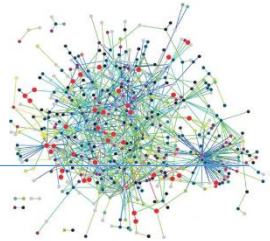
Data Mining and Ware housing

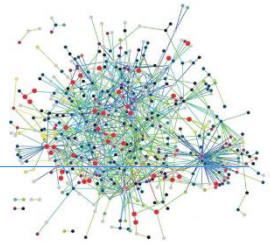
Er. Jeewan Rai

Terminology



- **Data** : Raw piece of information
- **Information**: refined form of data
- **Knowledge**: application of **information;** awareness or understanding on the subject acquired from education or experience of a person.
- **Database**: organized collection of such data
- **Data warehouse**: organizes all the data available in an organization
- **Data Mining**: discovering meaningful patterns and trends often previously unknown
- **Data mart**: meet the particular demands of a specific group of users within the organization
- **Metadata**: summarizes basic information about data

The Foundations of Data Mining

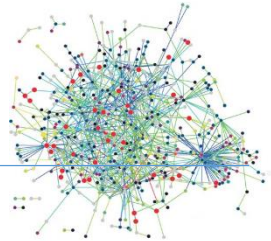


Data Mining and Ware housing

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective (looking back), static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases (RDBMS), Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	prospective (expecting in future) and proactive/ active information delivery

Approaches of Data Mining



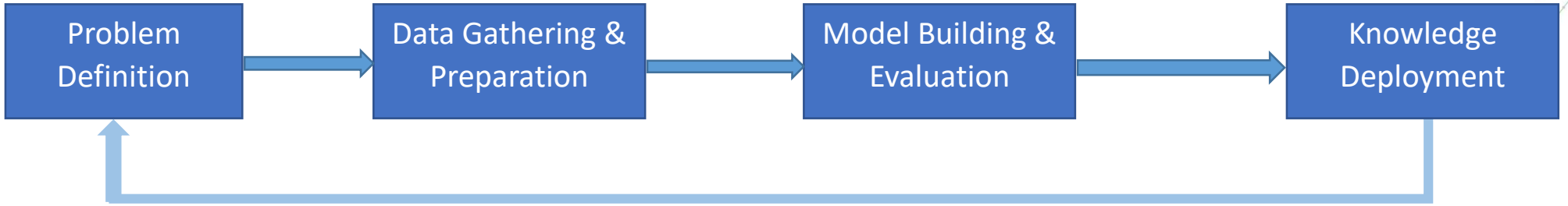
1) **Predictive Data Mining:** Prediction of trends and behaviours

- Targeted marketing: Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings.
- Forecasting economic failure and other forms of default, and identifying segments of a population likely to respond similarly to given events.

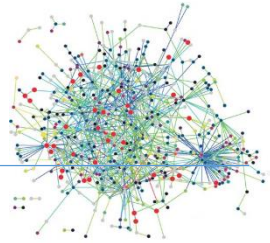
2) Descriptive Data Mining: Discovery of previously unknown patterns

- Analysis of retail sales data, to identify likely unrelated products that are often purchased together.
- Detecting fraudulent/ fake credit card transactions and identifying anomalous/ abnormal data that could represent data entry keying errors.
- It characterizes the general properties of data in the database.
- It finds patterns in data the user determinants which ones are important

Data Mining Process

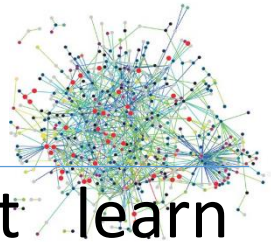


- **Problem Definition:** project objectives and requirements. Business and data understanding
- **Data Gathering and Preparation:** need to be selected, cleaned, constructed and formatted
- **Model Building and Evaluation: model** for the prepared dataset.
- **Knowledge Deployment:** needs to be presented in such a way that stakeholders can use it when they want it



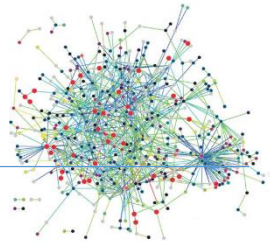
Data Mining Vs. Query Tools

Techniques used in data mining



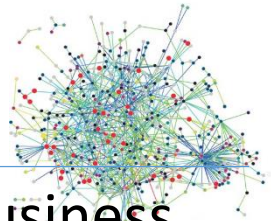
- **Artificial Neural Networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- **Decision Trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. E.g. Classification
- **Genetic Algorithms**: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the ***concepts of evolution***.
- **Nearest Neighbour Method**: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset. Sometimes called the k-nearest neighbour technique.
- **Rule Induction**: The extraction of useful if-then rules from data based on statistical significance.

Major issues in Data Mining

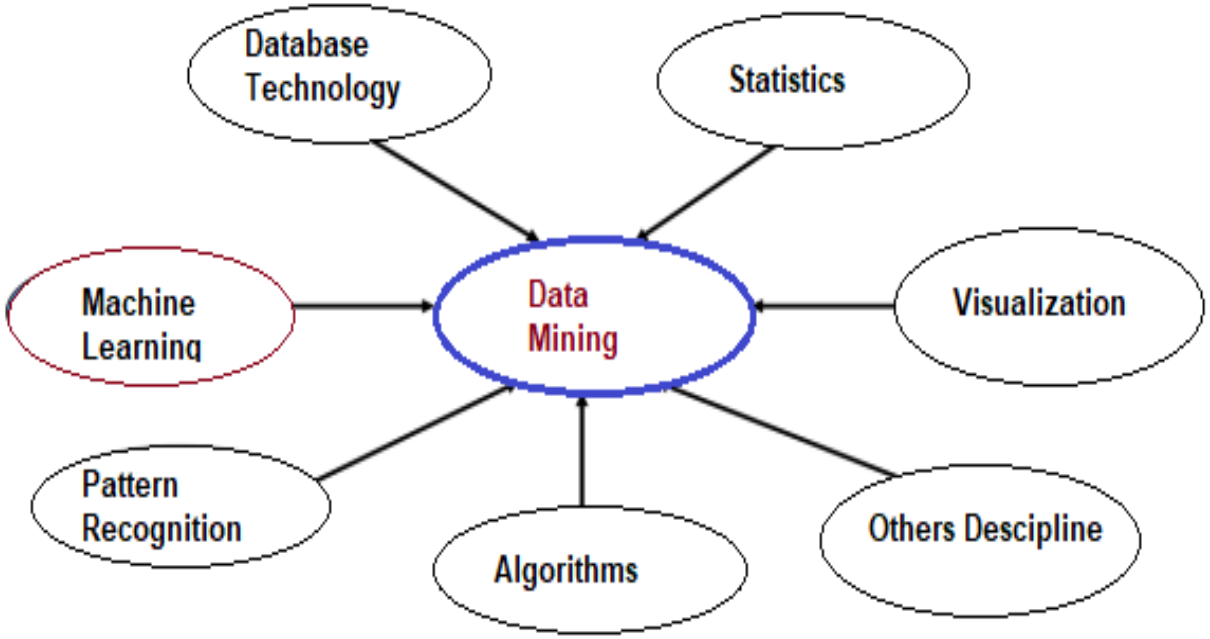


- **Mining Methodology**
  - Mining various and new kinds of knowledge; Mining knowledge in multidimensional space
  - Data Mining-an interdisciplinary effort
  - Boosting the power of discovery in a networked environment; Handling uncertainty, noise, or incompleteness of data
  - Patter evaluation and pattern or constraint-guided mining
- **User Interaction**
  - Interactive mining: dynamically change the focus of a search; Incorporation of background knowledge
  - Ad hoc data mining and data mining query languages
  - Presentation and visualization of data mining results
- **Efficiency and Scalability**
  - Efficiency and scalability of data mining algorithms; Parallel, distributed and incremental mining algorithms
- **Diversity of Database types**
  - Handling complex types of data; Mining dynamic, networked and global data repositories
- **Data Mining and Society**
  - social impacts of data mining: should address individual privacy and data protection rights
  - privacy-preserving data mining: to observe data sensitivity and preserve people's privacy while performing successful data mining
  - invisible data mining: hidden data mining algorithms e.g. web search engines

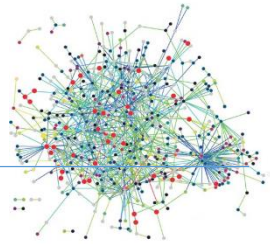
Applications of Data Mining



- **BI (Business Intelligence):** provides historical, current and predictive views of business operations. E.g. reporting, OLAP, business performance management, competitive intelligence, predictive analytics.
- **Web Search Engines**: searches information from web by crawling, indexing and searching techniques.
- **A medicine company**
- **Credit card company**
- **Transportation company**
- **Supermarket or CRM**
- **Intrusion Detection**
- **Telecommunication Industry**
- **Medical: Biological Data Analysis**
- **Financial Data Analysis**
- **Retail Industry**



Characteristics of Data Warehouse



A data warehouse should be…

- **Time – Frame:** information can then be sourced according to period.
- **Non-Volatile:** never updated, but used only for queries.
- **Subject Oriented:** Organized around major subjects, such as customer, product, sales.
- **Integrated:** Constructed by integrating multiple, heterogeneous data sources

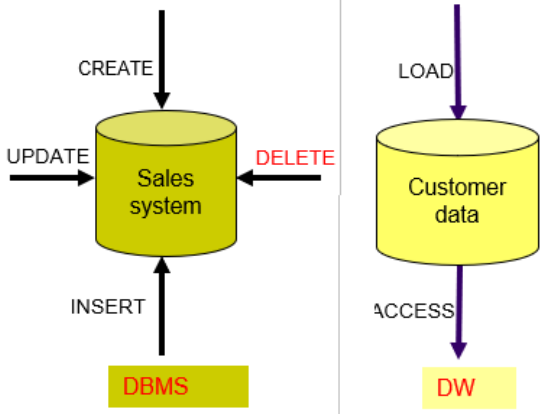


Fig. Non-Volatile

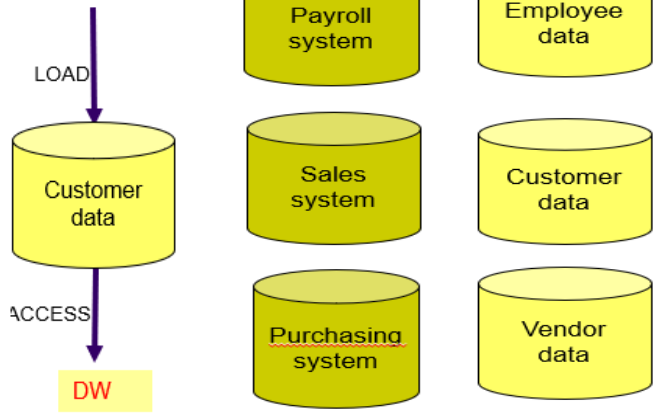


Fig. Subject-Oriented

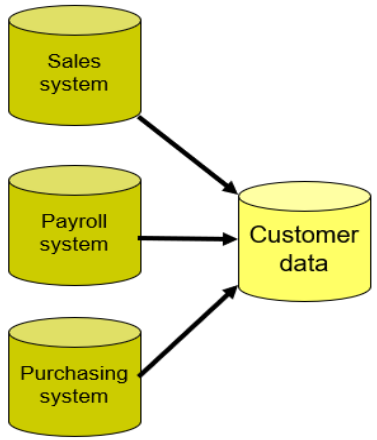
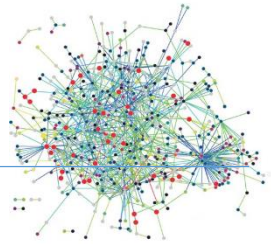
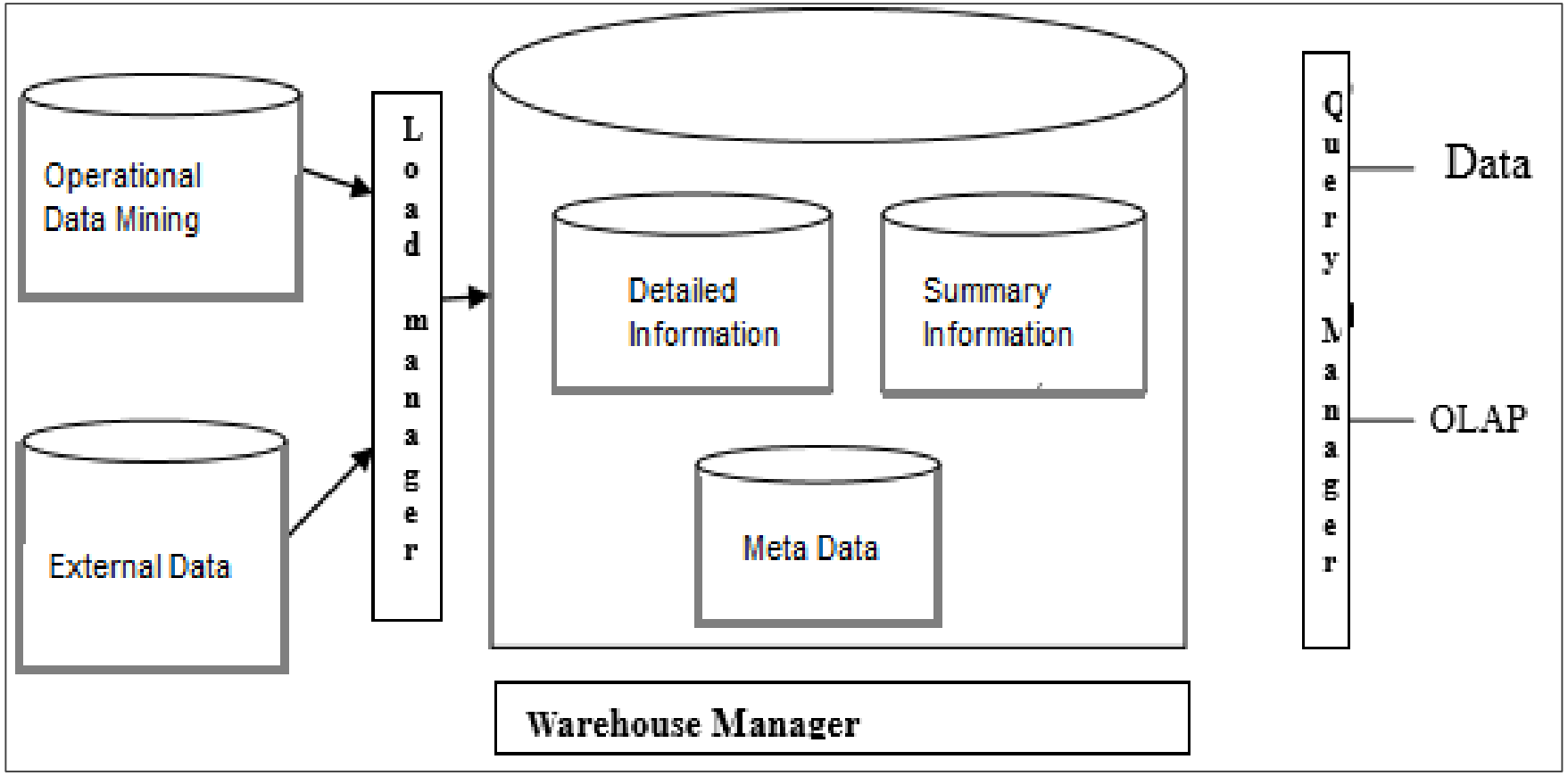


Fig. Integrated

Architecture of a Data Warehouse



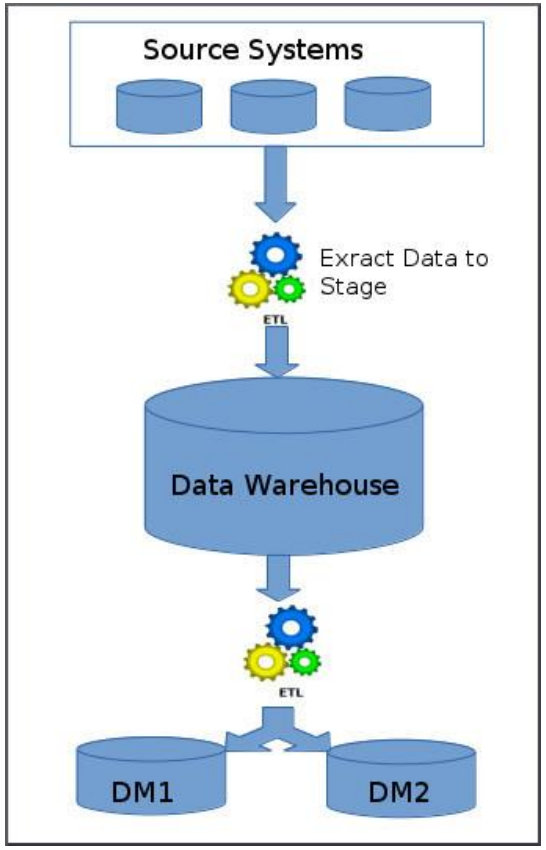


Detailed Information: Data warehouse schemas; Fact data; Dimension data; Partitioning data

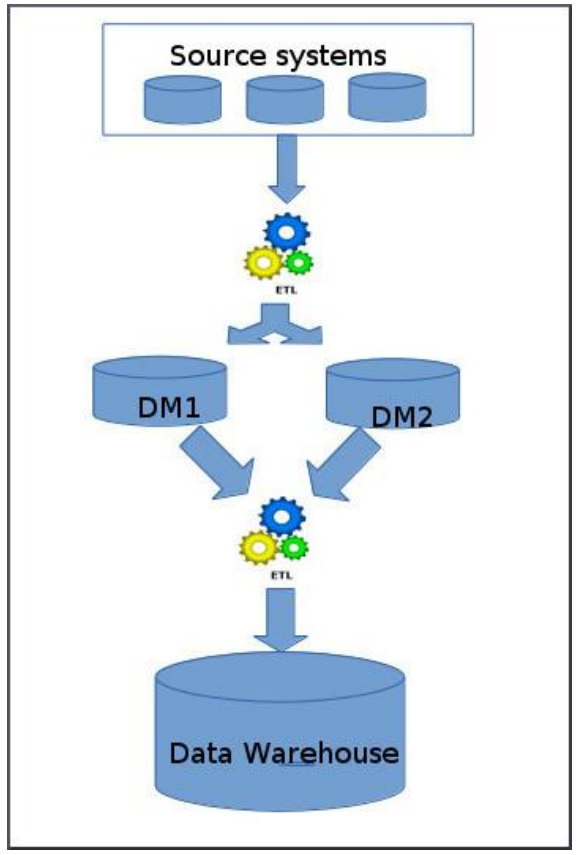
Summary Information: predefined aggregations e.g. performance, operation cost

Approaches to Data Warehousing

1) **Top-Down Approach**: *First DW is built then data is loaded into Data Mart*
2) **Bottom-Up Approach:** *First Data Marts are created, then data is loaded into DW*

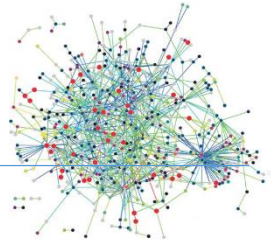


Fig(a): Top-Down Approach



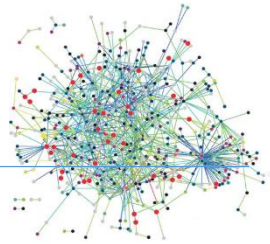
Fig(b): Bottom-Up Approach

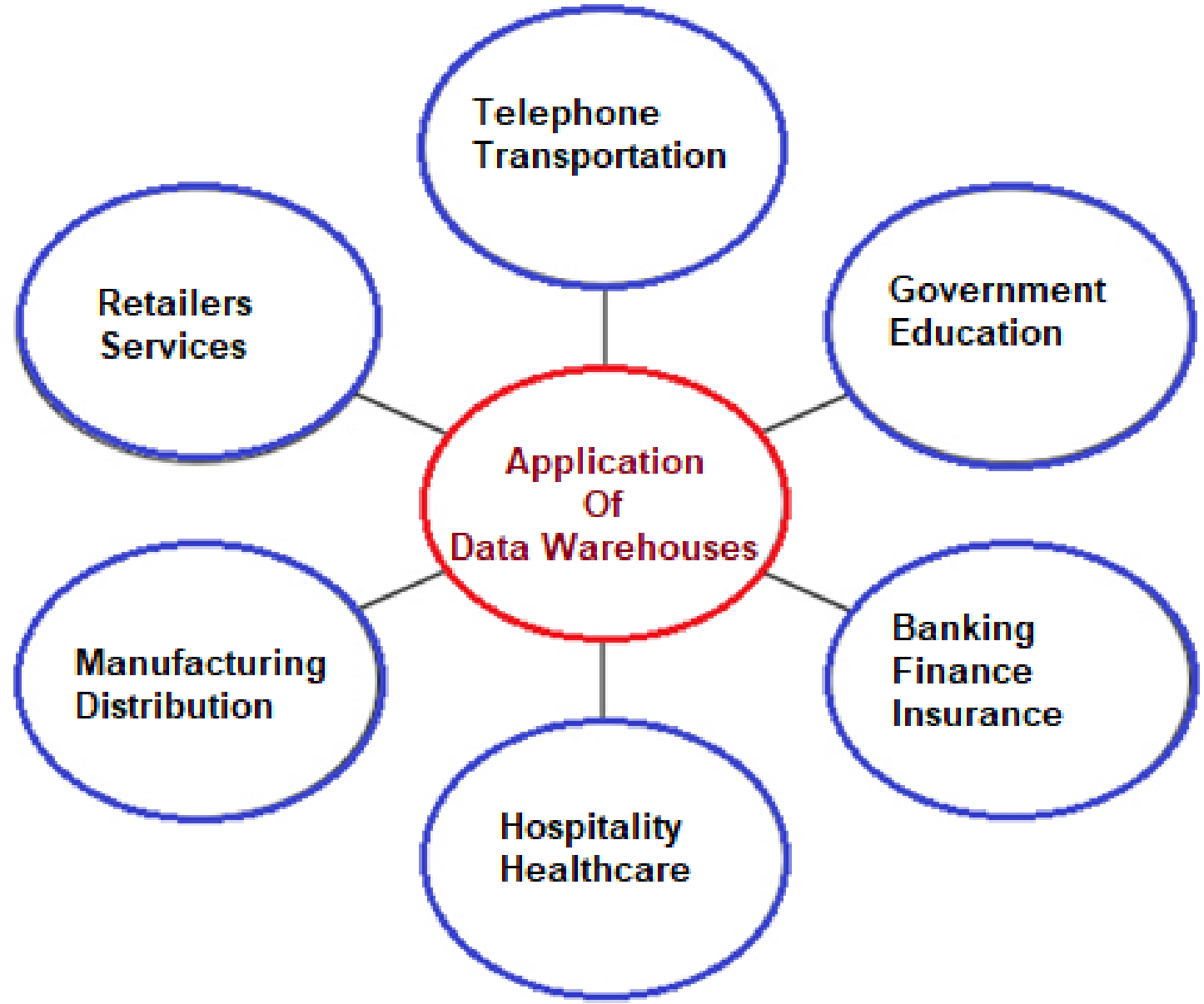
Discussion



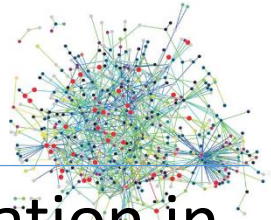
- Data Warehouse vs Data Mart
- Data Warehouse vs Database

Application of Data Warehouse



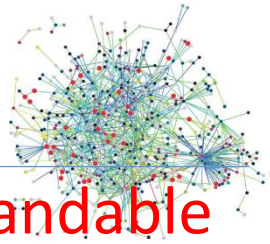


Data Pre-Processing: Terminology



- **Population** - the collection of all individuals or items under consideration in a statistical study
- **Sample** - that part of the population from which information is collected
- **Parameter** – statistical description of the population
- **Variable** – characteristic that varies from one item to another e.g. Quantitative (numerical) Qualitative (categorical)
- **Data**: Observing the values of the variables yield data
- **Observation** – individual piece of data
- **Data matrix** – collection of observations for variable Data matrix k variables measured in sample with the size of n

Data Pre-Processing



Data mining technique that involves transforming raw data into an understandable format

Data Pre-processing is required because: Real world data are generally:

* **Incomplete:** Missing - attribute values, certain attributes of importance, or having only aggregate data **e.**g. occupation = “”

* **Noisy:** Containing errors or outliers e.g., Age=“-10”

* **Inconsistent/ Varying:** Containing discrepancies in codes or names e.g., Age=“42” Birthday=“03/07/1997” , discrepancy between duplicate records

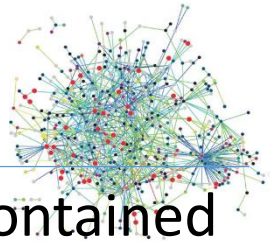
Why Is Data Pre-processing Important?

- **No quality** data, no quality mining results! (garbage in garbage out!) : Quality decisions must be based on quality data

e.g., duplicate or missing data may cause incorrect or even misleading statistics.

- **Data preparation, cleaning, and transformation** comprises the majority of the work in a data mining application

Data Types and Attributes



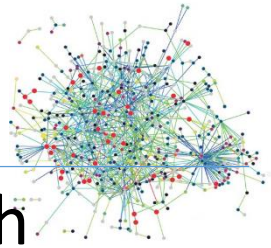
Data Mining and Ware housing

Attribute type can be compound, list or whatever, D**ata type** is the type of data contained in that data structure

Data Type	Used for	Example
String	Alphanumeric characters	hello world, Alice, Bob123
Integer	Whole numbers	7, 12, 999
Float (floating point)	Number with a decimal point	3.15, 9.06, 00.13
Character	Encoding text numerically	97 (in ASCII, 97 is a lower case 'a')
Boolean	Representing logical values	TRUE, FALSE

Approach	Attribute Type	Used for	Example
1	Quantitative	Can measure	Height, width
	Qualitative	Characteristics and Description	Smells, taste, color
2	Nominal	By names	Gender: Male, Female
	Ordinal	In order object	Grade: A,B,C
	Binary	Two options	Result: Yes, No
	Interval	Equal intervals	Calendar dates, time on clock
	Ratio	Same as interval but 0 does not exist	Age 0 does not exits
3	Discrete	Distinct can take only certain value	No. of student in class; Gender:M,F
	Continuous	Data can take any value	Time in race; temperature
4	Character	String value	Male, Good
	Number	Numeric value	1, 5.5

Dataset Types



Data Mining and Ware housing

1) **Record** - Data that consists of a collection of records, each of which consists of a fixed set of attributes

a) **Data Matrix**: same fixed set of numeric attributes; represented by an m by n matrix

b) **Document Data:** text documents: term-frequency vector

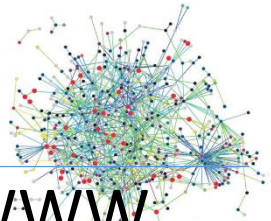
c) **Transaction Data** - each record (transaction) involves a set of items.

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

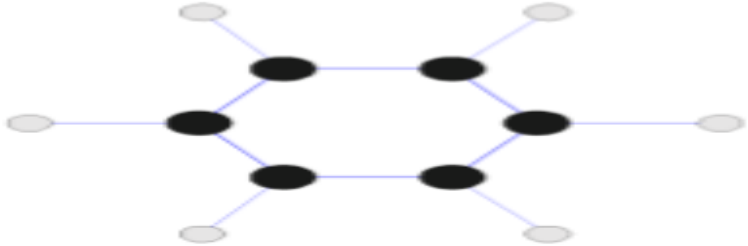
	team	coach	play	ball	score	game	win	lost	timeout	Season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

…Dataset Types



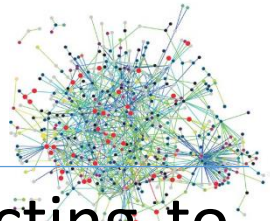
2) Graph and Network: Contain nodes and connecting vertices. E.g. WWW, Social or information networks, Molecular Networks



3) Ordered: Has Sequences of transactions

- **Spatial Data**: maps, Spatial data or geospatial data, is information about a physical object that can be represented by numerical values in a geographic coordinate system.
- **Temporal Data**: time series, A temporal data denotes the evolution of an object characteristic over a period of time. Eg d=f(t).
- **Sequential Data:** transaction sequence, Data arranged in sequence. E.g. Video data – sequence of images

Characteristics of Structured Data



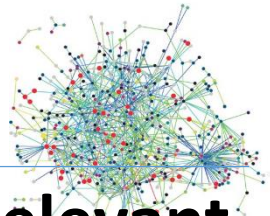
- **Dimensionality: A** Data Dimension is a set of data attributes affecting to something of interest to a business. Dimensions are things like "customers", "products", "stores" and "time".

i. **Curse of Dimensionality:** When dimensionality increases, data becomes increasingly light in the space that it occupies. Avoid curse/ weak points of dimensionality. Reduce amount of time and memory required by data mining algorithms

ii. **Dimensionality Reduction**: Goal: to reduce dimensionality of data. process of reducing the number of random variables under consideration by obtaining a set of principal variables

a) **Principle Component Analysis (PCA):** find a projection that captures the largest amount of variation in data. Construct a neighborhoods graph.

…Characteristics of Structured Data



a) Feature Subset Selection: **Redundant features (**Duplicate) **or Irrelevant features(**no information)

Techniques:

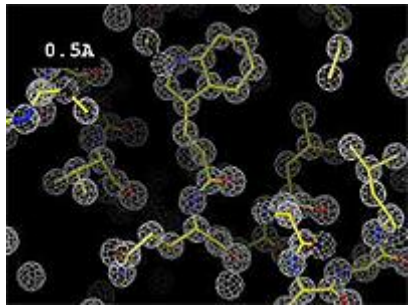
- **Brute-force approach:-** Try all possible feature subsets as input to data mining algorithm
- **Embedded approaches:** - Feature selection occurs naturally as part of the data mining algorithm
- **Filter approaches:-** Features are selected before data mining algorithm is run
- **Wrapper approaches:** - Use the data mining algorithm as a black box to find best subset of attributes.

…Characteristics of Structured Data

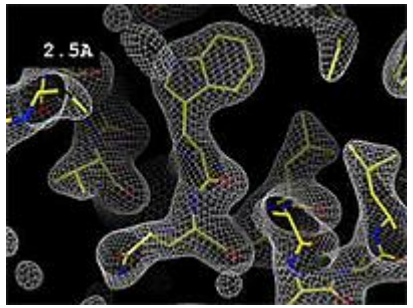
- **Sparsity and Density:** Describe the number of cells in a table that are empty (sparsity) and that contain information (density).

A table that is 10% dense has 10% of its cells populated with non-zero values. It is therefore 90% sparse – meaning that 90% of its cells are either not filled with data or are zeros.

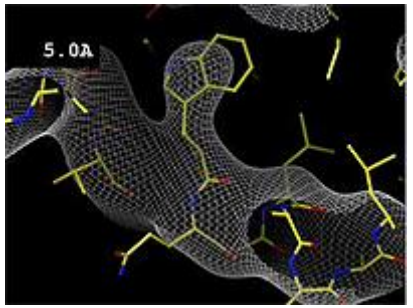
- **Resolution:** if two objects are closer than this distance, they appear as one combined blob rather than two separate objects.



Resolution at 0.5 (Excellent)
backbone and most sidechains very clear

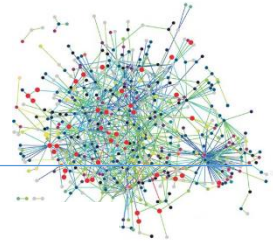


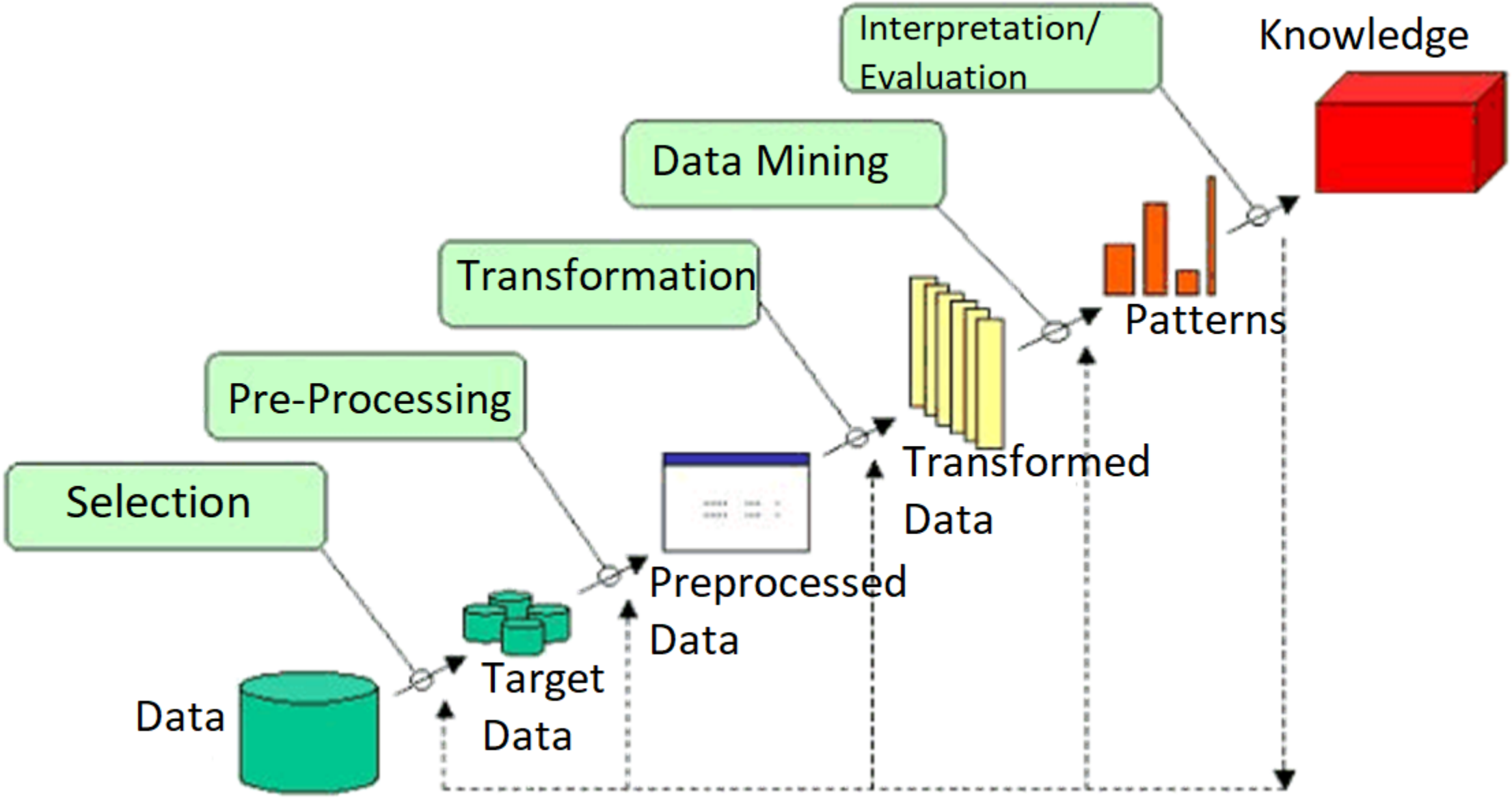
Resolution at 2.5A (good)
backbone and many sidechains clear



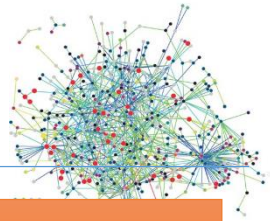
Resolution at 5.0A (poor)
backbone mostly clear; sidechains not clear

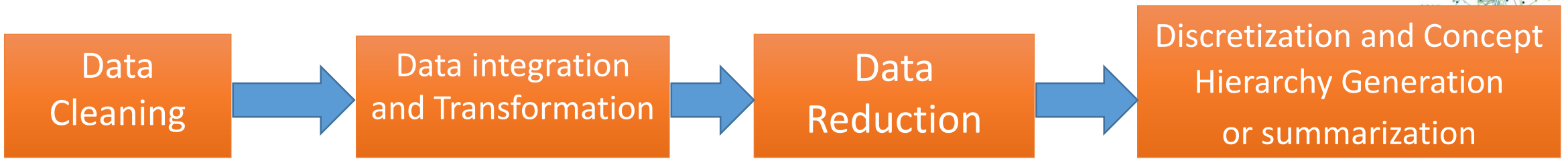
KDD (Knowledge Discovery in Databases)





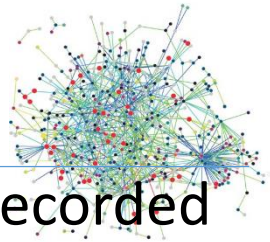
Data Pre-Processing Process





- **Data cleaning: is** a process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.
  - i. Fill-in missing values
  - ii. Identify outliers and smooth out noisy data
  - iii. Correct inconsistent data
  - iv. Eliminate duplicate data

Data Pre-Processing Process: (1)Data Cleaning



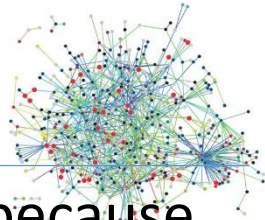
a) Missing Data -Data is not always available because many tuples may not have recorded values for several attributes such as age, income. Missing data may be due to: -

- Equipment Malfunction.
- Inconsistent with other recorded data and thus deleted.
- Data not entered due to misunderstanding.
- Certain data may not be considered important at the time of entry.
- Not register history or changes of the data.

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing. Not effective when the percentage of missing values per attribute varies considerably.
- Fill-in missing values manually: Tedious and infeasible task.
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean for all samples belonging to the same class: smarter
  - the most probable value: inference-based such as Bayesian formula or decision tree

…Data Pre-Processing Process: (1)Data Cleaning



b) Noisy Data – Meaningless data. e.g. Salary=”-10” Noisy data is a form of error because of random error in a measured variable. Incorrect attribute values may be due to:

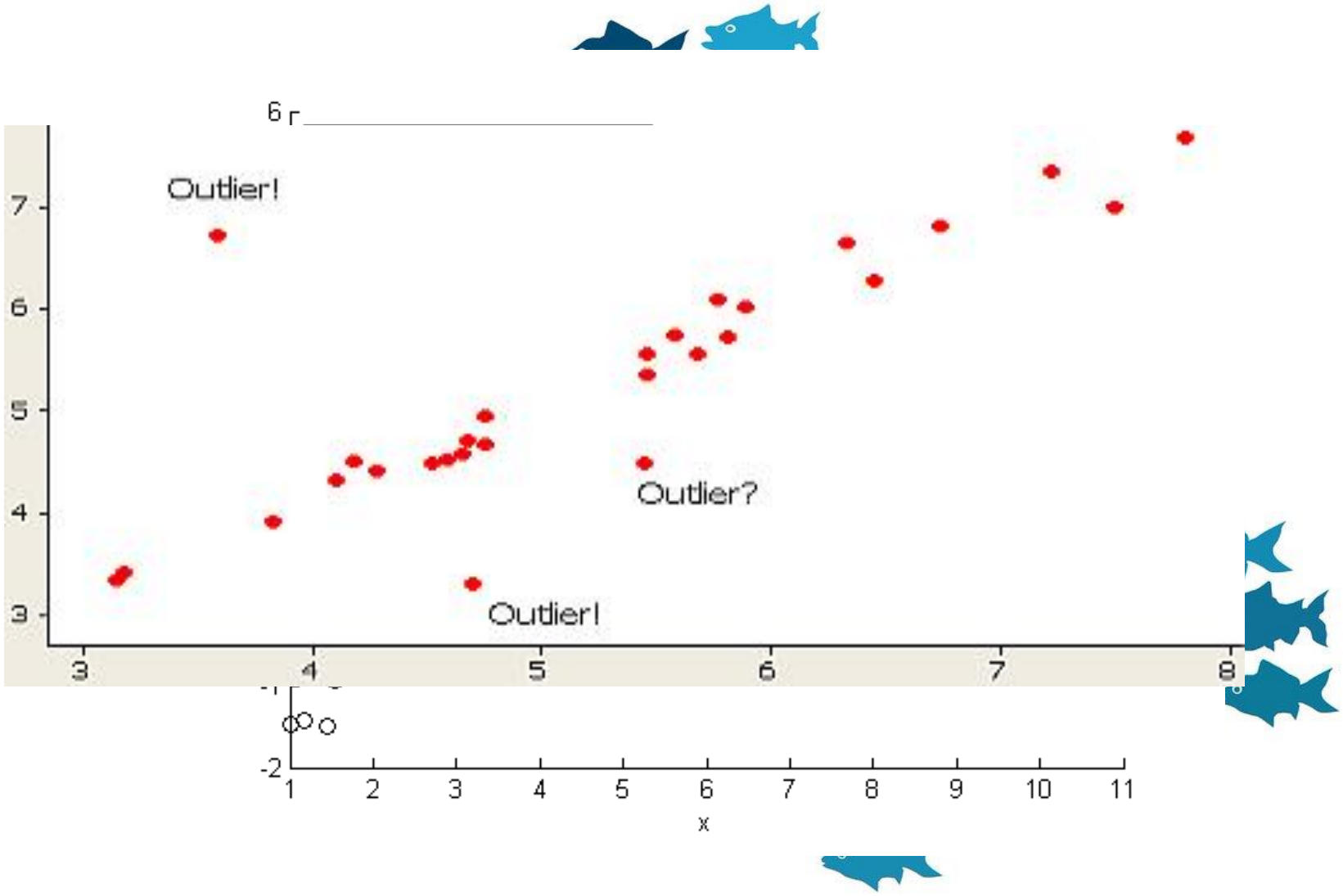
- Faulty data collection instruments .
- Data entry problem .
- Data transmission problem .
- Technology limitation .
- Inconsistency in naming convention
- Duplicate records, incomplete data(e.g. Occupation = “ ” (missing data)), inconsistent data (e.g. Was rating “1, 2, 3”, now “A, B, C”)

How to Handle Noisy Data

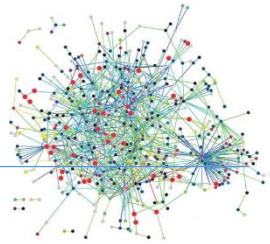
- **Clustering**: Detect and remove outliers

- **Regression**: Smooth by fitting the data into regression function- Linear or Multi-Linear regression

- **Binning Method**: first sort data and partition into (equal-frequency) bins, then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.


Outlier!
Outlier?
Outlier!
x

…Data Pre-Processing Process: (1)Data Cleaning



Binning Method

Example: Sorted data for price: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* **Partition** into equal-frequency (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34

* Smoothing by **bin** means:
  - Bin 1 (36/4 = 9) : 9, 9, 9, 9
  - Bin 2 (91/4 = 22.75) : 23, 23, 23, 23
  - Bin 3 (117/4 = 29.25): 29, 29, 29, 29

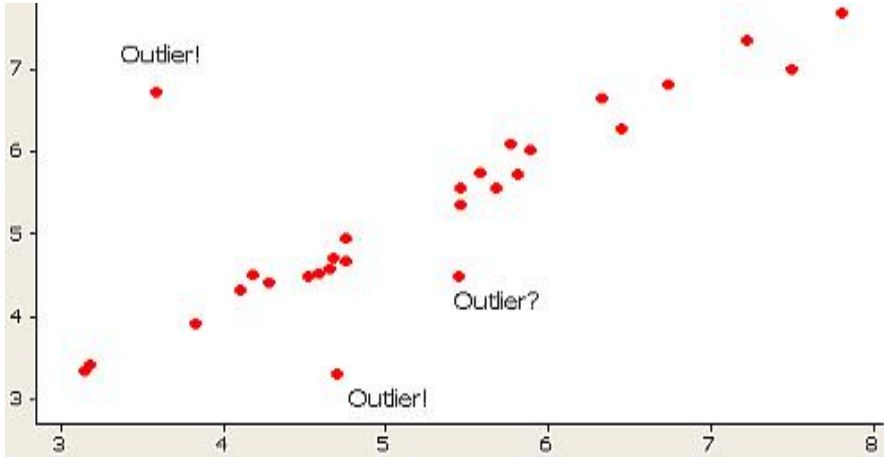
4+8+9+15=36
21+21+24+25=91
26+28+29+34=117

* Smoothing by **bin** boundaries:
  - Bin 1: 4, 4, 4, 15 boundary is (4,15) where 8, 9 are closure with 4 than 15
  - Bin 2: 21, 21, 25, 25 boundary is (21,25) where 21 is closure with 21 but 24 are closure with 25
  - Bin 3: 26, 26, 26, 34 boundary is (26,34) where 28, 29 are closure with 26 than 34

… Data Pre-Processing Process: (1) Data Cleaning

c) Outliers - Outliers are a set of data points that are considerably dissimilar or inconsistent with the remaining data that lie outside normal experience. In most of the cases they are inference of noise while in some cases they may actually carry valuable information. Outliers can occur because of:

- Transient malfunction of data measurement.
- Error in data transmission or transcription or data entry
- Changes in system behaviour.
- The data are wrong scaled;
- Fault in assumed theory



How to Handle Outliers?

Fundamental approaches to the problem of outlier's detection

a. Type 1 – Unsupervised Learning: Determine the outliers with no prior knowledge of data or neither classified nor labeled. E.g. Association or Clustering

b. Type 2 – Supervised Learning: Model with normality and abnormality. E.g. Classification or Regression

c. Type 3 – Semi-Supervised Learning: Model with normality.

Data Pre-Processing Process: (2)Data Integration and Transformation

Data integration: Combines data from multiple sources into a coherent data store e.g. data warehouse

Issues in data integration:

- Schema integration: e.g ***A.cust-id=B.cust-#.***
- Detecting and resolving data value conflicts: For the same real-world entity, attribute values from different sources are different. **e.g., metric vs. British units**
- Redundant data occur often when integration of multiple databases:

Data Transformation: Transformation process deals with rectifying any inconsistency.

- each old value can be identified with one of the new values
- One of the most common transformation **issues** is 'Attribute Naming Inconsistency'.

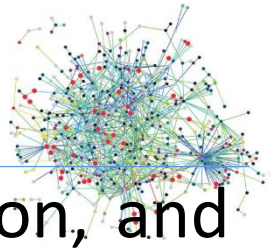
*E.g. Employee Name may be **EMP_NAME** in one database, **ENAME** in the other.*

- Once all the data elements have right names, they must be converted to common formats.

-2, 32,100, 59, 48 ⟶ 0.02, 0.32, 1.00, 0.59, 0.48

Data Transformation

(2)Data transformation Techniques



- **Smoothing**, remove noise from the data, include binning, regression, and clustering.
- **Aggregation**, sometimes "LESS IS MORE" i.e. summarization of data, where summary or aggregation operations are applied to the data e.g. data cube construction. Combining of two or more objects into a single object.

***For example**, the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple granularities.*

- **Generalization** of the data, Hierarchy climbing of data where low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies.

***For example**, categorical attributes, like street, can be generalized to higher-level concepts, like city or country. Similarly, values for numerical attributes, like age, may be mapped to higher-level concepts, like youth, middle-aged, and senior.*

…(2)Data transformation Techniques: Normalization

Normalization, Mainly, when data not following normality assumptions we transform it to get normality. Scaled to fall within a small and specified range, where the attribute data are scaled so as to fall within a small specified range, such as−1.0 to 1.0, or 0.0 to 1.0.

- **Min-Max Normalization:** transforms the data set from one range to another. Transform the data from measured units to a new interval from minA to maxA for feature A:

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

where v is the current value of feature A.

Example: Suppose that the minimum and maximum values for the feature income are Rs. 12,000 and Rs. 98,000, respectively. We would like to map income to the range [0.0, 1.0]. By min-max normalization, a value of Rs. 73,600 for income is transformed to:

Here;
V=73,600
minA=12,000
maxA=98,000
new_maxA=1.0
new_mina=0.0

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0.0) + 0 = 0.716$$

…(2)Data transformation Techniques: Normalization

- **Z-Score (Zero-Score) Normalization:** Transform the data by converting the values to a common scale with an average of zero and a standard deviation of one. A value, v, of A is normalized to v' by computing: $v' = \frac{v - \bar{F}}{\sigma_F}$ where $\bar{F}$ and σ_F are the mean and standard deviation of feature F, respectively.

Example: Suppose that the mean and standard deviation of the values for the feature income are $54,000 and $16,000, respectively. With z-score normalization, a value of $73,600 for income is transformed to (73,600 – 54,000)/ 16,000 = 1.225

- **Normalization by decimal scaling:** Transform the data by moving the decimal points of values of feature F. The number of decimal points moved depends on the maximum absolute value of F. A value v of F is normalized to v' by computing:

V'= V/ 10^j where j is the smallest integer such that max (|V'|) <1

Example: Suppose that the recorded values of F range from – 986 to 917. The maximum absolute value of F is 986. To normalize by decimal scaling, we therefore divide each value by 1,000 (i.e., j = 3) so that –986 normalizes to –0.986 and 917 normalizes to 0.917.

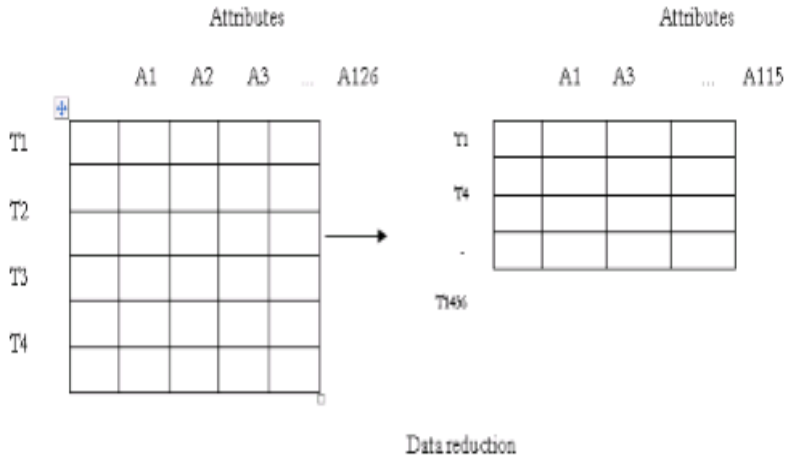
- **Features/ Attributes construction** (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.

Data Pre-Processing Process: (3)Data Reduction

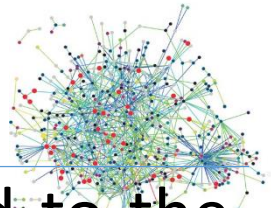
- Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Process of minimizing the amount of data that needs to be stored in a data storage environment.
- Data reduction can increase storage efficiency and reduce costs.

Need for data reduction:

- Reducing the number of attributes
- Reducing the number of attribute values
- Reducing the number of tuples

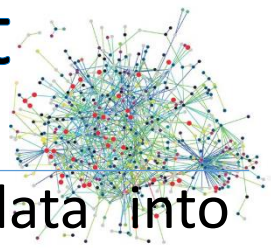


… (3)Data Reduction: Strategies



1) **Data cube aggregation**, where aggregation operations are applied to the data in the construction of a data cube. store multidimensional aggregated information.
2) **Attribute subset selection**, where irrelevant, weakly relevant, or redundant attributes or dimensions may be detected and removed. By Redundant attributes and Irrelevant Attributes
3) **Dimensionality reduction**, where encoding mechanisms are used to reduce the data set size. "compressed" representation of the original data.
4) **Numerosity reduction**, reduce data volume where the data are replaced or estimated by alternative, smaller form of data representations. By parametric models (store only the model parameters instead of the actual data) or nonparametric (clustering, sampling, and the use of histograms)
5) **Data Sampling:** It is one of main method for data selection i.e. sampling is the main technique employed for data selection. By Random sampling, Stratified Sampling

Data Pre-Processing Process: (4) Discretization and Concept Hierarchy Generation (or summarization):



- Discretization convert continuous data into discrete data and Partition data into different classes.
- **Discretization**: Reduce the number of values for a given continuous attribute by divide the range of a continuous attribute into intervals. Interval labels can then be used to replace actual data values.
- **Concept Hierarchies**: Reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute "age") by higher level concepts (such as young, middle-aged or senior).

Approaches:

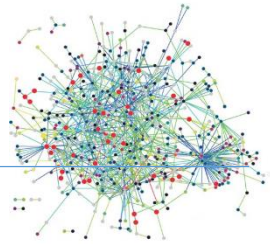
1) Equal width (distance) partitioning:

- It divides the range into N intervals of equal size.
- If A and B are the lowest and the highest values of the attribute, the width of interval will be - **W = (A – B)/N.**

2)Equal depth (frequency) partitioning:

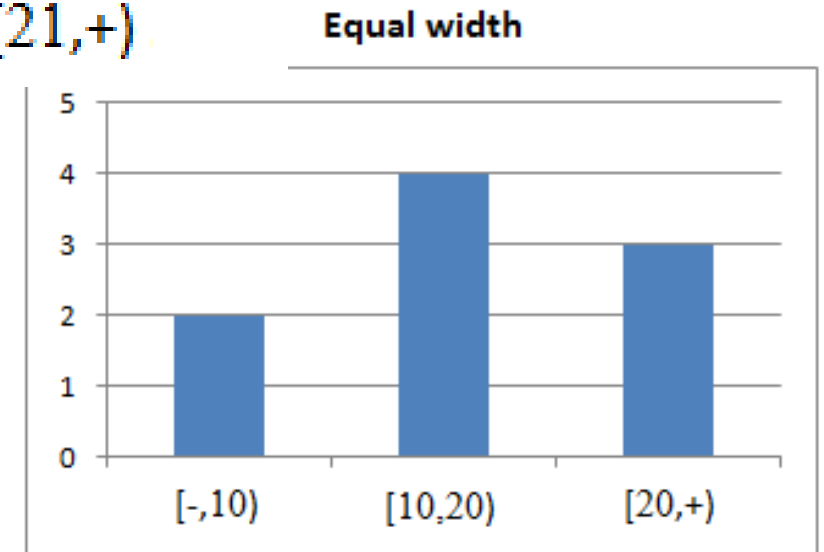
- It divides the range into N intervals, each containing approximately same number of samples.

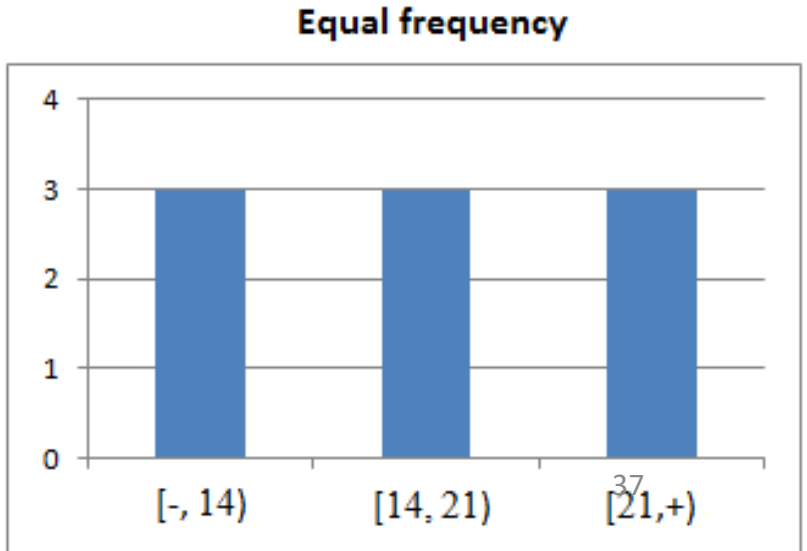
Example: Equal Width and Equal Frequency



- **Data :** 0, 4, 12, 16, 16, 18, 24, 26, 28
- **Equal width**
  - Bin 1: 0, 4 [-,10) < 10
  - Bin 2: 12, 16, 16, 18 [10,20) 10 - 20
  - Bin 3: 24, 26, 28 [20,+) 20 >
- **Equal frequency**
  - Bin 1: 0, 4, 12 [-, 14) 16 + 12 /2 = 14
  - Bin 2: 16, 16, 18 [14, 21) 18+24 /2 = 21
  - Bin 3: 24, 26, 28 [21,+)

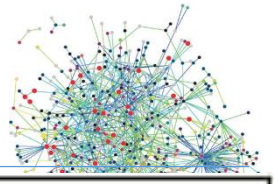
Max-min/k = 28-0/3 = 9.33 ~ 10



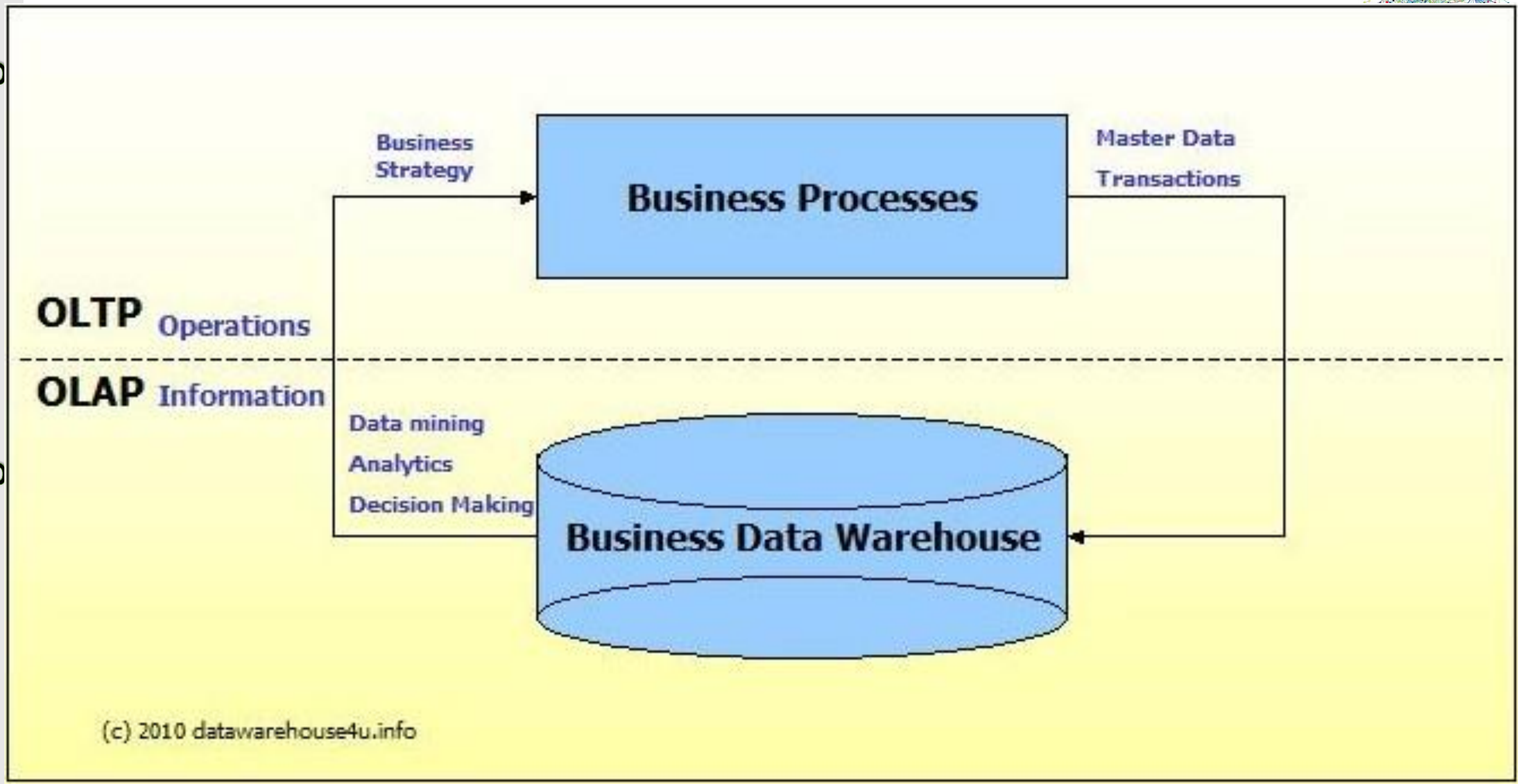


Data Minir

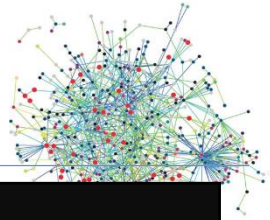
OLAP and OLTP



Data Mining and Ware housing

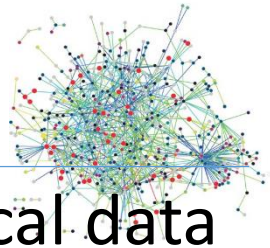


OLAP vs OLTP

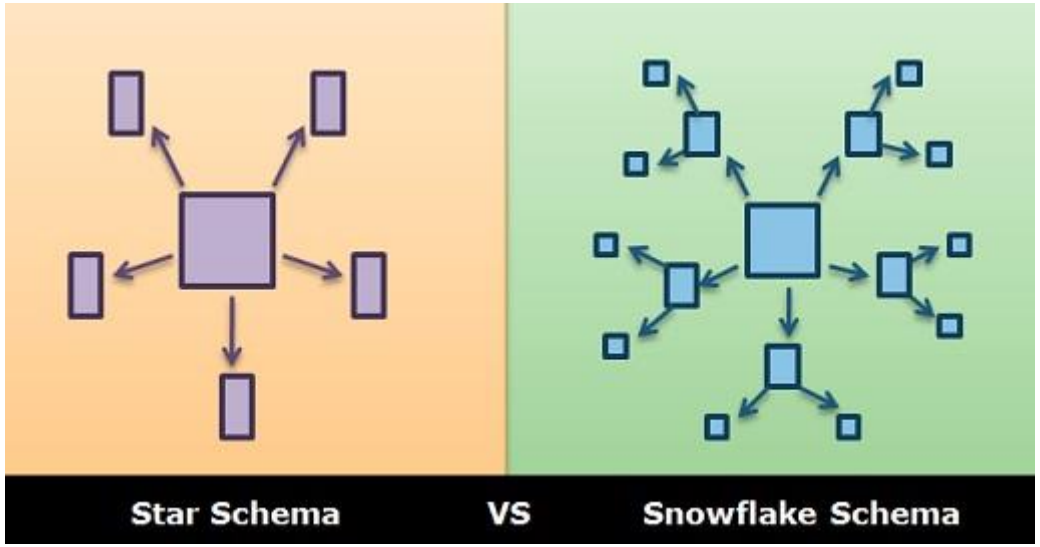


Attributes	OLTP	OLAP
Type of users	clerk, IT professional	knowledge worker
Function	day to day operations	decision support
DB design	application-oriented	subject-oriented
Data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
Usage	repetitive	ad-hoc
Access	read/write index/hash on prim. key	lots of scans
Unit of work	short, simple transaction	complex query
Records accessed	tens	millions
No of users	thousands	hundreds
DB size	100MB-GB	100GB-TB
Metric	transaction throughput	query throughput, response

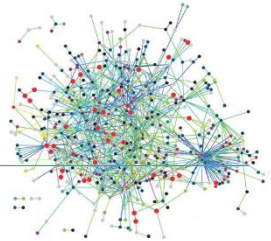
OLAP & Multidimensional Data Analysis



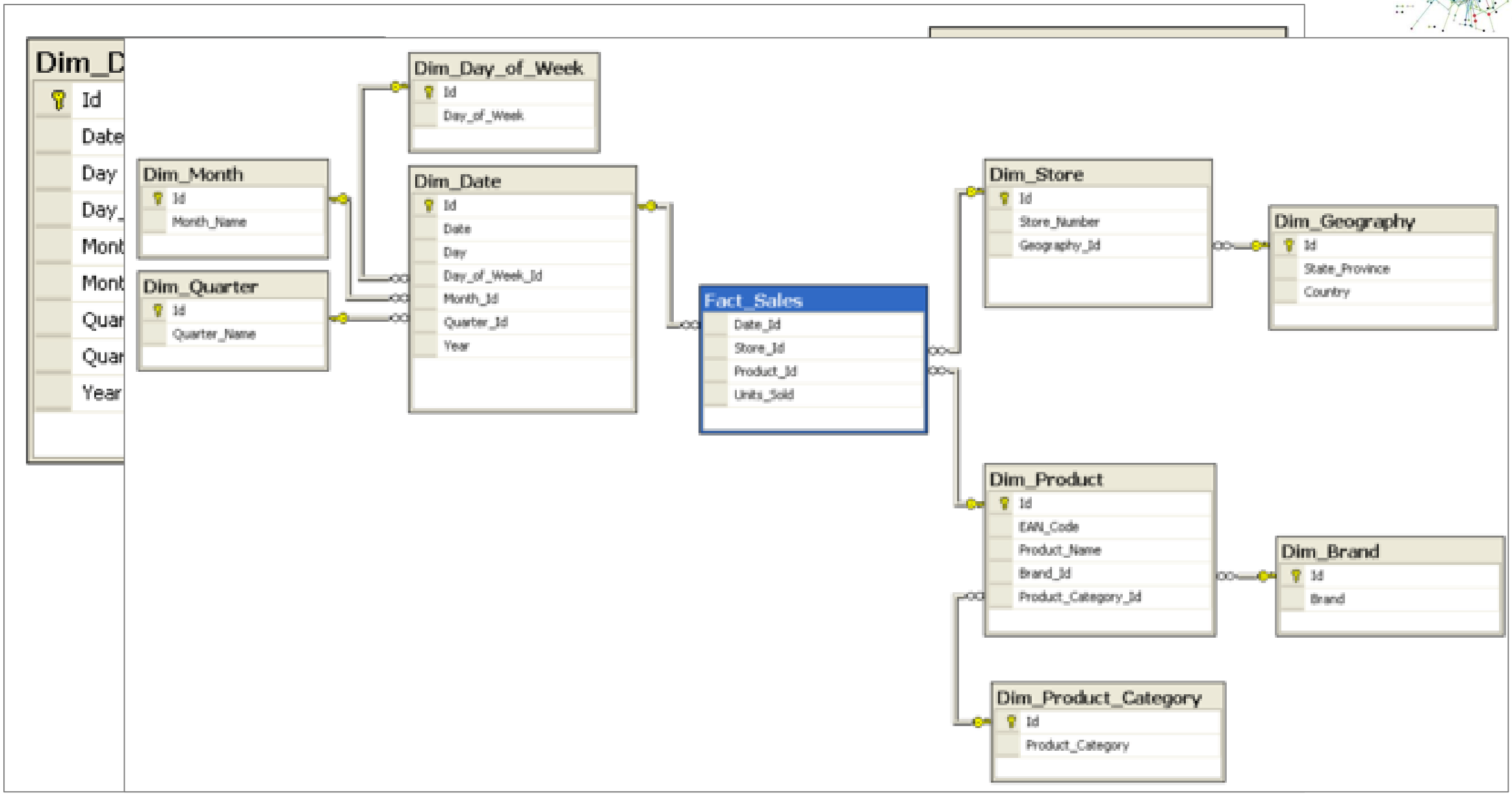
- OLAP is a design pattern, a way to seek information out of the physical data store.
- OLAP is all about summary.
- It aggregates information from multiple systems, and stores it in a multi-dimensional format.
- Format could be a star, snowflake or a hybrid kind of a schema.



Example: Star and Snowflake Schemas



Data Mining and Ware housing

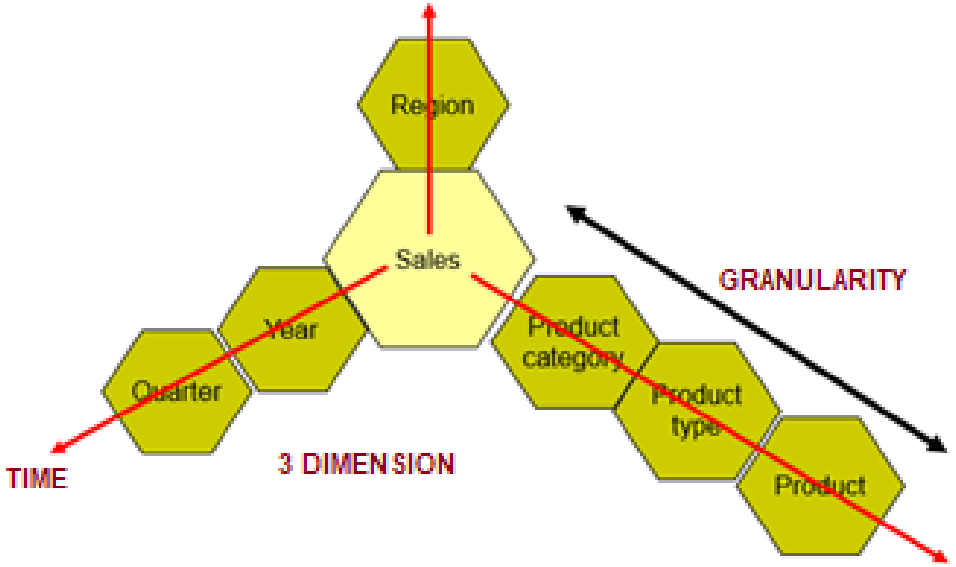


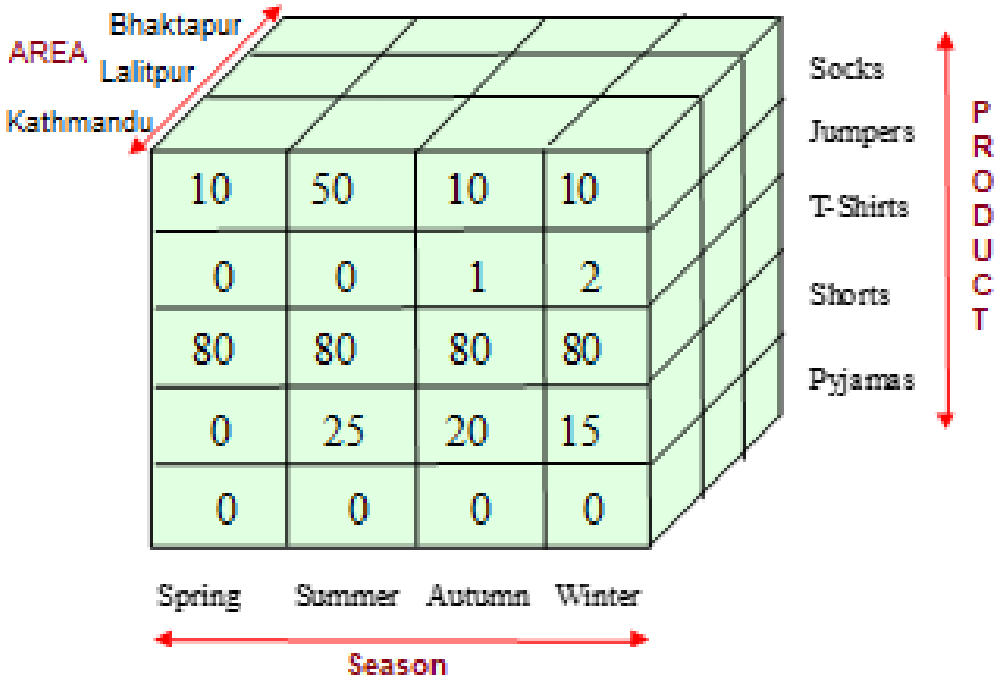
Types of OLAP system

- **ROLAP- Star Schema Based**: Relational OLAP work primarily from the data that resides in a relational database, where the base data and dimension tables are stored as relational tables
- **MOLAP- Cube Based**: Multidimensional OLAP where data are pre-summarized and are stored in an optimized format in a multidimensional cube, instead of in a relational database. In this type of model, data are structured into proprietary formats in accordance with a client's reporting requirements with the calculations pre-generated on the cubes.
- **HOLAP**: Hybrid OLAP attempt to incorporate the best features of MOLAP and ROLAP into a single architecture

Multi-Dimensional Analysis

- Informational Analysis on data which takes into account many different relationships, each of which represents a dimension
- Dimension refers to a structural attribute of a data cube. The dimension is composed or related and hierarchical members.

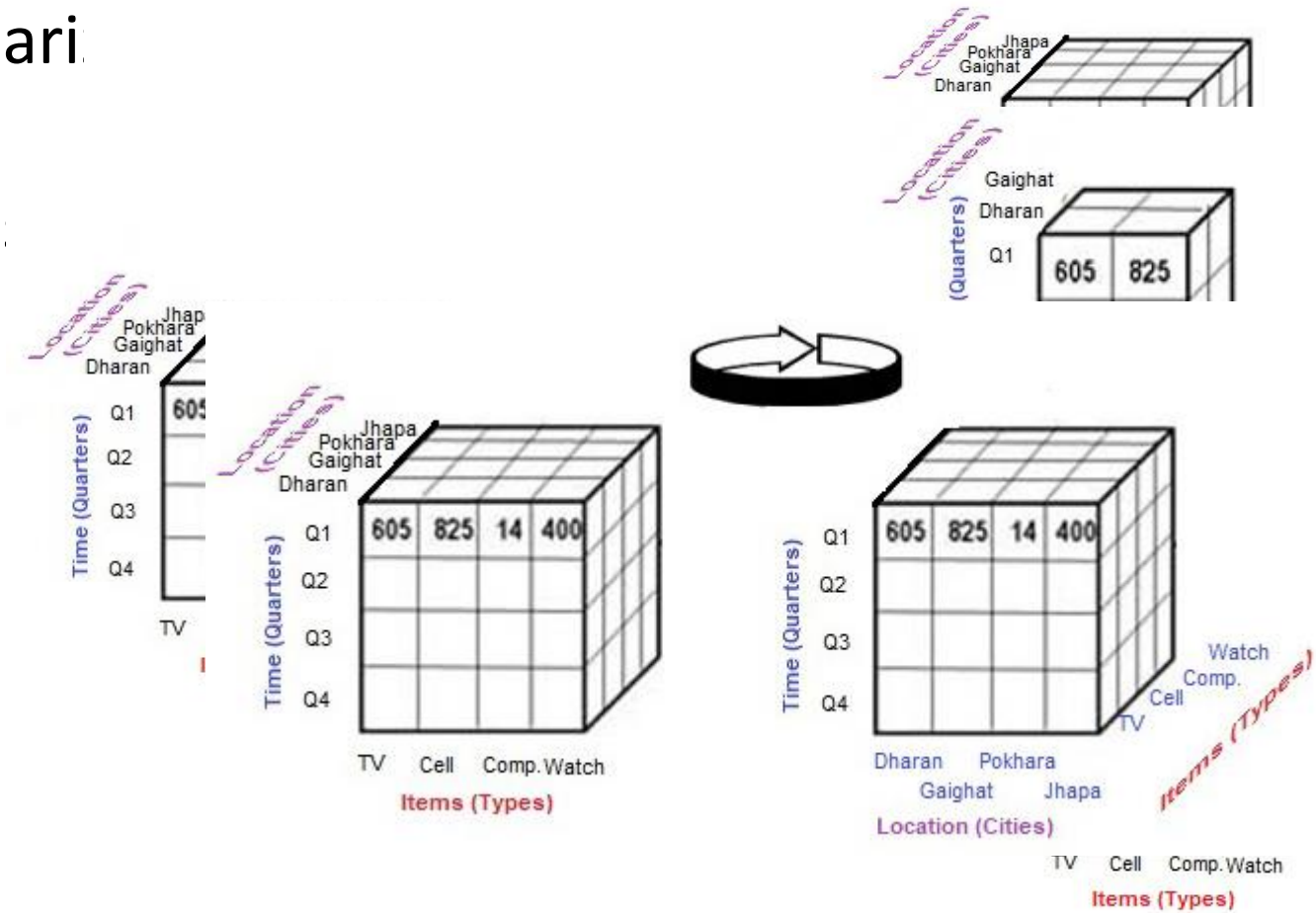


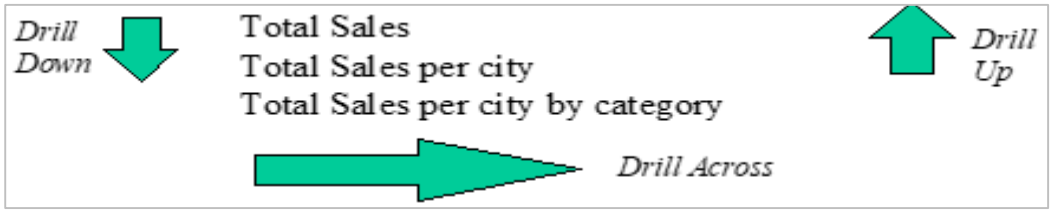


Eg. What is the total socks selling in Kathmandu are in Summer season?

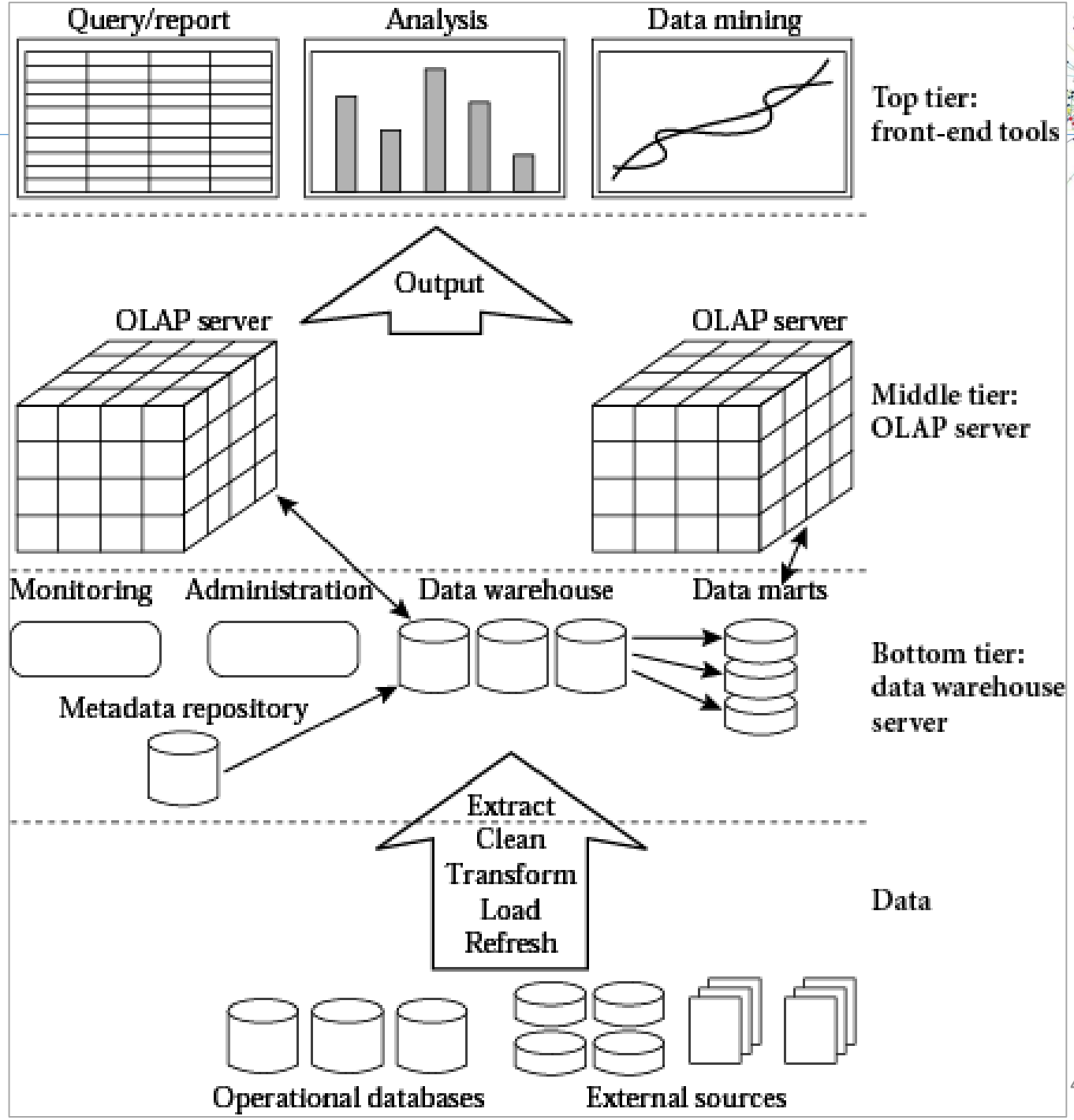
OLAP Operations

1) **Roll-Up (Drill-Up):** Summarize data
2) **Drill-Down (Roll Down):** Summari hierarchy,
3) **Slicing:** selection on one dimen
4) **Dicing:** a sub-cube by performing a selection of one or more dimensions.
1) **Pivoting (Rotate):** Rotates the data axis to view the data from different perspectives.





OLAP Architecture



Data Cube Computation