# TRINITY COLLEGE FOR WOMEN NAMAKKAL

## Department of Computer Science

## DATA MINING AND WAREHOUSING

## 19UAC   -EVEN Semester

Presented by

Mrs.S.USHARANI

Assistant Professor

Department of Computer Science

http://www.trinitycollegenkl.edu.in/

# What is Data Mining

➢ Data mining is the process of finding anomalies, patterns and correlations within large data sets to predict outcomes.

➢ Data mining (knowledge discovery from data)
Extraction of interesting (<u>non-trivial,</u> <u>implicit,</u> <u>previously unknown</u> and <u>potentially useful)</u> patterns or knowledge from huge amount of data.

## Alternative names
Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
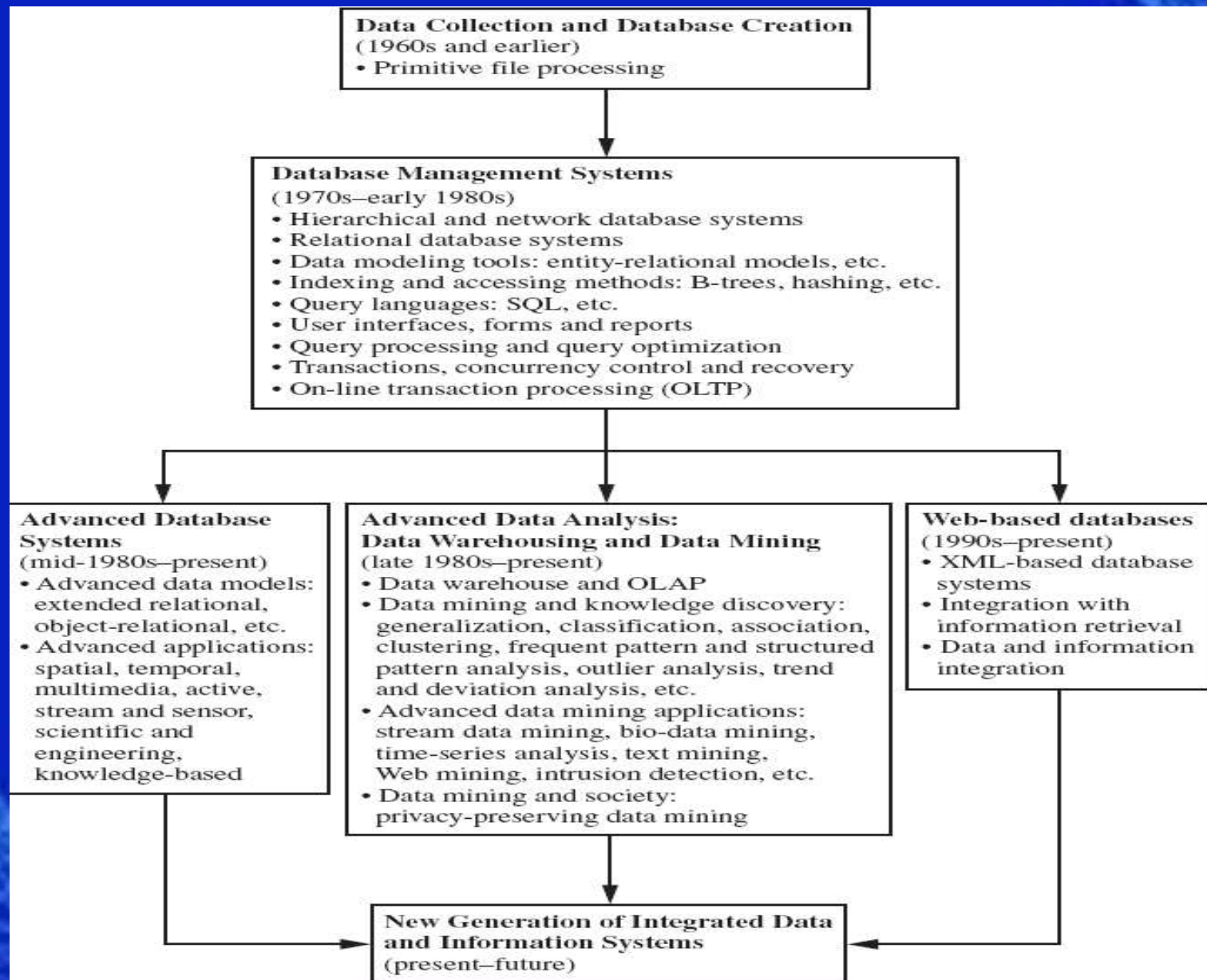
# Why Data Mining?

- The Explosive Growth of Data: from terabytes($1000^4$) to yottabytes($1000^8$)
  - Data collection and data availability
    - Automated data collection tools, database systems, web
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: bioinformatics, scientific simulation, medical research …
    - Society and everyone: news, digital cameras, …
- Data rich but information poor!
  - What does those data mean?
  - How to analyze data?
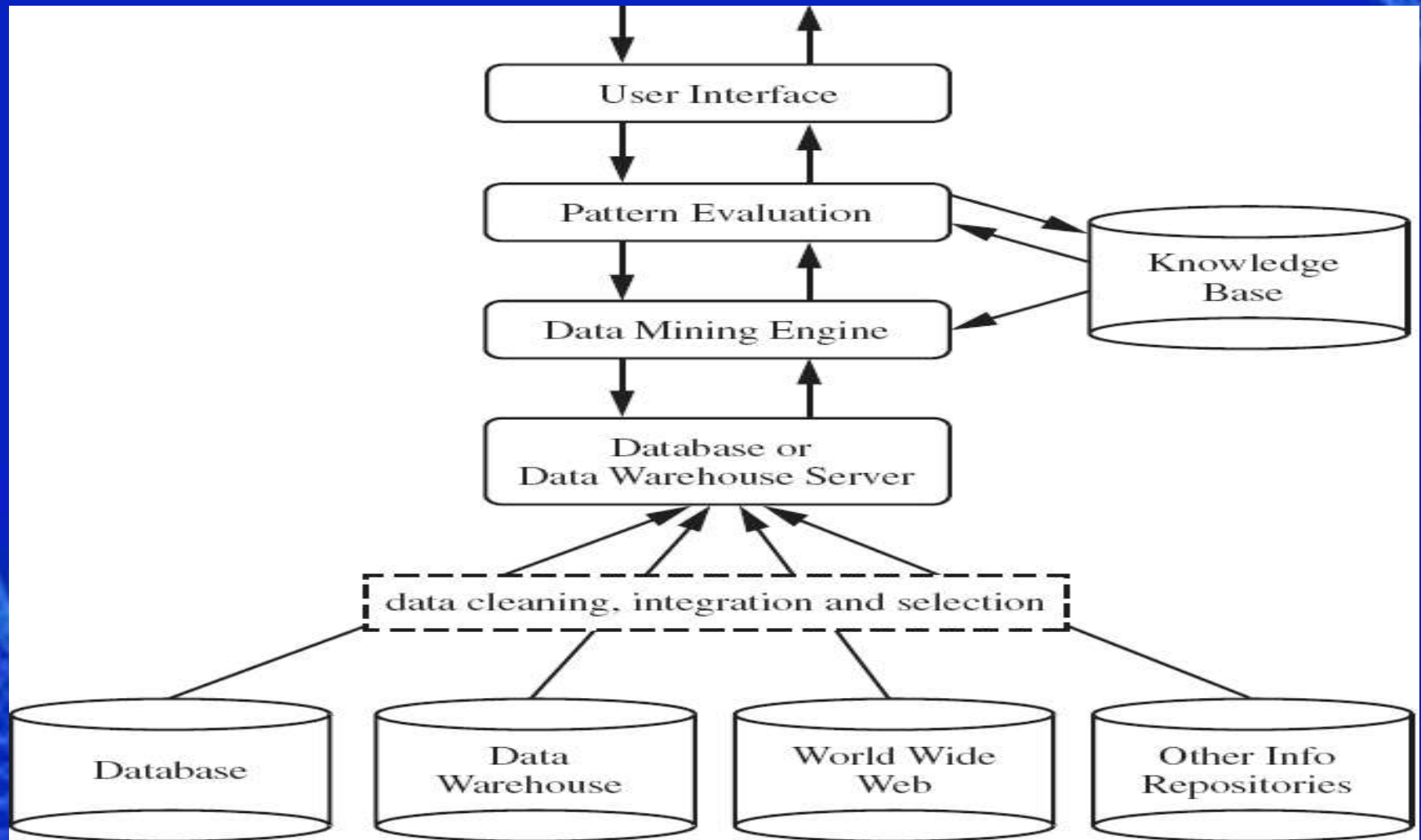- Data mining — Automated analysis of massive data sets

# Potential Applications

- 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
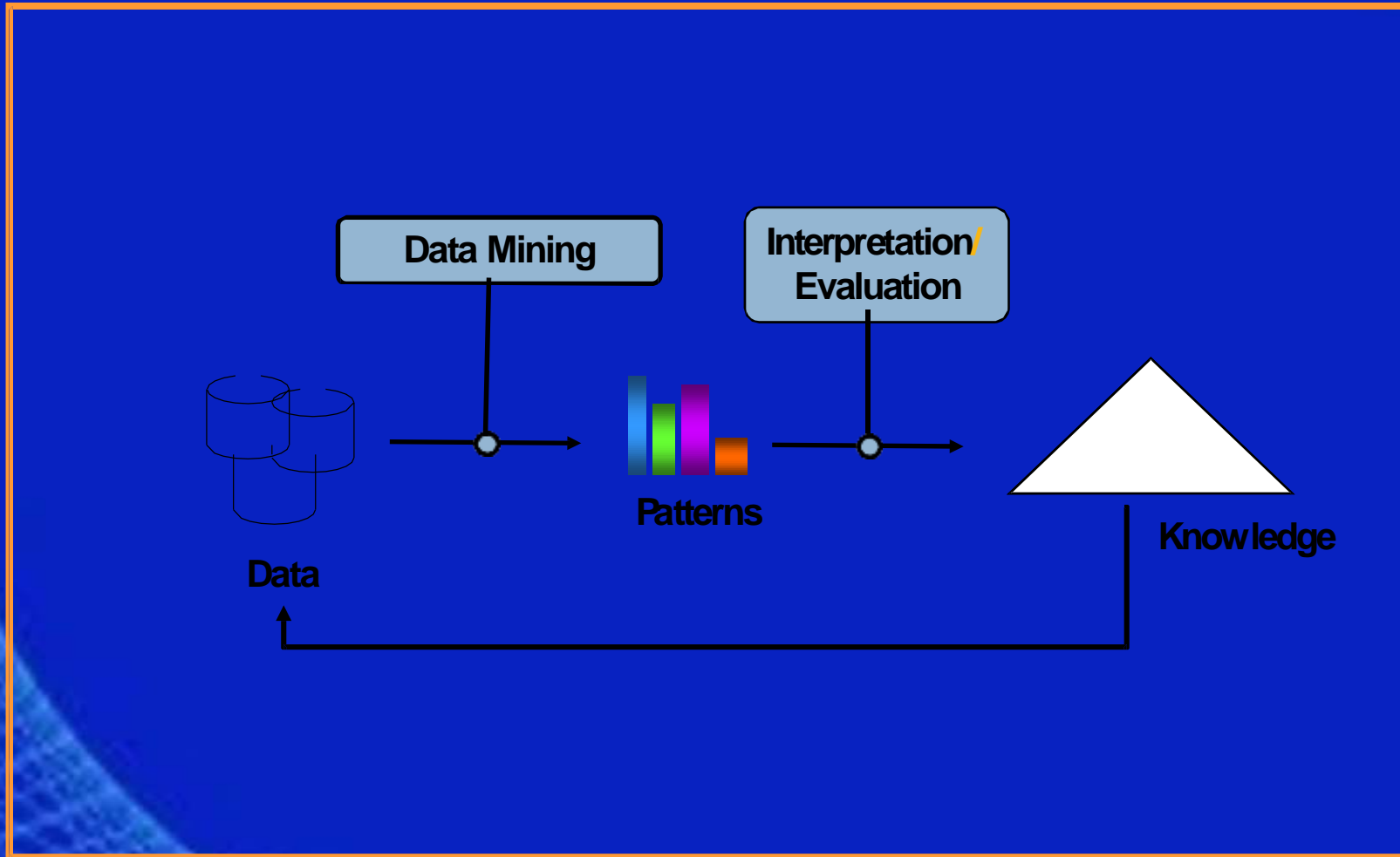  - Bioinformatics and bio-data analysis

# Evolution of Database Technology

**Data Collection and Database Creation**
(1960s and earlier)
• Primitive file processing

**Database Management Systems**
(1970s–early 1980s)
• Hierarchical and network database systems
• Relational database systems
• Data modeling tools: entity-relational models, etc.
• Indexing and accessing methods: B-trees, hashing, etc.
• Query languages: SQL, etc.
• User interfaces, forms and reports
• Query processing and query optimization
• Transactions, concurrency control and recovery
• On-line transaction processing (OLTP)

**Advanced Database Systems**
(mid-1980s–present)
• Advanced data models: extended relational, object-relational, etc.
• Advanced applications: spatial, temporal, multimedia, active, stream and sensor, scientific and engineering, knowledge-based

**Advanced Data Analysis:**
**Data Warehousing and Data Mining**
(late 1980s–present)
• Data warehouse and OLAP
• Data mining and knowledge discovery: generalization, classification, association, clustering, frequent pattern and structured pattern analysis, outlier analysis, trend and deviation analysis, etc.
• Advanced data mining applications: stream data mining, bio-data mining, time-series analysis, text mining, Web mining, intrusion detection, etc.
• Data mining and society: privacy-preserving data mining

**Web-based databases**
(1990s–present)
• XML-based database systems
• Integration with information retrieval
• Data and information integration

**New Generation of Integrated Data and Information Systems**
(present–future)

# A typical DM System Architecture

# Knowledge Discovery in Data: Process

# Knowledge Discovery in Data: Challenges

# Data Come from Everywhere



**Grocery Markets**

**E-Commerce**

**Stock Exchange**

## But, they have different form



**Hospital**

**Weather Station**

**Social Media**

# What is Data?

- Collection of records and their attributes

- An attribute is a characteristic of an object

- A collection of attributes describe an object

Objects

| 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 |

# Typesof Data

- Record Data
  - Transactional Data

- Temporal Data
  - Time Series Data
  - Sequence Data

- Spatial & Spatial-Temporal Data
  - Spatial Data
  - Spatial-Temporal Data
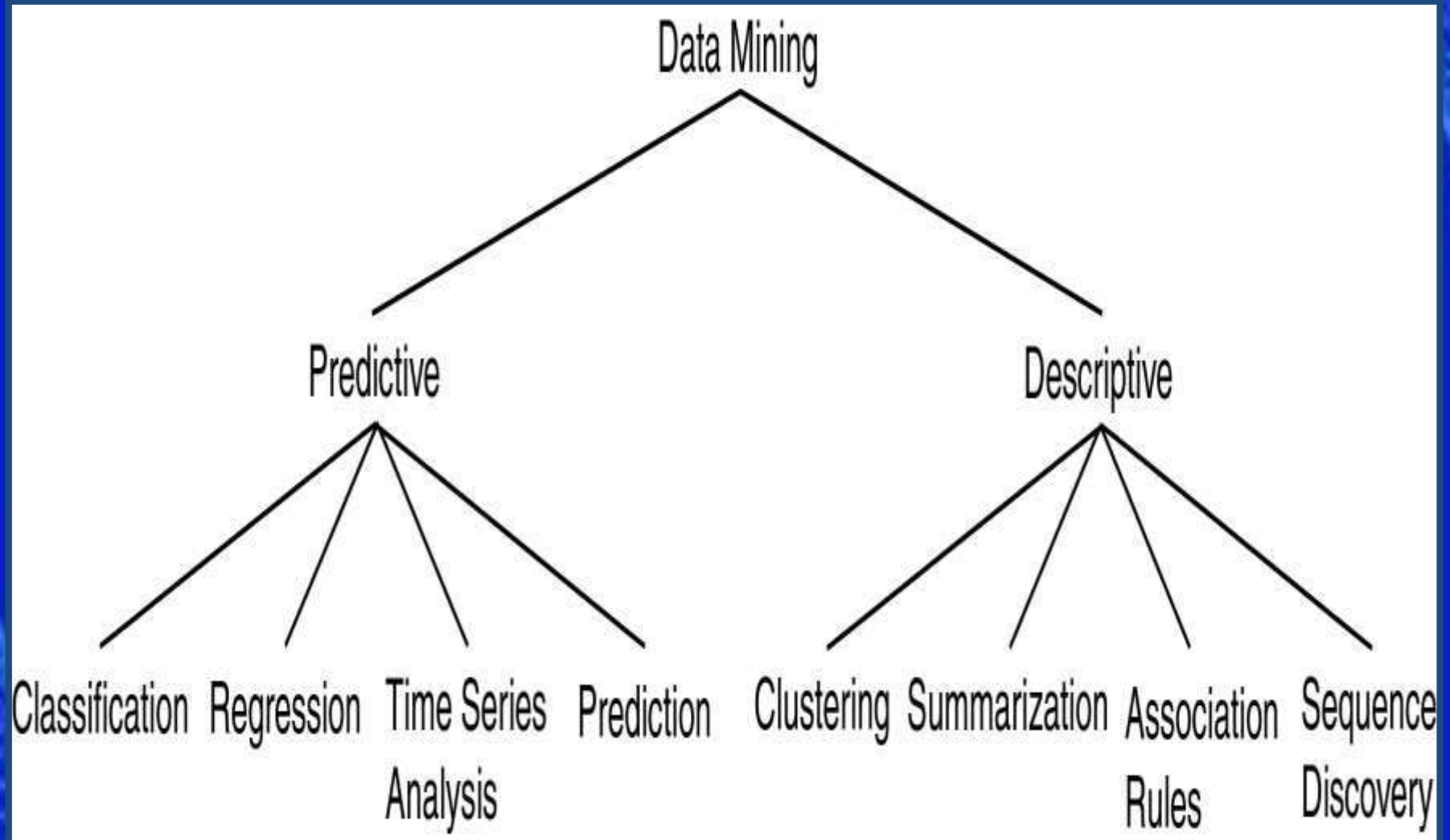
- Graph Data
  - Transactional Data

- UnStructured Data
  - Twitter Status Message
  - Review, news article

- Semi-Structured Data
  - Paper Publications Data
  - XML format

# Data Mining Models and Tasks

# Decisions in Data Mining

- **<u>Databases to be mined</u>**
  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.

- **<u>Knowledge to be mined</u>**
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Multiple/integrated functions and mining at multiple levels
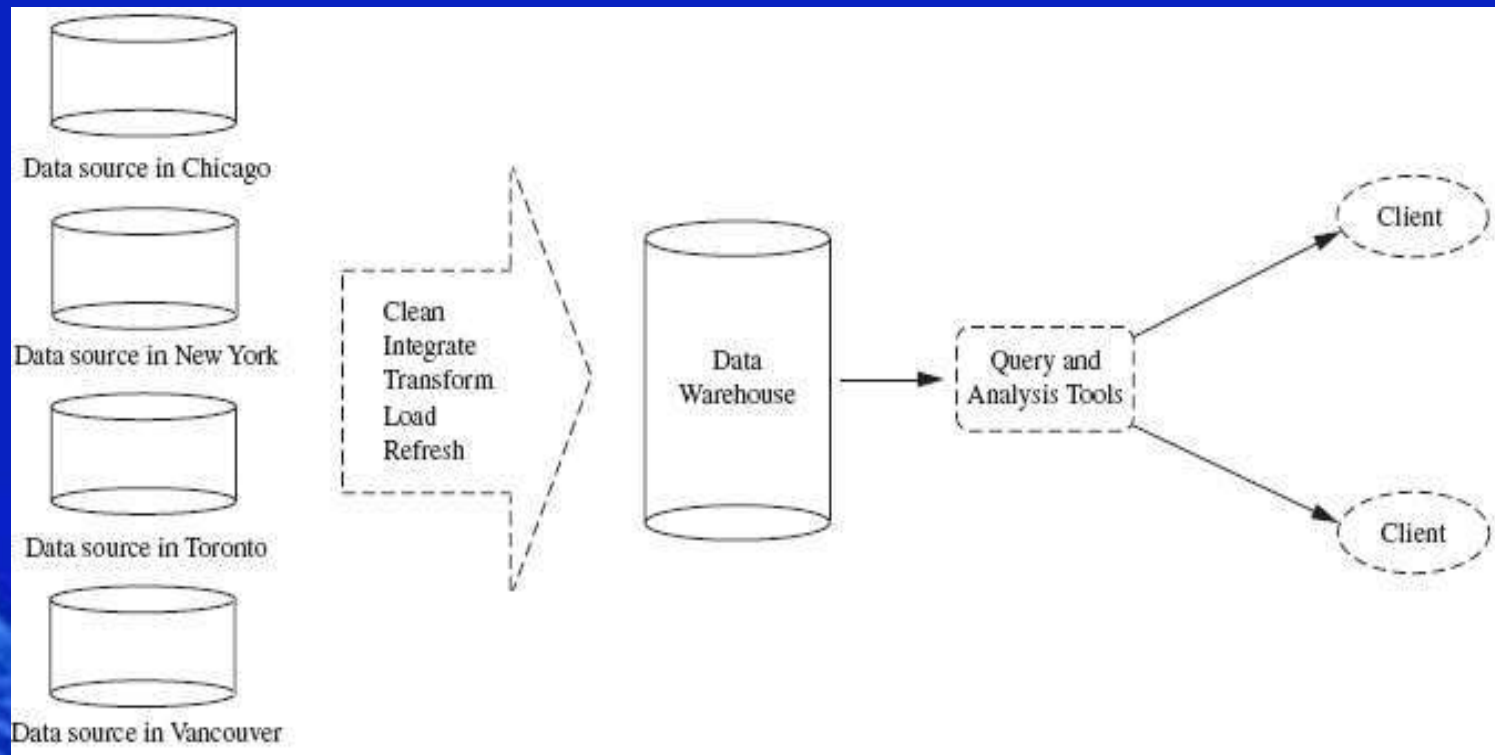
- **<u>Techniques utilized</u>**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.

- **<u>Applications adapted</u>**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

# Data Warehouses

- A repository of information collected from multiple sources, stored under a unified schema, and that usually resides at a single site.
- Constructed via a process of data cleaning, data integration, data transformation, data loading and periodic data refreshing.

# THANK YOU

http://www.trinitycollegenkl.edu.in/