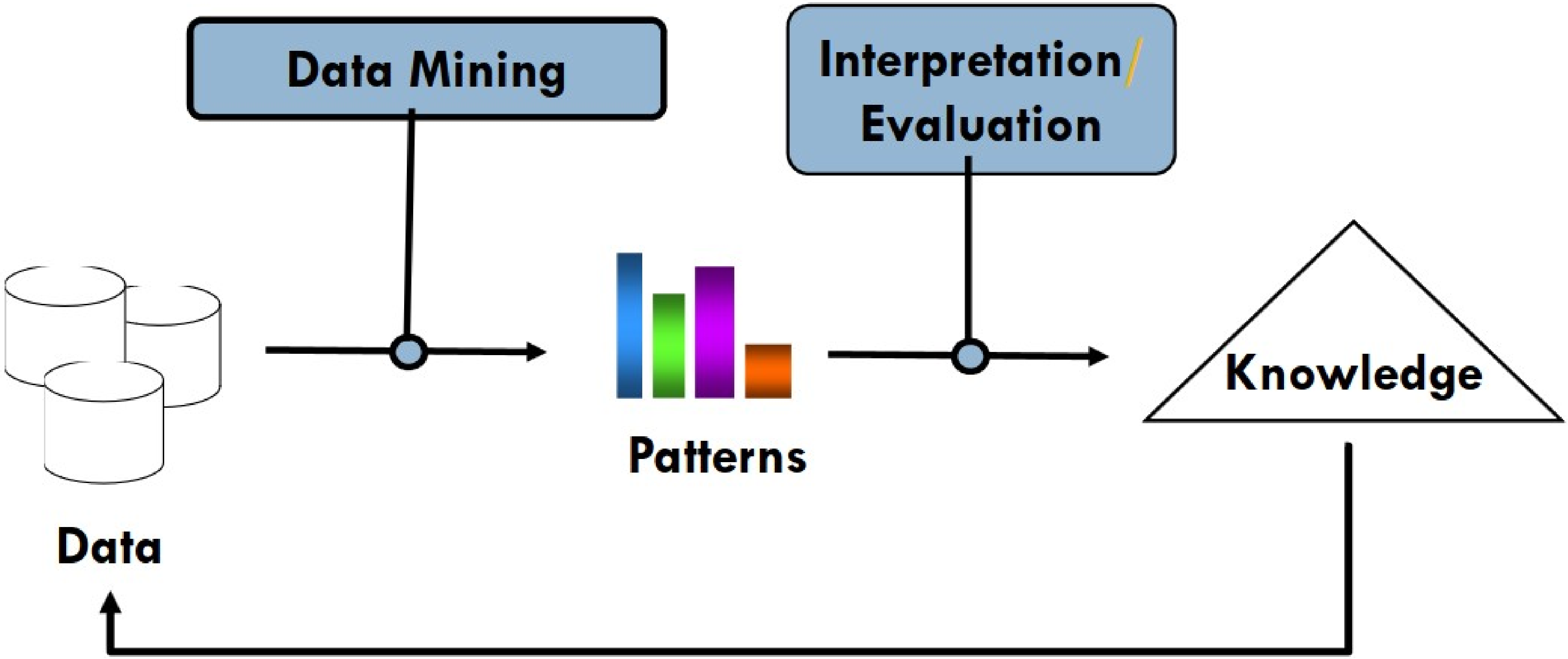
Virtual Warehouse

- **Virtual data warehouse** (VDW) is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized.
- A virtual warehouse is easy to build but requires excess capacity on operational database servers.
- The VDW acts as a logical view of the data, providing a unified view of multiple data sources without the need for physically storing the data in a single location

What is data mining?

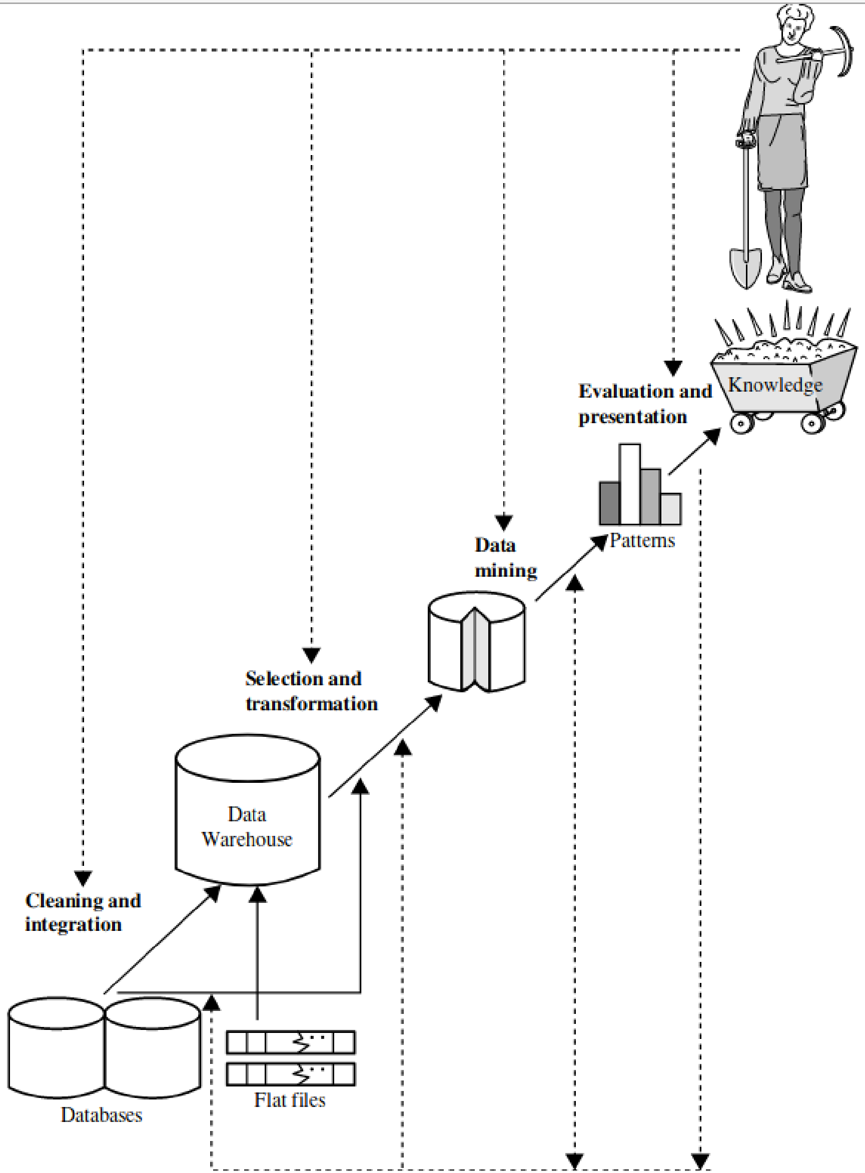
- Data mining is
  - Data mining is also called knowledge discovery in Data (KDD)
  - Extraction of useful patterns from data sources, e.g., databases, texts, web, image.
  - Patterns must be:
    - Valid
    - Novel
    - Potentially useful
    - Understandable

Knowledge Discovery in Data: Process



KDD – 7 steps

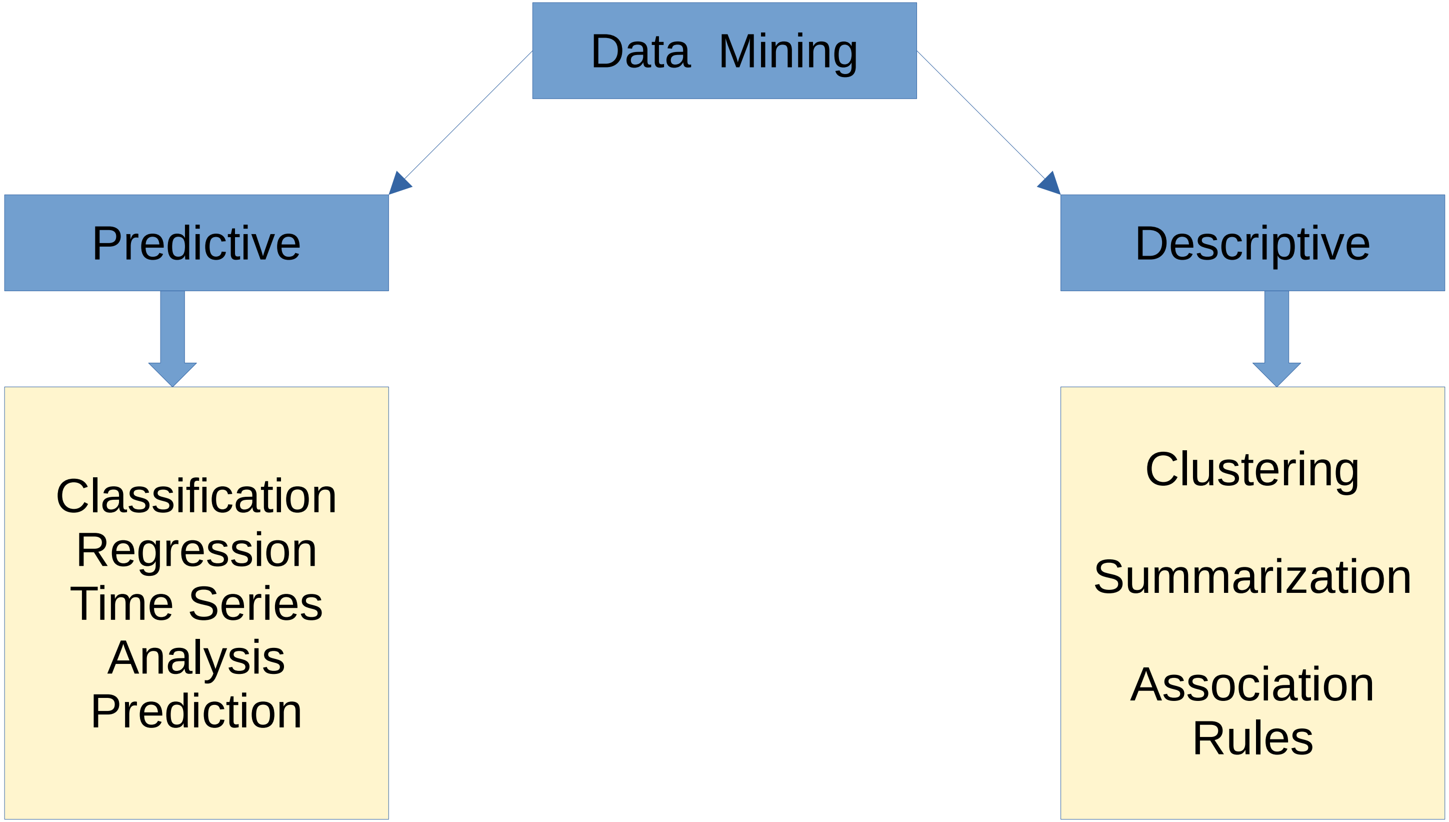
- Data Cleaning
- Data Integration
- Data Selection
- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge Representation



Data Mining Techniques

- The two "high-level" primary goals of data mining, in practice, are prediction and description.
  - **Prediction** involves using some variables or fields in the database to predict unknown or future values of other variables of interest.
  - **Description** focuses on finding human-interpretable patterns describing the data.
- Clustering
- Classification
- Association

Data Mining Techniques



Data Mining

Predictive

Descriptive

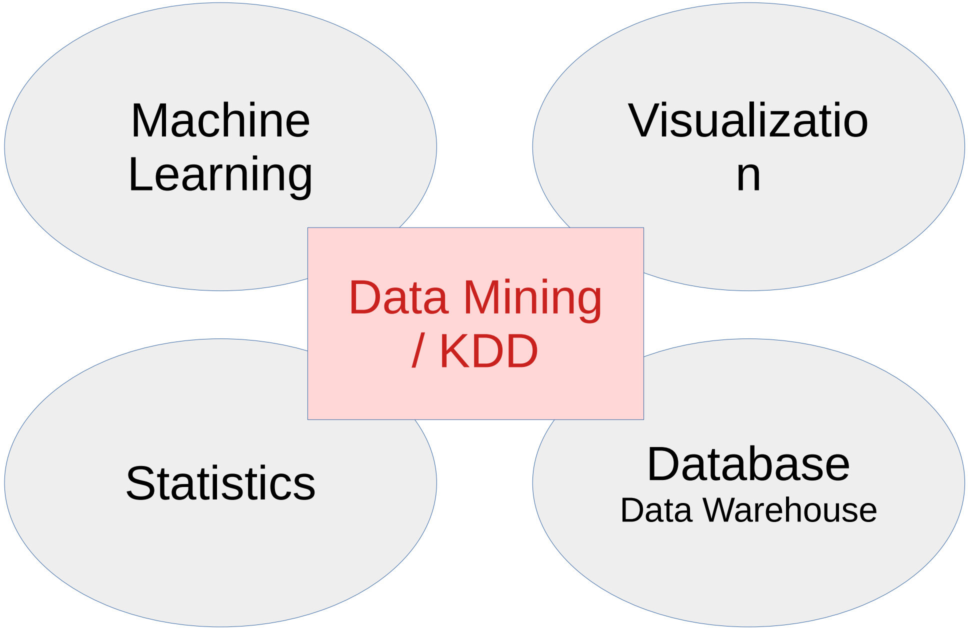
Classification
Regression
Time Series Analysis
Prediction

Clustering

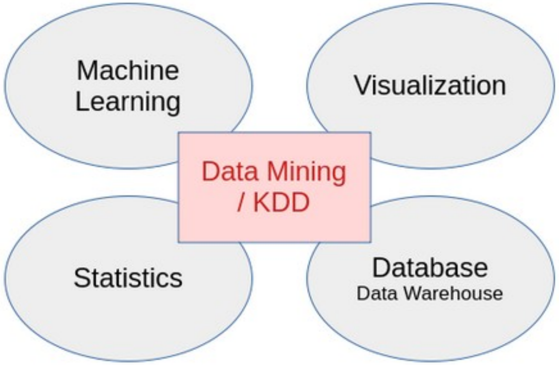
Summarization

Association Rules

Related Fields



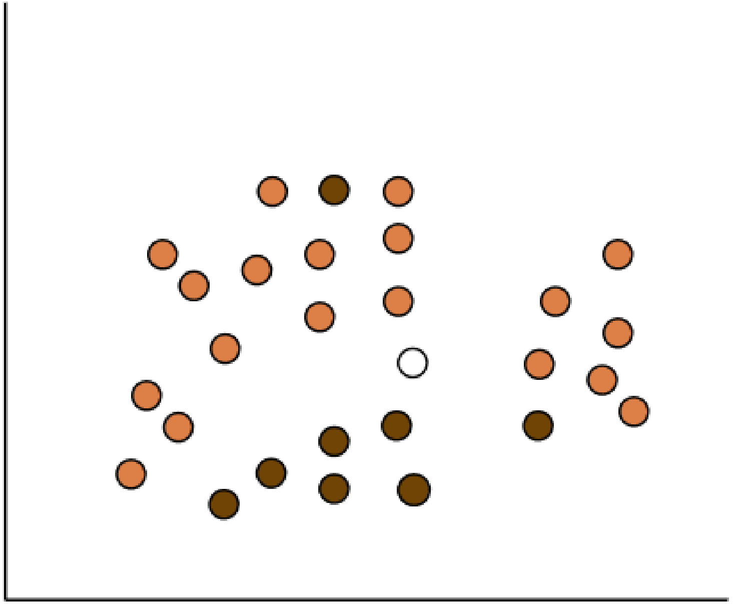
Related Fields



- Statistics
  - more theory-based
  - more focused on testing hypotheses
- Machine learning
  - more heuristic
  - focused on improving performance of a learning agent
  - also looks at real-time learning and robotics – areas not part of data mining
- Data Mining and Knowledge Discovery
  - integrates theory and heuristics
  - focus on the entire process of knowledge discovery, including data cleaning,learning, and integration and visualization of results

Classification

Learn a method for predicting the instance class from pre-labeled (classified) instances

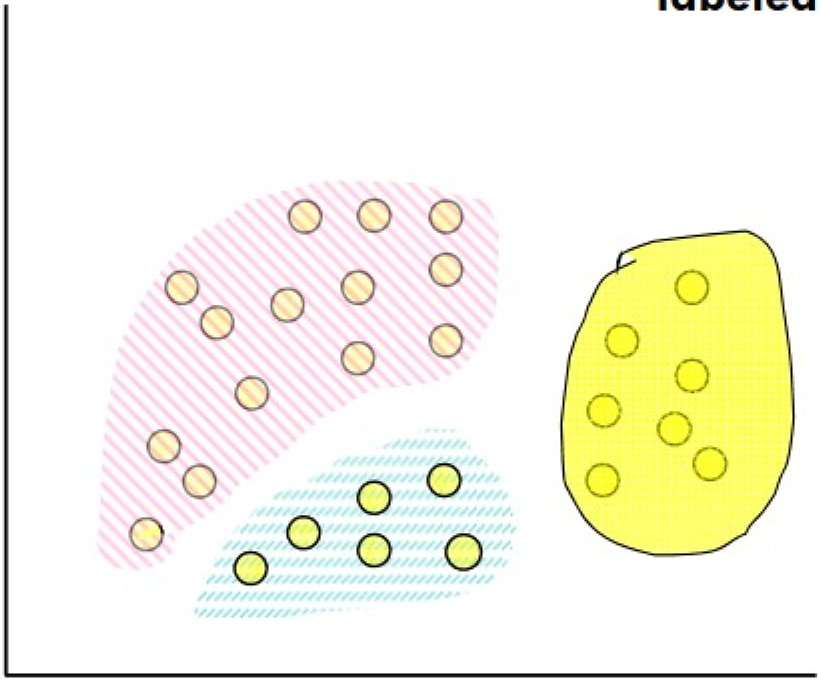


Many approaches: Statistics, Decision Trees, Neural Networks,

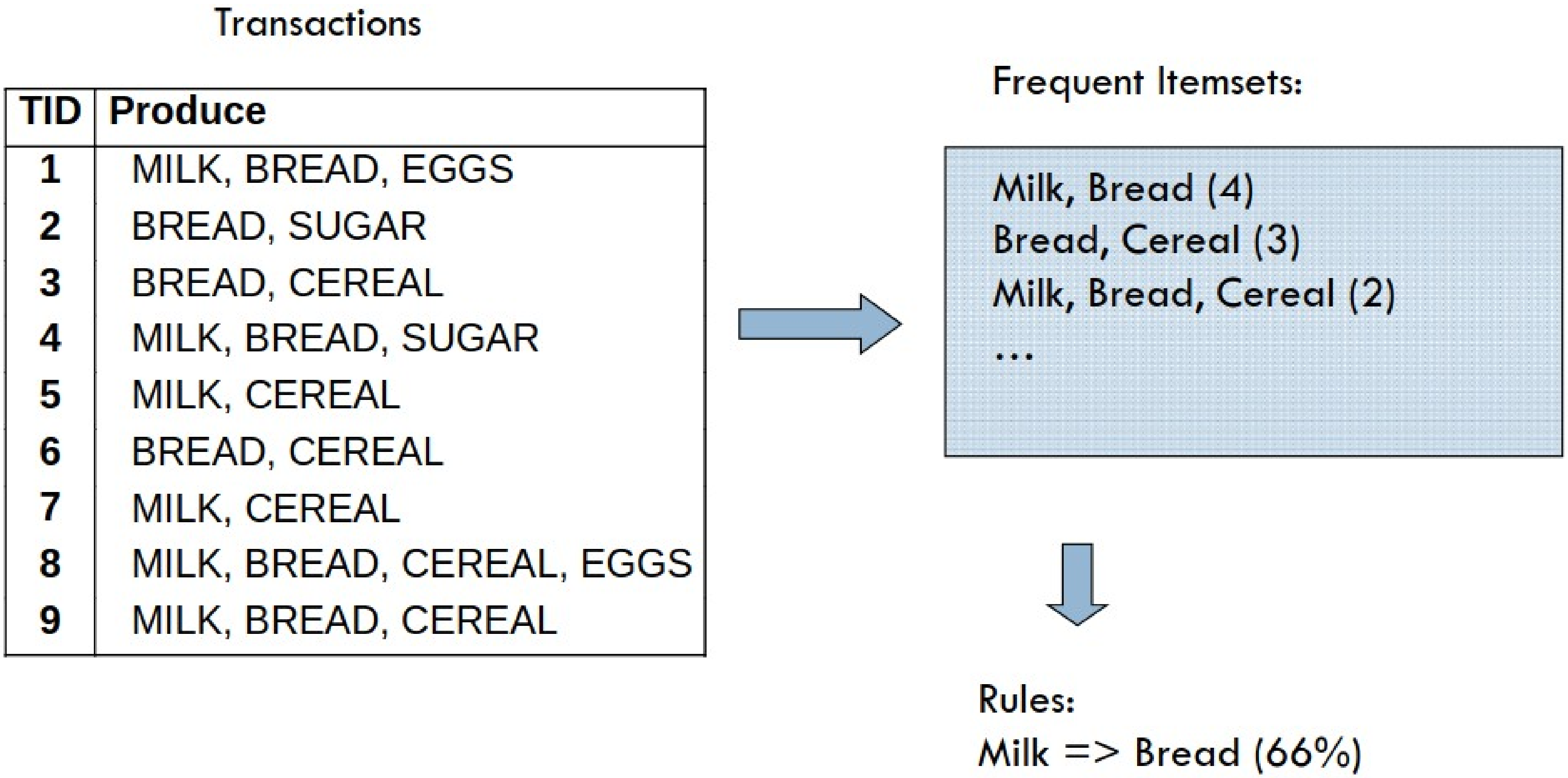
...

Clustering

Find “natural” grouping of instances given un-labeled data



Association Rules & Frequent Itemsets



Transactions

TID	Produce
1	MILK, BREAD, EGGS
2	BREAD, SUGAR
3	BREAD, CEREAL
4	MILK, BREAD, SUGAR
5	MILK, CEREAL
6	BREAD, CEREAL
7	MILK, CEREAL
8	MILK, BREAD, CEREAL, EGGS
9	MILK, BREAD, CEREAL

Frequent Itemsets:

Milk, Bread (4)
Bread, Cereal (3)
Milk, Bread, Cereal (2)
…

Rules:
Milk => Bread (66%)

Types of data

- Relational Data
- Graph Data
- Temporal Data
  - Time Series Data
  - Sequence Data
- Spatial Data
  - – location data , GPS, Coordinates, Map etc.
- Spatial-Temporal Data
  - Location with Time components
- Unstructured data
  - Text, review, comments etc.
- Semi-Structured Data
  - Published data, json, xml, html data etc.

Data Quality

- We need quality data for increasing the accuracy
- Pre-processing techniques can be used to enhance the data quality
  - Heterogeneous Data
    - Inconsistencies
    - Data format
  - Noise
  - Outliers
  - Redundancy

Data pre-processing Steps

- Exploratory Data Analysis
- Deal with Missing
- Deal with Duplicates and Outliers
- Encode Categorical Features
- Split dataset into training and test set
- Deal with Imbalanced Data

Measuring Similarity

- The data similarity can be computed using distance
  - Euclidean
  - Manhattan
  - Minkowski
  - Cosine
  - Pearson
  - Jaccard
  - Levenshtein
  - Hamming

Data Visualization

- Types of data visualization
  - Distribution plot
  - Box and Whisker Plot
  - Line Plot
  - Bar Plot
  - Scatter Plot
  - Histogram
  - Pie chart
  - Heatmap etc.

Thank you