Advanced Database and Data Mining

CS-513

Faculty-Dr Aruna Malik

Introduction

Outline

- 1.1 Motivation: Why data mining?
- 1.2 What is data mining?
- 1.3 Data Mining: On what kind of data?
- 1.4 Data mining functionality: What kinds of Patterns Can Be Mined?
- 1.5 Are all the patterns interesting?
- 1.6 Classification of data mining systems
- 1.7 Data Mining Task Primitives
- 1.8 Integration of data mining system with a DB and DW System
- 1.9 Major issues in data mining

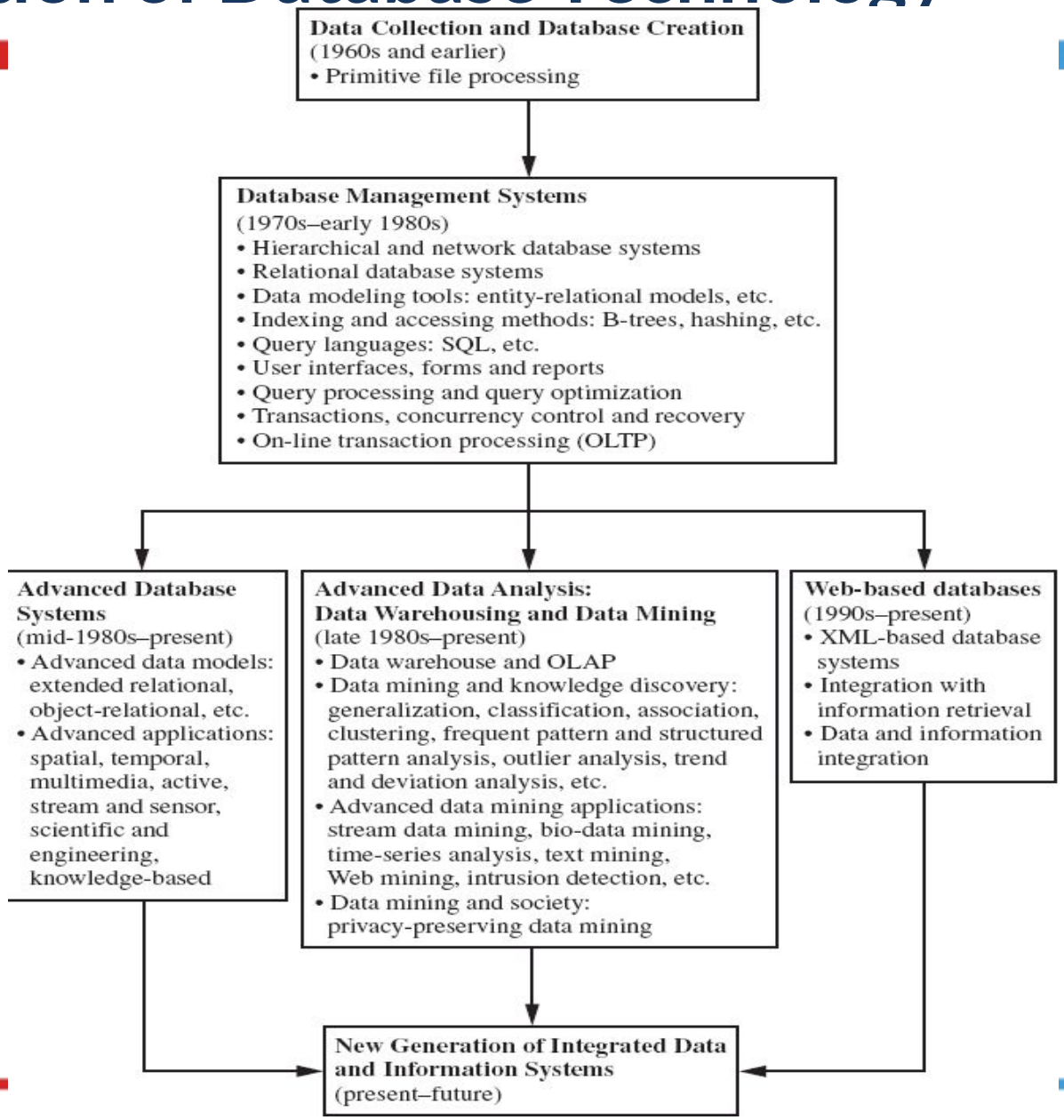
1.1 Why Data Mining?

- The Explosive Growth of Data: from terabytes(1000^4) to petabytes
  - 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

1.2 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.

Evolution of Database Technology



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

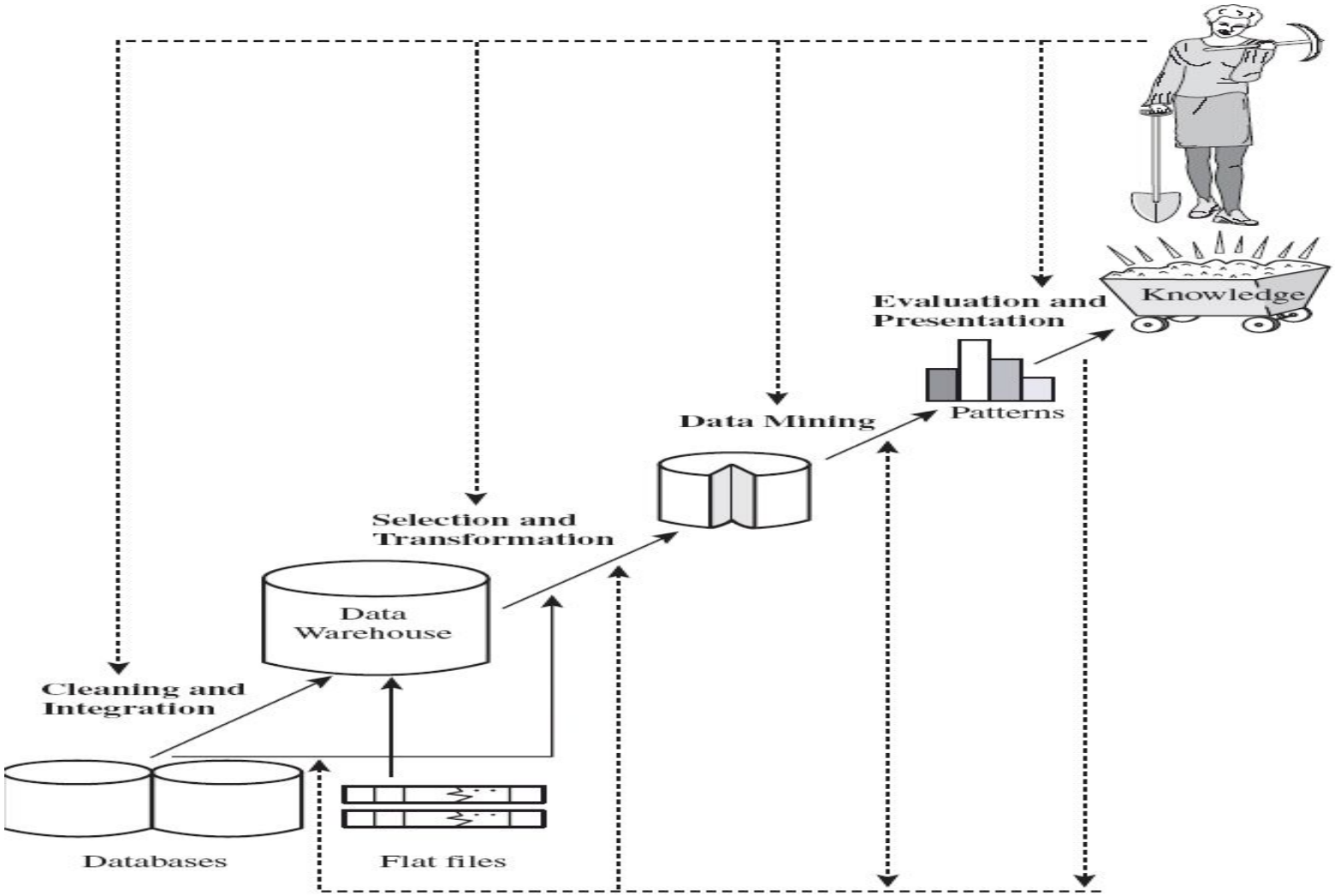
Ex.: Market Analysis and Management

- 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

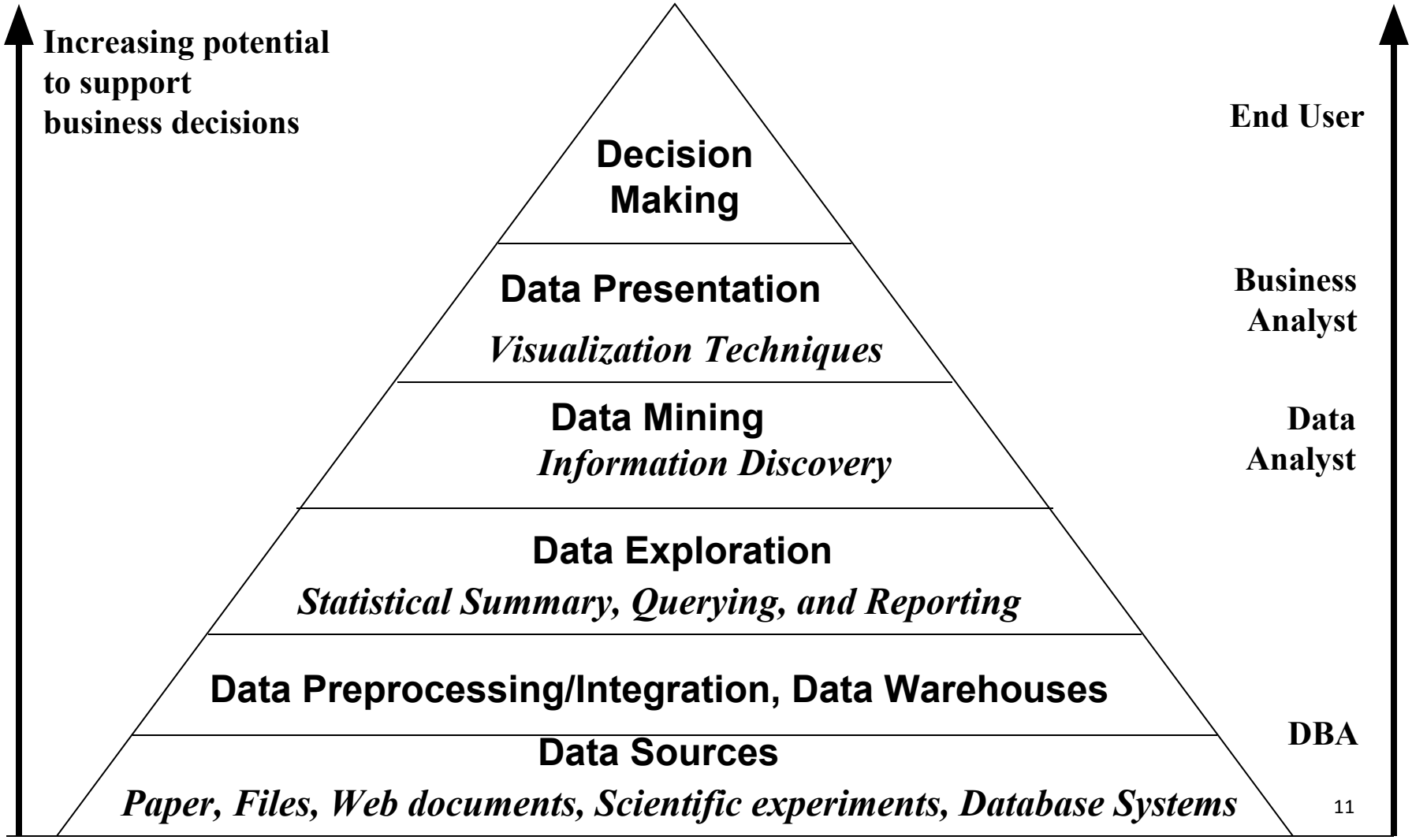
Knowledge Discovery (KDD) Process



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** (data transformed or consolidated into forms appropriate for mining)
    (Done with data preprocessing)
  - **Data mining** (an essential process where intelligent methods are applied to extract data patterns)
  - **Pattern evaluation** (indentify the truly interesting patterns)
  - **Knowledge presentation** (mined knowledge is presented to the user with visualization or representation techniques)
- Use of discovered knowledge

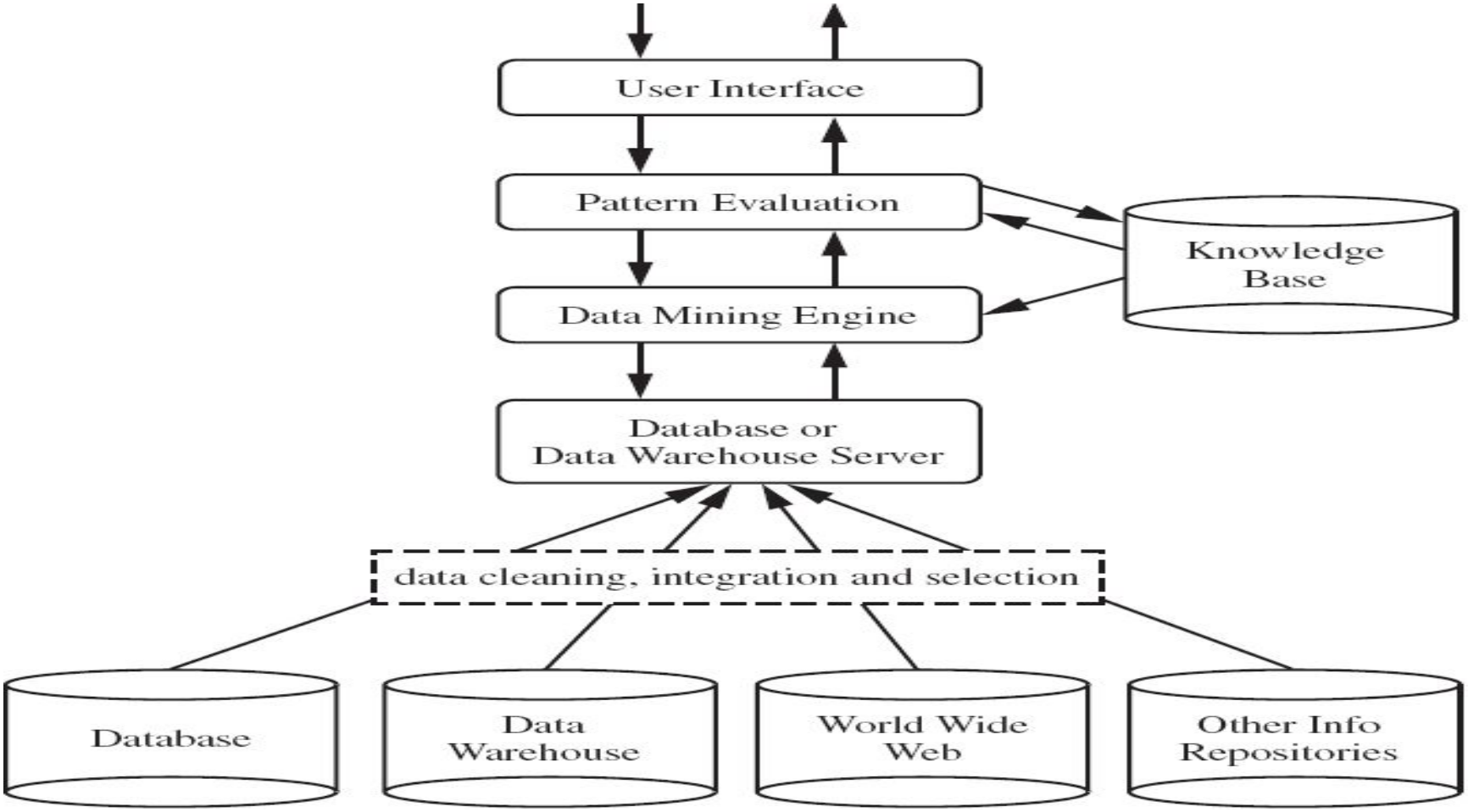
Data Mining and Business Intelligence



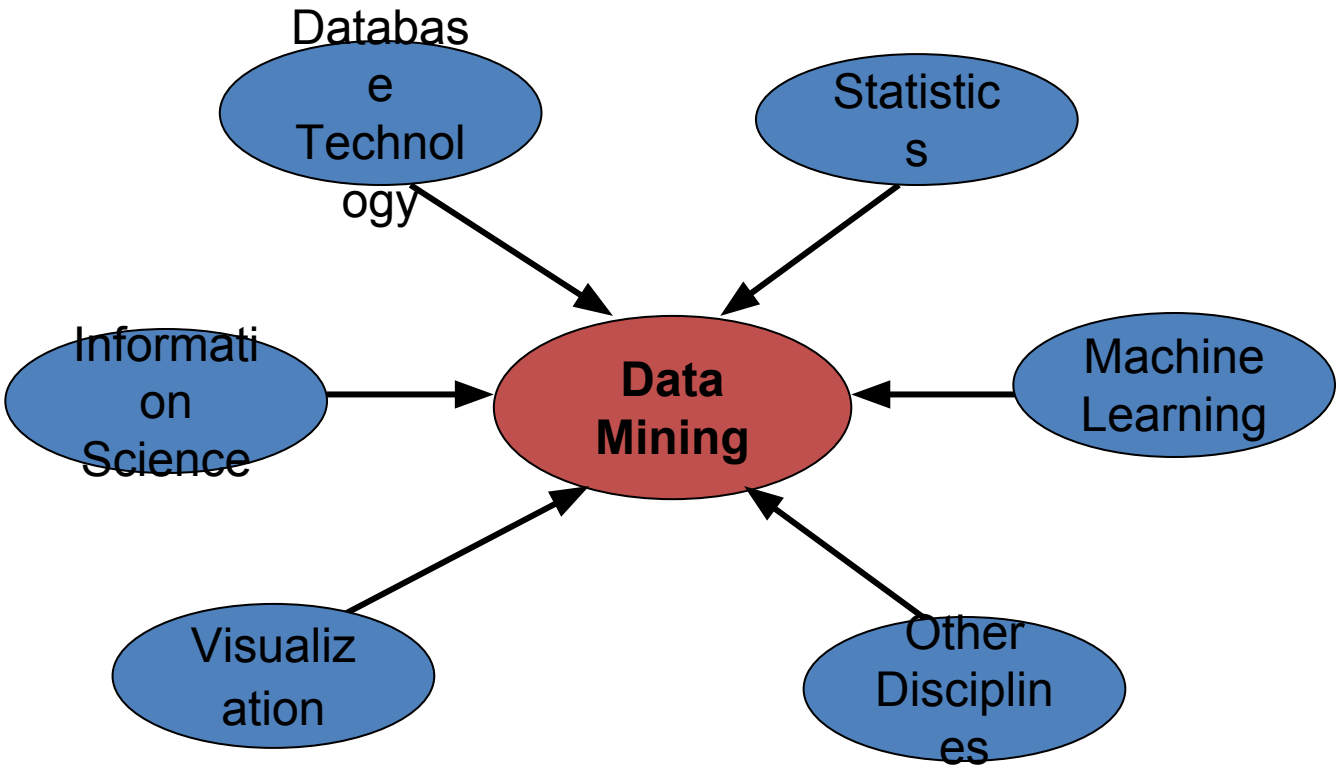
A typical DM System Architecture

- 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…)

1.3 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

Relational Databases

- DBMS – database management system, contains a collection of interrelated databases
  e.g. Faculty database, student database, publications database
- Each database contains a collection of tables and functions to manage and access the data.
  e.g. student_bio, student_graduation, student_parking
- Each table contains columns and rows, with columns as attributes of data and rows as records.
- Tables can be used to represent the relationships between or among multiple tables.

Relational Databases (2) – *AllElectronics* store

customer

cust_ID	*name*	*address*	*age*	*income*	*credit_info*	*category*	...
C1	Smith, Sandy	1223 Lake Ave., Chicago, IL	31	$78000	1	3	...
...	...	...	...	...	...	...	...

item

item_ID	*name*	*brand*	*category*	*type*	*price*	*place_made*	*supplier*	*cost*
I3	hi-res-TV	Toshiba	high resolution	TV	$988.00	Japan	NikoX	$600.00
I8	Laptop	Dell	laptop	computer	$1369.00	USA	Dell	$983.00
...	...	...	...	...	...	...	...	...

employee

empl_ID	*name*	*category*	*group*	*salary*	*commission*
E55	Jones, Jane	home entertainment	manager	$118,000	2%
...	...	...	...	...	...

branch

branch_ID	*name*	*address*
B1	City Square	396 Michigan Ave., Chicago, IL
...	...	...

purchases

trans_ID	*cust_ID*	*empl_ID*	*date*	*time*	*method_paid*	*amount*
T100	C1	E55	03/21/2005	15:45	Visa	$1357.00
...	...	...	...	...	...	...

items_sold

trans_ID	*item_ID*	*qty*
T100	I3	1
T100	I8	2
...	...	...

works_at

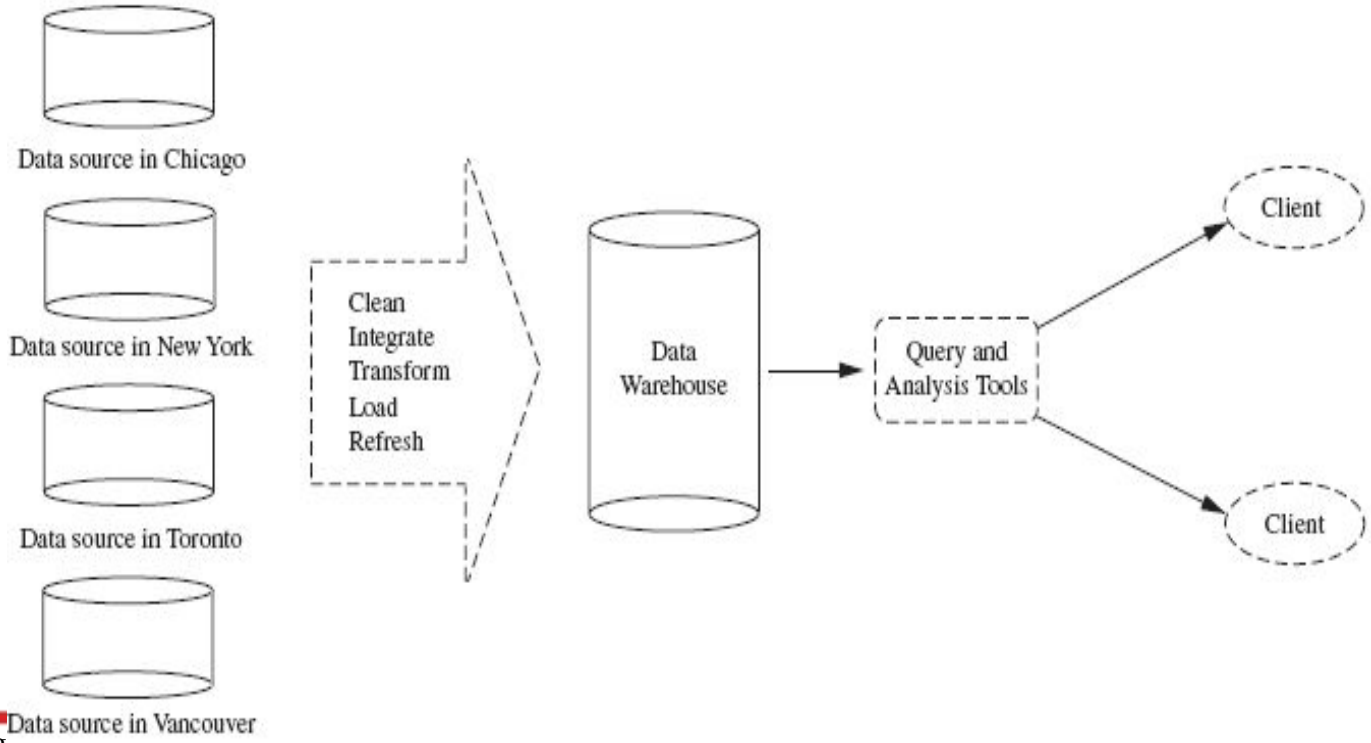
empl_ID	*branch_ID*
E55	B1
...	...

Relational Databases (3)

- With a relational query language, e.g. SQL, we will be able to find answers to questions such as:
  - How many items were sold last year?
  - Who has earned commissions higher than 10%?
  - What is the total sales of last month for Dell laptops?
- When data mining is applied to relational databases, we can search for trends or data patterns.
- Relational databases are one of the most commonly available and rich information repositories, and thus are a major data form in our study.

Data Warehouses

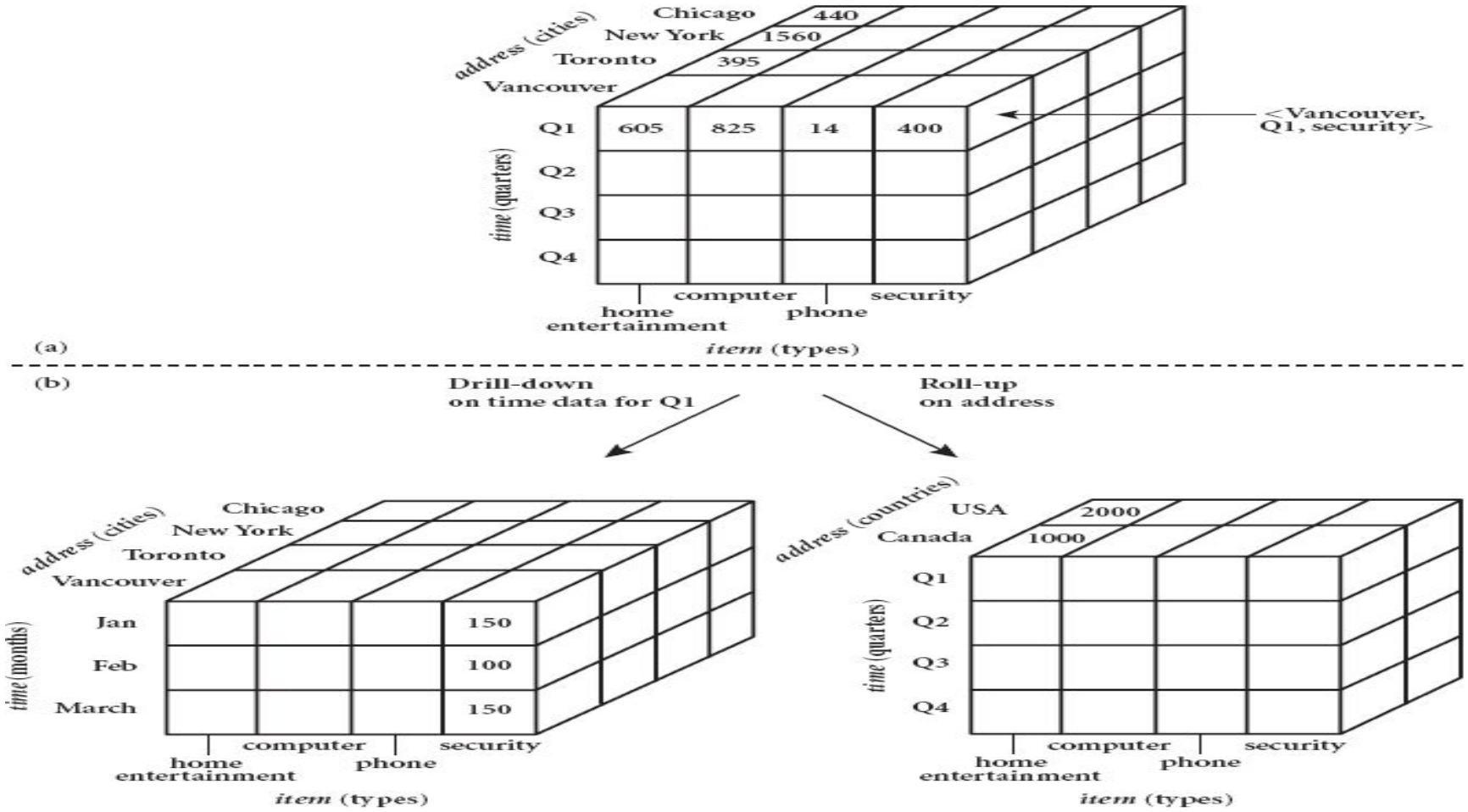
- A repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.
- Constructed via a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing.



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

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

- 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.

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

- 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

1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

– 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

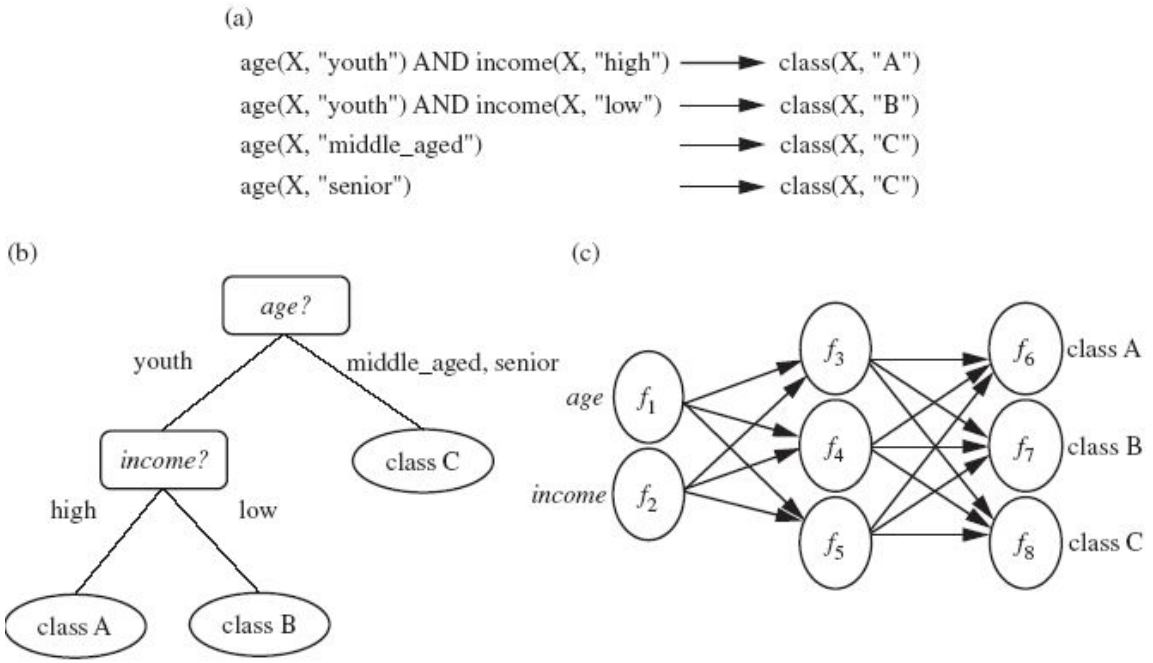
1.4 Data Mining Functionalities

- What kinds of patterns can be mined?

- 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

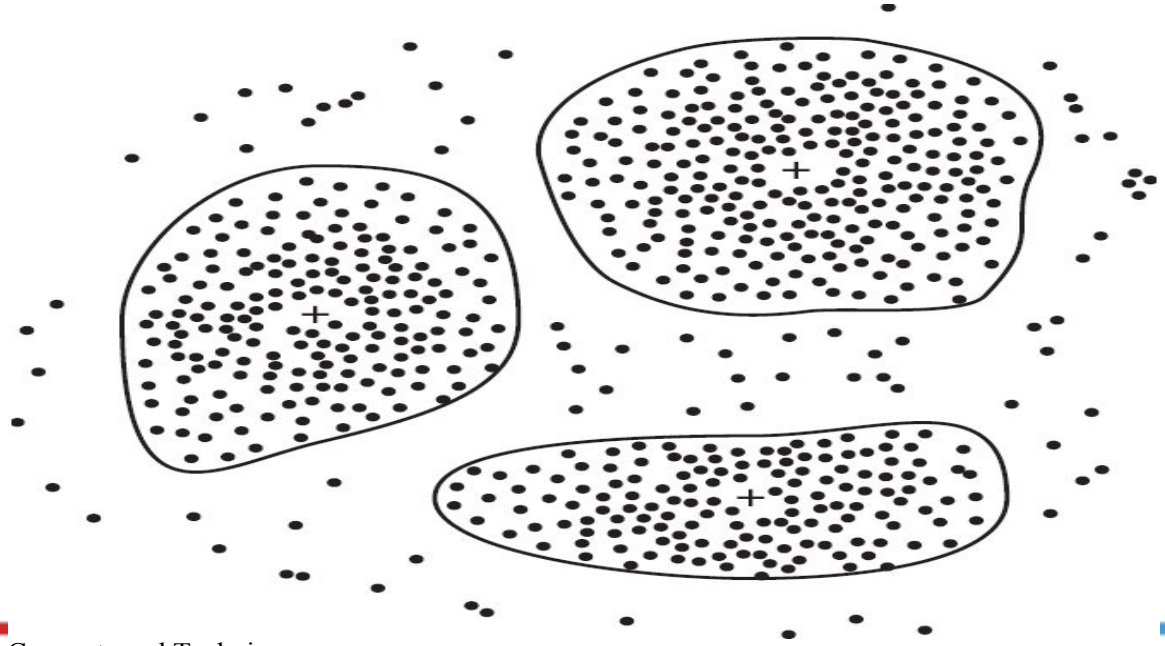
1.4 Data Mining Functionalities

- What kinds of patterns can be mined?



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.

1.5 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* !

1.6 Classification of data mining systems

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

1.9 Major Issues in Data Mining

- Mining methodology and User interaction
  - 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
  - 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

1.9 Major Issues in Data Mining

- 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 bases on user beliefs or expectations

1.9 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
- 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