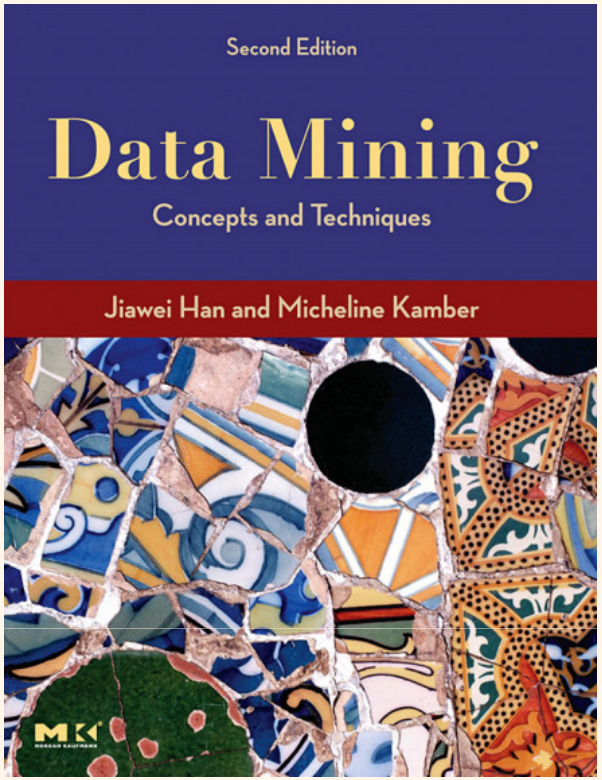


Introduction to Data Mining

Jiawei Han and Micheline Kamber

Department of Computer Science

University of Illinois at Urbana-Champaign

Agenda

- Motivation: Why data mining?
- What is data mining?
- Why mine data?
- Data mining functionality
- Are all patterns interesting?
- Classification of data mining systems
- Data mining tasks
- Integration of data mining system with a DB & DW System
- Major issues in data mining

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
- 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

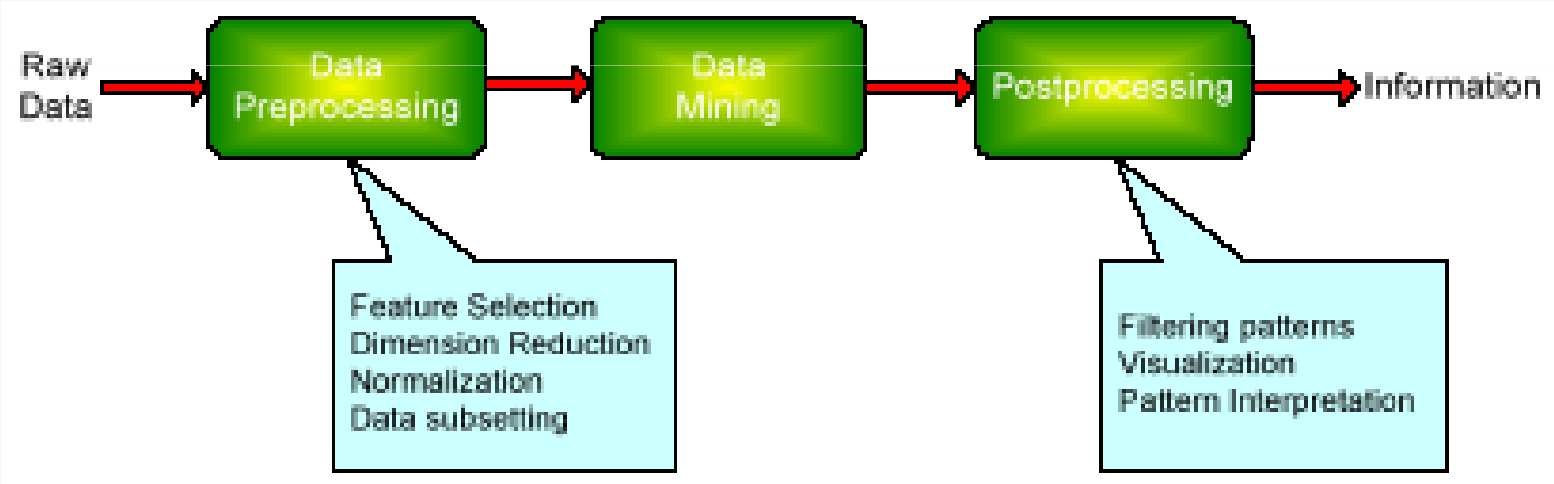
- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s:
  - Stream data management and mining
  - Data mining with a variety of applications
  - Web technology and global information systems

Agenda

- Motivation: Why data mining?
- What is data mining?
- Why mine data?
- Data mining functionality
- Are all patterns interesting?
- Classification of data mining systems
- Data mining tasks
- Integration of data mining system with a DB & DW System
- Major issues in data mining

What is Data Mining ?

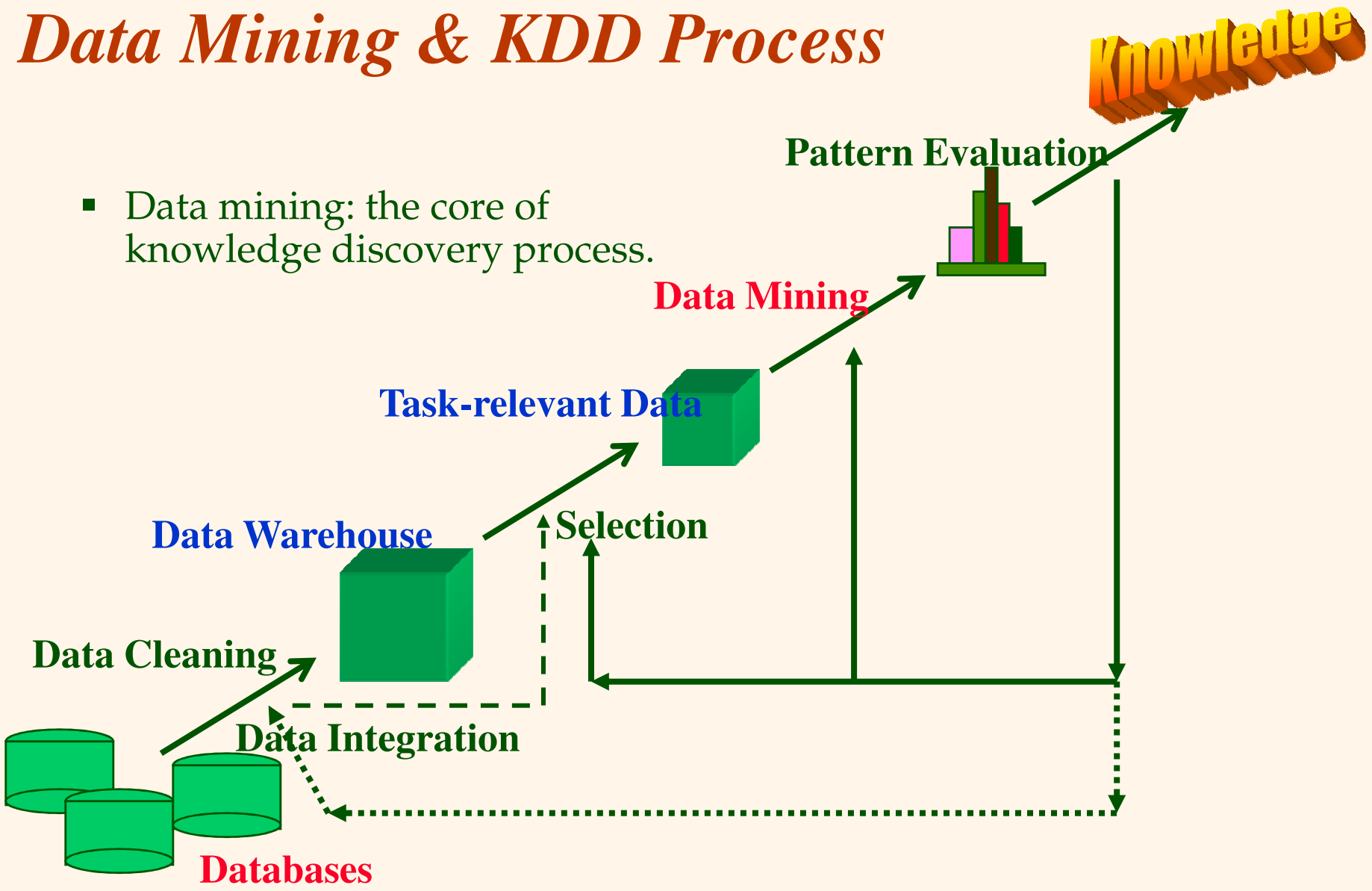
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Data mining is an integral part of Knowledge Discovery in Databases (KDD)



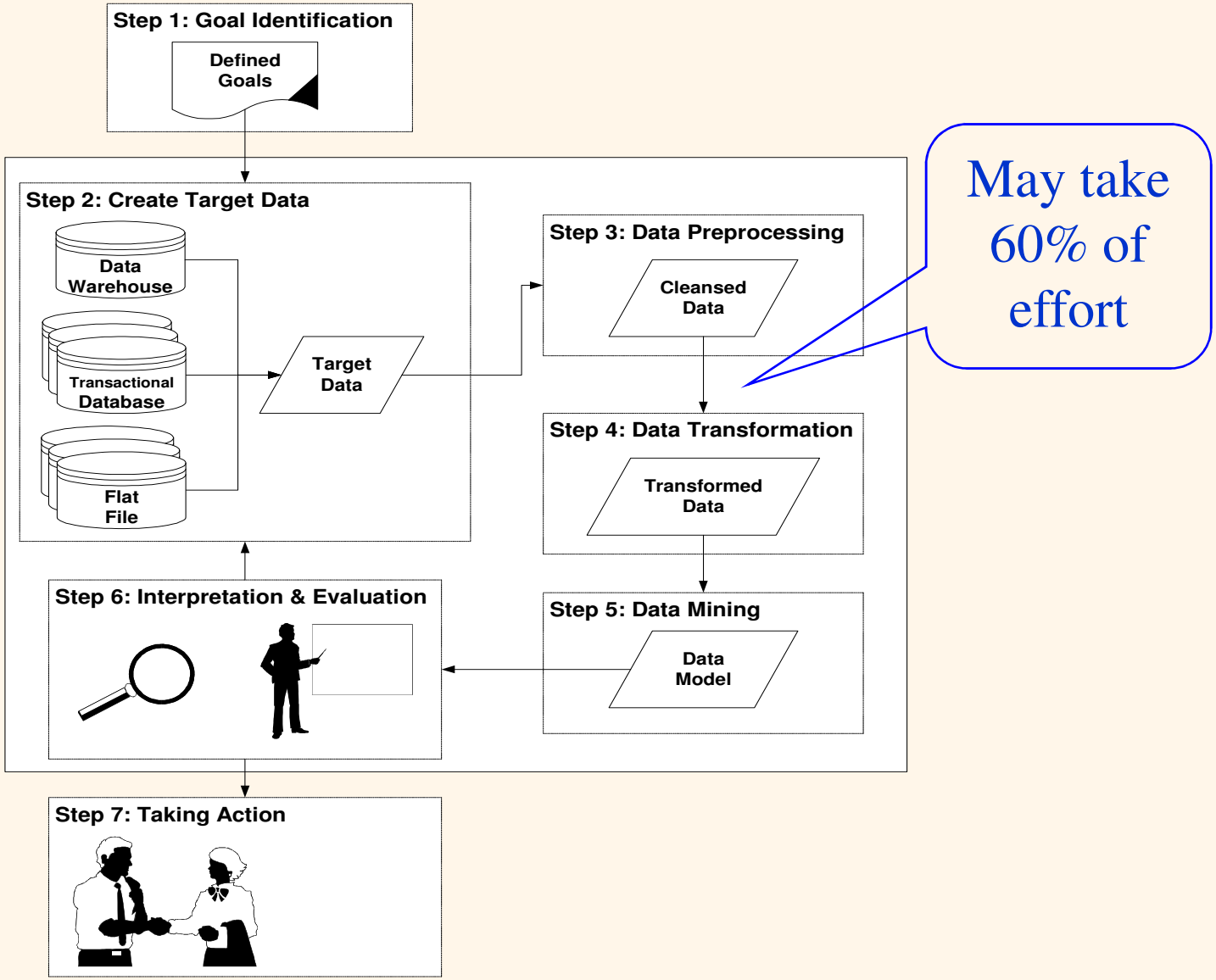
Knowledge Discovery in Databases (KDD)

Data Mining & KDD Process

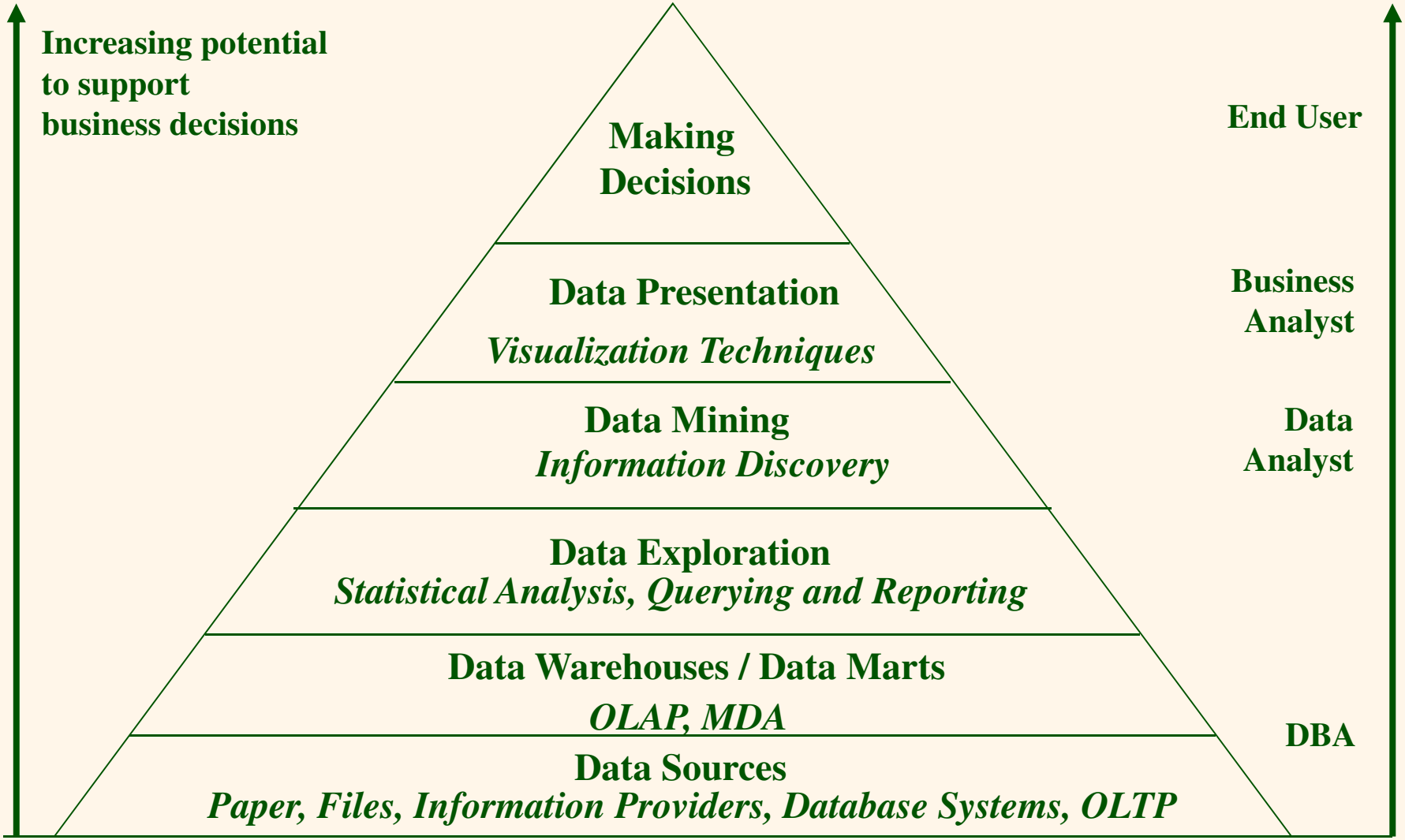
- Data mining: the core of knowledge discovery process.



KDD Process


Step 1: Goal Identification
Defined Goals
Step 2: Create Target Data
Data Warehouse
Transactional Database
Flat File
Target Data
Step 3: Data Preprocessing
Cleansed Data
Step 4: Data Transformation
Transformed Data
Step 5: Data Mining
Data Model
Step 6: Interpretation & Evaluation
Step 7: Taking Action
May take 60% of effort

Data Mining & Business Intelligence

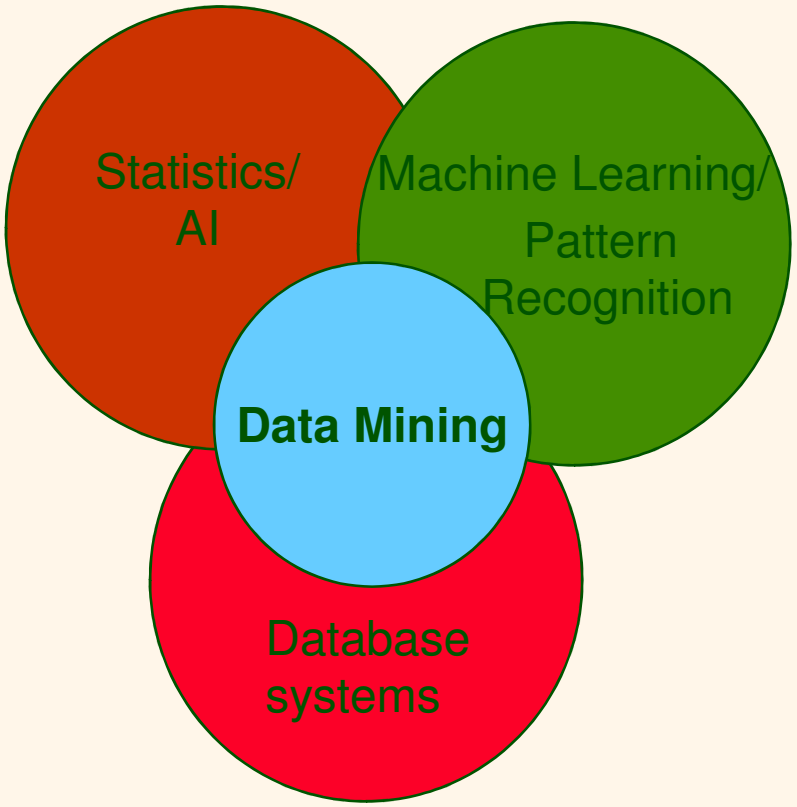


Data Mining: On What Kinds of Data

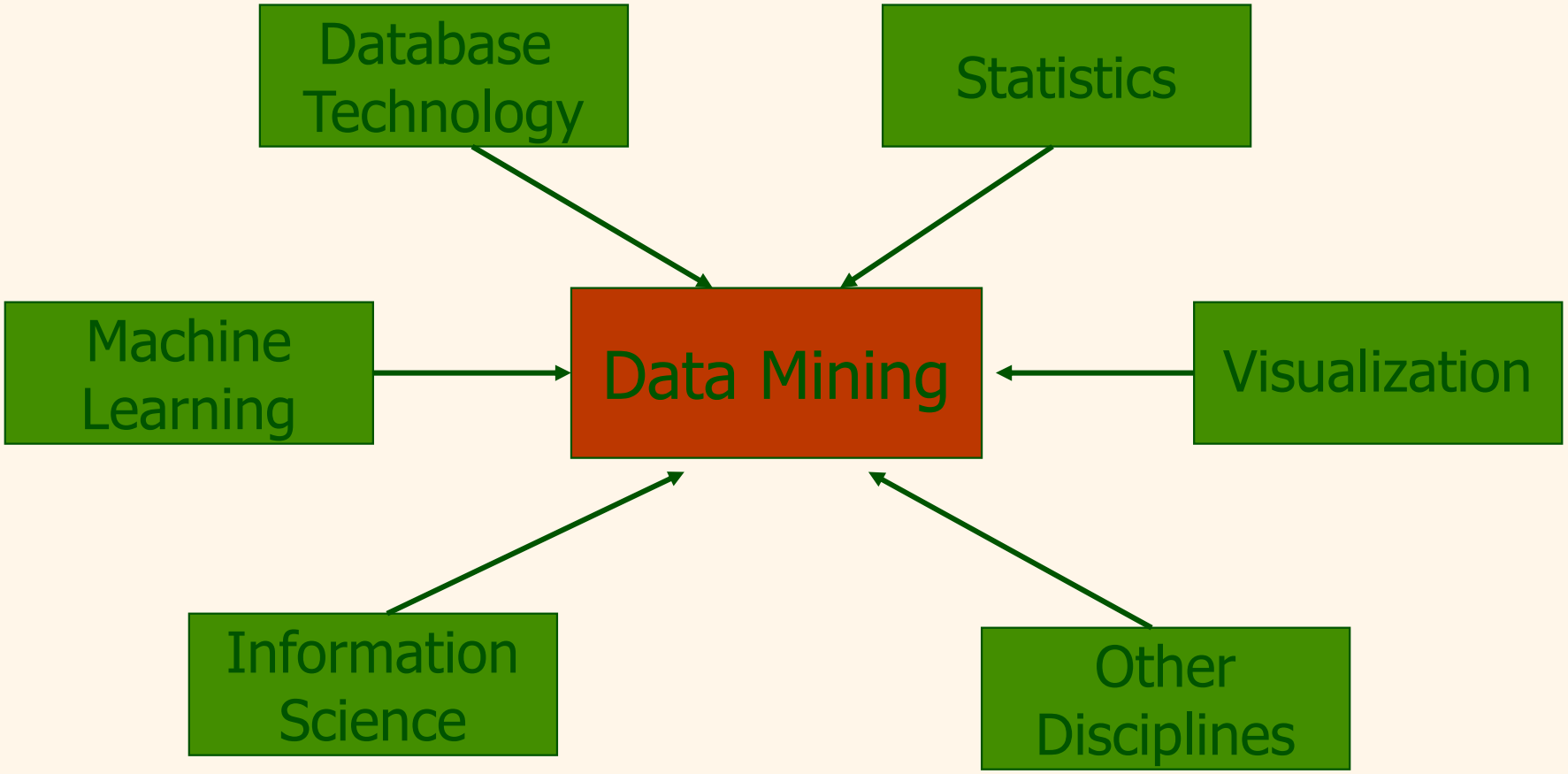
- Traditional database and appllications
  - Relational database, data warehouse, transactional database
- Advanced database and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data (incl. biosequences)
  - Structure data, graphs, social networks and link databases
  - Object-relational databases
  - Heterogeneous databases and legacy databases
  - Spatial data and spatiotemporal data
  - Multimedia database
  - Text databases
  - The World-Wide Web

Origins of Data Mining

- **Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems**
- **Traditional Techniques may be unsuitable due to**
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



Origins of Data Mining: Confluence of Multiple Disciplines



Agenda

- Motivation: Why data mining?
- What is data mining?
- Why mine data?
- Data mining functionality
- Are all patterns interesting?
- Classification of data mining systems
- Data mining tasks primitives
- Integration of data mining system with a DB & DW System
- Major issues in data mining

Data Explosion

"We are drowning in data, but starving for knowledge"

"The amount of data stored in various media has doubled in three years, from 1999 to 2002. The amount of data put into storage in 2002, five exabytes (one quintillion bytes), was equal to the contents of a half a million new libraries, each containing a digitised version of the print collection of the entire US Library of Congress" (Lyman and Varian, UC Berkeley, 2003)

Scale of Data

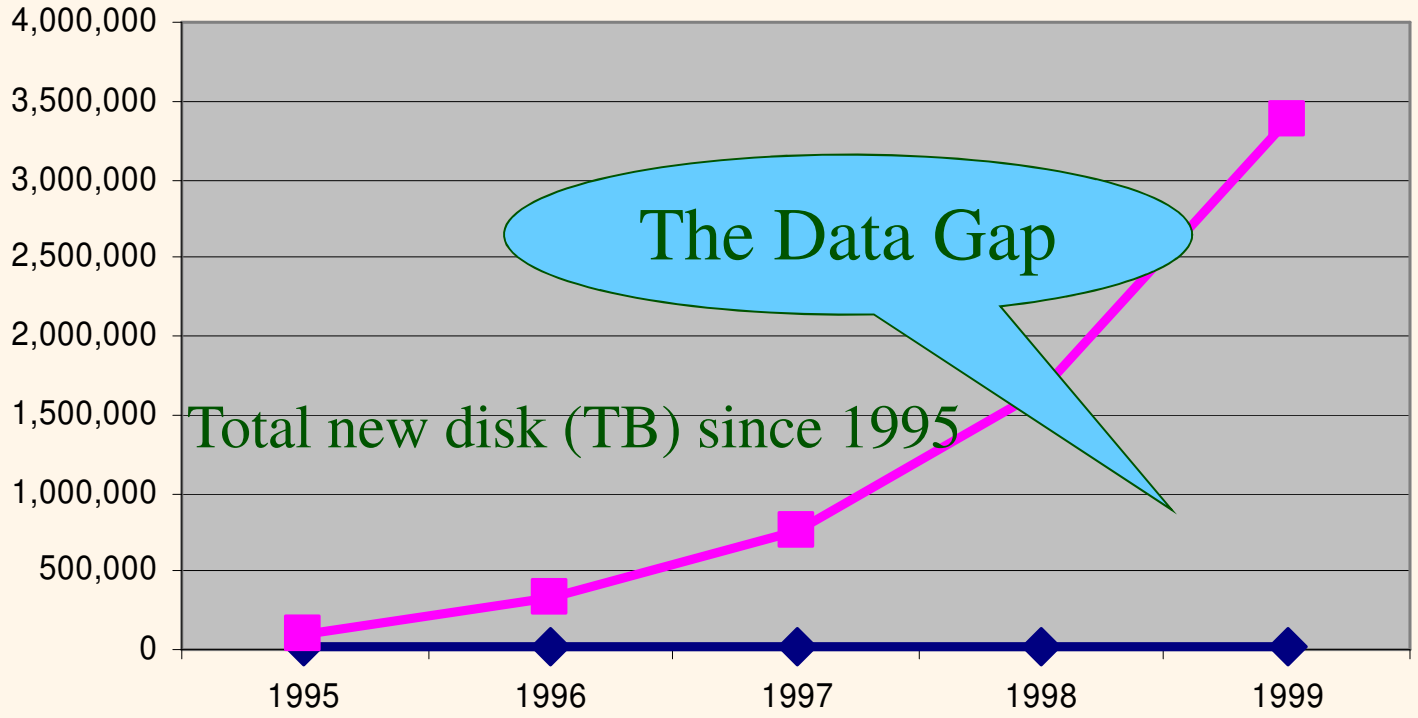
Organization	Scale of Data
Walmart	~ 20 million transactions/day
Google	~ 8.2 billion Web pages
Yahoo	~ 10 GB Web data/hr
NASA satellites	~ 1.2 TB/day
NCBI GenBank	~ 22 million genetic sequences
France Telecom	29.2 TB
UK Land Registry	18.3 TB
AT&T Corp	26.2 TB

"The great strength of computers is that they can reliably manipulate vast amounts of data very quickly. Their great weakness is that they don't have a clue as to what any of that data actually means"

(S. Cass, IEEE Spectrum, Jan 2004)

Why Mine Data ? - Motivation

- There is often information "*hidden*" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

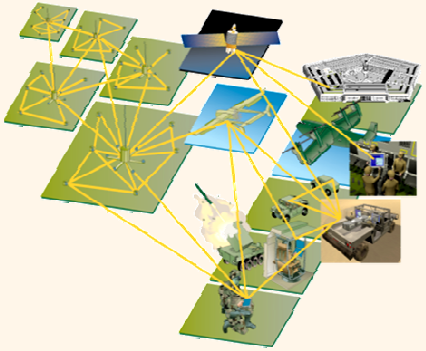
Why Mine Data? - Commercial Viewpoint

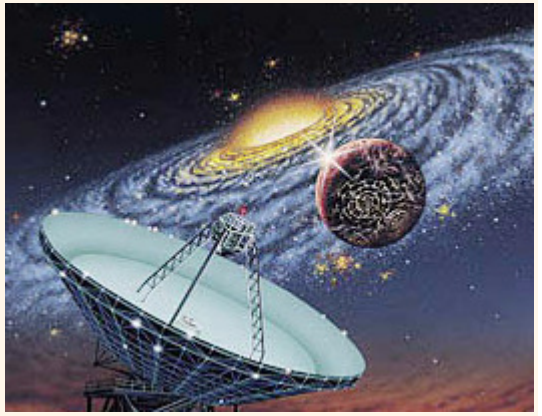
- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions

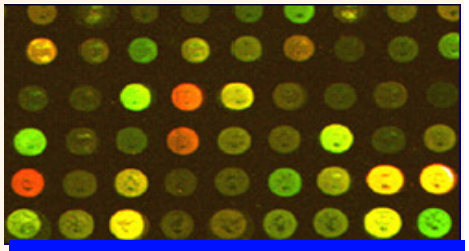
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)

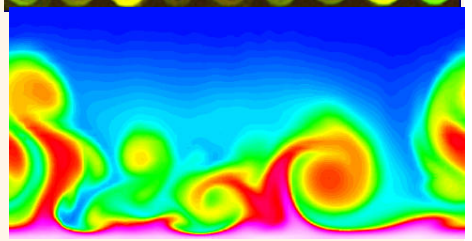
Why Mine Data? - Scientific Viewpoint

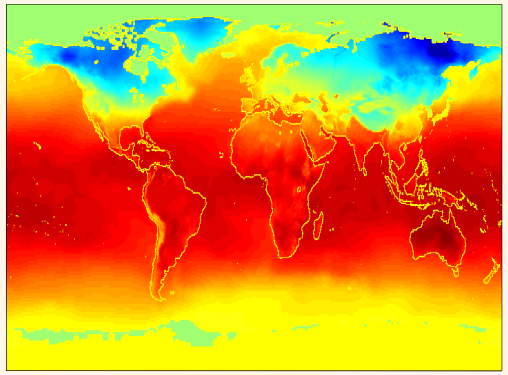
- Data collected and stored at enormous speeds (GB/hour)
  - Remote sensors on a satellite
  - Telescopes scanning the skies
  - Micro-arrays generating gene expression data
  - Scientific simulations generating terabytes of data
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - In classifying and segmenting data
  - In Hypothesis Formation











Agenda

- Motivation: Why data mining?
- What is data mining?
- Why mine data?
- Data mining functionality & applications
- Are all patterns interesting?
- Classification of data mining systems
- Data mining tasks
- Integration of data mining system with a DB & DW System
- Major issues in data mining

Data Mining Functionalities (1)

- Multidimensional concept description: Characterization and discrimination
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions
- Frequent patterns, association, correlation and causality
  - Diaper → Beer [0.5%, 75%] (Correlation or causality?)
- Classification and prediction
  - Construct models (functions) that describe and distinguish classes or concepts for future prediction
    - E.g., classify countries based on climate, or classify cars based on gas mileage
  - Predict some unknown or missing numerical values

Data Mining Functionalities (2)

- Cluster analysis
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Maximizing intra-class similarity & minimizing interclass similarity
- Outlier analysis
  - Outlier: Data object that does not comply with the general behavior of the data
  - Noise or exception? No! useful in fraud detection, rare events analysis
- Trend and evolution analysis
  - Trend and deviation: e.g., regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- Other pattern-directed or statistical analyses

Data Mining Applications (1)

Application	Input	Output
Business Intelligence	Customer purchase history, credit card information	What products are frequently bought together by customers
Collaborative Filtering	User-provided ratings for movies, or other products	Recommended movies or other products
Network Intrusion Detection	TCPdump trace or Cisco NetFlow logs	Anomaly score assigned to each network connection
Web search	Query provided by user	Documents ranked based on their relevance to user input
Medical Diagnosis	Patient history, physiological, and demographic data	Diagnosis of patient as sick or healthy
Climate Research	Measurements from sensors aboard NASA Earth observing satellites	Relationships among Earth Science events, trends in time series, etc
Process Mining	Event-based data from workflow logs	Discrepancies between prescribed models and actual process executions

Data Mining Applications (2)

- **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

Major Issues in Data Mining

- Mining methodology
  - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel, distributed and incremental mining methods
  - Integration of the discovered knowledge with existing one: knowledge fusion
- User interaction
  - Data mining query languages and ad-hoc mining
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction
- Applications and social impacts
  - Domain-specific data mining & invisible data mining
  - Protection of data security, integrity, and privacy