

# 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.) 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** is the process of identifying valid, novel, useful and understandable patterns in data.
- Also known as **KDD** (**K**nowledge **D**iscovery in **D**atabases).
- Data Mining refers to *extracting or “mining” knowledge from large amounts of data.*

# Data Mining

- The non-trivial extraction of novel, implicit, and actionable knowledge from large datasets.
  - Extremely large datasets
  - Discovery of the non-obvious
  - Useful knowledge that can improve processes
  - Can not be done manually
- Technology to enable data exploration, data analysis, and data visualization of very large databases at a high level of abstraction, without a specific hypothesis in mind.
- Sophisticated data search capability that uses statistical algorithms to discover patterns and correlations in data.

# Data Mining

- Data Mining is a step of Knowledge Discovery in Databases (KDD) Process
  - Data Warehousing
  - Data Selection
  - Data Preprocessing
  - Data Transformation
  - Data Mining
  - Interpretation/Evaluation
- Data Mining is sometimes referred to as KDD and DM and KDD tend to be used as synonyms.

# Data Mining: Classification Schemes

- Decisions in data mining
  - Kinds of databases to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted
- Data mining tasks
  - Descriptive data mining
  - Predictive data mining

# Decisions in Data Mining

- **Databases to be mined**
  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.
- **Knowledge to be mined**
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.



# Data Mining Tasks

- Prediction Tasks
  - Use some variables to predict unknown or future values of other variables.
  - Predictive mining tasks perform inference on the current data in order to make predictions.
- Description Tasks
  - Find human-interpretable patterns that describe the data.
  - Descriptive mining tasks characterize the general properties of the data in the database.

## Common data mining tasks

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

# Classification: Definition

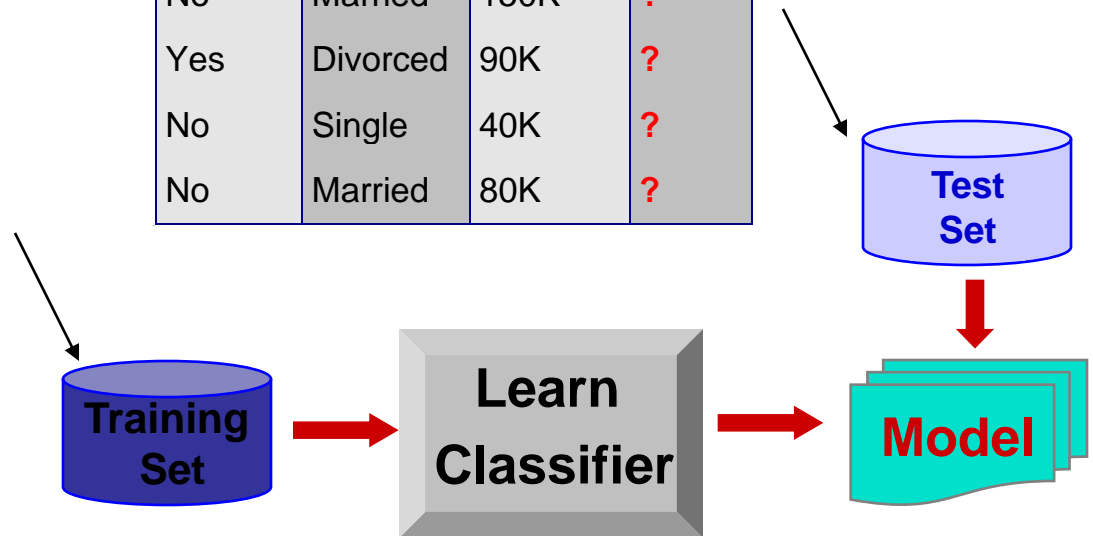
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- *Classification is the process of finding a model (or function) that describes and distinguishes data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data*
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

# Classification Example

*categorical*  
*categorical*  
*continuous*  
*class*

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



# Classification: Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
    - Use the data for a similar product introduced before.
    - We know which customers decided to buy and which decided otherwise. This *{buy, don't buy}* decision forms the *class attribute*.
    - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
      - Type of business, where they stay, how much they earn, etc.
    - Use this information as input attributes to learn a classifier model.

# Classification: Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

# Clustering Definition

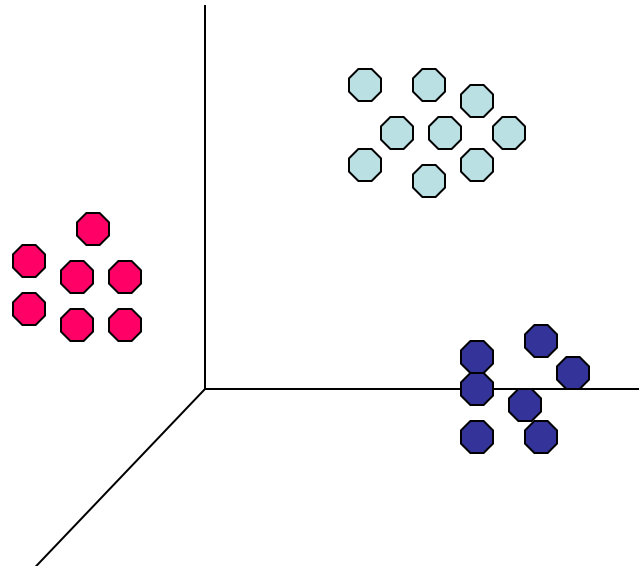
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

# Illustrating Clustering

x Euclidean Distance Based Clustering in 3-D space.

Intracluster distances  
are minimized

Intercluster distances  
are maximized



# Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.



# Clustering: Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

# Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection;
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

**{Milk} --> {Coke}**

**{Diaper, Milk} --> {Beer}**

# Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be  
 $\{Bagels, \dots\} \rightarrow \{Potato Chips\}$
  - Potato Chips as consequent => Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent => Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent and Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

# Association Rule Discovery

- Suppose, as a marketing manager of *AllElectronics*, you would like to determine which items are frequently purchased together within the same transactions.
- An example of such a rule, mined from the *AllElectronics transactional database*, is

$buys(X, \text{"computer"}) \Rightarrow buys(X, \text{"software"})$  [support = 1%, confidence = 50%]

- where  $X$  is a variable representing a **customer**.
- A **confidence**, or certainty, of 50% means that if a customer buys a computer, there is a 50% chance that she will buy software.
- A 1% **support** means that 1% of all of the transactions under analysis showed that computer and software were purchased together.

This association rule involves a single attribute or predicate (i.e., *buys*) that repeats. Association rules that contain a single predicate are referred to as **single-dimensional association rules**. Dropping the predicate notation, the rule can be written simply as “

“*computer*  $\Rightarrow$  *software* [1%, 50%]”.

# Sequential Pattern Discovery: Definition

Given is a set of *objects*, with each object associated with its own *timeline of events*, find rules that predict strong **sequential dependencies** among different events:

- In telecommunications alarm logs,
  - (Inverter\_Problem Excessive\_Line\_Current)  
(Rectifier\_Alarm) --> (Fire\_Alarm)
- In point-of-sale transaction sequences,
  - Computer Bookstore:  
(Intro\_To\_Visual\_C) (C++\_Primer) -->  
(Perl\_for\_dummies,Tcl\_Tk)
  - Athletic Apparel Store:  
(Shoes) (Racket, Racketball) --> (Sports\_Jacket)

# Regression

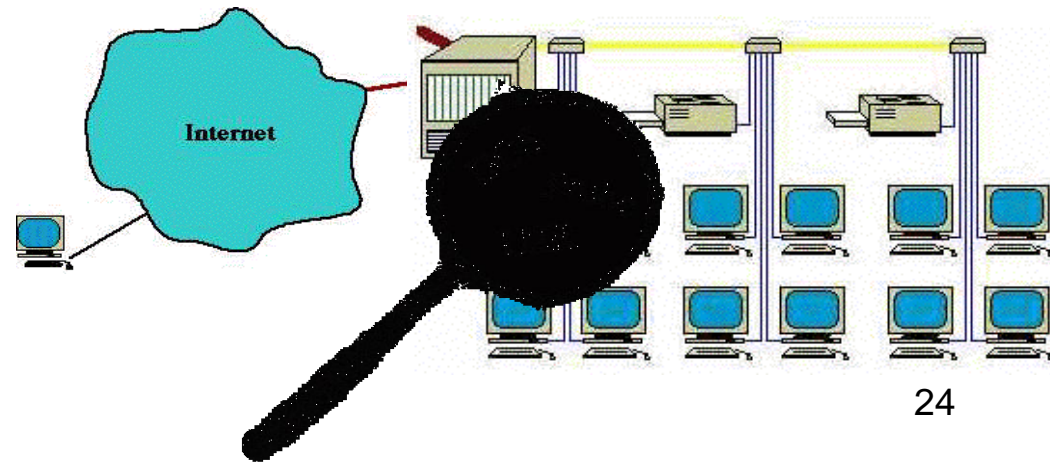
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Regression analysis is a statistical methodology that is most often used for numeric prediction, although other methods exist as well. Prediction also encompasses the identification of distribution *trends based on the available data*.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

# Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection



- Network Intrusion Detection





# The Data Mining Process

- Understanding domain, prior knowledge, and goals
- Data integration and selection
- Data cleaning and pre-processing
- Modeling and searching for patterns
- Interpreting results
- Consolidating and deploying discovered knowledge
- Loop

# Why Data Mining? — Potential 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
  - Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.
- Determine customer purchasing patterns over time
  - Conversion of single to a joint bank account: marriage, etc.
- Cross-market analysis
  - Associations/co-relations between product sales
  - Prediction based on the association information

# Market Analysis and Management

- Customer profiling
  - data mining can tell you what types of customers buy what products (clustering or classification)
- 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)

# Corporate Analysis and 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

# 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

- Detecting inappropriate medical treatment
  - Australian Health Insurance Commission identifies that in many cases blanket screening tests were requested (save Australian \$1m/yr).
- 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

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



# Data Mining Applications: Retail

- Performing basket analysis
  - Which items customers tend to purchase together. This knowledge can improve stocking, store layout strategies, and promotions.
- Sales forecasting
  - Examining time-based patterns helps retailers make stocking decisions. If a customer purchases an item today, when are they likely to purchase a complementary item?
- Database marketing
  - Retailers can develop profiles of customers with certain behaviors, for example, those who purchase designer labels clothing or those who attend sales. This information can be used to focus cost-effective promotions.
- Merchandise planning and allocation
  - When retailers add new stores, they can improve merchandise planning and allocation by examining patterns in stores with similar demographic characteristics. Retailers can also use data mining to determine the ideal layout for a specific store.

# Data Mining Applications: Banking

- Card marketing
  - By identifying customer segments, card issuers and acquirers can improve profitability with more effective acquisition and retention programs, targeted product development, and customized pricing.
- Cardholder pricing and profitability
  - Card issuers can take advantage of data mining technology to price their products so as to maximize profit and minimize loss of customers. Includes risk-based pricing.
- Fraud detection
  - Fraud is enormously costly. By analyzing past transactions that were later determined to be fraudulent, banks can identify patterns.
- Predictive life-cycle management
  - DM helps banks predict each customer's lifetime value and to service each segment appropriately (for example, offering special deals and discounts).

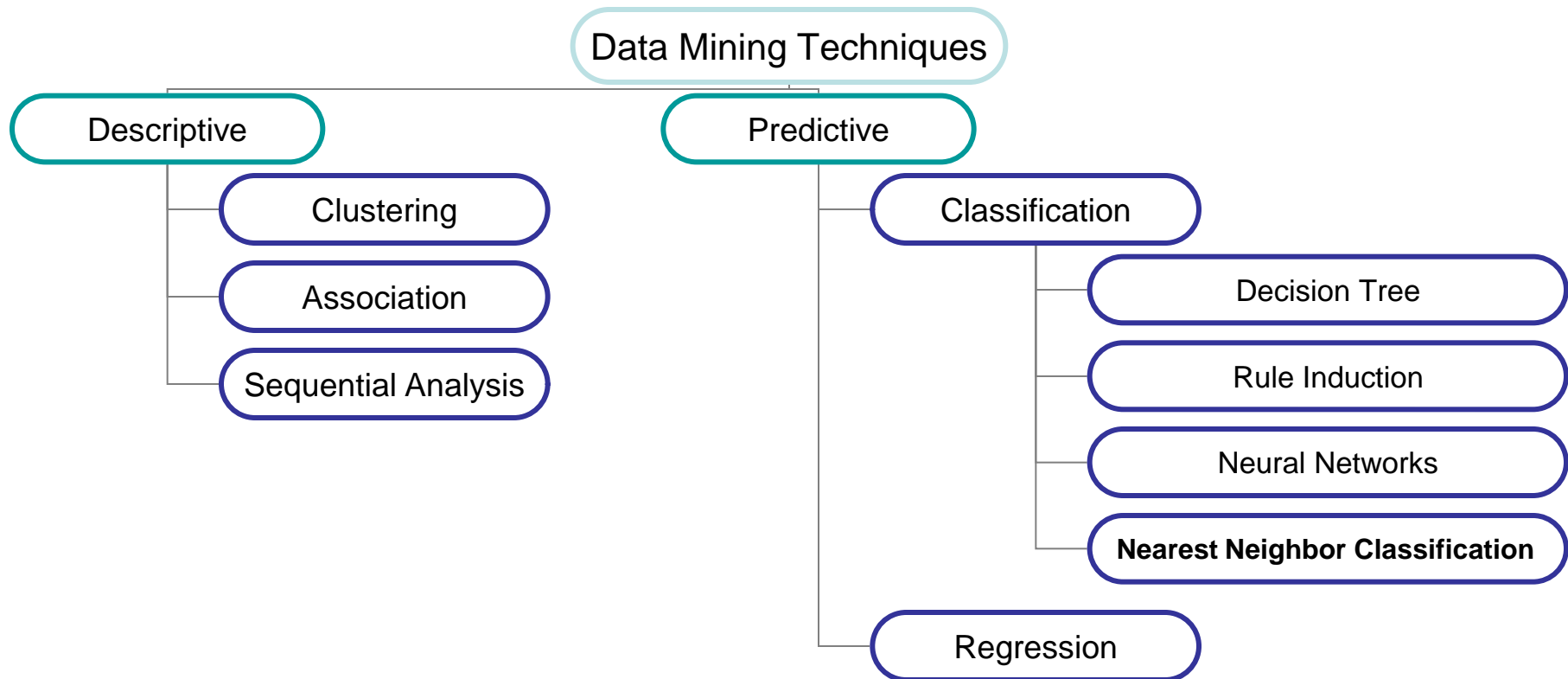
# Data Mining Applications: Telecommunication

- Call detail record analysis
  - Telecommunication companies accumulate detailed call records. By identifying customer segments with similar use patterns, the companies can develop attractive pricing and feature promotions.
- Customer loyalty
  - Some customers repeatedly switch providers, or “churn”, to take advantage of attractive incentives by competing companies. The companies can use DM to identify the characteristics of customers who are likely to remain loyal once they switch, thus enabling the companies to target their spending on customers who will produce the most profit.

# Data Mining Applications: Other Applications

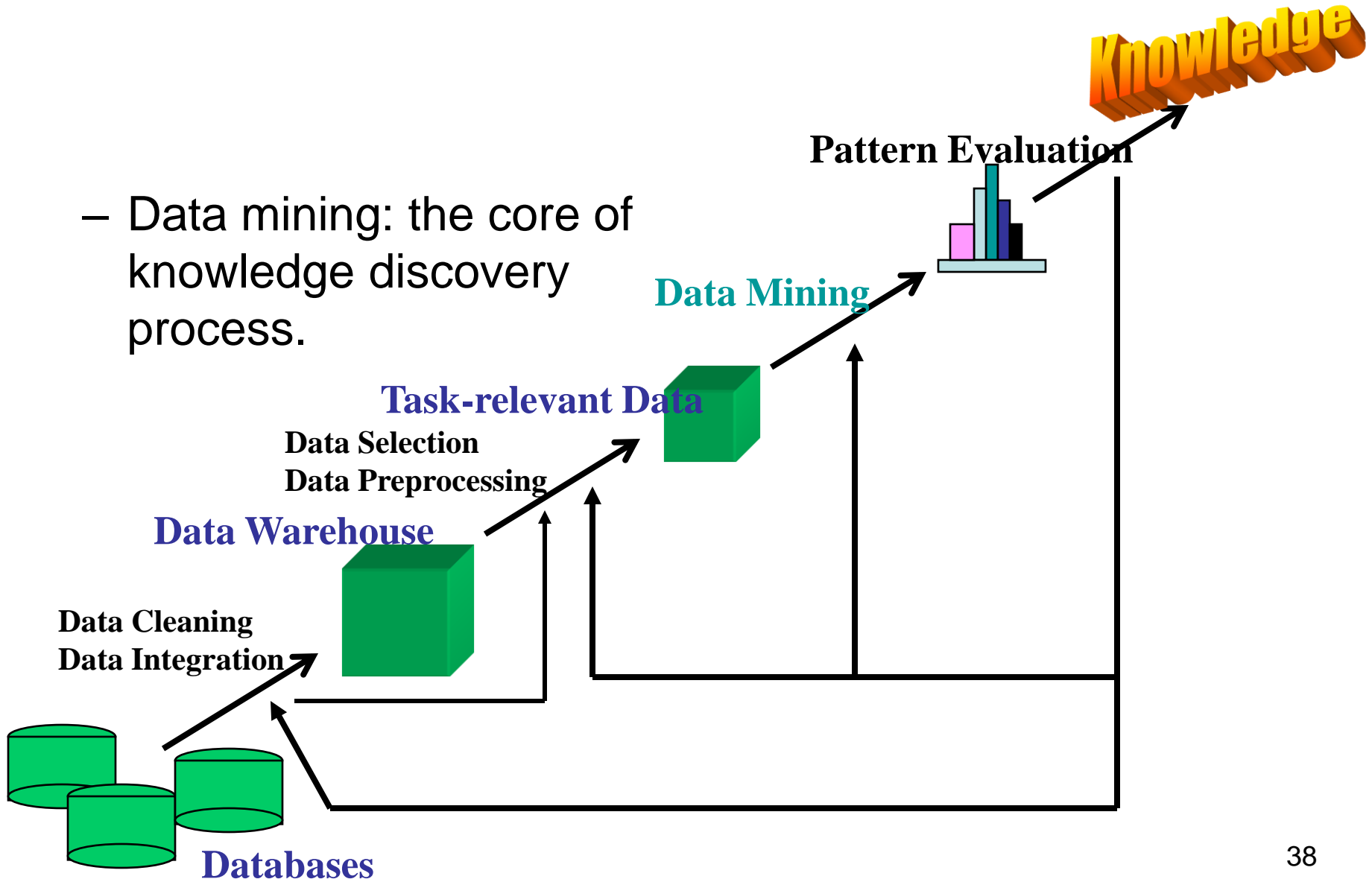
- Customer segmentation
  - All industries can take advantage of DM to discover discrete segments in their customer bases by considering additional variables beyond traditional analysis.
- Manufacturing
  - Through choice boards, manufacturers are beginning to customize products for customers; therefore they must be able to predict which features should be bundled to meet customer demand.
- Warranties
  - Manufacturers need to predict the number of customers who will submit warranty claims and the average cost of those claims.
- Frequent flier incentives
  - Airlines can identify groups of customers that can be given incentives to fly more.

# Data Mining Techniques



# Data Mining: A KDD Process

- Data mining: the core of knowledge discovery process.



# Steps of a KDD Process

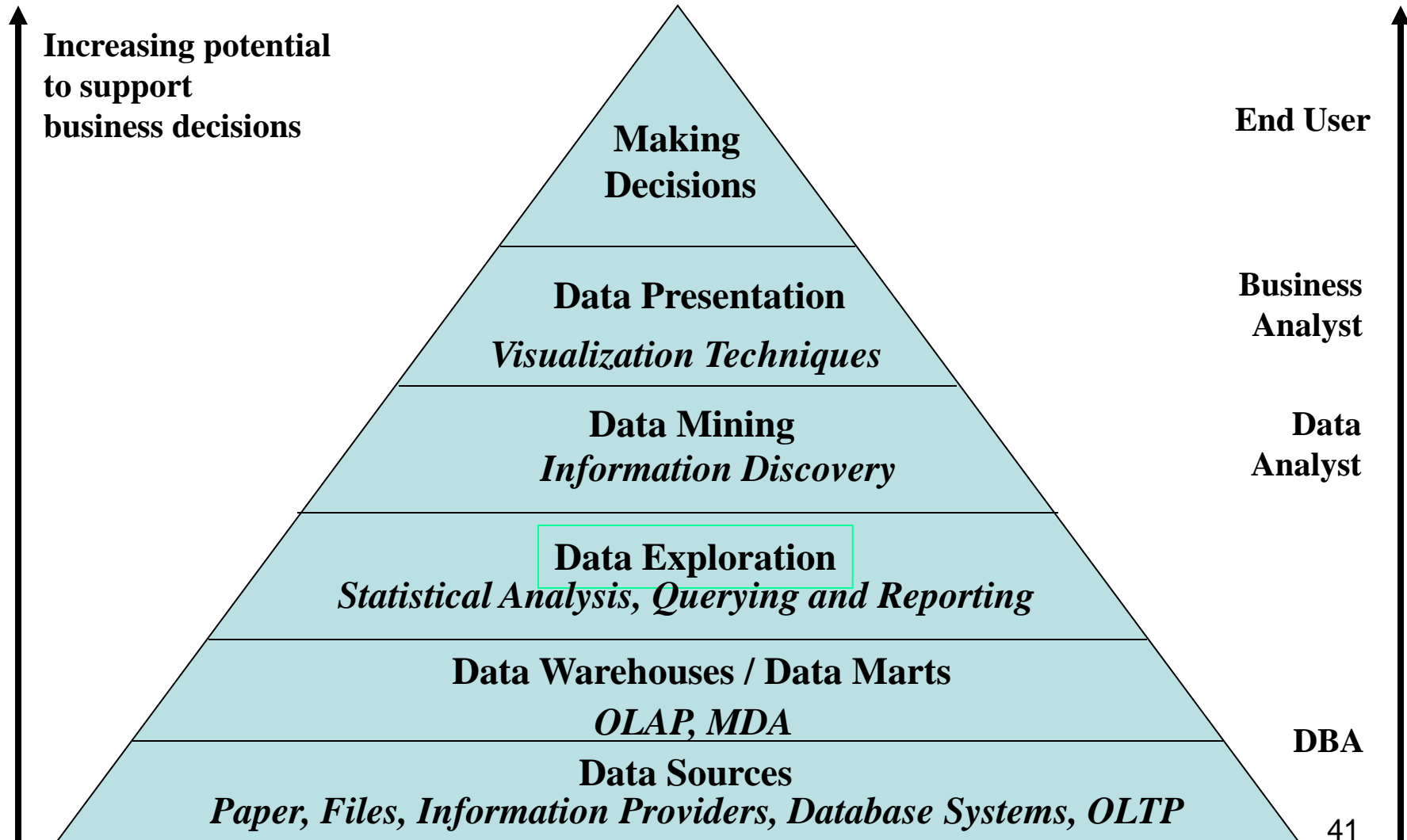
- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- **Data cleaning** and preprocessing: (may take 60% of effort!)
- **Data reduction and transformation:**
  - Find useful features, dimensionality/variable reduction, 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

# Steps of a KDD Process

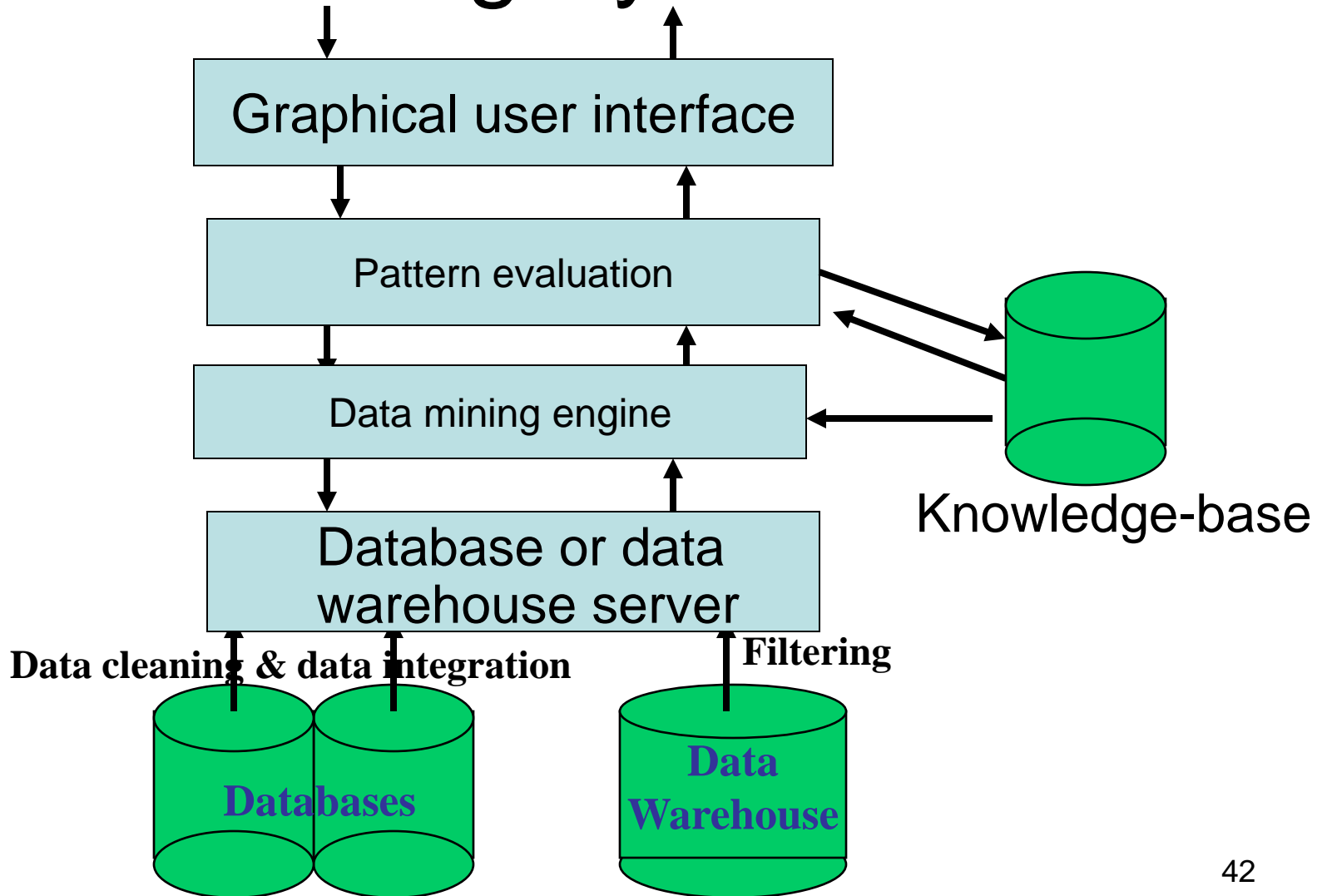
- **Data cleaning** (to remove noise and inconsistent data)
- **Data integration** (where multiple data sources may be combined)
- **Data selection** (where data relevant to the analysis task are retrieved from the database)
- **Data transformation** (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
- **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



# Architecture of a Typical Data Mining System



# Architecture of a Typical Data Mining System

- **Database, data warehouse, WorldWideWeb, or other information repository:** This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.
- **Database or data warehouse server:** The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.
- **Knowledge base:**
  - concept hierarchies
  - user beliefs
  - asses pattern's interestingness
  - other thresholds

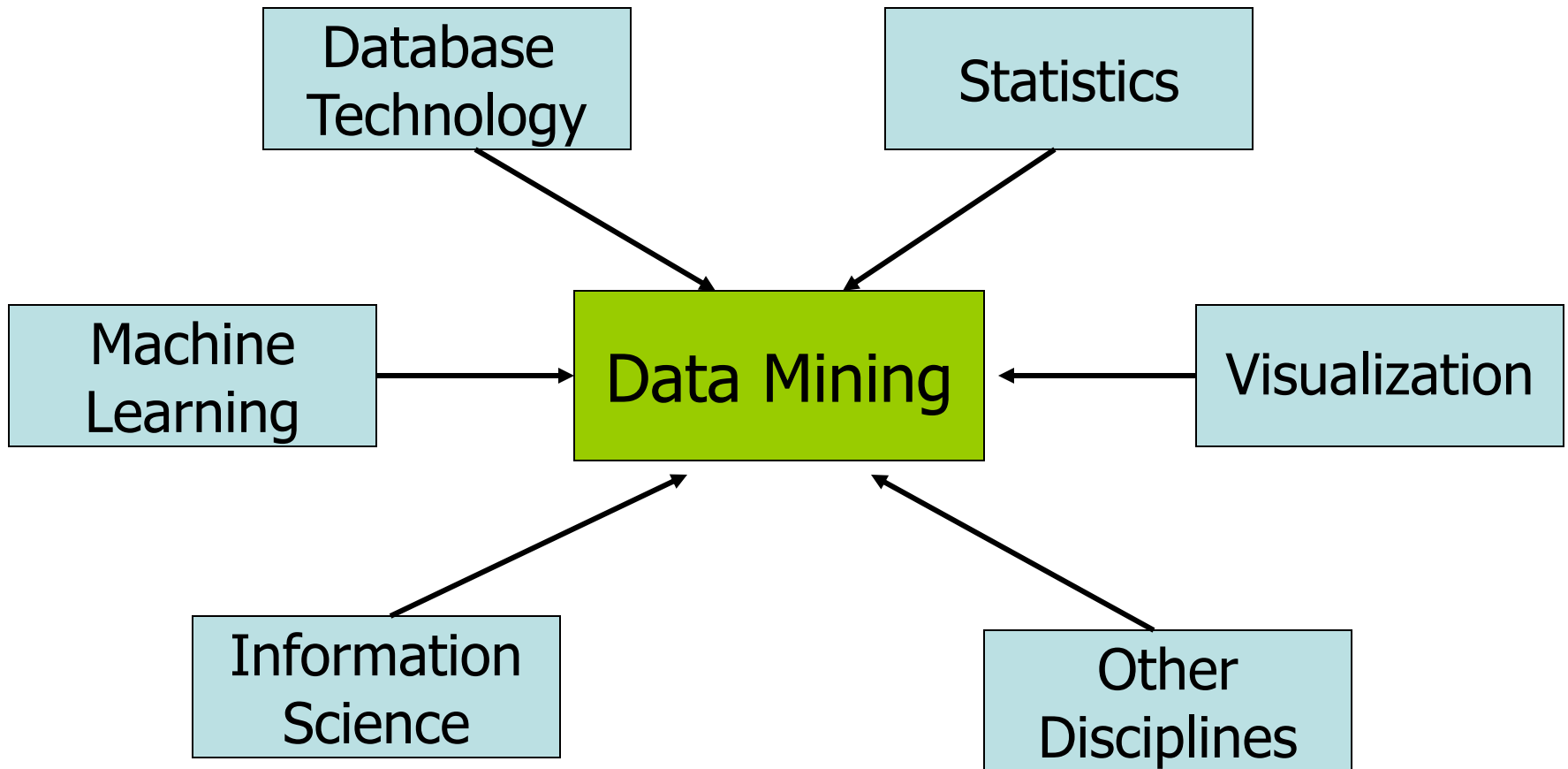
# Architecture of a Typical Data Mining System

- **Data mining engine:**
  - functional modules
    - characterization, association, classification, cluster analysis, evolution and deviation analysis
- **Pattern evaluation module:** This component typically employs interestingness measures and interacts with the data mining modules so as to *focus the* search toward interesting patterns.
- **Graphical user interface:** This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results.

# Data Mining: On What Kind of Data?

- Relational databases
- Data warehouses
- Transactional databases
- Advanced DB and information repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - WWW

# Data Mining: Confluence of Multiple Disciplines



# Data Mining vs. Statistical Analysis

## Statistical Analysis:

- Ill-suited for Nominal and Structured Data Types
- Completely data driven - incorporation of domain knowledge not possible
- Interpretation of results is difficult and daunting
- Requires expert user guidance

## Data Mining:

- Large Data sets
- Efficiency of Algorithms is important
- Scalability of Algorithms is important
- Real World Data
- Lots of Missing Values
- Pre-existing data - not user generated
- Data not static - prone to updates
- Efficient methods for data retrieval available for use

# Data Mining vs. DBMS

- Example DBMS Reports
  - Last months sales for each service type
  - Sales per service grouped by customer sex or age bracket
  - List of customers who lapsed their policy
- Questions answered using Data Mining
  - What characteristics do customers that lapse their policy have in common and how do they differ from customers who renew their policy?
  - Which motor insurance policy holders would be potential customers for my House Content Insurance policy?



# Data Mining and Data Warehousing

- Data Warehouse: a centralized data repository which can be queried for business benefit.
- Data Warehousing makes it possible to
  - extract archived operational data
  - overcome inconsistencies between different legacy data formats
  - integrate data throughout an enterprise, regardless of location, format, or communication requirements
  - incorporate additional or expert information
- OLAP: On-line Analytical Processing
- Multi-Dimensional Data Model (Data Cube)
- Operations:
  - Roll-up
  - Drill-down
  - Slice and dice
  - Rotate

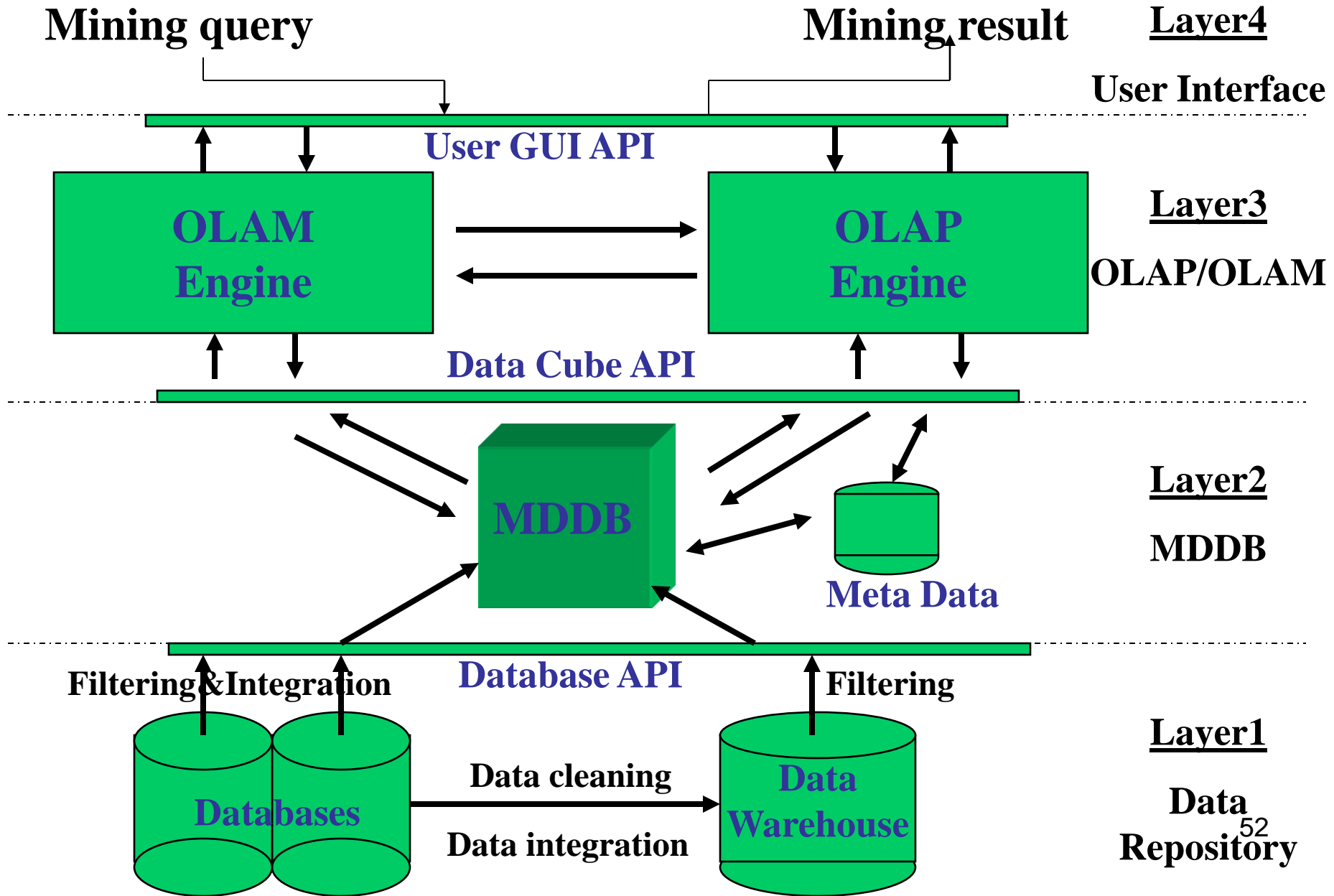
# From On-Line Analytical Processing to On Line Analytical Mining (OLAM)

- Why online analytical mining?
  - High quality of data in data warehouses
    - DW contains integrated, consistent, cleaned data
  - Available information processing structure surrounding data warehouses
    - ODBC, OLEDB, Web accessing, service facilities, reporting and OLAP tools
  - OLAP-based exploratory data analysis
    - mining with drilling, dicing, pivoting, etc.
  - On-line selection of data mining functions
    - integration and swapping of multiple mining functions, algorithms, and tasks.

# From On-Line Analytical Processing to On Line Analytical Mining (OLAM)

- **On-line analytical mining (OLAM)** (also called OLAP mining) integrates on-line analytical processing (OLAP) with data mining and mining knowledge in multidimensional databases

# An OLAM Architecture



# An OLAM Architecture

- An OLAM server performs analytical mining in data cubes in a similar manner as an OLAP server performs on-line analytical processing.
- An integrated OLAM and OLAP architecture is shown, where the OLAM and OLAP servers both accept user on-line queries (or commands) via a graphical user interface API and work with the data cube in the data analysis via a cube API.
- A metadata directory is used to guide the access of the data cube. The data cube can be constructed by accessing and/or integrating multiple databases via an MDDDB API and/or by filtering a data warehouse via a database API that may support OLE DB or ODBC connections.
- Since an OLAM server may perform multiple data mining tasks, such as concept description, association, classification, prediction, clustering, time-series analysis, and so on, it usually consists of multiple integrated data mining modules and is more sophisticated than an OLAP server.

# Importance of OLAM

**OLAM is particularly important for the following reasons:**

- **High quality of data in data warehouses:** Most data mining tools need to work on integrated, consistent, and cleaned data, which requires costly data cleaning, data integration, and data transformation as preprocessing steps
- **Available information processing infrastructure surrounding data warehouses:** Comprehensive information processing and data analysis infrastructures will be systematically constructed surrounding data warehouses, which include accessing, integration, consolidation, and transformation of multiple heterogeneous databases, ODBC/OLE DB connections, Web-accessing and service facilities, and reporting and OLAP analysis tools.
- **OLAP-based exploratory data analysis:** Effective data mining needs exploratory data analysis. A user will often want to traverse through a database, select portions of relevant data, analyze them at different granularities, and present knowledge/results in different forms.
- **On-line selection of data mining functions:** By integrating OLAP with multiple data mining functions, on-line analytical mining provides users with the flexibility to select desired data mining functions and swap data mining<sup>54</sup> tasks dynamically.

# DBMS, OLAP, and Data Mining

	<b>DBMS</b>	<b>OLAP</b>	<b>Data Mining</b>
<b>Task</b>	Extraction of detailed and summary data	Summaries, trends and forecasts	Knowledge discovery of hidden patterns and insights
<b>Type of result</b>	Information	Analysis	Insight and Prediction
<b>Method</b>	Deduction (Ask the question, verify with data)	Multidimensional data modeling, Aggregation, Statistics	Induction (Build the model, apply it to new data, get the result)
<b>Example question</b>	Who purchased mutual funds in the last 3 years?	What is the average income of mutual fund buyers by region by year?	Who will buy a mutual fund in the next 6 months and why?

# Major Issues in Data Warehousing and Mining

- Mining methodology and user interaction
  - Mining different kinds of knowledge in databases
  - Interactive mining of knowledge at multiple levels of abstraction
  - Incorporation of background knowledge
  - Data mining query languages and ad-hoc data mining
  - Expression and visualization of data mining results
  - Handling noise and incomplete data
  - Pattern evaluation: the interestingness problem
- Performance and scalability
  - Efficiency and scalability of data mining algorithms
  - Parallel, distributed and incremental mining methods



# Major Issues in Data Warehousing and Mining

- Issues relating to the diversity of data types
  - Handling relational and complex types of data
  - Mining information from heterogeneous databases and global information systems (WWW)
- Issues related to applications and social impacts
  - Application of discovered knowledge
    - Domain-specific data mining tools
    - Intelligent query answering
    - Process control and decision making
  - Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
  - Protection of data security, integrity, and privacy