KDD: (Knowledge discovery in databases)

→ It is a process that involves the extraction of useful, previously unknown, and potentially valuable information from large datasets.

→ steps involved in KDD:

① selection – Relevant subset of the data for Analysis

② Preprocessing – clean and transform the data. it is ready for analysis
Tasks are data normalization, missing value handling, & data Integration

③ Transformation

④ Data Mining

⑤ Interpretation

⑥ Evaluation

⑦ Deployment

Transformation

Transform the data into suitable format
e.g Graph

Data Analytics

→ Apply DA techniques and alg to extract useful info & insights.

→ clustering, classification, Association rule mining

Interpretation

→ Interpret the results and extract knowledge from the data.

→ Tasks included are

Visualizing the results

Evaluating the quality of the discovered patterns

identifying relationships association among the data.

Evaluation

→ Ensure that the extracted knowledge is useful, accurate and Meaningful.

Deployment:

→ use the discovered knowledge to solve the business problem and make decision

Why we need DA:

→ better decision making, bec'z day to day data are increasing.

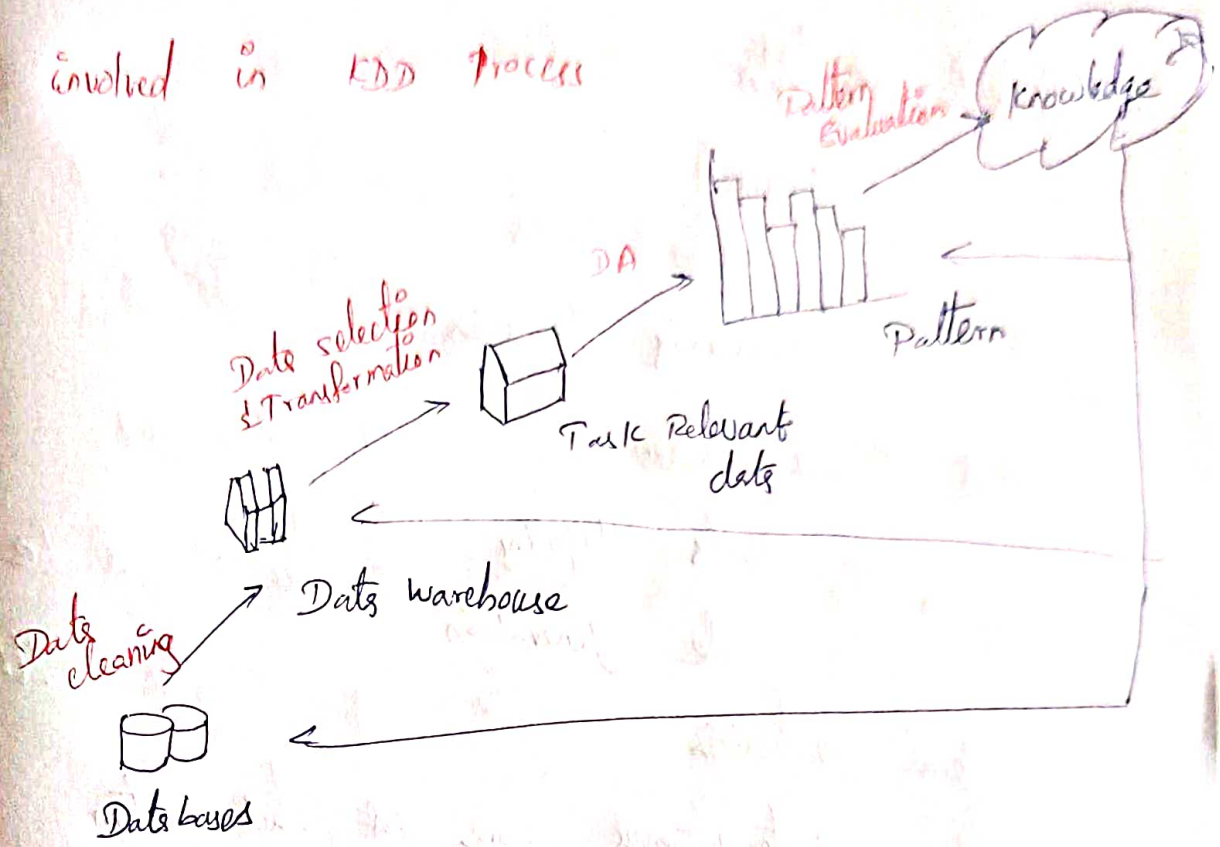
Why DA used in Business:

→ Automatically summarization of data

→ Extracting essence of info stored

→ Discovering patterns in raw data.

Steps involved in KDD Process



Data cleaning:

→ Remove Noisy data & irrelevant data from collection.

① cleaning in case of Missing values

② " Noisy data, where noise is a random or variance error.

③ cleaning with data discrepancy detection and data transformation tools

Data Integration.

→ defined as heterogeneous data from multiple sources combined in a common source.

① Data Migration Tools.

② " synchronization "

③ Extract Load transformation.

Data selection:

Neural N/w

Decision Trees

clustering, Regression

Data Transformation:

Data Mapping

Code generation

Pattern Evaluation

Interestingness score of each pattern

Summarization & Visualization

Knowledge Representation

Generate Reports

Tables

classification rules.

→ KDD is an iterative process

→ Preprocessing of db consists of Data cleaning and Data integration

Adv:

Improve DM.

Increased efficiency

Better customer service

Fraud detection

Predictive modeling

Disadv:

Privacy Concerns

Complexity

Data quality

High Cost

overfitting

KDD used?

→ used for Machine learning, db, AS Pattern Matching & enterprise

→ KDD is the systematic process of identifying valid, practical and understandable patterns in massive and complicated data sets.