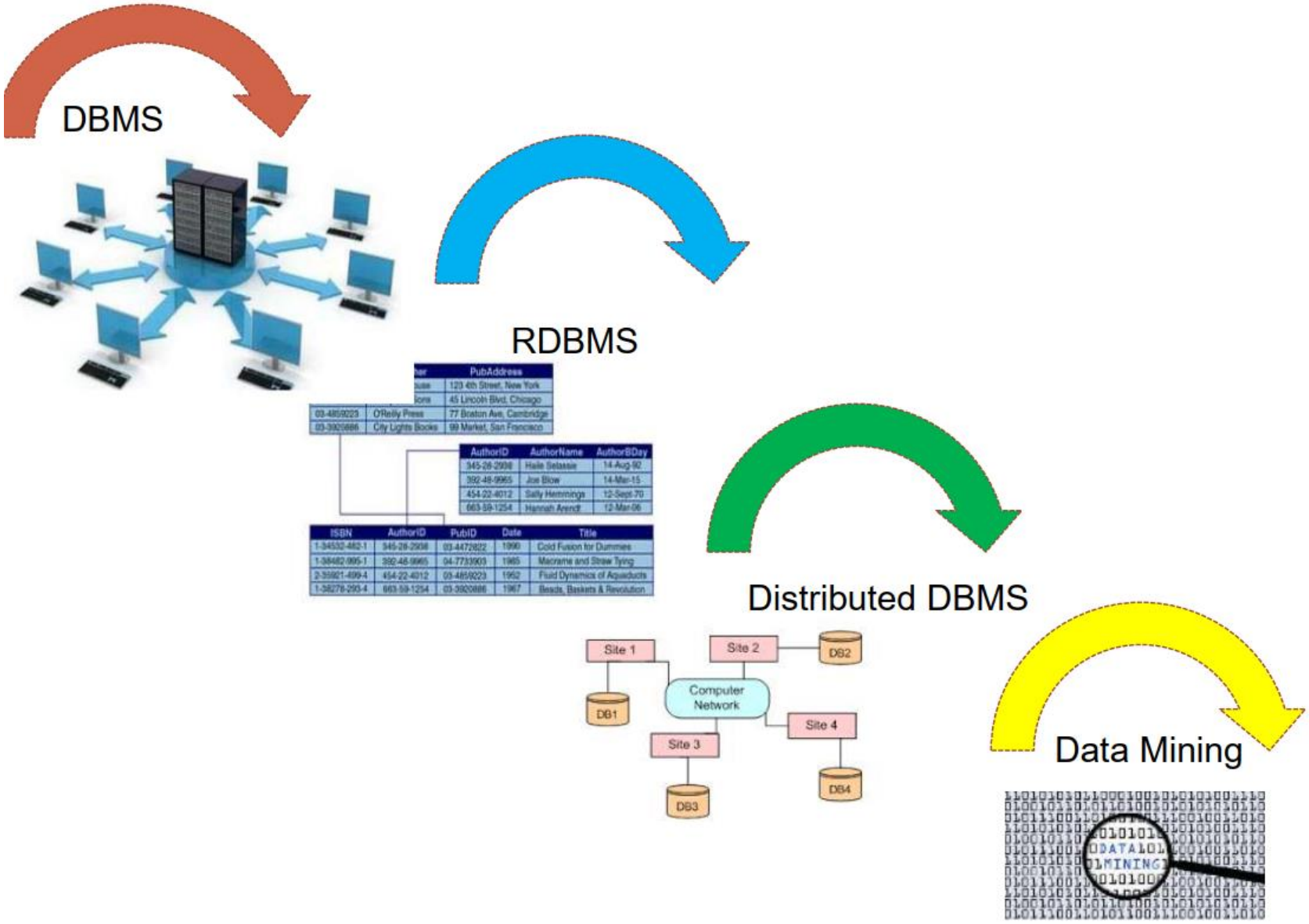
E-Commerce Applications Development

Data Mining in Ecommerce business

Data at different stages



Data Mining

- Data Mining: It deals with the discovery of hidden knowledge, unexpected patterns and new rules from large data sets.

Data Mining

- There is huge amount of data available in the information industry. This data is of no use until it is converted into useful information. It is necessary to analyse this data and extract useful information from it.
- Extraction of information is not only the single process, data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining , Pattern evaluation.
- Once all these processes are over, we would be able to use this information in many applications such as Fraud detection, Market-Basket analysis etc

Data Mining

What is Data Mining?

- Extraction of interesting Patterns or Knowledge from huge amount of data.
- Knowledge Discovery from Data (KDD).

Need for Data Mining:

- The explosive growth of data: from terabytes to peta bytes
- We are drowning in data but starving for knowledge.

Data Mining

Examples of Information extracted using query language:

- List customer who use credit cards to purchase products of worth 5000 rupees and above.
- List employees who have taken loans.
- List patients who had at least one heart attack.

Data Mining

Examples of what data mining is used for:

- Develop a general profile of credit card customers.
- Determine patients whose lifestyle is prone to getting a hearth attack in near future.
- Differentiate employees who have taken loan for any purpose.

Data Mining

- Data mining differs from usual query from normal relational database in following manner:

	Query Processing	Data Mining
Query	Wel formed as Select... From... Where.....	Query is not well formed. What is found out that is usually hidden
Data	Data from online transaction processing systems generally in table formats	Data is integrated from various sources. Huge amount of data
Output	Subset of databases	Not only subset but also in analyzed and in terms of patterns

Data Mining

- Data Mining (Knowledge discovery from data)

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.
- Alternative names used for data mining:
  - Knowledge discovery(mining) in database(KDD), knowledge extraction, data/pattern analysis, data archelogy, data dredging, information harvesting, business intelligence etc

Data Mining

- Knowledge discovery in database (KDD) is a multistep process of finding useful information and patterns in data while Data Mining is one of the steps in KDD of using algorithms for extraction of patterns.

Steps of KDD:

1. Selection:

Data Extraction: Obtaining data from heterogenous data sources, Databases, Data warehouses, World wide web or other information repositories.'

2. Pre-processing:

Data Cleaning: Incomplete, noisy inconsistent data to be cleaned, missing data may be ignored or predicted, erroneous data may be deleted or corrected.

Data Mining

3. Transformation:

Data integration: combines data from multiple sources into a coherent store, data can be encoded in common formats , normalized , reduced.

4. Data Mining:

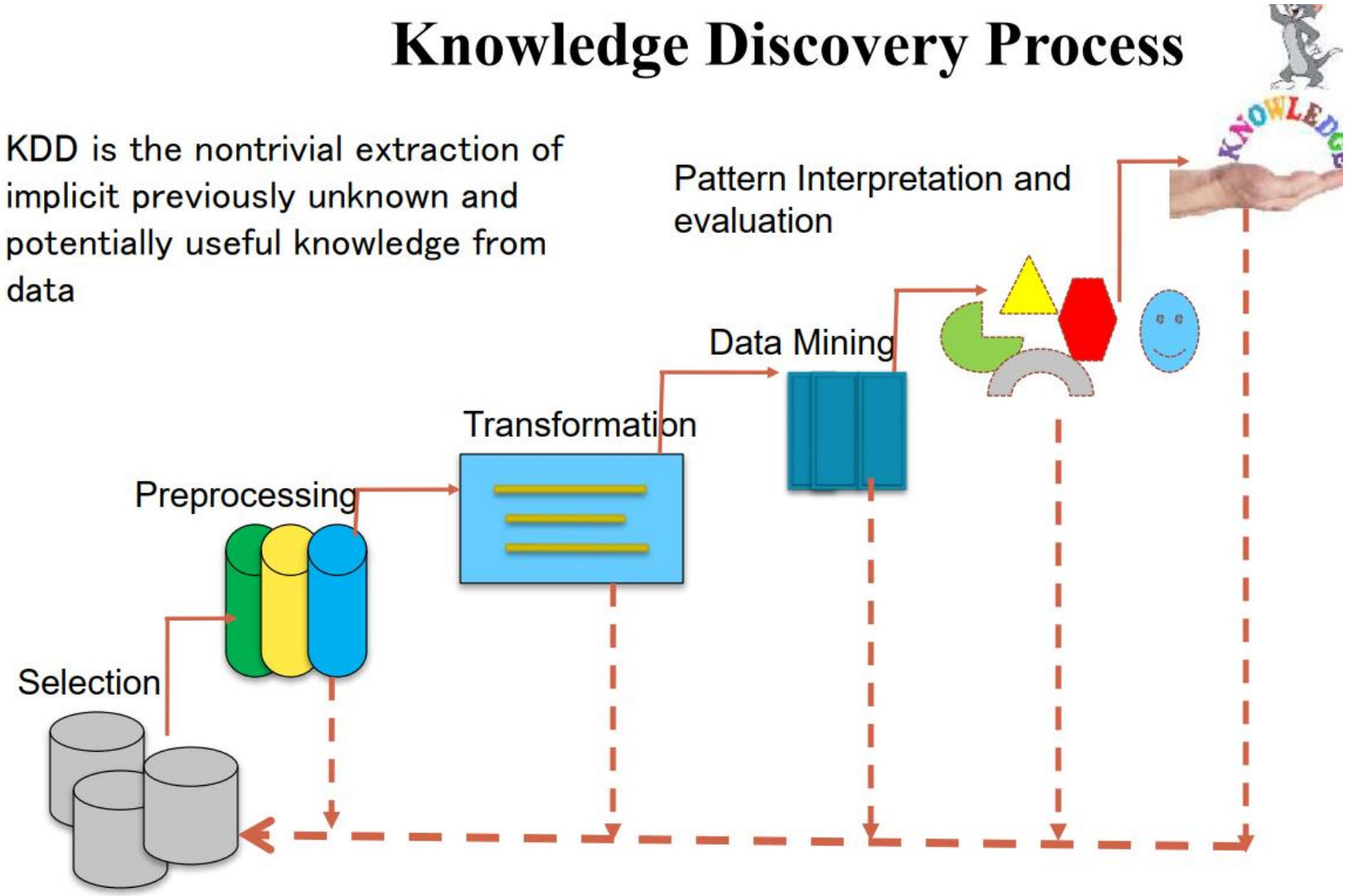
Apply algorithms to transformed data an extract patterns.

5. Pattern Interpretation/ evaluation:

Pattern evaluation: Evaluate the interestingness of resulting patterns or apply various measures to filter out discovered patterns.

Knowledge presentation: present the mined knowledge visualization techniques can be used.

Data Mining

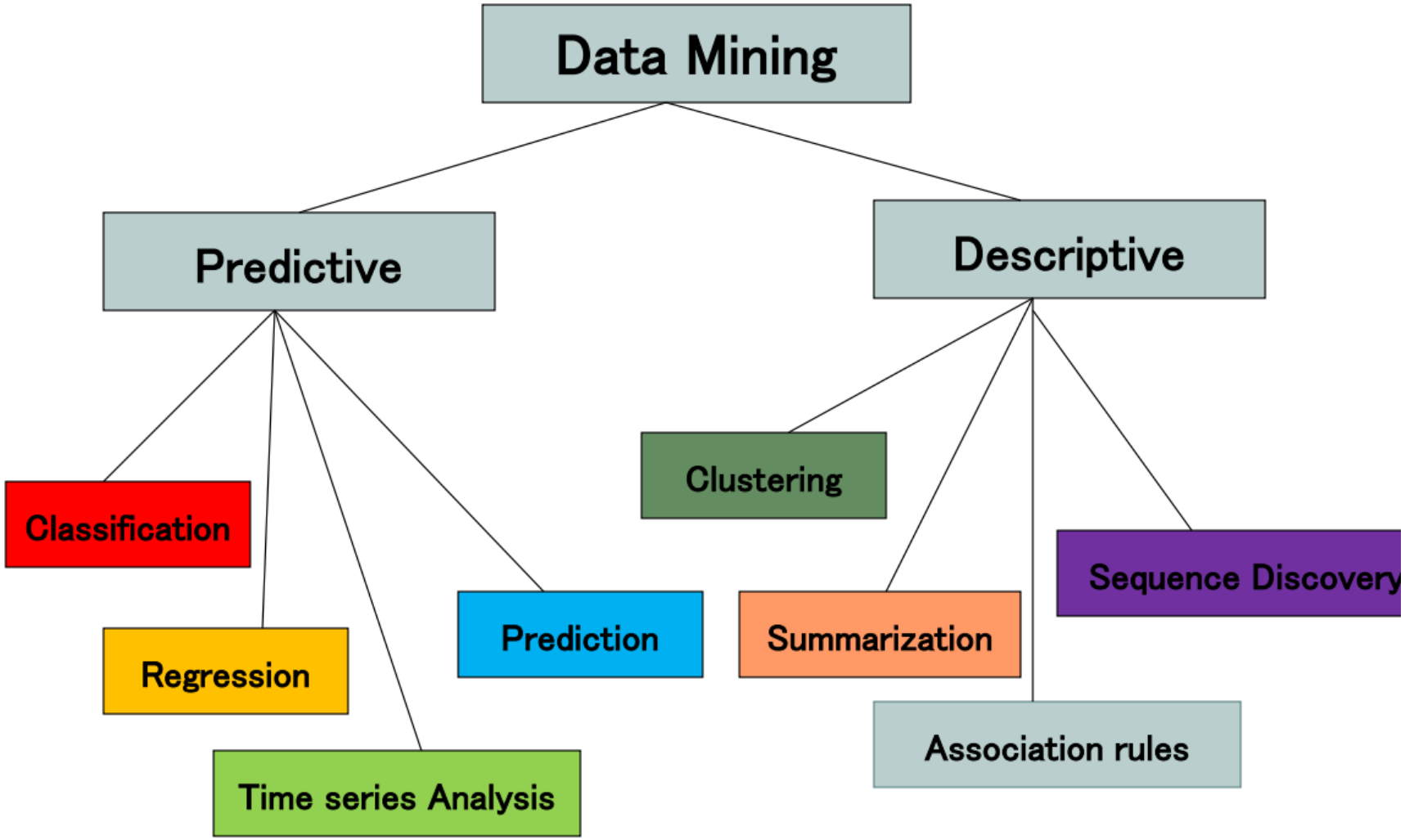

Knowledge Discovery Process
KDD is the nontrivial extraction of implicit previously unknown and potentially useful knowledge from data
Pattern Interpretation and evaluation
Data Mining
Transformation
Preprocessing
Selection

Data Mining

Different forms of data:

- Relational databases:
  - Collection of tables
- Data warehouses:
  - Data from different sources
- Transactional databases:
  - Consists of a file where each record represent transactions.
- Advanced Data and Applications
  - Multimedia and spatial data

Data Mining


Data Mining
Predictive
Descriptive
Classification
Regression
Time series Analysis
Prediction
Clustering
Summarization
Association rules
Sequence Discovery

Data Mining

Predictive techniques:
Predict values of data by making use of known results from a different set of sample data.

Descriptive techniques:
Enables you to determine patterns and relationships in sample data.

Data Mining

Common Data Mining Techniques:

- Classification
- Clustering
- Regression
- Association Rules
- Prediction

Data Mining

Classification:

- Maps data into predefined groups or classes.
- It uses supervised learning.
- The algorithm uses learning phase to build a classifier using training data set containing data attributes and associated class labels.

Example:

Result of a student: If a student gets a total of 85 out of 100 then he will secure grade A.

Pattern recognition is type of classification where input pattern is classified into several classes based on its similarity to predefined classes.

Example:

identifying different users' behavior on a website and then working on front end layout for improvements.

Data Mining: Classification

To understand classification let's look at an example of a child trying to classify the fruit kept in front of him into either apple or orange. Now based on certain features such as color, shape, smell etc. the child will categorize the fruit into apple or orange.

Now when a model in data mining tries to classify the target based on certain features from the dataset into various classes or categories that is called classification

Data Mining

Classification: Grading of coal furnace

Grade	Useful Heat Value(kcal/kg)
A	>6200
B	5601 – 6200
C	4941 – 5600
D	4201 – 4940
E	3361 – 4200
F	2401 – 3360
G	1301 – 2400

Data Mining

Clustering:

- Find similarities between data according to the characteristics found in the data and grouping similar data objects into clusters.

- Unsupervised learning: no predefined classes/ Labels
- Interpretability and usability- results should be comprehensible and usable-domain expert is required.

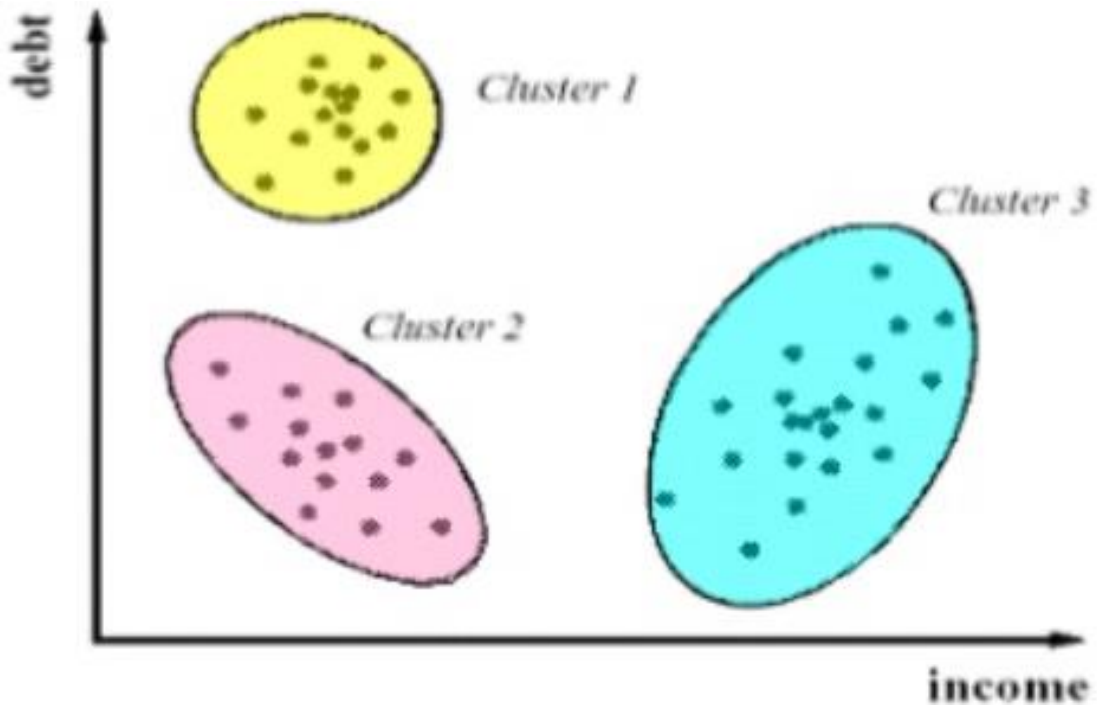
Example:

  - Students are clustered among various attributes like good academics, area in which they live, age, height, body mass index, extra, extra curricular activities.
- Cluster do not have specific and shape.

Data Mining

Clustering:

For example: similar data is grouped in same cluster.



Data Mining

Regression:

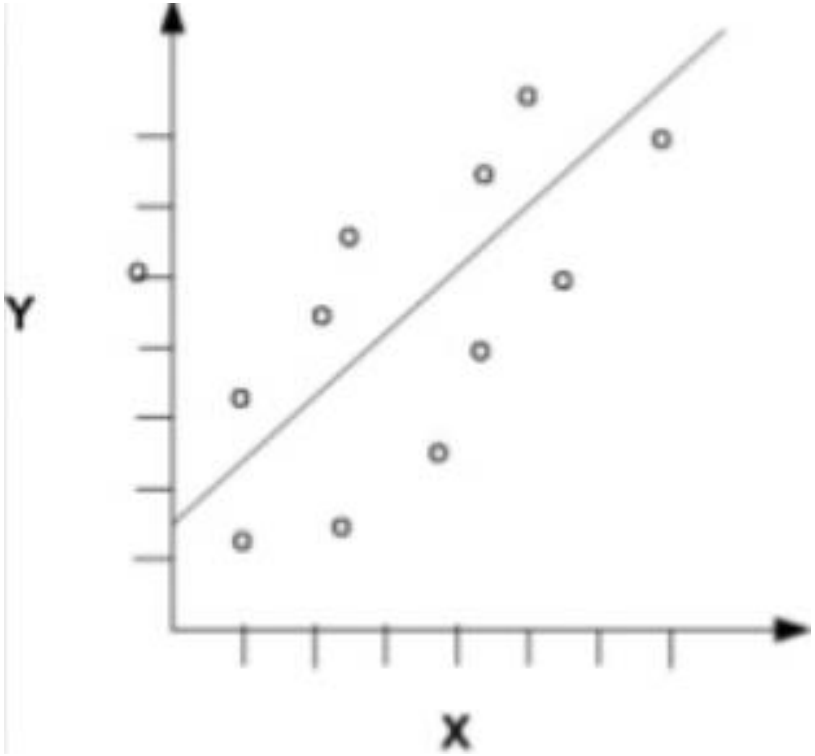
Regression deals with the prediction of a value rather than a class.

- It maps data into real-valued prediction variable.
- Algorithm tries to find best function (linear, non-linear that fits into some function)

Data Mining

Regression:

- Regression is used to predict value of y, given value of x.
- For Example: a model could be used to predict children's height given their weight, age and other factors



Data Mining

Regression:

Bike Damaged example: In the following table attributes are given such as color, type, origin and subject can be yes or no.

Bike No	Color	Type	Origin	Damaged?
10	Blue	Moped	Indian	Yes
20	Blue	Moped	Indian	No
30	Blue	Moped	Indian	Yes
40	Red	Moped	Indian	No
50	Red	Moped	Japanese	Yes
60	Red	Sports	Japanese	No
70	Red	Sports	Japanese	Yes
80	Red	Sports	Indian	No
90	Blue	Sports	Japanese	No
100	Blue	Moped	Japanese	Yes

- Results are determined using probabilities and likelihood.

Data Mining

Association:

- An association algorithm creates rules that describe how often events have occurred together.
- It discovers relationship among data – used in Market-basket analysis to find item frequently purchased together.

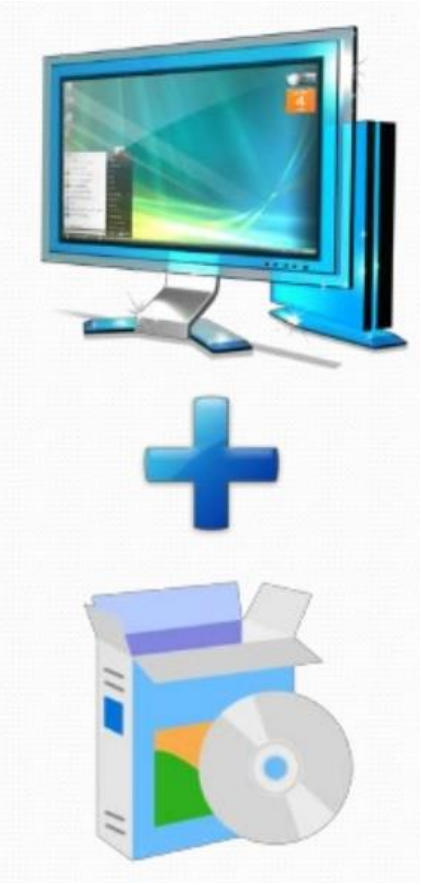
Example:

- person buying a sugar in the mall also buys milk. The thing which person buy together will always kept together.
- When a customer buys a computer, then 90% of the time they will buy software disks.

Data Mining

Association rules:

- Market-Basket analysis is very common in online ecommerce stores, research has shown that through this technique customers mostly perform extra purchases.



Data Mining

Prediction:

- predicts future values using regression, time series analysis or other approaches.
- Example: to find out flood prediction of river depending on water level, rain amount time humidity, Sensors at different locations are placed in the river area which will monitor flood conditions and flood prediction can be done.

  - Weather prediction / forecasting
  - Pollution analysis

Data Mining

Depending on data mining approach, techniques from other disciplines may be applied such as:

- Information Retrieval
- Artificial Intelligence
- Neural network
- Fuzzy set theory
- Knowledge representation
- Logic programming
- High performance computing

Data Mining

Data mining Metrics:

How to measure the effectiveness of data mining process?

-KDD process is expensive , Return on Investment will be the saving due to decision process using the results.

Difficult to measure and quantify.

Social Implications of Data mining:

There are two aspects related to this,

1. Data mining can be used to improve customer service and satisfaction.
2. Data mining can be used to confront any one's right to privacy and their personal information can be used against their consent.

Data Mining

Data mining should follow certain guidelines:

- Purpose specification and use limitation (defined scope)
- Openness
- Security safeguards (prioritize security of information)
- Privacy preserving data mining

Data Mining

Applications of Data Mining:

1. Market Basket Analysis:

It is modelling technique based upon a theory that if you buy a certain group of item you are more likely to buy another grpup of items.

This information may help the retailer to know the buyer's needs and retailer can enhance the store 's layout.

2. Bio-Informatics:

Mining biological data helps to extract useful knowledge from massive datasets gathered in biology and in other related life sciences areas.

Applications of data mining to bioinformatics include gene finding, protein function inference ,, disease diagnosis disese treatment.

Data Mining

3. Education:

Data mining can be used by an institution to take accurate decisions and also to predict the results of the student.

Learning pattern of the students can be captured and used to develop techniques to teach them.

4. CRM Customer Relationships Management:

To maintain a proper relationship with a customer a business need to collect data and analyse the information

With data mining technologies the collected data can eb used for analysis.

Data Mining

- Analysing Consumer behaviour using Association rule mining:

- An observant at Walmart discovered that there is a strong association between bread and eggs.
- Analysis of purchases revealed that whenever bread was purchased, most of the time eggs were also purchased.

- This knowledge was used and these products were placed next to each other.

Data Mining

Market-Basket Analysis:

- Identifies customer purchasing habits.
- It provides insight into combination of products within a customers basket.
- We often compare all orders associated with a single customer.

Data Mining

- What is association rule mining?

- Recent research has positioned association mining as one of the most popular tools in retail analytics.
- A data mining technique which generates rules in the form of X => Y where X and Y are two non-overlapping discrete sets.

Association rules-applications

Association rules-applications

Association rules-applications

Association rules-applications

- Developing this understanding enables businesses to promote their most profitable products.
- It can also encourage customers to buy items that might have otherwise been overlook or missed.
- Almost all ecommerce stores use this knowledge to increase their sales. For instance, look how amazon store promotes other similar products(cross-selling) to customers:

Market-basket illustration

ID	ITEMS
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

market basket transactions

{Diapers, Beer} Example of a frequent itemset

{Diapers} → {Beer} Example of an association rule

Market-basket illustration

- In 60% cases, diaper and beer are sold together. In 75% of cases when someone purchases a diaper, a beer is also purchased.

A => B [Support, Confidence]

ANTECEDENT CONSEQUENT

{DIAPER} -> {BEER} [Support = 60% , Confidence = 75 %]

Association rules

- Rule: If a basket contains X , it is likely to contain Y
- X-apples , Y-bread , so the rule will be if a customer buys apples he is likely to buy bread.
- The rule is depicted as {apples} --> {bread}
- Apriori algorithm is used to find confidence and support between products and then ultimately define association rules.

Data Warehousing

- A data warehouse is a large collection of business data used to help an organization make decisions. The data comes from different sources such as internal applications like marketing, sales data, finance, customer-side, and other external systems

- Data warehousing is combining data form multiple sources into one comprehensive and easily manipulated database.

- The primary aim for data warehousing is to provide businesses with analytics results from data mining, OLAP and reporting.

Data Warehousing

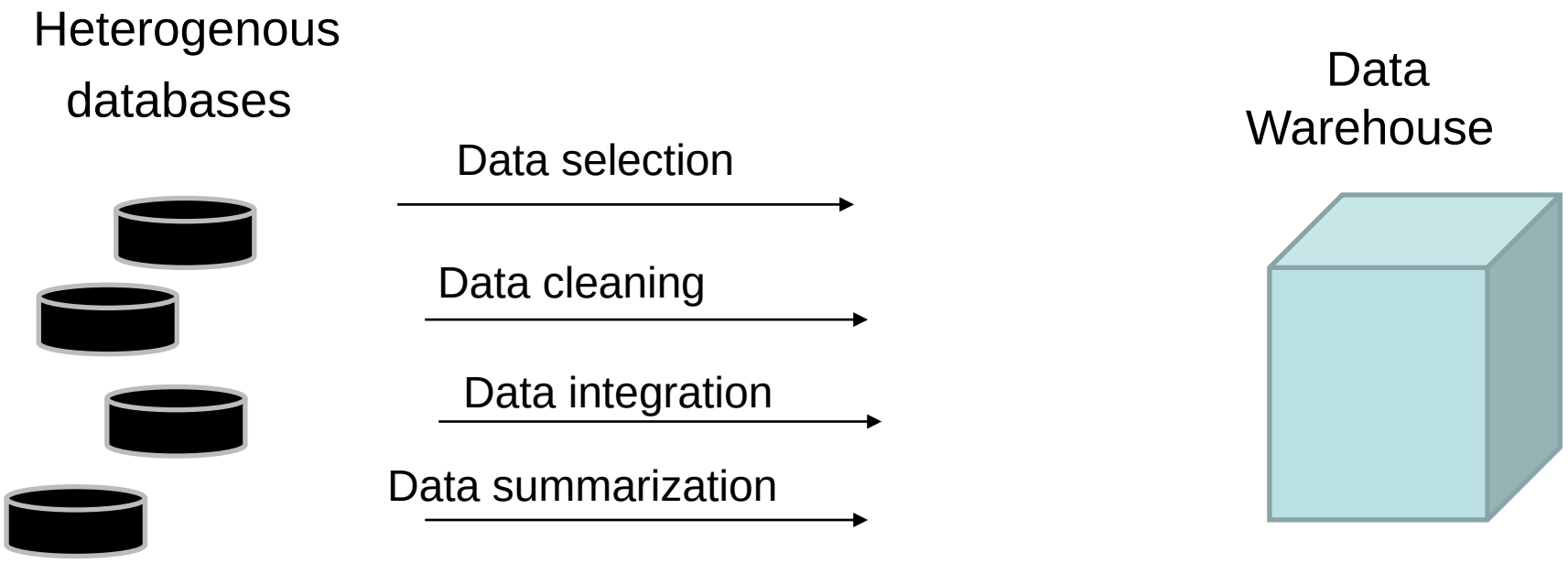
Data warehousing includes:

- Retrieving data
- Analysing data
- Extracting data
- Loading data
- Transforming data
- Managing data

Data Warehousing

Why Data Warehousing?

- Data warehousing can be considered as an important pre-processing step for data mining



- A data warehouse also provides On-Line Analytical processing OLAP tools for interactive multidimensional data analysis.

Data Warehousing

- A multidimensional model views data in the form of a data-cube.
- A data cube enables data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.
- For example, a shop may create a sales data warehouse to keep records of the store's sales for the dimension time, item, and location.

Data Warehousing

Business advantages of Data warehousing:

- It provides business users with a “customer-centric” view of the company’s heterogenous data by helping to integrate data form sales, service, manufacturing and distribution and other customer related business systems.

- It provides added value to the company’s customers by allowing them to access better information when data warehousing is coupled with internet technology.

Data Warehousing

- It consolidates data about individual customers and provides a repository of all customer contacts for segmentation modelling, customer retention planning and cross sales analysis.

- It reports on trends across multidivisional, multinational operating units, including trends or relationships in areas such as merchandising, production planning etc.

Data Warehousing

Data Marts

- A data mart is a scaled down version of a data warehouse that focuses on a particular subject area.
- A data mart is a subset of an organizational data store, usually oriented to a specific purpose or major data subject.
- Data marts are analytical data stores designed to focus on specific business functions for a specific community within an organization

Reason to use data mart:

- Easy access to frequently needed data
- Ease of creation in less time
- Improves end user response
- Lower cost than implementing a full data warehouse

Thank you