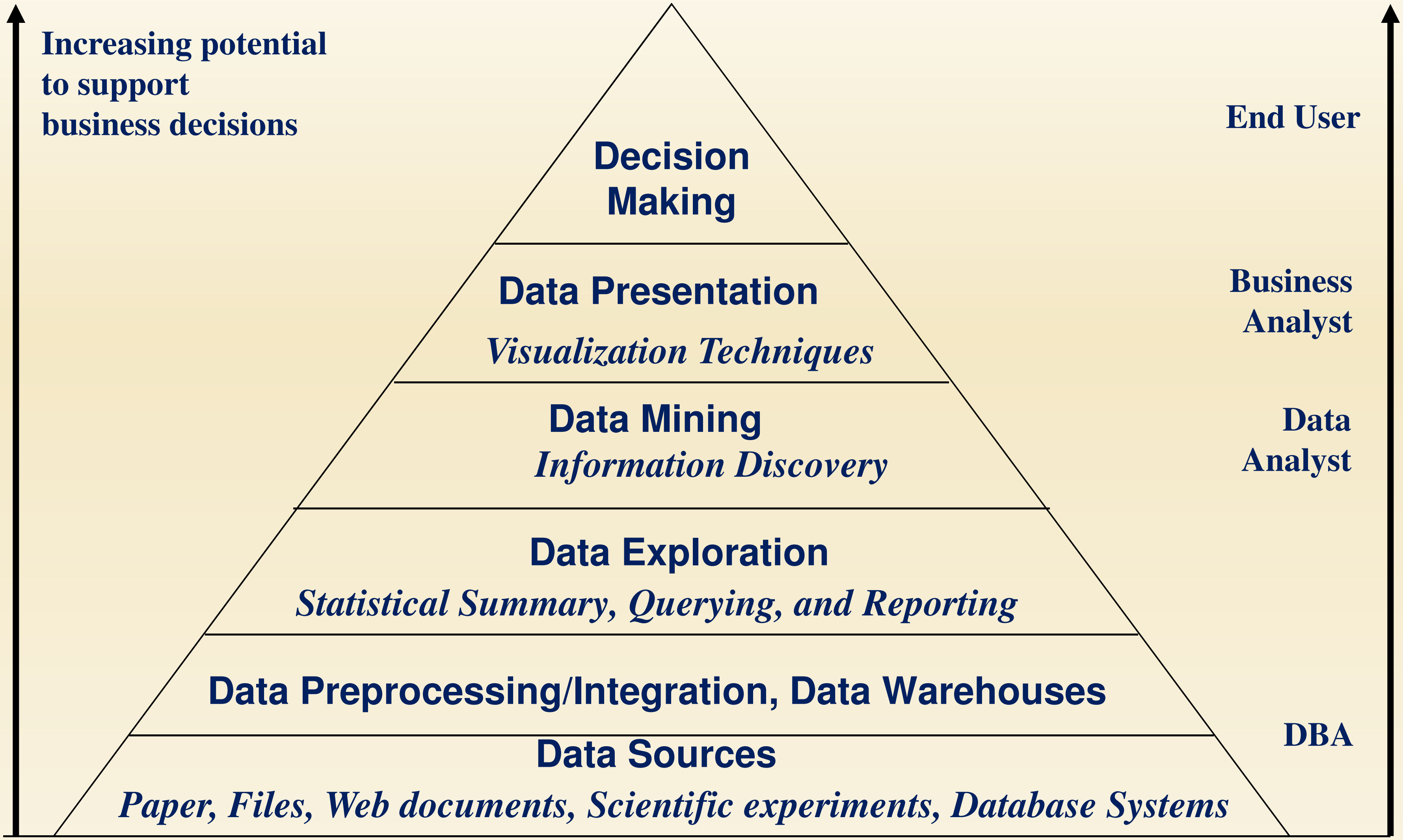
Data Mining

Background & Applications

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

Data Mining and Business Intelligence



Increasing potential
to support
business decisions

Decision Making

Data Presentation

Visualization Techniques

Data Mining

Information Discovery

Data Exploration

Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources

Paper, Files, Web documents, Scientific experiments, Database Systems

End User

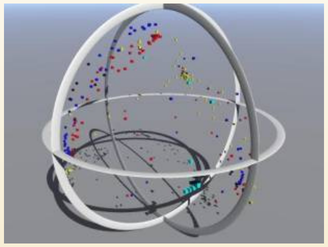
Business Analyst

Data Analyst

DBA

Why Not Traditional Data Analysis?

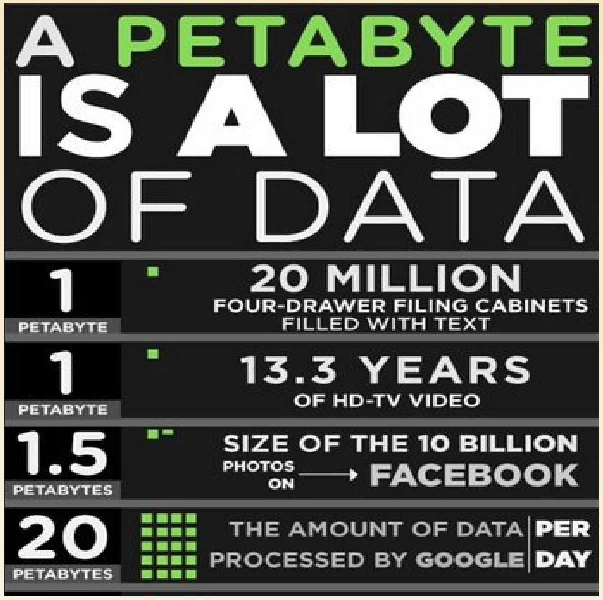
- **High-dimensionality of data**
  - Micro-array may have thousands of dimensions



- **High complexity of data**
  - Time-series data, stream data, sequence data
  - Graphs, social networks
  - Heterogeneous databases
  - Spatial, multimedia, text and Web data
  - Software programs, scientific simulations.

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.



A PETABYTE IS A LOT OF DATA

1 PETABYTE	20 MILLION FOUR-DRAWER FILING CABINETS FILLED WITH TEXT
1 PETABYTE	13.3 YEARS OF HD-TV VIDEO
1.5 PETABYTES	SIZE OF THE 10 BILLION PHOTOS ON → FACEBOOK
20 PETABYTES	THE AMOUNT OF DATA PROCESSED BY GOOGLE PER DAY

Why Data Mining?

- **Major sources of abundant data:**
  - Business: Web, e-commerce, transactions, stocks, …
  - Science: Remote sensing, bioinformatics, scientific simulation, …
  - Society: digital news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets.

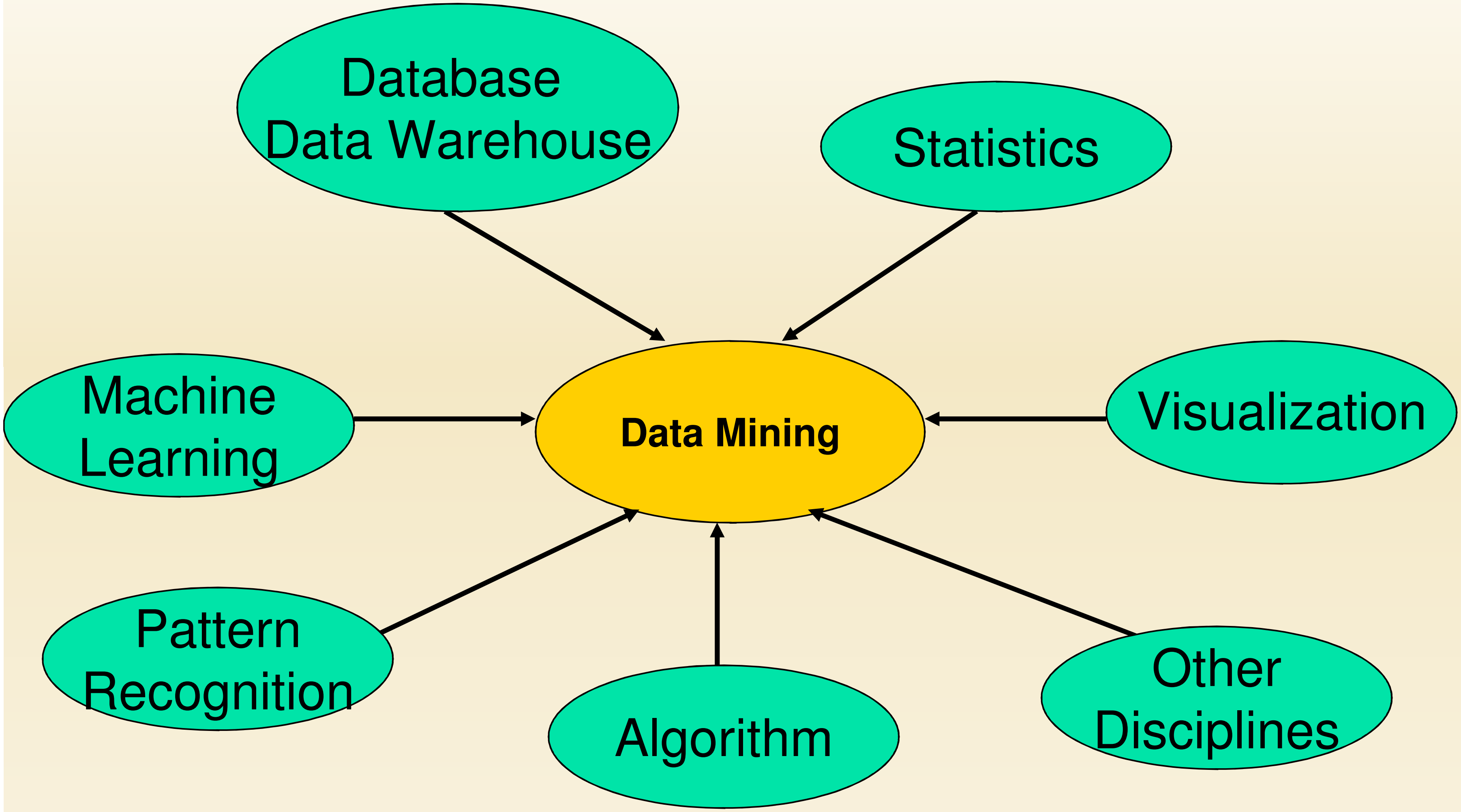
Evolution of Database Technology

- 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.)
- 1990s:
  - Data mining, data warehousing, multimedia databases
- 2000s:
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration)

What Is Data Mining?

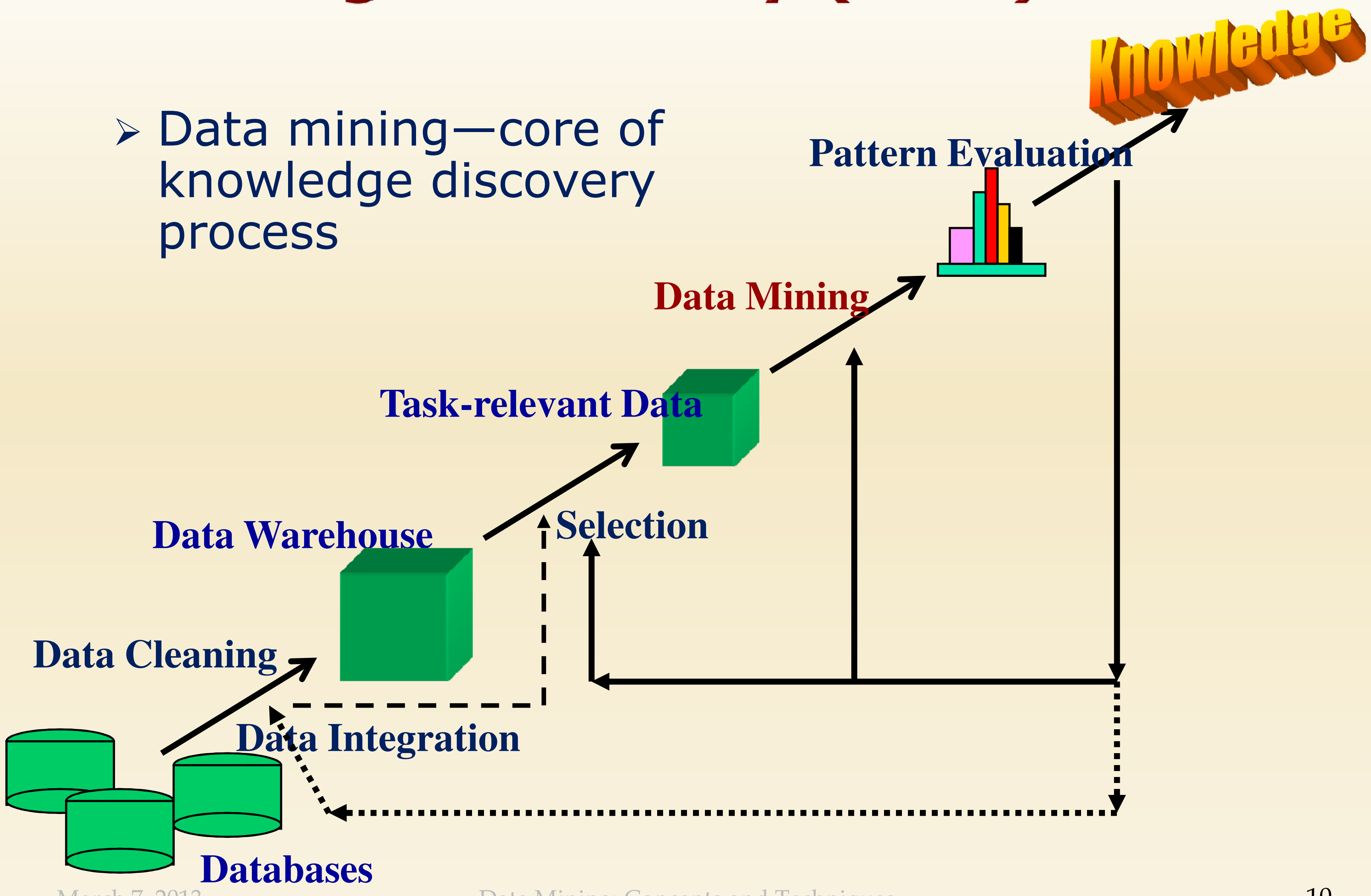
- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names:
  - Knowledge discovery in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems.

Data Mining: Confluence of Multiple Disciplines



Knowledge Discovery (KDD) Process

- Data mining—core of knowledge discovery process



KDD Process: Several Key Steps

- Learning the application domain
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and transformation
  - Find useful features, dimensionality/variable reduction, invariant representation
- Choosing functions of data mining
  - summarization, classification, regression, association, clustering
- Choosing the mining algorithm(s)
- Data mining: search for patterns of interest
- Pattern evaluation and knowledge presentation
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge

Top-10 Algorithms Selected at ICDM'06

- ❑ #1: C4.5 (61 votes)
- ❑ #2: K-Means (60 votes)
- ❑ #3: SVM (58 votes)
- ❑ #4: Apriori (52 votes)
- ❑ #5: EM (48 votes)
- ❑ #6: PageRank (46 votes)
- ❑ #7: AdaBoost (45 votes)
- ❑ #7: kNN (45 votes)
- ❑ #7: Naive Bayes (45 votes)
- ❑ #10: CART (34 votes)

International Conference on Data Mining

These are classification, clustering and association rule mining algorithms

Major Issues in Data Mining

- Mining methodology: Which data mining technique to employ?
  - Mining different kinds of knowledge from diverse data types
  - Performance: efficiency, effectiveness, and scalability
  - Pattern evaluation: the interestingness problem
  - Incorporation of background knowledge
  - Handling noise and incomplete data
  - Parallel and distributed mining
  - Integration of the discovered knowledge with existing one

- User interaction
  - Expression and visualization of data mining results
  - Interactive mining of knowledge at multiple levels of abstraction

- Applications and social impacts
  - Protection of data security, integrity, and privacy

Why Data Mining?—Potential Applications

- Market analysis and management
  - Target marketing, customer relationship management (CRM), market basket analysis, market segmentation
- Risk analysis and management
  - Forecasting, improved underwriting, quality control, competitive analysis
- Fraud detection and detection of unusual patterns (outliers)

41st Parameter

Proven Fraud Detection and Intervention

prevent fraud before it happens.

find out how we can help >

News E-commerce Fraud: The Latest Criminal Schemes

41st Parameter

Data Mining Applications

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

Dynamic Trader ver 5.0.84.112 - (indexes) - FTSE100 - [FTSE-100 Index Cadj ZX6500 D-D]
File Edit Workspaces Chart Indicators Reports Statistics Tools Matrix Cycles Help
Wave Bands
Fri 16-May-2008 6288.0 6360.0 6272.5 6293.0 46.0
Retracement Swing 2
left shoulder
Head
28 traded days
Swing 1
28 days including holidays
Predicted Swing 3
Tue 18-Mar-2008 O 5501.0 H 5621.0 L 5496.0 C 5604.5 R 125.0 160.5
SW1
SW2

Ex. 1: Market Analysis and Management

- Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls

- Target marketing: Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.

- 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

Ex. 2: Corporate Analysis & Risk Management

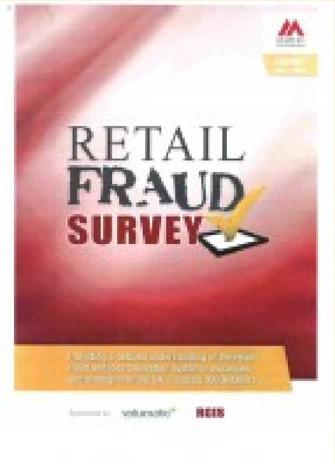
- Finance planning and asset evaluation
  - Cash flow analysis and prediction
  - Contingent claim analysis to evaluate assets
- Resource planning
  - Summarize and compare the resources and spending
- Competition
  - Monitor competitors and market directions
  - Group customers into classes and a class-based pricing procedure
  - Set pricing strategy in a highly competitive market.

Ex. 3: Fraud Detection & Mining Unusual Patterns

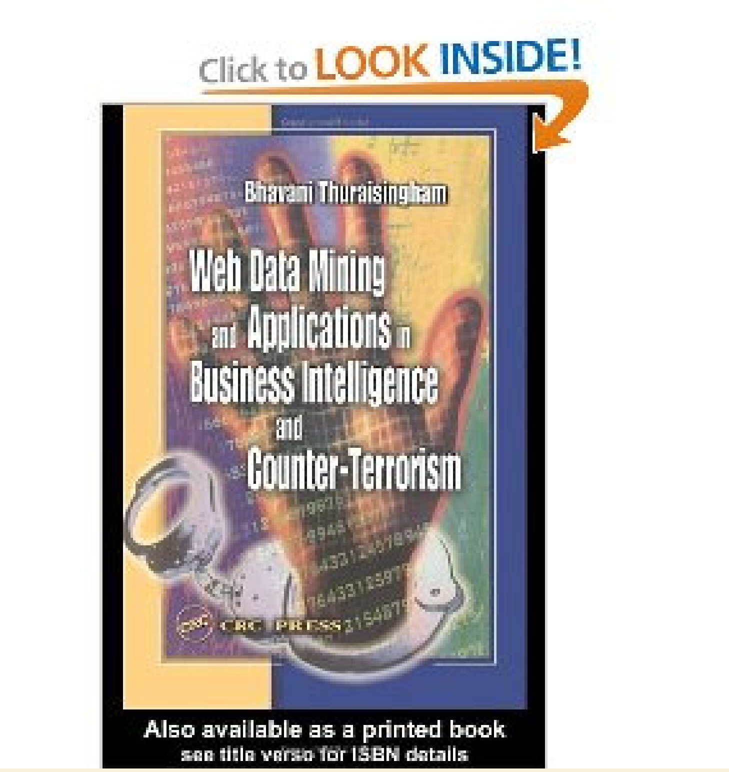
- Money laundering: suspicious monetary transactions
- Medical insurance
  - Professional patients
  - Unnecessary screening tests
- Telecommunications: phone-call fraud
  - Phone call model: destination of the call, duration, time of day or week. Analyze patterns that deviate from an expected norm

Ex: 3

- Retail industry
  - Analysts estimate that 38% of retail shrink is due to dishonest employees
- ATMs
- Anti-terrorism



RETAIL FRAUD SURVEY



Click to LOOK INSIDE!

Bhavani Thuraisingham

Web Data Mining and Applications in Business Intelligence and Counter-Terrorism

CRC PRESS

Also available as a printed book
see title verso for ISBN details

Are All the "Discovered" Patterns Interesting?

- Data mining may generate thousands of patterns: Not all of them are interesting
  - Suggested approach: Human-centered, query-based, focused mining
- **Interestingness measures**
  - 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, or validates some hypothesis that a user seeks to confirm
- **Objective vs. subjective interestingness measures**
  - Objective: based on statistics and structures of patterns, e.g., support, confidence, etc.
  - Subjective: based on user's belief in the data, e.g., unexpectedness, novelty, actionability, etc.

Find All and Only Interesting Patterns?

- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns? Do we need to find all of the interesting patterns?

- Search for only interesting patterns: An optimization problem
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First generate all the patterns and then filter out the uninteresting ones
    - Generate only the interesting patterns—mining query optimization

Other Pattern Mining Issues

- Precise patterns vs. approximate patterns
  - Association: find sets of possible precise patterns
  - But approximate patterns can be more compact and sufficient (in the presence of noise)
  - Gene sequence mining: approximate patterns are inherent
  - How to derive efficient approximate pattern mining algorithms??
- Constrained vs. non-constrained patterns
  - Why constraint-based mining?
  - What are the possible kinds of constraints? How to push constraints into the mining process?

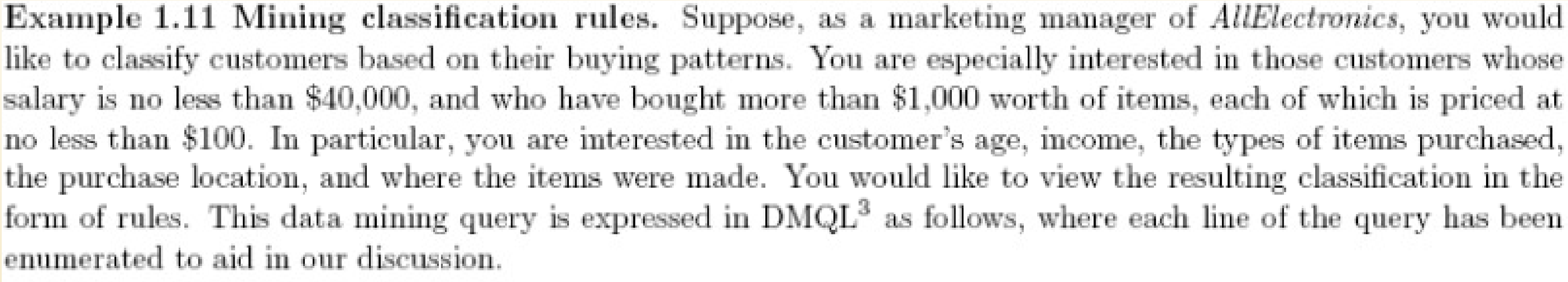
Why Data Mining Query Language?

- Automated vs. query-driven?
  - Finding all the patterns autonomously in a database?—unrealistic because some could be uninteresting
- Data mining should be an interactive process
- Users must be provided with a set of primitives to be used to communicate with the data mining system
- Incorporating these primitives in a data mining query language
  - More flexible user interaction
  - Standardization of data mining industry and practice

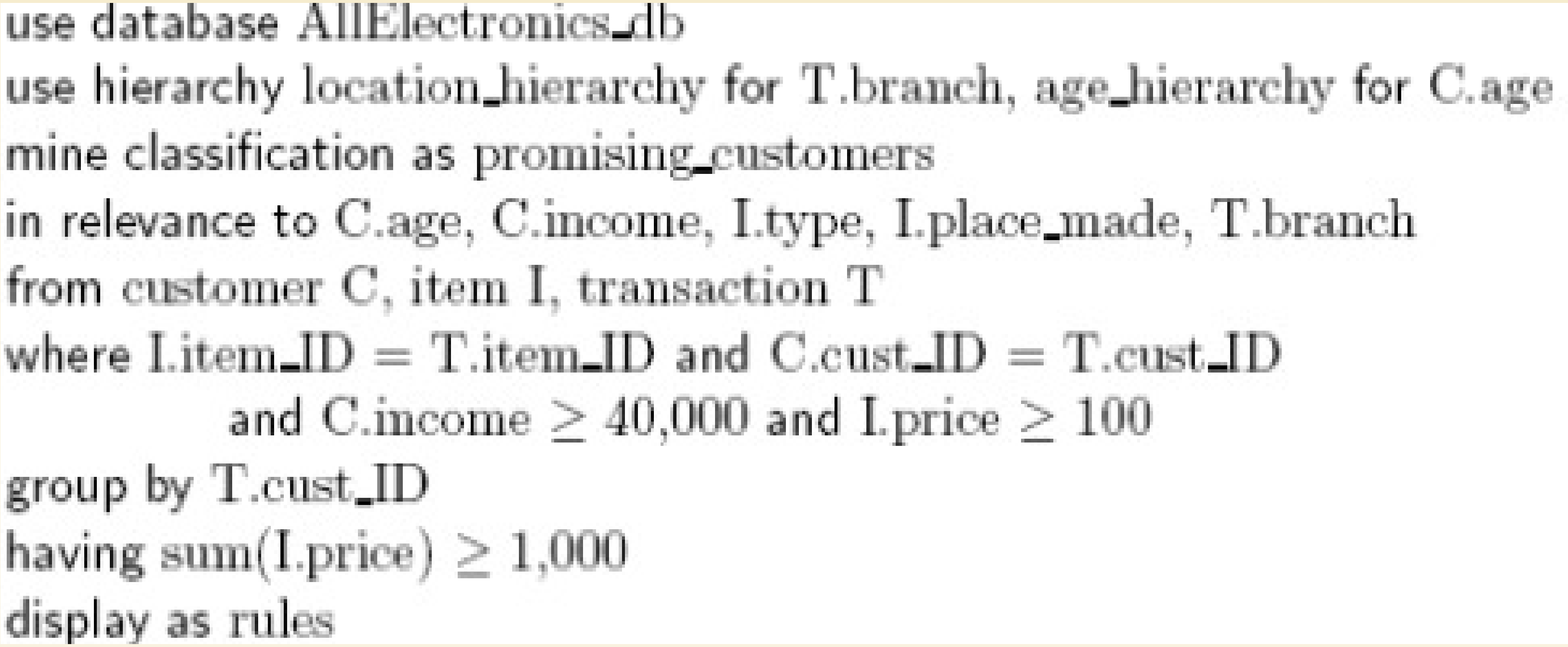
DMQL—A Data Mining Query Language

- Motivation
  - A DMQL can provide the ability to support ad-hoc and interactive data mining
  - By providing a standardized language like SQL
    - Hope to achieve a similar effect like that SQL has on relational database
    - Foundation for system development and evolution
    - Facilitate information exchange, technology transfer, commercialization and wide acceptance
- Design
  - DMQL is designed with the primitives described earlier.

An Example Query in DMQL



Example 1.11 Mining classification rules. Suppose, as a marketing manager of *AllElectronics*, you would like to classify customers based on their buying patterns. You are especially interested in those customers whose salary is no less than $40,000, and who have bought more than $1,000 worth of items, each of which is priced at no less than $100. In particular, you are interested in the customer's age, income, the types of items purchased, the purchase location, and where the items were made. You would like to view the resulting classification in the form of rules. This data mining query is expressed in DMQL[3] as follows, where each line of the query has been enumerated to aid in our discussion.



```
use database AllElectronics_db
use hierarchy location_hierarchy for T.branch, age_hierarchy for C.age
mine classification as promising_customers
in relevance to C.age, C.income, I.type, I.place_made, T.branch
from customer C, item I, transaction T
where I.item_ID = T.item_ID and C.cust_ID = T.cust_ID
          and C.income ≥ 40,000 and I.price ≥ 100
group by T.cust_ID
having sum(I.price) ≥ 1,000
display as rules
```

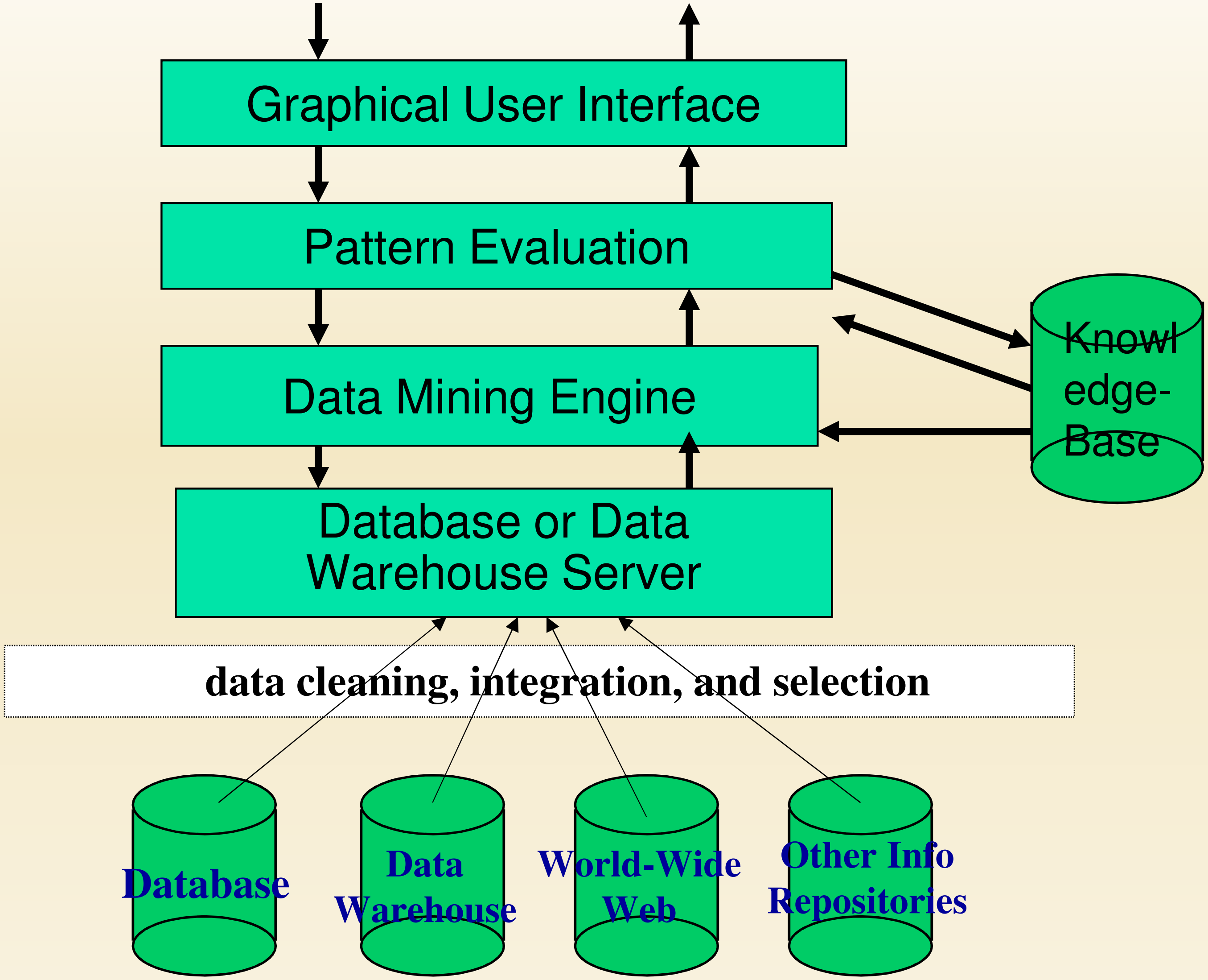
Other Data Mining Languages & Standardization Efforts

- Association rule language specifications
  - MSQL (Imielinski & Virmani'99)
  - MineRule (Meo Psaila and Ceri'96)
  - Query flocks based on Datalog syntax (Tsur et al'98)
- OLEDB for DM (Microsoft'2000) and recently DMX (Microsoft SQLServer 2005)
  - Based on OLE, OLE DB, OLE DB for OLAP, C#
  - Integrating DBMS, data warehouse and data mining
- DMML (Data Mining Mark-up Language) by DMG (www.dmg.org)
  - Providing a platform and process structure for effective data mining
  - Emphasizing on deploying data mining technology to solve business problems

Coupling Data Mining with DB/DW Systems

- No coupling—flat file processing, not recommended
- Loose coupling
  - Fetching data from DB/DW
- Semi-tight coupling—enhanced DM performance
  - Provide efficient implementation of a few data mining primitives in a DB/DW system, e.g., sorting, indexing, aggregation, histogram analysis, multi-way join, pre-computation of some stat functions
- Tight coupling—A uniform information processing environment
  - DM is smoothly integrated into a DB/DW system, mining query is optimized based on indexing, query processing methods, etc.

Architecture: Typical Data Mining System



Open Source DM Tools

- Rapid miner
- Weka
- Orange
- KNIME

- SPSS

http://www.kdnuggets.com/

Questions