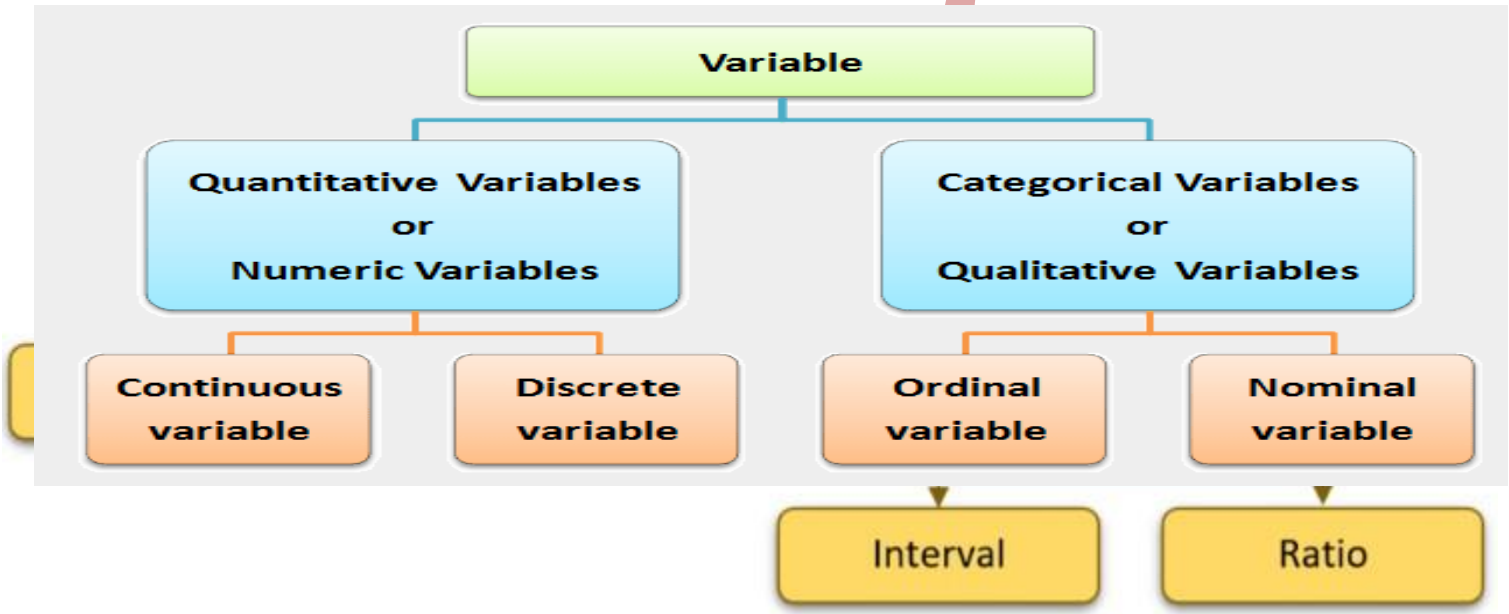
UNIT 1
DESCRIBING DATA

Types of Data - Types of Variables -Describing Data with Tables and Graphs –Describing Data with Averages - Describing Variability - Normal Distributions and Standard (z) Scores

APEC

Types of Variables



Numerical Variable

- Numerical Variable is also called a Quantitative Variable.
- Numerical data is either an integer or a real value.
- It has a mathematical meaning.
- There are two types of Numerical Variables. They are,
  - Discrete Variable
  - Continuous Variable
- Discrete
- Quantitative discrete variables are variables for which the values it can take are countable and have a finite number of possibilities. The values are often (but not always) integers. Here are some examples of discrete variables:

- Number of children per family
- Number of students in a class
- Number of citizens of a country

- **Continuous**
- On the other hand, **quantitative continuous** variables are variables for which the values are **not countable** and have an **infinite number of possibilities**. For example:
- Age
- Weight
- Height
- However, a 28-year-old man could actually be 28 years, 7 months, 16 days, 3 hours, 4 minutes, 5 seconds, 31 milliseconds, 9 nanoseconds old.

APEC

- Interval
- An interval scale is one where there is order and the difference between two values is meaningful.

- Examples of interval variables include:

- temperature (Farenheit), temperature (Celcius), pH, SAT score (200-800), credit score (300-850).

- Ratio
- A ratio variable, has all the properties of an interval variable, and also has a clear definition of 0.0. When the variable equals 0.0, there is none of that variable.

- Examples of ratio variables include:

- enzyme activity, dose amount, reaction rate, flow rate, concentration, pulse, weight, length, temperature in Kelvin (0.0 Kelvin really does mean “no heat”), survival time.

APEC

Data Mining & Data Warehousing

Why Data mining?

- I am not able to find the data I need.
  (Data is dispersed over the network)

- The data I have access to is poorly documented.
  (Proper information is missing)

- I am not able to use the data I have.
  (Unexpected results)

What is Data mining?

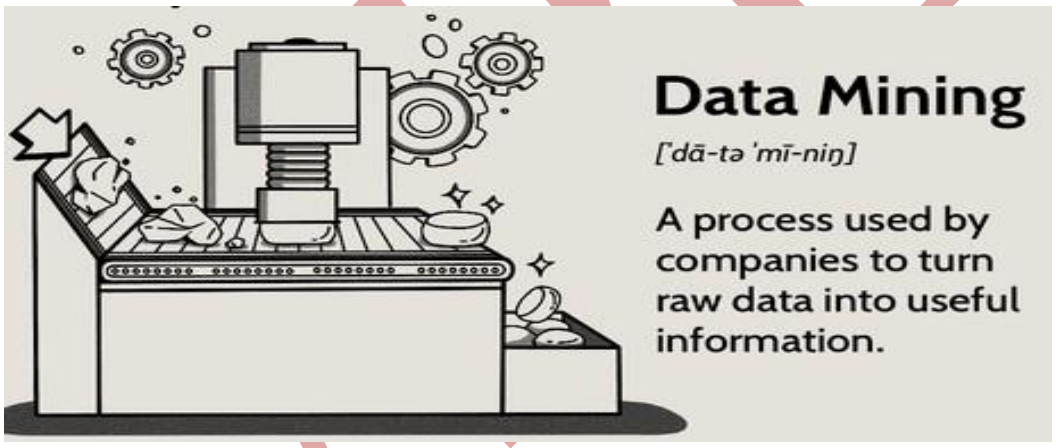
- Data mining refers to extracting or “mining” Knowledge from large amount of data.

  - By Jiawei Han, Micheline Kamber,



Data Mining

- Data mining is the process of sorting large data sets to identify patterns and relationships that helps to solve business problems

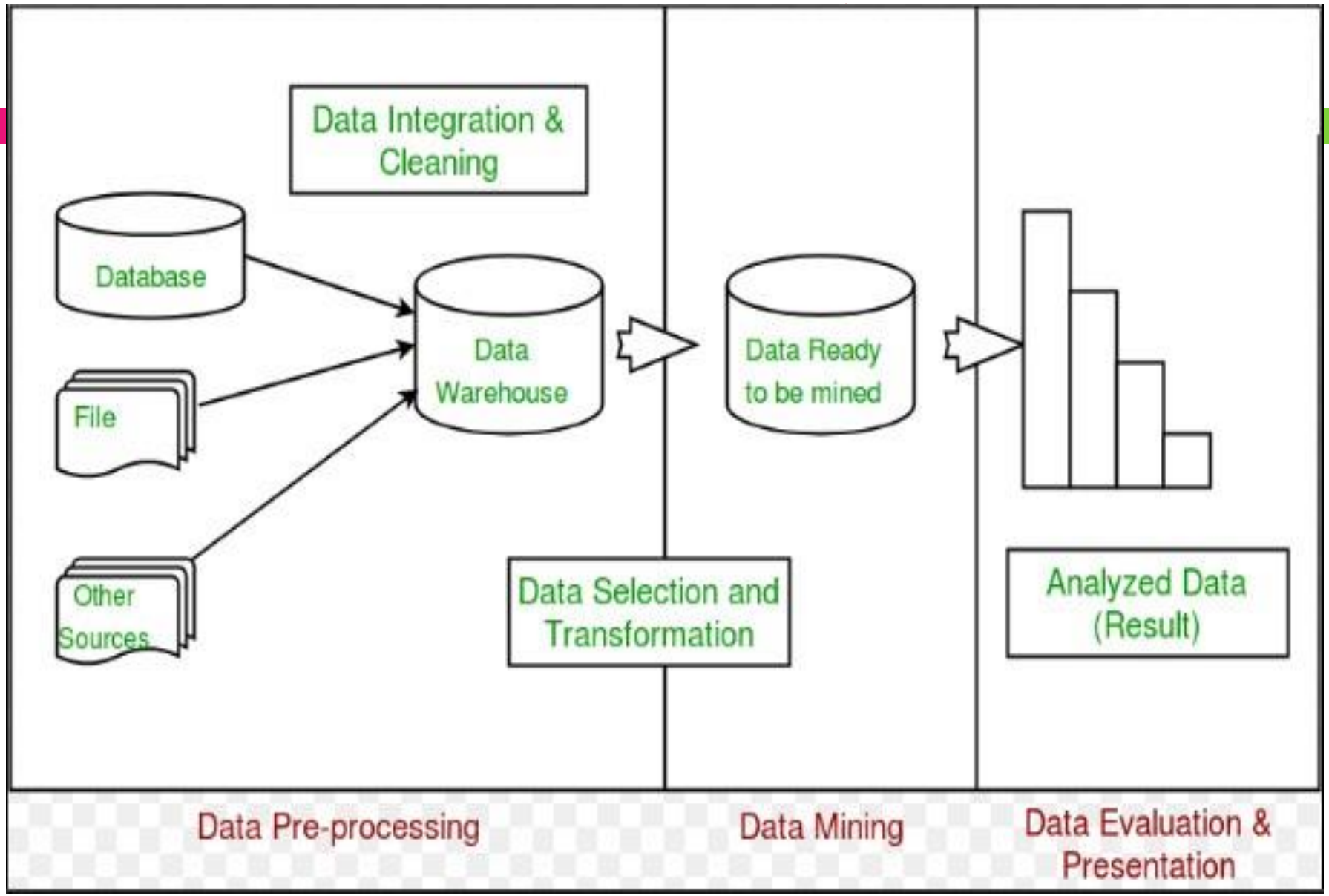


Data Mining → Process of Knowledge Discovery or Knowledge Extraction

Brief History of Data Mining:

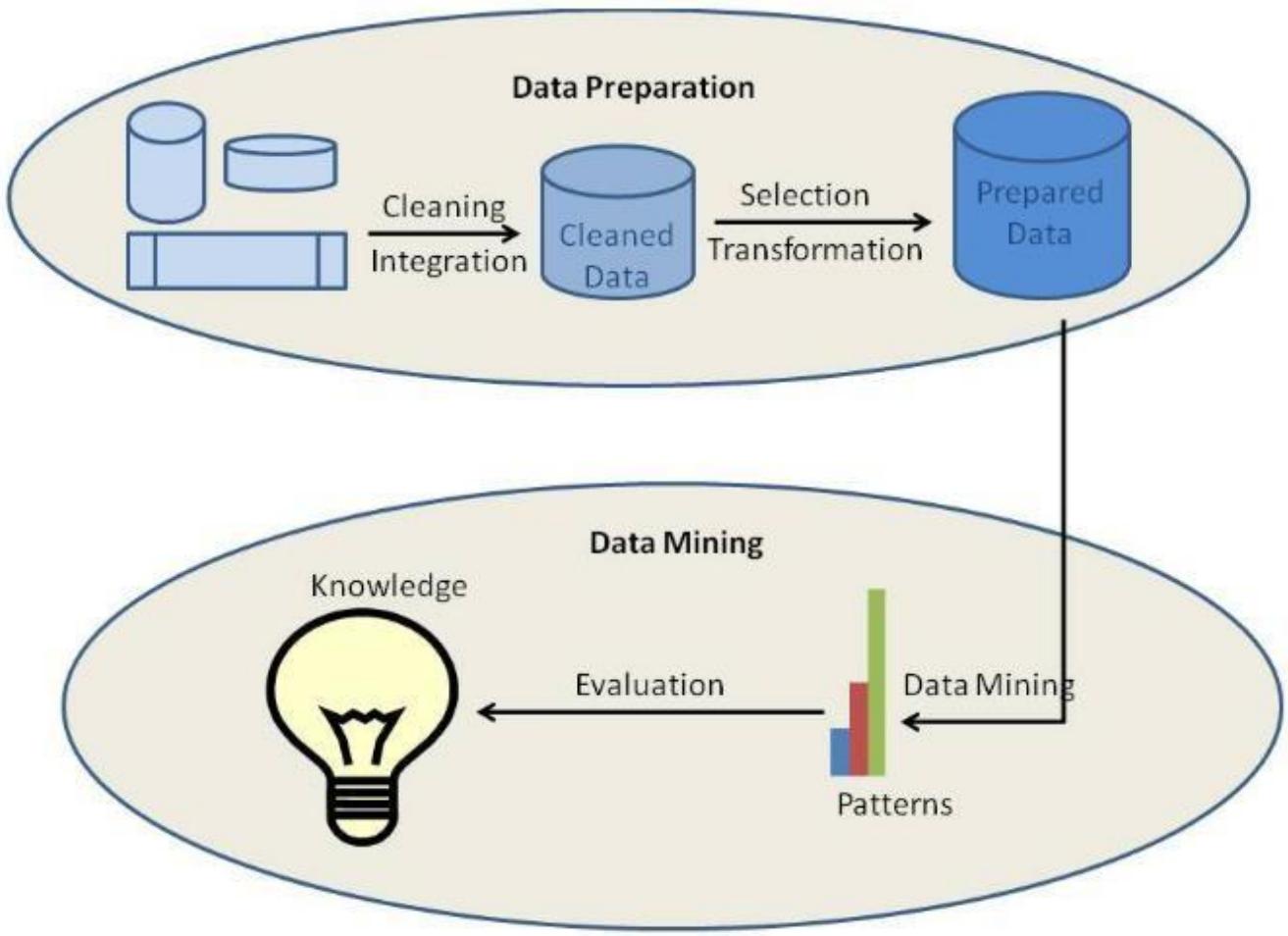
- Data mining first appeared in the 1970's and peaked around 2002
- Predictive analytics appeared in the 2000's, but has yet to catch on with the general public.
  - It is mostly used by businesses and government agencies.
- In 2010, experts claimed data science had five purposes. The scientist's job is to obtain data, scrub it, explore it, model it, and interpret it.
  - Social media platforms are now the top locations for data mining and analytics

APEC


Data Integration & Cleaning
Database
File
Other Sources
Data Warehouse
Data Selection and Transformation
Data Ready to be mined
Analyzed Data (Result)
Data Pre-processing
Data Mining
Data Evaluation & Presentation


APEC

Data Mining Process



Steps In The Data Mining Process

- The data mining process is divided into two parts i.e. Data Preprocessing and Data Mining.
- **Data Preprocessing involves**
  - Data cleaning
  - Data integration
  - Data reduction
  - Data transformation.
- **Data mining involves**
  - Data mining
  - Pattern evaluation
  - Knowledge representation of data.

APEC

Why do we preprocess the data?

- There are many factors that determine the usefulness of data such as
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness.
- Thus preprocessing is crucial in the data mining process.

APEC

Data Cleaning

- Data cleaning is the first step in data mining. This step involves the **removal of noisy or incomplete data** from the collection.
- This steps involved in cleaning are:
- **(i) Fill The Missing Data:**
  - Missing data can be filled by methods such as:
    - Ignoring the tuple.
    - Filling the missing value manually.
    - Use the measure of central tendency, median or
    - Filling in the most probable value.

APEC

Contd...

Student ID	Student Name	Age	GPA	Classification
100122014	Joseph	21	3.5	Junior
100232015	Patrick	200	3.2	Sophomore
100122012	Seller	24	3.0	Senior
100342013	Roger	23	234	Senior
100942012	Davis	2.8	3.7	Sophomore
	Travis	23	3.4	Sr
100982015	Alex	27		Sophomore
100982013	Trevor	-22	4.0	Senior
AUC2016XC	Aman	30	3.5	Jr

Missing Data

Inconsistent Data

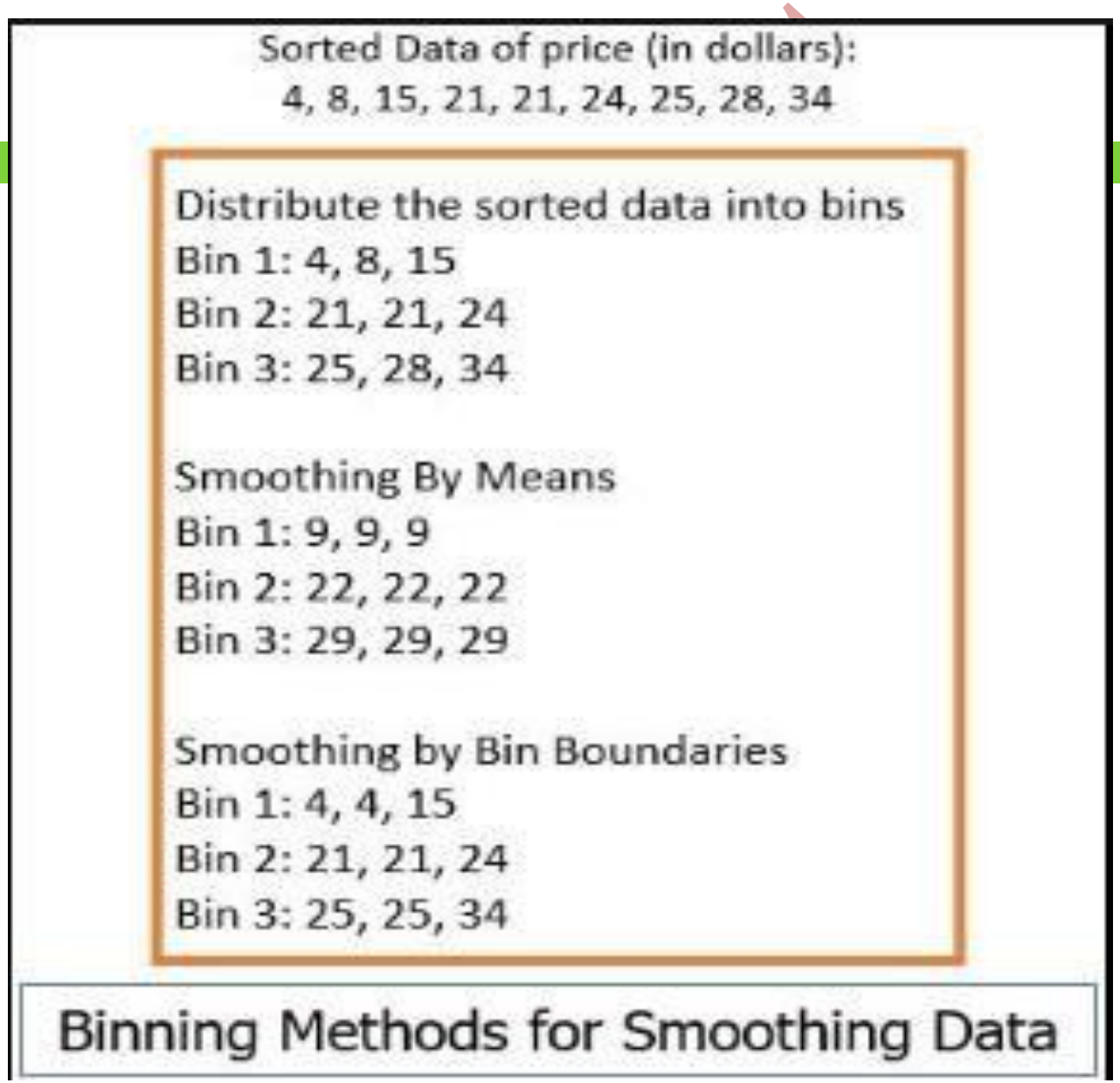
Noisy Data

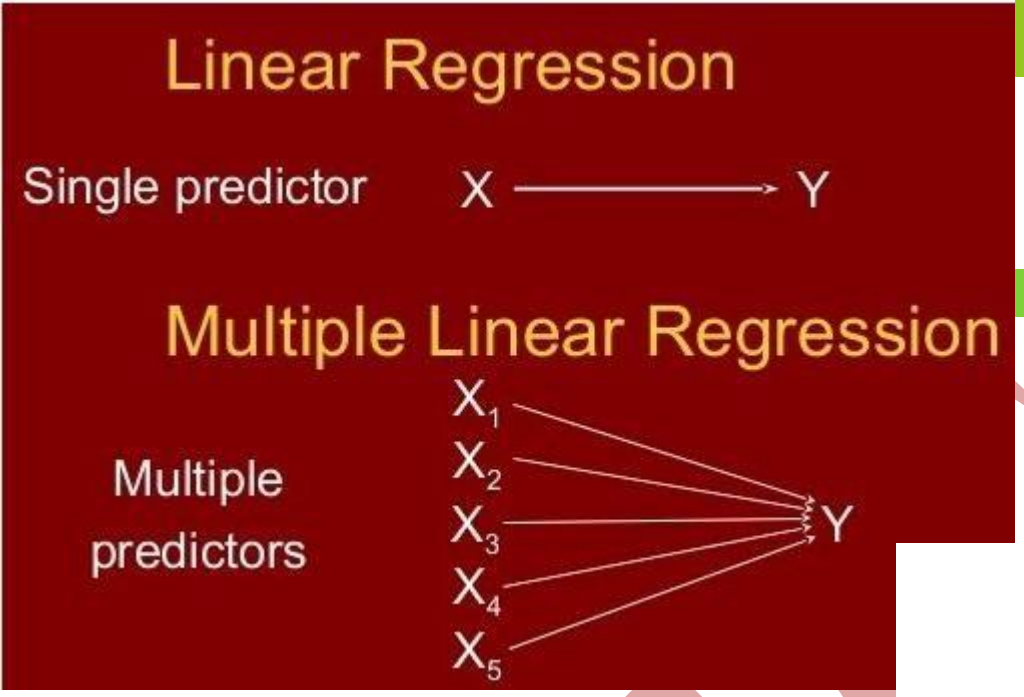
Data Cleaning

- **(ii) Remove The Noisy Data:** Random error is called noisy data. Methods to remove noise are :
  - Binning
  - Regression
  - Classification
  - Identifying the Outliers
  - Inconsistent Data

  - **Binning:**
    - Binning methods are applied by sorting values into buckets or bins. Smoothening is performed by consulting the neighboring values.
    - Smoothing by a median
      - Each bin value is replaced by a bin median.
    - Smoothing by bin boundaries
      - The minimum and maximum values in the bin are bin boundaries.

APEC

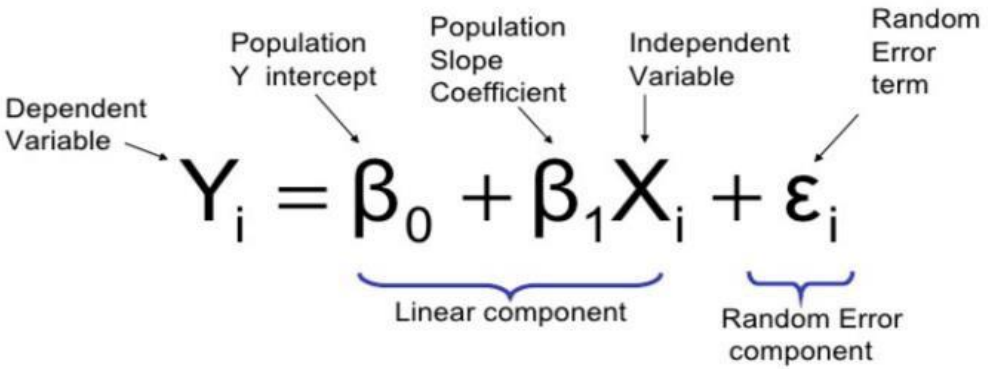

Sorted Data of price (in dollars):
4, 8, 15, 21, 21, 24, 25, 28, 34
Distribute the sorted data into bins
Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34
Smoothing By Means
Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29
Smoothing by Bin Boundaries
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34
Binning Methods for Smoothing Data

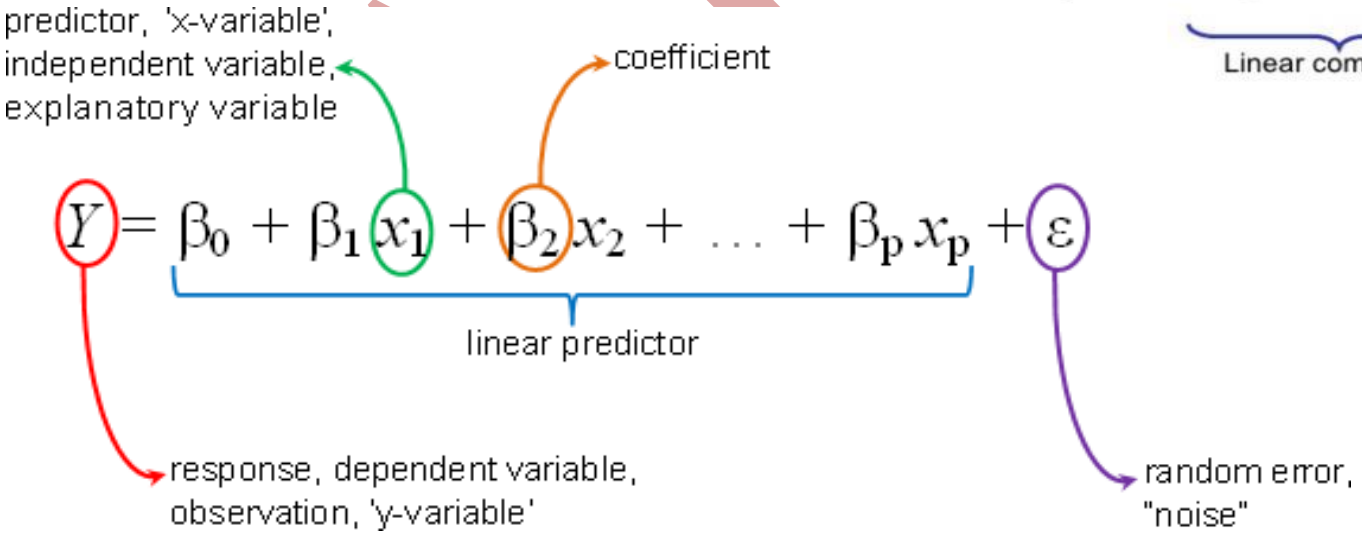


Regression

Statistical technique that relates a dependent variable to one or more independent (explanatory) variables.

Predicting the chance of winning, or predicting annual income of a person based on the education, age, profession etc.

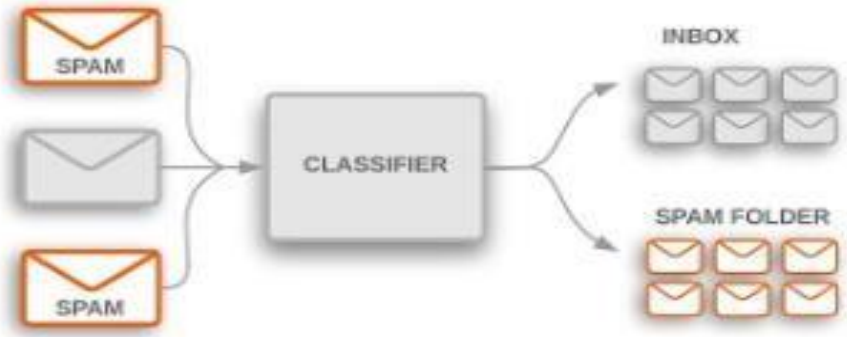


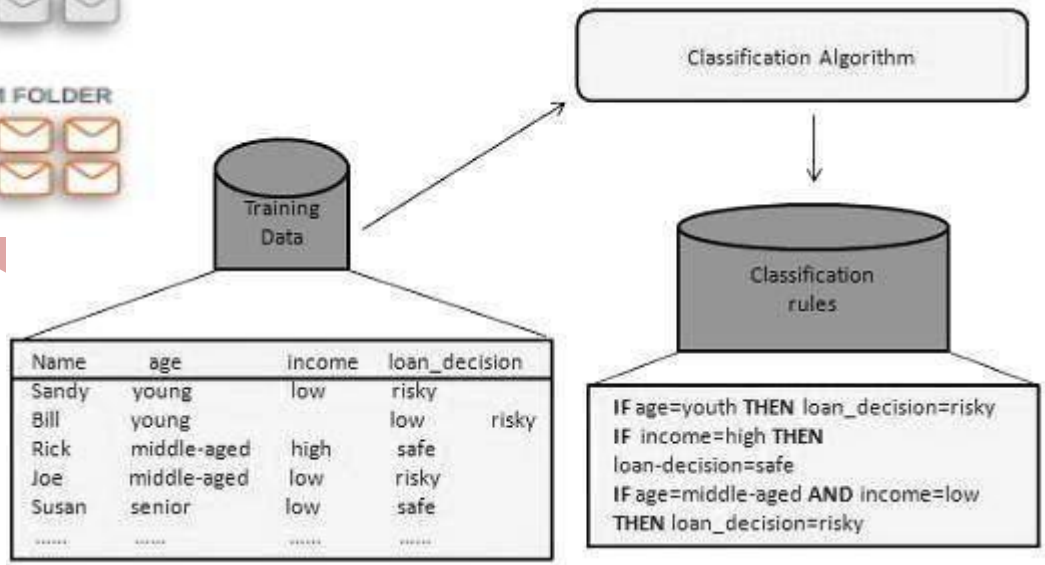


Classification is a process of categorizing a given set of data into classes.

What is Classification

Example :
Based on age, predicting an Adult.
if AGE >60 then
ADULT
else
YOUTH





Regression Vs Classification

Regression



What will be the temperature tomorrow?



Fahrenheit

Classification



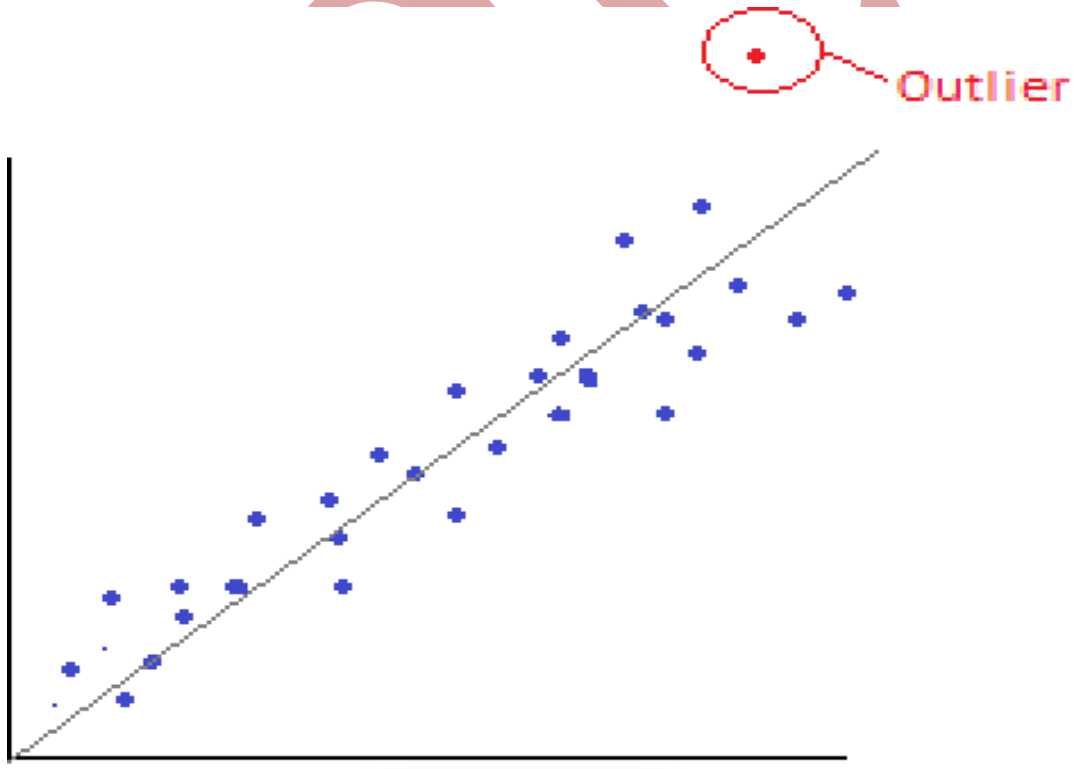
Will it be hot or cold tomorrow?



Fahrenheit

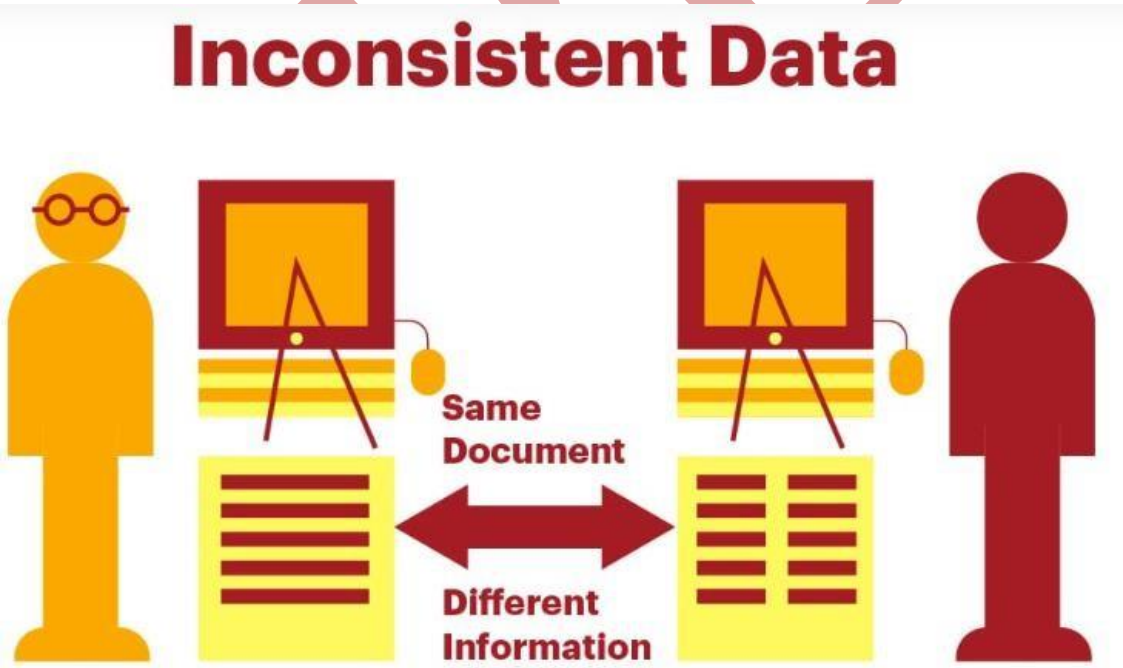
Finding the outliers

- An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution errors.



Data Inconsistency

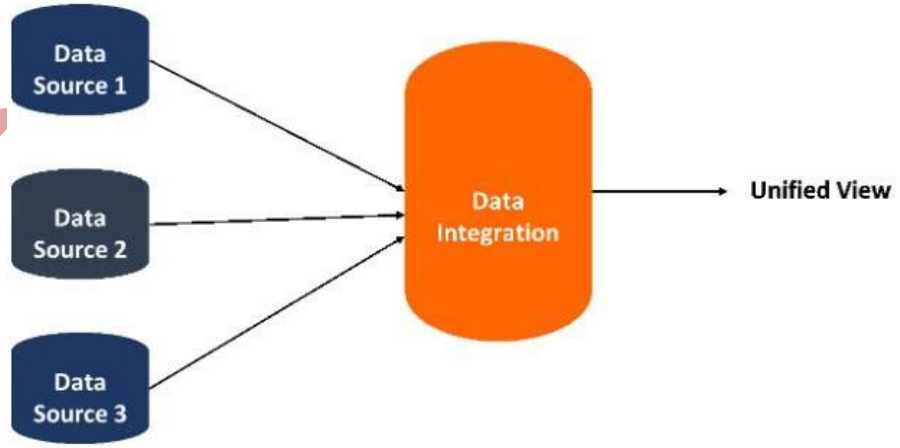
- When the same data exists in different formats in multiple tables.



Data Integration

- When **multiple heterogeneous data sources** such as databases, data cubes or files are combined for analysis, this process is called data integration.
  - This helps in improving the accuracy and speed of the data mining process.

Tools:
Oracle Data Integrator tool
SQL Server Integration Services



APEC

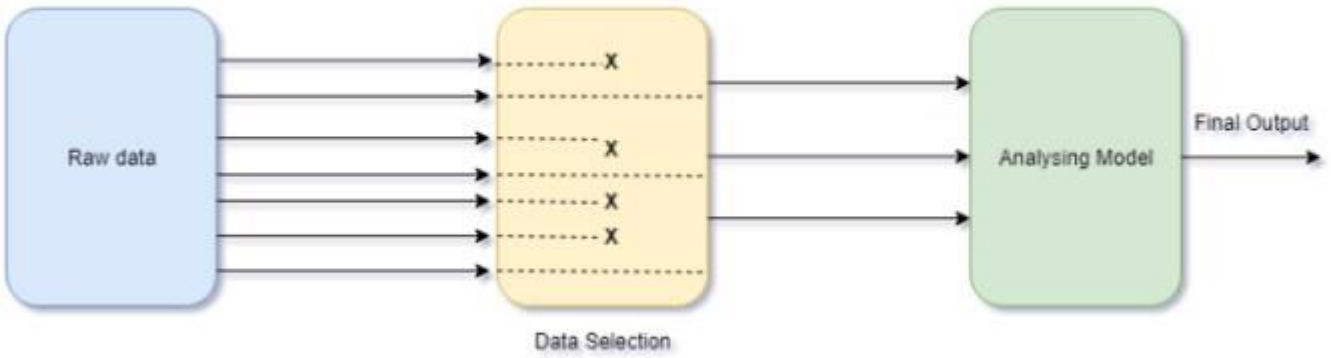
Data Reduction

- Data reduction is a process that reduces the volume of original data and represents it in a much smaller volume.
- **Main thing: Maintain Integrity while reducing data.**
- Some strategies of data reduction are:
  - **Dimensionality Reduction:** Reducing the number of attributes in the dataset.
  - **Numerosity Reduction:** Replacing the original data volume by smaller forms of data representation.
  - **Data Compression:** Compressed representation of the original data.

APEC

Dimensionality Reduction

Feature selection techniques



Feature selection process

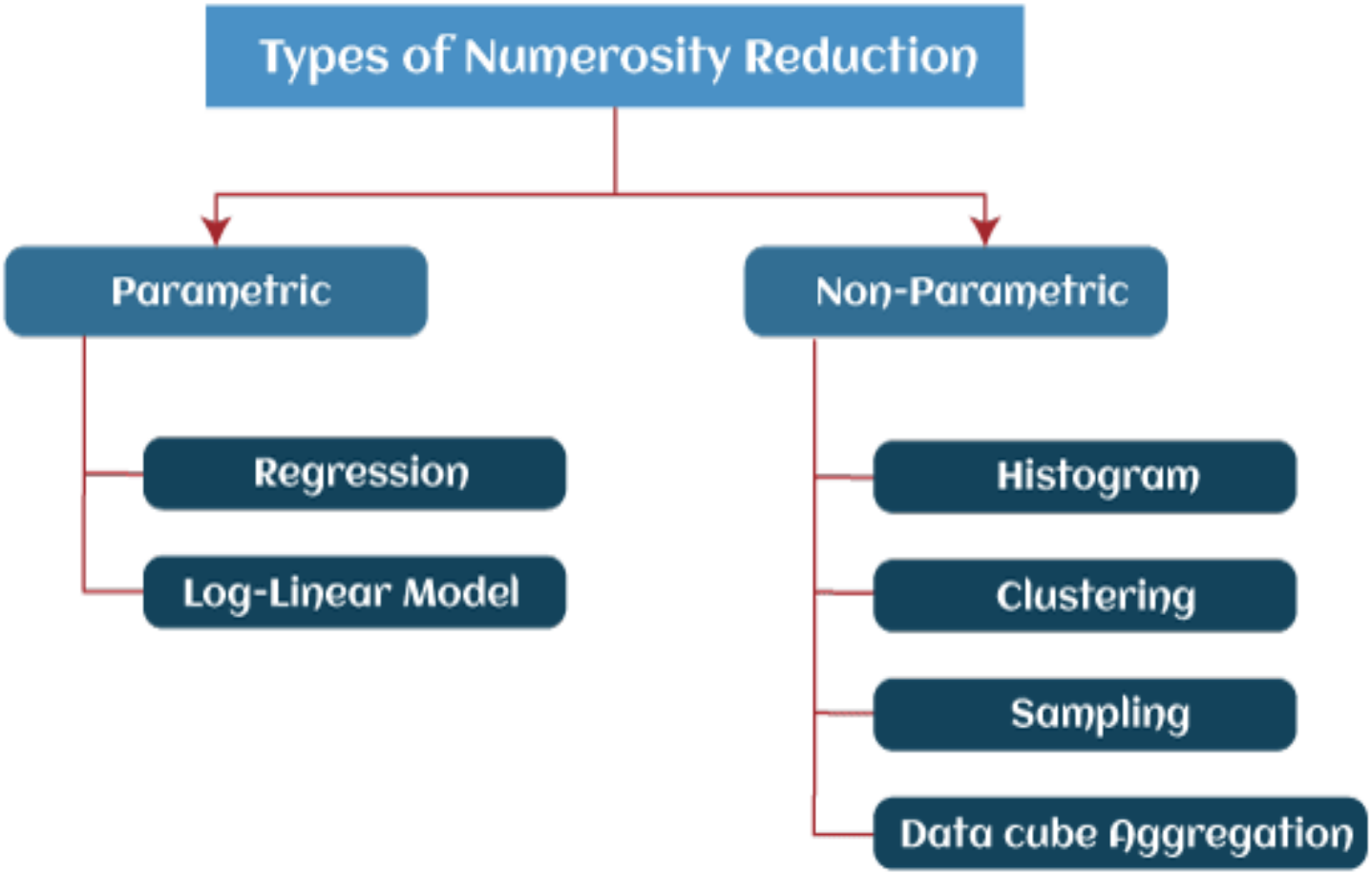
RollNo	Name	Mobile Number	Mobile Network
T4Tutorials1	Sameed	+92 302 XX XXX XX	Mobilink
T4Tutorials1	Ali	+92 333 XX XXX XX	Ufone

Figure 1: Before Dimension reduction

RollNo	Name	Mobile Number
T4Tutorials1	Sameed	+92 302 XX XXX XX
T4Tutorials1	Ali	+92 333 XX XXX XX

Figure 2: After Dimension reduction

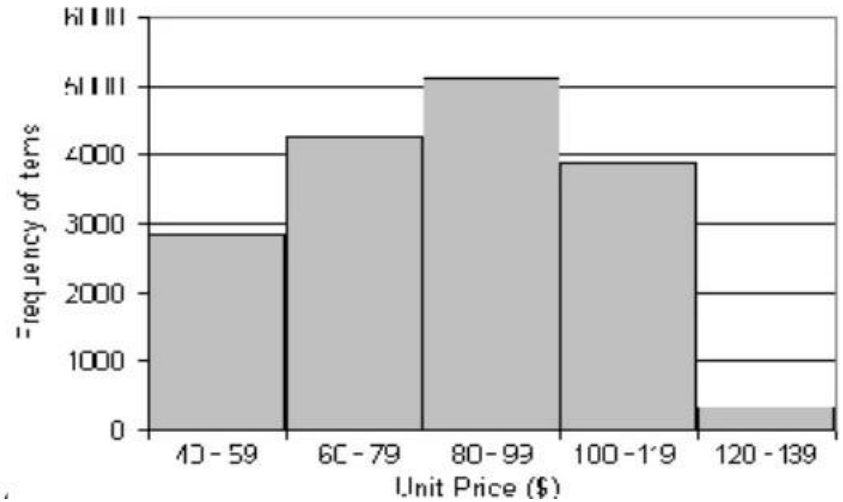
Numerosity Reduction



Histogram

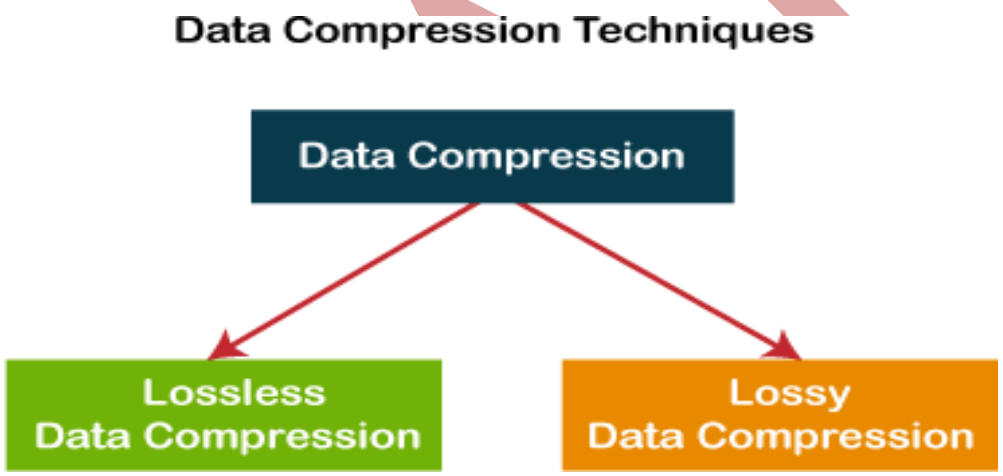
- A histogram is a graphical representation of data points organized into user-specified ranges.

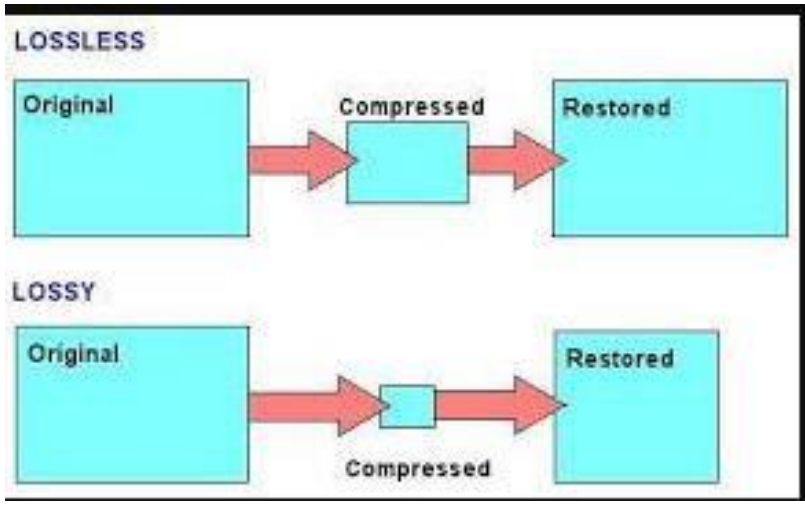
  - Frequency histograms
    - A univariate graphical method
    - Consists of a set of rectangles that reflect the counts or frequencies of the classes present in the given data



Data compression

- Data compression is the process of modifying data in order to reduce its size.
  - Two types of data compression techniques
    - Lossless Data Compression
    - Lossy Data Compression





Lossless data compression

- **Lossless data compression** is used to compress the files without losing an original file's quality and data.
  - In lossless data compression, file size is reduced, but the quality of data remains the same.
  - Advantage
    - We can restore the original data in its original form after the decompression.
  - Lossless data compression mainly used in the sensitive documents, confidential information.

APEC

Lossy data compression

- Lossy data compression is used to compress larger files into smaller files.
  - In this compression technique, some specific amount of **data and quality are removed (loss) from the original file**
    - This technique is generally useful for us when the quality of data is not our first priority.

APEC

Data Transformation

- Data transformation is a technique used to convert the raw data into a suitable format which is needed for data mining process

- **Strategies for data transformation are:**
  - **Smoothing:** Removing noise from data using clustering, regression techniques, etc.
  - **Aggregation:** Summary operations are applied to data.
  - **Normalization:** Scaling of data to fall within a smaller range.
  - **Discretization:** Raw values of numeric data are replaced by intervals. **For Example,** Age.

APEC

Data Mining process

- **Data Mining**
  - Data Mining is a process to identify interesting patterns and knowledge from a large amount of data.
    - In these steps, intelligent patterns are applied to extract the data patterns.
    - The data is represented in the form of patterns and models are structured using classification and clustering techniques.
- **Pattern Evaluation**
  - This step involves identifying interesting patterns representing the knowledge based on interestingness measures. D
    - Data summarization and visualization methods are used to make the data understandable by the user.
- **Knowledge Representation**
  - Knowledge representation is a step where data visualization and knowledge representation tools are used to represent the mined data.
    - Data is visualized in the form of reports, tables, etc.

APEC

KDD
(Knowledge Discovery in Databases)

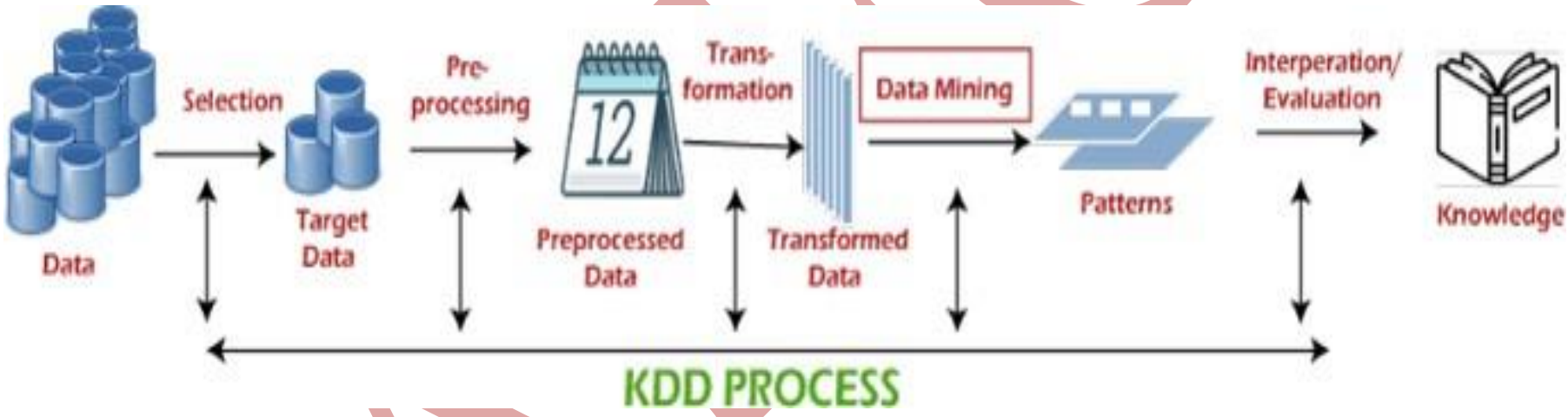
APEC

KDD
(Knowledge Discovery in Databases)

- KDD (Knowledge Discovery in Databases) is a field of computer science, which includes the tools and theories to help humans in extracting useful and previously unknown information (i.e., knowledge) from large collections of digitized data.
  - KDD consists of several steps, and Data Mining is one of them.

APEC

KDD process


Selection
Pre-
processing
Trans-
formation
Data Mining
Interperation/
Evaluation
Data
Target
Data
Preprocessed
Data
Transformed
Data
Patterns
Knowledge
KDD PROCESS

- **Goal identification**:Identify the KDD process's goal from the customer perspective.
- **Creating a target data set:** Selecting the data set or or data samples on which the discovery to be done.
- **Data cleaning and preprocessing:** Basic operations like removing noise, handling missing data fields etc.
- **Data reduction and projection:** Finding useful features to represent the data depending on the purpose of the task.
- **Matching process objectives:** KDD with step 1 a method of mining particular. For example, summarization, classification, regression, clustering, and others.
- **Modeling and exploratory analysis and hypothesis selection:** Choosing the algorithms and selecting the methods to search for data patterns.
  - This process includes deciding which model and parameters may be appropriate and the matching of data mining methods for KDD

APEC

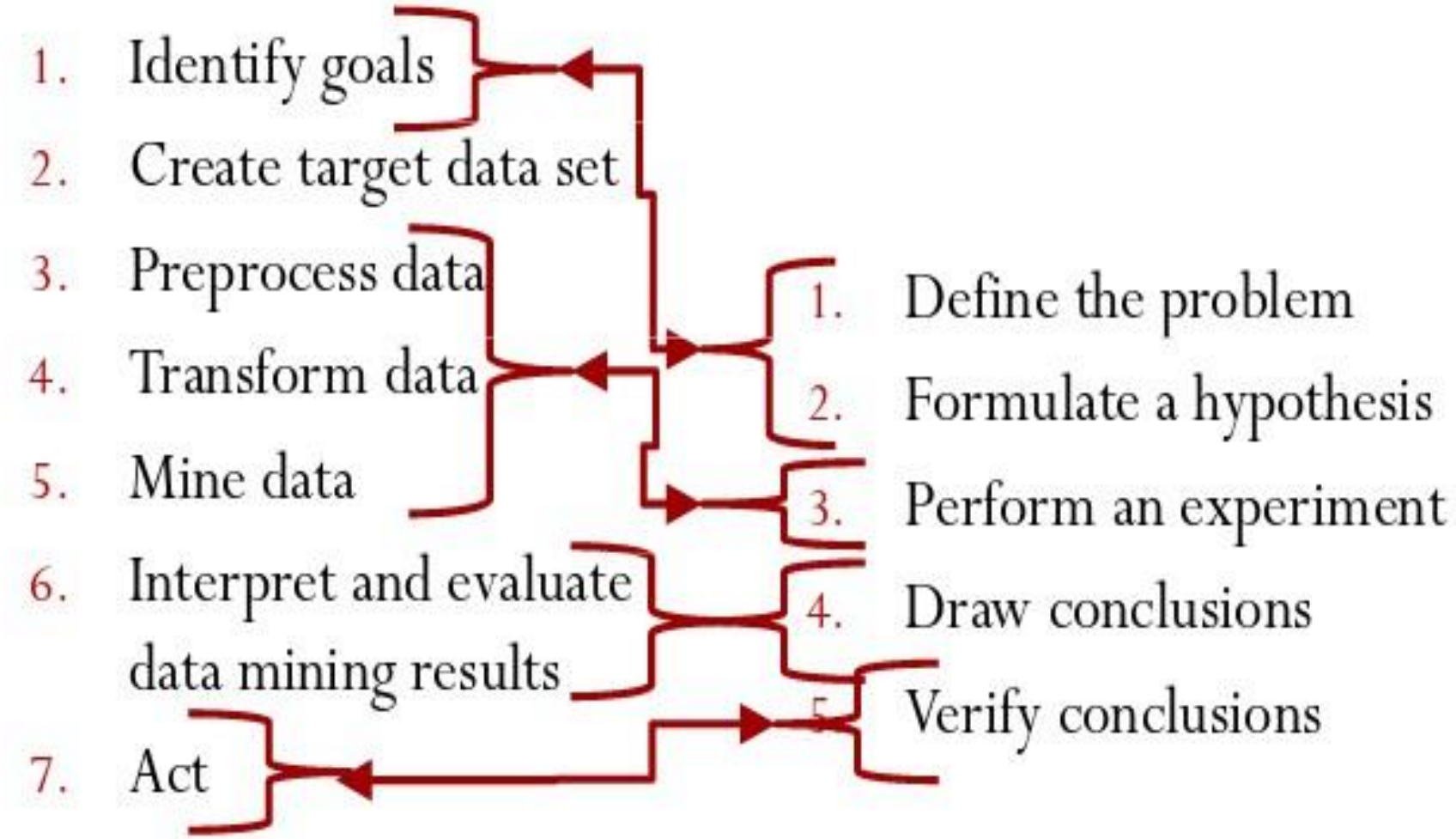
- **Data Mining:** The search for patterns of interest in a particular representational form like classification rules or trees, regression, and clustering.

- **Presentation and evaluation:** Interpreting mined patterns. This step may involve the visualization of the extracted patterns, models.

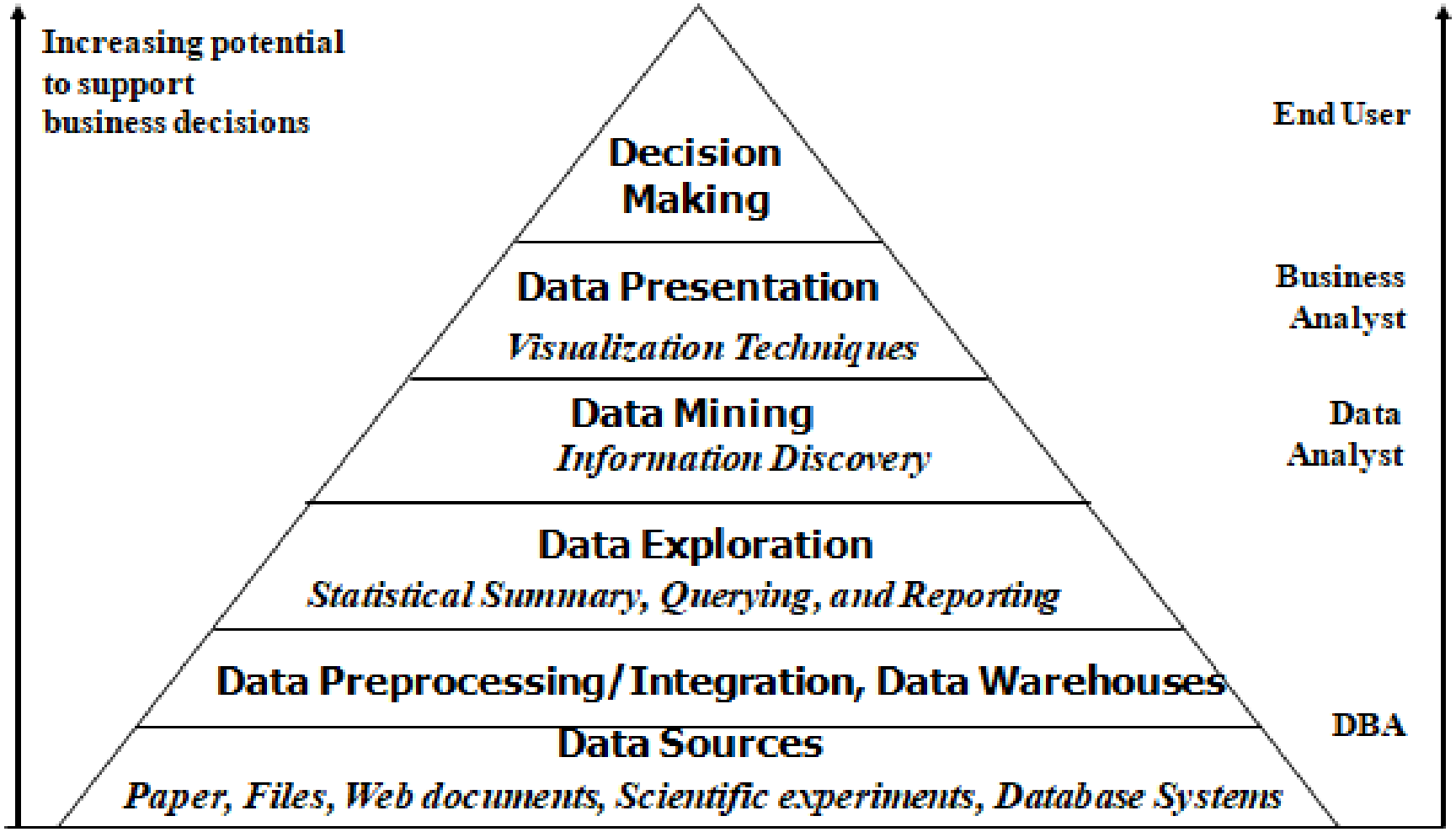
- **Taking action on the discovered knowledge:** Using the knowledge directly, incorporating the knowledge in another system for further action, or simply documenting and reporting to stakeholders.

APEC

7-Step KDD Process



Example: Data Mining in Business Intelligence



What type of data can be mined

- Flat Files
- Relational Databases
- DataWarehouse
- Transactional Databases
- Multimedia Databases
- Spatial Databases
- Time Series Databases
- World Wide Web(WWW)

APEC

What Kinds of Patterns Can Be Mined?

- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks.
  - It can be classified into two categories: descriptive and predictive.

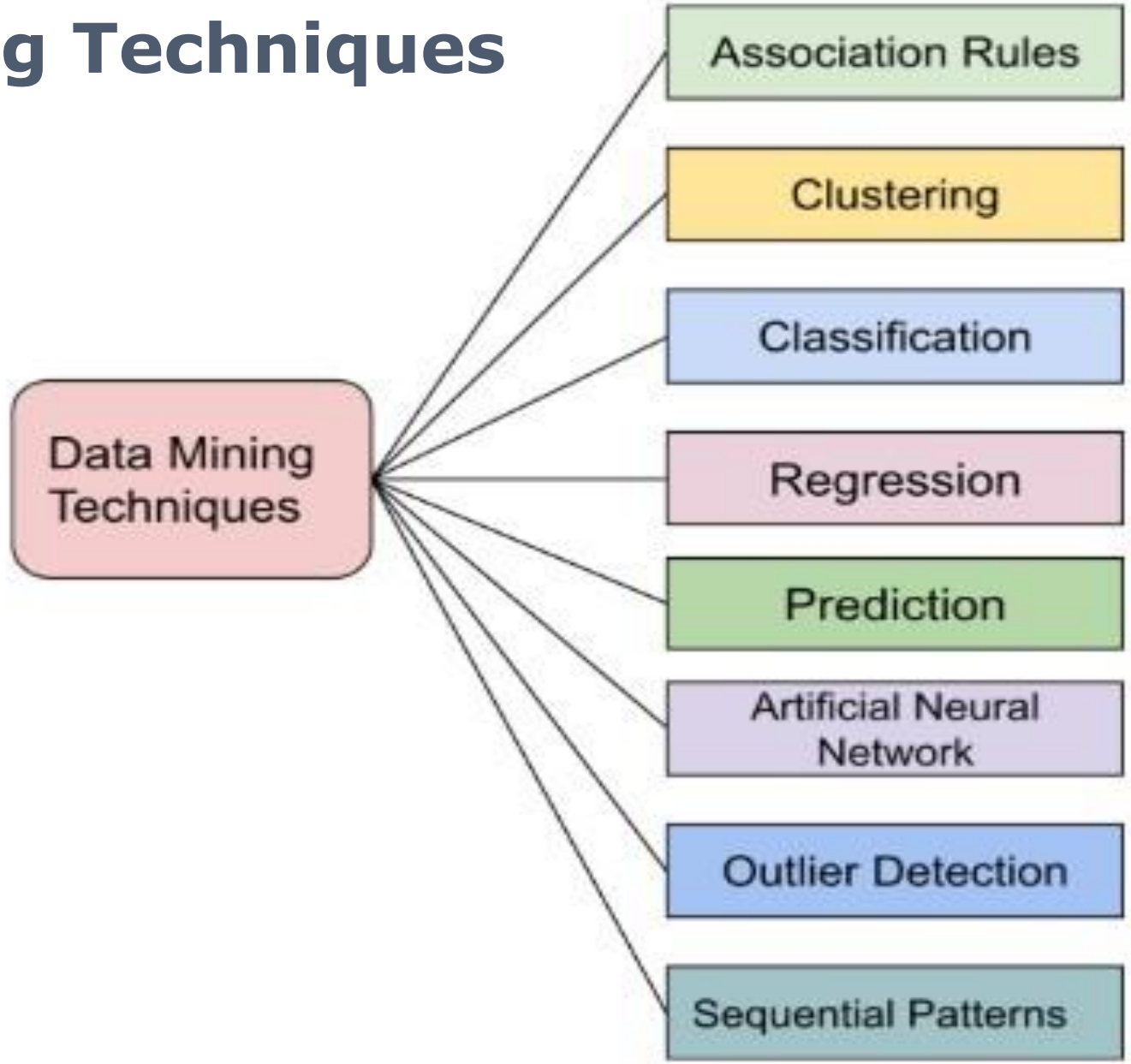
- **Descriptive Mining**, as the name implies,

  "describes" the data.
  - You convert the data into a human-readable format once it has been collected.

- **Predictive Data Mining** is the Analysis done to predict a future event or other data or trends, as the term 'Predictive' means to predict something.
  - Business Analysts can use Predictive Data Mining to make better decisions

APEC

Data Mining Techniques



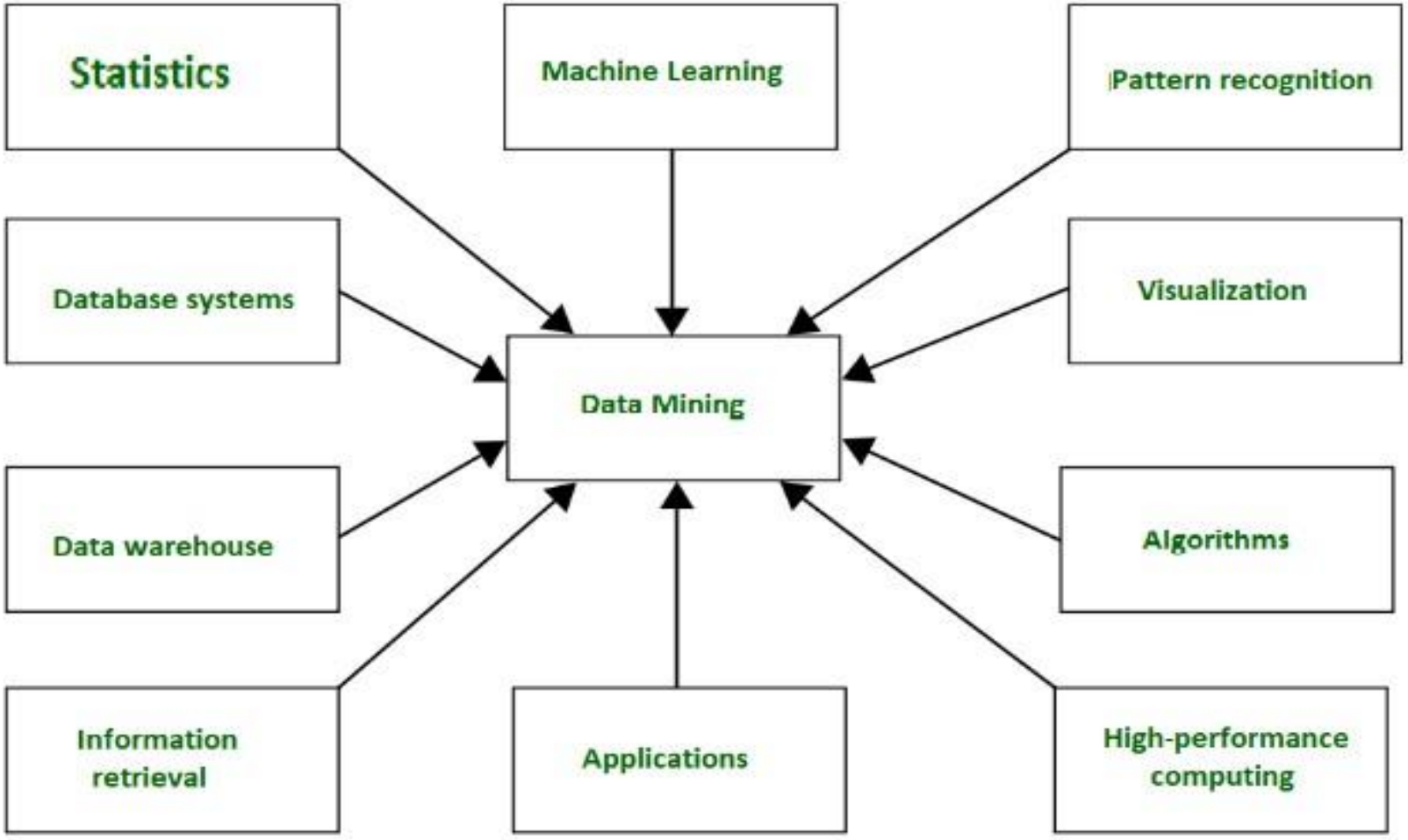
- **Association Rules**
  - Identify relationship between two or more things.
    - Example: “If a customer buys bread, he is 70% likely to buy jam”
- **Classification**
  - Allocates objects in a dataset to desired groups or classes.
- **Clustering**
  - It detects similar data. Clustering helps in grouping similar data together.
- **Regression analysis**
  - The term "regression" refers to a data mining approach for predicting numeric values in a data set..
- **Prediction**
  - The prediction uses several other data mining techniques, and so on, to forecast a future event.
- **Artificial Neural Network**
  - ANN handles information in the same manner that the human brain does.
- **Outlier Detection**
  - It identify data elements that do not match an expected pattern or behavior.

APEC

- **Sequential Patterns**
  - Discover sequential patterns by examining sequential data.

APEC

Data mining has been integrated with many other disciplines

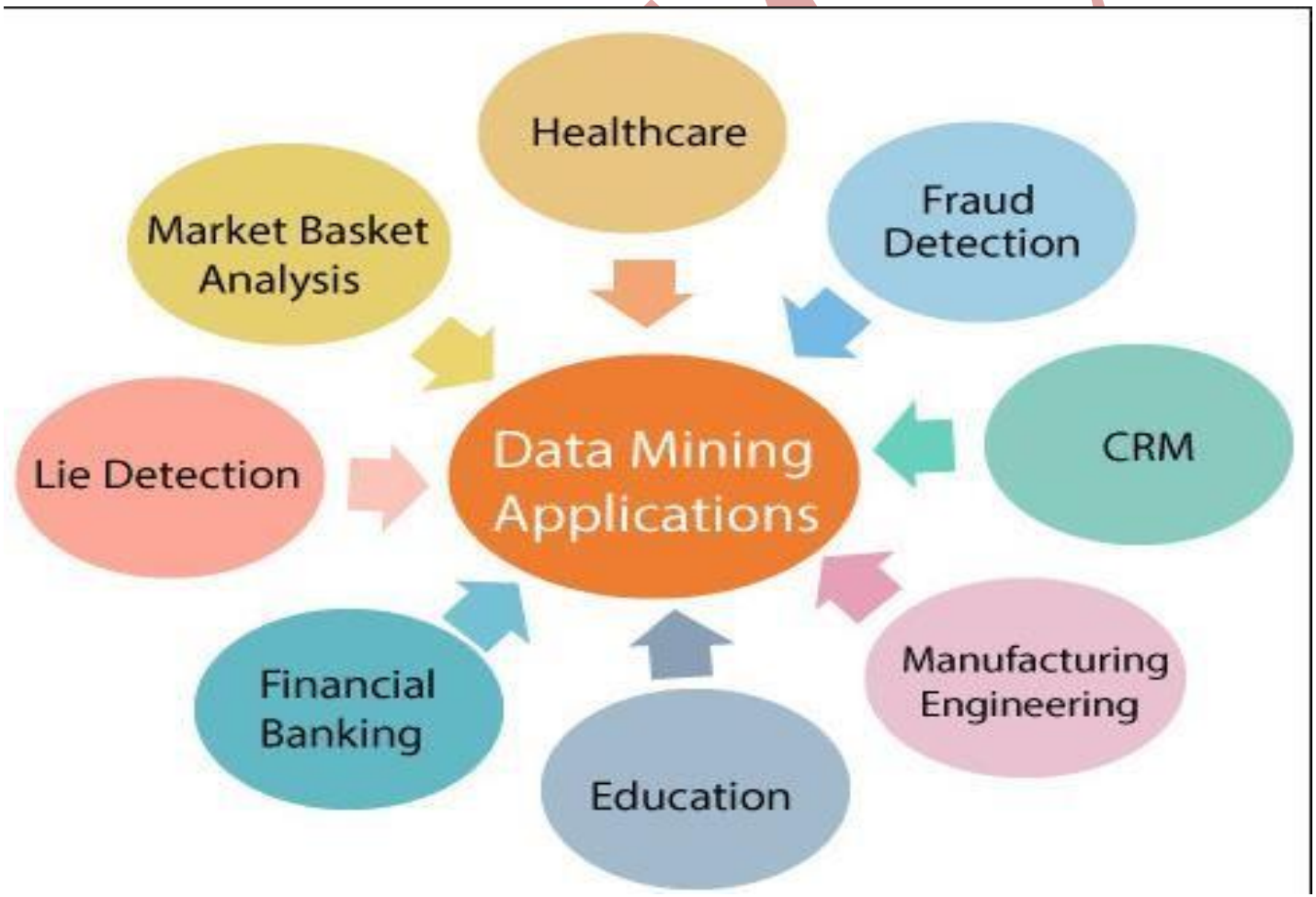


Contd…

- **Statistics** studies the collection, analysis, explanation, and presentation of data.
- **Machine learning** investigates how computers can learn (or improve their performance) based on data.
- **Database systems** research focuses on the creation, maintenance, and use of databases for organizations and end- users.
- **A data warehouse** integrates data originating from multiple sources and various timeframes.
- **Information retrieval (IR)** is the science of searching for documents or information in documents.

APEC

Data Mining Applications



- **Data Mining in Healthcare:**
  - Data mining in healthcare has excellent potential to improve the health system. It uses data and analytics for better insights and to identify best practices that will enhance health care services and reduce costs.
- **Data Mining in Market Basket Analysis:**
  - Market basket analysis is a modeling method based on a hypothesis. If you buy a specific group of products, then you are more likely to buy another group of products. This technique may enable the retailer to understand the purchase behavior of a buyer..
- **Data Mining in Manufacturing Engineering:**
  - Data mining tools can be beneficial to find patterns in a complex manufacturing process. It can also be used to forecast the product development period, cost, and expectations among the other tasks.

AEC

Data mining in Education:

- Education data mining is a newly emerging field, concerned with developing techniques that explore knowledge from the data generated from educational Environments.
  - An organization can use data mining to make precise decisions and also to predict the results of the student. With the results, the institution can concentrate on what to teach and how to teach.

Data Mining in CRM (Customer Relationship Management):

- Customer Relationship Management (CRM) is all about obtaining and holding Customers, also enhancing customer loyalty and implementing customer-oriented strategies.
  - To get a decent relationship with the customer, a business organization needs to collect data and analyze the data. With data mining technologies, the collected data can be used for analytics.

ABC

Data Mining in Fraud detection:

- Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are a little bit time consuming and sophisticated. An ideal fraud detection system should protect the data of all the users.

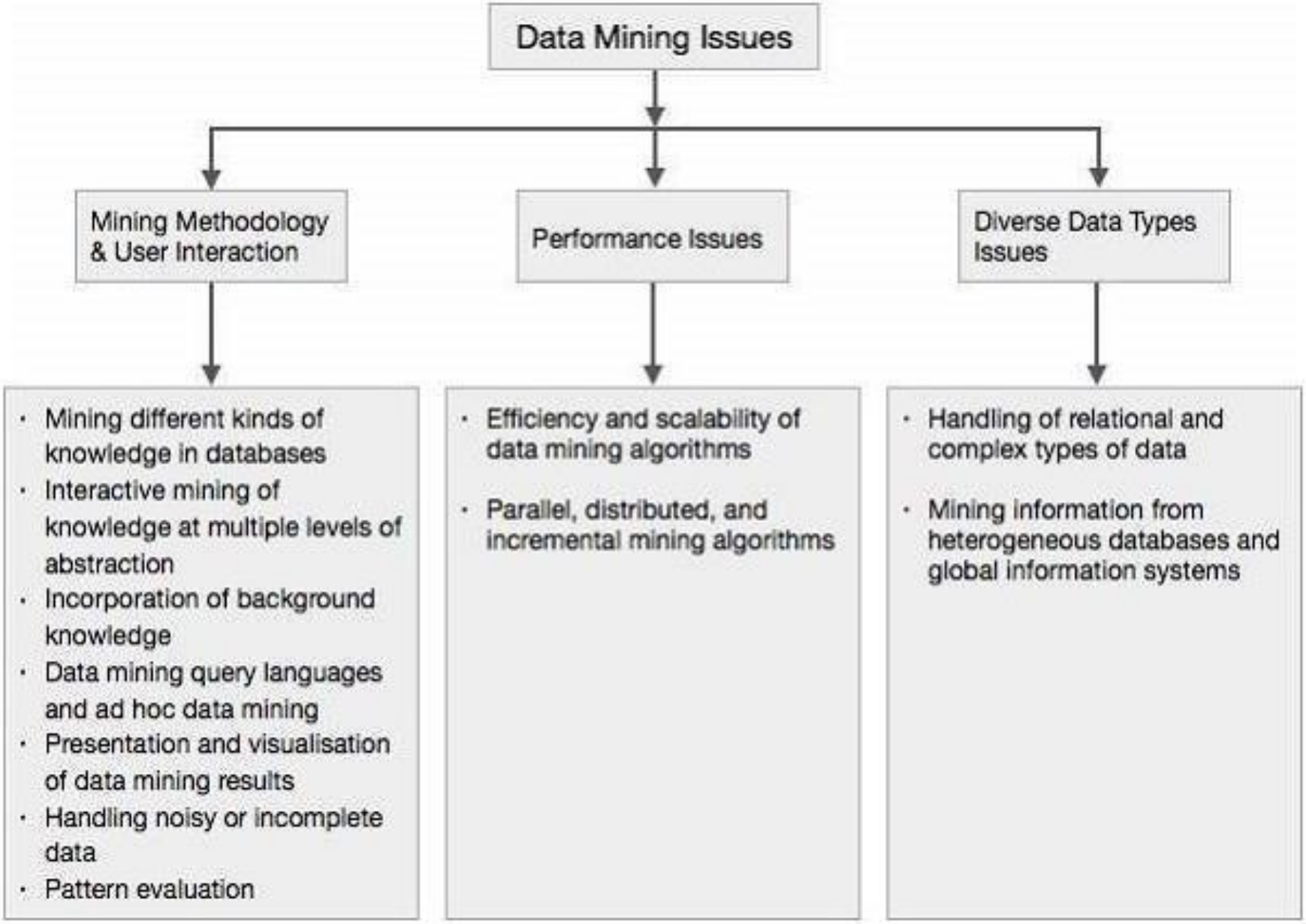
Data Mining in Lie Detection:

- Law enforcement may use data mining techniques to investigate offenses, monitor terrorist communications, etc.

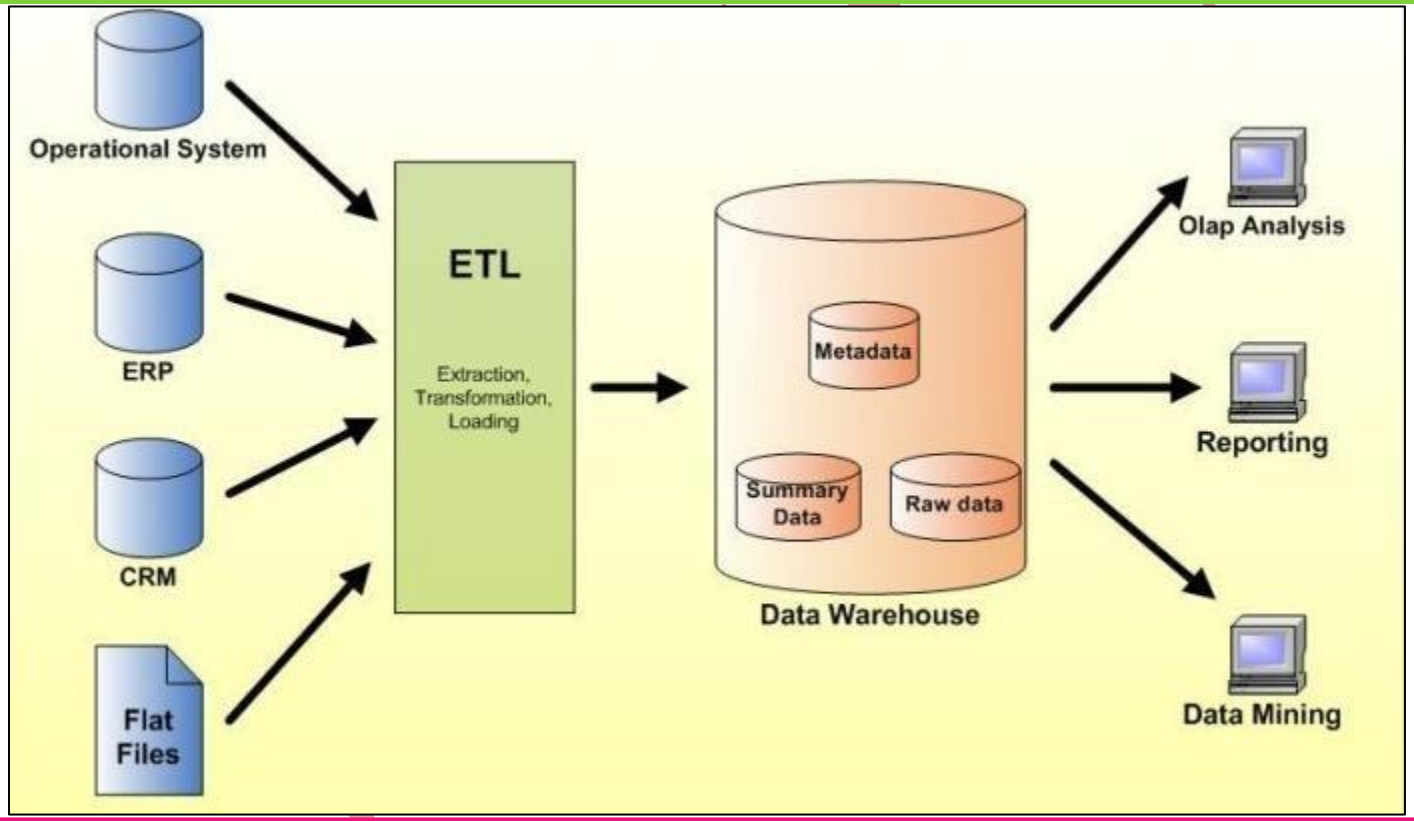
Data Mining Financial Banking:

- The Digitalization of the banking system is supposed to generate an enormous amount of data with every new transaction.
  - The data mining technique can help bankers by solving business-related problems in banking and finance by identifying trends, casualties, and correlations in business information and market costs

APEC


Data Mining Issues
Mining Methodology & User Interaction
Performance Issues
Diverse Data Types Issues
Mining different kinds of knowledge in databases
Interactive mining of knowledge at multiple levels of abstraction
Incorporation of background knowledge
Data mining query languages and ad hoc data mining
Presentation and visualisation of data mining results
Handling noisy or incomplete data
Pattern evaluation
Efficiency and scalability of data mining algorithms
Parallel, distributed, and incremental mining algorithms
Handling of relational and complex types of data
Mining information from heterogeneous databases and global information systems

Data Warehouse



- Data Warehousing (DW) is process for collecting and managing data from varied sources to provide meaningful business insights.

Operational Database	Data Warehouse
Relational databases are created for on-line transactional Processing (OLTP)	Data Warehouse designed for on-line Analytical Processing (OLAP)
Less Number of data accessed.	Large Number of data accessed.
Operational systems are usually concerned with current data.	Data warehousing systems are usually concerned with historical data.
It is designed for real-time business dealing and processes.	It is designed for analysis of business measures by subject area, categories, and attributes.
It is optimized for a simple set of transactions, generally adding or retrieving a single row at a time	It is optimized for high, complex queries that access many rows
It supports thousands of concurrent clients.	It supports a few concurrent clients relative to OLTP.
Operational systems are widely process-oriented.	Data warehousing systems are widely subject-oriented

APJ

Operational systems are usually optimized to perform fast inserts and updates of small volumes of data.	Data warehousing systems are usually optimized to perform fast retrievals of high volumes of data.

APEC

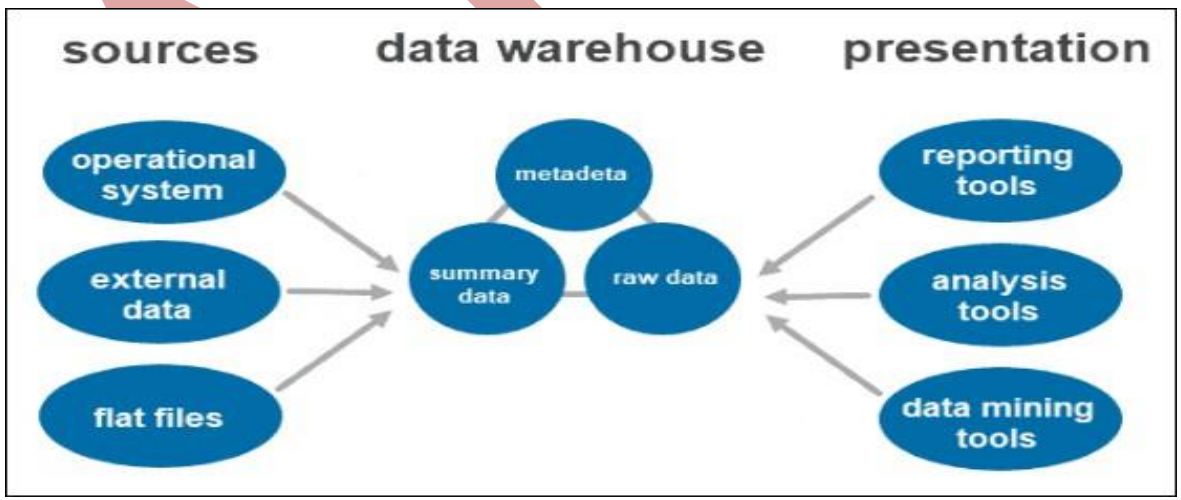
Data Warehouse Architecture

- There are three ways to construct a data warehouse system. These approaches are classified by the number of tiers in the architecture. They are:
  - Single-tier architecture
  - Two-tier architecture
  - Three-tier architecture

APEC

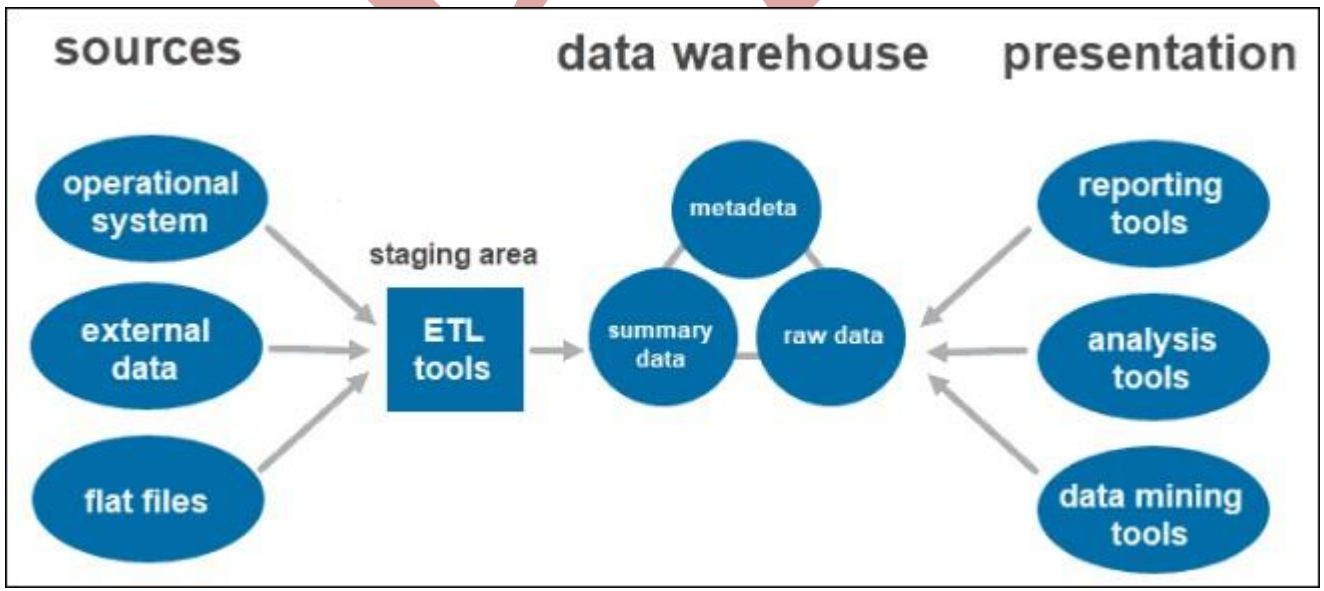
Single-tier Data Warehouse Architecture

- The single-tier architecture is not a frequently practiced approach.
  - The main goal is to remove redundancy by minimizing the amount of data stored.
- Disadvantage:
  - Doesn't have a component that separates analytical and transactional processing.



Two-tier Data Warehouse Architecture

- A two-tier architecture includes a staging area for all data sources, before the data warehouse layer.
  - By adding a staging area between the sources and the storage repository, we ensure all data loaded into the warehouse is cleansed and in the appropriate format.

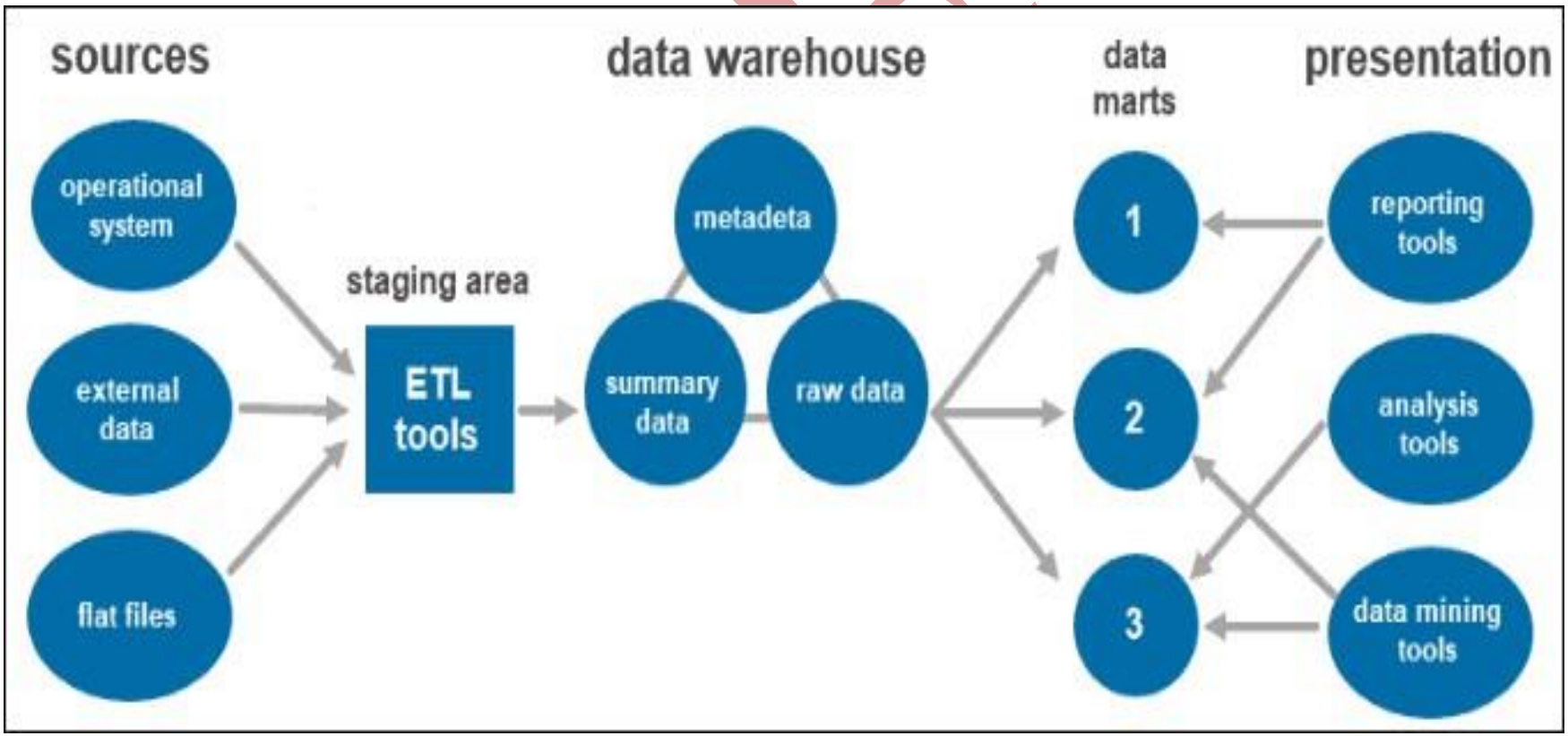


Three-tier Data Warehouse Architecture

- The three-tier approach is the most widely used architecture for data warehouse systems. Essentially, it consists of three tiers:
  - **The bottom tier** is the database of the warehouse, where the cleansed and transformed data is loaded.
  - **The middle tier** is the application layer giving an abstracted view of the database. It arranges the data to make it more suitable for analysis. This is done with an OLAP server.
  - **The top-tier** is where the user accesses and interacts with the data. It represents the front-end client layer. It uses reporting tools, query, analysis or data mining tools.

APEC

Three-tier Data Warehouse Architecture – Contd…



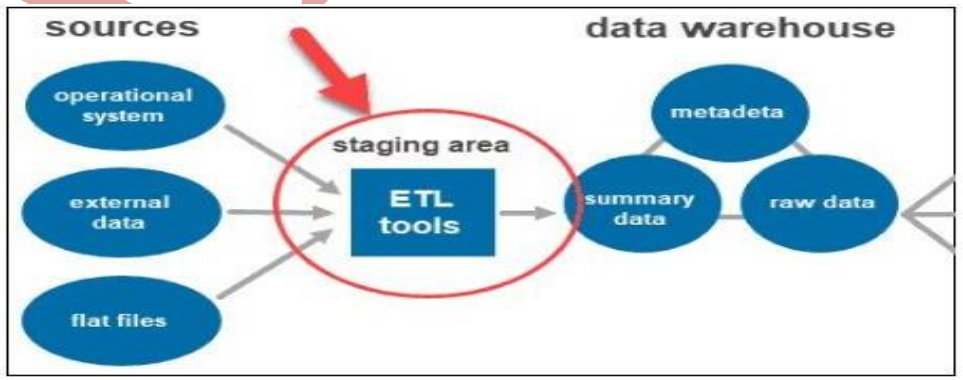
Data Warehouse Components

- ETL Tools
- Database
- Data
- Access Tools

APEC

ETL Tools

- The data coming from the data source layer can come in a variety of formats.
- ETL stands for **Extract**, **Transform**, **Load**.
  - The staging layer uses **ETL tools** to extract the needed data from various formats and checks the quality before loading it into the data warehouse.



Database

- The most crucial component and the heart of each architecture is the database.
  - The warehouse is where the data is stored and accessed.
- Four types of databases are:
  - **Relational databases** (row-centered databases).
  - **Analytics databases** (developed to sustain and manage analytics).
  - **Data warehouse applications** (software for data management and hardware for storing data offered by third-party dealers).
  - Cloud-based databases (hosted on the cloud).

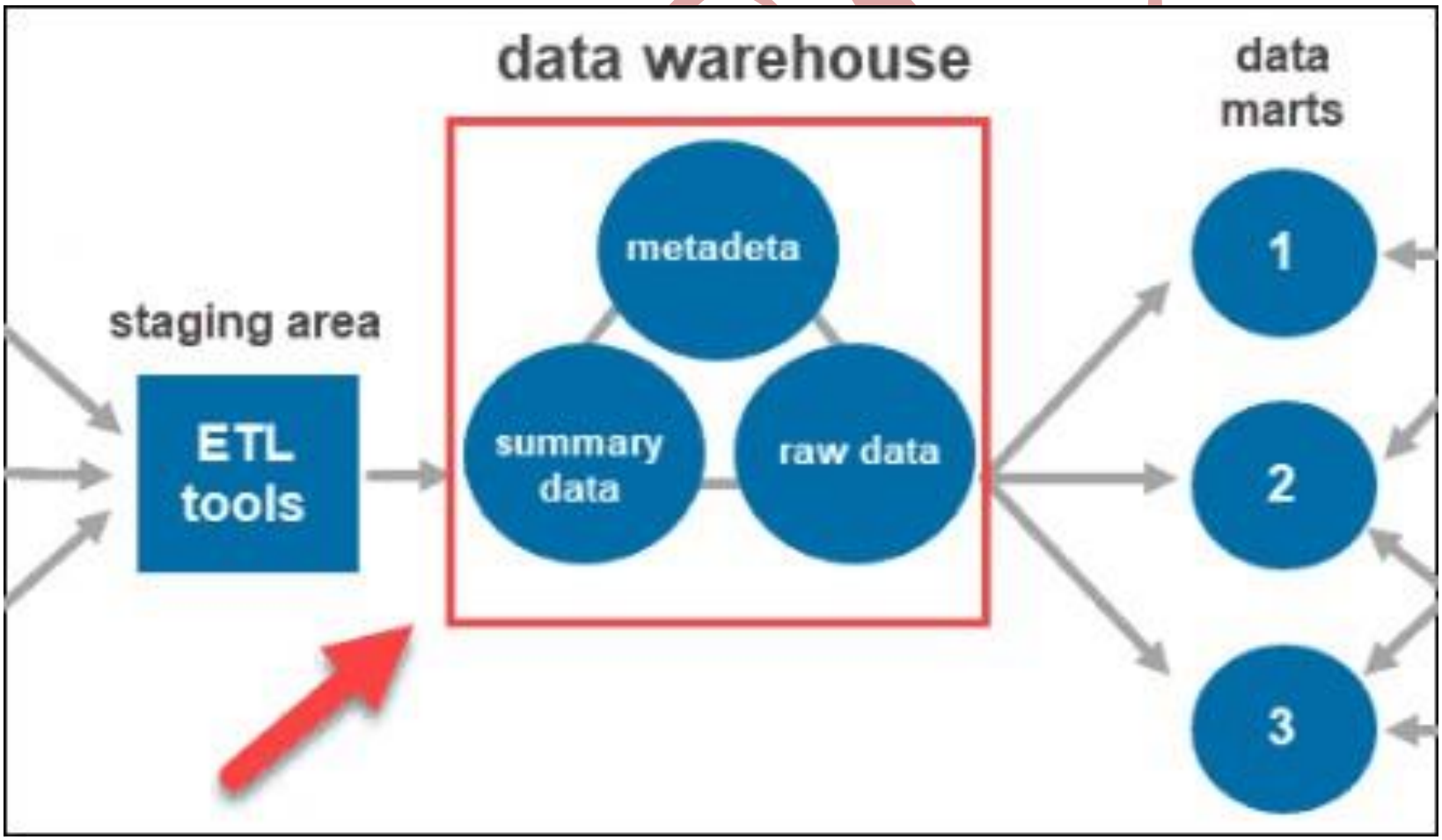
APEC

Data

- Once the system cleans and organizes the data, it stores it in the data warehouse.
  - The data warehouse represents the central repository that stores metadata, summary data, and raw data coming from each source.
  - **Metadata** is the information that defines the data. It allows data analysts to classify, locate, and direct queries to the required data.
  - **Summary data** is generated by the warehouse manager. It updates as new data loads into the warehouse. This component can include lightly or highly summarized data.
  - **Raw data** is the actual data loading into the repository, which has not been processed. Raw data needs further processing and analysis.

APEC

Data Contd…


data warehouse
data marts
metadeta
staging area
ETL tools
summary data
raw data
1
2
3

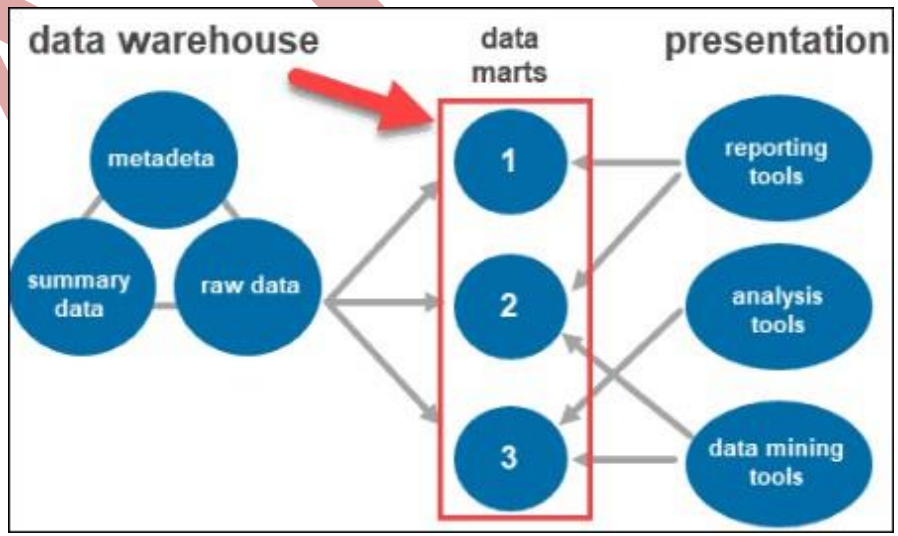
Access Tools

- Users interact with the gathered information through different tools and technologies.
  - They can analyse the data, gather insight, and create reports.
- Some of the tools used include:
  - **Reporting tools.**
    - Reporting tools include visualizations such as graphs and charts showing how data changes over time.
  - **OLAP tools.**
    - Online analytical processing tools which allow users to analyze multidimensional data from multiple perspectives.
  - **Data mining tools.**
    - Examine data sets to find patterns within the warehouse and the correlation between them.

APEC

Data Marts

- Data marts allow you to have multiple groups within the system by segmenting the data in the warehouse into categories.
  - It partitions data, producing it for a particular user group.
    - Ex: Sales group, finance group, account group.



Properties of Data Warehouse Architectures

- **1. Separation:**
  - Analytical and transactional processing should be keep apart as much as possible.
- **2. Scalability:**
  - Hardware and software architectures should be simple to upgrade the data volume,
- **3. Extensibility:**
  - The architecture should be able to perform new operations and technologies without redesigning the whole system.
- **4. Security:**
  - Monitoring accesses are necessary because of the strategic data stored in the data warehouses.
- **5. Administerability:**
  - Data Warehouse management should not be complicated.

APEC