Lecture-1

Data Mining

Introduction
(Basic Concepts)

Department of Computer Science and Engineering
ITER, Siksha 'O' Anusandhan University.

Content

- Motivation
- Evolution of Database Technology
- What is Data Mining?
- Applications of Data Mining
- KDD as a Process.
- Data Mining and Business Intelligence
- Summary

Motivation: "Necessity is the Mother of Invention"

- Data explosion problem
  - Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories

Motivation [Cont..]

- We are drowning in data, but starving for knowledge!
- Solution: Data warehousing and data mining
  - Data warehousing and on-line analytical processing.
  - Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases.

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.) and application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s—2000s:
  - Data mining and data warehousing, multimedia databases, and Web databases

What Is Data Mining??

- Data mining (knowledge discovery in databases):
  - Extracting or "mining" of interesting (non-trivial, implicit, previously unknown and potentially useful) knowledge from large amounts of data.
- Alternative names and their "inside stories":
  - Data mining: a misnomer?
  - Knowledge discovery(mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- What is not data mining?
  - Query processing.
  - Expert systems or small ML/statistical programs

Applications

- **Database analysis and Decision Support**
  - **Market analysis and management**
    - target marketing, customer relation management, market basket analysis, cross selling, market segmentation
  - **Risk analysis and management**
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - **Fraud Detection and Management**
- **Other Applications**
  - **Text mining (news group, email, documents) and Web analysis**.
  - **Intelligent query answering**

Market Analysis and Management

- Where are the data sources for analysis?
  - Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies

➢Target Marketing

- A target market is a group of people considered likely to buy a product or service.
- Find clusters of “model” customers who share the same characteristics like age, location interest, income level, spending habits and lifestyle.
- Determine customer purchasing patterns over time.

Market Analysis and Management [Cont..]

➢Customer Relation Management (CRM)

- Customer relationship management (CRM) is a term that refers to practices, strategies and technologies that companies use to manage and analyze customer interactions and data throughout the customer lifecycle, with the goal of improving business relationships with customers, assisting in customer retention and driving sales growth.

➢Market Basket Analysis

- Market Basket Analysis is a modelling technique based upon the theory that if you buy a certain group of items, you are more (or less) likely to buy another group of items.

Market Analysis and Management [Cont..]

➢Cross Selling

- Cross-selling is the action or practice of selling an additional product or service to an existing customer.
- Example:
  - A Life Insurance company suggesting its customer sign up for car or health insurance.
  - A laptop seller offering a customer a mouse, pen-drive, and/or accessories.

➢Market Segmentation

- It is the process of dividing a broad consumer or business market, normally consisting of existing and potential customers, into sub-groups of consumers (known as *segments*) based on some type of shared characteristics.

Market Analysis and Management [Cont..]

- Cross-market analysis
  - Associations/co-relations between product sales.
  - Prediction based on the association information.
- Customer profiling
  - data mining can tell you what types of customers buy what products (clustering or classification).

Market Analysis and Management [Cont..]

- Identifying customer requirements
  - Identifying the best products for different customers.
  - Use prediction to find what factors will attract new customers.
- Provides summary information
  - Various multidimensional summary reports.
  - Statistical summary information (data central tendency and variation).

Risk Analysis and Management

- Finance planning and asset evaluation
  - cash flow analysis and prediction
  - contingent claim analysis to evaluate assets
  - cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)

- Resource planning:
  - summarize and compare the resources and spending

Risk Analysis and Management [Cont..]

➢Forecasting

- Forecasting is the process of making predictions of the future based on past and present data and most commonly by analysis of trends.

➢Customer Retention

- Customer retention refers to the activities and actions companies and organizations take to reduce the number of customer defections.
- The goal of customer retention programs is to help companies retain as many customers as possible, often through customer loyalty and brand loyalty initiatives.

Risk Analysis and Management [Cont..]

➢Improved Underwriting

- Underwriting services are provided by some large specialist financial institutions, such as banks, insurance or investment houses, whereby they guarantee payment in case of damage or financial loss and accept the financial risk for liability arising from such guarantee.

➢Quality Control

- An aspect of the quality assurance process that consists of activities employed in detection and measurement of the variability in the characteristics of output attributable to the production system, and includes corrective responses.

Risk Analysis and Management [Cont..]

➢Competitive Analysis

- Identifying your competitors and evaluating their strategies to determine their strengths and weaknesses relative to those of your own product or service .

  ✓Monitor competitors and market directions
  ✓Group customers into classes and a class-based pricing procedure
  ✓Set pricing strategy in a highly competitive market

- With this evaluation, we can establish what makes our product or service unique

Risk Analysis and Management [Cont..]

- Attributes in order to attract your target market are:-
  - Who are your competitors?
  - What products or services do they sell?
  - What is each competitor's market share?
  - What are their past strategies?
  - What are their current strategies?
  - What type of media are used to market their products or services?
  - How many hours per week do they purchase to advertise through the media used in this market?
  - What are each competitor's strengths and weaknesses?
  - What potential threats do your competitors pose?
  - What potential opportunities do they make available for you?

Fraud Detection and Management

- Applications
  - Widely used in health care, retail, credit card services, telecommunications (phone card fraud), etc.
- Approach
  - Use historical data to build models of fraudulent behavior and use data mining to help identify similar instances
- Examples
  - **Auto insurance**: detect a group of people who stage accidents to collect on insurance
  - **Money laundering**: detect suspicious money transactions (US Treasury's Financial Crimes Enforcement Network)
  - **Medical insurance**: detect professional patients and ring of doctors and ring of references

Fraud Detection and Management [Cont..]

- Detecting inappropriate medical treatment
  - Australian Health Insurance Commission identifies that in many cases blanket screening tests were requested.
- Detecting telephone fraud
  - Telephone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm.
  - British Telecom identified discrete groups of callers with frequent intra-group calls, especially mobile phones, and broke a multimillion dollar fraud.
- Retail
  - Analysts estimate that 38% of retail shrink is due to dishonest employees.

Other Applications

➢Text mining

- Text mining, also referred to as text data mining, roughly equivalent to text analytics.
- It is the process of deriving high-quality information from text.
- High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning.

➢Sports

- IBM Advanced Scout analyzed NBA game statistics (shots blocked, assists, and fouls) to gain competitive advantage for New York Knicks and Miami Heat

Other Applications

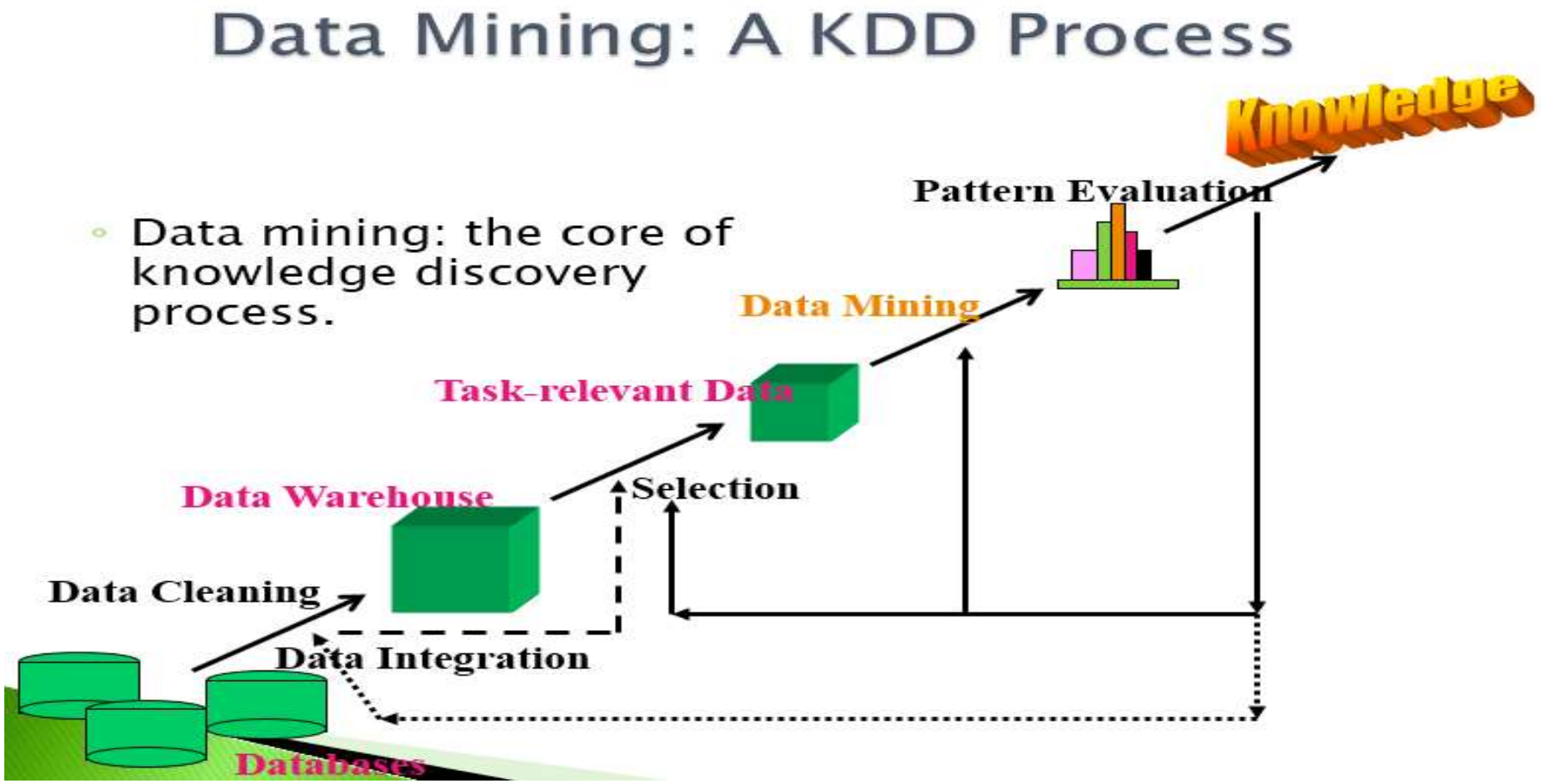
➢Astronomy

- JPL and the Palomar Observatory discovered 22 quasars with the help of data mining

➢Internet Web Surf-Aid

- IBM Surf-Aid applies data mining algorithms to Web access logs for market-related pages to discover customer preference and behavior pages, analyzing effectiveness of Web marketing, improving Web site organization, etc.

KDD Process


Data Mining: A KDD Process
Data mining: the core of knowledge discovery process.
Knowledge
Pattern Evaluation
Data Mining
Task-relevant Data
Data Warehouse
Selection
Data Cleaning
Data Integration
Databases

Steps of a KDD Process

- Data cleaning

  To remove noise and inconsistent

- Data integration

  Where multiple data sources may be combined

- Data selection

  Where data relevant to the analysis task are retrieved from the database

- Data transformation

  Where data are transformed or consolidated into forms appropriate

  for mining by performing summary or aggregation operations, for instance

Steps of a KDD Process

- Data mining

  An essential process where intelligent methods are applied in order to extract data patterns

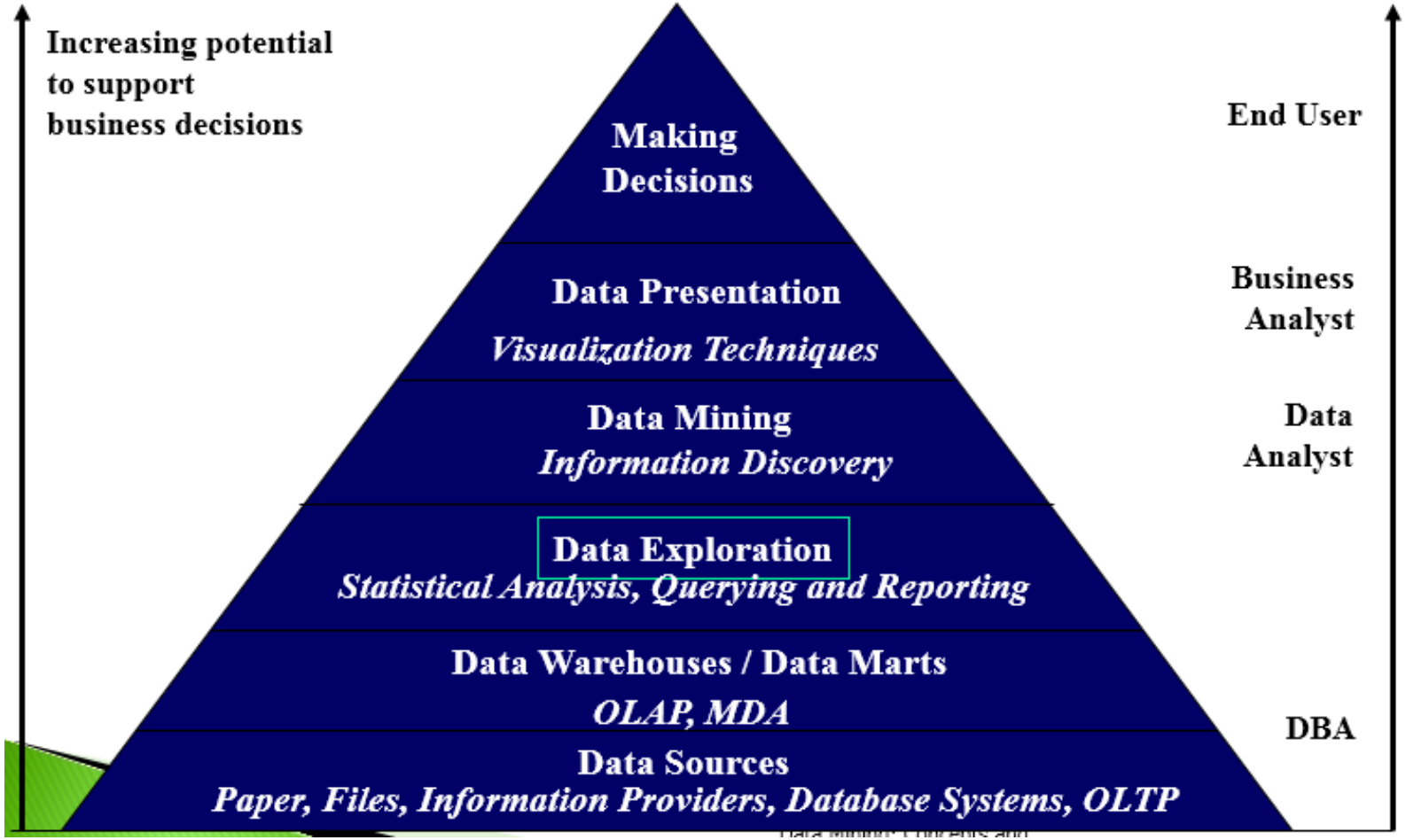
- Pattern evaluation

  To identify the truly interesting patterns representing knowledge based on some interestingness measures

- Knowledge presentation

  Where visualization and knowledge representation techniques are used to present the mined knowledge to the user

Data Mining and Business Intelligence



Summary

- Data mining is the task of discovering interesting patterns from large amounts of data, where the data can be stored in databases, datawarehouses, or other information repositories.
- It is a young interdisciplinary field, drawing from areas such as database systems, data warehousing, statistics, machine learning, data visualization, information retrieval, and high-performance computing.
- A knowledge discovery process includes data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and knowledge presentation.

Thank You