Data Mining and Data Warehousing

— Introduction —

Source: Slides from Data Mining: Concepts and Techniques

-Jiawei Han and Micheleline Kamber

Chapter 1. Introduction

- Motivation: Why data mining?

- What is data mining?

- Data Mining: On what kind of data?

- Data mining functionality

- Classification of data mining systems

- Top-10 most popular data mining algorithms

- Major issues in data mining

- Overview of the course

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
- <u>We are drowning in data, but starving for knowledge!</u>
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

Information and knowledge gained can be used for Applications:

- Market Analysis
- Fraud detection
- Customer retention
- Production control
- Science Exploration
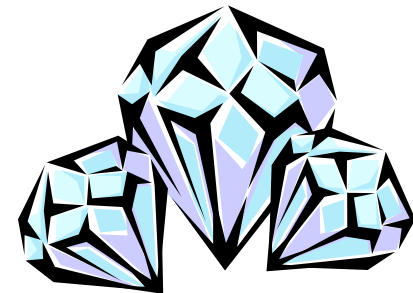
Evolution of Database Technology

- 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
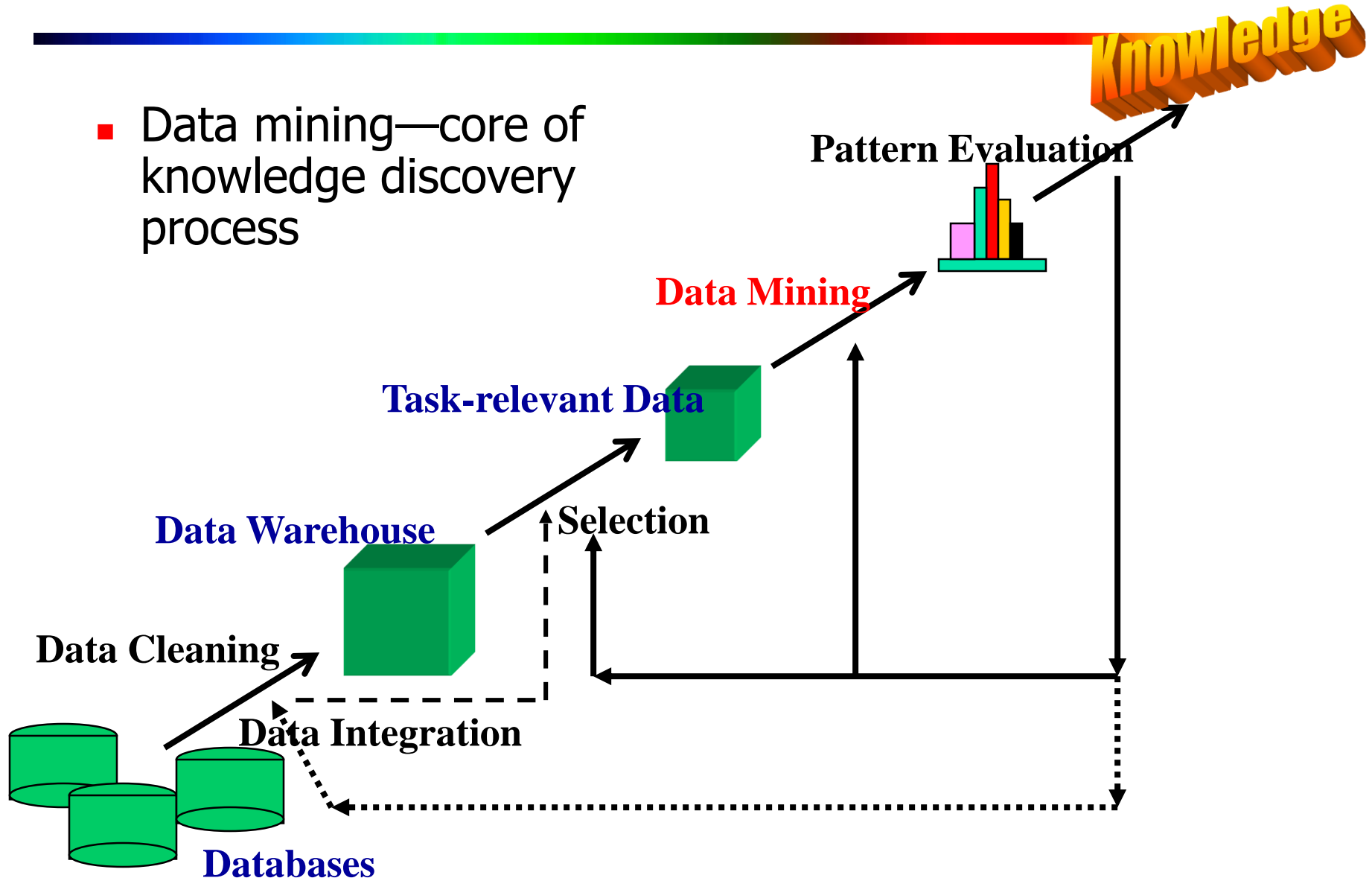
What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (<u>non-trivial</u>, <u>implicit</u>, <u>previously unknown</u> and <u>potentially useful)</u> 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.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

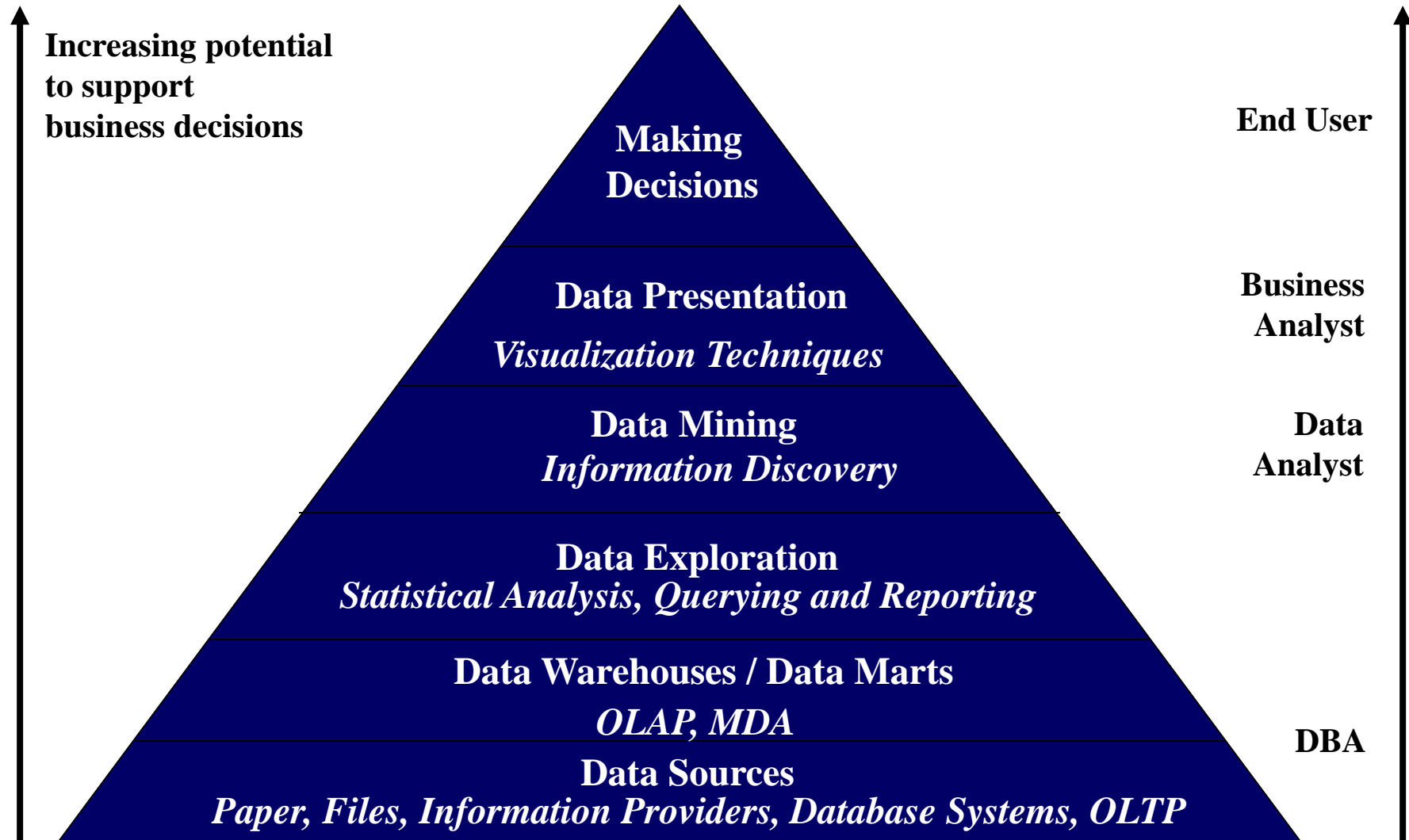# Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process

Knowledge

Pattern Evaluation

Data Mining

Task-relevant Data

Data Warehouse

Selection

Data Cleaning

Data Integration

Databases

Data Mining and Business Intelligence

Increasing potential to support business decisions

End User

Making Decisions

Business Analyst

Data Presentation
Visualization Techniques

Data Analyst

Data Mining
Information Discovery

Data Exploration
Statistical Analysis, Querying and Reporting

Data Warehouses / Data Marts
OLAP, MDA

DBA

Data Sources
Paper, Files, Information Providers, Database Systems, OLTP

Why Not Traditional Data Analysis?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

Ex. 1: 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
- Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association
- Customer profiling—What types of customers buy what products (clustering or classification)
- Customer requirement analysis
  - Identify the best products for different groups of customers
  - Predict what factors will attract new customers
- Provision of summary information
  - Multidimensional summary reports
  - Statistical summary information (data central tendency and variation)

Ex. 2: Corporate Analysis & Risk 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

- Competition

  - monitor competitors and market directions

  - group customers into classes and a class-based pricing procedure

  - set pricing strategy in a highly competitive market

Ex. 3: Fraud Detection & Mining Unusual Patterns

- Approaches: Clustering & model construction for frauds, outlier analysis
- Applications: Health care, retail, credit card service, telecomm.
  - Auto insurance: ring of collisions
  - Money laundering: suspicious monetary transactions
  - Medical insurance
    - Professional patients, ring of doctors, and ring of references
    - Unnecessary or correlated screening tests
  - Telecommunications: phone-call fraud
    - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm
  - Retail industry
    - Analysts estimate that 38% of retail shrink is due to dishonest employees
  - Anti-terrorism