SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE

www.sathyabama.ac.in

SIT1301 -Data Mining and Warehousing

UNIT II

What is a Data Warehouse?

- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon

Data warehousing:

  - The process of constructing and using data warehouses

Why Data Mining

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

Evolution of Database Technology

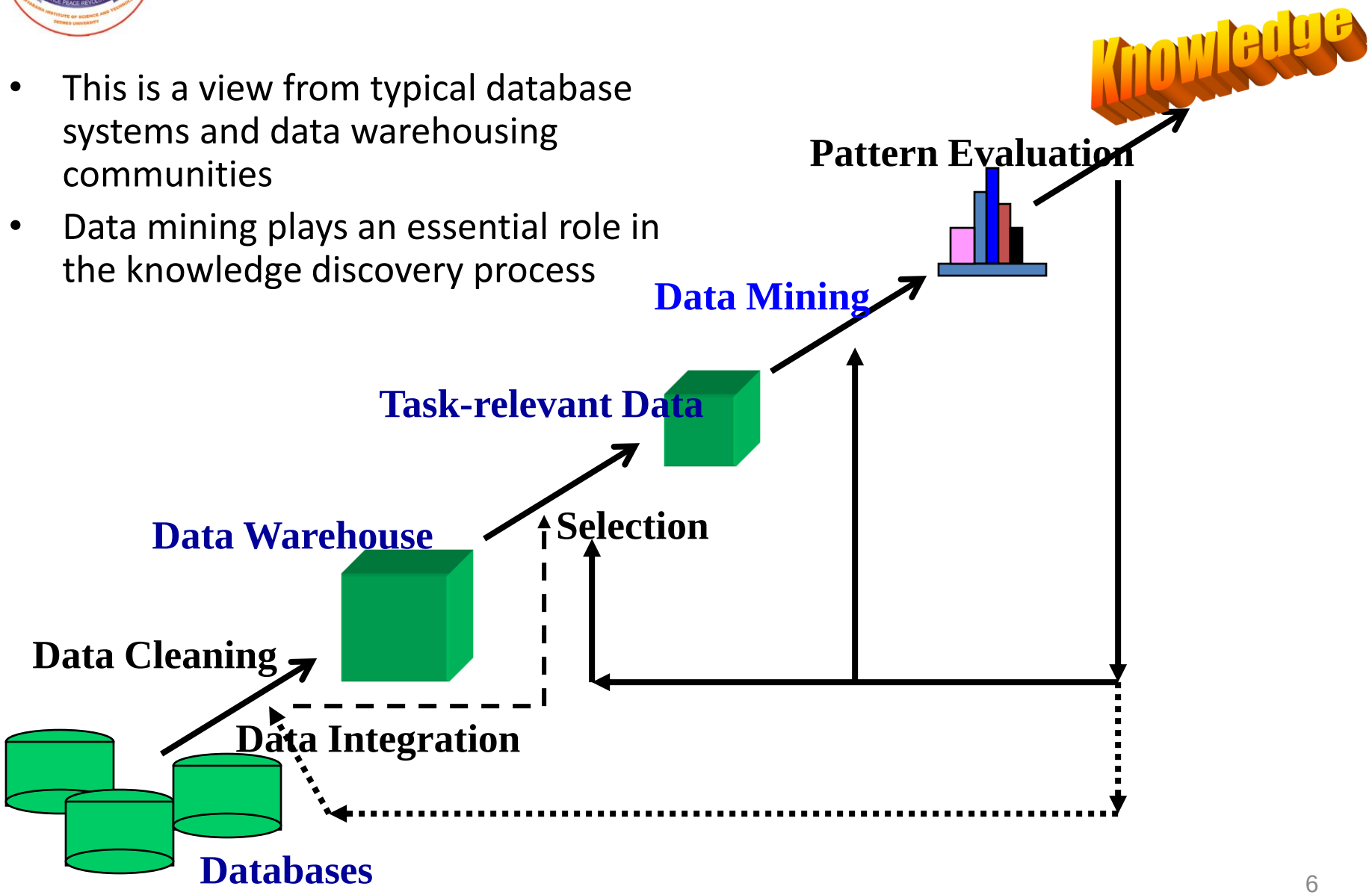
- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems

What is Data Mining

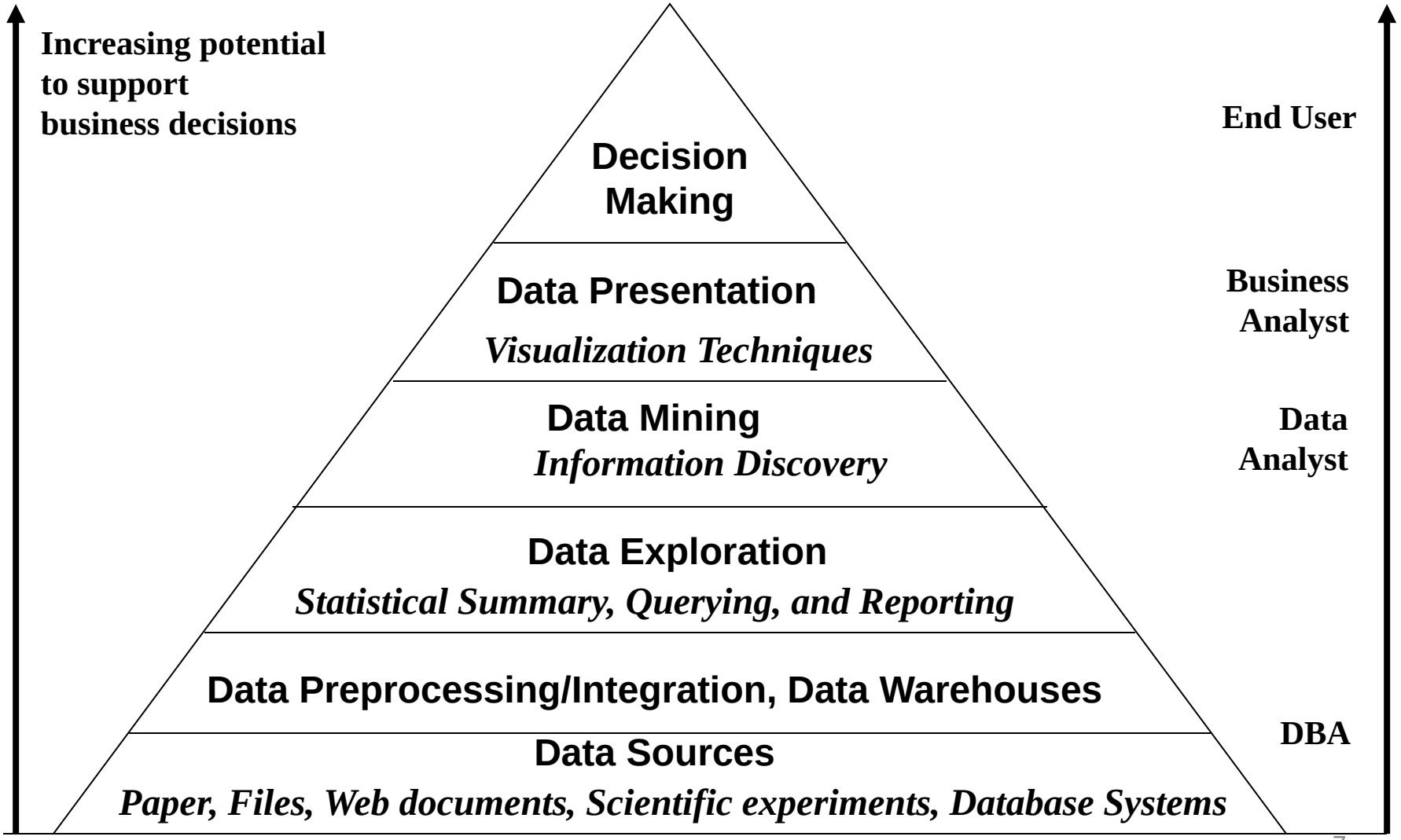
- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?
- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge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

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



Data Mining in Business Intelligence



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

Data Preprocessing

- Why Data Preprocessing?
- Why Is Data Dirty?
- Why Is Data Preprocessing Important?
- Multi-Dimensional Measure of Data Quality
- Major Tasks in Data Preprocessing
- Forms of Data Preprocessing
- Mining Data Descriptive Characteristics

Data Preprocessing

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking
- certain attributes of interest, or containing
- only aggregate data e.g., occupation=“ ”
- noisy: containing errors or outliers e.g., Salary=“-10”
- inconsistent: containing discrepancies in codes or names
  - e.g., Age=“42” Birthday=“03/07/1997”
  - e.g., Was rating “1,2,3”, now rating “A, B, C”
  - e.g., discrepancy between duplicate records

Why Is Data Dirty?

- **Incomplete data may come from**
  - "Not applicable" data value when collected
  - Different considerations between the time when the data was
  - collected and when it is analyzed.
  - Human/hardware/software problems
- **Noisy data (incorrect values) may come from**
  - Faulty data collection instruments
  - Human or computer error at data entry
  - Errors in data transmission
- **Inconsistent data may come from**
  - Different data sources
  - Functional dependency violation (e.g., modify some linked data)
- **Duplicate records also need data cleaning**

Why Is Data Preprocessing Important?

- **No quality data, no quality mining results!**
  - Quality decisions must be based on quality data
    - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
  - Data warehouse needs consistent integration of quality data
- **Data extraction, cleaning, and transformation comprises the majority of the work of building a data warehouse**

Multi-Dimensional Measure of Data Quality

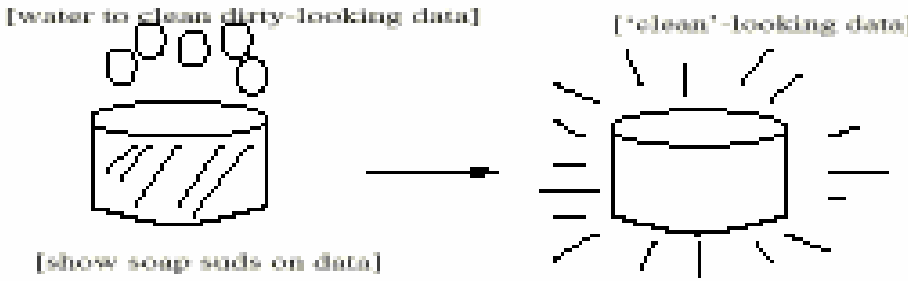
- **A well-accepted multidimensional view:**
  - Accuracy
  - Completeness
  - Consistency
  - Timeliness
  - Believability
  - Value added
  - Interpretability
  - Accessibility
- **Broad categories:**
  - Intrinsic, contextual, representational, and accessibility

Major Tasks in Data Preprocessing

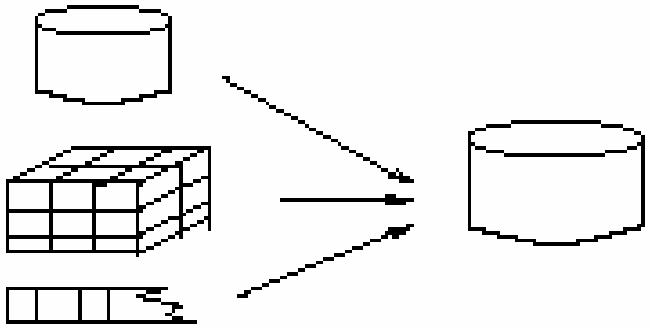
- **Data cleaning**
  - Fill in missing values, smooth noisy data, identify or remove
  - outliers, and resolve inconsistencies
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data transformation**
  - Normalization and aggregation
- **Data reduction**
  - Obtains reduced representation in volume but produces the same or similar analytical results
- **Data discretization**
  - Part of data reduction but with particular importance, especially for numerical data

Forms of Data Preprocessing

Data Cleaning



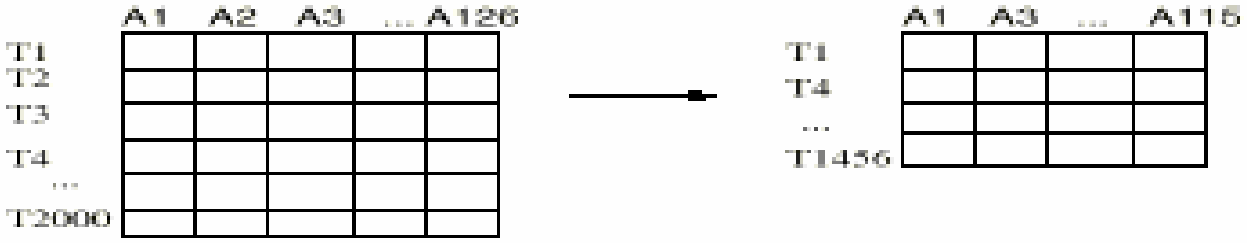
Data Integration



Data Transformation

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

Data Reduction



Mining Data Descriptive Characteristics

- **Motivation**
  - To better understand the data: central tendency, variation and spread
- **Data dispersion characteristics**
  - median, max, min, quantiles, outliers, variance, etc.
- **Numerical dimensions correspond to sorted intervals**
  - Data dispersion: analyzed with multiple granularities of
- precision
  - Boxplot or quantile analysis on sorted intervals
- **Dispersion analysis on computed measures**
  - Folding measures into numerical dimensions
  - Boxplot or quantile analysis on the transformed cube

Data Cleaning

- Importance
  - "Data cleaning is one of the three biggest problems in data warehousing"—Ralph Kimball
  - "Data cleaning is the number one problem in datawarehousing"—DCI survey
- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

Missing Data

- **Data is not always available**
  - E.g., many tuples have no recorded value for several
  - attributes, such as customer income in sales data
- **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
- **Missing data may need to be inferred.**

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (assuming the tasks in classification—not effective when the percentage of missing values per attribute varies considerably.
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - 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

Noisy Data

- **Noise: random error or variance in a measured variable**
- **Incorrect attribute values may due to**
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- **Other data problems which requires data cleaning**
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- **Binning**
  - 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.
- **Regression**
  - smooth by fitting the data into regression functions
- **Clustering**
  - detect and remove outliers
- **Combined computer and human inspection**
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Simple Discretization Methods: Binning

- **Equal-width (distance) partitioning**
  - Divides the range into N intervals of equal size: uniform grid
  - if A and B are the lowest and highest values of the attribute, the width of intervals will be: W = (B –A)/N.
  - The most straightforward, but outliers may dominate presentation
  - Skewed data is not handled well
- **Equal-depth (frequency) partitioning**
  - Divides the range into N intervals, each containing approximately
- same number of samples
  - Good data scaling
  - Managing categorical attributes can be tricky

Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 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: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries:
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Data Cleaning as a Process

- **Data discrepancy detection**
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check field overloading
  - Check uniqueness rule, consecutive rule and null rule
  - Use commercial tools
- **Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections**
- **Data auditing: by analyzing data to discover rules and relationship to detect**
- **Data migration and integration**
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface Integration of the two processes
  - Iterative and interactive (e.g., Potter's Wheels)

Data Integration

- **Data integration:**
  - Combines data from multiple sources into a coherent store
- **Schema integration**: e.g., A.cust-id º B.cust-#
  - Integrate metadata from different sources
- **Entity identification problem:**
  - Identify real world entities from multiple data sources, e.g., Bill Clinton = William Clinton
- **Detecting and resolving data value conflicts**
  - For the same real world entity, attribute values from
- **different sources are different**
  - Possible reasons: different representations, different scales, e.g., metric vs. British units

Handling Redundancy in Data Integration

- **Redundant data occur often when integration of multiple databases**
  - **Object identification**: The same attribute or object may have different names in different databases
- **Derivable data:** One attribute may be a "derived" attribute in another table, e.g., annual revenue
- **Redundant attributes may be able to be detected by correlation analysis**
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Data Transformation

- **Smoothing**: remove noise from data
- **Aggregation:** summarization, data cube construction
- **Generalization:** concept hierarchy climbing
- **Normalization:** scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- **Attribute/feature construction**
  - New attributes constructed from the given ones

Data Transformation

- **Smoothing**: remove noise from data
- **Aggregation:** summarization, data cube construction
- **Generalization:** concept hierarchy climbing
- **Normalization:** scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- **Attribute/feature construction**
  - New attributes constructed from the given ones

Data Transformation

- **Smoothing**: remove noise from data
- **Aggregation:** summarization, data cube construction
- **Generalization:** concept hierarchy climbing
- **Normalization:** scaled to fall within a small, specified range
  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling
- **Attribute/feature construction**
  - New attributes constructed from the given ones

Data Transformation: Normalization

- Min-max normalization: to [new_min_A, new_max_A]

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

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,600 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000}(1.0 - 0) + 0 = 0.716$
- Z-score normalization (μ: mean, σ: standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. The $\frac{73,600 - 54,000}{16,000} = 1.225$
- Normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

Where j is the smallest integer such that $\text{Max}(|v'|) < 1$

Data Reduction Strategies

- Why data reduction?
  - A database/data warehouse may store terabytes of data
  - Complex data analysis/mining may take a very long time to run on the complete data set
- 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
- Data reduction strategies
  - Data cube aggregation:
  - Dimensionality reduction — e.g., remove unimportant attributes
  - Data Compression
  - Numerosity reduction — e.g., fit data into models
  - Discretization and concept hierarchy generation

Data Cube Aggregation

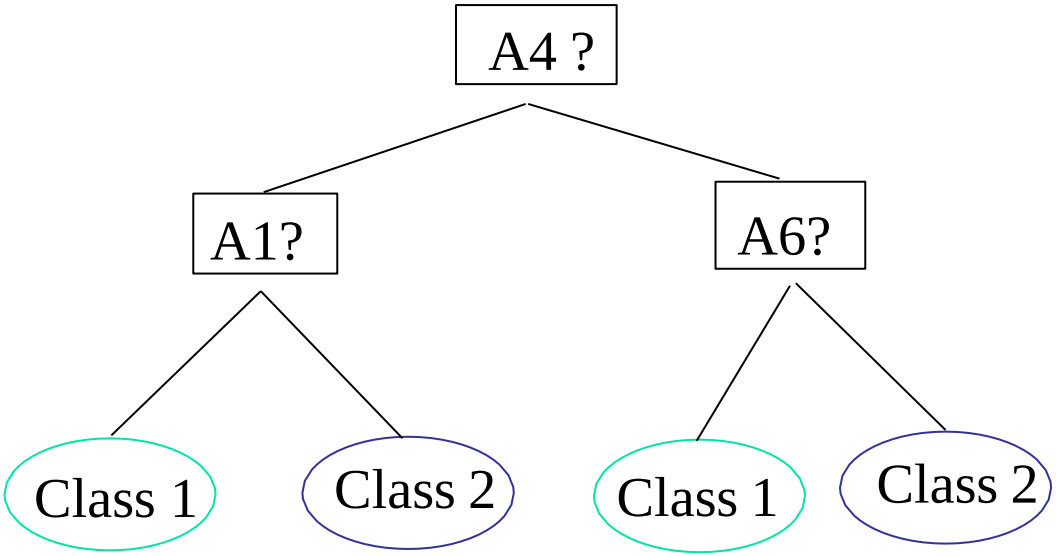
- The lowest level of a data cube (base cuboid)
  - The aggregated data for an individual entity of interest
  - E.g., a customer in a phone calling data warehouse
- Multiple levels of aggregation in data cubes
  - Further reduce the size of data to deal with
- Reference appropriate levels
  - Use the smallest representation which is enough to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible

Attribute Subset Selection

- Feature selection (i.e., attribute subset selection):
  - Select a minimum set of features such that the probability distribution of different classes given the values for those features is as close as possible to the original distribution given the values of all features
  - reduce # of patterns in the patterns, easier to understand
- Heuristic methods (due to exponential # of choices):
  - Step-wise forward selection
  - Step-wise backward elimination
  - Combining forward selection and backward elimination
  - Decision-tree induction

Example of Decision Tree Induction

Initial attribute set:
{A1, A2, A3, A4, A5, A6}

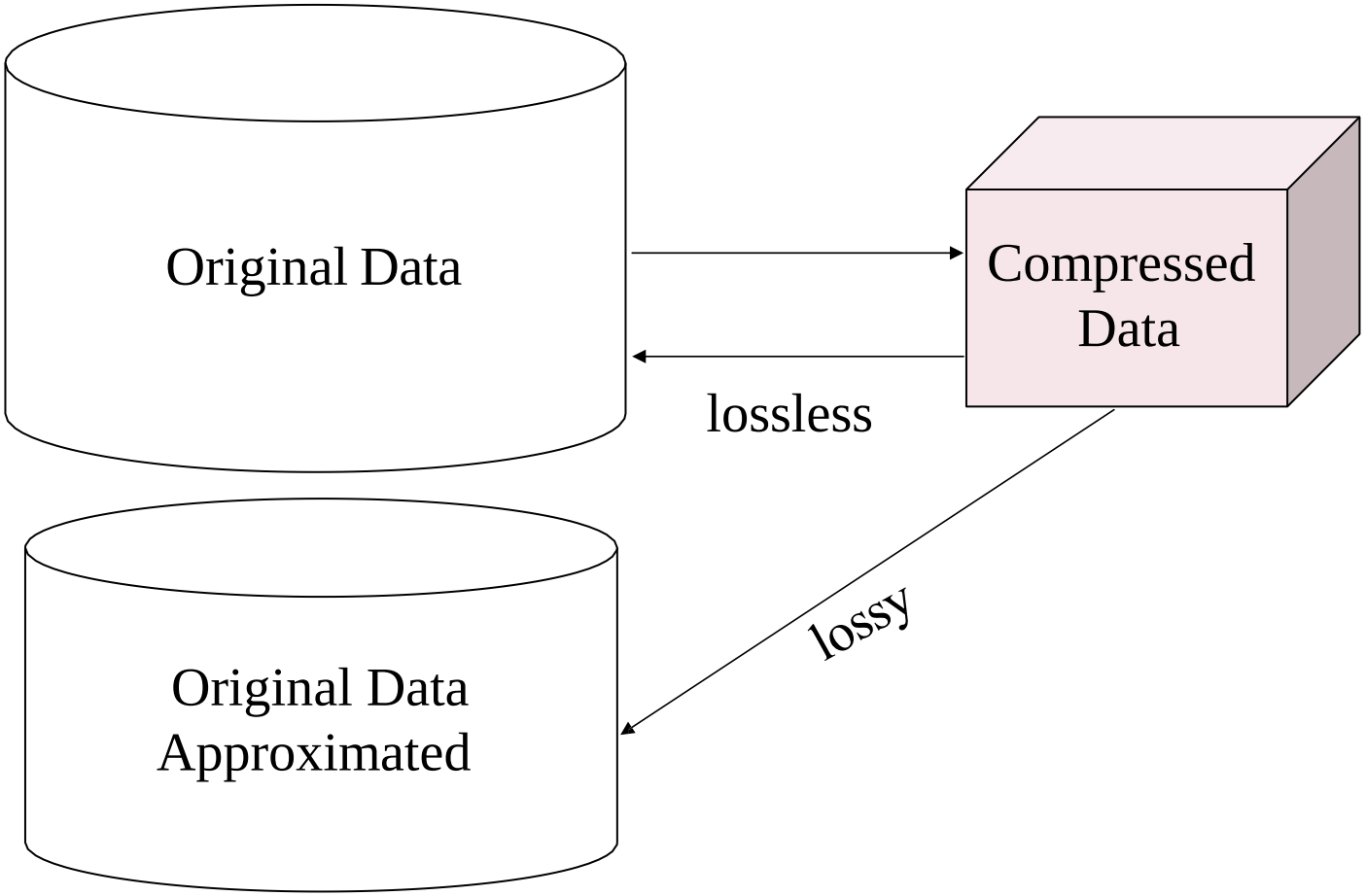


----> Reduced attribute set: {A1, A4, A6}

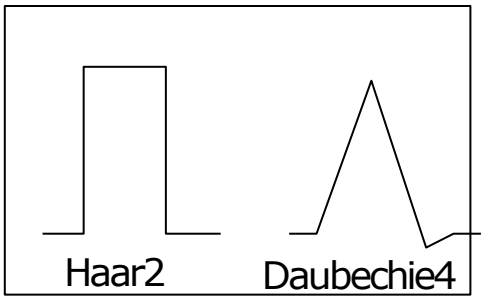
Data Compression

- String compression
  - There are extensive theories and well-tuned algorithms
  - Typically lossless
  - But only limited manipulation is possible without expansion
- Audio/video compression
  - Typically lossy compression, with progressive refinement
  - Sometimes small fragments of signal can be reconstructed without reconstructing the whole
- Time sequence is not audio
  - Typically short and vary slowly with time

Data Compression



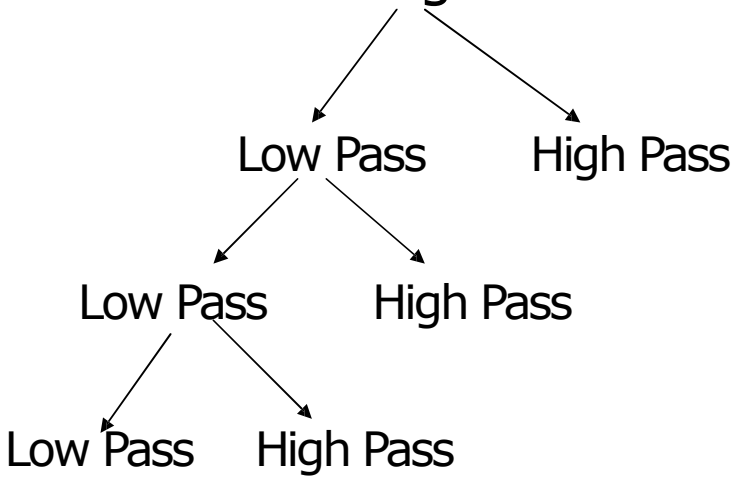
Dimensionality Reduction: Wavelet Transformation

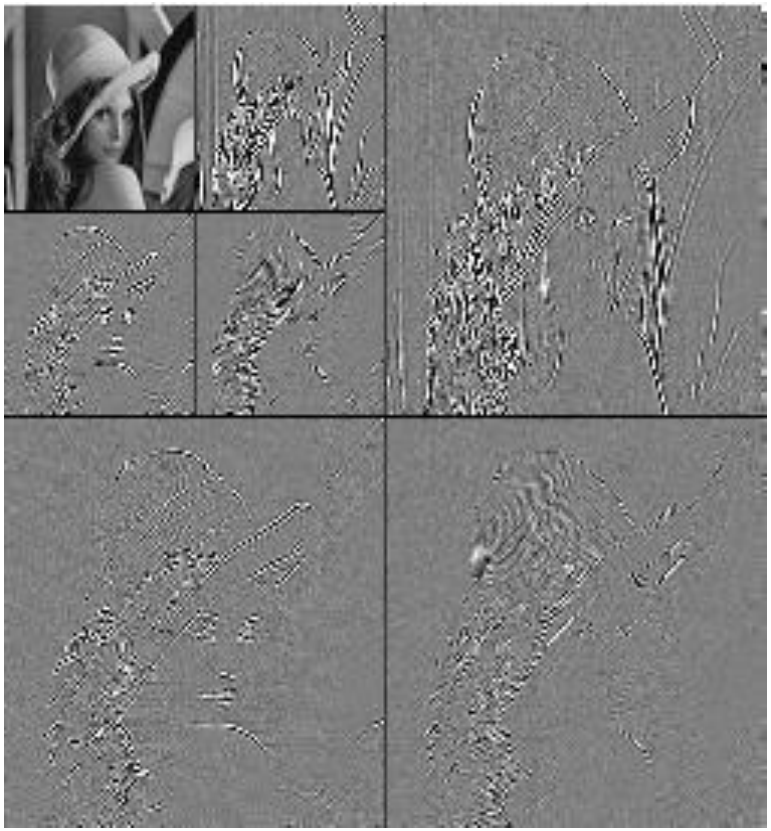


- Discrete wavelet transform (DWT): linear signal processing, multi-resolutional analysis
- Compressed approximation: store only a small fraction of the strongest of the wavelet coefficients
- Similar to discrete Fourier transform (DFT), but better lossy compression, localized in space
- Method:
  - Length, L, must be an integer power of 2 (padding with 0's, when necessary)
  - Each transform has 2 functions: smoothing, difference
  - Applies to pairs of data, resulting in two set of data of length L/2
  - Applies two functions recursively, until reaches the desired length

DWT for Image Compression

- Image

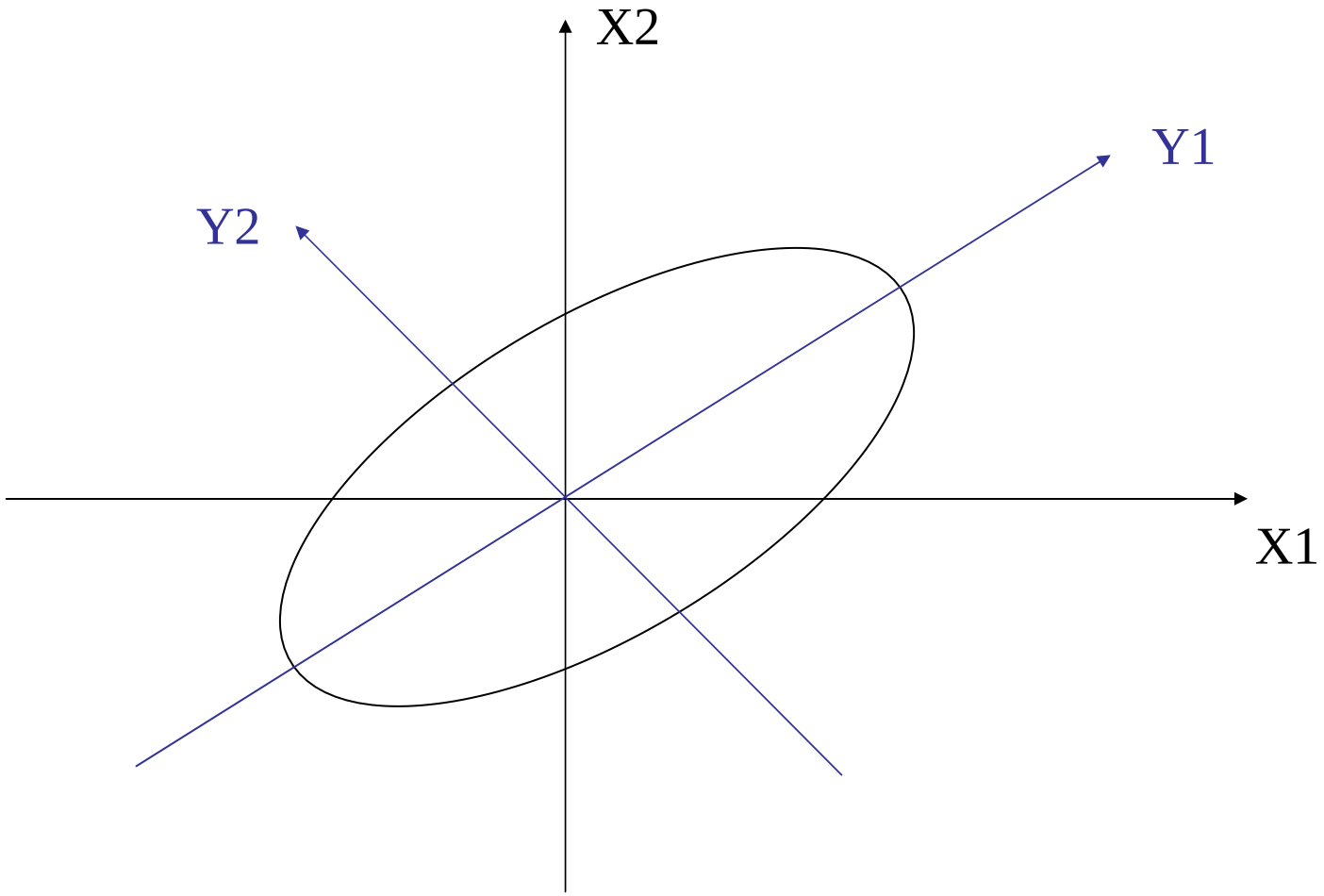




Dimensionality Reduction: Principal Component Analysis (PCA)

- Given N data vectors from n-dimensions, find $k \leq n$ orthogonal vectors (*principal components*) that can be best used to represent data
- Steps
  - Normalize input data: Each attribute falls within the same range
  - Compute k orthonormal (unit) vectors, i.e., *principal components*
  - Each input data (vector) is a linear combination of the k principal component vectors
  - The principal components are sorted in order of decreasing "significance" or strength
  - Since the components are sorted, the size of the data can be reduced by eliminating the weak components, i.e., those with low variance. (i.e., using the strongest principal components, it is possible to reconstruct a good approximation of the original data
- Works for numeric data only
- Used when the number of dimensions is large

Principal Component Analysis



Numerosity Reduction

- Reduce data volume by choosing alternative, smaller forms of data representation
- Parametric methods
  - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
  - Example: Log-linear models—obtain value at a point in m-D space as the product on appropriate marginal subspaces
- Non-parametric methods
  - Do not assume models
  - Major families: histograms, clustering, sampling

Data Reduction Method (1): Regression and Log-Linear Models

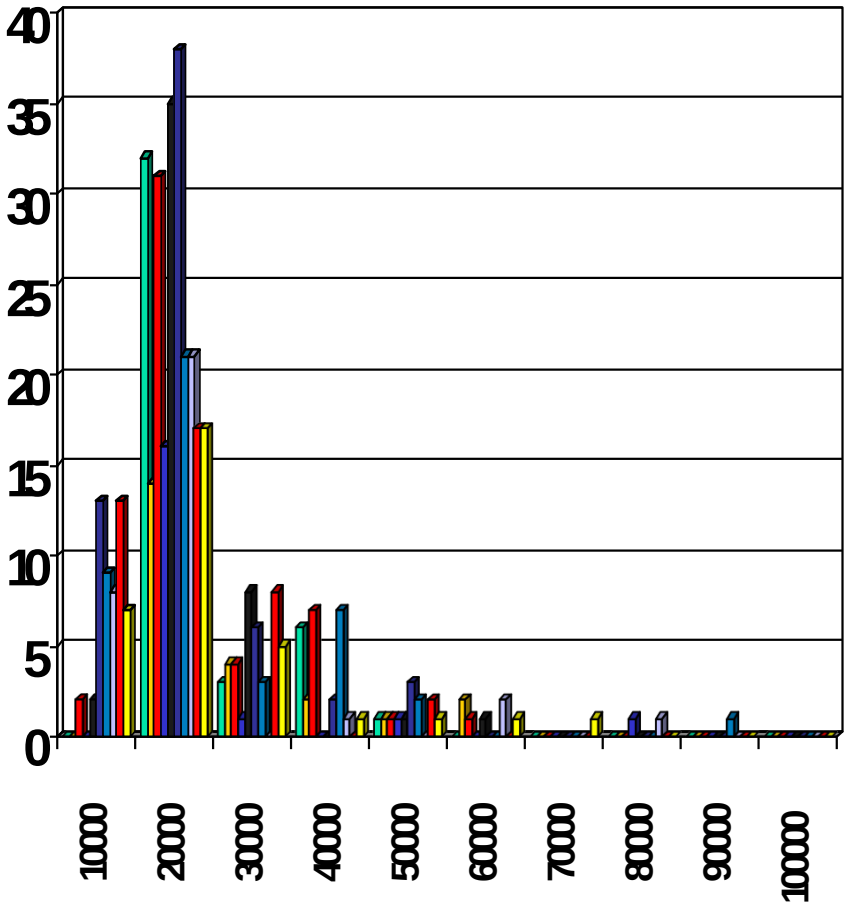
- Linear regression: Data are modeled to fit a straight line
  - Often uses the least-square method to fit the line
- Multiple regression: allows a response variable Y to be modeled as a linear function of multidimensional feature vector
- Log-linear model: approximates discrete multidimensional probability distributions

Regress Analysis and Log-Linear Models

- Linear regression: $Y = w X + b$
  - Two regression coefficients, w and b, specify the line and are to be estimated by using the data at hand
  - Using the least squares criterion to the known values of $Y1, Y2, \ldots, X1, X2, \ldots$.
- Multiple regression: $Y = b0 + b1 X1 + b2 X2$.
  - Many nonlinear functions can be transformed into the above
- Log-linear models:
  - The multi-way table of joint probabilities is approximated by a product of lower-order tables
  - Probability: $p(a, b, c, d) = \square ab \square ac \square ad \square bcd$

Data Reduction Method (2): Histograms

- Divide data into buckets and store average (sum) for each bucket
- Partitioning rules:
  - Equal-width: equal bucket range
  - Equal-frequency (or equal-depth)
  - V-optimal: with the least *histogram variance* (weighted sum of the original values that each bucket represents)
  - MaxDiff: set bucket boundary between each pair for pairs have the β–1 largest differences



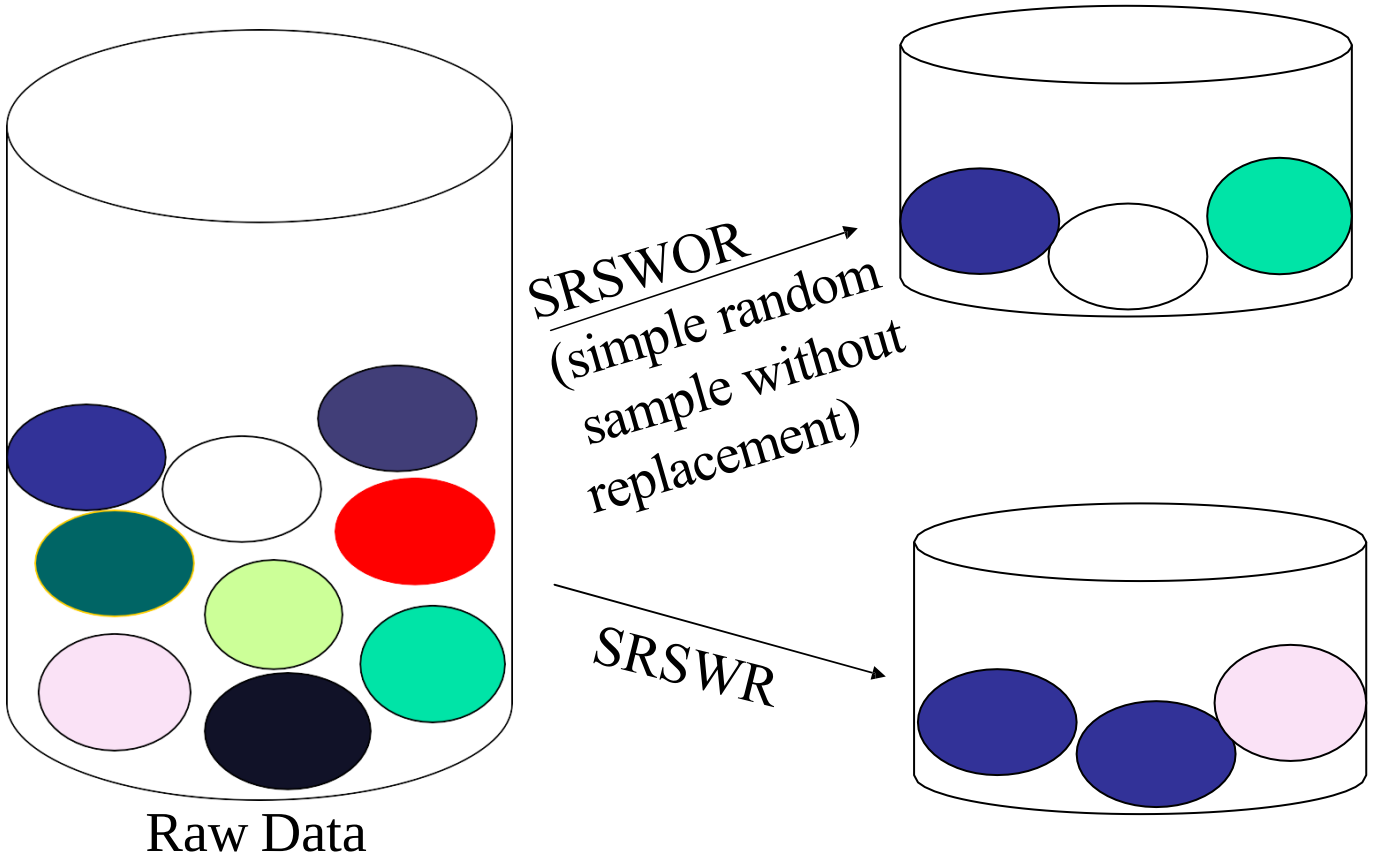
Data Reduction Method (3): Clustering

- Partition data set into clusters based on similarity, and store cluster representation (e.g., centroid and diameter) only
- Can be very effective if data is clustered but not if data is "smeared"
- Can have hierarchical clustering and be stored in multi-dimensional index tree structures
- There are many choices of clustering definitions and clustering algorithms
- Cluster analysis will be studied in depth in Chapter 7

Data Reduction Method (4): Sampling

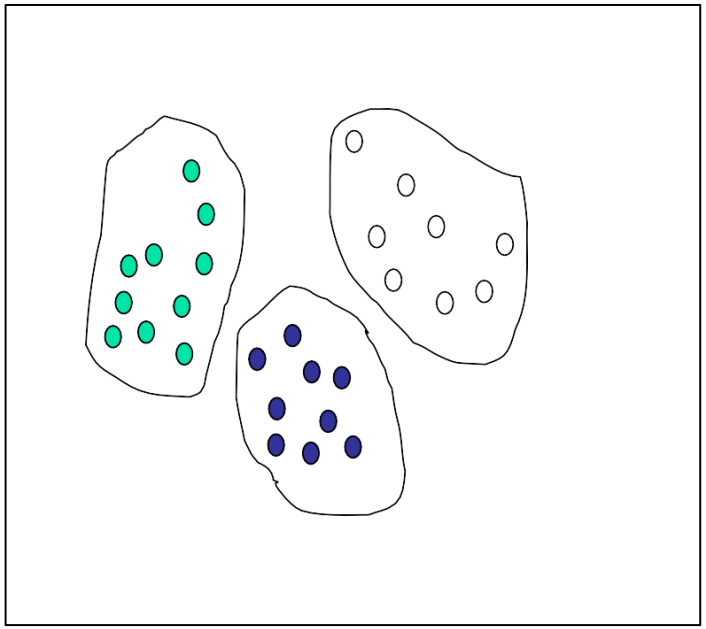
- Sampling: obtaining a small sample *s* to represent the whole data set *N*
- Allow a mining algorithm to run in complexity that is potentially sub-linear to the size of the data
- Choose a representative subset of the data
  - Simple random sampling may have very poor performance in the presence of skew
- Develop adaptive sampling methods
  - Stratified sampling:
    - Approximate the percentage of each class (or subpopulation of interest) in the overall database
    - Used in conjunction with skewed data
- Note: Sampling may not reduce database I/Os (page at a time)

Sampling: with or without Replacement

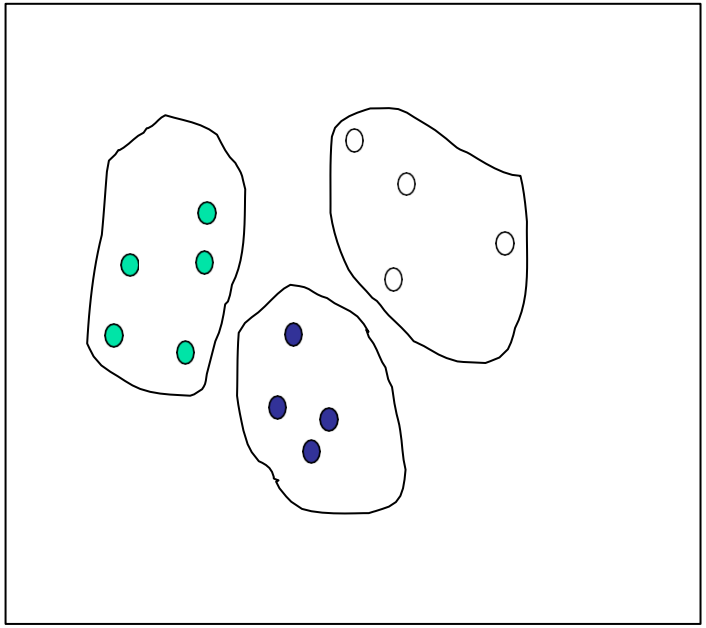


Sampling: Cluster or Stratified Sampling

Raw Data



Cluster/Stratified Sample



Discretization

- Three types of attributes:
  - Nominal — values from an unordered set, e.g., color, profession
  - Ordinal — values from an ordered set, e.g., military or academic rank
  - Continuous — real numbers, e.g., integer or real numbers
- Discretization:
  - Divide the range of a continuous attribute into intervals
  - Some classification algorithms only accept categorical attributes.
  - Reduce data size by discretization
  - Prepare for further analysis

Discretization and Concept Hierarchy

- Discretization
  - Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals
  - Interval labels can then be used to replace actual data values
  - Supervised vs. unsupervised
  - Split (top-down) vs. merge (bottom-up)
  - Discretization can be performed recursively on an attribute
- Concept hierarchy formation
  - Recursively reduce the data by collecting and replacing low level concepts (such as numeric values for age) by higher level concepts (such as young, middle-aged, or senior)

Data Mining Functionalities

Data Mining functionalities are used to specify the kind of patterns to be found in data mining tasks.

Data Mining tasks can be classified into two categories

- **Descriptive**: Characterize general properties of data in the database
- **Predictive**: perform inference on data to make predictions

Data Mining Functionalities: - Generalization Characterization and Discrimination

Data can be associated with **classes or concepts** that can be described in summarized, concise, and yet precise, terms.

Such descriptions of a concept or class are called **class/concept descriptions.**

These descriptions can be derived via

- Data Characterization
- Data Discrimination

Data Mining Functionalities: Characterization and Discrimination

Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a query.

ex: Description of all users who spent more than $10,000 a year at *AllElectronics?* A general profile of all customers, such as age, salary, location and credit ratings. Among all the customers meeting target condition (spent > $10,000), 10% are “Youth”, 60% are “Adults” and 30% are “Seniors”.

The output of data characterization can be presented in pie charts, bar charts, multidimensional data cubes, and multidimensional tables. They can also be presented in rule form.

Data Mining Functionalities Characterization and Discrimination

Data discrimination is a comparison of the target class data objects against the objects from one or multiple contrasting classes with respect to customers that share specified generalized feature(s).

ex: compare change in sales of software products for customers with given generalized feature: 40% of “Youth” have sales that increased by more 10% from last year; 10% of “Youth” have sales that decreased by at least 30% during the same period; the remaining 50% of “Youth” change in sales the fell in-between. “Youth” describes the generalized tuple, while increase in sales by > 10% is the target class. The other two amounts of change in sales are the contrasting classes.

The forms of output presentation are similar to those for characteristic descriptions, although discrimination descriptions should include comparative measures that help to distinguish between the target and contrasting classes.

Data Mining Functionalities
Association and Correlation Analysis

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association, correlation vs. causality
  - A typical association rule
    - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Data Mining Functionalities: Mining Frequent Patterns

Frequent patterns are the patterns that occur frequently in the data. Patterns can include itemsets, sequences and subsequences.

A frequent itemset refers to a set of items that often appear together in a transactional data set.

ex: bread and milk

Data Mining Functionalities: Mining Frequent Patterns

Association Rules

buys(X, “computer”)=>buys(X, “software”) [support =1%, confidence = 50%]

if a customer buys a computer, there is a 50% chance that he will buy software as well

Single Dimension Association Rule

1% of all the transactions under analysis show
that computer and software are purchased together

age(X, “20..29”)^income(X, “40K..49K”)=>buys(X, “laptop”)

[support = 2%, confidence = 60%]

Multi-Dimension Association Rule

Data Mining Functionalities Classification

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on (climate), or classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, …
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, …

Data Mining Functionalities: Classification and Prediction

Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts. The model is derived based on the analysis of a set of training data and is used to predict the class label of objects for which the the class label is unknown.

Representation of Derived model

IF-THEN Rules

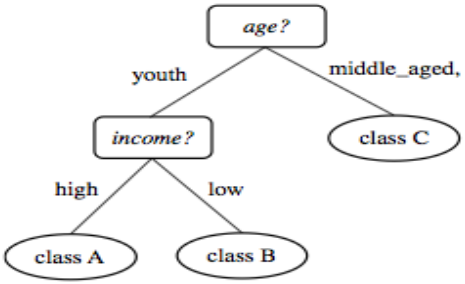
$age(X, \text{“youth”}) \; AND \; income(X, \text{“high”}) \rightarrow class(X, \text{“A”})$

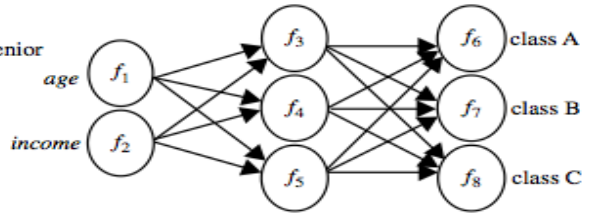
$age(X, \text{“youth”}) \; AND \; income(X, \text{“low”}) \rightarrow class(X, \text{“B”})$

$age(X, \text{“middle\_aged”}) \rightarrow class(X, \text{“C”})$

$age(X, \text{“senior”}) \rightarrow class(X, \text{“C”})$

Decision Tree





Data Mining Functionalities: Classification and Prediction

Prediction values continuous valued functions, i.e. it is used to predict missing or unavailable numeric data values rather than class labels.

Prediction can be used for both numeric prediction and class label prediction.

Regression analysis is a statistical method used numeric prediction.

Classification and regression may need to be preceded by relevance analysis, which attempts to identify attributes that are significantly relevant to the classification and regression process. Such attributes will be selected for the classification and regression process. Other attributes, which are irrelevant, can then be excluded from consideration

Data Mining Functionalities Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications

Data Mining Functionalities: Cluster Analysis

Clustering analyzes data objects without consulting class labels.

Clustering can be used to generate class labels for a group of data which did not exist at the beginning.

The objects are clustered or grouped based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity.

Data Mining Functionalities: Outlier Analysis

Outliers are data objects that do not comply with the general behavior or model of data.

Many data mining techniques discard outliers or exceptions as noise.

However, in some events these kind of events are more interesting. This analysis of outlier data is referred to as outlier analysis

ex: fraud detection.

Data Mining Functionalities
Evolution Analysis

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time.

This may include characterization, discrimination, association and correlation analysis, classification, prediction or clustering of time related data.

Distinct features of such data include time series data analysis, sequence or periodicity pattern matching and similarity based data analysis.

DM- Task Primitives

A data mining query is defined in terms of the following primitives

1. **Task-relevant data:**
   - Typically interested in only a subset of the entire database
   - Specify
     - the name of database/data warehouse (AllElectronics_db)
     - names of tables/data cubes containing relevant data (item, customer, purchases, items_sold)
     - conditions for selecting the relevant data (purchases made in Canada for relevant year)
     - relevant attributes or dimensions (name and price from item, income and age from customer)

Example,

suppose that you are a manager of All Electronics in charge of sales in the United States and Canada. In particular, you would like to study the buying trends of customers in Canada. Rather than mining on the entire database. These are referred to as relevant attributes

2. The kinds of knowledge to be mined:

- This specifies the data mining functions to be performed, such as characterization, discrimination, association, classification, clustering, or evolution analysis.
- For instance, if studying the buying habits of customers in Canada, you may choose to mine associations between customer profiles and the items that these customers like to buy

3. Background knowledge:

- Users can specify background knowledge, or knowledge about the domain to be mined.
- This knowledge is useful for guiding the knowledge discovery process, and for evaluating the patterns found.

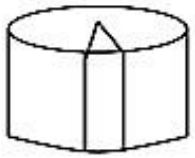
4. Interestingness measures:

- These functions are used to separate uninteresting patterns from knowledge.
- They may be used to guide the mining process, or after discovery, to evaluate the discovered patterns.
- Different kinds of measures are simplicity, certainty, utility, novelty.

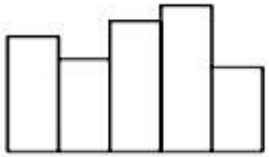
5. Presentation and visualization of discovered patterns:

- This refers to the form in which discovered patterns are to be displayed.
- Users can choose from different forms for knowledge presentation, such as rules, tables, charts, graphs, decision trees, and cubes.

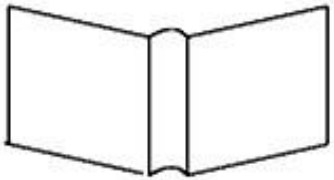
Figure : Primitives for specifying a data mining task.



Task-relevant data
- database or data warehouse name
- database tables or data warehouse cubes
- conditions for data selection
- relevant attributes or dimensions
- data grouping criteria

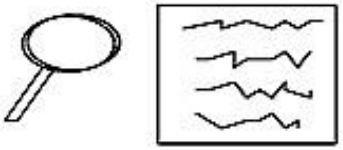


Knowledge type to be mined
- characterization
- discrimination
- association
- classification/prediction
- clustering



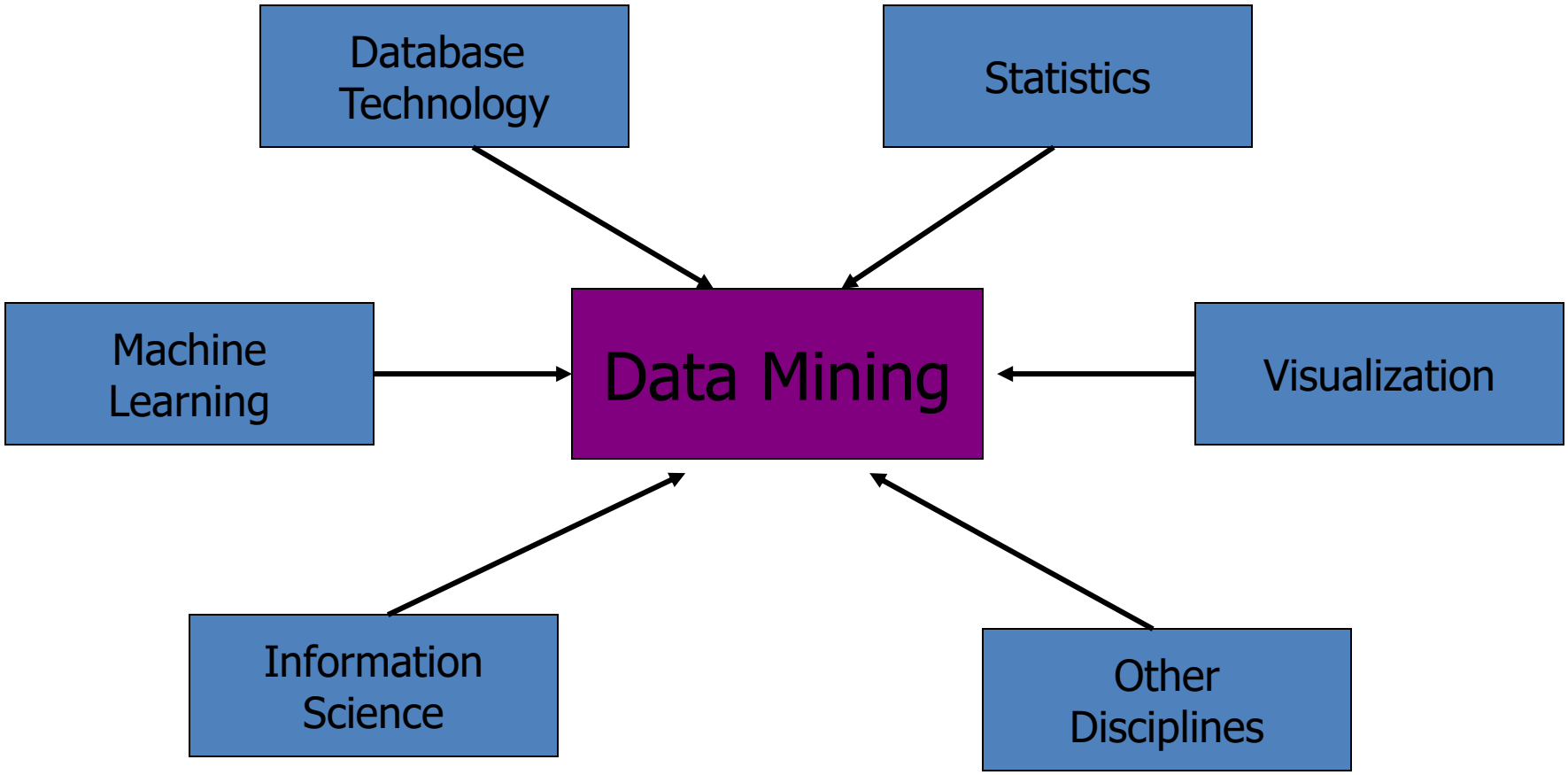
Background knowledge
- concept hierarchies
- user beliefs about relationships in the data

Pattern interestingness measurements
- simplicity
- certainty (e.g., confidence)
- utility (e.g., support)
- novelty



Visualization of discovered patterns
- rules, tables, reports, charts, graphs, decisison trees, and cubes
- drill-down and roll-up

Data Mining: Confluence of Multiple Disciplines



Data Mining: Classification Schemes

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views, different classifications
  - Kinds of databases to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted

A Multi-Dimensional View of Data Mining Classification

- **Databases to be mined**
  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

INTEGRATION OF DATAMINING SYSTEM WITH DATA WAREHOUSE

DB andDW systems, possible integration schemes include

- no coupling,
- loose coupling,
- semitight coupling, and
- tight coupling.

1.No coupling:

- *No coupling* means that a DM system will not utilize any function of a DB or DW system.
- It may fetch data from a particular source (such as a file system), process data using some data mining algorithms, and then store the mining results in another file.

2.Loose coupling:

- *Loose coupling* means that a DM system will use some facilities of a DB or DW system, fetching data from a data repository ,performing data mining, and then storing the mining results either in a file or in a designated place in a database or data Warehouse.
- Loose coupling is better than no coupling because it can fetch any portion of data stored in databases or data warehouses by using query processing, indexing, and other system facilities.
- It is difficult to achieve high scalability and good performance with large data sets.

3.Semitight coupling:

- *Semitight coupling* means that besides linking a DM system to a DB/DW system, a few essential data mining primitives can be provided in the DB/DW system.
- These primitives can include sorting, indexing, aggregation, histogram analysis, multi way join, and precomputation of some essential statistical measures, such as sum, count, max, min ,standard deviation.

4.Tight coupling:

- *Tight coupling* means that a DM system is smoothly integrated into the DB/DW system.
- The data mining subsystem is treated as one functional component of information system.
- Data mining queries and functions are optimized based on mining query analysis, data structures, indexing schemes, and query processing methods of a DB or DW system.

Applications of Data Mining

- Web page analysis: from web page classification, clustering to PageRank & HITS algorithms
- Collaborative analysis & recommender systems
- Basket data analysis to targeted marketing
- Biological and medical data analysis: classification, cluster analysis (microarray data analysis), biological sequence analysis, biological network analysis
- Data mining and software engineering (e.g., IEEE Computer, Aug. 2009 issue)
- From major dedicated data mining systems/tools (e.g., SAS, MS SQL-Server Analysis Manager, Oracle Data Mining Tools) to invisible data mining

Issues in Data Mining

- Mining Methodology
  - Mining various and new kinds of knowledge
  - Mining knowledge in multi-dimensional space
  - Data mining: An interdisciplinary effort
  - Boosting the power of discovery in a networked environment
  - Handling noise, uncertainty, and incompleteness of data
  - Pattern evaluation and pattern- or constraint-guided mining
- User Interaction
  - Interactive mining
  - Incorporation of background knowledge
  - Presentation and visualization of data mining results

Issues in Data Mining

- Efficiency and Scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed, stream, and incremental mining methods
- Diversity of data types
  - Handling complex types of data
  - Mining dynamic, networked, and global data repositories
- Data mining and society
  - Social impacts of data mining
  - Privacy-preserving data mining
  - Invisible data mining