### Introduction:

* 1. DBMS= [Advanced Database Systems + Advanced Data Analysis].
  2. Steps:
     1. Data cleaning (to remove noise and inconsistent data).
     2. Data integration (where multiple data sources may be combined)
     3. Data selection (where data relevant to the analysis task are retrieved from the database)
     4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
     5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
     6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on interestingness measures—see Section 1.4.6)
     7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users).
  3. Data Types:
     1. Relational Data
     2. Data Warehouse(collect from different places to ease query).
        1. Data cube used--each dimension is an attribute or a set of attributes.
     3. Transactional data
     4. Other:
        1. Sequence / Time related.
        2. Data streams(sensors etc)
        3. Spatial data(maps)
        4. Multimedia
        5. Web
        6. Temporal data(state in time eg land-use/rainfall etc).
  4. Patterns Types that can be mined:
     1. Descriptive
        1. Classification
        2. Clustering
        3. Outlier analysis
        4. Characterisation
           1. Pie charts etc.
     2. Predictive
        1. Pattern mining, Association, Correlation
           1. Frequent itemsets
           2. Frequent sequences
           3. Frequent substructure(like tree etc)
           4. Support[this much % transactions confirm items occurrence together], confidence[chance that they will occur together in real scenario/probability].
           5. Single dimensional rule:

Having one predicate/attribute:

buys(X,”computer”)=>buys(X,”keyboard”)

* + - * 1. Multidimensional Rule:

Having multiple predicates(3 here):

age(X,49) && income(X,40-45K)=>buys(X,”computer”)

* + 1. Classification & Regression for predictive analysis
       1. Classification:
          1. Decision Trees
          2. Neural Network
          3. Naive Bayes Classification
          4. SVM
          5. kNN
          6. Classification predicts categorical(discrete, unordered) labels.
          7. Regression predicts numerical(continuous, ordered) data /value to predict.
    2. Cluster Analysis
       1. maximizing the intraclass similarity and minimizing the inter-class similarity
       2. Taxonomy Formation
       3. After clustering, we can mine the rules
    3. Outlier Analysis/ Anomaly Mining
       1. Statistical methods
       2. Distance methods
       3. Density based methods
    4. Are patterns interesting?
       1. NO
       2. Objective interestingness measures
          1. Support
          2. Confidence
          3. Accuracy
       3. Subjective interestingness measures
          1. If patterns are unexpected or actionable
       4. They are important to increase efficiency of pattern discovery/clustering
       5. Research areas:
          1. Can it generate only interesting patterns(bewakoofi hai ye!)
          2. Can it generate all interesting patterns.(Optimization problem in data mining)
  1. Technologies used:
     1. Statistics
     2. Machine learning
        1. Supervised----classification
        2. Unsupervised-----clustering
        3. Semi-supervised------some are labeled some are not so the ones not labeled are used to refine the decision boundary.
     3. DBMS systems(query languages, query processing and optimization methods, data storage, and indexing and accessing methods), Data warehouses
     4. Information retrieval(science of searching for documents or information in documents)
  2. Applications:
     1. Business intelligence
     2. Web search engines
  3. Data Mining Issues:
     1. Mining method
        1. Multiple things like kind of data, pattern evaluation, pattern or constraint guided evaluation, handling noise,uncertainity etc.
     2. User Interaction
        1. Interactive mining
        2. Incorporation of background knowledge
        3. Ad hoc data mining and data mining query languages
        4. Presentation and visualization of data mining results
     3. Efficiency & Scalability
     4. Diversity of data types
     5. Mining & Society