



# RAMA UNIVERSITY

www.ramauniversity.ac.in

FACULTY OF ENGINEERING

## DATA MINING & WAREHOUSEING LECTURE-18

MR. DHIRENDRA

ASSISTANT PROFESSOR

RAMA UNIVERSITY

# OUTLINE

- ❖ **MOTIVATION**
- ❖ **DATA MINING**
- ❖ **THREE LEVELS OF TESTING**
- ❖ **EVOLUTION OF DATABASE TECHNOLOGY**
- ❖ **WHAT IS DATA MINING**
- ❖ **DATA MINING ALGORITHM**
- ❖ **DATA MINING PROCESS**
- ❖ **MCQ**
- ❖ **REFERENCES**

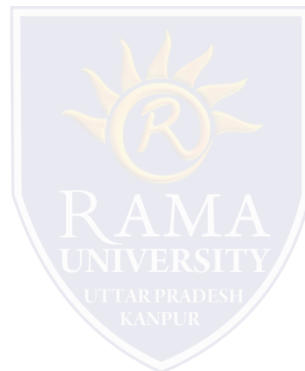


# Motivation

- ❑ In real world applications data can be inconsistent incomplete and or noisy.

## Errors can happen:

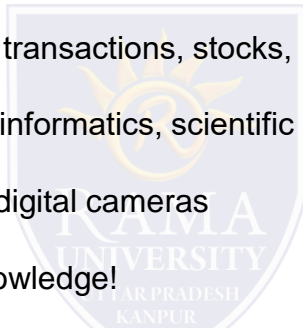
- ❑ Faulty data collection instruments
- ❑ Data entry problems.
- ❑ Human misjudgment during data entry
- ❑ Data transmission problems.
- ❑ Technology limitations
- ❑ Discrepancy in naming conventions



## Results:

- ❑ Duplicated records
- ❑ Incomplete data
- ❑ Contradictions in data.

- ❑ 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
- ❑ We are drowning in data, but starving for knowledge!
- ❑ “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets



# Evolution of Database Technology

## ❑ 1960s:

Data collection, database creation, IMS (Information Management System) and network DBMS

## ❑ 1970s:

Relational data model, relational DBMS implementation

## ❑ 1980s:

RDBMS, advanced data models (extended-relational, OO, deductive, etc.)

Application-oriented DBMS (spatial, scientific, engineering, etc.)

## ❑ 1990s:

Data mining, data warehousing, multimedia databases, and Web databases

## ❑ 2000s

Stream data management and mining

Data mining and its applications

Web technology (XML, data integration) and global information systems

# What Is Data Mining?

## ❑ Data mining (knowledge discovery from data)

- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Data mining: a misnomer?

❑ The exploration and analysis, by Automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns.

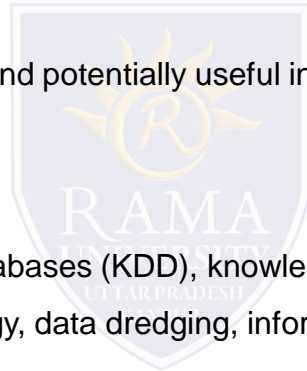
❑ The extraction of implicit, previously unknown, and potentially useful information from data or the process of discovery advantages patterns in data.

## ❑ Alternative names

- 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”?

- Simple search and query processing
- (Deductive) expert systems



# Data Mining Algorithm

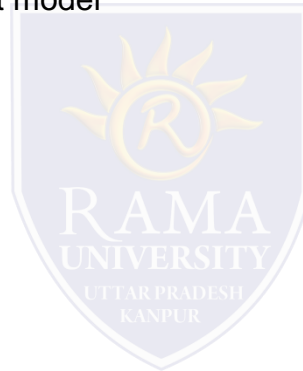
## ❑ **Objective:** Fit Data to a Model

- Descriptive (characterize the general properties of the data in the database)
- Predictive (perform inference on the current data in order to make prediction)

## ❑ **Preference** – Technique to choose the best model

## ❑ **Search** – Technique to search the data

- **“Query”**



# Data Mining Process

❑ **Define & Understanding the Problem.**

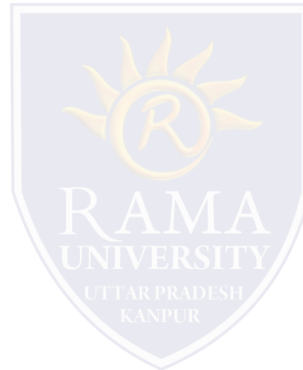
❑ **Data Warehousing**

- Collect / Extract data
- Clean Data
- Data Engineering

❑ **Algorithm selection / Engineering**

❑ **Run Mining Algorithm**

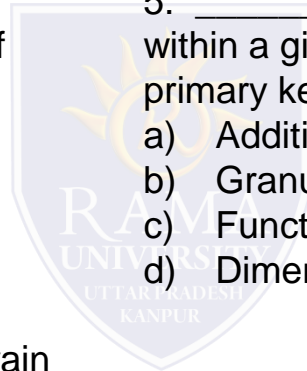
❑ **Analyze the Results**





# Multiple Choice Question

1. The dimension tables describe the \_\_\_\_\_.
  - a) entities
  - b) facts
  - c) keys
  - d) units of measures.
2. The granularity of the fact is the \_\_\_\_\_ of detail at which it is recorded.
  - a) transformation
  - b) summarization
  - c) level
  - d) transformation and summarization.
3. Which of the following is not a primary grain in analytical modeling?
  - a) Transaction
  - b) Periodic snapshot.
  - c) Accumulating snapshot.
  - d) All of the above.
4. Granularity is determined by \_\_\_\_\_.
  - a) number of parts to a key.
  - b) granularity of those parts.
  - c) both A and B.
  - d) none of the above.
5. \_\_\_\_\_ of data means that the attributes within a given entity are fully dependent on the entire primary key of the entity.
  - a) Additivity
  - b) Granularity
  - c) Functional dependency.
  - d) Dimensionality.



# REFERENCES

- [https://www.tutorialspoint.com/dwh/dwh\\_overview.htm](https://www.tutorialspoint.com/dwh/dwh_overview.htm)
- <http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf> DATA MINING BOOK WRITTEN BY Micheline Kamber
- <https://www.javatpoint.com/three-tier-data-warehouse-architecture>
- M.H. Dunham, “ Data Mining: Introductory & Advanced Topics” Pearson Education
- Jiawei Han, Micheline Kamber, “ Data Mining Concepts & Techniques” Elsevier
- Sam Anahory, Dennis Murray, “ data warehousing in the Real World: A Practical Guide for Building Decision Support Systems, “ Pearson Education
- Mallach, “ Data Warehousing System”, TMH
- R. Agrawal, A. Gupta, and S. Sarawagi. Modeling multidimensional databases. ICDE’97 S. Chaudhuri and U. Dayal. An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26:65-74, 1997
- S. Agarwal, R. Agrawal, P. M. Deshpande, A. Gupta, J. F. Naughton, R. Ramakrishnan, and S. Sarawagi. On the computation of multidimensional aggregates. VLDB’96 D. Agrawal, A. E. Abbadi, A. Singh, and T. Yurek. Efficient view maintenance in data warehouses. SIGMOD’97
- E. F. Codd, S. B. Codd, and C. T. Salley. Beyond decision support. Computer World, 27, July 1993.
- J. Gray, et al. Data cube: A relational aggregation operator generalizing group-by, cross-tab and sub-totals. Data Mining and Knowledge Discovery, 1:29-54, 1997.