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, …

Why Data Mining?

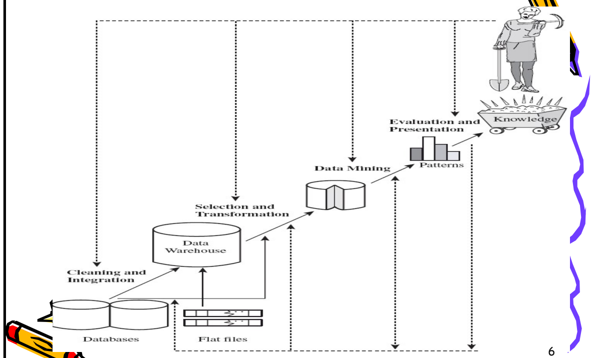
- Data rich but information poor!
  - What does those data mean?
  - How to analyze data?
- Data mining — Automated analysis of massive data sets

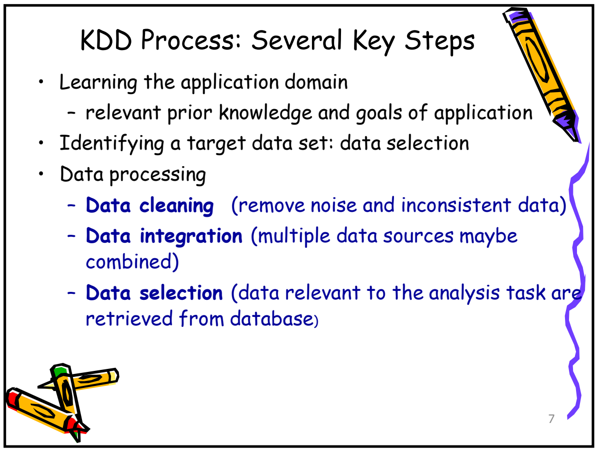
What Is Data Mining?

- 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.

- Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems.
- It primarily turns raw data into useful information.

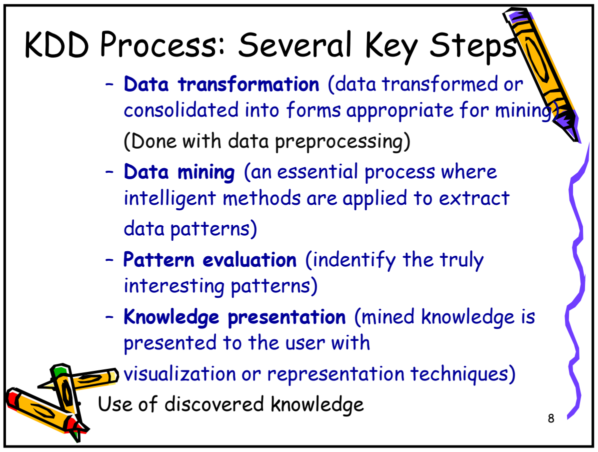
Knowledge Discovery (KDD) Process





KDD Process: Several Key Steps

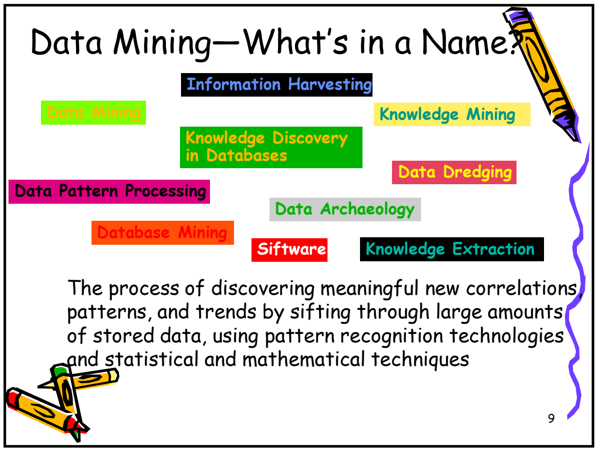
- Learning the application domain
  - relevant prior knowledge and goals of application
- Identifying a target data set: data selection
- Data processing
  - **Data cleaning** (remove noise and inconsistent data)
  - **Data integration** (multiple data sources maybe combined)
  - **Data selection** (data relevant to the analysis task are retrieved from database)



KDD Process: Several Key Steps

- **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)

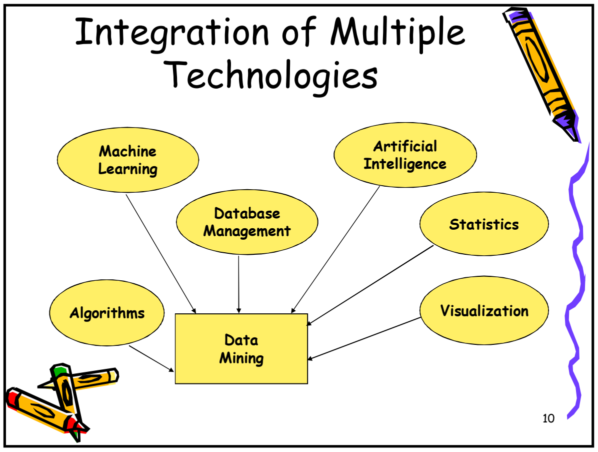
Use of discovered knowledge



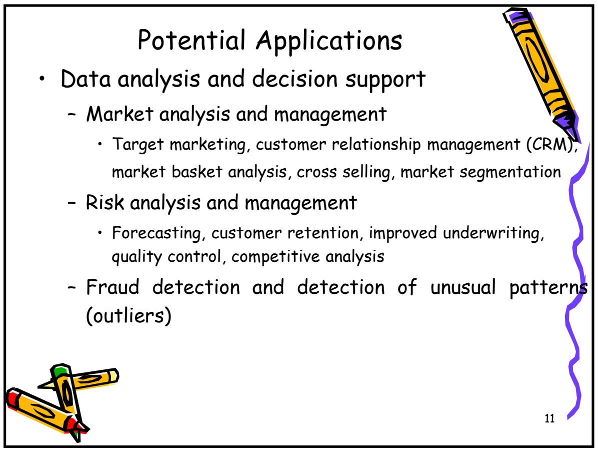
Data Mining—What's in a Name?

Information Harvesting
Data Mining
Knowledge Mining
Knowledge Discovery in Databases
Data Dredging
Data Pattern Processing
Data Archaeology
Database Mining
Siftware
Knowledge Extraction

The process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of stored data, using pattern recognition technologies and statistical and mathematical techniques

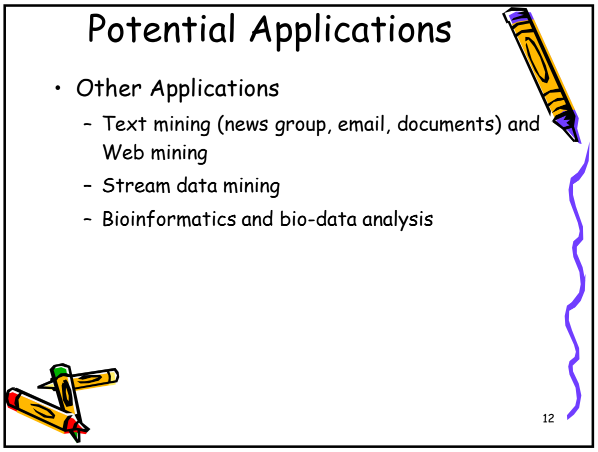


Integration of Multiple Technologies



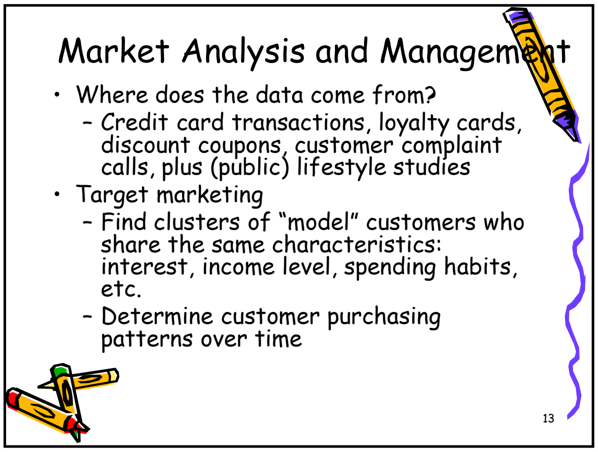
Potential Applications

- 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)



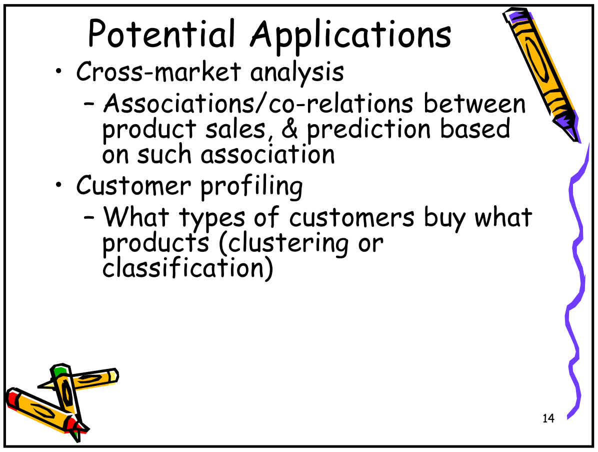
Potential Applications

- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis



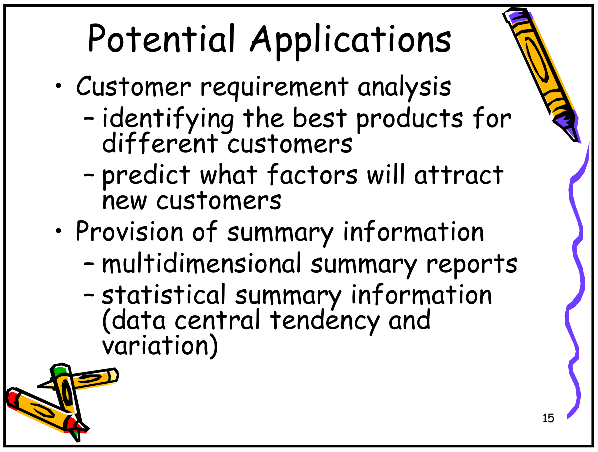
Market Analysis and Management

- Where does the data come from?
  - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine customer purchasing patterns over time



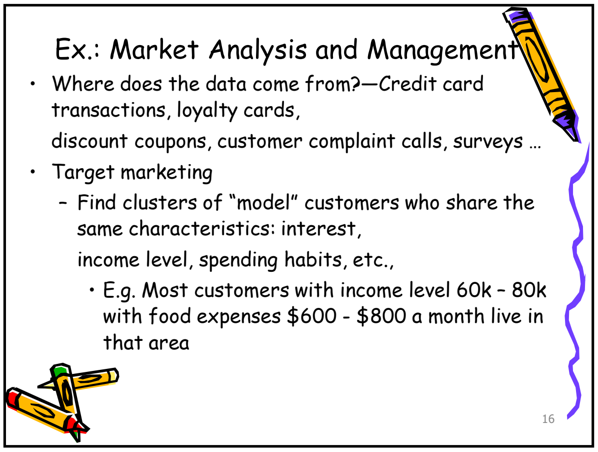
Potential Applications

- Cross-market analysis
  - Associations/co-relations between product sales, & prediction based on such association
- Customer profiling
  - What types of customers buy what products (clustering or classification)



Potential Applications

- Customer requirement analysis
  - identifying the best products for different customers
  - predict what factors will attract new customers
- Provision of summary information
  - multidimensional summary reports
  - statistical summary information (data central tendency and variation)

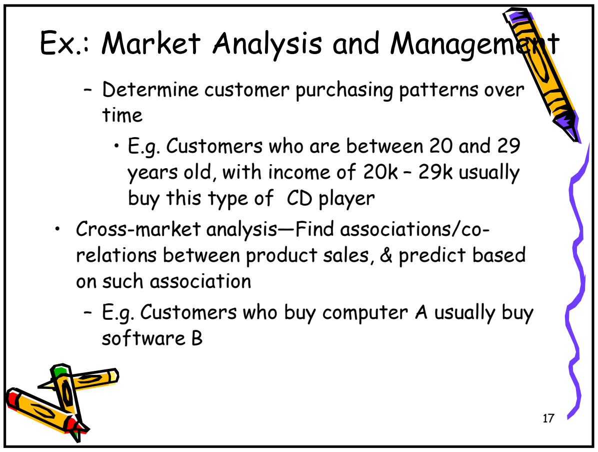


Ex.: Market Analysis and Management

- Where does the data come from?—Credit card transactions, loyalty cards,

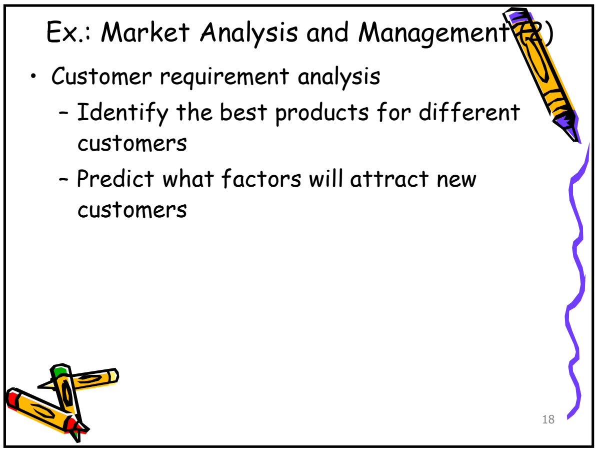
  discount coupons, customer complaint calls, surveys …
- Target marketing
  - Find clusters of "model" customers who share the same characteristics: interest,

    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



Ex.: Market Analysis and Management

  - 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

Market Analysis and Management

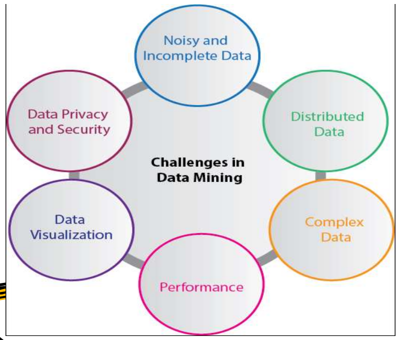
- Provision of summary information
  - Multidimensional summary reports
    - E.g. Summarize all transactions of the first quarter from three different branches.
    - Summarize all transactions of last year from a particular branch.
    - Summarize all transactions of a particular product

  - 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

Challenges of Implementation in Data mining

- Although data mining is very powerful, it faces many challenges during its execution.
- Various challenges could be related to performance, data, methods, and techniques, etc.
- The process of data mining becomes effective when the challenges or problems are correctly recognized and adequately resolved.

Challenges of Implementation in Data mining

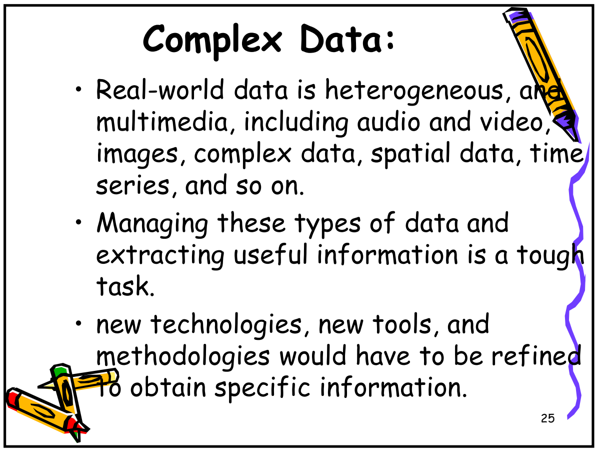

Noisy and Incomplete Data
Data Privacy and Security
Distributed Data
Challenges in Data Mining
Data Visualization
Complex Data
Performance

Incomplete and noisy data:

- The process of extracting useful data from large volumes of data is data mining.
- The data in the real-world is heterogeneous, incomplete, and noisy.
- Data in huge quantities will usually be inaccurate or unreliable.
- These problems may occur due to data measuring instrument or because of human errors.

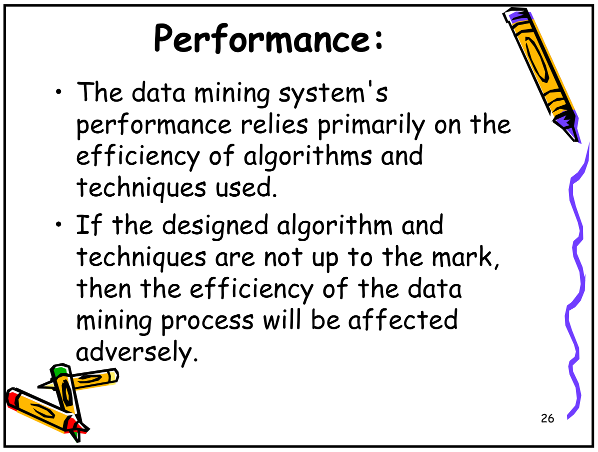
Data Distribution:

- Real-worlds data is usually stored on various platforms in a distributed computing environment.
- It might be in a database, individual systems, or even on the internet..
- Therefore, data mining requires the development of tools and algorithms that allow the mining of distributed data.



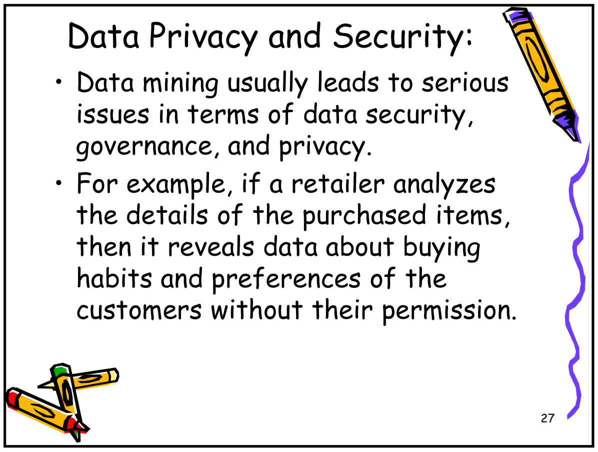
Complex Data:

- Real-world data is heterogeneous, and multimedia, including audio and video, images, complex data, spatial data, time series, and so on.
- Managing these types of data and extracting useful information is a tough task.
- new technologies, new tools, and methodologies would have to be refined to obtain specific information.



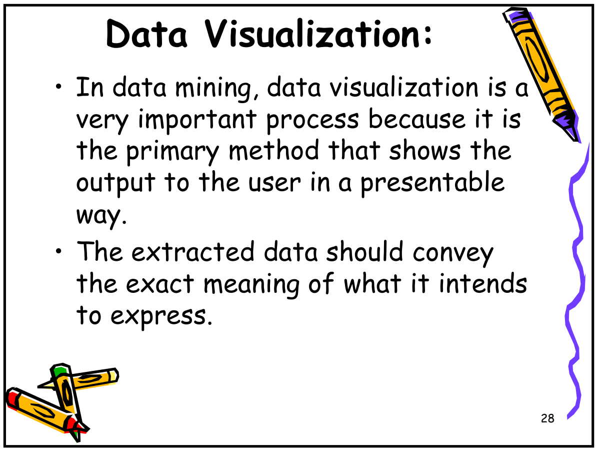
Performance:

- The data mining system's performance relies primarily on the efficiency of algorithms and techniques used.
- If the designed algorithm and techniques are not up to the mark, then the efficiency of the data mining process will be affected adversely.



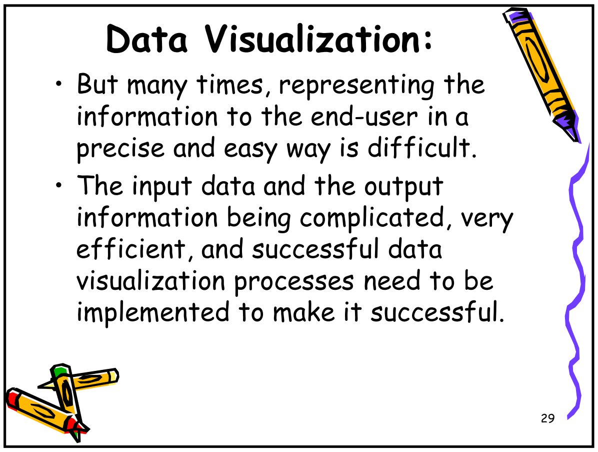
Data Privacy and Security:

- Data mining usually leads to serious issues in terms of data security, governance, and privacy.
- For example, if a retailer analyzes the details of the purchased items, then it reveals data about buying habits and preferences of the customers without their permission.



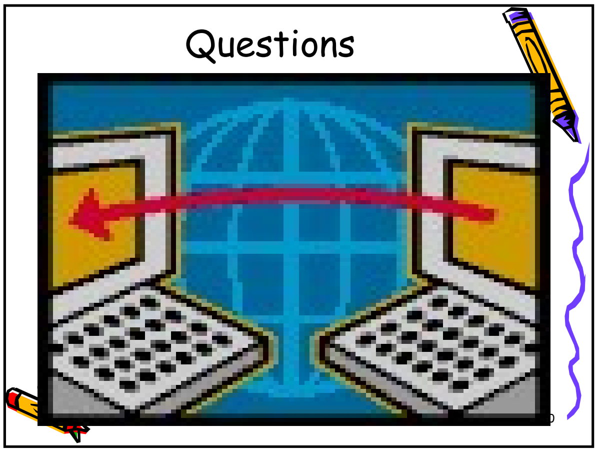
Data Visualization:

- In data mining, data visualization is a very important process because it is the primary method that shows the output to the user in a presentable way.
- The extracted data should convey the exact meaning of what it intends to express.



Data Visualization:

- But many times, representing the information to the end-user in a precise and easy way is difficult.
- The input data and the output information being complicated, very efficient, and successful data visualization processes need to be implemented to make it successful.



Questions