Data Mining: Introduction

CONTACT SESSION - 1

By

Dr. D.VENKATA SUBRAMANIAN

GUEST FACULTY – BITS

CHENNAI CENTER

dvsubramanian@wilp.bits-pilani.ac.in

+91 9941562171

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

AGENDA … TOPICS TO BE DISCUSSED

1. Brief overview of Data Warehouse, Datamart
2. Brief overview of Datamarts, OLTP Vs OLAP
3. Brief overview of DWH and ETL architecture
4. What is Data mining ?
5. Why Data mining?
6. List of functionalities of Data mining!
7. Data Mining Process
8. Issues with Data Mining
9. Applications of Data Mining

What is a Data Warehouse ?

Can I see credit report from Accounts, Sales from marketing and open order report from order entry for this customer

Data from multiple sources is integrated for a subject

A data warehouse is a subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management's decisions.

– WH Inmon

Identical queries will give same results at different times. Supports analysis requiring historical data

Data stored for historical period. Data is populated in the data warehouse on daily/weekly basis depending upon the requirement.

Subject-Oriented

- Data is arranged and optimized to provide answer to questions from diverse functional areas
  - ✓ Data is organized and summarized by topic
    - Sales / Marketing / Finance / Distribution / Etc.

Integrated

- The data warehouse is a centralized, consolidated database that integrated data derived from the entire organization
  - ✓ Multiple Sources
  - ✓ Diverse Sources
  - ✓ Diverse Formats

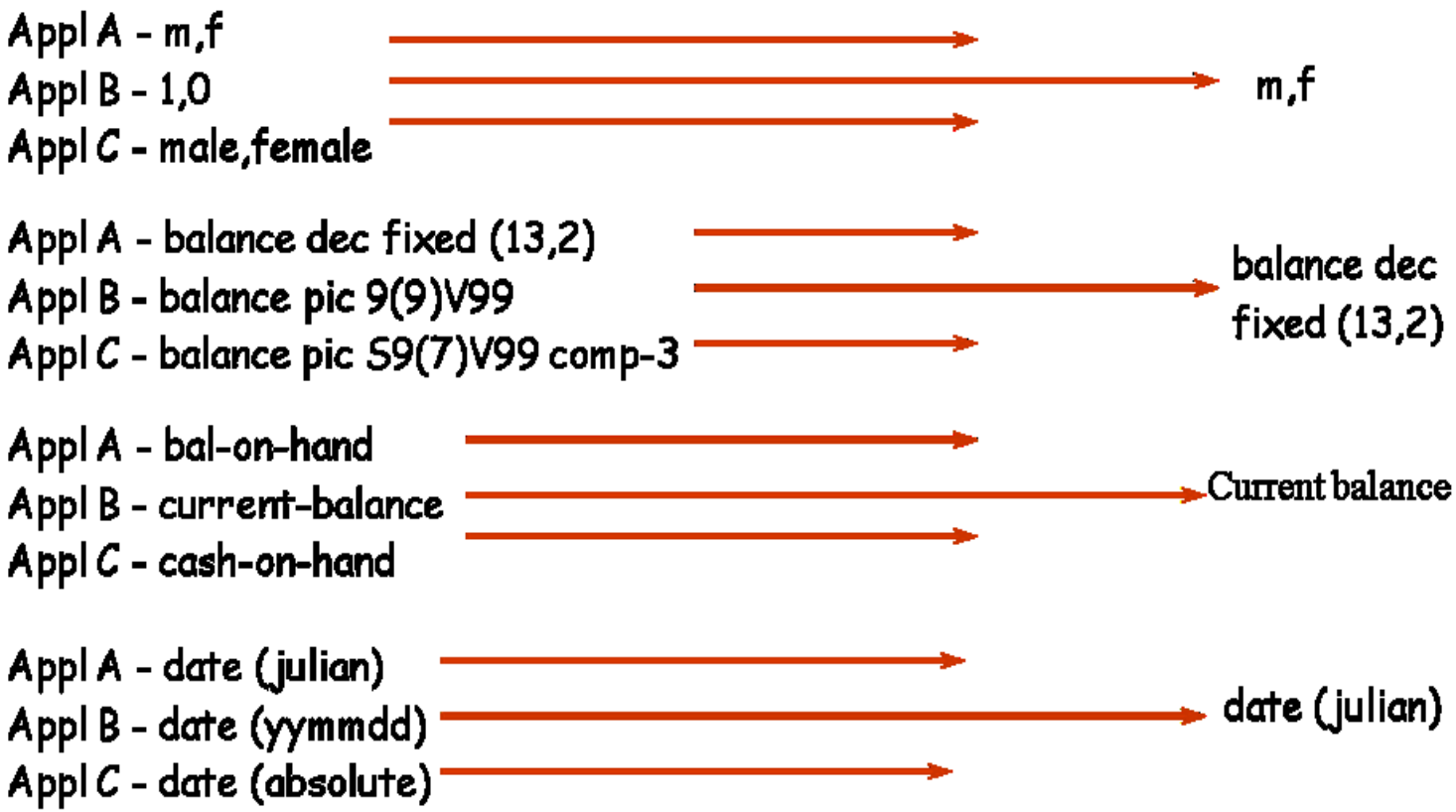
Nonvolatile

- Once data is entered it is NEVER removed
- Represents the company's entire history
  - ✓ Near term history is continually added to it
  - ✓ Always growing
  - ✓ Must support terabyte databases and multiprocessors
- Read-Only database for data analysis and query processing

Time-Variant

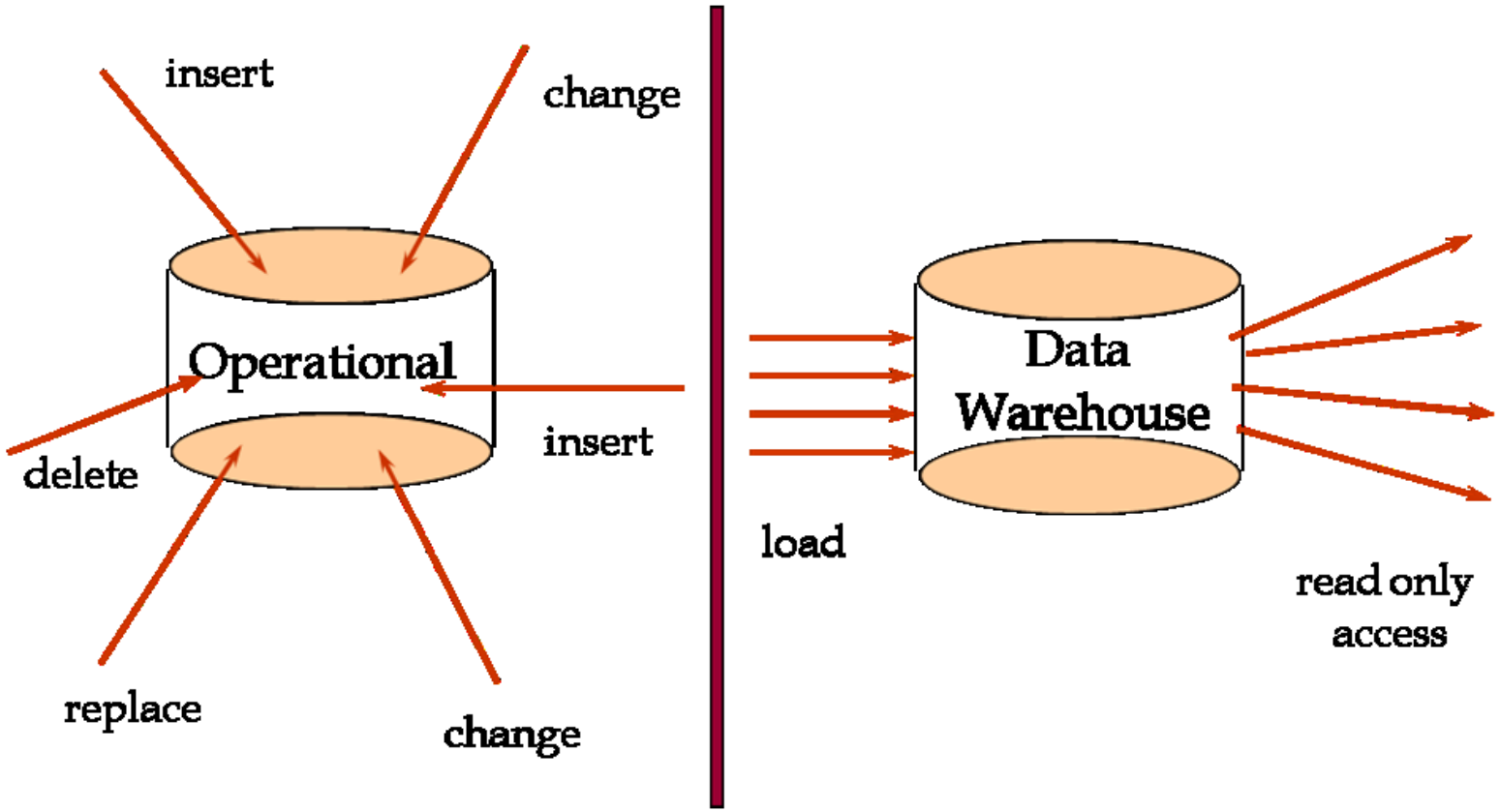
- The Data Warehouse represents the flow of data through time
- Can contain projected data from statistical models
- Data is periodically uploaded then time-dependent data is recomputed

Integrated - Characteristics of a Data Warehouse



Integrated View Is The Essence Of A Data Warehouse

Non-volatile - Characteristics of a Data Warehouse



Integrated View Is The Essence Of A Data Warehouse

12 Rules of a Data Warehouse

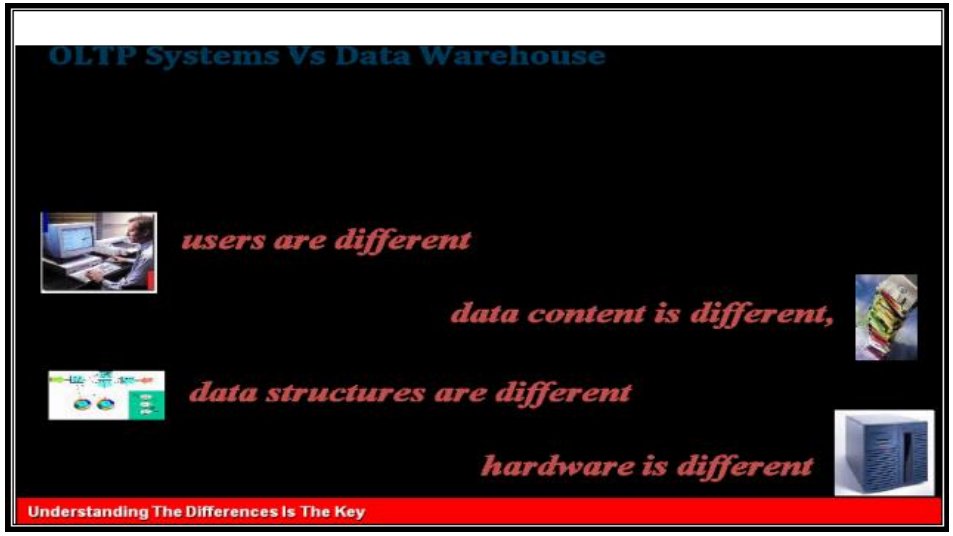
- Data Warehouse and Operational Environments are Separated
- Data is integrated
- Contains historical data over a long period of time
- Data is a snapshot data captured at a given point in time
- Data is subject-oriented

- Mainly read-only with periodic batch updates
- Development Life Cycle has a data driven approach versus the traditional process-driven approach
- Data contains several levels of detail
  - ✓ Current, Old, Lightly Summarized, Highly Summarized

- Environment is characterized by Read-only transactions to very large data sets
- System that traces data sources, transformations, and storage
- Metadata is a critical component
  - ✓ Source, transformation, integration, storage, relationships, history, etc
- Contains a chargeback mechanism for resource usage that enforces optimal use of data by end users

OLTP Vs Warehouse

Operational System	Data Warehouse
Transaction Processing	Query Processing
Time Sensitive	History Oriented
Operator View	Managerial View
Organized by transactions (Order, Input, Inventory)	Organized by subject (Customer, Product)
Relatively smaller database	Large database size
Many concurrent users	Relatively few concurrent users
Volatile Data	Non Volatile Data
Stores all data	Stores relevant data
Not Flexible	Flexible



Data Marts

- Small Data Stores
- More manageable data sets
- Targeted to meet the needs of small groups within the organization
- Small, Single-Subject data warehouse subset that provides decision support to a small group of people

- Enterprise wide data warehousing projects have a very large cycle time
- Getting consensus between multiple parties may also be difficult
- Departments may not be satisfied with priority accorded to them
- Sometimes individual departmental needs may be strong enough to warrant a local implementation
- Application/database distribution is also an important factor

f

Data Warehouse and Data Marts

	Data Warehouse	Data Marts
Scope	●Application Neutral ●Centralized, Shared ●Cross LOB/enterprise	●Specific Application Requirement ●LOB, department ●Business Process Oriented
Data Perspective	●Historical Detailed data ●Some summary	●Detailed (some history) ●Summarized
Subjects	●Multiple subject areas	●Single Partial subject ●Multiple partial subjects

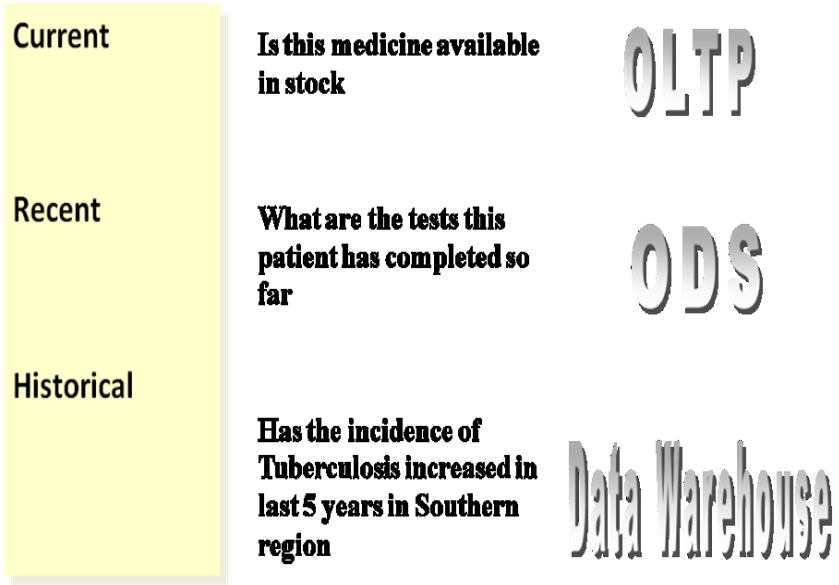
Data Warehouse and Data Marts

	Data Warehouse	Data Marts
Data Sources	●Many ●Operational/ External Data	●Few ●Operational, external data
Implement Time Frame	●9-18 months for first stage ●Multiple stage implementation	●4-12 months
Characteristics	●Flexible, extensible ●Durable/Strategic ●Data orientation	●Restrictive, non extensible ●Short life/tactical ●Project Orientation

Functions of an ODS

- Converts Data,
- Decides Which Data of Multiple Sources Is the Best,
- Summarizes Data,
- Decodes/encodes Data,
- Alters the Key Structures,
- Alters the Physical Structures,
- Reformats Data,
- Internally Represents Data,
- Recalculates Data.

Different kinds of Information Needs



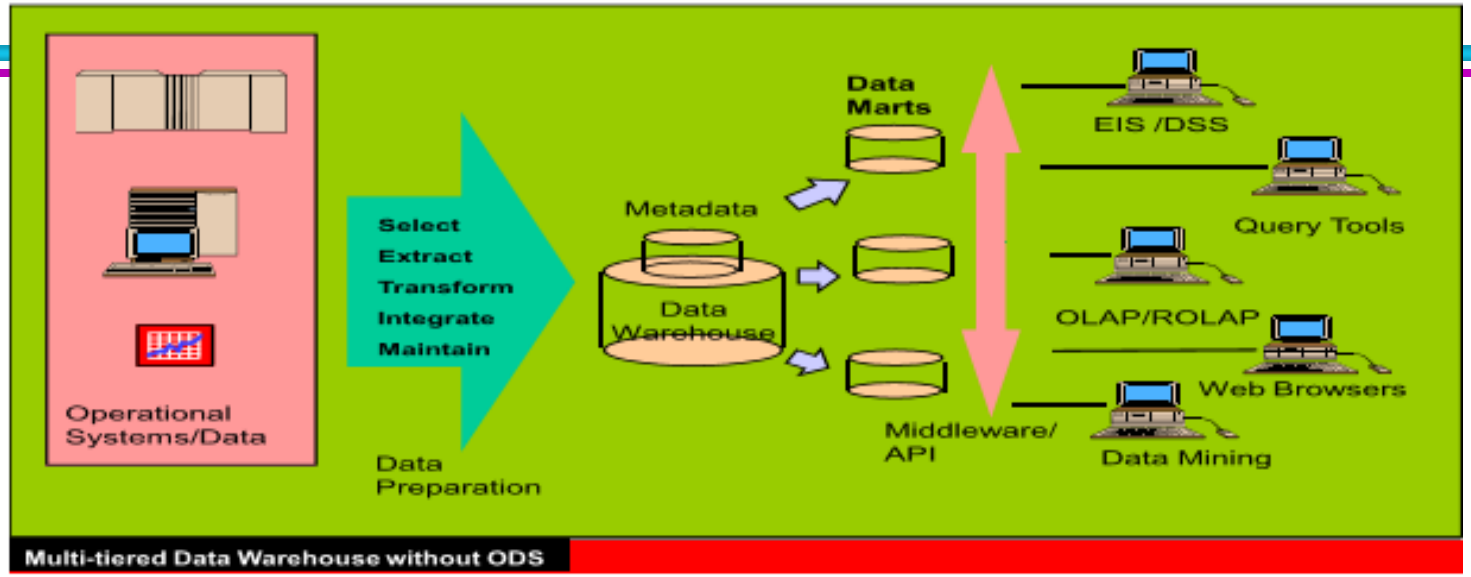
OLTP Vs ODS Vs DWH

Characteristic	OLTP	ODS	Data Warehouse
Audience	Operating Personnel	Analysts	Managers and analysts
Data access	Individual records, transaction driven	Individual records, transaction or analysis driven	Set of records, analysis driven
Data content	Current, real-time	Current and near-current	Historical
Data Structure	Detailed	Detailed and lightly summarized	Detailed and Summarized
Data organization	Functional	Subject-oriented	Subject-oriented
Type of Data	Homogeneous	Homogeneous	Vast Supply of very heterogeneous data

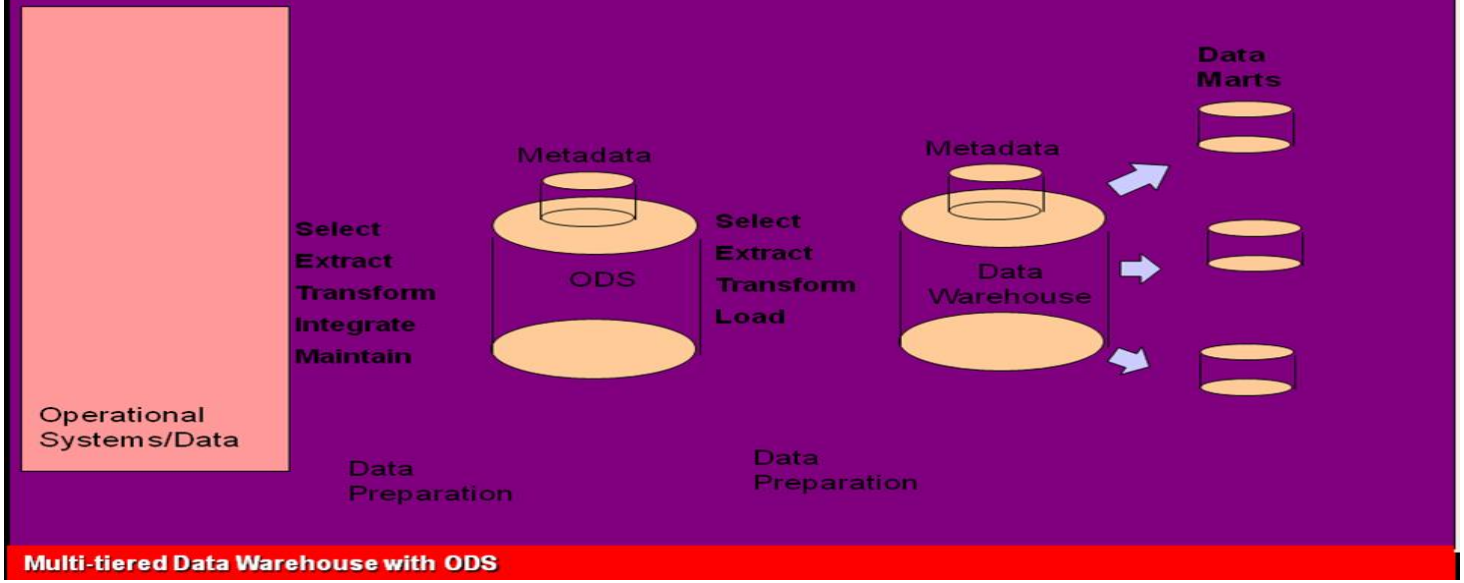
OLTP Vs ODS Vs DWH

Characteristic	OLTP	ODS	Data Warehouse
Data redundancy	Non-redundant within system; Unmanaged redundancy among systems	Somewhat redundant with operational databases	Managed redundancy
Data update	Field by field	Field by field	Controlled batch
Database size	Moderate	Moderate	Large to very large
Development Methodology	Requirements driven, structured	Data driven, somewhat evolutionary	Data driven, evolutionary
Philosophy	Support day-to-day operation	Support day-to-day decisions & operational activities	Support managing the enterprise

Typical Data Warehouse Architecture



MULTI-TIERED DATAWAREHOUSE WITH ODS



ETL- Extract, Transform and Load

As the name suggests, ETL process covers the following phases :

- Extraction of data from data sources.
- Transforming the extracted data to meet business requirements.
- Loading the data in to the target warehouse/database.

- Data Extract
  - ✓ Get Data from source
- Data Transformation
  - ✓ Data Cleansing - Data Quality Assurance
  - ✓ Data Scrubbing - Removing errors and inconsistencies
  - ✓ Processing Calculations
  - ✓ Applying Business Rules
  - ✓ Changing Data Types
  - ✓ Making the Data More Readable
  - ✓ Replacing Codes with Actual Values
  - ✓ Summarizing the Data
- Data Load
  - ✓ Load data into Warehouse

Extraction

- The first part of the ETL process.
- Data under consideration is being extracted from the different data sources.
- The source data may use a different data organization/format.
- Some of the common data sources are :
  - ✓ Databases
  - ✓ Flat files

Transform

- It involves applying a series of rules to the data extracted from the source to derive the data to load the target.
- Depending on the requirement of the target, the transformation rules may be simple or complex.
- Transformation may involve :
  - ✓ Selecting only certain columns to load
  - ✓ Filtering
  - ✓ Sorting
  - ✓ Combining data from multiple sources
  - ✓ Generating Surrogate keys, etc.

Load

- The last step of the ETL process
- The load phase loads the transformed data to the end target.
- Depending on the requirement, the load phase may be:
  - ✓ Full Load
  - ✓ Incremental Load

Importance of ETL

- Data of an organization spread across multiple geographies and domains.
- Data organized in different format in different sources.
- Consolidation of the data to make it more meaningful.
- Applying Business rules enriches the value the data provides.
- Identifying the inconsistencies and providing a unified view.
- Improving the data quality.

Sample ETL Tools

- Teradata Warehouse Builder from Teradata
- DataStage from IBM
- SAS System from SAS Institute
- Power Mart/Power Center from Informatica
- Sagent Solution from Sagent Software
- Hummingbird Genio Suite from Hummingbird Communications

INFORMTICA
OWB ODI
DATA STAGE
ABINITIO
CRYSTAL REPORTS
BI
COGNOS. BO
TERRADATA
SQLSERVER BI

Data Access and Analysis

- It is the process of timely access and analysis of data
- It is the means by which the End Users 'see' the data warehouse or the ODS or the Operational Systems

Data Access and Analysis - Terminologies

Reporting

- A category of data access solution in which the information is presented in the form of reports
- Reporting tools are also referred as Query and reporting tools

OLAP (On-Line Analytic Processing)

- Defined as "Fast Analysis of Multidimensional Information" by the OLAP council
- Used interchangeably with 'BI'
- OLAP tools are synonymous with Multidimensional tools or applications

DSS tools that use multidimensional data analysis techniques

- Support for a DSS data store
- Data extraction and integration filter
- Specialized presentation interface

Data Access and Analysis - Terminologies

Data Mining

- A process that uses a variety of statistical and artificial intelligence frameworks to discover patterns and relationships in data
- Used to make valid predictions in data analysis problems where the exact sequence and nature of queries/questions to be written/asked against the data to make the prediction is not known and the number of variables involved in the analysis is too large to be intuitively handled by structured querying or OLAP tools

Web Access

- A category of data access solutions in which information is viewed through a web browser

Importance of Data Access

Businesses today face challenges like

- Large volume of data
- User demands of flexible and timely access to information
- Extracting value from key business data

Data Access is the 'last mile' that enables decision makers to

- Reach the database infrastructure

Prompt, reliable data access

- Lowers operating costs
- Reduces error
- Increases productivity.

ONLINE ANALYICAL PROCESSING

Need for More Intensive Decision Support

4 Main Characteristics

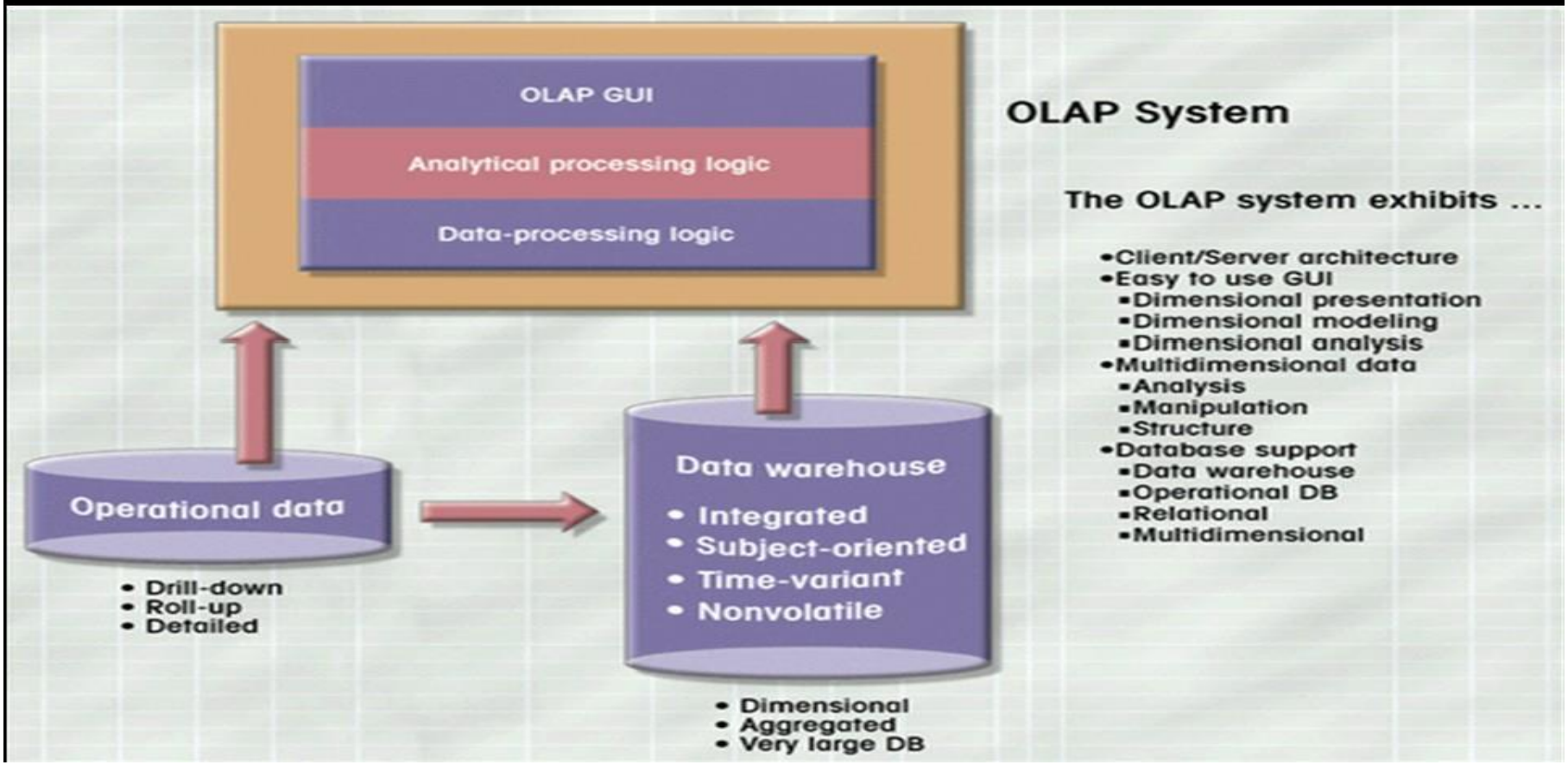
- Multidimensional data analysis
- Advanced Database Support
- Easy-to-use end-user interfaces
- Support Client/Server architecture

Multidimensional Data Analysis Techniques

Advanced Data Presentation Functions

- 3-D graphics, Pivot Tables, Crosstabs, etc.
- Compatible with Spreadsheets & Statistical packages
- Advanced data aggregations, consolidation and classification across time dimensions
- Advanced computational functions
- Advanced data modeling functions

OLAP Client/Server Architecture

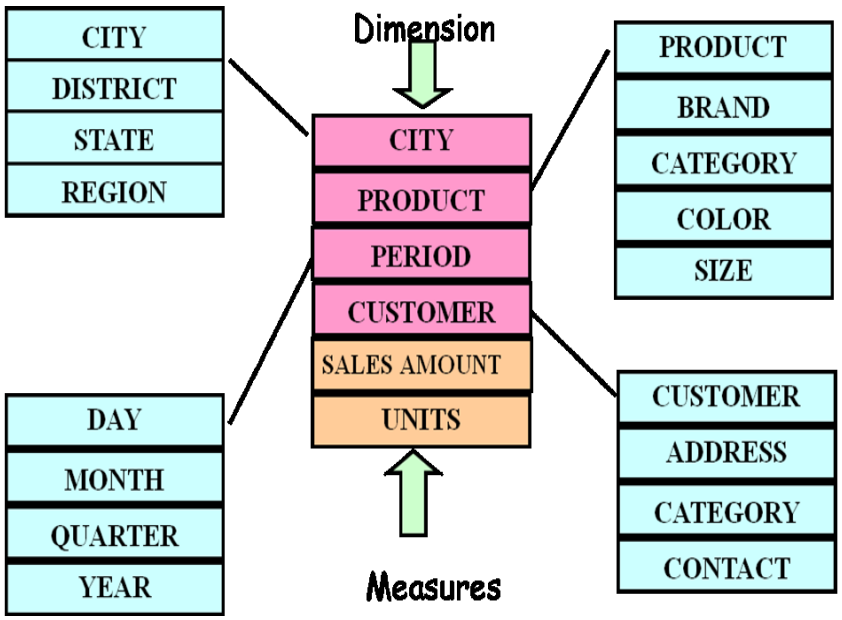


Relational OLAP

- **Relational Online Analytical Processing**
  - ✓ **OLAP functionality using relational database and familiar query tools to store and analyze multidimensional data**
- **Multidimensional data schema support**
- **Data access language & query performance for multidimensional data**
- **Support for Very Large Databases**

Star Schema

- The fact table is always the largest table in the star schema
- Each dimension record is related to thousand of fact records
- Star Schema facilitated data retrieval functions
- DBMS first searches the Dimension Tables before the larger fact table



Dimensional Modeling - Basic Concepts

- Represents the data in a standard, intuitive framework that allows for high-performance access;
- Schema designed to process large, complex, adhoc and data intensive queries.
- No concern for concurrency, locking and insert/update/delete performance
- Every dimensional model is composed of one table with a multipart key, called the fact table, and a set of smaller tables called dimension tables.
- This characteristic "star-like" structure is often called a star join.

Star Schema Representation

- Fact and Dimensions are represented by physical tables in the data warehouse database
- Fact tables are related to each dimension table in a Many to One relationship (Primary/Foreign Key Relationships)
- Fact Table is related to many dimension tables
  - ✓ The primary key of the fact table is a composite primary key from the dimension tables
- Each fact table is designed to answer a specific DSS question

Fact Tables

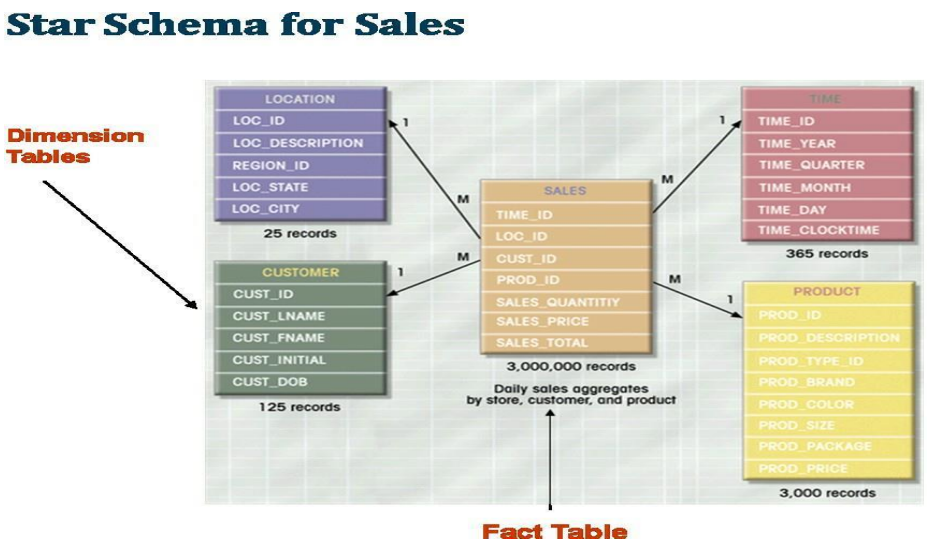
- The most useful facts in a fact table are numeric and additive
- Typically represents a business transaction, or event that can be used in analyzing business process
- By nature fact tables are sparse
- Usually very large - billions of records

Dimension Tables

- Each dimension table has a single-part primary key that corresponds exactly to one of the components of the multipart key in the fact table.
- Dimension tables, most often contain descriptive textual information
- Determine contextual background for facts
- Examples :
  - Time
  - Location/Region
  - Customers

Measures

- A numeric attribute of a fact
- Represents performance or behavior of the business relative to the dimensions
- The actual numbers are called variables
- Occupy very little space compared to Fact Tables
- Examples :
  - Quantity supplied
  - Transaction amount
  - Sales volume



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
- 2000 onwards
  - Big Data, Semi-structured, Unstructured data, data streams, etc

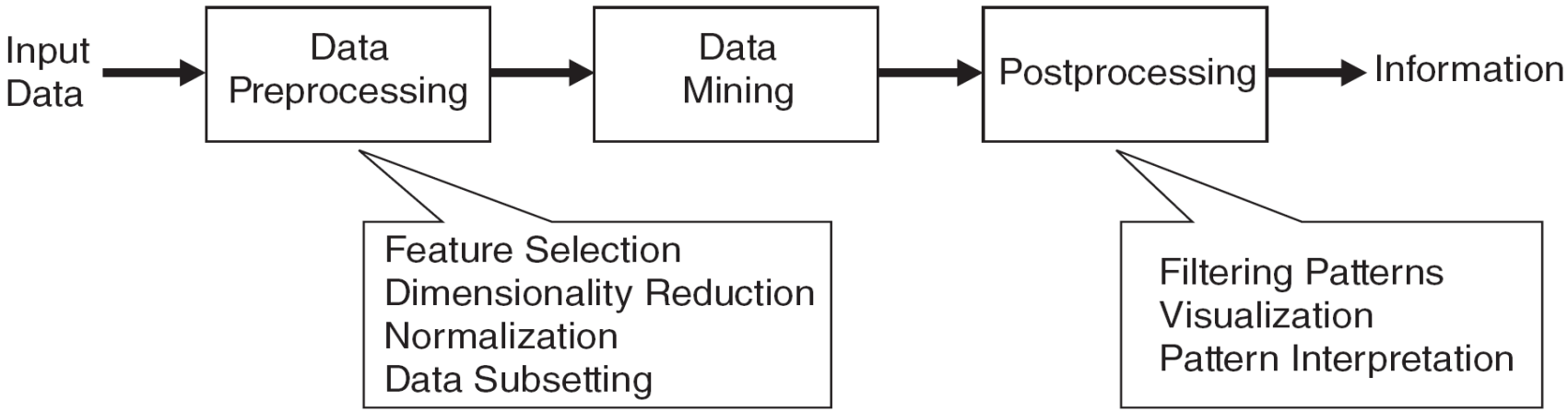
What Is Data Mining?

- Data mining (knowledge discovery in databases):
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from data in large databases
- 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?
  - (Deductive) query processing.
  - Expert systems or small ML/statistical programs

What is Data Mining? …. Many Definitions

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover
  meaningful patterns
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

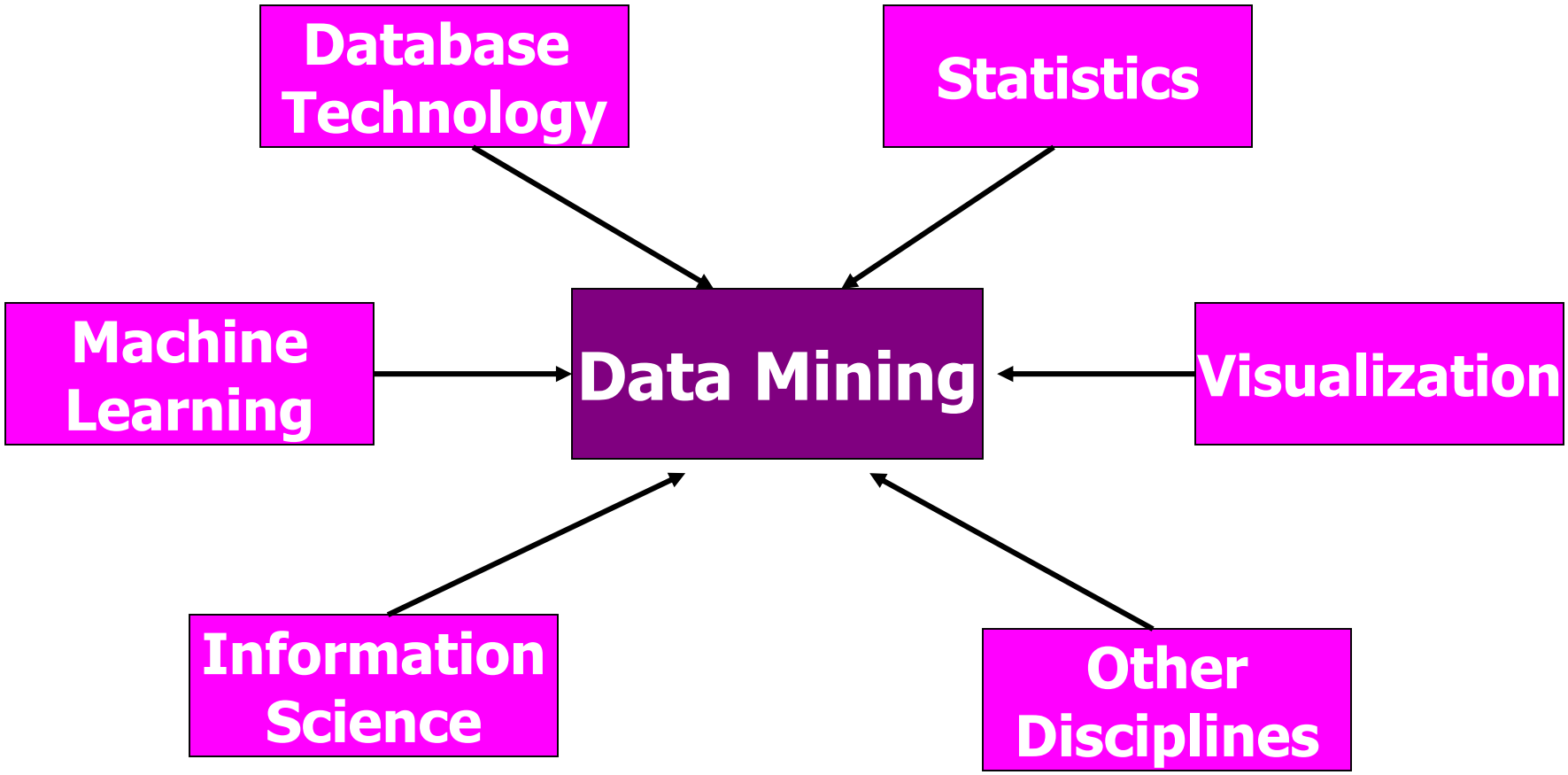
Watch out: Is everything "data mining"?



What is (not) Data Mining?

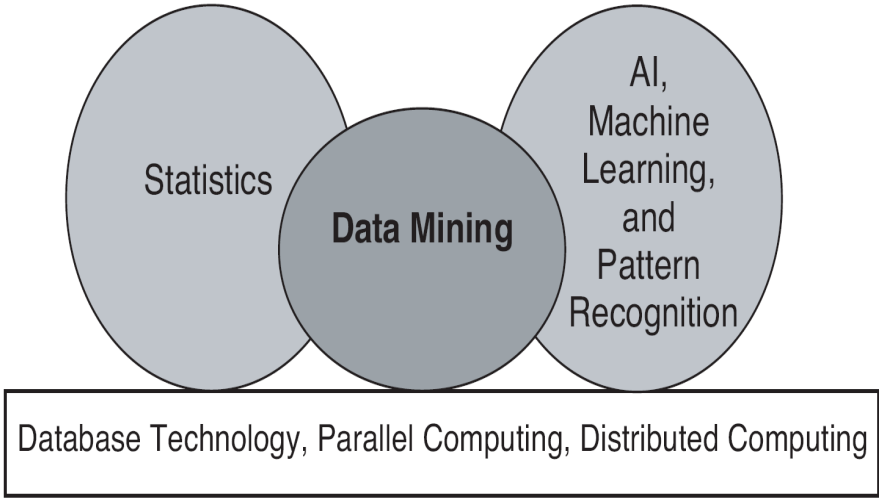
- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”

- **What is Data Mining?**
  - Certain names are more prevalent in certain US locations (O’Brien, O’Rourke, O’Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g., Amazon rainforest, Amazon.com)

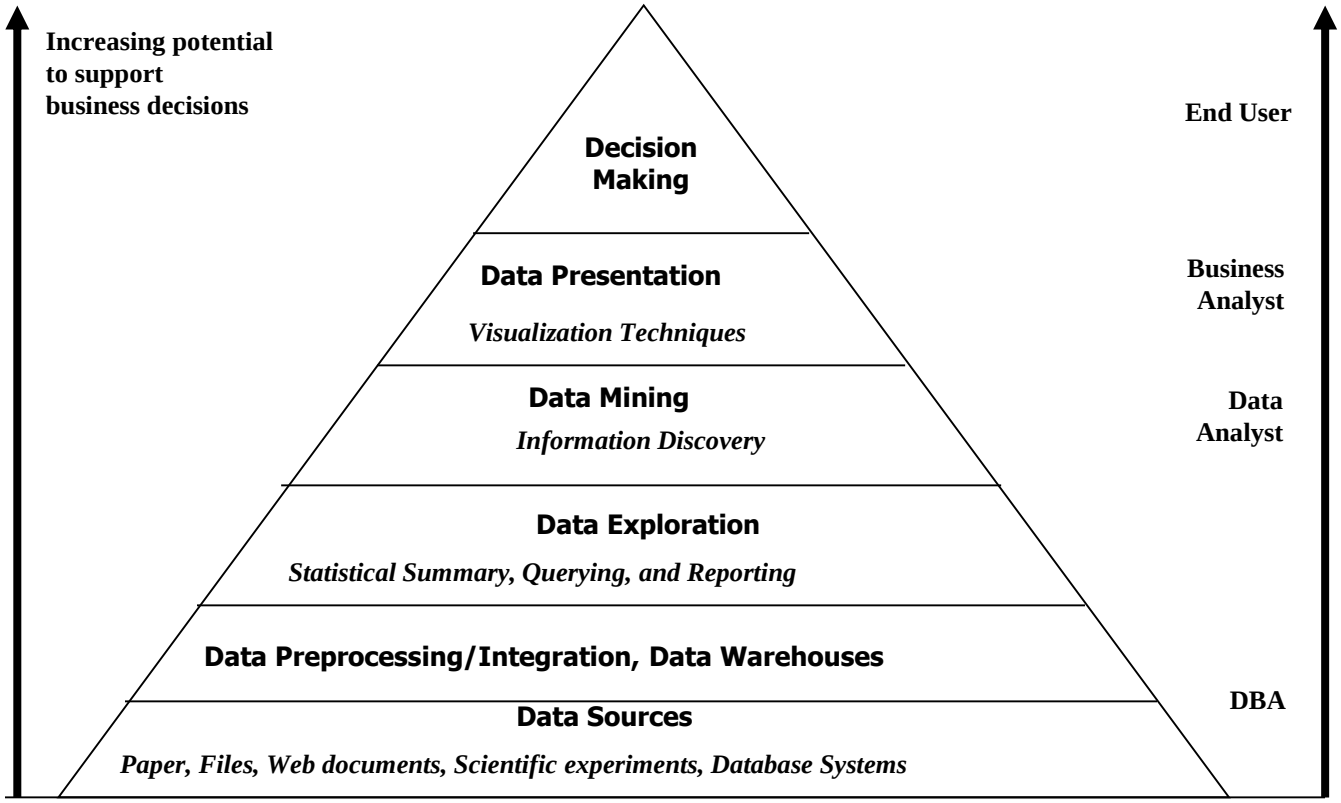

Database Technology
Statistics
Machine Learning
Data Mining
Visualization
Information Science
Other Disciplines

Origins of Data Mining

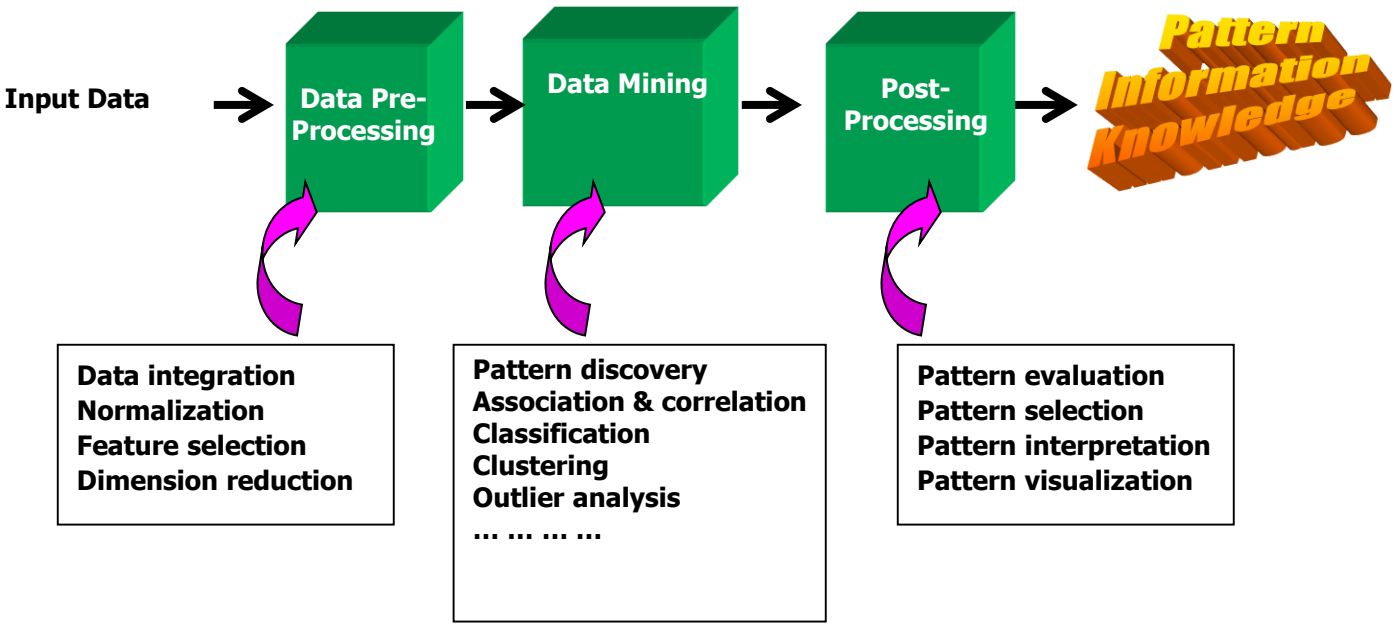
- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



Data Mining in Business Intelligence

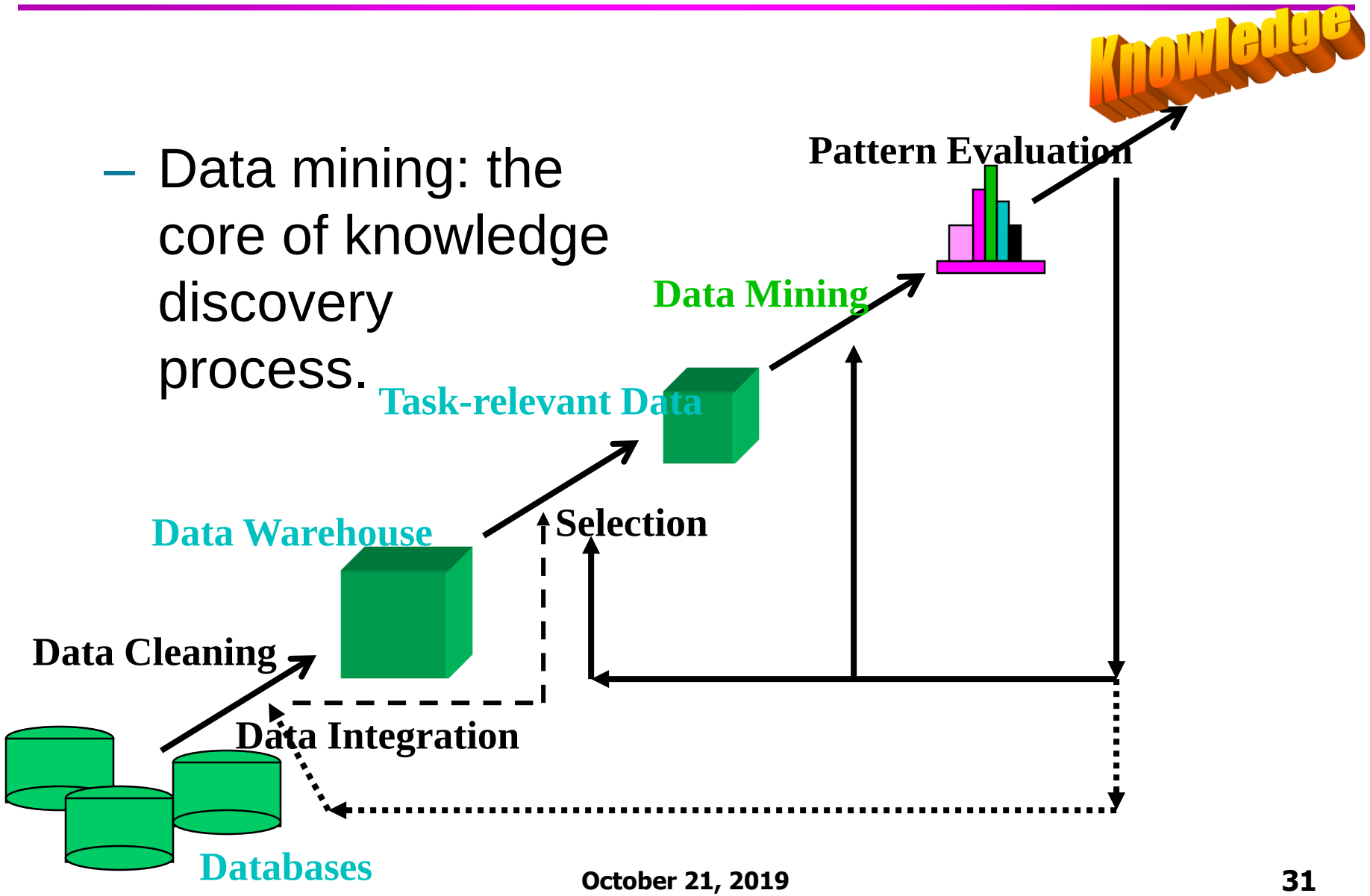


Data Mining/KDD Process



Data Mining: A KDD Process

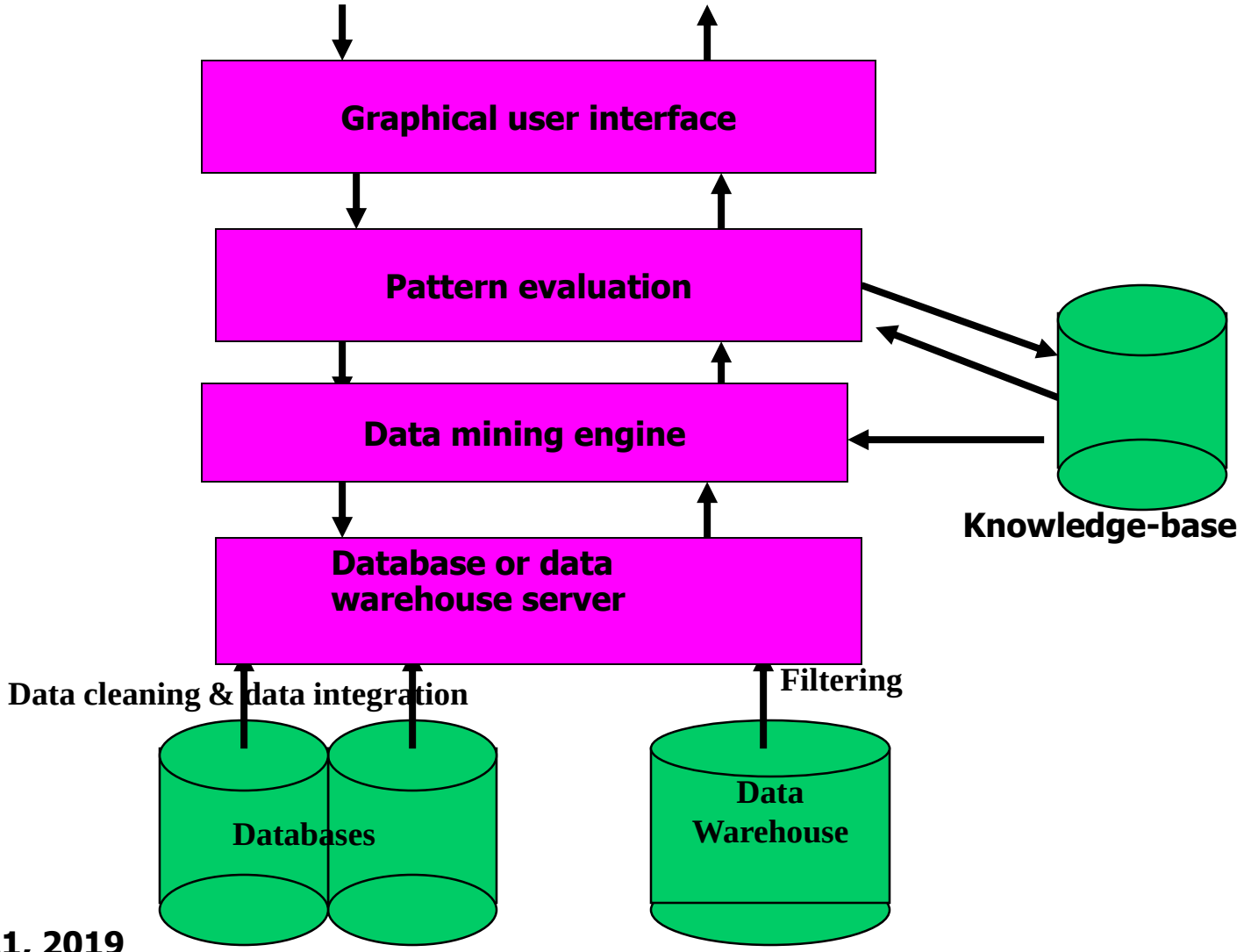
- Data mining: the core of knowledge discovery process.



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

Architecture of a Typical Data Mining System



Applications… and

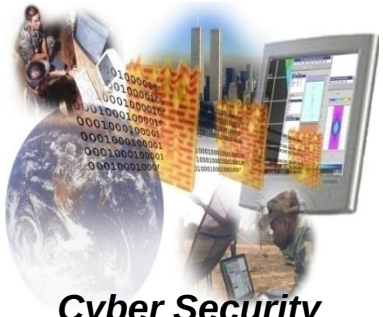
Opportunities……

Why Data Mining?

- **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

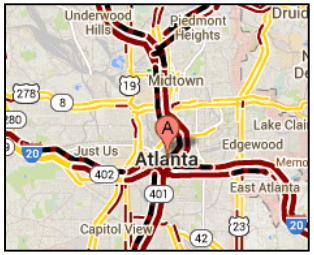
Large-scale Data is Everywhere!

- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

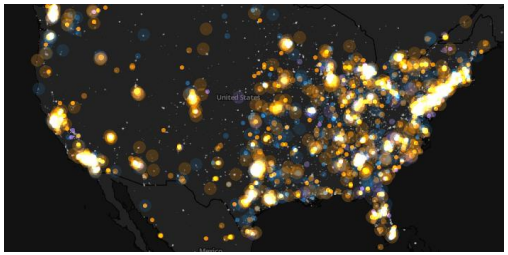


Cyber Security

E-Commerce

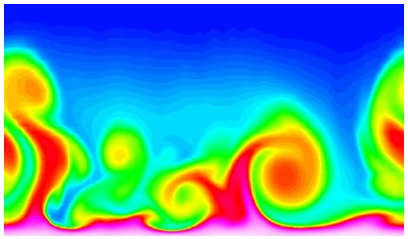


Traffic Patterns



Social Networking: Twitter

Sensor Networks



Computational Simulations

Why Data Mining – simple example/case study

A search engine (e.g., Google) receives hundreds of millions of queries every day. Each query can be viewed as a transaction where the user describes her or his information need.

What novel and useful knowledge can a search engine learn from such a huge collection of queries collected from users over time? Some patterns found in user search queries can disclose invaluable knowledge that cannot be obtained by reading individual data items alone.

For example, Google's *Flu Trends* uses specific search terms as indicators of flu activity. It found a close relationship between the number of people who search for flu-related information and the number of people who actually have flu symptoms.

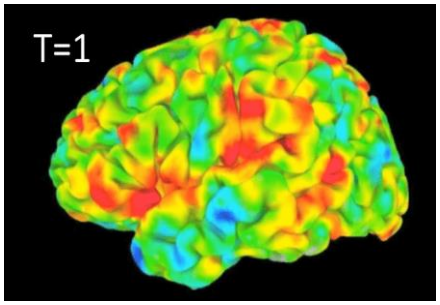
A pattern emerges when all of the search queries related to flu are aggregated. Using aggregated Google search data, *Flu Trends* can estimate flu activity up to two weeks faster than traditional systems can. This example shows how data mining can turn a large collection of data into knowledge that can help meet a current global challenge.

Why Data Mining? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data
    - Yahoo has Peta Bytes of web data
    - Facebook has billions of active users
  - purchases at department/ grocery stores, e-commerce
    - Amazon handles millions of visits/day
  - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- **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, YouTube**
- **We are drowning in data, but starving for knowledge!**
- **Competitive Pressure is Strong**
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

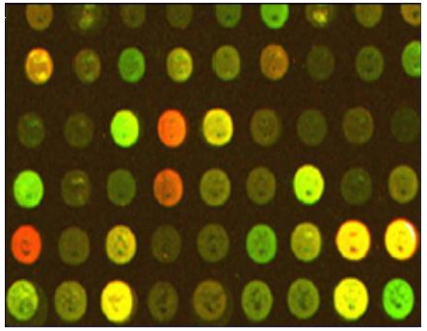
Why Data Mining? Scientific Viewpoint

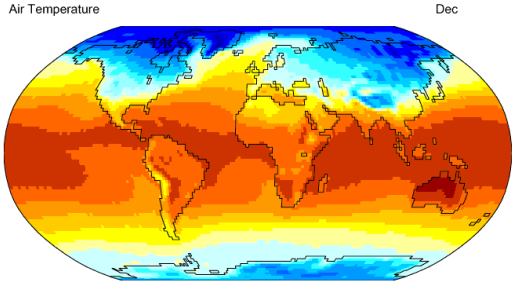
- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours



fMRI Data from Brain


Sky Survey Data


Gene Expression Data



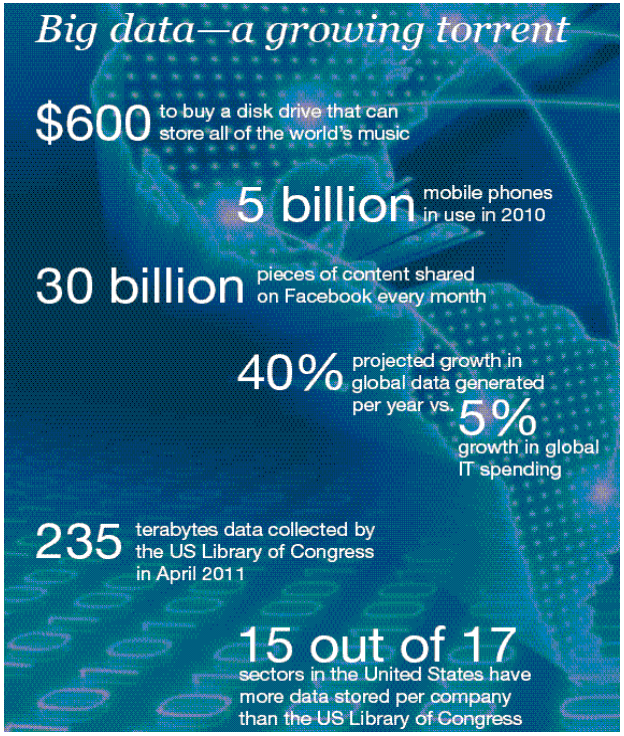
Surface Temperature of Earth

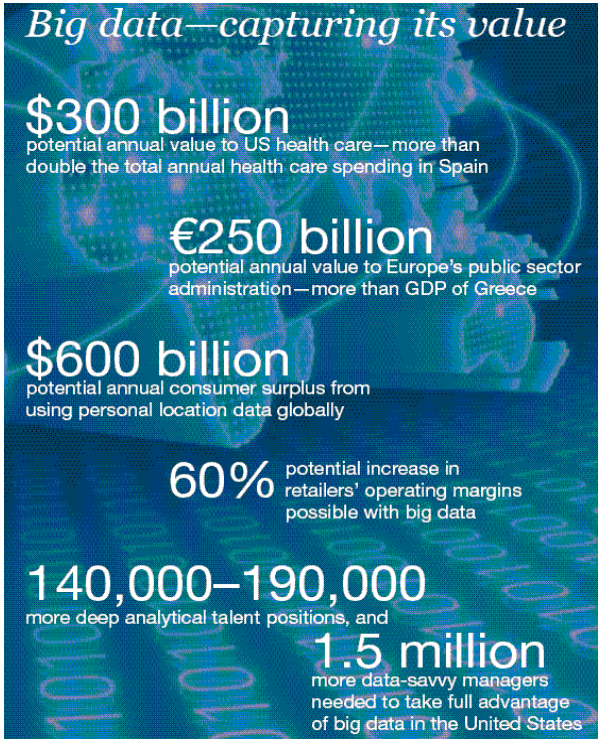
- Data mining helps scientists
  - in automated analysis of massive datasets
  - In hypothesis formation

Great opportunities to improve productivity in all walks of life

McKinsey Global Institute

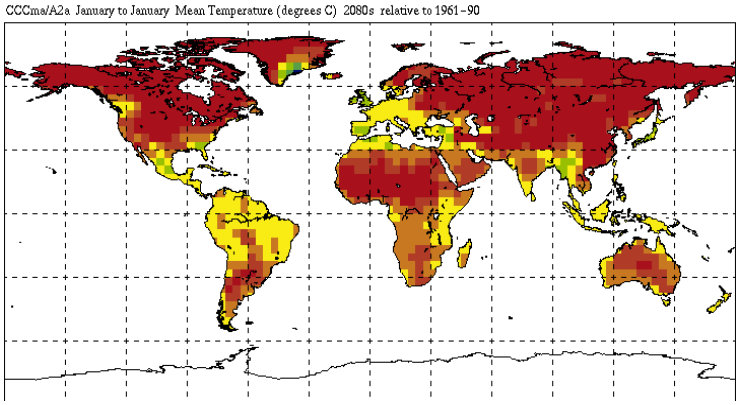
Big data: The next frontier for innovation, competition, and productivity





Great Opportunities to Solve Society's Major Problems

Improving health care and reducing costs



Predicting the impact of climate change

Finding alternative/ green energy sources

Reducing hunger and poverty by increasing agriculture production

Market Analysis and Management (1)

- 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 (2)

- 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 (1)

- 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 (2)

- 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: 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*
  - *Others…*

Data Mining Functionalities (1)

- **Concept description: Characterization and discrimination**
  - Generalize, summarize, and contrast data characteristics, e.g., dry vs. wet regions

- **Association (correlation and causality)**
  - Multi-dimensional vs. single-dimensional association
  - age(X, “20..29”) ^ income(X, “20..29K”) → buys(X, “PC”) [support = 2%, confidence = 60%]
  - contains(T, “computer”) → contains(x, “software”) [1%, 75%]

Data Mining Functionalities (2)

- **Classification and Prediction**
  - Finding 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
  - Presentation: decision-tree, classification rule, neural network
  - Prediction: Predict some unknown or missing numerical values
- **Cluster analysis**
  - Class label is unknown: Group data to form new classes, e.g., cluster houses to find distribution patterns
  - Clustering based on the principle: maximizing the intra-class similarity and minimizing the interclass similarity

Data Mining Functionalities (3)

- **Outlier analysis**
  - Outlier: a data object that does not comply with the general behavior of the data
  - It can be considered as noise or exception but is quite useful in fraud detection, rare events analysis
- **Trend and evolution analysis**
  - Trend and deviation: regression analysis
  - Sequential pattern mining, periodicity analysis
  - Similarity-based analysis
- **Other pattern-directed or statistical analysis**

Are All the "Discovered" Patterns Interesting?

- A data mining system/query 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.

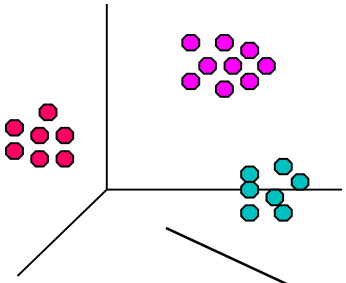
Can We Find All and Only Interesting Patterns?

- Find all the interesting patterns: Completeness
  - Can a data mining system find all the interesting patterns?
  - Association vs. classification vs. clustering
- Search for only interesting patterns: Optimization
  - Can a data mining system find only the interesting patterns?
  - Approaches
    - First general all the patterns and then filter out the uninteresting ones.
    - Generate only the interesting patterns—mining query optimization

Data Mining: Classification Schemes or Types

- General functionality
  - Descriptive data mining
  - Predictive data mining
- Different views, different classifications
  - Kinds of databases to be mined
  - Kinds of knowledge to be discovered
  - Kinds of techniques utilized
  - Kinds of applications adapted

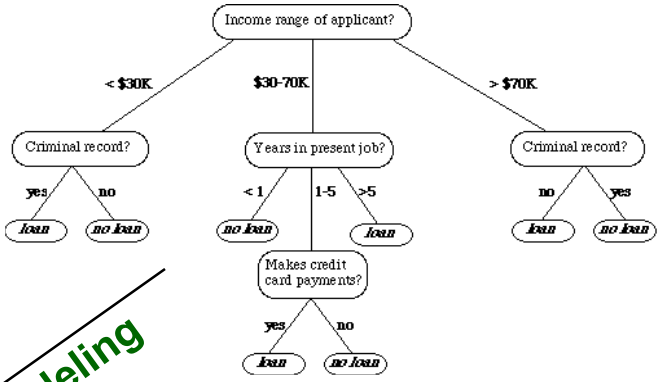
Data Mining Tasks ...



Clustering

Data

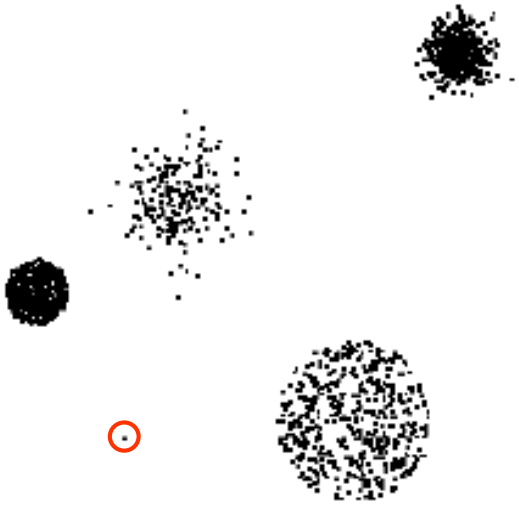
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes
11	No	Married	60K	No
12	Yes	Divorced	220K	No
13	No	Single	85K	Yes
14	No	Married	75K	No
15	No	Single	90K	Yes



Predictive Modeling

Association Rules

Anomaly Detection



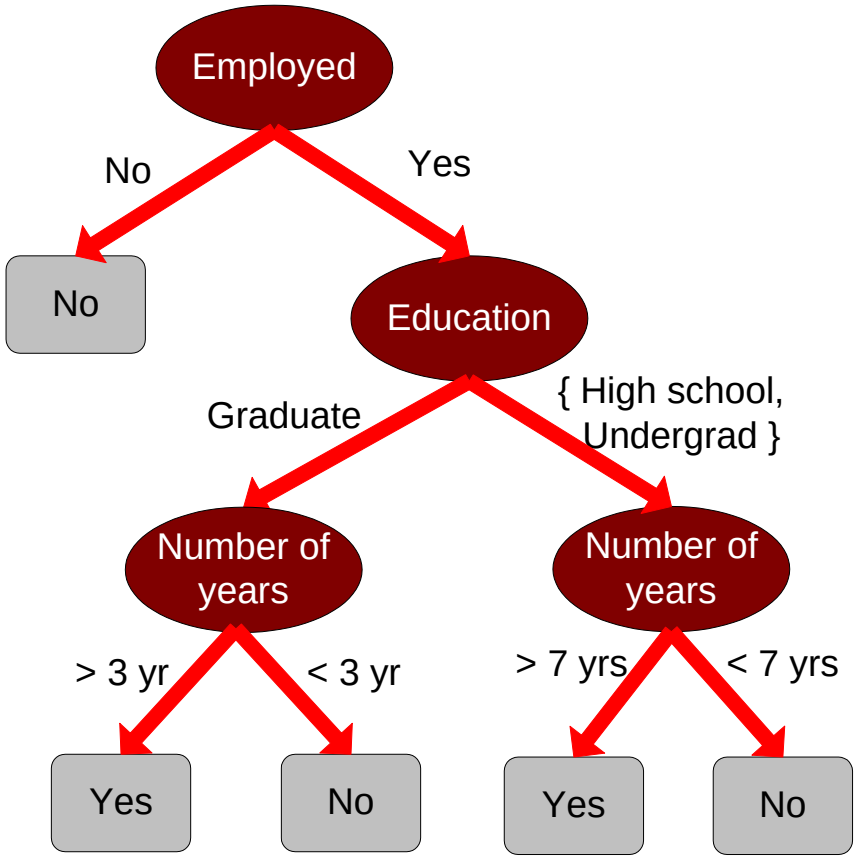
Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes

Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Model for predicting credit worthiness

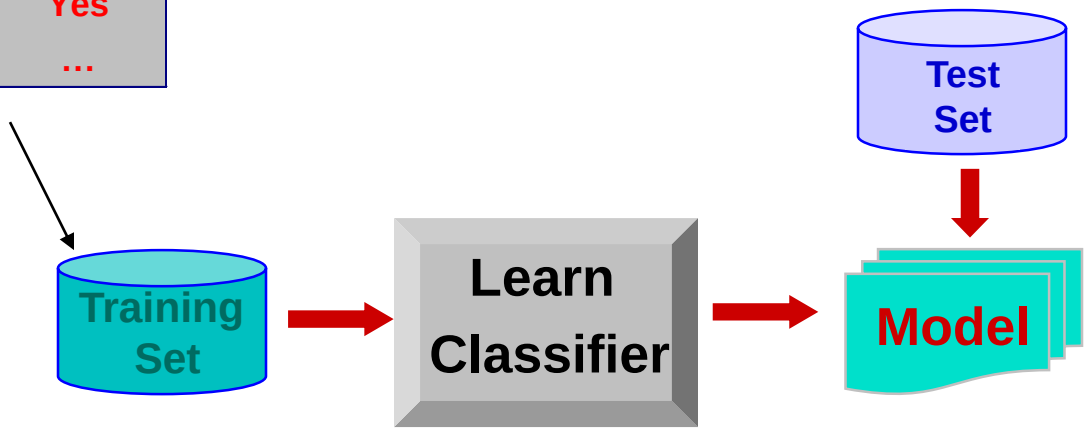


Classification Example

categorical categorical quantitative class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…

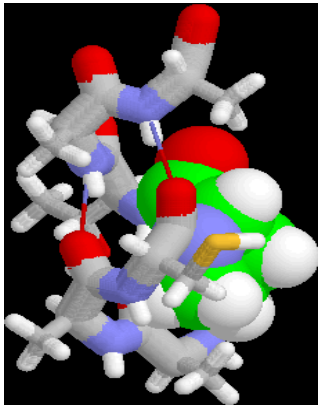


Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent

- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data

- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant



- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil

Classification: Application 1

- 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.

Classification: Application 2

- Churn prediction for telephone customers
  - **Goal:** To predict whether a customer is likely to be lost to a competitor.
  - **Approach:**
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

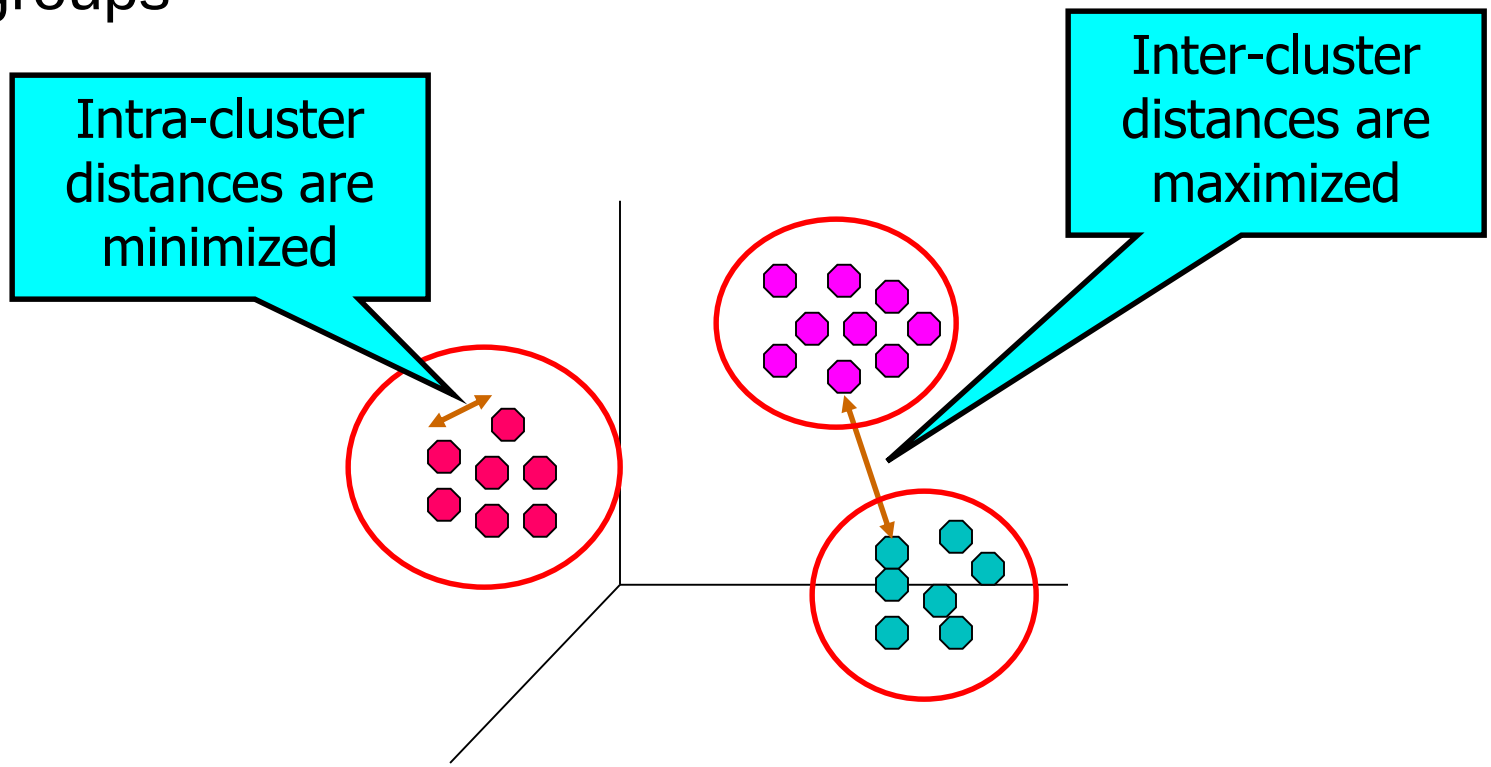
From [Berry & Linoff] Data Mining Techniques, 1997

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.
- Extensively studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



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.

TID	*Items*
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}

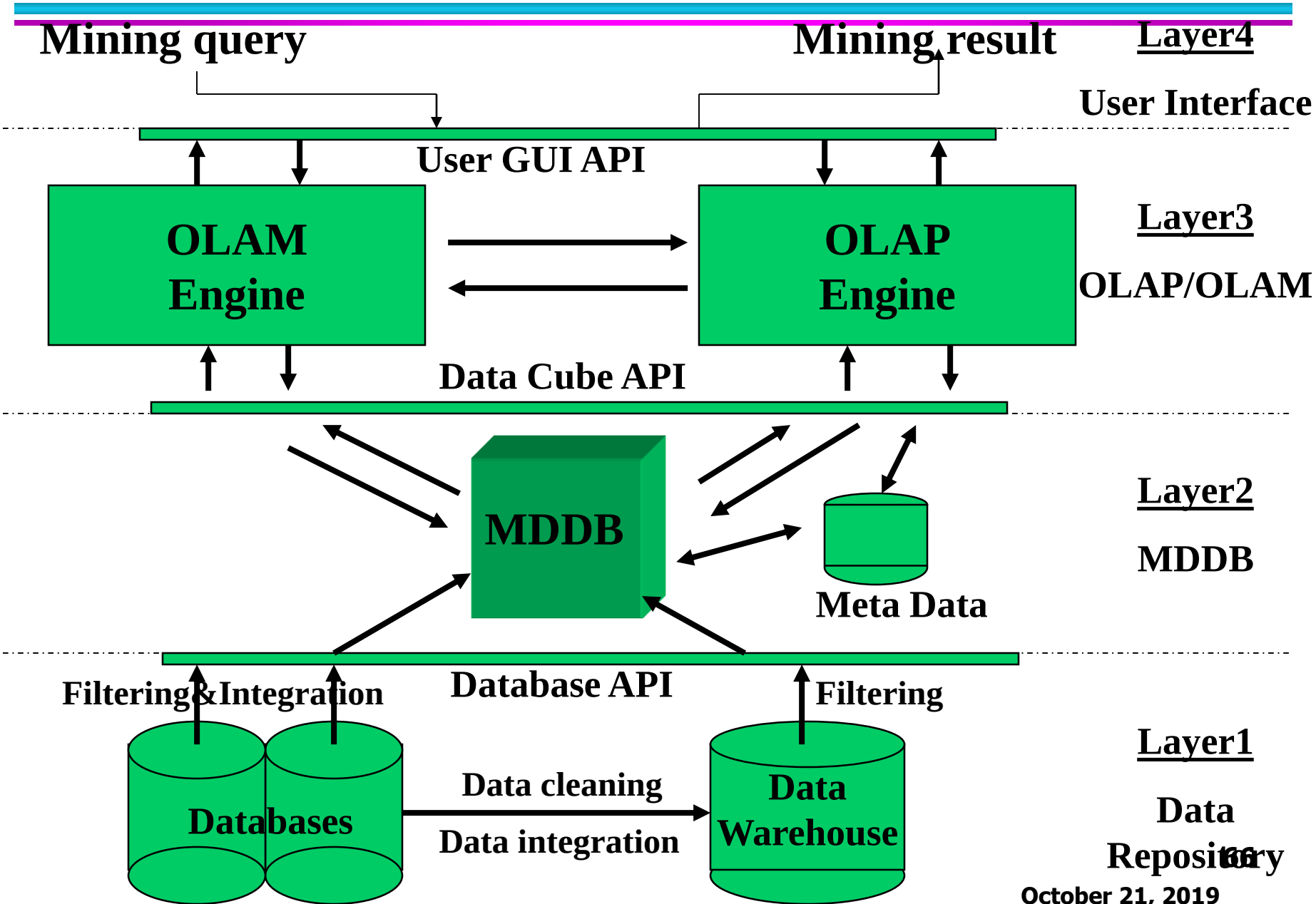
A Multi-Dimensional View of Data Mining Classification

- **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.

OLAP Mining: An Integration of Data Mining and Data Warehousing

- **Data mining systems, DBMS, Data warehouse systems coupling**
  - No coupling, loose-coupling, semi-tight-coupling, tight-coupling
- **On-line analytical mining data**
  - integration of mining and OLAP technologies
- **Interactive mining multi-level knowledge**
  - Necessity of mining knowledge and patterns at different levels of abstraction by drilling/rolling, pivoting, slicing/dicing, etc.
- **Integration of multiple mining functions**
  - Characterized classification, first clustering and then association

An OLAM Architecture



Major Issues in Data Mining (1)

- 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 Mining (2)

- 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

Data Mining & Machine Learning

According to Tom M. Mitchell, Chair of Machine Learning at Carnegie Mellon University and author of the book *Machine Learning* (McGraw-Hill),

> A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with the experience E.
>
> We now have a set of objects to define machine learning:
>
> > Task (T), Experience (E), and Performance (P)
>
> With a computer running a set of tasks, the experience should be leading to performance increases (to satisfy the definition)

Many data mining tasks are executed successfully with help of machine learning

Prior Knowledge

- Data Mining tools/solutions identify hidden patterns.
  - Generally we get many patterns
  - Out of them many could be false or trivial.
  - Filtering false patterns requires domain understanding.
- Understanding how the data is collected, stored, transformed, reported, and used is essential.
- Causation vs. Correlation
  - A bank may decide interest rate based on credit score. Looking at data, credit score moves as per interest rate. It does not make sense to derive credit score from interest rate.

DM Issues/Challenges – Mining Methodology

- **Mining various and new kinds of knowledge**: Data mining covers a wide spectrum of data analysis and knowledge discovery tasks, from data characterization and discrimination to association and correlation analysis, classification, regression, clustering, outlier analysis, sequence analysis, and trend and evolution analysis.

- **Mining knowledge in multidimensional space**: When searching for knowledge in large data sets, we can explore the data in multidimensional space. That is, we can search for interesting patterns among combinations of dimensions (attributes) at varying levels of abstraction. Data can be aggregated or viewed as a multidimensional data cube.

- **Data mining—an interdisciplinary effort**: For example, to mine data with natural language text, it makes sense to fuse data mining methods with methods of information retrieval and natural language processing, e.g. consider the mining of software bugs in large programs, known as bug mining, benefits from the incorporation of software engineering knowledge into the data mining process.

DM Issues/Challenges – Mining Methodology

- **Boosting the power of discovery in a networked environment**: Most data objects reside in a linked or interconnected environment, whether it be the Web, database relations, files, or documents. Semantic links across multiple data objects can be used to advantage in data mining.

- **Handling uncertainty, noise, or incompleteness of data**: Data often contain noise, errors, exceptions, or uncertainty, or are incomplete. Errors and noise may confuse the data mining process, leading to the derivation of erroneous patterns. Data cleaning, data preprocessing, outlier detection and removal, and uncertainty reasoning are examples of techniques that need to be integrated with the data mining process.

- **Pattern evaluation and pattern- or constraint-guided mining**: What makes a pattern interesting may vary from user to user. Therefore, techniques are needed to assess the interestingness of discovered patterns based on subjective measures. These estimate the value of patterns with respect to a given user class, based on user beliefs or expectations.

DM Issues/Challenges – User Interaction

The user plays an important role in the data mining process. Interesting areas include how to interact with a data mining system, how to incorporate a user's background knowledge in mining, and how to visualize and comprehend data mining results.

- **Interactive mining**: The data mining process should be highly interactive. Thus, it is important to build flexible user interfaces and an exploratory mining environment, facilitating the user's interaction with the system. A user may like to first sample a set of data, explore general characteristics of the data, and estimate potential mining results. Interactive mining should allow users to dynamically change the focus of a search, to refine mining requests based on returned results, and to drill, dice, and pivot through the data and knowledge space interactively, dynamically exploring "cube space" while mining.

- **Incorporation of background knowledge**: Background knowledge, constraints, rules, and other information regarding the domain under study should be incorporated into the knowledge discovery process. Such knowledge can be used for pattern evaluation as well as to guide the search toward interesting patterns.

DM Issues/Challenges – User Interaction

- **Ad hoc data mining and data mining query languages**: Query languages (e.g., SQL) have played an important role in flexible searching because they allow users to pose ad hoc queries. Similarly, high-level data mining query languages or other high-level flexible user interfaces will give users the freedom to define ad hoc data mining tasks. This should facilitate specification of the relevant sets of data for analysis, the domain knowledge, the kinds of knowledge to be mined, and the conditions and constraints to be enforced on the discovered patterns. Optimization of the processing of such flexible mining requests is another promising area of study.

- **Presentation and visualization of data mining results**: How can a data mining system present data mining results, vividly and flexibly, so that the discovered knowledge can be easily understood and directly usable by humans? This is especially crucial if the data mining process is interactive. It requires the system to adopt expressive knowledge representations, user-friendly interfaces, and visualization techniques.

DM Issues/Challenges - Efficiency and Scalability

Efficiency and scalability are always considered when comparing data mining algorithms. As data amounts continue to multiply, these two factors are especially critical.

- **Efficiency and scalability of data mining algorithms**:

Data mining algorithms must be efficient and scalable in order to effectively extract information from huge amounts of data in many data repositories or in dynamic data streams. In other words, the running time of a data mining algorithm must be predictable, short, and acceptable by applications. Efficiency, scalability, performance, optimization, and the ability to execute in real time are key criteria that drive the development of many new data mining algorithms.

DM Issues/Challenges - Efficiency and Scalability

- **Parallel, distributed, and incremental mining algorithms**: The humongous size of many data sets, the wide distribution of data, and the computational complexity of some data mining methods are factors that motivate the development of parallel and distributed data-intensive mining algorithms. Such algorithms first partition the data into "pieces." Each piece is processed, in parallel, by searching for patterns. The parallel processes may interact with one another. The patterns from each partition are eventually merged.

- **Cloud computing and cluster computing**, which use computers in a distributed and collaborative way to tackle very large-scale computational tasks, are also active research themes in parallel data mining. In addition, the high cost of some data mining processes and the incremental nature of input promote incremental data mining, which incorporates new data updates without having to mine the entire data "from scratch." Such methods perform knowledge modification incrementally to amend and strengthen what was previously discovered.

DM Issues/Challenges - Diversity of Database Types

The wide diversity of database types brings about challenges to data mining.

Handling complex types of data: Diverse applications generate a wide spectrum of new data types, from structured data such as relational and data warehouse data to semi-structured and unstructured data; from stable data repositories to dynamic data streams; from simple data objects to temporal data, biological sequences, sensor data, spatial data, hypertext data, multimedia data, software program code, Web data, and social network data. It is unrealistic to expect one data mining system to mine all kinds of data, given the diversity of data types and the different goals of data mining. Domain- or application-dedicated data mining systems are being constructed for in-depth mining of specific kinds of data. The construction of effective and efficient data mining tools for diverse applications remains a challenging area.

Mining dynamic, networked, and global data repositories: Multiple sources of data are connected by the Internet and various kinds of networks, forming gigantic, distributed, and heterogeneous global information systems and networks. The discovery of knowledge from different sources of structured, semi-structured, or unstructured yet interconnected data with diverse data semantics poses great challenges to data mining. Mining such gigantic, interconnected information networks may help disclose many more patterns and knowledge in heterogeneous data sets than can be discovered from a small set of isolated data repositories. Web mining, multisource data mining, and information network mining have become challenging and fast-evolving data mining fields.

DM Issues/Challenges - Society

How does data mining impact society? What steps can data mining take to preserve the privacy of individuals? Do we use data mining in our daily lives without even knowing that we do?

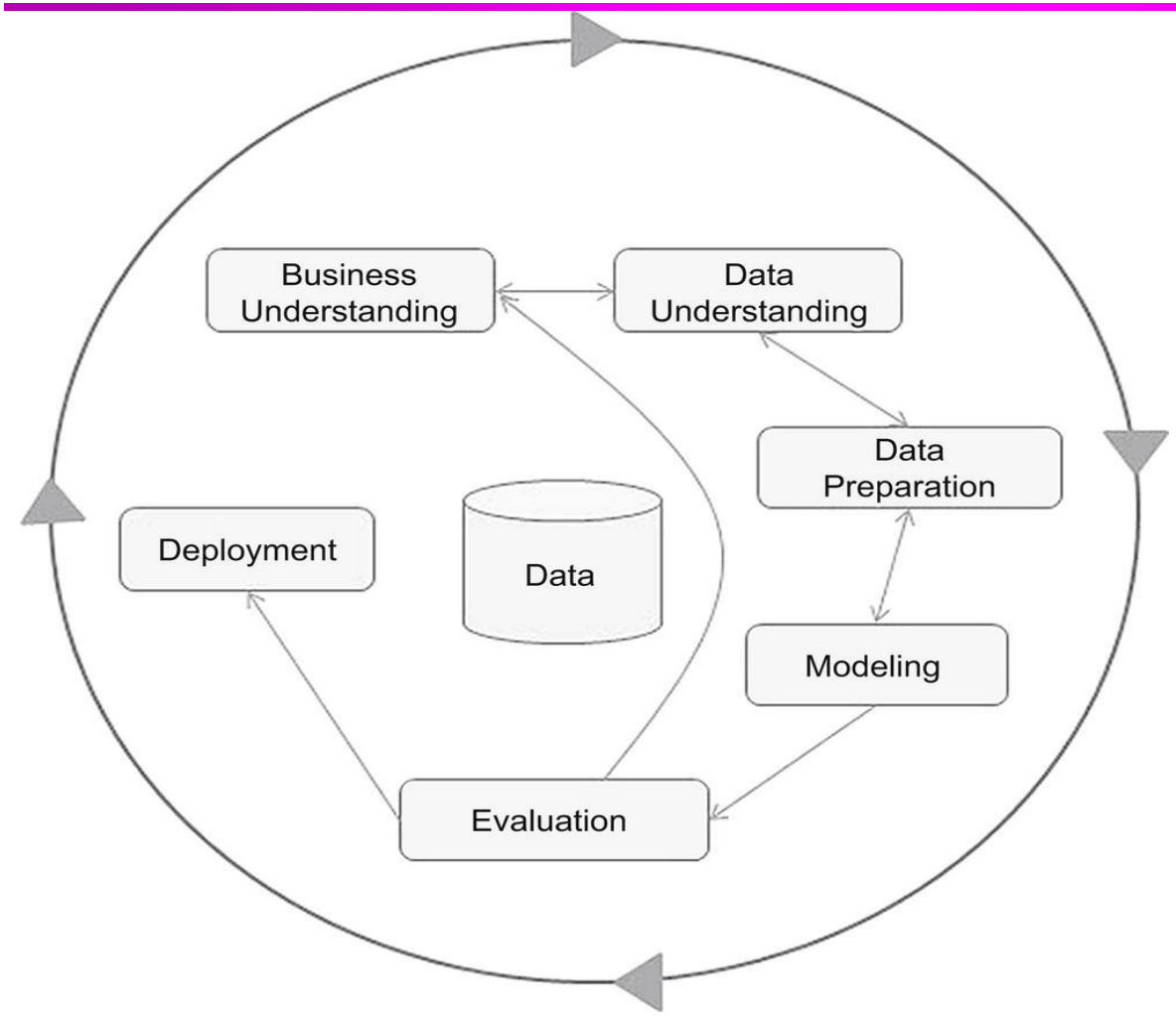
Social impacts of data mining: With data mining penetrating our everyday lives, it is important to study the impact of data mining on society. How can we use data mining technology to benefit society? How can we guard against its misuse? The improper disclosure or use of data and the potential violation of individual privacy and data protection rights are areas of concern that need to be addressed.

DM Issues/Challenges - Society

Privacy-preserving data mining: Data mining will help scientific discovery, business management, economy recovery, and security protection (e.g., the real-time discovery of intruders and cyberattacks). However, it poses the risk of disclosing an individual's personal information. Studies on privacy-preserving data publishing and data mining are ongoing. The philosophy is to observe data sensitivity and preserve people's privacy while performing successful data mining.

Invisible data mining: We cannot expect everyone in society to learn and master data mining techniques. More and more systems should have data mining functions built within so that people can perform data mining or use data mining results simply by mouse clicking, without any knowledge of data mining algorithms. Intelligent search engines and Internet-based stores perform such invisible data mining by incorporating data mining into their components to improve their functionality and performance. This is done often unbeknownst to the user. For example, when purchasing items online, users may be unaware that the store is likely collecting data on the buying patterns of its customers, which may be used to recommend other items for purchase in the future.

CRISP data mining framework



CRISP is the most popular methodology for analytics, data mining, and data science projects, with 43% share as per 2014 KDnuggets Poll.

CRISP-DM was conceived in 1996. In 1997 it got underway as a European Union project, led by SPSS, Teradata, Daimler AG, NCR Corporation and OHRA.

Data Quality, Preparation & Data Preprocessing…. More in Next class….

- Data needs to be understood. It requires descriptive statistics such as mean, median, mode, standard deviation, and range for each attribute
- Data quality is an ongoing concern wherever data is collected, processed, and stored.
  - The data cleansing practices include elimination of duplicate records, quarantining outlier records that exceed the bounds, standardization of attribute values, substitution of missing values, etc.
  - it is critical to check the data using data exploration techniques in addition to using prior knowledge of the data and business before building models to ensure a certain degree of data quality
- Missing Values
  - Need to track the data lineage of the data source to find right solution
- Data Types and Conversion
  - The attributes in a data set can be of different types, such as continuous numeric (interest rate), integer numeric (credit score), or categorical
  - data mining algorithms impose different restrictions on what data types they accept as inputs
- Transformation
  - Can go beyond type conversion, may include dimensionality reduction or numerosity reduction
- Outliers are anomalies in the data set
  - May occur legitimately or erroneously.
- Feature Selection
  - Many data mining problems involve a data set with hundreds to thousands of attributes, most of which may not be helpful. Some attributes may be correlated, e.g. sales amount and tax.
- Data Sampling may be adequate in many cases