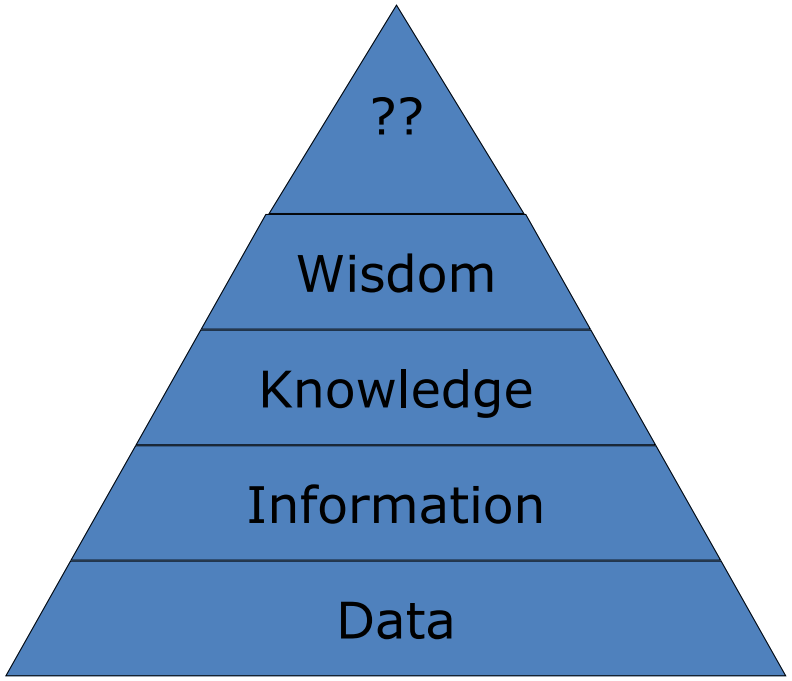
Unit 1: Introduction

- Evolution of data-ware house
- Concept of Data Mining and Data warehousing, client server and Data Warehouse
- Benefit of data mining.
- Data warehousing, Mining , Architecture.
- KDD Process
- Application of datamining

What is Data?

- A representation of facts, concepts, or instructions in a formal manner suitable for communication, interpretation, or processing by human beings or by computers.



Review of basic concepts of data warehousing and data mining

- The Explosive Growth of Data: from terabytes to petabytes
- Data accumulate and double every 9 months
- High-dimensionality of data
- High complexity of data
- New and sophisticated applications
- There is a big gap from stored data to knowledge; and the transition won’t occur automatically.
- Manual data analysis is not new but a bottleneck
- Fast developing Computer Science and Engineering generates new demands

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

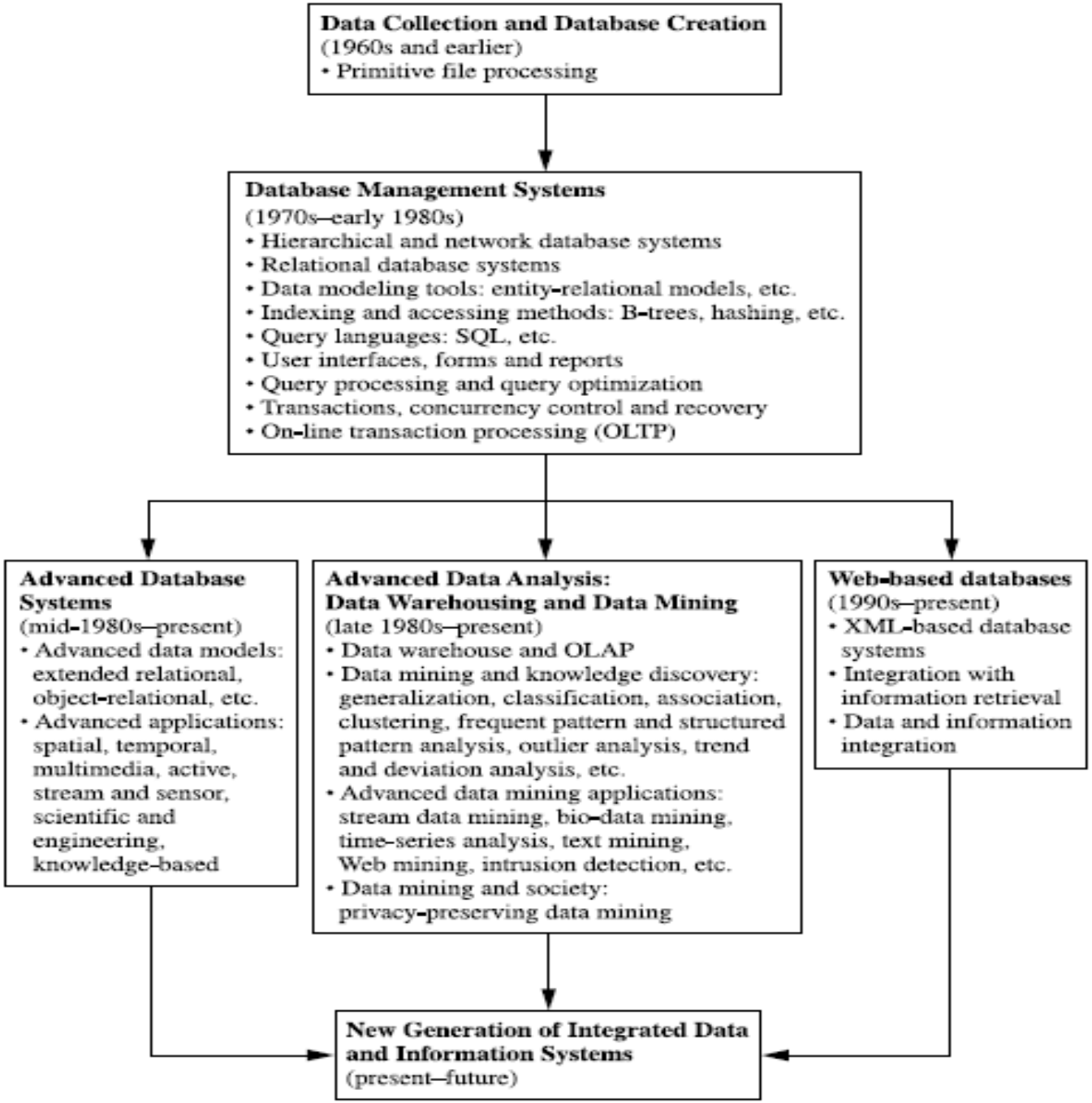


Figure: The evolution of database system technology

Very Large Databases

- Terabytes -- 10^12 bytes: Walmart -- 24 Terabytes
- Petabytes -- 10^15 bytes: Geographic Information Systems
- Exabytes -- 10^18 bytes: National Medical Records
- Zettabytes -- 10^21 bytes: Weather images
- Zottabytes -- 10^24 bytes: Intelligence Agency Videos

Data explosion problem

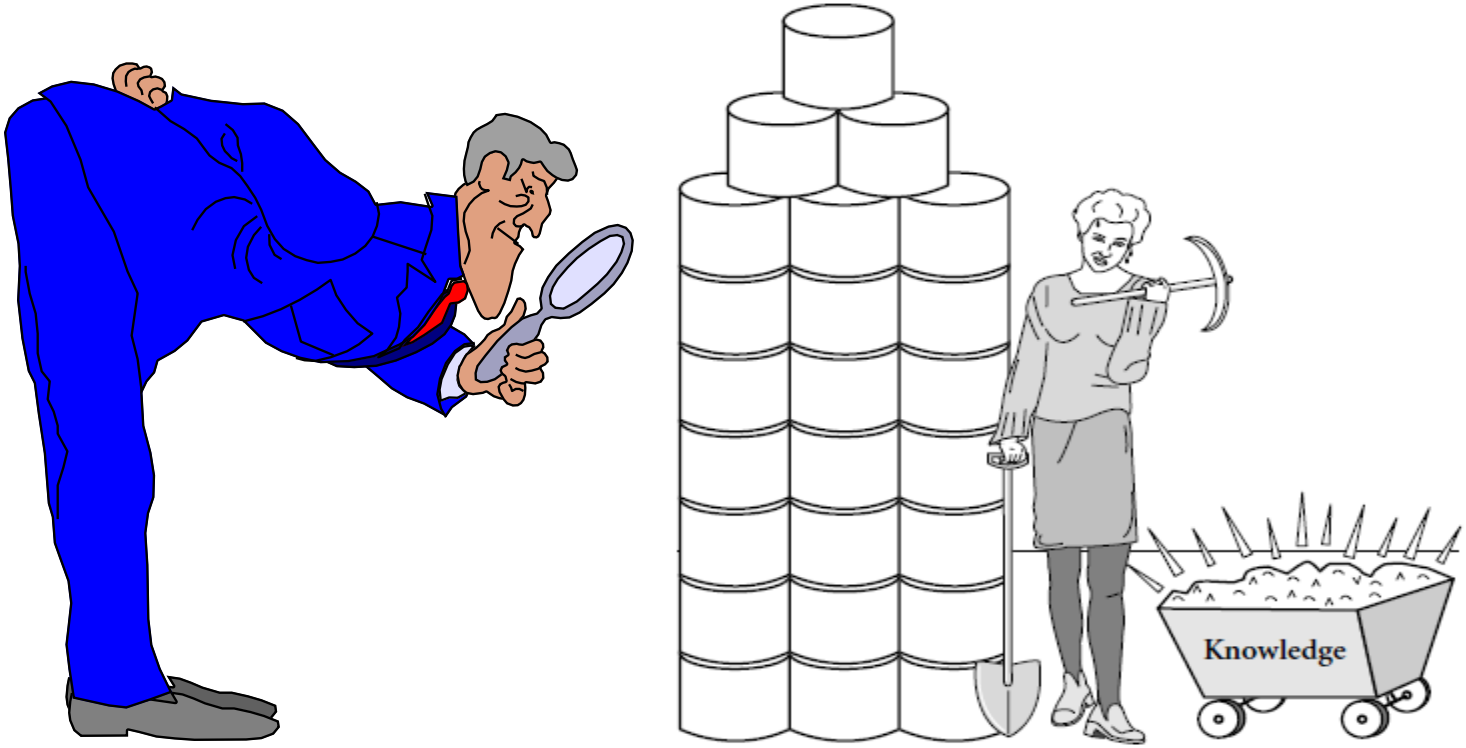
Automated data collection tools and mature database technology lead to tremendous amounts of data accumulated and/or to be analyzed in databases, data warehouses, and other information repositories

We are drowning in data, but starving for knowledge!

Solution:

"Necessity is the mother of invention"—**Data Warehousing and Data Mining**

What is Data Mining?

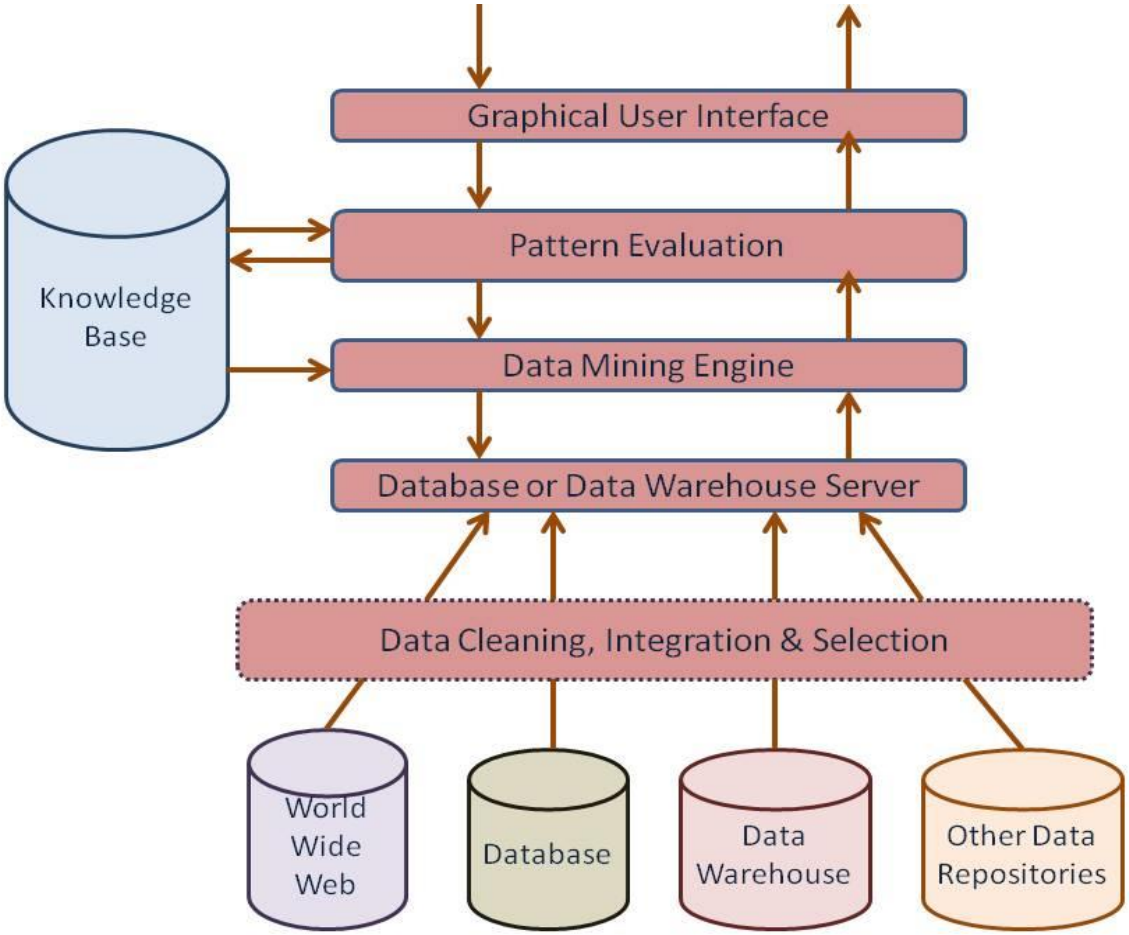


Art/Science of extracting non-trivial, implicit, previously unknown, valuable, and potentially Useful information from a large database

Data mining is

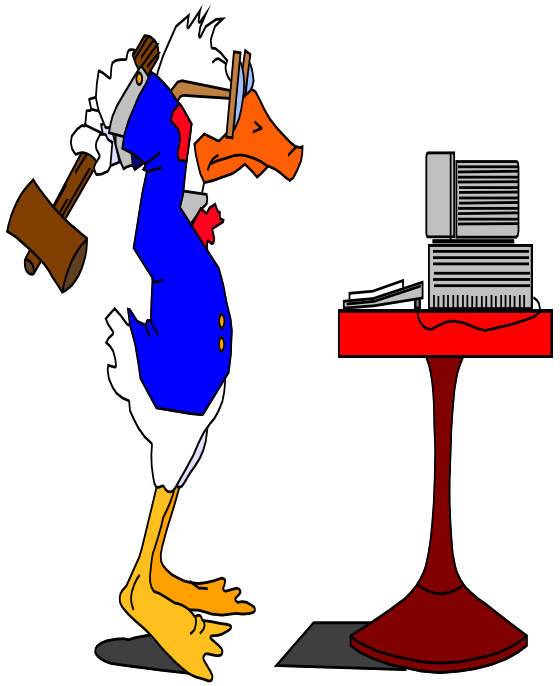
- A hot buzzword for a class of techniques that find patterns in data
- A user-centric, interactive process which leverages analysis technologies and computing power
- A group of techniques that find relationships that have not previously been discovered
- Not reliant on an existing database
- A relatively easy task that requires knowledge of the business problem/subject matter expertise

Data mining Overview



Data mining is not

- Brute-force crunching of bulk data
- "Blind" application of algorithms
- Going to find relationships where none exist
- Presenting data in different ways
- A difficult to understand technology requiring an advanced degree in computer science



Data mining is not

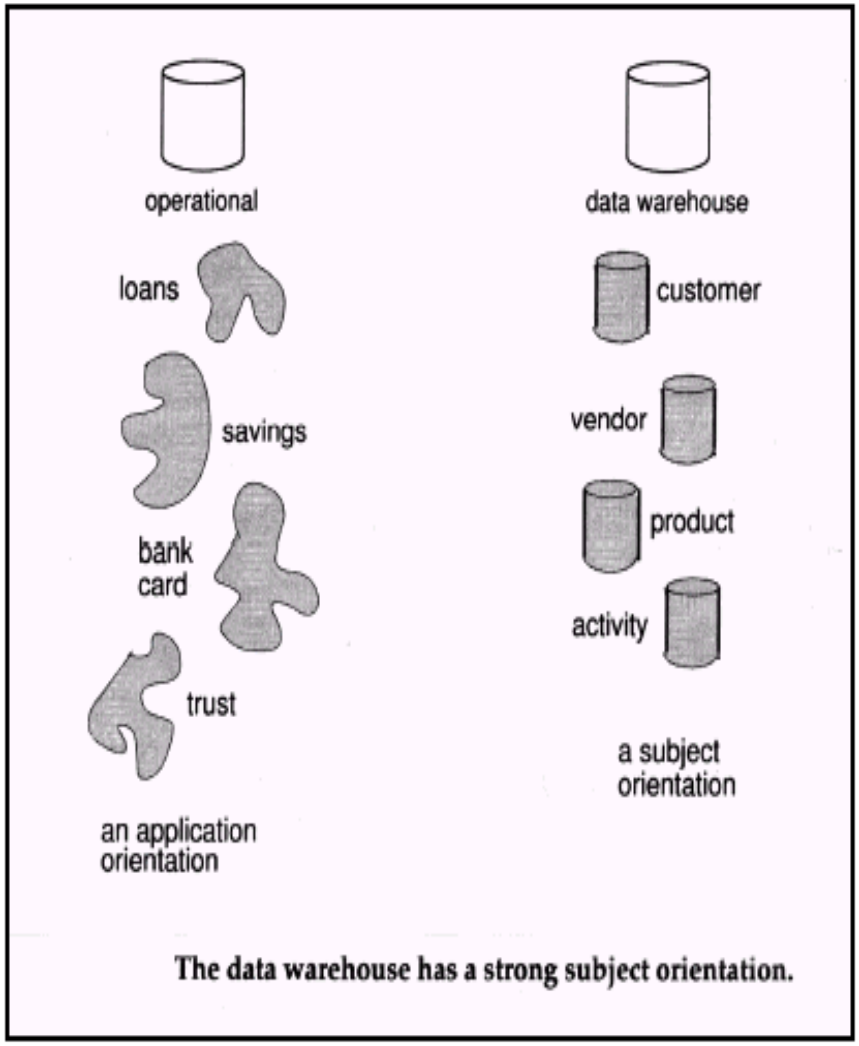
- A cybernetic magic that will turn your data into gold. It's the process and result of knowledge production, knowledge discovery and knowledge management.
- Once the patterns are found Data Mining process is finished.
- Queries to the database are not DM.

Data Warehouse

What is Data Warehouse?

- According to W. H. Inmon, a **data warehouse** is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decisions.
- "A data warehouse is a copy of transaction data specifically structured for querying and reporting" – Ralph Kimball
- **Data Warehousing** is the process of building a data warehouse for an organization.
- Data Warehousing is a process of transforming data into information and making it available to users in a timely enough manner to make a difference

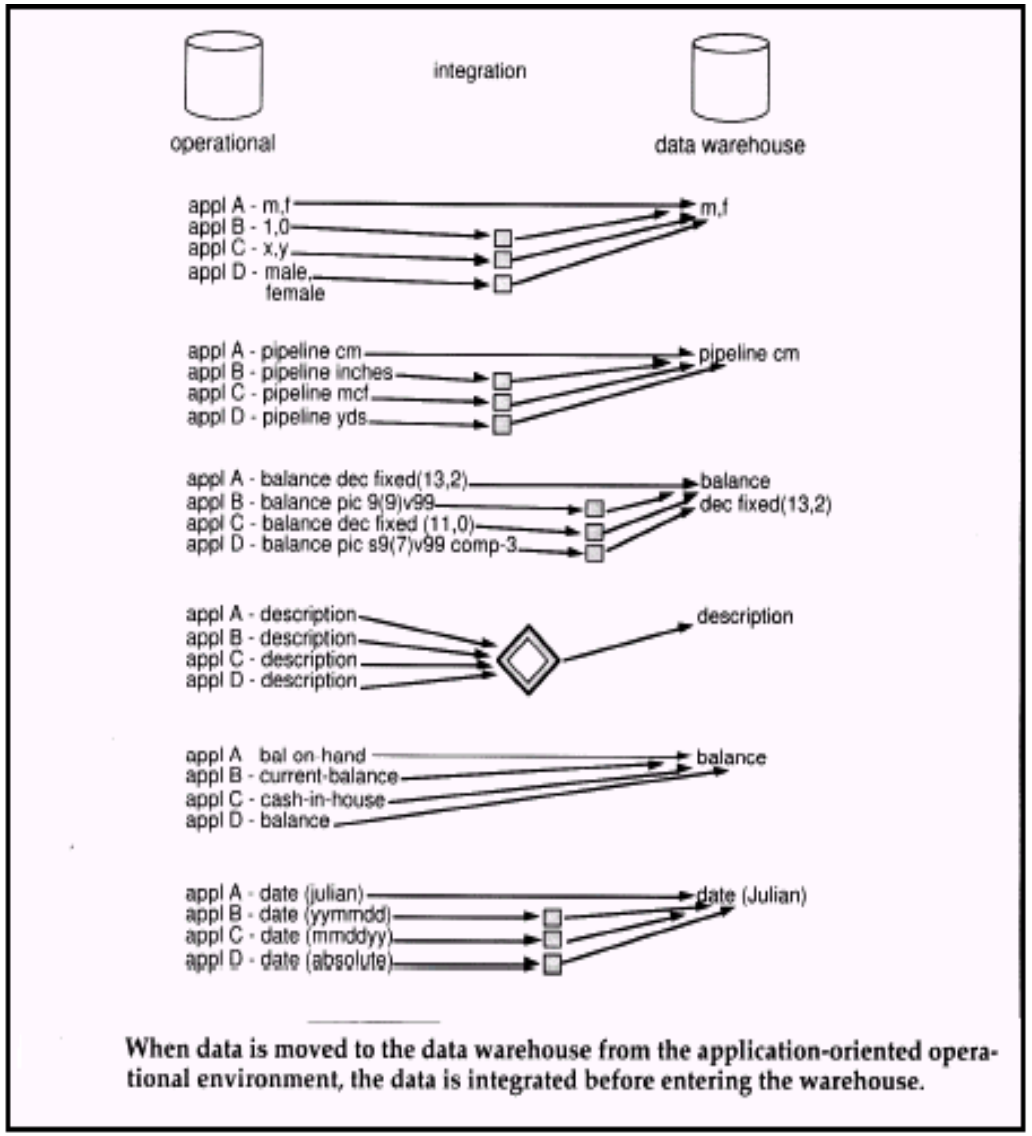
Subject Oriented



The data warehouse has a strong subject orientation.

- Focus is on Subject Areas rather than Applications
- Organized around major subjects, such as customer, product, sales.
- Provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.

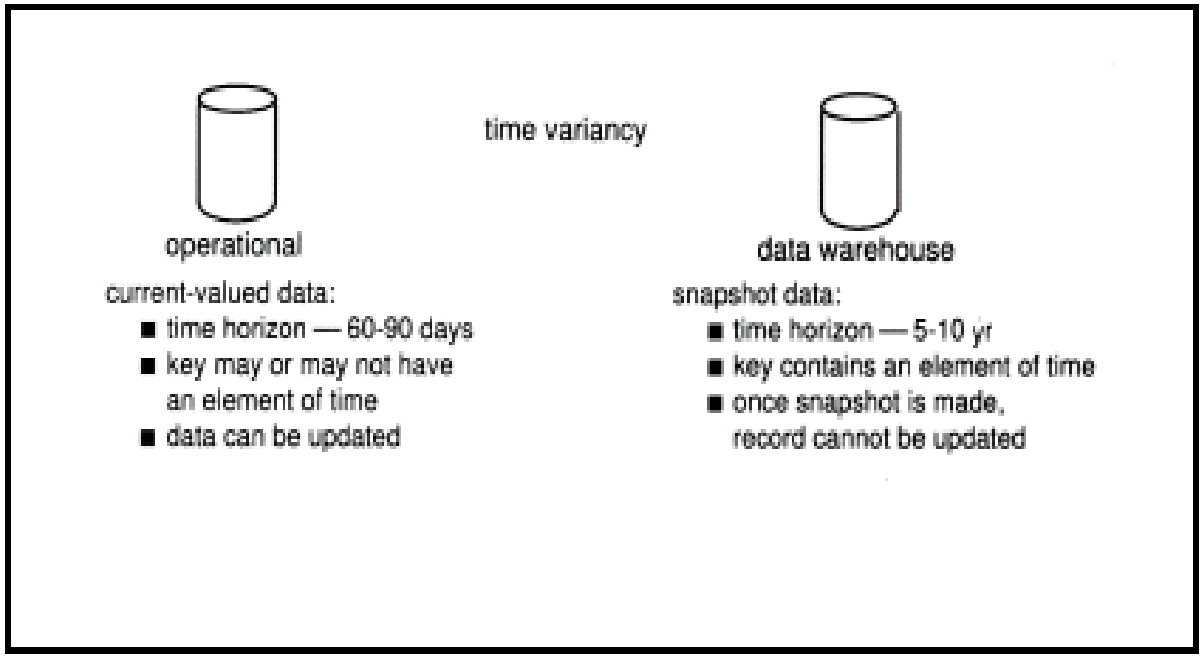
Integrated



When data is moved to the data warehouse from the application-oriented operational environment, the data is integrated before entering the warehouse.

- Constructed by integrating multiple, heterogeneous data sources
- Integration tasks handles naming conventions, physical attributes of data
- Must be made consistent.

Time Variant

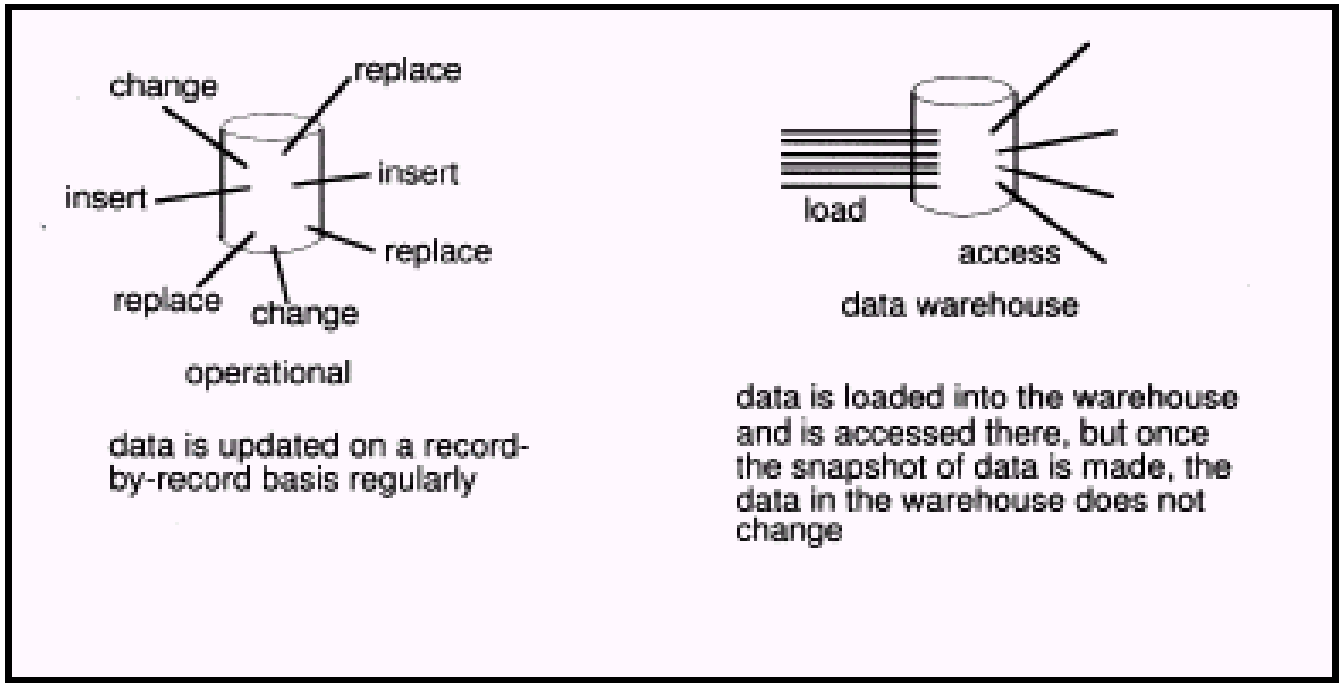


- Only accurate and valid at some point in time or over some time interval.
- The time horizon for the data warehouse is significantly longer than that of operational systems.

  Operational database provides current value data.

  Data warehouse data provide information from a historical perspective (e.g., past 5-10 years)

Non Volatile



- Data Warehouse is relatively Static in nature.
- Not updated in real-time but data in the data warehouse is loaded and refreshed from operational systems, it is not updated by end users.

Data warehousing helps business managers to :

- Extract data from various source systems on different platforms
- Transform huge data volumes into meaningful information
- Analyze integrated data across multiple business dimensions
- Provide access of the analyzed information to the business users anytime anywhere

OLTP vs. Data Warehouse

- Online Transaction Processing (OLTP) systems are tuned for known transactions and workloads while workload is not known a priori in a data warehouse
- OLTP applications normally automate clerical data processing tasks of an organization, like data entry and enquiry, transaction handling, etc. (access, read, update)
- Special data organization, access methods and implementation methods are needed to support data warehouse queries (typically multidimensional queries)
  - e.g., *average amount spent on phone calls between 9AM-5PM in Kathmandu during the month of March, 2012*

- OLTP
  - Application Oriented
  - Used to run business
  - Detailed data
  - Current up to date
  - Isolated Data
  - Repetitive access
  - Clerical User
- Data Warehouse
  - Subject Oriented
  - Used to analyze business
  - Summarized and refined
  - Snapshot data
  - Integrated Data
  - Ad-hoc access
  - Knowledge User (Manager)

- OLTP
  - Performance Sensitive
  - Few Records accessed at a time (tens)
  - Read/Update Access
  - No data redundancy
  - Database Size 100MB -100 GB
- Data Warehouse
  - Performance relaxed
  - Large volumes accessed at a time(millions)
  - Mostly Read (Batch Update)
  - Redundancy present
  - Database Size 100 GB - few terabytes

- OLTP
  - Transaction throughput is the performance metric
  - Thousands of users
  - Managed in entirety
- Data Warehouse
  - Query throughput is the performance metric
  - Hundreds of users
  - Managed by subsets

Why Data Mining?

Because it can improve customer service, better target marketing campaigns, identify high-risk clients, and improve production processes. In short, because it can help you or your company make or save money.

Data mining has been used to:

- Identify unexpected shopping patterns in supermarkets.
- Optimize website profitability by making appropriate offers to each visitor.
- Predict customer response rates in marketing campaigns.
- Defining new customer groups for marketing purposes.
- Predict customer defections: which customers are likely to switch to an alternative supplier in the near future.
- Distinguish between profitable and unprofitable customers.
- Identify suspicious (unusual) behavior, as part of a fraud detection process.

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis

Market Analysis and Management

- Where does the data come from?—Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies
- Target marketing
  - Find clusters of "model" customers who share the same characteristics: interest, income level, spending habits, etc.
  - Determine customer purchasing patterns over time
- Cross-market analysis—Find associations/co-relations between product sales, & predict based on such association
- Customer profiling—What types of customers buy what products (clustering or classification)
- Customer requirement analysis
  - Identify the best products for different groups of customers
  - Predict what factors will attract new customers
- Provision of summary information
  - Multidimensional summary reports
  - Statistical summary information (data central tendency and variation)

Corporate Analysis & Risk Management

- Finance planning and asset evaluation
  - cash flow analysis and prediction
  - contingent claim analysis to evaluate assets
  - cross-sectional and time series analysis (financial-ratio, trend analysis, etc.)
- Resource planning
  - summarize and compare the resources and spending
- Competition
  - monitor competitors and market directions
  - group customers into classes and a class-based pricing procedure
  - set pricing strategy in a highly competitive market

Fraud Detection & Mining Unusual Patterns

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

Knowledge Discovery in Databases Process

- Data selection
- Cleaning
- Enrichment
- Coding
- Data Mining
- Reporting

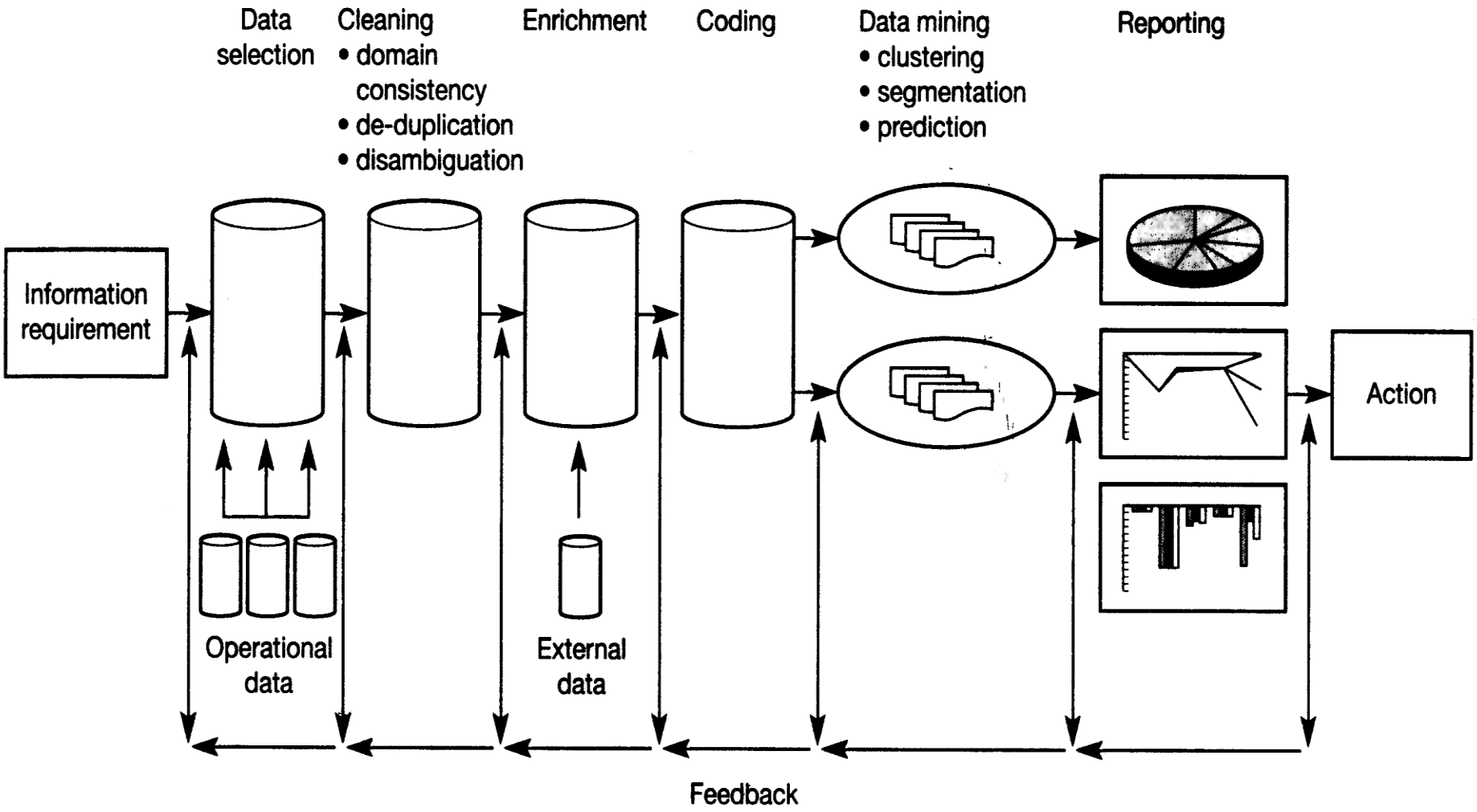


Figure: Knowledge Discovery in Databases (KDD) Process

Data Selection

Once you have formulated your informational requirements, the nest logical step is to collect and select the data you need. Setting up a KDD activity is also a long term investment.

A data environment will need to download from operational data on a regular basis, therefore investing in a data warehouse is an important aspect of the whole process.

Client number	Name	Address	Date purchase made	Magazine purchased
23003	Johnson	1 Downing Street	04–15–94	car
23003	Johnson	1 Downing Street	06–21–93	music
23003	Johnson	1 Downing Street	05–30–92	comic
23009	Clinton	2 Boulevard	01–01–01	comic
23013	King	3 High Road	02–30–95	sports
23019	***Jonson***	1 Downing Street	01–01–01	house

Figure: Original Data

Cleaning

Almost all databases in large organizations are polluted and when we start to look at the data from a data mining perspective, ideas concerning consistency of data change. Therefore, before we start the data mining process, we have to clean up the data as much as possible, and this can be done automatically in many cases.

Client number	Name	Address	Date purchase made	Magazine purchased
23003	Johnson	1 Downing Street	04–15–94	car
23003	Johnson	1 Downing Street	06–21–93	music
23003	Johnson	1 Downing Street	05–30–92	comic
23009	Clinton	2 Boulevard	***01–01–01***	comic
23013	King	3 High Road	02–30–95	sports
23003	Johnson	1 Downing Street	***01–01–01***	house

Figure: De-duplication

Client number	Name	Address	Date purchase made	Magazine purchased
23003	Johnson	1 Downing Street	04–15–94	car
23003	Johnson	1 Downing Street	06–21–93	music
23003	Johnson	1 Downing Street	05–30–92	comic
23009	Clinton	2 Boulevard	NULL	comic
23013	King	3 High Road	02–30–95	sports
23003	Johnson	1 Downing Street	12–20–94	house

Figure: Domain Consistency

Enrichment

Matching the information from bought-in databases with your own databases can be difficult. A well-known problem is the reconstruction of family relationships in databases. In a relational environment, we can simply join this information with our original data.

Client name	Date of birth	Income	Credit	Car owner	House owner
Johnson	04-13-76	$18,500	$17,800	no	no
Clinton	10-20-71	$36,000	$26,600	yes	no

Figure: Enrichment

Client number	Name	Date of birth	Income	Credit	Car owner	House owner	Address	Date purchase made	Magazine purchased
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	04-15-94	car
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	06-21-93	music
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	05-30-92	comic
23009	Clinton	10-20-71	$36,000	$26,600	yes	no	2 Boulevard	NULL	comic
23013	King	NULL	NULL	NULL	NULL	NULL	3 High Road	02-30-95	sports
23003	Johnson	04-13-76	$18,500	$17,800	no	no	1 Downing Street	12-20-94	house

Figure: Enriched Table

Coding

We can apply following coding technique:

(1) Address to regions

(2) Birthdate to age

(3) Divide income by 1000

(4) Divide credit by 1000

(5) Convert cars yes/no to 1/0

(6) Convert purchased date to months numbers

Client number	Age	Income	Credit	Car owner	House owner	Region	Month of purchase	Magazine purchased
23003	20	18.5	17.8	0	0	1	52	car
23003	20	18.5	17.8	0	0	1	42	music
23003	20	18.5	17.8	0	0	1	29	comic
23009	25	36.0	26.6	1	0	1	NULL	comic
23003	20	18.5	17.8	0	0	1	48	house

Figure: After Coding Stage

							Magazines purchased				
Client number	Age	Income	Credit	Car owner	House owner	Region	car magazine	house magazine	sports magazine	music magazine	comic magazine
23003	20	18.5	17.8	0	0	1	1	1	0	1	1
23009	25	36.0	26.6	1	0	1	0	0	0	0	1

Figure: Final Table

Data Mining

- It is a discovery stage in KDD process.
- Data mining refers to extracting or “mining” knowledge from large amounts of data.
- Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Database, or KDD.
- Alternatively, others view data mining as simply an essential step in the process of knowledge discovery.

Some Alternative names to data mining are:

- Knowledge discovery (mining) in databases (KDD)
- Knowledge extraction
- Data/pattern analysis
- Data archeology
- Data Dredging
- Information Harvesting
- Business intelligence, etc.

	Average
Age	46.9
Income	20.8
Credit	34.9
Car owner	0.59
House owner	0.59
car magazine	0.329
house magazine	0.702
sports magazine	0.447
music magazine	0.146
comic magazine	0.081

Figure: Averages

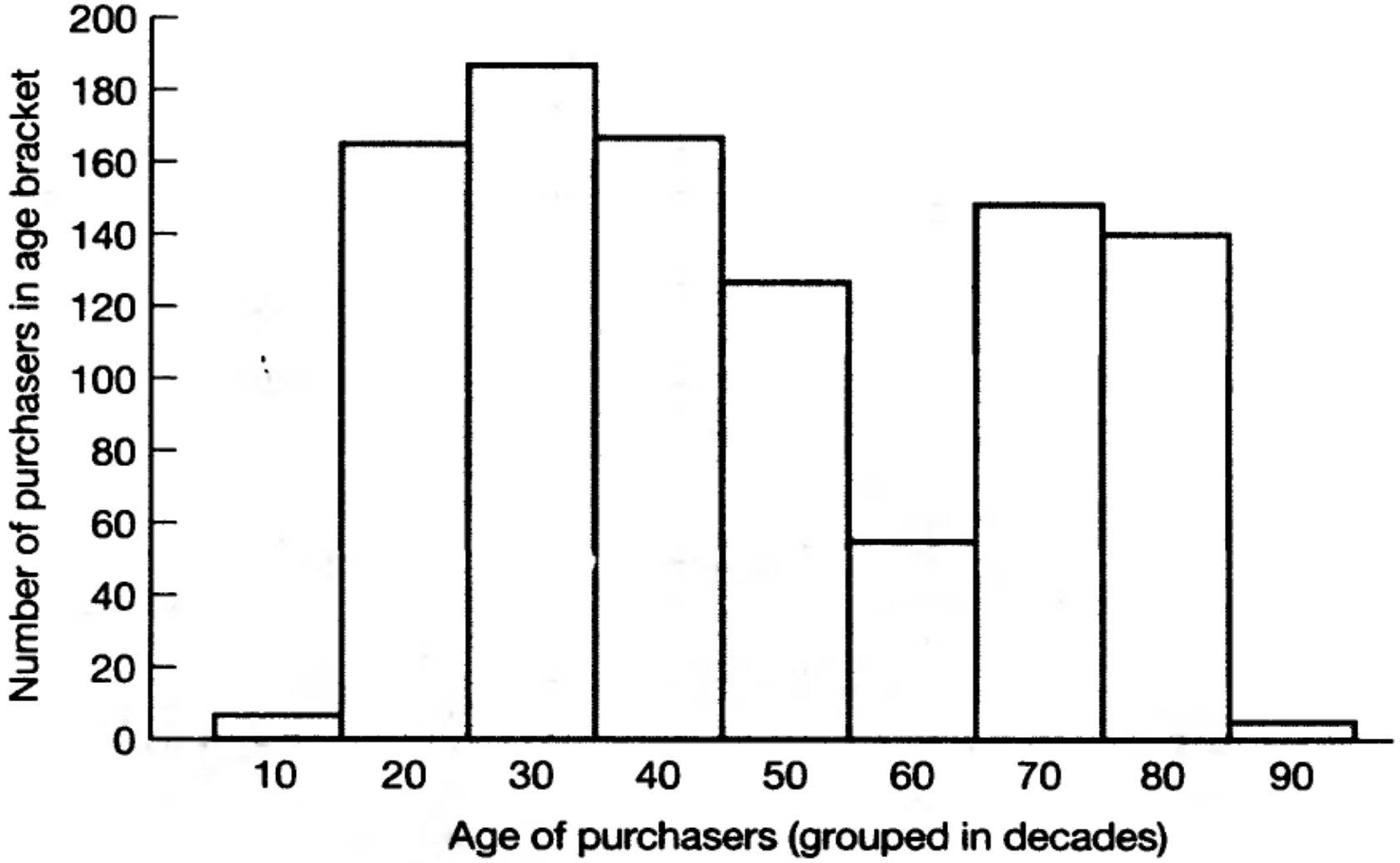


Figure: Age distribution of readers

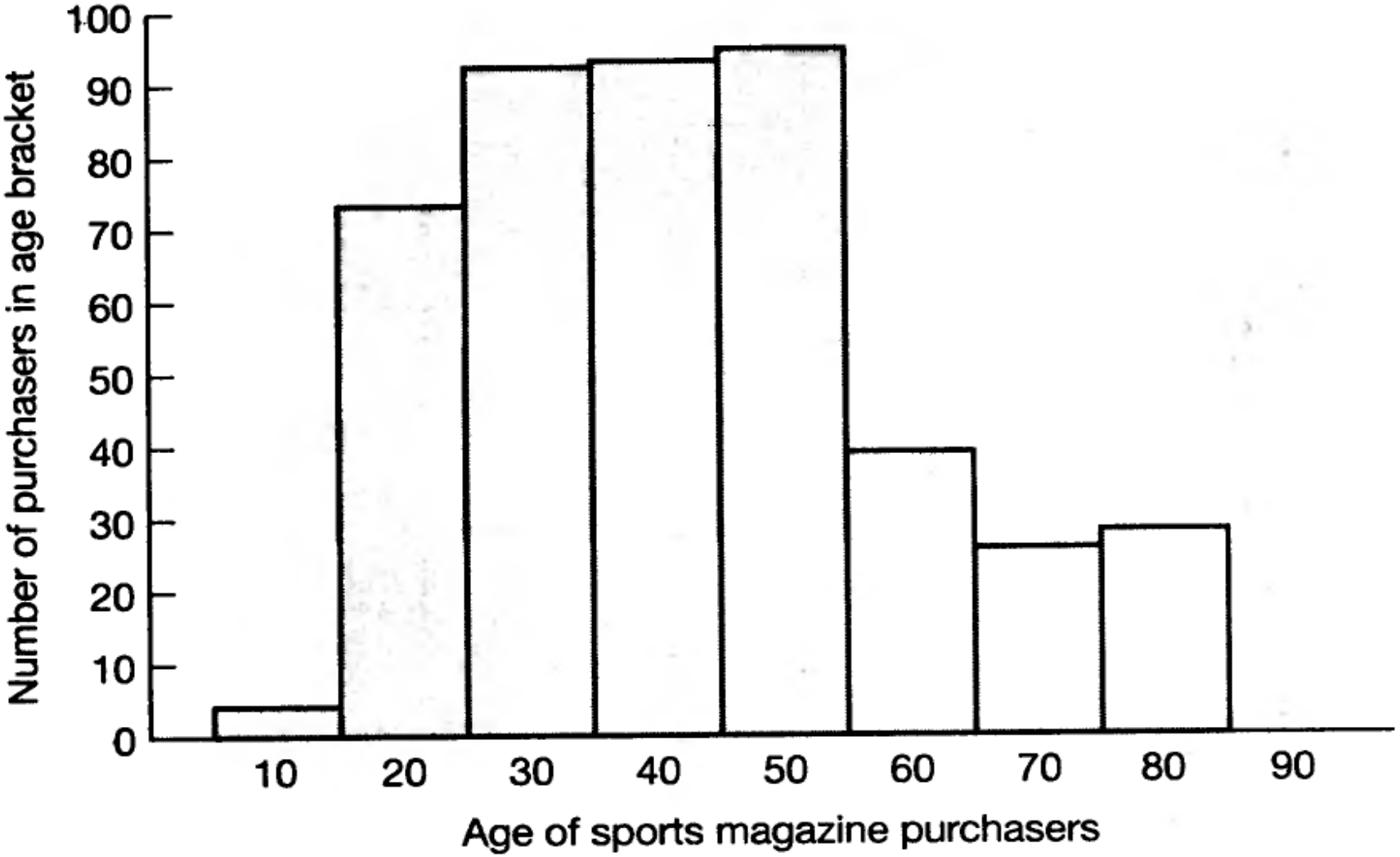


Figure: Age distribution of readers of sports magazines

Reporting

- It uses two functions:
  1. Analysis of the results
  2. Application of results
- Visualization and knowledge representation techniques are used to present the mined knowledge to the user.

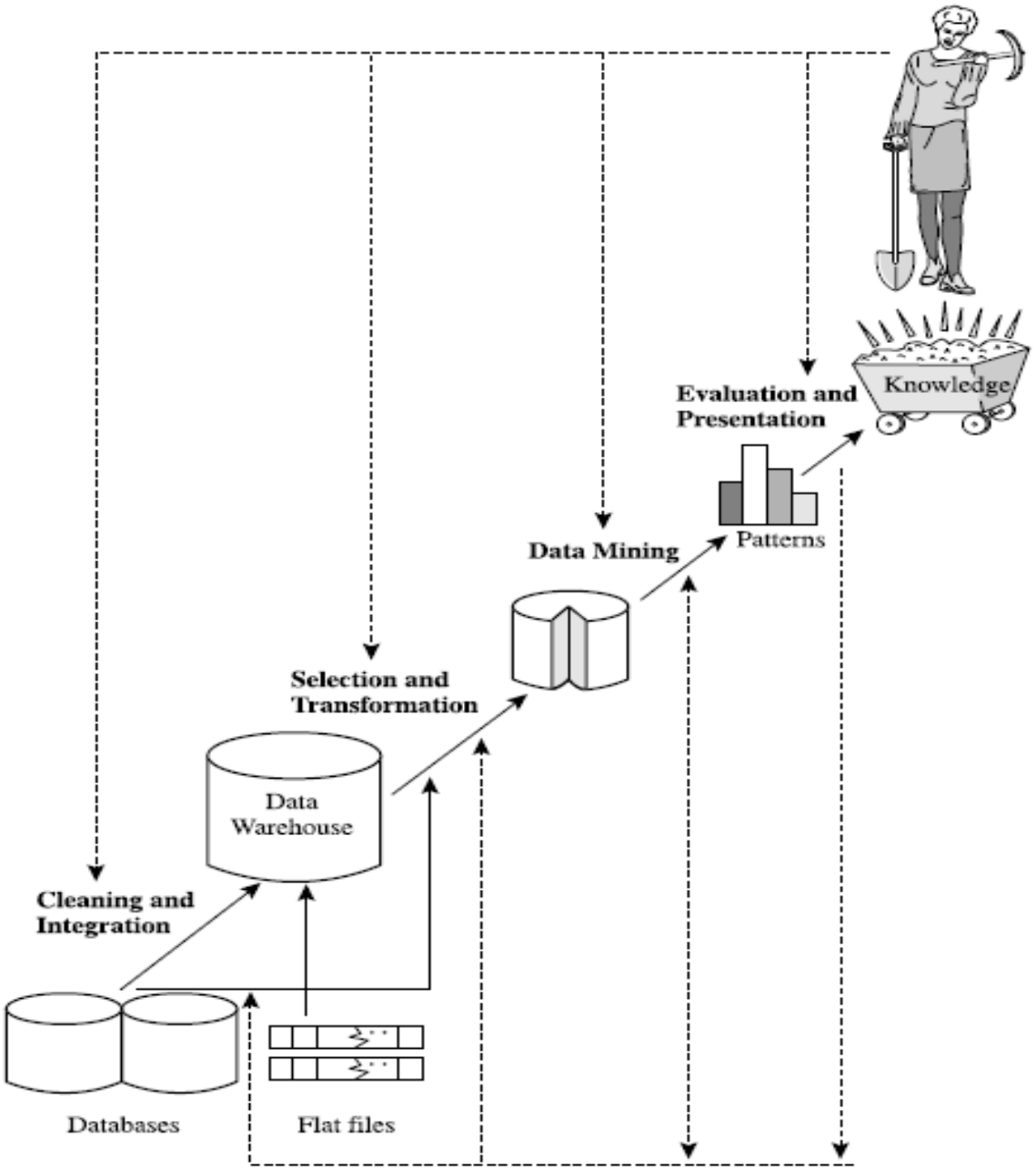
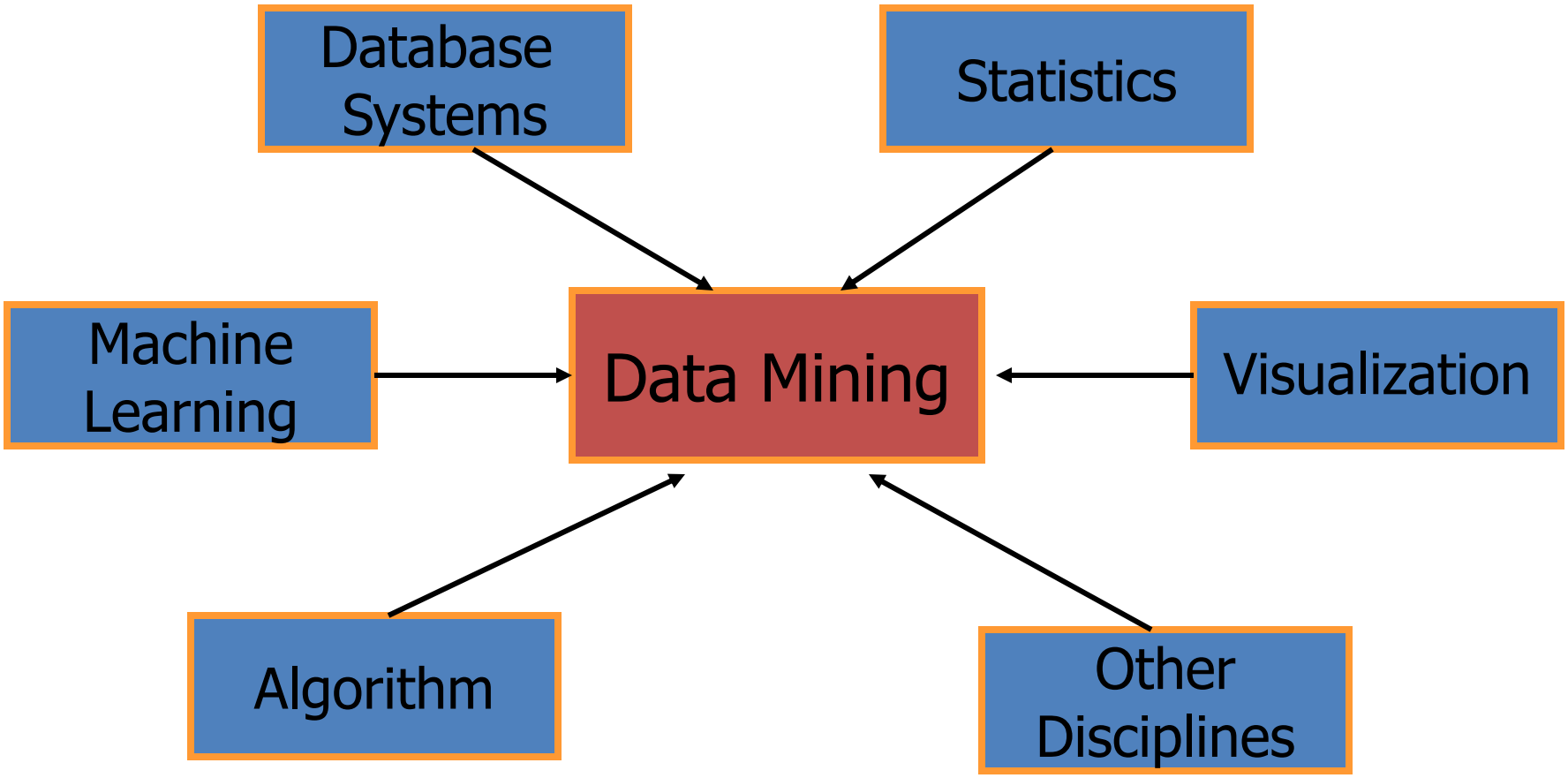
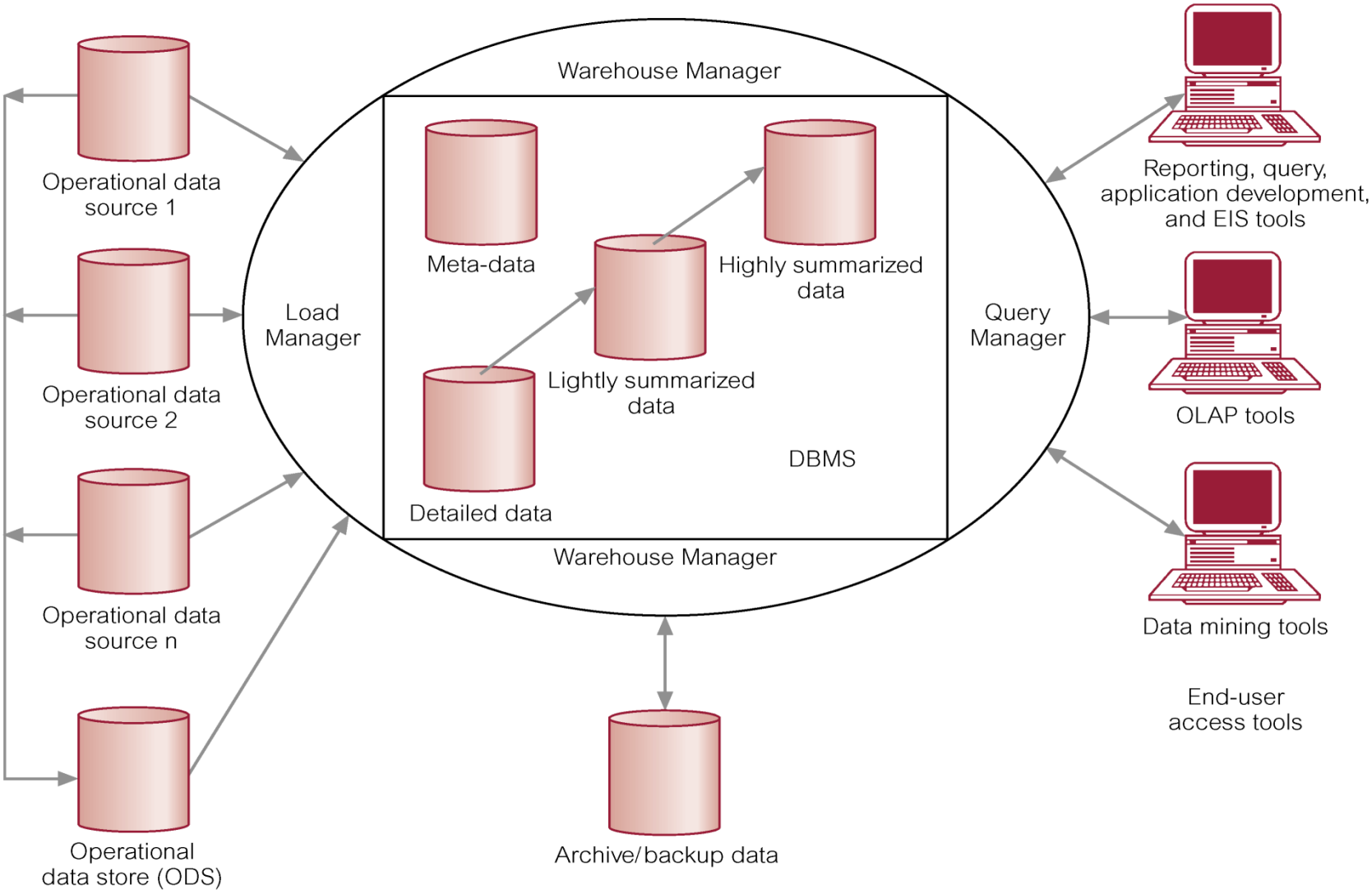


Figure: Data mining as a step in the process of knowledge discovery.

Data Mining: Confluence of Multiple Disciplines



Data Warehouse Architecture



Operational Data Sources: It may include:

- Network databases.
- Departmental file systems and RDBMSs.
- Private workstations and servers.
- External systems (Internet, commercially available databases).

Operational Data Store (ODS): It is a repository of current and integrated operational data used for analysis.

- Often structured and supplied with data in same way as DW.
- May act simply as staging area for data to be moved into the warehouse.
- Provides users with the ease of use of a relational database while remaining distant from decision support functions of the DW.

Warehouse Manager (Data Manager):

- Operations performed include:
  - Analysis of data to ensure consistency.
  - Transformation/merging of source data from temp storage into DW
  - Creation of indexes.
  - Backing-up and archiving data.

Query Manager (Manages User Queries):

- Operations include:
  - directing queries to the appropriate tables and
  - scheduling the execution of queries.
- In some cases, the query manager also generates query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate.

Meta Data: This area of the DW stores all the meta-data (data about data) definitions used by all the processes in the warehouse.

- Used for a variety of purposes:
  - Extraction and loading processes
  - Warehouse management process
  - Query management process
- End-user access tools use meta-data to understand how to build a query.
- Most vendor tools for copy management and end-user data access use their own versions of meta-data.

Lightly and Highly Summarized Data: It stores all the pre-defined lightly and highly aggregated data generated by the warehouse manager.

- The purpose of summary info is to speed up the performance of queries.
- Removes the requirement to continually perform summary operations (such as sort or group by) in answering user queries.

Archive/Backup Data: It stores detailed and summarized data for the purposes of archiving and backup.

- May be necessary to backup online summary data if this data is kept beyond the retention period for detailed data.
- The data is transferred to storage archives such as magnetic tape or optical disk.

End-User Access Tools:

- The principal purpose of data warehousing is to provide information to business users for strategic decision-making.
- Users interact with the warehouse using end-user access tools.
- There are three main groups of access tools:
  1. Data reporting, query tools
  2. Online analytical processing (OLAP) tools *(Discussed later)*
  3. Data mining tools *(Discussed later)*

Benefits of Data Warehousing

- Queries do not impact Operational systems
- Provides quick response to queries for reporting
- Enables Subject Area Orientation
- Integrates data from multiple, diverse sources
- Enables multiple interpretations of same data by different users or groups
- Provides thorough analysis of data over a period of time
- Accuracy of Operational systems can be checked
- Provides analysis capabilities to decision makers

- Increase customer profitability
- Cost effective decision making
- Manage customer and business partner relationships
- Manage risk, assets and liabilities
- Integrate inventory, operations and manufacturing
- Reduction in time to locate, access, and analyze information (Link multiple locations and geographies)
- Identify developing trends and reduce time to market
- Strategic advantage over competitors

- Potential high returns on investment
- Competitive advantage
- Increased productivity of corporate decision-makers
- Provide reliable, High performance access
- Consistent view of Data: Same query, same data. All users should be warned if data load has not come in.
- Quality of data is a driver for business re-engineering.

Applications of Data Mining

- Data mining is an interdisciplinary field with wide and diverse applications
  - There exist nontrivial gaps between data mining principles and domain-specific applications
- Some application domains
  - Financial data analysis
  - Retail industry
  - Telecommunication industry
  - Biological data analysis

Data Mining for Financial Data Analysis

- Financial data collected in banks and financial institutions are often relatively complete, reliable, and of high quality
- Design and construction of data warehouses for multidimensional data analysis and data mining
  - View the debt and revenue changes by month, by region, by sector, and by other factors
  - Access statistical information such as max, min, total, average, trend, etc.
- Loan payment prediction/consumer credit policy analysis
  - feature selection and attribute relevance ranking
  - Loan payment performance
  - Consumer credit rating

- Classification and clustering of customers for targeted marketing
  - multidimensional segmentation by nearest-neighbor, classification, decision trees, etc. to identify customer groups or associate a new customer to an appropriate customer group
- Detection of money laundering and other financial crimes
  - integration of from multiple DBs (e.g., bank transactions, federal/state crime history DBs)
  - Tools: data visualization, linkage analysis, classification, clustering tools, outlier analysis, and sequential pattern analysis tools (find unusual access sequences)

Data Mining for Retail Industry

- Retail industry: huge amounts of data on sales, customer shopping history, etc.
- Applications of retail data mining
  - Identify customer buying behaviors
  - Discover customer shopping patterns and trends
  - Improve the quality of customer service
  - Achieve better customer retention and satisfaction
  - Enhance goods consumption ratios
  - Design more effective goods transportation and distribution policies

- Example 1. Design and construction of data warehouses based on the benefits of data mining
  - Multidimensional analysis of sales, customers, products, time, and region
- Example 2. Analysis of the effectiveness of sales campaigns
- Example 3. Customer retention: Analysis of customer loyalty
  - Use customer loyalty card information to register sequences of purchases of particular customers
  - Use sequential pattern mining to investigate changes in customer consumption or loyalty
  - Suggest adjustments on the pricing and variety of goods
- Example 4. Purchase recommendation and cross-reference of items

Data Mining for Telecommunication Industry

- A rapidly expanding and highly competitive industry and a great demand for data mining
  - Understand the business involved
  - Identify telecommunication patterns
  - Catch fraudulent activities
  - Make better use of resources
  - Improve the quality of service
- Multidimensional analysis of telecommunication data
  - Intrinsically multidimensional: calling-time, duration, location of caller, location of callee, type of call, etc.

- Fraudulent pattern analysis and the identification of unusual patterns
  - Identify potentially fraudulent users and their typical usage patterns
  - Detect attempts to gain fraudulent entry to customer accounts
  - Discover unusual patterns which may need special attention
- Multidimensional association and sequential pattern analysis
  - Find usage patterns for a set of communication services by customer group, by month, etc.
  - Promote the sales of specific services
  - Improve the availability of particular services in a region
- Use of visualization tools in telecommunication data analysis

Biomedical Data Analysis

- DNA sequences: 4 basic building blocks (nucleotides): adenine (A), cytosine (C), guanine (G), and thymine (T).
- Gene: a sequence of hundreds of individual nucleotides arranged in a particular order
- Humans have around 30,000 genes
- Tremendous number of ways that the nucleotides can be ordered and sequenced to form distinct genes
- Semantic integration of heterogeneous, distributed genome databases
  - Current: highly distributed, uncontrolled generation and use of a wide variety of DNA data
  - Data cleaning and data integration methods developed in data mining will help

- Similarity search and comparison among DNA sequences
  - Compare the frequently occurring patterns of each class (e.g., diseased and healthy)
  - Identify gene sequence patterns that play roles in various diseases
- Association analysis: identification of co-occurring gene sequences
  - Most diseases are not triggered by a single gene but by a combination of genes acting together
  - Association analysis may help determine the kinds of genes that are likely to co-occur together in target samples
- Path analysis: linking genes to different disease development stages
  - Different genes may become active at different stages of the disease
  - Develop pharmaceutical interventions that target the different stages separately
- Visualization tools and genetic data analysis

Problems in Data Warehousing

- Underestimation of resources for data loading
- Hidden problems with source systems
- Required data not captured
- Increased end-user demands
- Data homogenization
- High demand for resources
- Data ownership
- High maintenance
- Long duration projects
- Complexity of integration

Major Challenges in Data Warehousing

- Data mining requires single, separate, clean, integrated, and self-consistent source of data.
  - A DW is well equipped for providing data for mining.
- Data quality and consistency is essential to ensure the accuracy of the predictive models.
  - DWs are populated with clean, consistent data
- Advantageous to mine data from multiple sources to discover as many interrelationships as possible.
  - DWs contain data from a number of sources.
- Selecting relevant subsets of records and fields for data mining
  - requires query capabilities of the DW.
- Results of a data mining study are useful if can further investigate the uncovered patterns.
  - DWs provide capability to go back to the data source.

- The largest challenge a data miner may face is the sheer volume of data in the data warehouse.
- It is quite important, then, that summary data also be available to get the analysis started.
- A major problem is that this sheer volume may mask the important relationships the data miner is interested in.
- The ability to overcome the volume and be able to interpret the data is quite important.

Major Challenges in Data Mining

- Efficiency and scalability of data mining algorithms
- Parallel, distributed, stream, and incremental mining methods
- Handling high-dimensionality
- Handling noise, uncertainty, and incompleteness of data
- Incorporation of constraints, expert knowledge, and background knowledge in data mining
- Pattern evaluation and knowledge integration
- Mining diverse and heterogeneous kinds of data: e.g., bioinformatics, Web, software/system engineering, information networks
- Application-oriented and domain-specific data mining
- Invisible data mining (embedded in other functional modules)
- Protection of security, integrity, and privacy in data mining

Warehouse Products

- Computer Associates -- CA-Ingres
- Hewlett-Packard -- Allbase/SQL
- Informix -- Informix, Informix XPS
- Microsoft -- SQL Server
- Oracle -- Oracle7, Oracle Parallel Server
- Red Brick -- Red Brick Warehouse
- SAS Institute -- SAS
- Software AG -- ADABAS
- Sybase -- SQL Server, IQ, MPP

Data Mining Products

Candidate Products	Vendor
Enterprise Miner	SAS
Intelligent Miner	IBM
SPSS Clementine	SPSS
Teradata Miner	NCR

References

1. “Knowledge Discovery Nuggets”: http://www.kdnuggets.com/

2. T. Bhavani. “Data Mining: Technologies, Techniques, Tools and Trends”, CRC Press 1999.

3. H. David, M. Heikki and S. Padhraic. “*Principles of Data Mining”, MIT Press 2001*.

4. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. “Advances in Knowledge Discovery and Data Mining”, *AAAI/MIT Press, 1996*.

5. Sam Anahory, Dennis Murray, “Data warehousing In the Real World”, Pearson Education.

6. Adriaans, P. and D. Zatinge, “ Data Mining” , Addison Wesley, 1996

7. Kimball, R. “The Data Warehouse Toolkit”, Wiley, 1996.

8. “Data Mining Concepts and Techniques”, Morgan Kaufmann J. Han, M Kamber Second Edition ISBN : 978-1-55860-901-3

9. U. Fayyad, G. Grinstein, and A. Wierse. “*Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann”, 2001.*

10. H. Witten and E. Frank. “Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations”, Morgan Kaufmann, 2001.