# Syllabus

22 September 2022 07:05 PM

**UNIT-1** **Introduction to Data Mining**

Related Technologies-Machine Learning, DBMS, OLAP and Statistics, Data Mining Goals, Stages of Data Mining Process, Data Warehouse and Multidimensional Data Model, OLAP Operations.

**UNIT-2** **Data Preprocessing and Data Visualization**

Data Cleaning, Data Transformation, Data Reduction, Discretization and Concept hierarchy, Exploratory Data Analysis Tools(Plots, Graphs, Summary Statistics, histograms, heat maps)

**UNIT-3** **Data Mining algorithms:-Association Rules and Regression Analysis**

Item sets, Frequent Patterns, Interestingness measures(support, confidence and lift),Co-relation analysis,

Apriori and Frequent Growth Algorithms, Linear Regression and Models

10

**UNIT-4** **Data Mining Algorithms: Classification**

Decision Trees, Random Forests, Bayesian Networks, Nearest Neighbour Algorithms.

**UNIT-5** **Data Mining Algorithms:Clustering and Model Evaluation**:

Partitioning Methods:-K-means,k-Medeoids,ExpectationMaximization,HierarchicalMethods:Distance-Basedagglomertive clustering and divisible clustering,Training And Testing(Cross-Validsation),Combining multiple models(bagging,boosting)

Data mining is the extraction of intresting patterns or knowledge from a huge amount of data

Data mining can be applied to any kind of data as long as the data are meaningful for terget application

Stages of Data Mining
28 September 2022 05:08 AM

# Stages of Data mining

1. **Data Gathering** (Data fetching) : Relevant data for analytical application is assembled from various sources {mix of structured & unstructured data}

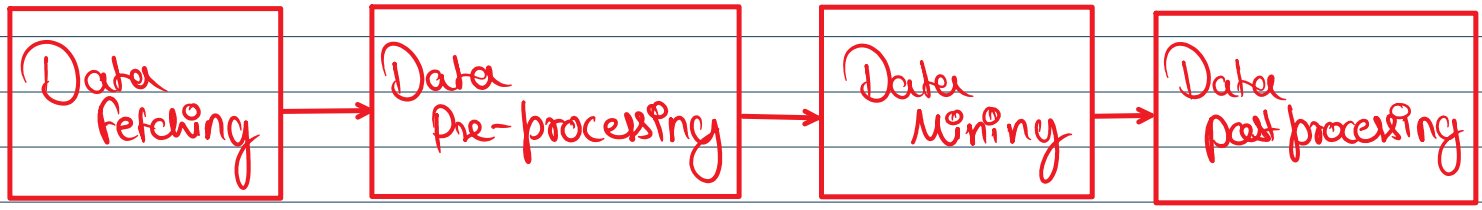
2. **Data preparation** (Data pre processing) ; This stage includes set of steps to ready the data for mining which includes;
   - data Exploration
   - profiling
   - pre processing
   - data cleansing
   - data transformation

3. **Mining of data** (Data mining) ; Prepared data is mined according to appropriate mining techniques chosen by data scientist

   Various techniques / algorithms are applied

4. **Data Analysis & Interpretation** (Data post processing)

   Mining results are used to create analytical models (which helps in business & decision making)

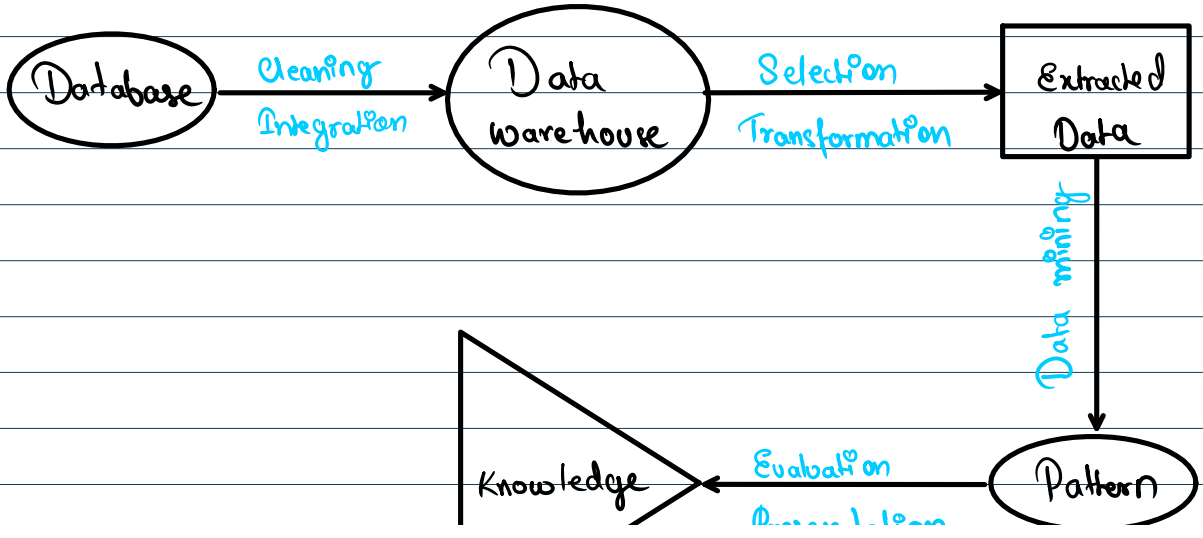

Data fetching
Data Pre-processing
Data Mining
Data post processing

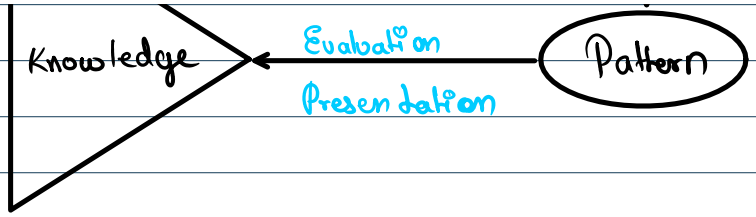
! IDK??

25 September 2022 12:32 PM

# Data mining includes following iterative steps

1. Data Cleaning (removal of noise & inconsistant data)
2. Data Integration (compilation & combination of data from multiple data sources)
3. Data Selection (selection of relevant data for analysis from DB)
4. Data transformation (Data is converted into forms appropriate for mining)
5. Data mining (essential process where intelligent methods are applied to extract data patterns)
6. Pattern Evaluation (Identification of pattern representing knowledge)
7. Knowledge presentation (Visualization of knowledge & their techniques.

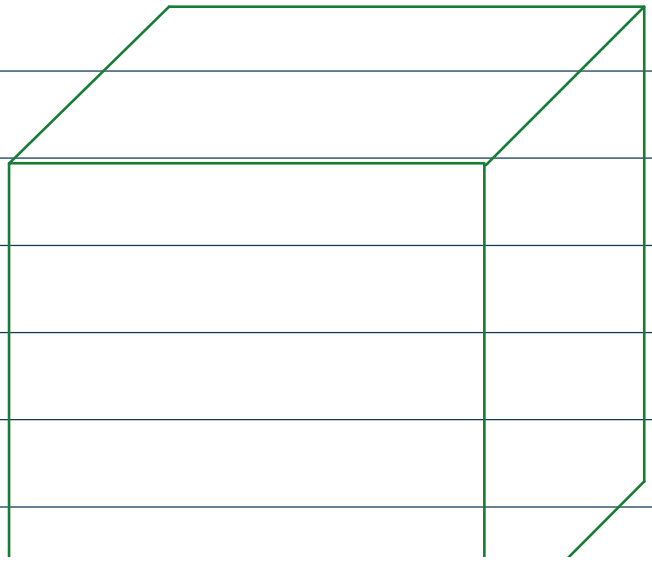



Knowledge
Evaluation
Presentation
Pattern

# Multi-Dimensional View of Data Mining

- **Data to be mined** ∩ Data
  - Database data (extended-relational, object-oriented, heterogeneous, legacy), data warehouse, transactional data, stream, spatiotemporal, time-series, sequence, text and web, multi-media, graphs & social and information networks
- **Knowledge to be mined (or: Data mining functions)** ∩ Knowledge
  - Characterization, discrimination, association, classification, clustering, trend/deviation, outlier analysis, etc.
  - Descriptive vs. predictive data mining
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized** ∩ Technology
  - Data-intensive, data warehouse (OLAP), machine learning, statistics, pattern recognition, visualization, high-performance, etc.
- **Applications adapted** ∩ Application
  - Retail, telecommunication, banking, fraud analysis, bio-data mining, stock market analysis, text mining, Web mining, etc.

15



!Data Warehousing

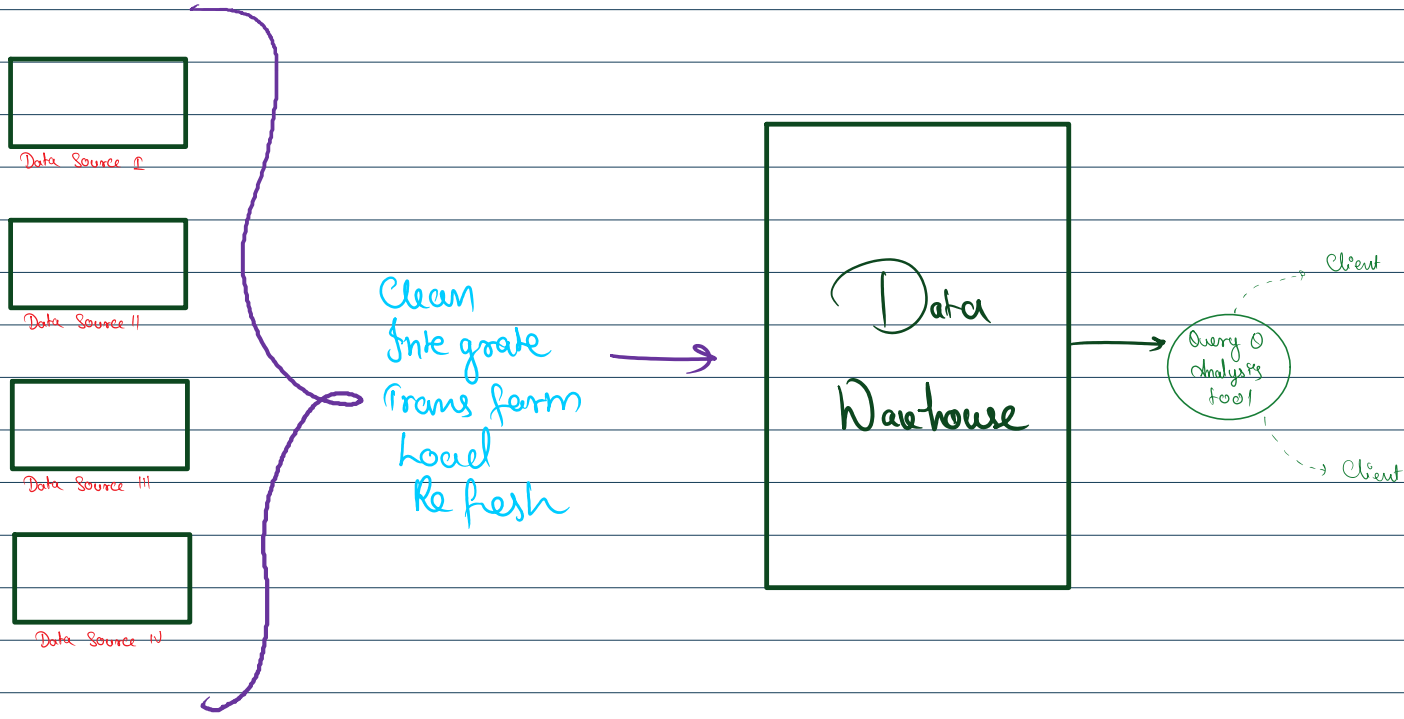
28 September 2022 05:33 AM

**Data Warehouse** : It is a repository for long-term storage of data from multiple sources

- The data is stored under a unified schema & typically summarized
- Data warehousing allows multidimensional data analysis capability

⇓

Online analytical process (OLAP)



OLAP vs OLTP

28 September 2022 06:56 AM

| OLAP | OLTP |
| --- | --- |
| ◦ Consists of long term historical data collected from various databases | ◦ Consists of short term, currently operational data & transactional data |
| ◦ Makes use of data warehouse | ◦ Makes use of Standard DBMS |
| ◦ The data is used in long term planning & decision making | ◦ Data is used to perform day to day fundamental operation |
| ◦ Astronomical size of data | ◦ Relatively small size of data |
| ◦ Not often updated | ◦ data integrity & files are constantly being updated |
| ◦ Managed by CEO, GM | ◦ Managed by clerks, Manages |
| ◦ Only read operation is required | ◦ Both read write required |
| ◦ Snow flake or Star Schema | ◦ ER diagram |

?Attribute

28 September 2022 07:35 AM

# Type of Attributes

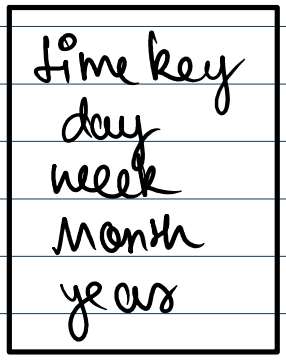
1. Nominal attr.
2. Binary
3. Ordinal
4. Numeric

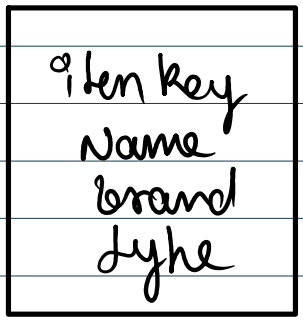
# !Types of Data Sets

28 September 2022 07:14 AM

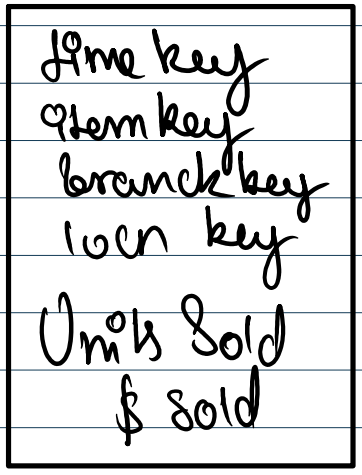
# Schema Diagram

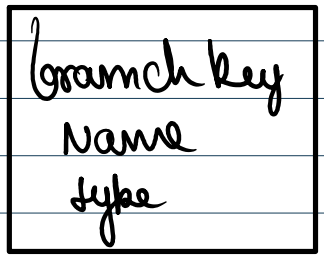
## Star Schema

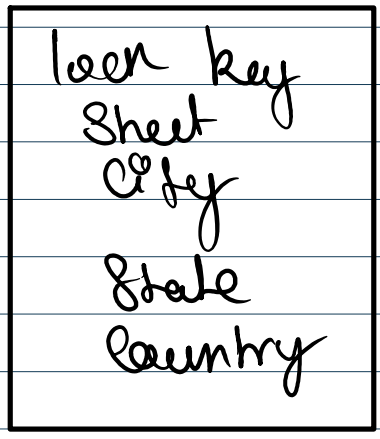



fact table




Other than the fact table rest are called the dimension table

If any dimension table were to be normalized then

# Advantages of Decision trees

1. It doesn't require any domain knowledge
2. Classification steps of decision tree are simple & fast
3. Missing values in data doesn't affect output
4. A decision tree model is automatic & doesn't need Standardization of data