Data Mining and Analytics

Information - collection of data

Data :- Unprocessed things / raw data.

Mining :- Extraction of information.

Knowledge :- Useful information converting from the raw data.

Pattern :- Crenerate patterns / insights from available data.

Analysis :-

Market Basket Analysis ::

⇒

Extract useful & meaningful information / patterns from the data.

clustering ⇒ grouping, No class / label, group based on similarity.

classification ⇒ labeling, putting under a label or class.

Note :-

classification is based on the label.

Clustering → classification.

Prediction :- (V-2)

Predicting with the already existing data

eg :- weather prediction.
Stock market
Health care - Cancer prediction
House rent prediction.

Business sector
Education

Association Rule Mining :- (V-2)

If - then rule.

If (18 < age-group < 22) {

then college student.

Pre-processing :-

⇒ Removal of unwanted data = Data cleaning

⇒ Data transformation

⇒ Data Integration

⇒ Data reduction

DBMS :-

Quering ⇒ Fetching

Database vs Data warehouse

Database :-

⇒ Small amount of data (transactional data)

⇒ Online Transaction processing

OLTP

Data warehouse :-

⇒ Large amount of data

⇒ can't update, but more data can be included.

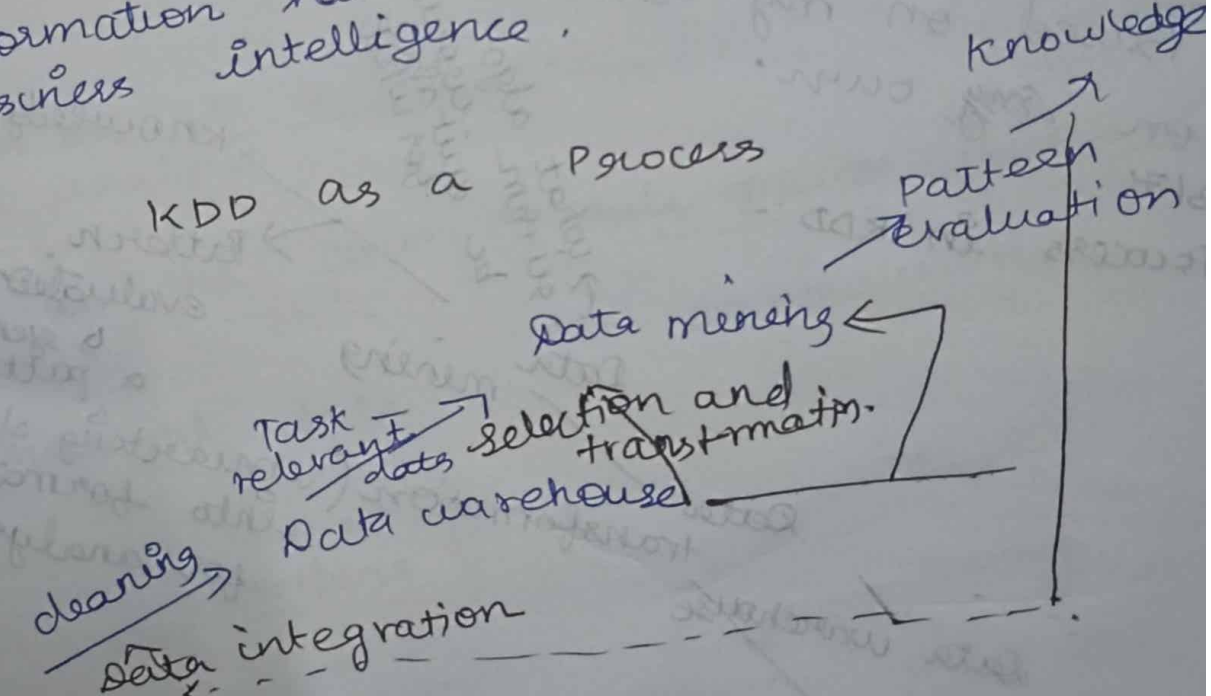
⇒ OLAP ⇒ Online Analytical processing

Machine learning :-

Data mining :-
Extracting previously unknown and potentially useful patterns from large amounts of data.

Knowledge Discovery in DataBase (KDD) :-
knowledge extraction
data / pattern analysis
data archeology
data dredging
information harvesting
business intelligence.

KDD as a Process



Data cleaning :-
1. Dealing with missing values.

Data lake ⇒ source of data
Large amt, structured / unstructured,
place where raw data is stored.

Unstructured data :-

Video / audio files.

Semi - structured.

Excel + video files.

Supervised :- Labelled data

eg :- Classification

Unsupervised :- Un - labelled data

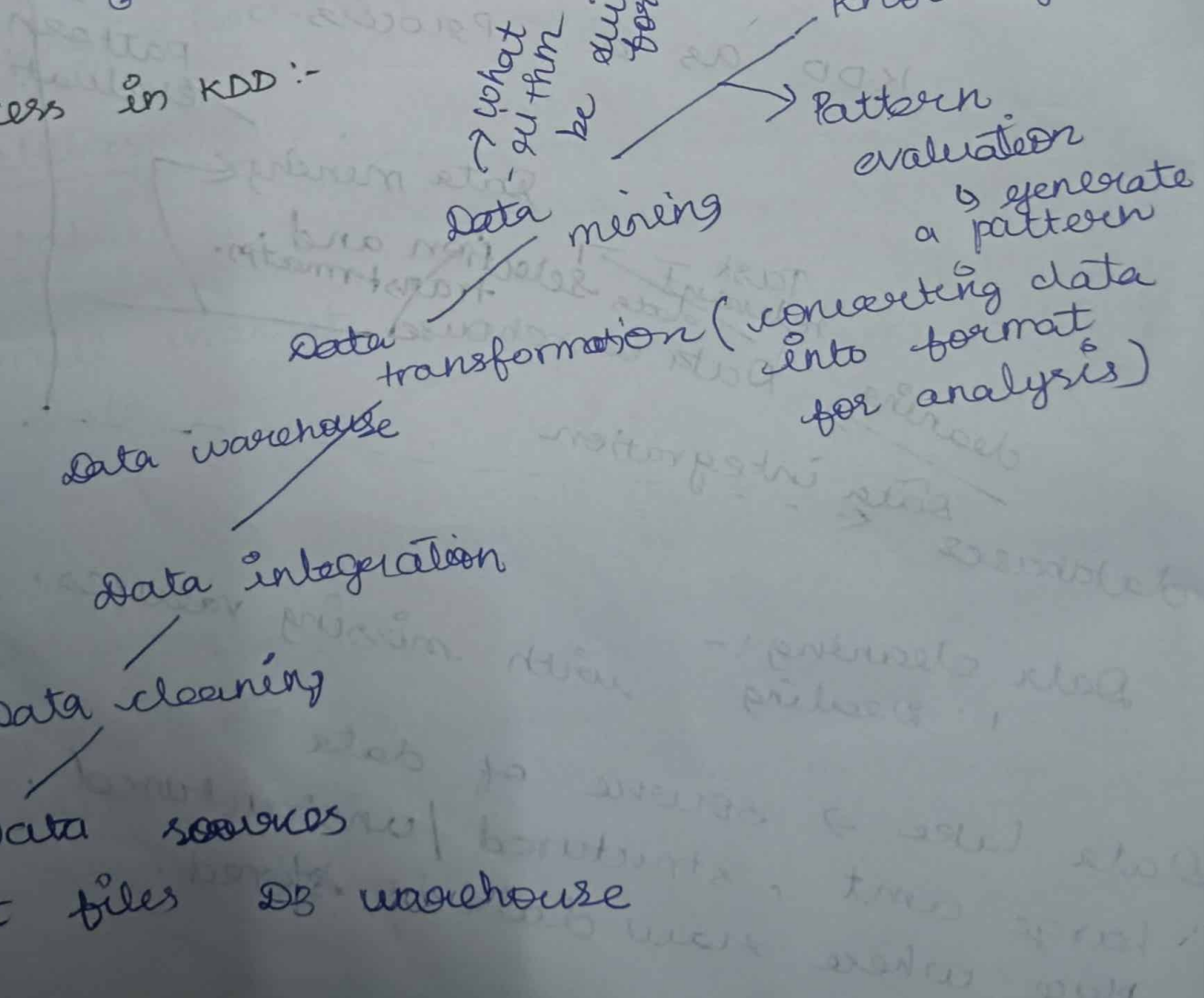
eg :- Clustering

Re-inforcement learning :-

Based on my interest I'm learning it on ~~my~~ own.

13/7

Process in KDD :-



Extraction of Knowledge ⇒ Data Mining.

Technologies related to DM :-

* DBMS
* ML
* OLAP
* statistics

DBMS ⇒ software that's used to manage the data

DB ⇒ storage space of the data.

tuples ⇒ row column ⇒ attributes.

Note :-
DM is used when it involves more than 2 or 3 join queries.

ML :-
used in DM for better accuracy and prediction
Outlier analysis

OLAP :-
Data in the form of data cubes.
↓
3D structure

3-dimensional data.

OLAP operations

- Roll up
  - ↓
  - Lower version of data converted to higher eg:- city → country
- Drill down
  - ↓
  - opposite of Roll up
  - eg:- year use ↓ month wise
- Slice
  - ↓
- Dice
  - ↓
  - Taking away
  - ↓
  - dice from the entire cube
- pivot (or) rotate
  - ↓
  - rotating the entire cubes.