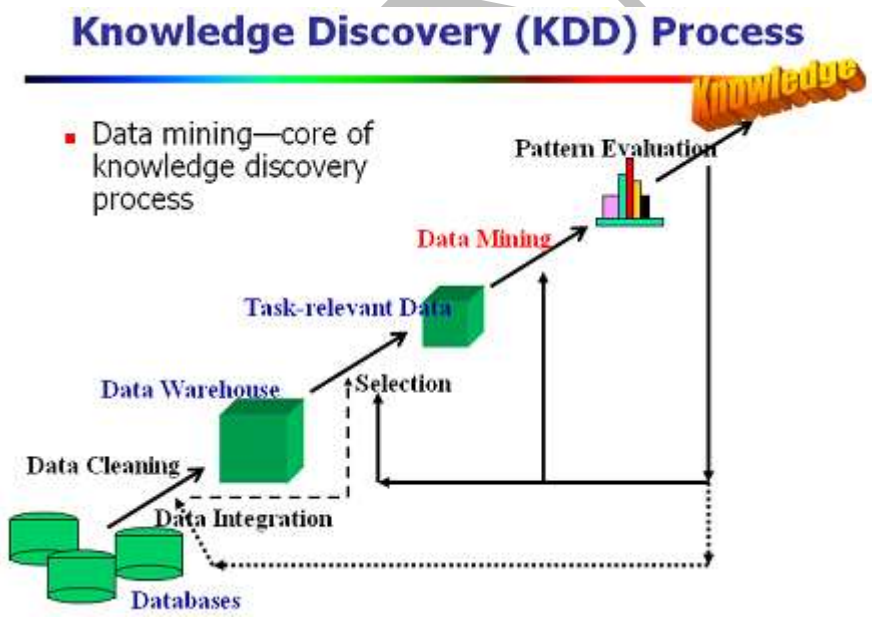
## 1.1 Relation to Statistics

**What motivated data mining? Why it is important?**

* Huge Volume of data
* Major Sources of Abundant data:
  - Business – Web, E-commerce, Transactions, Stocks
  - Science – Remote Sensing, Bio informatics, Scientific Simulation
  - Society and Everyone – News, Digital Cameras, You Tube
* Need for turning data into knowledge – Drowning in data, but starving for knowledge
* Applications that use data mining:
  - Market Analysis - Fraud Detection - Customer Retention
  - Production Control - Scientific Exploration

**What is Data Mining?**

- Extracting and 'Mining' knowledge from large amounts of data.
- "Gold Mining from rock or sand" is same as "Knowledge mining from data"
- Other terms for Data Mining:
  - Knowledge Mining
  - Knowledge Extraction
  - Pattern Analysis
  - Data Archeology
  - Data Dredging
- Data Mining is not same as KDD (Knowledge Discovery from Data)
- Data Mining is a step in KDD

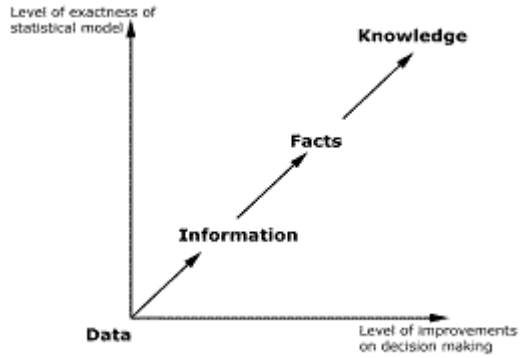



Data Cleaning – Remove noisy and inconsistent data
Data Integration – Multiple data sources combined
Data Selection – Data relevant to analysis retrieved
Data Transformation – Transform into form suitable for Data Mining (Summarized / Aggregated)
Data Mining – Extract data patterns using intelligent methods
Pattern Evaluation – Identify interesting patterns

Knowledge Presentation – Visualization / Knowledge Representation
– Presenting mined knowledge to the user

**Relation to Statistics:**

- Statistics – "Learning from Data" or "Turning data into information".
- Data – Crude Information – Does not makes sense – What we capture & store
- e.g. customer data, store data, demographical data, geographical data
- Information – relates items of data – relevant to the decision problem
- e.g. X lives in Z; S is Y years old; X and S moved; W has money in Z
- Facts – Information becomes facts when data can support it
- Knowledge – What we know or infer – relates items of information
- e.g. a quantity Q of product A is used in region Z; customers of class L use N% of C in period D




- Data Mining interface between statistics, computer science, artificial intelligence, machine learning, database management, data visualization,…
- "*Data mining is the application of statistics to reveal patterns and trends in very large data sets*."
- Data mining can learn from Statistics. Statistics is fundamental to data mining.
- Data mining will not become knowledge discovery without statistical thinking.
- Statistics will not be able to succeed on massive and complex datasets without data mining.

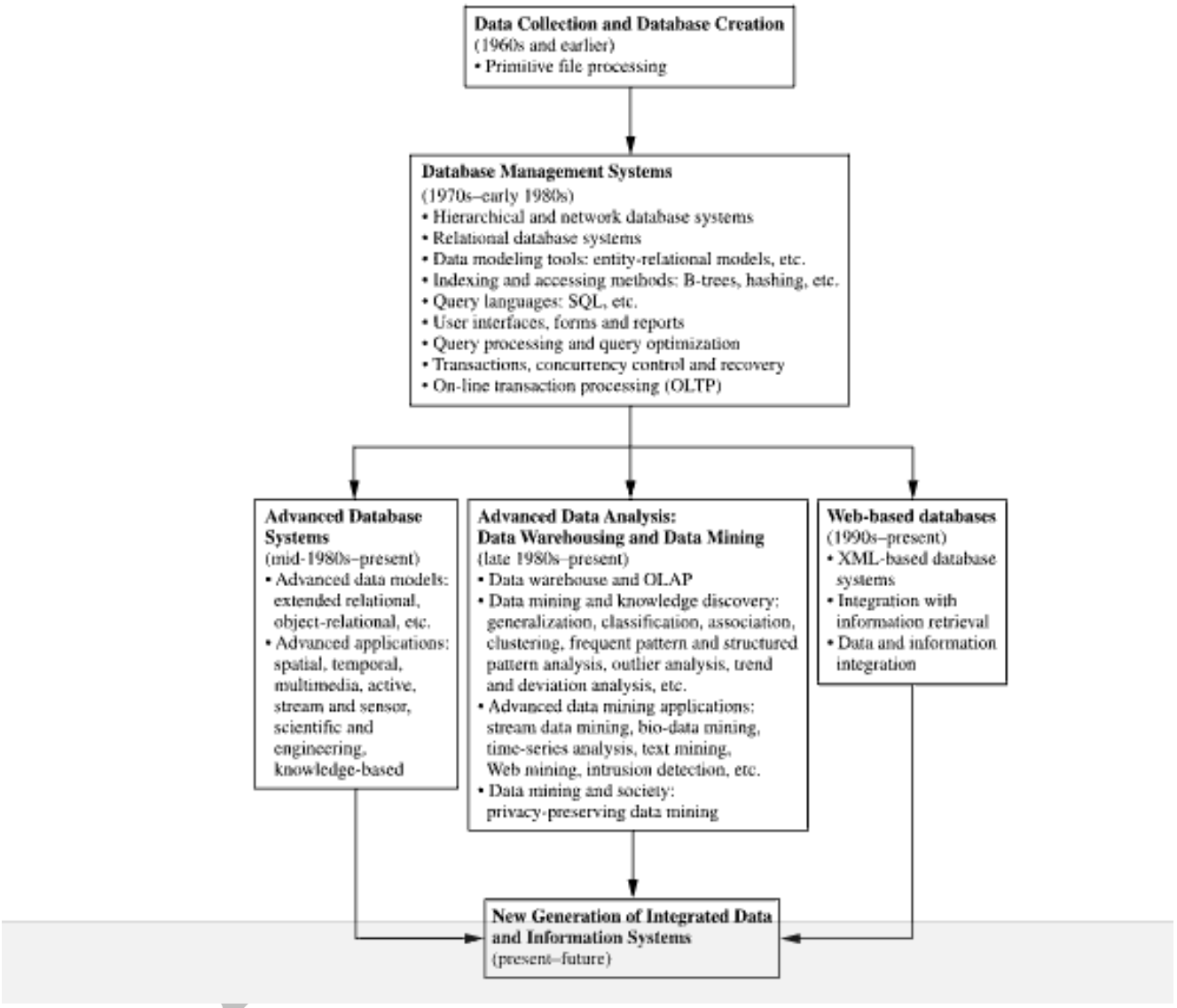
## 1.2 Databases

**Database System – Evolutionary Path:**

- Progress in Hardware technology -> Led to powerful, affordable computers
- New data repository architecture -> Led to Data warehouses (multiple heterogeneous data sources in single schema)
    - Warehouses facilitates management decision making
    - Warehouse includes data cleaning, data integration, OLAP (Online Analytical Processing)
        - OLAP consists of Summarization, Consolidation, Aggregation, Different angle view / Multidimensional Analysis for decision making
- However, in-depth analysis requires additional data analysis tools
- Data rich and information poor situation

KLNCIT

- Expert systems – rely on domain experts for decision making - using their knowledge intuition
    - Time consuming, costly, error prone, biased
- So the solution is to use Data Mining tools – performs data analysis, finds data patterns
    - Contributes to business strategies, knowledge bases, scientific & medical research




**Data Mining – Confluence of Multiple Disciplines**

1. Databases 2. Data Warehousing 3. Statistics 4. Machine Learning
5. Information Retrieval 6. Image and Signal Processing 7. Pattern Recognition
8. Neural Networks 9. Data Visualization 10. Spatial / Temporal Data Analysis

**Data Mining – On What Kinds of Data?**

- Database-oriented data sets and applications
    - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
    - Data streams and sensor data
    - Time-series data, temporal data, sequence data (incl. bio-sequences)
    - Structure data, graphs, social networks and multi-linked data

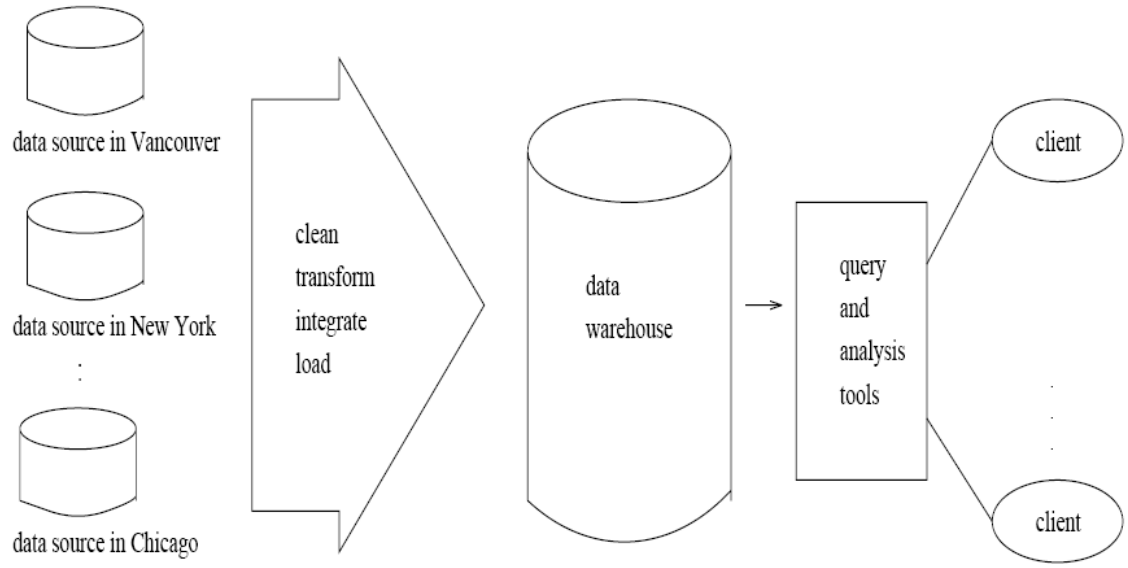
- Object-relational databases o Heterogeneous databases and legacy databases
- Spatial data and spatiotemporal data o Multimedia database
- Text databases o The World-Wide Web

**Relational Databases:**

- Consists of Database (inter related data) and set of software programs to manage and access data.
- Collection of tables
- Each table has a set of attributes (columns / fields) and large set of tuples (records or rows)
- Unique key – describes a tuple.
- Data Model used is ER Data Model – Set of entities and its relationships
- Accessed by Database queries - written in SQL / using GUI
- Query – transformed to join, selection and projection operations
- Query Optimized for efficient processing
- SQL includes aggregate functions (Group by) sum, avg, count, max, min
- Apply data mining on databases
  - Searches for trends or data patterns
    - Predict credit risk of new customers based on income, age & prev. credit info
  - Detects deviations
    - Sales of particular items – sales < expected comparison with previous year
  - Deviations further investigated for reason
    - Increase in price, change in packing

**Data Warehouses:**

Data spread in several databases – physically located at numerous sites
Data warehouse – repository of multiple DBs in single schema; resides at single site.
Data warehousing processes
1. Data Cleaning 2. Data Integration 3. Data Transformation
4. Data Loading 5. Periodic data refreshing




- Data in a data warehouse are organized around major subjects

- Data provide information on historical perspective – summarized on periodic dimension
- Eg. Sales of an item for a region in a period
- Data warehouse model – multidimensional database structure / data cube
- Dimensions – Attributes / set of attributes
- Facts – Aggregated measures (Count / Sales amount)

**Data Mart & Data Warehouse – Difference**

- Data Mart – Department subset of data warehouse
- Data Warehouse – Enterprise wide scope, suited for OLAP
- DWH - Presents data at different levels of abstraction; accommodates different user views
- OLAP Operations – Drill Down (Data at Month Level) & Roll Up (Data at Country Level)
- If in-depth analysis required – use Data Mining tolls and techniques

**Transactional Databases**

- Consists of a file with records where each record is a transaction
- Each transaction has a unique transaction ID and list of items that make up transactions
- Transactional database may have additional tables associated with it.
- These tables contain other reference information regarding sales, date of transactions, customer ID, Sales Person Id, Branch of Sales.
- Eg. "Show all items purchased by John"; "How many transactions include item number I3"
- These queries requires full scan of transactional database.
- But deeper analysis required in real time.
- Eg. "Which items sold well together" → Market Basket Data Analysis → Groups / bundles items together as a strategy for maximizing sales.
- Eg. "Printers and Computers sold together" → Printers can be given at discount rate
- Such queries not answered by transactional database
- So apply Data Mining techniques to identify frequent item set patterns.

**Advanced Data and Information Systems and Advanced Applications**

- Spatial Data (maps) o WWW (Internet)
- Engineering Data (Building design, circuit design, system components)
- Hyper text and Multi Media Data (text, image, video, audio)
- Time-related data (Historical records or stock exchange data)
- Stream Data (Video, Surveillance, Sensor – data flows in & out)
- All these advanced data types requires complex DB Schema structure with dynamic changes
- Hence we have advanced & application oriented DB systems:
    - Object-relational DB systems
    - Temporal and Time-series DB systems
    - Heterogeneous and Legacy DB systems
    - Data Stream Management Systems
    - Web-Based information systems
- These raise many challenging research and implementation issues for data mining

KLNCIT

**Object-Relational Databases**

- Extended relational model on handling complex objects – Popular in industry applications
- Each object includes:
  - Attributes / variables o Messages – Communicates with other objects
  - Methods – Code to implement a message – receives a value in return
    - Eg. Message get_photo(employee) → returns photo of employee object
- Object Class – Objects with common properties
- Each object = instance of a class
- Class → Subclass ; Employee Class → Sales-person Sub class
- Inheritance – subclass inherits the properties of its class + additional properties specific to the subclass (Eg. Commission – property specific to sub-class)
- Data Mining technique to be developed for handling complex object structures, class, subclass, inheritance etc.

**Temporal Databases, Sequence Databases and Time-Series Databases**

Temporal Databases

- Stores time related attributes & has time stamps with different semantics.

Sequence Databases

- Sequences of ordered events with or without time.
- Eg. Customer Buying Sequence; Web Logs; Biological sequences

Time-Series Databases

- Sequences of ordered events over time.
- Eg. Stock Exchange, Inventory Control, Temperature, Wind,…
- Data Mining techniques used to find trend of changes of objects in such databases.
- Used in decision making and strategy planning
- Based on multiple granularity of time
- Eg. Banking Data Mining → Customer traffic prediction
- Eg. Stock Exchange Data Mining → Investment Planning Strategy

**Spatial Databases and Spatiotemporal Databases:**

Spatial Databases:

- Spatial related information
- Eg. Geographic (map) dbs, medical image dbs, Satellite image dbs
- Spatial data represented in Raser Format → in n-dimensional bit maps
- Eg. 2D Satellite image – represented in Raser Data – each pixel represents rainfall in a given area
- Another representation is that Maps are represented in Vector Format → Roads, Bridges, Buildings and Lakes represented as points or lines or polygons
- Eg. Applications → Forestry, Ecology Planning
- What kind of data mining can be applied on Spatial Databases?
  - "Houses located near a park"
  - "Climate of hill areas at different altitudes"
  - "Poverty rate based on city distances from major highways"
  - "Spatial Data Cubes" can be constructed with multi dimensional hierarchies
    - – drill-down & roll-up

Spatio-temporal Databases:



KLNCIT

- Database with Spatial objects that change with time.
- Eg. Outbreak of flu based on geographic location with respect to time.

**Text Databases and Multimedia Databases:**

Text Databases:

- Database consists of Word descriptions, keywords, sentences, paragraphs.
- Eg. Product description, bug report, warning messages, summary reports, notes, documents.
- Text databases can be Unstructured / Semi-structured (E-mail messages, HTML/XML Web Pages) / Well structured (library databases – can use relational databases)
- What can Data Mining on text databases uncover?
    - Key word and content association
    - Clusters within text objects
    - Data Mining & Information Retrieval techniques can be integrated for better results
    - Uses hierarchies such as dictionaries and thesaurus

Multimedia Databases:

- Stores images, audio, video data. O Also called as Continuous Media Data
- Applications – picture retrieval system, voice-mail system, video on demand systems, WWW, Speech based user interfaces
- Storage and search techniques can be integrated with data mining methods for efficiency
- Can construct multimedia data cubes for similarity based pattern matching

**Heterogeneous Databases and Legacy Databases:**

Heterogeneous Databases:

- Database has autonomous components that are interconnected where components communicate.
- Objects in different components differ – hence difficult to understand semantics.

Legacy Databases:

- Has long history of information. O Different hardware and operating systems.
- Group of heterogeneous databases connected by intra or inter-computer networks
- Information exchange across such databases is difficult because of diverse semantics.
- Data mining techniques provide solution by performing statistical data distribution and correlation analysis.

**Data Streams:**

- Data Flows in & out of the platform (or window) dynamically.
- Unique Features:
    - Huge volume o Dynamically changing o Flows in & out in a fixed order
    - Small number of scans o Demanding fast response time
    - Eg. Power supply, network traffic, telecommunications, web logs, video surveillance
    - Ongoing research on data stream management systems.
    - In these systems we use Continuous Query Model that has pre-defined queries
    - Multidimensional, on-line analysis and mining can be performed on stream data

KLNCIT

**World Wide Web:**

- Online Information Services
- Eg. Yahoo, Google, Wikepedia
- Data objects are linked together to facilitate interactive access.
- Users traverse from one object to another via links
- Such Data allows more challenging opportunity for Data Mining
- Understanding user access patterns allows better web design and better marketing strategy.
- This is called as Web Usage Mining or Web Log Mining
- Web pages are highly unstructured
- Web Page Analysis – Ranks web pages – helps in easy retrieval of relevant content pages when key word search is made.
- Web Page Clustering and Classification
- Web Community Analysis – hidden behavior of groups of web pages users.

## 1.3 Data Mining Functionalities

**Broad view of Data Mining Functionality:**

Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses or other information repositories.

**Data Mining Functionalities – What kinds of Patterns can be mined?**

- Two Categories – (i) Descriptive (ii) Predictive
- Descriptive – describes the general properties of the data in the database
- Predictive – Makes Predictions from the data
- Data Mining Functionalities should allow:
  - Mining of multiple kinds of patterns
    - that accommodates different user expectations and applications
  - Discover patterns at different levels of granularity
  - Hints / Specifications / Queries to focus the search for interesting patterns
- Each discovered pattern is measured for its "trustworthiness" based on data in the database.

- **1) Characterization:**
  - Concept description / Class description
  - Data is associated with concept or class
  - Eg. Classes of items for sale – (i) Computers (ii) Printers
  - Eg. Concepts of customers – (i) Big Spenders (ii) Budget Spenders
  - Class / Concept descriptions can be delivered via:
    - (1) Data Characterization
    - (2) Data Discrimination
    - (3) Data Characterization and Data Discrimination
  - Data Characterization
    - Summarization of general characteristics or features of a target class of data
    - Data specific to a class are collected by query
    - Types of data summarization:

- Summarization based on simple statistical measures
- Data summarization along a dimension – user controlled – OLAP rollup
- Attribute oriented induction – without user interaction.

- Output of data characterization can be represented in different forms:
  - Pie Charts, Bar Charts, Curves
  - Multidimensional Data cubes, Multidimensional tables
  - Generalized relations – in rule form – called as "Characteristics Rules"
- Eg. "Find Summarization of characteristics of customers who spend more than Rs. 50000 in shop S1 in a year"
- Result = "Customers are 40–50 years old, employed and have high credit rating"
- Users can drill down on any dimension – Eg. "Occupation of customers"

- Data Discrimination
  - Comparison of general features of target class data objects with the general features of objects from one or a set of contrasting classes.
  - Target and Contrasting classes are specified by the users
  - Data objects retrieved through database queries
  - Eg. "Users wants to compare general features of S/W products whose sales increased by 10% in the last year with those whose sales decreased by 30% during the same period."
  - Output of data discrimination is same as output of data characterization.
  - Rule form is called as "Discriminate Rules".
  - Eg. Compare two groups of customers.
    - Group1 – Shops frequently – at least 2 times a month
      **Vs**
    - Group 2 – Shops rarely – less than 3 times a year
  - Result = "80% of frequent shopping customers are between 20-40 years old & have university education." & "60% of infrequent shopping customers are seniors or youths with no university degree."
  - Users can drill down on income level dimension for better discriminative features between the two classes of customers.

- **2) Mining Frequent Patterns, Associations and Correlation:**
  - Frequent Patterns – Patterns that occur frequently in data.
  - Many kinds of Frequent Patterns exists:
    - (1) Itemsets (2) Subsequences (3) Substructures
    - Frequent Itemsets: (Simple)
      - Set if items that frequently appear together in a database.
      - Eg. Bread & Jam
    - Frequent Subsequences: (Advanced)
      - Frequent sequential patterns
      - Eg. Purchase PC → Purchase Digital Camera → Memory Card
    - Frequent Structured Patterns: (Advanced)
      - Structural forms that occur frequently
      - Structural forms – Graphs, Trees, Lattices

KLNCIT

- Result = Discovery of interesting associations and correlations within data.

- Eg. Association Analysis:
  - Example 1: "Find which items are frequently purchased together in the same transactions".
  - Buys (X. "Computer") => Buys (X, "Software") [Support = 1%, Confidence = 50%]
  - X is a variable representing customers
  - Confidence = % of chance that a customer buying computer buys a software
  - Support = % of transactions in the whole database that showed computers and software's were purchased together.
  - This association rule has a single repeated predicate "Buys"
  - Such association rules are called "Single Dimensional Association Rules"
  - Example 2:
  - Age (X, "20…29") ^ Income (X, "20K…29K") => Buys (X, "CD Player") [Support = 2%, Confidence = 60%]
  - Association Rule = "2% of total customers in the database are between 20-29 years of age and with income Rs.20000 to Rs.29000 and have purchased CD player." & "There is 60% probability that a customer in this age group and income group will purchase a CD player"
  - This is an association between more than one predicate (ie. Age, Income and Buys)
  - This is called as "Multidimensional Association Rule".
  - Association rules that do not satisfy minimum support threshold and minimum confidence threshold are discarded.

- **3) Classification and Prediction:**
  - Classification:
    - Process of finding a model that describes data classes or concepts
    - Based on a set of training data
    - This model can be represented in different forms
      - Classification Rules
      - Decision Trees
      - Mathematical Formulae
      - Neural Networks
    - Decision Trees
      - Flowchart like tree structure
      - Each Node = Test on the attribute value
      - Each Branch = Outcome of the test
      - Tree Leaves = Classes or class distributions
      - Decision trees can be converted into classification rules
    - Neural Networks
      - Collection of neuron-like processing units + weighted connections between the units.
    - Other methods of Classification
      - Naïve Bayesian Classification
      - Support Vector Machines

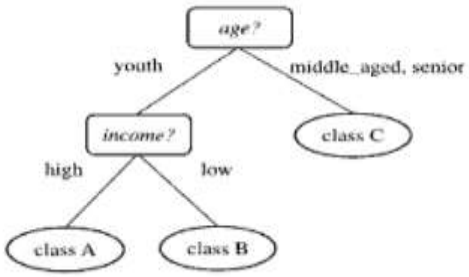
KLNCIT

- K-nearest neighbor Classification

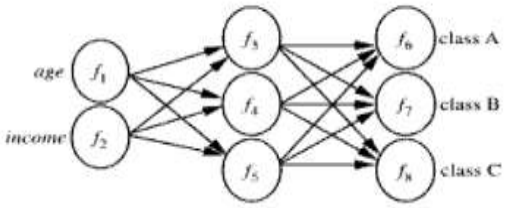
- Classification is used to predict missing or unavailable numeric data values => Prediction.
- Regression Analysis:- is a statistical methodology used for numeric prediction.
- Prediction also includes distribution trends based on the available data.
- Classification and Prediction may be used to be preceded by Relevance Analysis.
- Relevance Analysis:- attempts to identify attributes that do not contribute to the classification or prediction process which can be excluded.
- Example – Classification and Prediction:
  - **1) IF-THEN rules – Classification Model:**

age(X, "youth") AND income(X, "high") ⟶ class(X, "A")
age(X, "youth") AND income(X, "low") ⟶ class(X, "B")
age(X, "middle_aged") ⟶ class(X, "C")
age(X, "senior") ⟶ class(X, "C")

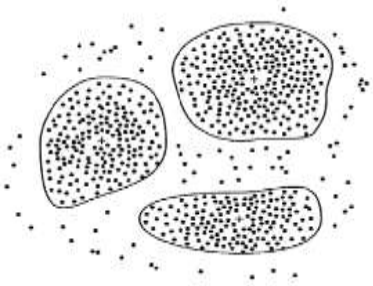
  - **2) A Decision Tree – Classification Model:**




  - **3) A Neural Network Classification Model:**




- **4) Cluster Analysis:**
  - Analyzes data objects without consulting a known class label.
  - The objects are clustered or grouped based on the principles of "Maximizing the intra-class similarity" and "Minimizing the inter-class similarity".
  - Objects within a cluster have high similarity compared to objects in other clusters.
  - Each cluster formed is a class of objects.
  - From this class of objects rules can be derived.
  - Clustering allows "Taxonomy Formation" → Hierarchy of classes that groups similar events together.
  - Eg. Customers with respect to customer locations in a city.



- 3 Data Clusters; Cluster center marked with a '+'

- **5) Outlier Analysis:**
  - Data that do not comply with general behavior of data are called as Outliers.
  - Most Data Mining methods discard outliers as noise or exceptions.
  - Some applications like fraud detection, rare events can be interesting than regular ones.
  - Analysis of such outliers is called as Outlier Analysis / Outlier Mining.
  - Outliers detected using:
    - Statistical Methods
    - Distance Measures
    - Deviation Based Methods
      - Difference in characteristics of an object in a group
  - Example – Outlier Analysis:
    - Fraudulent usage of credit cards by detecting purchase of extremely large amount for a given credit card account compared to its general charges incurred.
    - Same applies for Type of purchase, Place of purchase, Frequency of purchase.

- **6) Evolution Analysis:**
  - Describes the trends of data whose behavior change over time.
  - This step includes:
    - Characterization & Discrimination
    - Association & Correlation Analysis
    - Classification & Prediction
    - Clustering of time-related data
    - Time-series data analysis
    - Sequence or periodicity pattern matching
    - Similarity based data analysis
  - Example – Evolution Analysis:
    - Stock exchange data for past several years available.
    - You want to invest in TATA Steel Corp.
    - Data mining study / Evolution analysis on previous stock exchange data can help prediction of future trends in stock exchange prices.

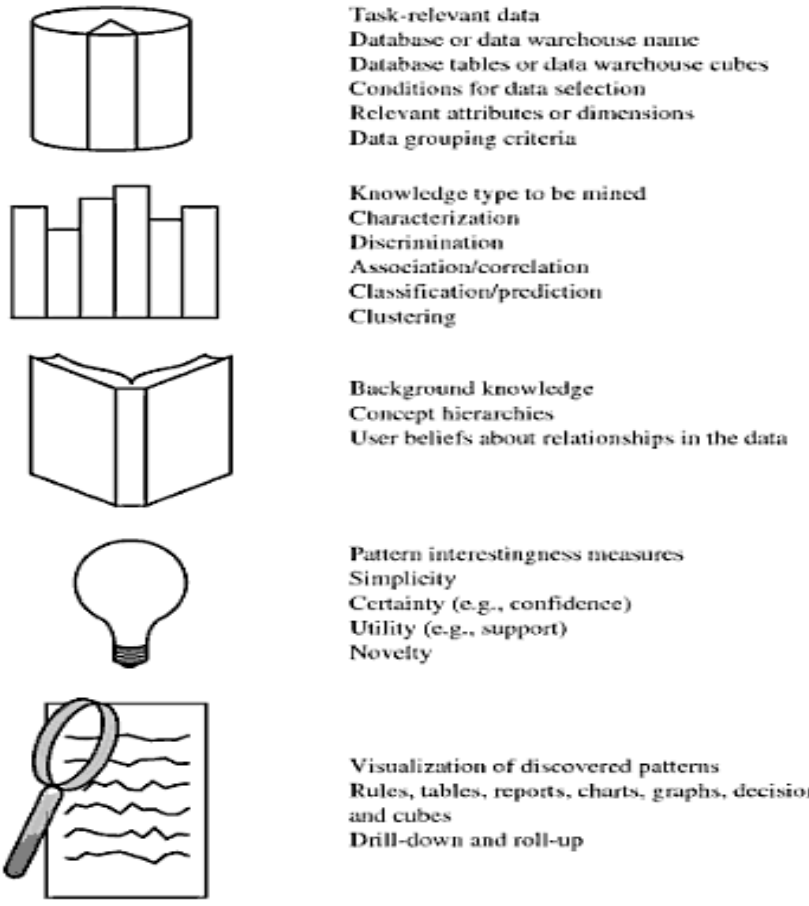
This will help in decision making in stock investment.

KLNCIT

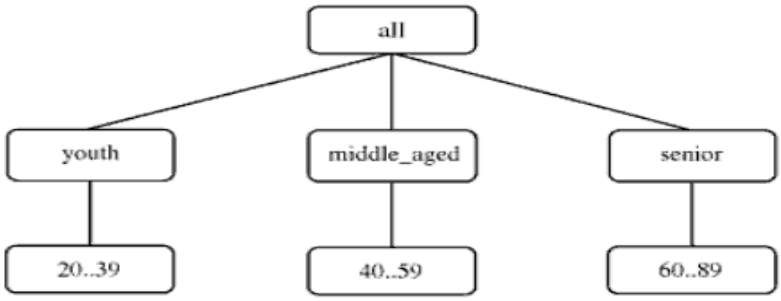
## 1.4 Steps in Data Mining Process

**Data Mining Task Primitives:**

- Data mining task can be specified in the form of a data mining query, which is input to the data mining system.
- Data mining query is defined in terms of data mining task primitives.
- Data Mining task primitives are:
    - Set of task relevant data to be mined. (relevant attributes / dimensions)
    - Kind of knowledge to be mined (kind of data mining functionality)
    - Background knowledge to be used in the discovery process. (knowledge base – concept hierarchy, user beliefs)
    - Interestingness measures and thresholds for pattern evaluation (Interestingness measure for association rules are 'support' and 'confidence')
    - Expected representation for visualizing the discovered patterns. (Tables, charts, graphs, decision trees, cubes)




Primitives for specifying a data mining task.




A concept hierarchy for the attribute (or dimension) *age*. The root node represents the most general abstraction level, denoted as all.

**Are all of the Patterns Interesting?**

- A Data Mining system can generate thousands or even millions of patterns or rules, but only few are interesting.
- What makes a pattern interesting?
  - A pattern is interesting if it is:
    - Easily understood by humans
    - Valid on a test data (with some degree of certainty)
    - Useful & Novel
    - Validates a user defined hypothesis
- Interesting pattern = Knowledge
- Objective measures of Pattern Interestingness:
  - Association rules of the form X => Y is Support
  - Support = "% of transactions from the database that satisfies the given rule"
  - Support (X => Y) = P (X U Y)
  - Confidence = "Assesses the degree of certainty of the association rule"
  - Confidence (X => Y) = P (X / Y) [i.e. Transaction containing X also contains Y]
  - Rules that do not satisfy a confidence threshold say 50% is uninteresting.
  - Rules below the threshold are noise / exceptions
  - Subjective Interestingness Measure: (Based on users belief in the data)
    - Unexpected (Contradicts user's belief)
    - Expected (Confirms user's belief)
    - Actionable (User can act on)
- Completeness of Data Mining Algorithms:
  - Can a data mining algorithm generate all of the interesting patterns?
    - If so it is unrealistic and inefficient.
  - Instead user provided specification can confine the search on interesting patterns
- If a data mining algorithm produces only interesting patterns → it is highly desirable & efficient, but it is a challenge in Data Mining Domain.

KLNCIT

## 1.5 Architecture of Typical Data Mining Systems

**Architecture of a typical Data Mining System – Major Components:**

Knowledge Base:

- Domain knowledge is used to guide search – used to evaluate interestingness of patterns.
- Includes concept hierarchies, user benefits, thresholds, metadata etc.

Database / Data warehouse Server:

- Responsible for fetching relevant data based on data mining request.

Data Mining Engine:

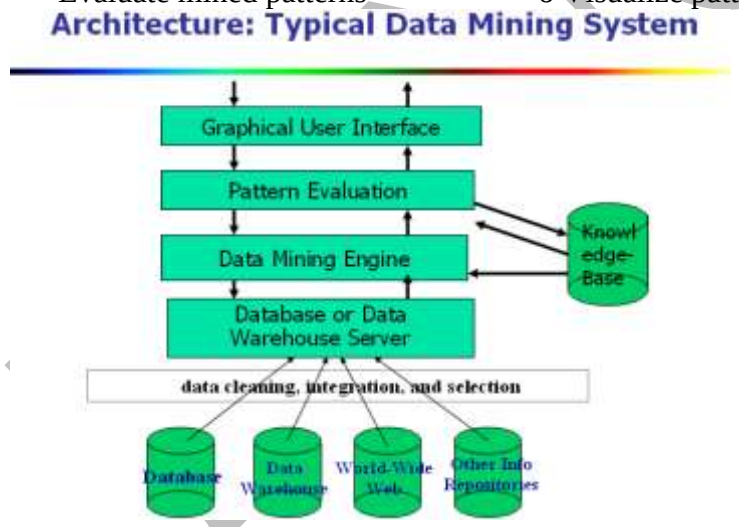
- Consists of modules for characterization, association, correlation analysis, classification, cluster analysis, prediction, outlier analysis and evolution analysis.

Pattern Evaluation Module:

- Interacts with data mining modules. O Focuses the search towards interesting patterns
- Pattern evaluation module may be integrated with mining module to confine the search.

User Interface:

- Communicates between users and data mining system
- Specifies data mining query – to focus search
- Uses intermediate data mining results to perform exploratory data mining.
- Browse database / data warehouse
- Evaluate mined patterns o Visualize patterns in different forms




## 1.6 Classification of Data Mining Systems

**Classification of Data Mining Systems**

- Data Mining is an inter-disciplinary field.
- It is a confluence of a set of disciplines:
  1. Database Systems
  2. Statistics
  3. Machine Learning
  4. Visualization
  5. Neural Networks
  6. Fuzzy Logic
  7. Rough Set Theory
  8. Knowledge Representation

9. Spatial Data Analysis 10. Information Retrieval
11. Pattern Recognition 12. Image Processing
13. Signal Processing 14. Computer Graphics
15. Web Technology 16. Economics
17. Business 18. Bio-informatics 19. Psychology

- Generates large variety of data mining systems.
- Need for classification of Data Mining Systems => users can choose based on their needs.
- Data Mining Systems Classification based on:

1) Classification according to the kinds of databases mined:

- This classification is based on different criteria:
  - Data Models
    - Relational mining systems
    - Transactional mining systems
    - Object-relational mining systems
    - Data Warehouse mining systems
  - Type of Data
    - Spatial Data
    - Time-Series Data
    - Text Data
    - Stream Data
  - Type of Application
    - Multimedia Data Mining Systems
    - World Wide Web mining Systems

2) Classification according to the kinds of knowledge mined:

- Based on the data mining functionalities:
  - Characterization & Discrimination
  - Association & Correlation Analysis
  - Classification & Prediction
  - Clustering, Outlier Analysis, Evolution Analysis
- Based on the granularity levels of abstraction of knowledge mined
  - Generalized knowledge (High Level of Abstraction)
  - Primitive Level Knowledge (Raw data level)
  - Knowledge at multiple levels
- Data mining systems that mine data regularities (Common data patterns) **Vs** Data mining systems that mine data irregularities (Outliers / Exceptions).

3) Classification according to the kinds of techniques utilized:

- Based on the degree of user interactions involved:
  - Autonomous Systems
  - Interactive Exploratory Systems
  - Query Driven Systems
- Based on the methods of data analysis involved:
  - Database oriented method of data analysis
  - Data warehouse oriented method of data analysis
  - Machine Learning > Statistics > Visualization
  - Pattern Recognition > Neural Networks
  - Sophisticated Data Mining System has effective integrated techniques

KLNCIT

4) Classification according to the applications adapted:

- Finance Telecommunication DNA
- Stock Market E-mail
- Integrated Application specific systems

## 1.7 Overview of Data Mining Techniques

**Major Issues in Data Mining:**

- Mining Methodology Issues:
  - Mining different kinds of knowledge in databases.
  - Incorporation of background knowledge
  - Handling noisy or incomplete data
  - Pattern Evaluation – Interestingness Problem
- User Interaction Issues:
  - Interactive mining of knowledge at multiple levels of abstraction
  - Data mining query languages and ad-hoc data mining.
  - Presentation and visualization of data mining results.
- Performance Issues:
  - Efficiency and Scalability of Data Mining Algorithms.
  - Parallel, distributed and incremental mining algorithms.
- Issues related to diversity of data types:
  - Handling of relational and complex types of data.
  - Mining information from heterogeneous databases and global information systems.

**Data Mining Query Language:**

- DMQL – adopts SQL like syntax o DMQL can be easily integrated with SQL
- Example: "Want to classify customers whose salary >= Rs.40,000
  And who have purchased for >= Rs.1,000 of which the price of each item >= Rs.100.
  We are interested in customer's age, income, types of items purchased,
  Purchase location, where items were made." → Resulting classification is in rule form.
  - **Use database** store_db
  - **Use hierarchy** location_hierarchy **for** T.branch, age_hierarchy **for** C.age
  - **Mine classification as** promising_customers
  - **In relevance to** C.age, C.income, I.type, I.place_made, T.branch
  - **From** customer C, item I, transaction T
  - **Where** I.item_ID = T.item_ID
    **And** C.cust_ID = T.cust_ID
    **And** C.income >= 40000
    **And** I.price >= 100
  - **Group By** T.cust_ID
  - **Having** sum(I.price) >= 1000
  - **Display as** rules
- We want to generate a classification model for "promising_customers" Vs "non_promising_customers"

KLNCIT

- Attribute may be specified as **Class Label Attribute** whose values explicitly represent the classes.
- Specified data are retrieved and assigned as "Promising_customers" and the remaining data in the database are assigned as "non-promising_customers"
- Other types of Data Mining languages are:
    - Microsoft's OLEDB for data mining includes DMX – XML styled data mining language.
    - PMML – Programming Model Markup Language
    - CRISP-DM – CRoss Industry Standard Process for Data Mining

**Integration of a Data Mining System with a Database or Data Warehouse System:**

- DMQL – adopts SQL like syntax

## Review Questions

**Two Marks:**

1. What motivated Data Mining? Why is it important?
2. What is Data Mining?
3. List the kinds of data upon which Data Mining can be done.
4. What is the difference between Data Warehouse and Data Mart?
5. Write about the relation of Statistics with Data Mining.
6. What makes a pattern interesting?
7. Write about the objective measures of pattern interestingness?

**Sixteen Marks:**

1. (i) Describe the Database System Evolutionary Path. (8)
   (ii) Explain the steps in the Knowledge Discovery Process. (8)
2. (i) Detail on the Architecture of Data Mining Systems with a suitable diagram. (8)
   (ii) Explain about the data warehousing process. (8)
3. Detail on the various kinds of data upon which data mining can be done. (16)
4. Explain about various Data Mining functionalities. (16)
5. (i) List down and explain the classifications of data mining systems. (6)
   (ii) Discuss about the major issues in data mining. (5)
   (iii) Write about the integration of data mining system with the database or data warehouse system. (8)

**Assignment Topic:**

1. Explain on the Data Mining Query Language.

KLNCIT