DATA MINING
UNIT-I : INTRODUCTION

DR. P. BALAMURUGAN, ASST. PROFESSOR
DEPARTMENT OF COMPUTERE SCIENCE
GOVERNMENT ARTS COLLEGE, COIMBATORE-18
spbalamurugan@rediffmail.com

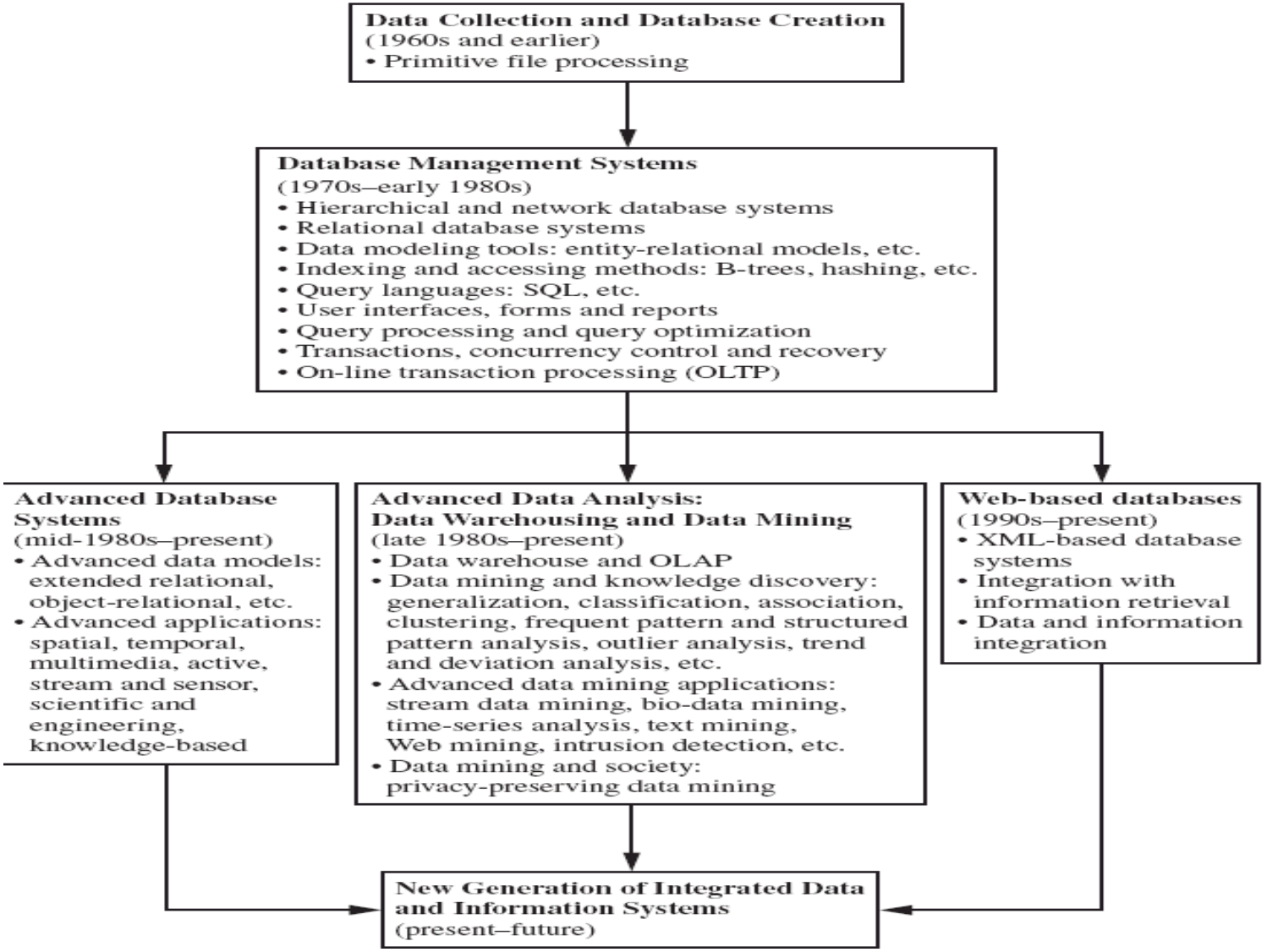
Outline

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data mining functionality
- Are all the patterns interesting?
- Classification of data mining systems
- Data Mining Task Primitives
- Integration of data mining system with a DB and DW System
- Major issues in data mining

Why Data Mining?

- The Explosive Growth of Data: from terabytes(1000^4) to yottabytes(1000^8)
  - Data collection and data availability
    - Automated data collection tools, database systems, web
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: bioinformatics, scientific simulation, medical research …
    - Society and everyone: news, digital cameras, …
- Data rich but information poor!
  - What does those data mean?
  - How to analyze data?
- Data mining — Automated analysis of massive data sets

Evolution of Database Technology



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.

Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis

Ex.: Market Analysis and Management

- Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, surveys …
- Target marketing
  - Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.,
    - E.g. Most customers with income level 60k – 80k with food expenses $600 - $800 a month live in that area
  - Determine customer purchasing patterns over time
    - E.g. Customers who are between 20 and 29 years old, with income of 20k – 29k usually buy this type of CD player
- Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association
  - E.g. Customers who buy computer A usually buy software B

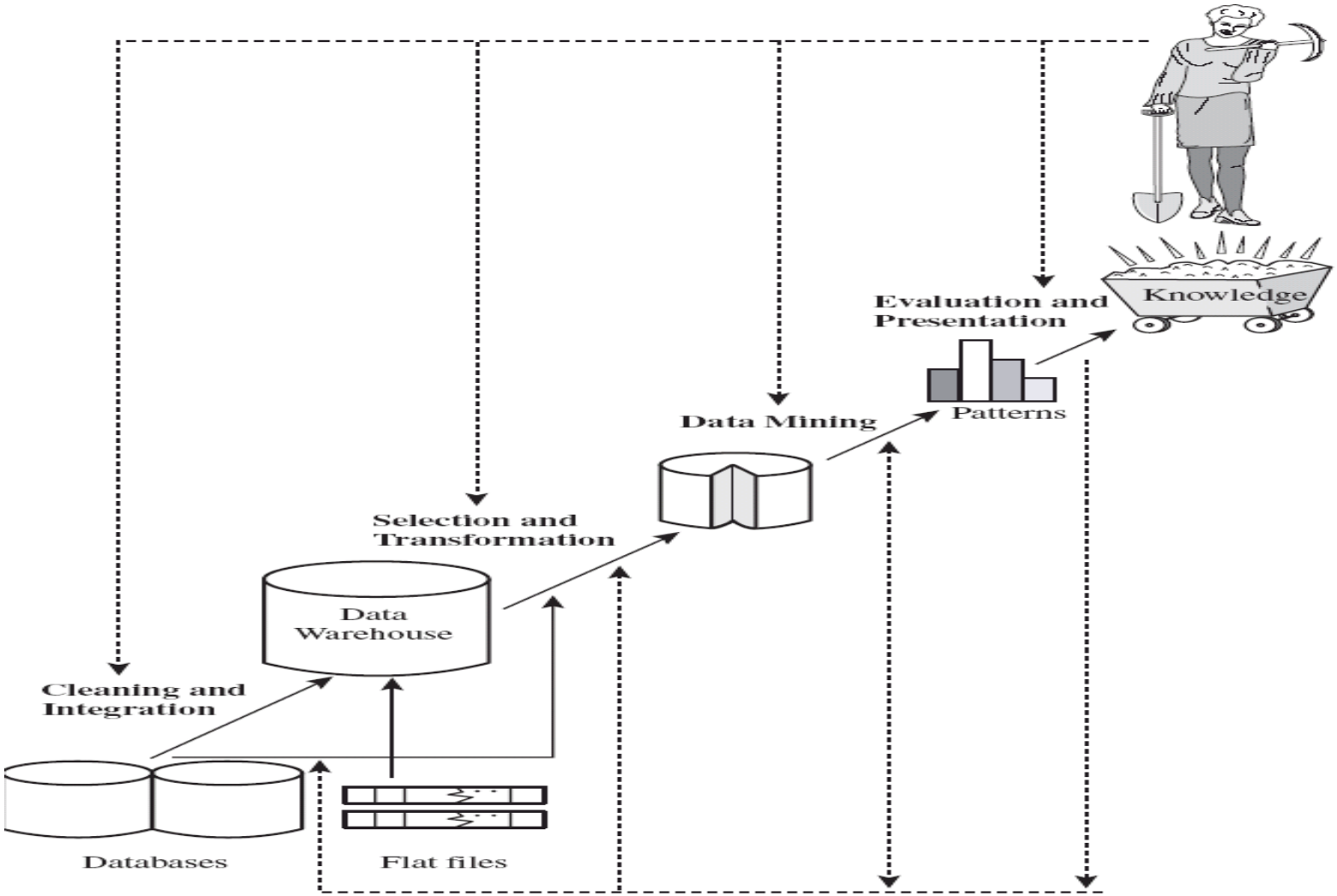
Ex.: Market Analysis and Management (2)

- Customer requirement analysis
  - Identify the best products for different customers
  - Predict what factors will attract new customers
- Provision of summary information
  - Multidimensional summary reports
    - E.g. Summarize all transactions of the first quarter from three different branches

      Summarize all transactions of last year from a particular branch

      Summarize all transactions of a particular product
  - Statistical summary information
    - E.g. What is the average age for customers who buy product A?
- Fraud detection
  - Find outliers of unusual transactions
- Financial planning
  - Summarize and compare the resources and spending

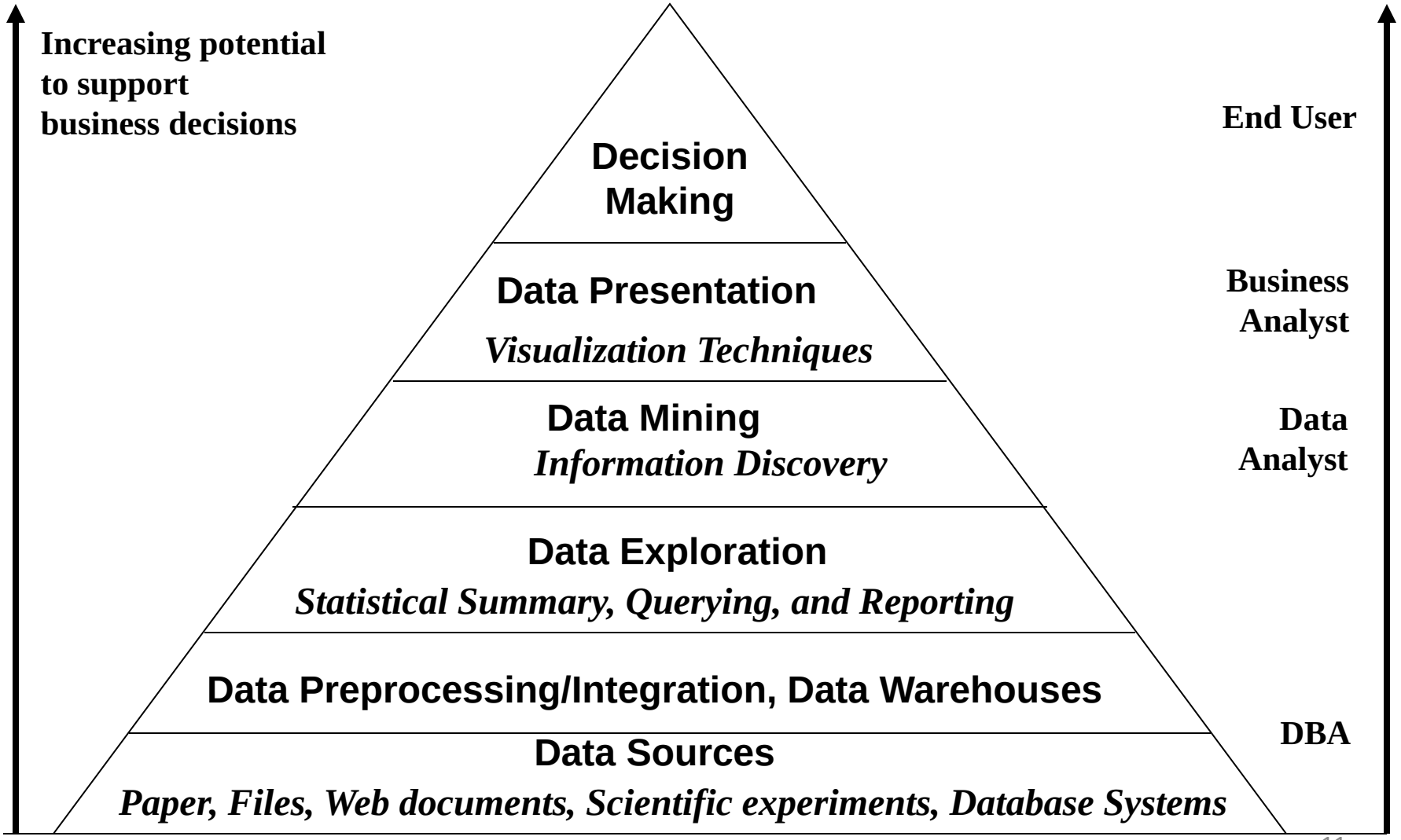
Knowledge Discovery (KDD) Process



KDD Process: Several Key Steps

- Learning the application domain
  - relevant prior knowledge and goals of application
- Identifying a target data set: data selection
- Data processing
  - **Data cleaning** remove noise and inconsistent data
  - **Data integration** multiple data sources maybe combined
  - **Data selection** data relevant to the analysis task are retrieved from database
  - **Data transformation** transformed/consolidated into forms appropriate for mining
  - **Data mining** extract data patterns
  - **Pattern evaluation** identify the truly interesting patterns
  - **Knowledge presentation** mined knowledge is presented to the user with visualization or representation techniques
- Use of discovered knowledge

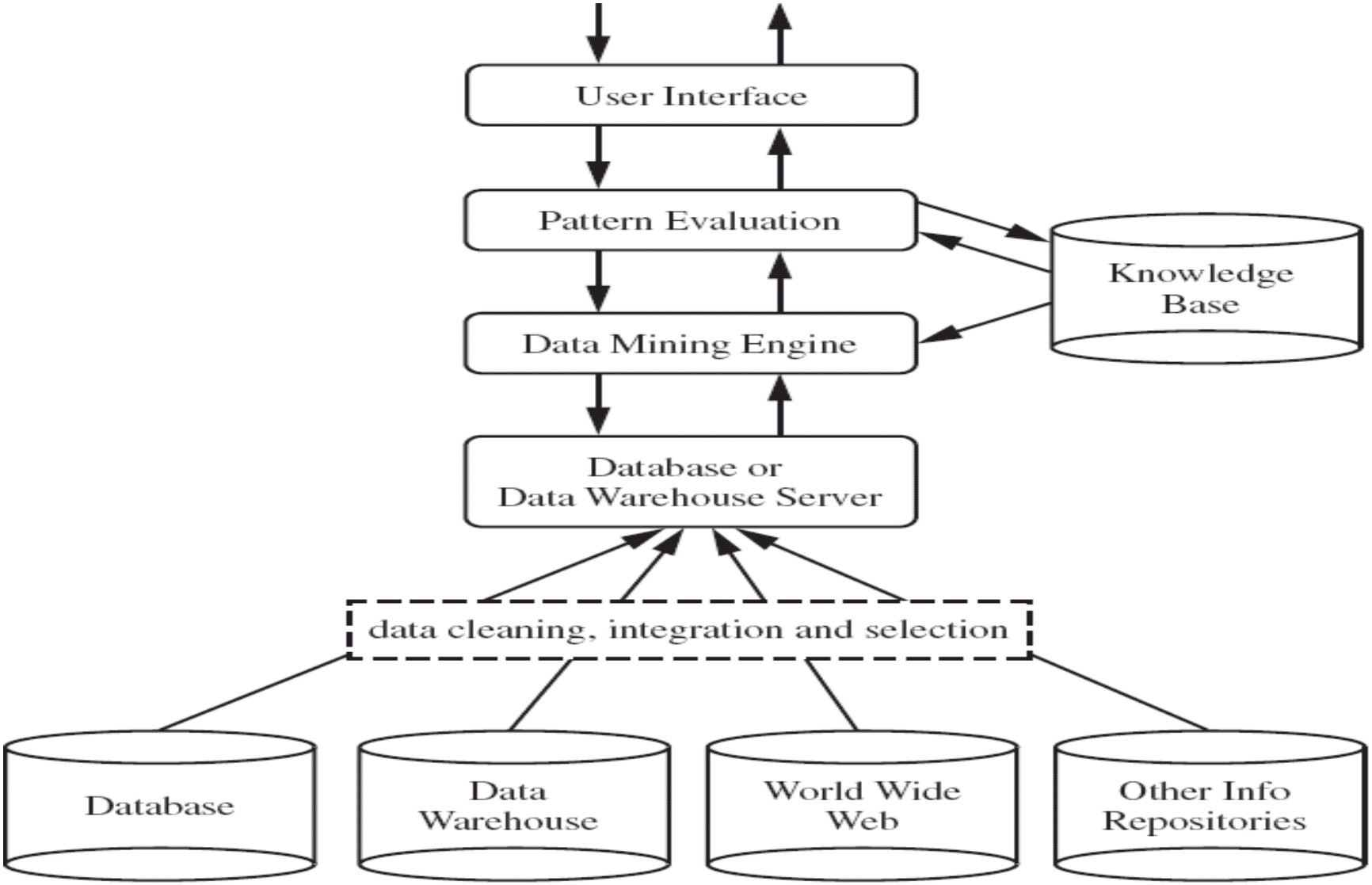
Data Mining and Business Intelligence



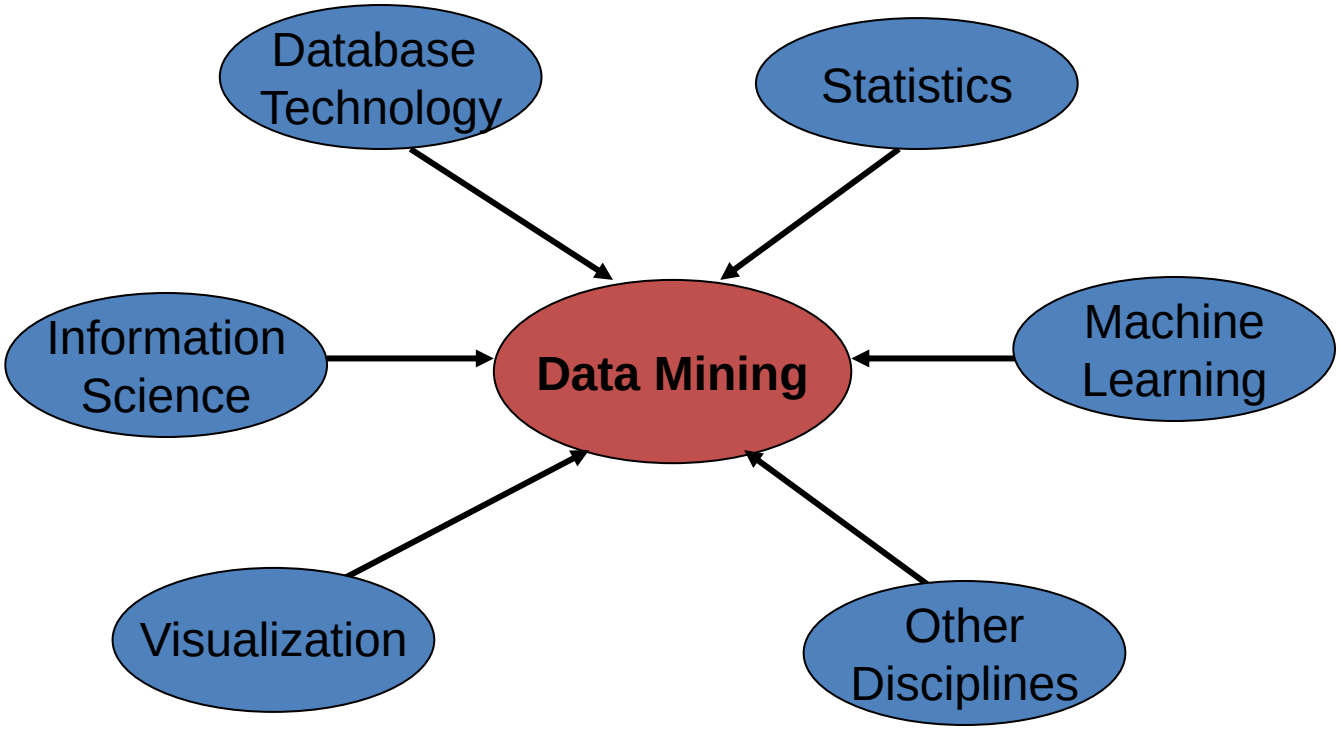
A typical DM System Architecture

- Database, data warehouse, WWW or other information repository (store data)
- Database or data warehouse server (fetch and combine data)
- Knowledge base (turn data into meaningful groups according to domain knowledge)
- Data mining engine (perform mining tasks)
- Pattern evaluation module (find interesting patterns)
- User interface (interact with the user)

A typical DM System Architecture (2)



Confluence of Multiple Disciplines



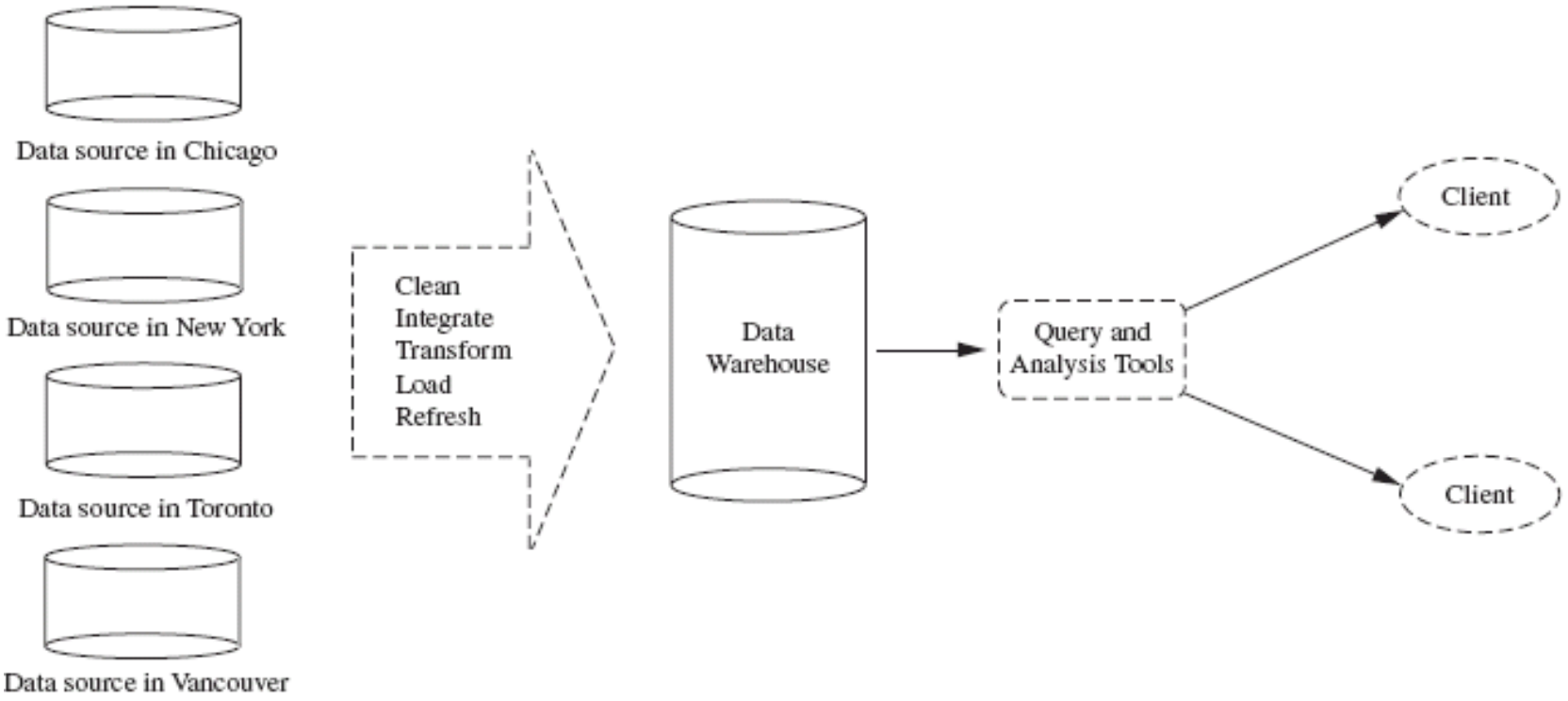
- Not all "Data Mining System" performs true data mining
  - machine learning system, statistical analysis (small amount of data)
  - Database system (information retrieval, deductive querying…)

On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Object-Relational Databases
  - Temporal Databases, Sequence Databases, Time-Series databases
  - Spatial Databases and Spatiotemporal Databases
  - Text databases and Multimedia databases
  - Heterogeneous Databases and Legacy Databases
  - Data Streams
  - The World-Wide Web

Data Warehouses

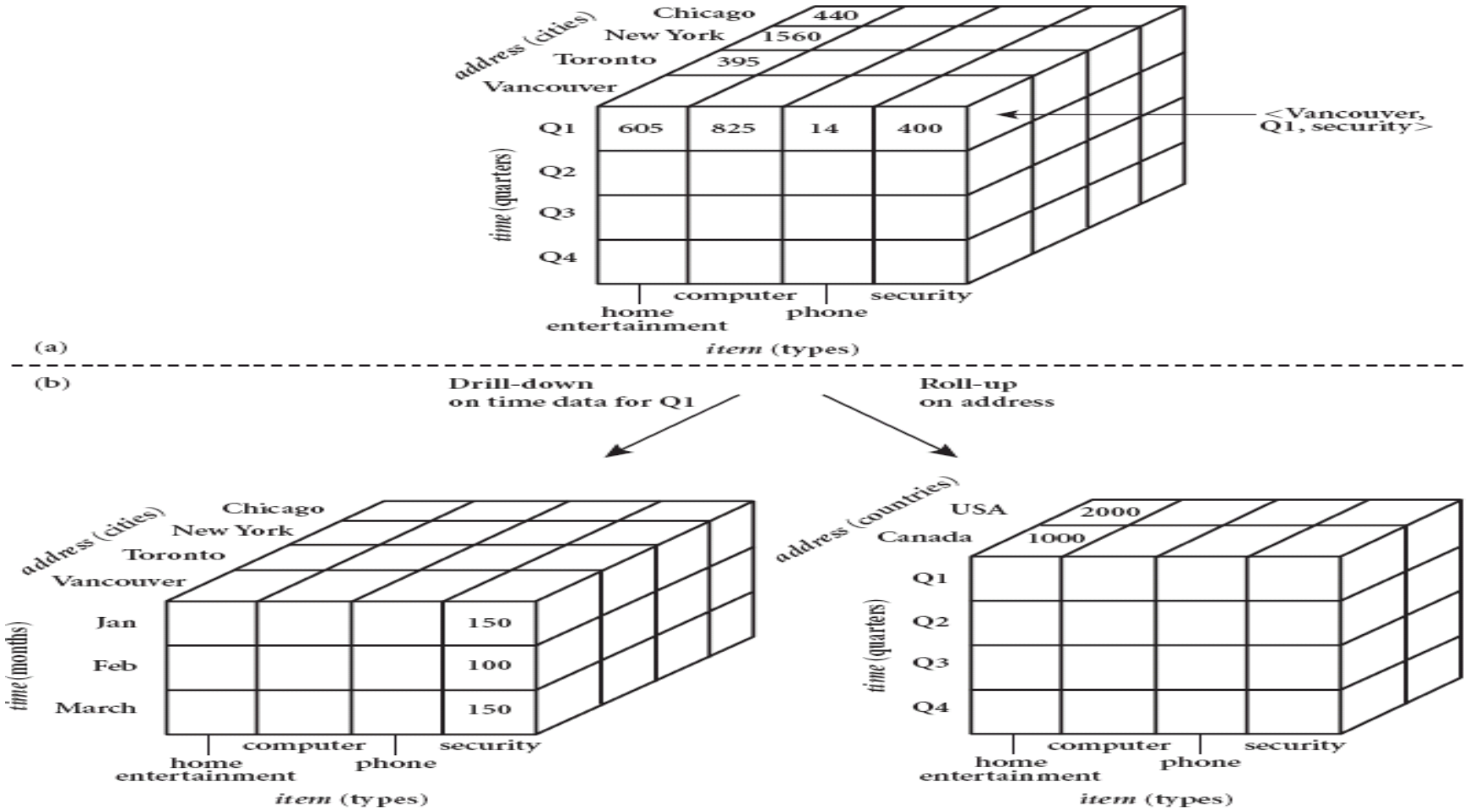
- A repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.
- Constructed via a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing.



Data Warehouses (2)

- Data are organized around major subjects, e.g. customer, item, supplier and activity.
- Provide information from a historical perspective (e.g. from the past 5 – 10 years)
- Typically summarized to a higher level (e.g. a summary of the transactions per item type for each store)
- User can perform drill-down or roll-up operation to view the data at different degrees of summarization

Data Warehouses (3)



Transactional Databases

- Consists of a file where each record represents a transaction
- A transaction typically includes a unique transaction ID and a list of the items making up the transaction.

trans_ID	*list of item_IDs*
T100	I1, I3, I8, I16
T200	I2, I8
. . .	. . .

- Either stored in a flat file or unfolded into relational tables
- Easy to identify items that are frequently sold together

Data Mining Functionalities

- Concept/Class Description: Characterization and Discrimination
  - Data can be associated with classes or concepts.

    E.g. classes of items – computers, printers, …

    concepts of customers – bigSpenders, budgetSpenders, …
  - How to describe these items or concepts?
  - Descriptions can be derived via Data characterization – summarizing the general characteristics of a target class of data.

    E.g. summarizing the characteristics of customers who spend more than $1,000 a year at *AllElectronics*.

    Result can be a general profile of the customers, such as 40 – 50 years old, employed, have excellent credit ratings. `

Data Mining Functionalities

- Data discrimination – comparing the target class with one or a set of comparative classes
  - E.g. Compare the general features of software products whole sales increase by 10% in the last year with those whose sales decrease by 30% during the same period
- Or both of the above

➢ Mining Frequent Patterns, Associations and Correlations

- Frequent itemset: a set of items that frequently appear together in a transactional data set (e.g. milk and bread)
- Frequent subsequence: a pattern that customers tend to purchase product A, followed by a purchase of product B

Data Mining Functionalities

- Association Analysis: find frequent patterns
  - E.g. a sample analysis result – an association rule:

    buys(X, “computer”) => buys(X, “software”)

    [support = 1%, confidence = 50%]

    (if a customer buys a computer, there is a 50% chance that she will buy software. 1% of all of the transactions under analysis showed that computer and software are purchased together.)
  - Associations rules are discarded as uninteresting if they do not satisfy both a minimum support threshold and a minimum confidence threshold.
- Correlation Analysis: additional analysis to find statistical correlations between associated pairs

Data Mining Functionalities

- Classification and Prediction
  - Classification
    - The process of finding a model that describes and distinguishes the data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown.
    - The derived model is based on the analysis of a set of training data (data objects whose class label is known).
    - The model can be represented in *classification (IF-THEN) rules,* decision trees, *neural networks,* etc.
  - Prediction
    - Predict missing or unavailable numerical data values

Data Mining Functionalities

(a)

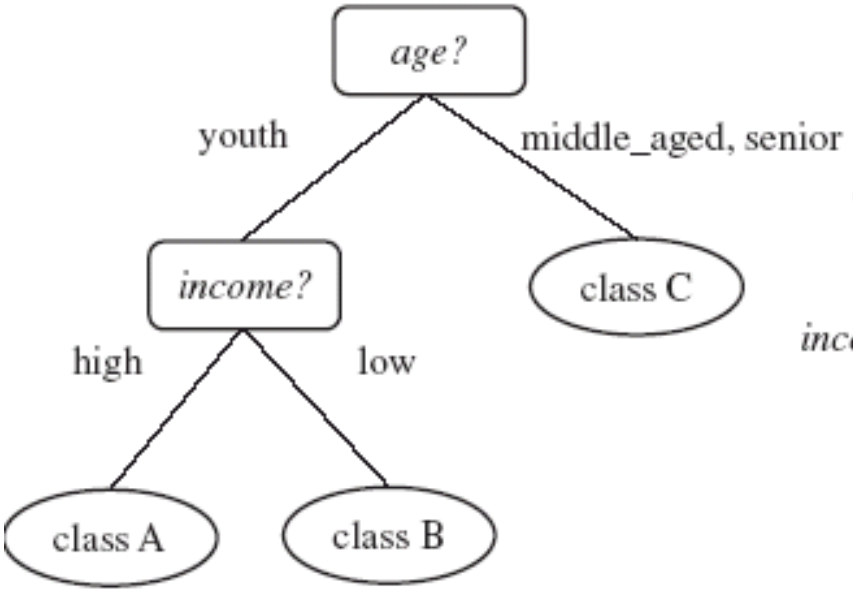
age(X, "youth") AND income(X, "high") ⟶ class(X, "A")

age(X, "youth") AND income(X, "low") ⟶ class(X, "B")

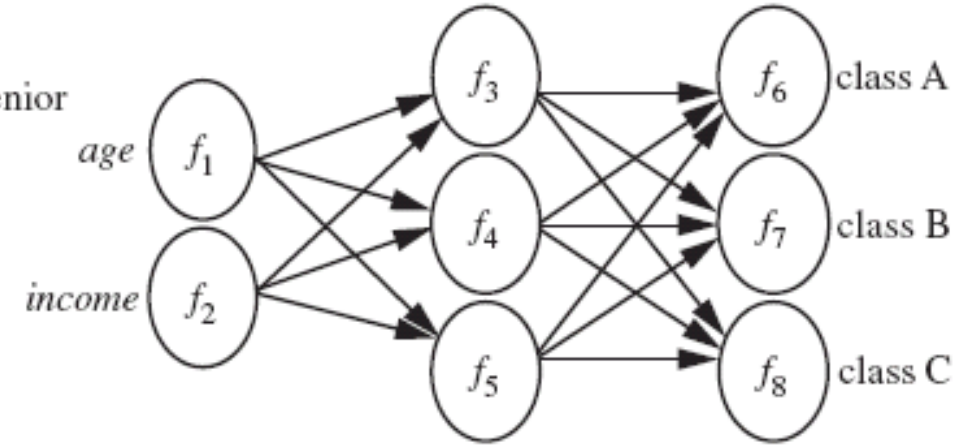
age(X, "middle_aged") ⟶ class(X, "C")

age(X, "senior") ⟶ class(X, "C")

(b)

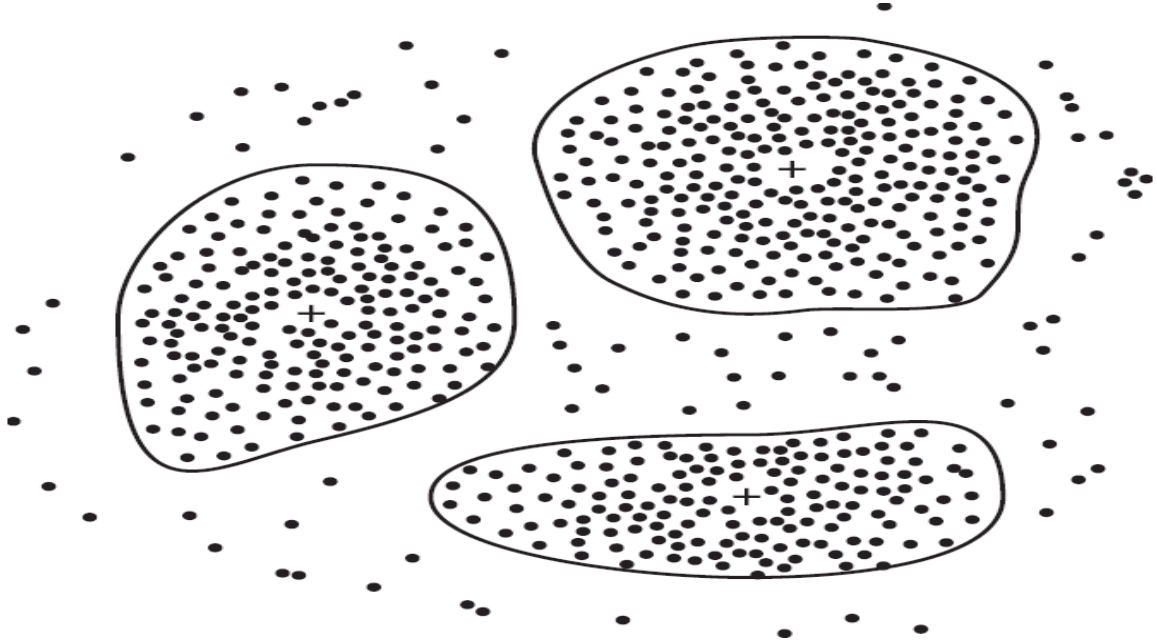


(c)



Data Mining Functionalities (2)

- Cluster Analysis
  - Class label is unknown: group data to form new classes
  - Clusters of objects are formed based on the principle of *maximizing intra-class similarity & minimizing interclass similarity*
    - E.g. Identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing.



Data Mining Functionalities (2)

- Outlier Analysis
  - Data that do no comply with the general behavior or model.
  - Outliers are usually discarded as noise or exceptions.
  - Useful for fraud detection.
    - E.g. Detect purchases of extremely large amounts
- Evolution Analysis
  - Describes and models regularities or trends for objects whose behavior changes over time.
    - E.g. Identify stock evolution regularities for overall stocks and for the stocks of particular companies.

PATTERN

- A **pattern** means that **data** are correlated that they have a relationship.
- Patterns are predictable.
- When you have a lack of **pattern**, you have true randomness.
- When you find a **pattern**, you can have a good idea when or where something will happen before it actually happens.
- A **pattern** is a series of **data** that repeats in a recognizable way. It can be identified in the history of the asset being evaluated or other assets with similar characteristics.

FREQUENT PATTERN

- Frequent patterns, as the name suggests, are patterns that occur frequently in data.
- Frequent patterns are item sets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold.
- For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set, is a frequent item set.
- A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern.

FREQUENT PATTERN

- A substructure can refer to different structural forms, such as subgraphs subtrees, or sub-lattices, which may be combined with item sets or subsequences.
- If a substructure occurs frequently, it is called a (frequent) structured pattern. Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

Are All of the Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
- A pattern is interesting if it is
  - easily understood by humans
  - valid on new or test data with some degree of certainty,
  - potentially useful
  - novel
  - validates some hypothesis that a user seeks to confirm
- An interesting measure represents *knowledge* !

Are All of the Patterns Interesting?

- Objective measures
  - Based on statistics and structures of patterns, e.g., support, confidence, etc. (Rules that do not satisfy a threshold are considered uninteresting.)
- Subjective measures
  - Reflect the needs and interests of a particular user.
    - E.g. A marketing manager is only interested in characteristics of customers who shop frequently.
  - Based on user's belief in the data.
    - e.g., Patterns are interesting if they are unexpected, or can be used for strategic planning, etc
- Objective and subjective measures need to be combined.

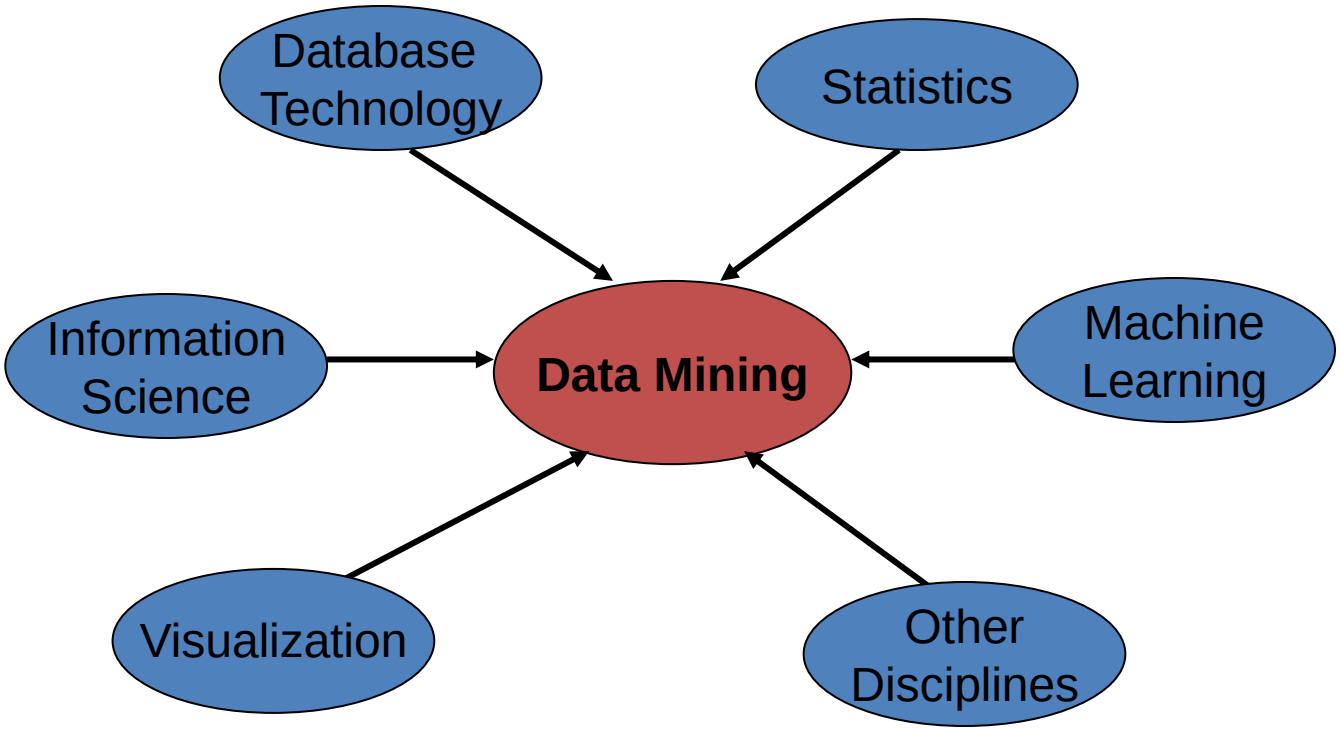
Are All of the Patterns Interesting?

- Find all the interesting patterns: Completeness
  - Unrealistic and inefficient
  - User-provided constraints and interestingness measures should be used
- Search for only interesting patterns: An optimization problem
  - Highly desirable
  - No need to search through the generated patterns to identify truly interesting ones.
  - Measures can be used to rank the discovered patterns according their interestingness.

Classification of data mining systems

- Data mining is an interdisciplinary field, the confluence of a set of disciplines.
- Because of the diversity of disciplines contributing to datamining, datamining research is expected to generate a large variety of data mining systems.

Classification of data mining systems



Classification of data mining systems

- **Classification according to Database**
  - Relational, data warehouse, transactional, stream, object-oriented/ relational, spatial, time-series, text data, multi-media, heterogeneous, WWW
- **Classification according to Knowledge**
  - Characterization, discrimination, association, classification, clustering, outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels(High level data, Raw level data, Multiple level data)
- **Classification according to Techniques utilized**
  - Degree of user interaction involved (Autonomous system, Query driven system, interactive exploratory system) or methods of data analysis involved (Database-oriented, data warehouse, machine learning, statistics, visualization, etc.)

Classification of data mining systems

- **Classification according to Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA, stock market analysis, text mining, Web mining, etc.
  - Different applications often require the integration of application specific methods. Therefore, a generic, all-purpose data mining system may not fit domain-specific mining tasks.

Data Mining Task Primitives

- Each user will have a data mining task in mind, that is, some form of data analysis that he or she would like to have performed.
- A data mining task can be specified in the form of a data mining query, which is input to the data mining system.
- A data mining query is defined in terms of data mining task primitives.
- How to construct a data mining query?
  - The primitives allow the user to interactively communicate with the data mining system during discovery to direct the mining process, or examine the findings

Data Mining Task Primitives

- The primitives specify:

  (1) The set of task-relevant data – which portion of the database to be used

    - Database or data warehouse name
    - Database tables or data warehouse cubes
    - Condition for data selection
    - Relevant attributes or dimensions
    - Data grouping criteria

Data Mining Task Primitives

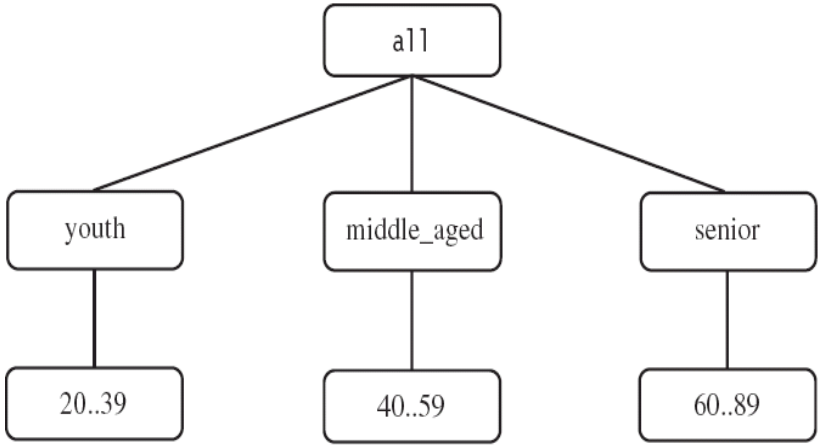
- The primitives specify:

  (2) The kind of knowledge to be mined – what DB functions to be performed

  - Characterization
  - Discrimination
  - Association
  - Classification/prediction
  - Clustering
  - Outlier analysis
  - Other data mining tasks

Data Mining Task Primitives

(3) The background knowledge to be used – what domain knowledge, concept hierarchies, etc.



(4) Interestingness measures and thresholds – support, confidence, etc.

(5) Visualization methods – what form to display the result, e.g. rules, tables, charts, graphs, …

DMQL

- DMQL – Data Mining Query Language
  - Designed to incorporate these primitives
  - Allow user to interact with DM systems
  - Providing a standardized language like SQL

Why Data Mining Query Language?

- Automated vs. query-driven?
  - Finding all the patterns autonomously in a database?—unrealistic because the patterns could be too many but uninteresting
- Data mining should be an interactive process
  - User directs what to be mined
- Users must be provided with a set of primitives to be used to communicate with the data mining system
- Incorporating these primitives in a data mining query language
  - More flexible user interaction
  - Foundation for design of graphical user interface
  - Standardization of data mining industry and practice

Major Issues in Data Mining

- Mining different kinds of knowledge
  - DM should cover a wide spectrum of data analysis and knowledge discovery tasks
  - Enable to use the database in different ways
  - Require the development of numerous data mining techniques
- Interactive mining of knowledge at multiple levels of abstraction
  - Difficult to know exactly what will be discovered
  - Allow users to focus the search, refine data mining requests
- Incorporation of background knowledge
  - Guide the discovery process
  - Allow discovered patterns to be expressed in concise terms and different levels of abstraction

Major Issues in Data Mining

- Data mining query languages and ad hoc data mining
  - High-level query languages need to be developed
  - Should be integrated with a DB/DW query language
- Presentation and visualization of results
  - Knowledge should be easily understood and directly usable
  - High level languages, visual representations or other expressive forms
  - Require the DM system to adopt the above techniques
- Handling noisy or incomplete data
  - Require data cleaning methods and data analysis methods that can handle noise
- Pattern evaluation – the interestingness problem
  - How to develop techniques to access the interestingness of discovered patterns, especially with subjective measures based on user beliefs or expectations

Major Issues in Data Mining

- Performance Issues
  - Efficiency and scalability
    - Huge amount of data
    - Running time must be predictable and acceptable
  - Parallel, distributed and incremental mining algorithms
    - Divide the data into partitions and processed in parallel
    - Incorporate database updates without having to mine the entire data again from scratch

Major Issues in Data Mining

- Diversity of Database Types
  - Other database that contain complex data objects, multimedia data, spatial data, etc.
  - Expect to have different DM systems for different kinds of data
  - Heterogeneous databases and global information systems
    - Web mining becomes a very challenging and fast-evolving field in data mining.

ASSOCIATION AND CORRELATIONS

- Associations and Correlations are called relationships among item sets. Ther e are some other interesting relationships used to find frequent item sets.

FREQUENT ITEM SET MINING METHODS

Apriori Algorithm

- Apriori algorithm was proposed by R. Agrawal and R. Srikant in 1994 for mining frequent itemsets for Boolean association rules.
- It uses prior knowledge of frequent itemset properties.
- Apriori employs an iterative approach known as a level-wise search, where k-itemsets are used to explore (k+1)-itemsets.
- Apriori property: All nonempty subsets of a frequent itemset must also be frequent.
- Apriori Property will help to improve the efficiency and reduce the search space.

FREQUENT ITEM SET MINING METHODS

Demonstration – Finding Frequent Item sets

- Sample - Transactional data for an AllElectronics branch.

TID List of item	IDs
T100	I1, I2, I5
T200	I2,I4
T300	I2,I3
T400	I1,I2,I4
T500	I1,I3
T600	I2,I3
T700	I1,I3
T800	I1,I2,I3,I5
T900	I1,I2,I3

FREQUENT ITEM SET MINING METHODS

Demonstration – Finding Frequent Item sets

1) Scan database (D) for count of each candidate(C_i). Compare candidate support count with minimum support count and find L_i, where i=1.

(C_1)

Itemset
{I1}
{I2}
{I3}
{I4}
{I5}

(C_1)

Itemset	Support Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

(L_1)

Itemset	Support Count
{I1}	6
{I2}	7
{I3}	6
{I4}	2
{I5}	2

FREQUENT ITEM SET MINING METHODS

Demonstration – Finding Frequent Item sets

1) Generate C_2 candidate from L_1 and scan D for counting each candidate.

Compare Candidate support with minimum support and find L_2

(C_2)

Itemset
{I1,I2}
{I1,I3}
{I1,I4}
{I1,I5}
{I2,I3}
{I2,I4}
{I2,I5}
{I3,I4}
{I3,I5}
{I4,I5}

(C_2)

Itemset	Support Count
{I1,I2}	4
{I1,I3}	4
{I1,I4}	1
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2
{I3,I4}	0
{I3,I5}	1
{I4,I5}	0

(L_2)

Itemset	Support Count
{I1,I2}	4
{I1,I3}	4
{I1,I5}	2
{I2,I3}	4
{I2,I4}	2
{I2,I5}	2

FREQUENT ITEM SET MINING METHODS

Demonstration – Finding Frequent Item sets

1) Generate C_3 candidate from L_2 and scan D for counting each candidate.

Compare Candidate support with minimum support and find L_3

(C_3)

Itemset
{I1,I2,I3}
{I1,I2,I5}
{I1,I2,I4}
{I1,I3,I5}
{I2,I3,I4}
{I2,I3,I5}
{I2,I4,I5}

(C_3)

Itemset	Support Count
{I1,I2,I3}	2
{I1,I2,I5}	2
{I1,I2,I4}	1
{I1,I3,I5}	1
{I2,I3,I4}	0
{I2,I3,I5}	1
{I2,I4,I5}	0

(L_3)

Itemset	Support Count
{I1,I2,I3}	2
{I1,I2,I5}	2