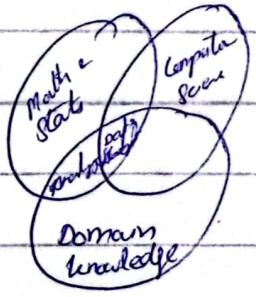
Lecture 15: Data Science

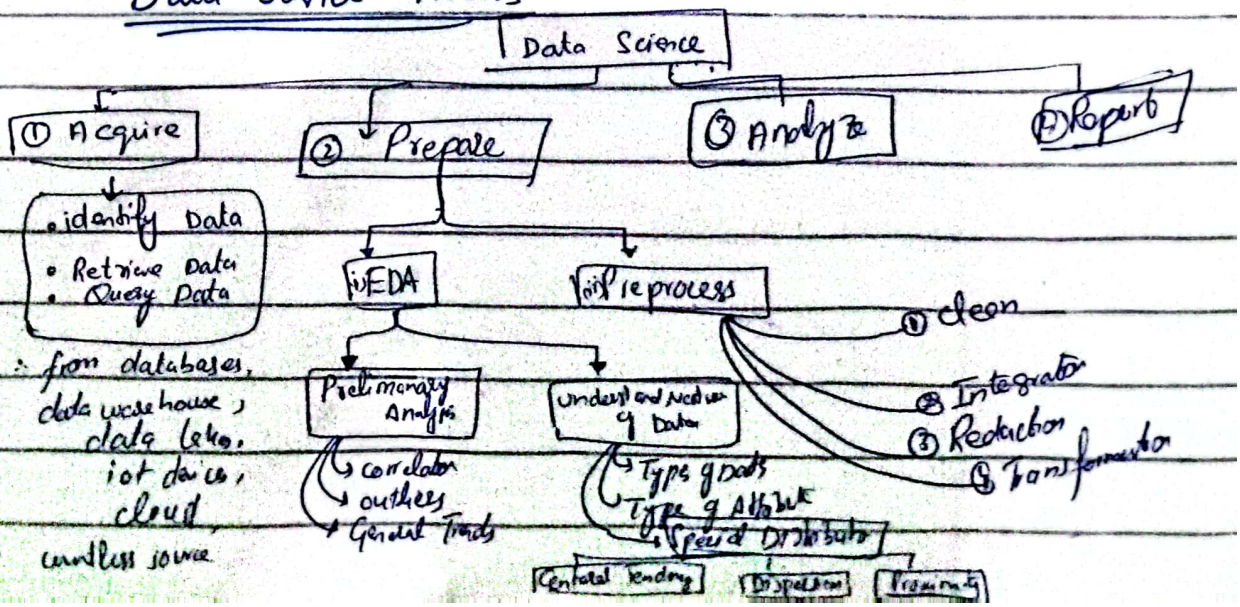
- multi-disciplinary data (statistics, computer science)
- domain knowledge where data is coming
- Extracting insights & knowledge from Data
- Today, data processing is backbone of all. (i.e. like social media → user preferences)
- Data is like "Gold Mine"

Data Science

" is the field of extracting knowledge and insights from data through a blend of statistics, programming, machine learning & domain expertise.



Data Science Process



- Types of Attribute ??
- what type of data attributes have ??
- Statistical Distribution

Structured Data

↳ organized & follow defined format

Attribute Types usually stored in databases / spreadsheets

→ Nominal data :- (Fi) / Categorical (no order)

→ Binary → (with two categories only)

↳ Ordinal (order ترتيب) :- smoker / non-smoker ✓ / ✗

like T-shirt size (S, M, L)

grades (A, B, C)

→ Numerical (Quantitative and measurable)

↳ can perform operation

With absolute zero (like experience) | No-absolute zero (like temperature)

discrete → absolute like 2, 51 Zero,

continuous → within range of values

Statistical distributs

① Central Tendency (مرکزی رجحان)

درمیانی و بلیو کیطرف رجحان

Three important measures

→ mean (average)

→ median (middle value of sorted data)

→ Mode (most frequent value)

② Measure of Dispersion (بکھراؤ)

(Raw data is spread from middle value)

important measure

(1) standard deviation & variance (if we use mean)

(2) inter quartile Range (IQR)

(if we use median)

③ Measure of proximity (قریب قریب)

(on a/more attributes usually)

(Raw values are close to each other)

important measures

① dot product

② Cosine Similarity

③ Co-relation

Interest → Relationship Strength (strong/weak)

→ Relationship direction (same/opposite)

Data Distribution → Symmetric

→ A-Symmetric (skewed) → Positively skewed

→ Negatively skewed

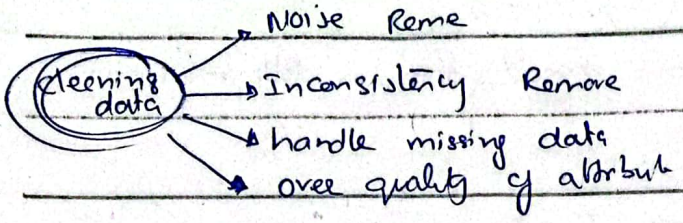
② Prepare Data

(ii) Pre-processing

① clean (To have quality data)

(accurate, consistent, no missing value, operate easily)

(manage)



② Data integration

(Techniques do merging data that are come from multiple sources)

③ Data Reduction

(i) (ii)

(duplicate colums, colums with similar information

↳ Remove → duplicate colums
→ duplicating rows
→ similar colums

Dimensionality Reduction

- To reduce similar data
- drop not relevant data

Technique ⇒ ① PCA (Principal Component Analysis)

↓ based on (i) eigen-decomposition,
(ii) singular value decomposition

② P-SNE

✓ Low quality data removal

④ Data Transformation