



**RAMA  
UNIVERSITY**

[www.ramauniversity.ac.in](http://www.ramauniversity.ac.in)

**FACULTY OF ENGINEERING**

**DATA MINING & WAREHOUSEING  
LECTURE-27**

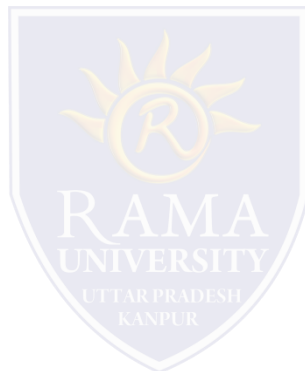
**MR. DHIRENDRA**

ASSISTANT PROFESSOR

RAMA UNIVERSITY

# OUTLINE

- ❖ DATA REDUCTION IN DATA MINING
- ❖ METHODS OF DATA REDUCTION
- ❖ DATA COMPRESSION
- ❖ NUMEROSITY REDUCTION
- ❖ DISCRETIZATION
- ❖ MCQ
- ❖ REFERENCES



# Data Reduction in Data Mining

The method of data reduction may achieve a condensed description of the original data which is much smaller in quantity but keeps the quality of the original data.

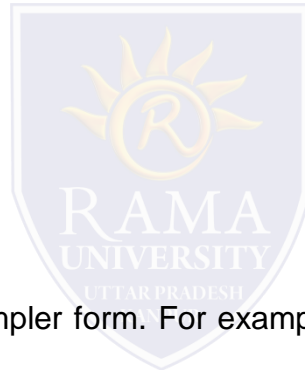
Methods of data reduction:

These are explained as following below.

- **Data Cube Aggregation:**

This technique is used to aggregate data in a simpler form. For example, imagine that information you gathered for your analysis for the years 2012 to 2014, that data includes the revenue of your company every three months.

They involve you in the annual sales, rather than the quarterly average, So we can summarize the data in such a way that the resulting data summarizes the total sales per year instead of per quarter. It summarizes the data.



# Methods of data reduction

## Dimension reduction:

Whenever we come across any data which is weakly important, then we use the attribute required for our analysis. It reduces data size as it eliminates outdated or redundant features.

### 1. Step-wise Forward Selection –

The selection begins with an empty set of attributes later on we decide best of the original attributes on the set based on their relevance to other attributes. We know it as a p-value in statistics.

Suppose there are the following attributes in the data set in which few attributes are redundant.

Initial attribute Set: {X1, X2, X3, X4, X5, X6}

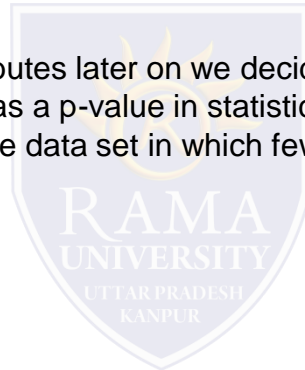
Initial reduced attribute set: { }

Step-1: {X1}

Step-2: {X1, X2}

Step-3: {X1, X2, X5}

Final reduced attribute set: {X1, X2, X5}



# Methods of data reduction

## •Step-wise Backward Selection –

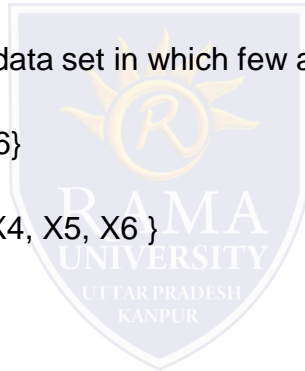
This selection starts with a set of complete attributes in the original data and at each point, it eliminates the worst remaining attribute in the set.

Suppose there are the following attributes in the data set in which few attributes are redundant.

Initial attribute Set: {X1, X2, X3, X4, X5, X6}

Initial reduced attribute set: {X1, X2, X3, X4, X5, X6}

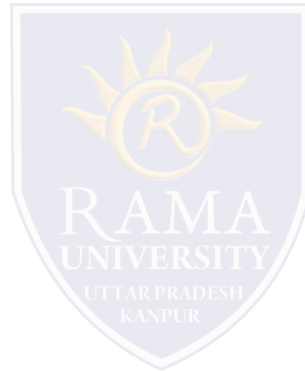
- Step-1: {X1, X2, X3, X4, X5}
- Step-2: {X1, X2, X3, X5}
- Step-3: {X1, X2, X5}
- Final reduced attribute set: {X1, X2, X5}



# Methods of data reduction

## Combination of forwarding and Backward Selection –

It allows us to remove the worst and select best attributes, saving time and making the process faster.



# Data Compression

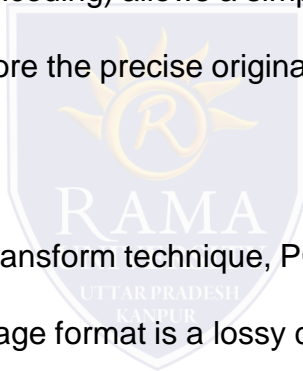
The data compression technique reduces the size of the files using different encoding mechanisms (Huffman Encoding & run-length Encoding). We can divide it into two types based on their compression techniques.

- **Lossless Compression –**

Encoding techniques (Run Length Encoding) allows a simple and minimal data size reduction. Lossless data compression uses algorithms to restore the precise original data from the compressed data.

- **Lossy Compression –**

Methods such as Discrete Wavelet transform technique, PCA (principal component analysis) are examples of this compression. For e.g., JPEG image format is a lossy compression, but we can find the meaning equivalent to the original the image. In lossy-data compression, the decompressed data may differ to the original data but are useful enough to retrieve information from them.



# Numerosity Reduction:

In this reduction technique the actual data is replaced with mathematical models or smaller representation of the data instead of actual data, it is important to only store the model parameter. Or non-parametric method such as clustering, histogram, sampling. For More Information on Numerosity Reduction Visit the link below:

- **Discretization & Concept Hierarchy Operation:**

Techniques of data discretization are used to divide the attributes of the continuous nature into data with intervals. We replace many constant values of the attributes by labels of small intervals. This means that mining results are shown in a concise, and easily understandable way.

- **Top-down discretization –**

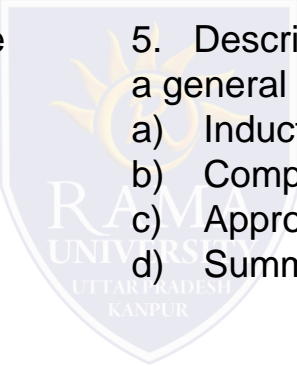
If you first consider one or a couple of points (so-called breakpoints or split points) to divide the whole set of attributes and repeat of this method up to the end, then the process is known as top-down discretization also known as splitting.

- **Bottom-up discretization –**

If you first consider all the constant values as split-points, some are discarded through a combination of the neighborhood values in the interval, that process is called bottom-up discretization.



# Multiple Choice Question

1. Various visualization techniques are used in \_\_\_\_\_ step of KDD.
    - a) selection
    - b) transformaion
    - c) data mining.
    - d) interpretation.
  
  2. Extreme values that occur infrequently are called as \_\_\_\_\_.
    - a) outliers
    - b) rare values.
    - c) dimensionality reduction.
    - d) All of the above.
  
  3. Box plot and scatter diagram techniques are \_\_\_\_\_.
    - a) Graphical
    - b) Geometric
    - c) Icon-based.
    - d) Pixel-based.
  
  4. \_\_\_\_\_ is used to proceed from very specific knowledge to more general information.
    - a) Induction
    - b) Compression.
    - c) Approximation.
    - d) Substitution.
  
  5. Describing some characteristics of a set of data by a general model is viewed as \_\_\_\_\_.
    - a) Induction
    - b) Compression
    - c) Approximation
    - d) Summarization
- 
- The watermark is a shield-shaped logo for Rama University. It features a stylized sun or flame symbol at the top, with the text 'RAMA UNIVERSITY' in the center and 'UTTAR PRADESH KANPUR' at the bottom.

# REFERENCES

- [https://www.tutorialspoint.com/dwh/dwh\\_overview.htm](https://www.tutorialspoint.com/dwh/dwh_overview.htm)
- <https://www.geeksforgeeks.org/>
- <http://myweb.sabanciuniv.edu/rdekharghani/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.