

2040233341
Data Mining
Unit 2 : Data-Preprocessing

By Rakesh Kharva

Why to pre-process data?

Section - 1

Why to pre-process data?

- ▶ Data pre-processing is a data mining technique that involves transforming raw data (real world data) into an understandable format.
- ▶ Real-world data is often incomplete, inconsistent, lacking in certain behaviors or trends and likely to contain many errors.
 - ↳ **Incomplete:** Missing attribute values, lack of certain attributes of interest, or containing only aggregate data.
 - E.g. Occupation = “ ”
 - ↳ **Noisy:** Containing errors or outliers.
 - E.g. Salary = “abcxy”
 - ↳ **Inconsistent:** Containing similarity in codes or names.
 - E.g. “Gujarat” & “Gujrat” (Common mistakes like spelling, grammar, articles)

Why to pre-process data?

No quality data, No quality
results

- ▶ It looks like Garbage In Garbage Out (GIGO).



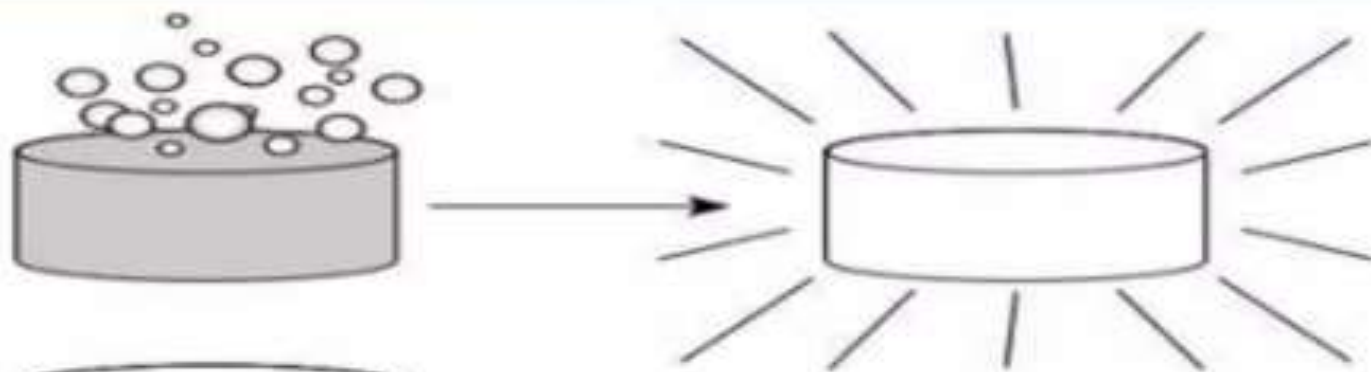
- Quality decisions must be based on quality data.
- Duplicate or missing data may cause incorrect or even misleading statistics.
- Data preprocessing **prepares** raw data for **further processing**.

Data preparation, cleaning and transformation are the **majority task (90%)** in data mining.

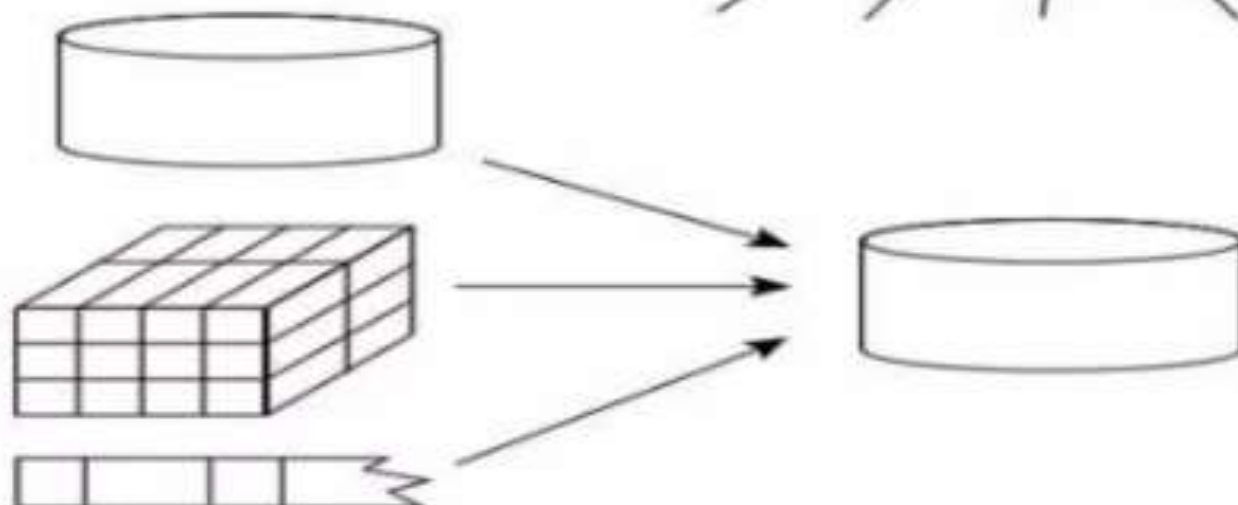
Major Task in Data-Preprocessing

- Data Cleaning :
- Data Integration:
- Data Reduction :
- Data Transformation :

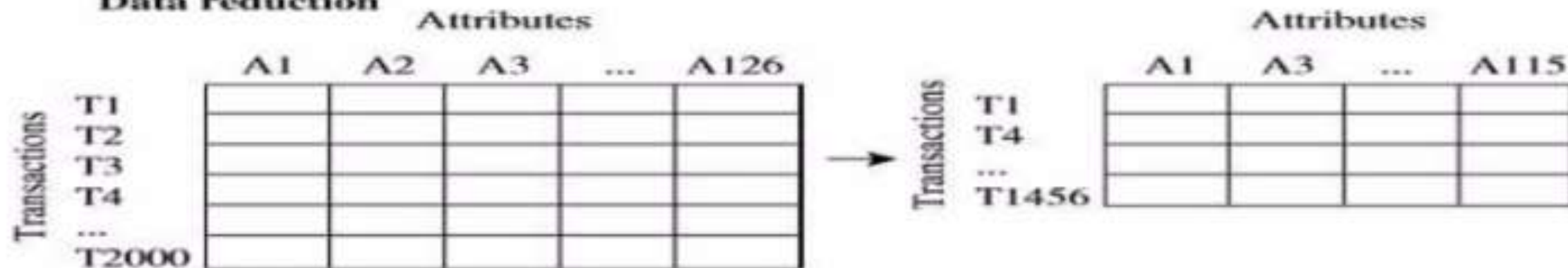
Data cleaning



Data integration



Data reduction



Data transformation

$-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

Data Cleaning

Section - 2

Data Cleaning

1. Fill in missing values

1. Ignore the tuple
2. Fill missing value manually
3. Fill in the missing value automatically
4. Use a global constant to fill in the missing value

2. Identify outliers and smooth out noisy data

1. Binning Method
2. Regression
3. Clustering

3. Correct inconsistent data

4. Resolve redundancy caused by data integration



- **Ignore the tuple (record/row):**
 - Usually done when **class label is missing**.
 - **Example**
 - The task is to distinguish between two types of emails, “spam” and “non-spam” (Ham).
 - Spam & non-spam are called as class label.
 - If an email comes to you, in which class label is missing then it is discarded.
- **Fill missing value manually:**
 - Use the **attribute mean (average)** to fill in the missing value and also use the attribute mean (average) for all samples belonging to the same class.
- **Fill in the missing value automatically:**
 - **Predict the missing value** by using a **learning algorithm**:
 - Consider the attribute with the missing value as a dependent variable and run a learning algorithm (usually Naive Bayes or Decision tree) to predict the missing value.
- **Use a global constant to fill in the missing value**
 - Replace **all missing attribute values** by the same constant such as a label like “Unknown”.

2) Identify outliers and smooth out noisy data

There are three data smoothing techniques as follows..

1. **Binning :**

↪ Binning methods smooth a sorted data value by consulting its “neighborhood” that is, the values around it.

2. **Regression :**

↪ It conforms data values to a function.

↪ Linear regression involves finding the “best” line to fit two attributes (or variables) so that one attribute can be used to predict the other.

3. **Outlier analysis :**

↪ Outliers may be detected by clustering for example, where similar values are organized into groups or “clusters”.

↪ In this, values that fall outside of the set of clusters may be considered as outliers.

1. Binning Method

- ▶ Binning method is a top-down splitting technique based on a specified number of bins.
- ▶ In this method the data is first sorted and then the sorted values are distributed into a number of buckets or bins.
- ▶ For example, attribute values can be discretized (separated) by applying equal-width or equal-frequency binning, and then replacing each value by the bin mean, median or boundaries.
- ▶ It can be applied recursively to the resulting partitions to generate concept hierarchies.
- ▶ It does not use class information, therefore it is called as unsupervised discretization technique.
- ▶ It used to minimize the effects of small observation errors.



Identify outliers and smooth out noisy data

There are basically two types of binning approaches..

1. **Equal width (or distance) binning :**

- The simplest binning approach is to partition the range of the variable into k equal-width intervals.
- The interval width is simply the range [Min, Max] of the variable divided by N,
- $\text{Width} = \text{Max} - \text{Min} / N$ (Number of Bins)

► Example

- **Data:** 5,10,11,13,15, 35, 50, 55, 72, 92, 204, 215
- As per above formula we have Max=215, Min=5, Number of Bins=3
 - $70+5=75$ (from 5 to 75) = **Bin 1:** 5,10,11,13,15, 35, 50, 55, 72
 - $70+75=145$ (from 75 to 145) = **Bin 2:** 92
 - $70+145=215$ (from 145 to 215) = **Bin 3:** 204, 215

2. **Equal depth (or frequency) binning :**

- In equal-frequency binning we divide the range [Max, Min] of the variable into intervals that contain (approximately) **equal number of points**; equal frequency may not be possible due to repeated values.

Binning Method Example – {Bin Means} Data Cleaning

▶ Given data: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

▶ Step: 1

▶ Partition into **equal-depth** [n=4]:

Bin 1: 4, 8, 9, 15

Bin 2: 21, 21, 24, 25

Bin 3: 26, 28, 29, 34

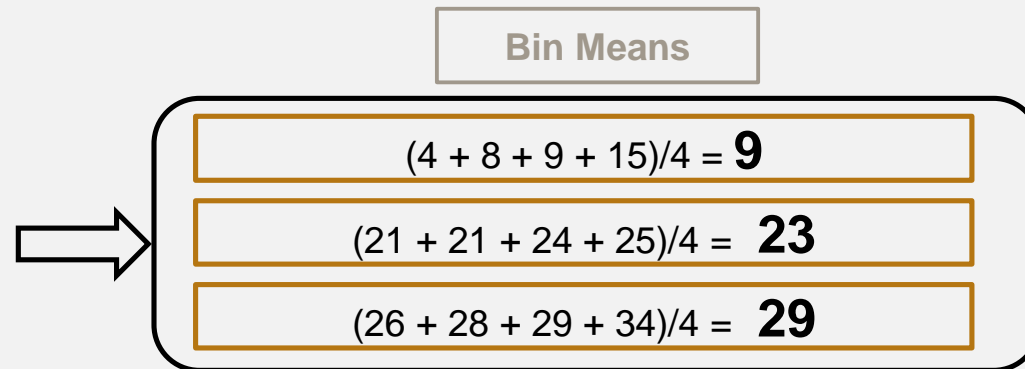
▪ Step: 2

- Smoothing by **bin means**:

Bin 1: 9, 9, 9, 9

Bin 2: 23, 23, 23, 23

Bin 3: 29, 29, 29, 29



Example – {Bin Boundaries}

▶ Given data: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

▶ Step: 1

▶ Partition into **equal-depth** [n=4]:

Bin 1: 4, 8, 9, 15

Bin 2: 21, 21, 24, 25

Bin 3: 26, 28, 29, 34

▪ Step: 2

• Smoothing by **bin boundaries**:

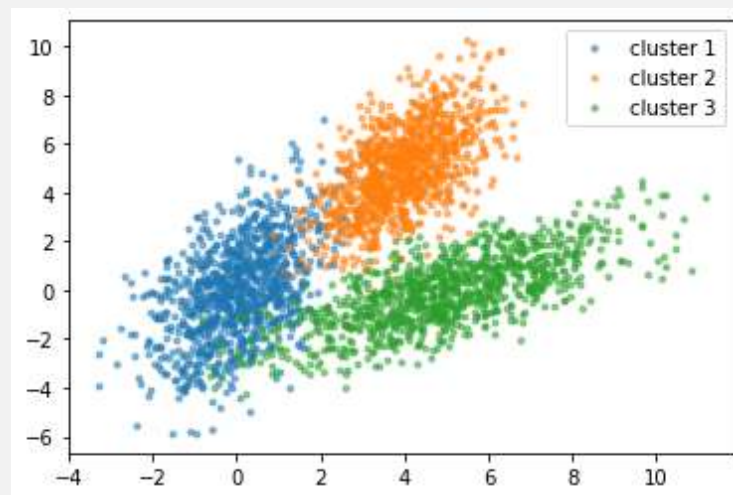
Bin 1: 4, 4, 4, 15

Bin 2: 21, 21, 25, 25

Bin 3: 26, 26, 26, 34

- ▶ Data smoothing can also be done by regression, a technique that conforms data values to a function.
- ▶ Regression analysis is a way to find trends in data & it is also called as mathematically describes the relationship between independent variables and the dependent variable.
- ▶ It can be divided into two categories..
 1. **Linear regression :**
 - It involves finding the “best” line to fit two attributes (or variables) so that one attribute can be used to predict the other.
 - In this, analysis on a single x variable for each dependent “y” variable. For example: (x_1, Y_1) .
 2. **Multiple linear regression :**
 - An extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.
 - It uses multiple “x” variables for each independent variable: $(x1)_1, (x2)_1, (x3)_1, Y_1$.

- ▶ **Cluster analysis** or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups (**clusters**).
- ▶ Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure.
- ▶ It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties.



3. Correct Inconsistent Data

- With larger datasets, it can be difficult to find all of the inconsistencies.
- It contains similarity in codes or names.
- We can manually solve common mistakes like spelling, grammar, articles or use other tools for it.

4. Resolve redundancy caused by data integration

- Data redundancy occurs in database systems **which have a field that is repeated in two or more tables.**
- When customer data is duplicated and attached with each product bought, then redundancy of data is known as **inconsistency.**
- So, the entity "customer" **might appear with different values.**
- Database **normalization** prevents redundancy and makes the best possible usage of storage.
- The proper use of **foreign keys** can minimize data redundancy and reduce the chance of destructive anomalies appearing.

Data Cleaning as a Process

The data cleaning method for data mining is demonstrated in the subsequent sections.

Monitoring the errors: Keep track of the areas where errors seem to occur most frequently. It will be simpler to identify and maintain inaccurate or corrupt information

Standardize the mining process: To help lower the likelihood of duplicity, standardize the place of insertion.

Validate data accuracy: Analyse the data and spend money on data cleaning software

Scrub for duplicate data: To save time when analyzing data, find duplicates.

Research on data: Our data needs to be vetted, standardized, and duplicate-checked before this action. There are numerous third-party sources, and these vetted and approved sources can extract data straight from our databases.

Communicate with the team: Keeping the group informed will help with client development and strengthening as well as giving more focused information to potential clients.

Data Integration

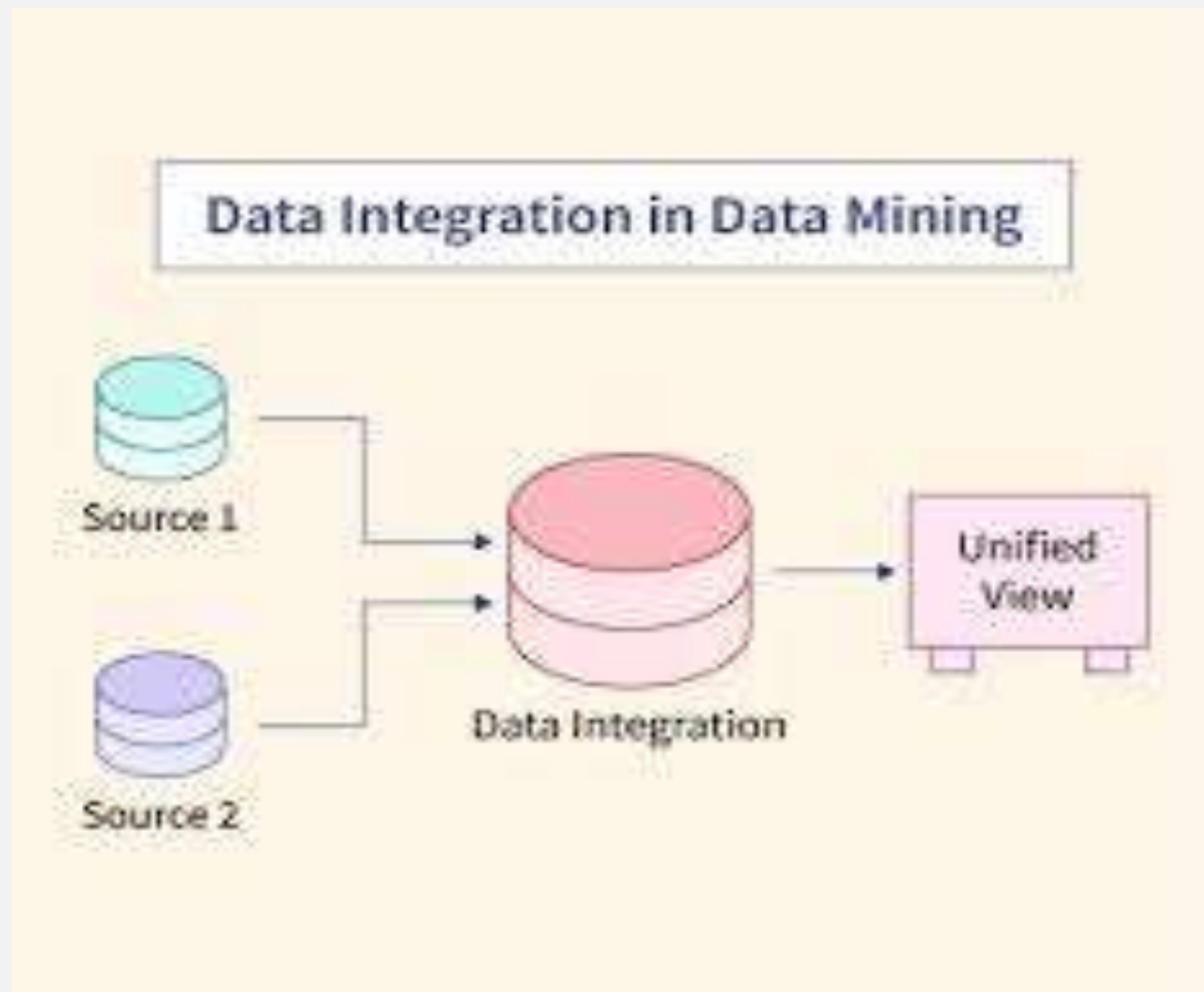
Section - 3

Data Integration

- ▶ Merging of data collected from multiple sources.
- ▶ Careful Integration can help reduce redundancies and inconsistencies in the resulting dataset.

- ▶ **Approaches in Data Integration :**

- ▶ 1. Entity Identification problem
- ▶ 2. Redundancy and Correlation analysis
- ▶ 3. Tuple Duplication
- ▶ 4. Data Value conflict Detection and Resolution



Data Integration

Entity identification problem

- ▶ Schema Integration and Object Matching are very important issues in Data Integration
- ▶ **Schema integration:** e.g., cust-id ,customer_id, cust_no,etc
 - ↳ Handling blank ,zero null values.
- ▶ **Object Matching:** Matching in structure of the data
 - ↳ e.g., Discount Issues, Currency type



Data Integration

Redundancy and Correlation Analysis

- ▶ **Redundancy :** An Attribute may be redundant if it can be “derived” from another attribute or set of attributes.

Ex – DOB, Age

Quarter Sales ,Year Sales

Name	DOB	Age

- ▶ **Correlation Analysis:** Given two attributes ,such analysis can measure how strongly one attribute implies the other, based on the available data.

Branch Id	Quarter total	Year Total

Data Integration

Tuple Duplication

- ▶ The use of denormalized tables (often done to improve performance by avoiding joins) is another source of data redundancy.
- ▶ Inconsistencies often arise between various duplicates, due to inaccurate data entry or updating some but not all data occurrences

Name	DOB	Branch	Occupation	Address
A	25	Tpg	Govt	Tpg
B	30	Tnk	Govt	Rjy
A	25	Tpg	Private	Tpg
C	30	tnk	Private	Rjy

Data Integration

Data Value conflict Detection and Resolution

- ▶ Attribute Values from different sources may differ. This may be due to differences in representation, scaling or encoding.
 - ↳ **Ex-** school curriculum (grading system)
- ▶ Attributes may also differ on the abstraction level. Where an attribute in one system is recorded at, say, a lower abstraction level than the “same” attribute in another
 - ↳ **Ex** – Monthly total sales in a store
 - ↳ & Monthly total sales from all stores in the region.

Data Reduction

Section - 4

Data Reduction

► Why Data Reduction?

- ↳ A database/data warehouse may store terabytes of data.
- ↳ Complex data analysis may take a very long time to run on the complete data set.

► What is Data Reduction?

- ↳ Data reduction process reduces the size of data and makes it suitable and feasible for analysis.
- ↳ In the reduction process, integrity of the data must be preserved and data volume is reduced.
- ↳ There are many techniques that can be used for data reduction like
 1. **Dimensionality reduction**
 2. **Numerosity reduction**
 3. **Data compression**

2. Numerosity Reduction

Data Reduction

- ▶ Numerosity Reduction is a data reduction technique which replaces the original data by smaller form of data representation.
- ▶ There are two techniques for numerosity reduction- **Parametric** and **Non-Parametric** methods.
- ▶ **Parametric Methods**
 - ↪ For parametric methods, data is represented using some model.
 - ↪ The model is used to estimate the data, so that only parameters of data are required to be stored, instead of actual data.
 - ↪ Regression and Log-Linear methods are used for creating such models.
- ▶ **Non-Parametric Methods**
 - ↪ These methods are used for storing reduced representations of the data include **histograms, clustering, sampling** and **data cube aggregation**.

► Histograms

- ↪ Histogram is the data representation in terms of frequency.
- ↪ It uses binning to approximate data distribution and is a popular form of data reduction.

► Clustering

- ↪ Clustering divides the data into groups/clusters, it partitions the whole data into different clusters.
- ↪ In data reduction, the cluster representation of the data are used to replace the actual data, It also helps to detect outliers in data.

► Sampling

- ↪ Sampling can be used for data reduction because it allows a large data set to be represented by a much smaller random data sample (or subset).

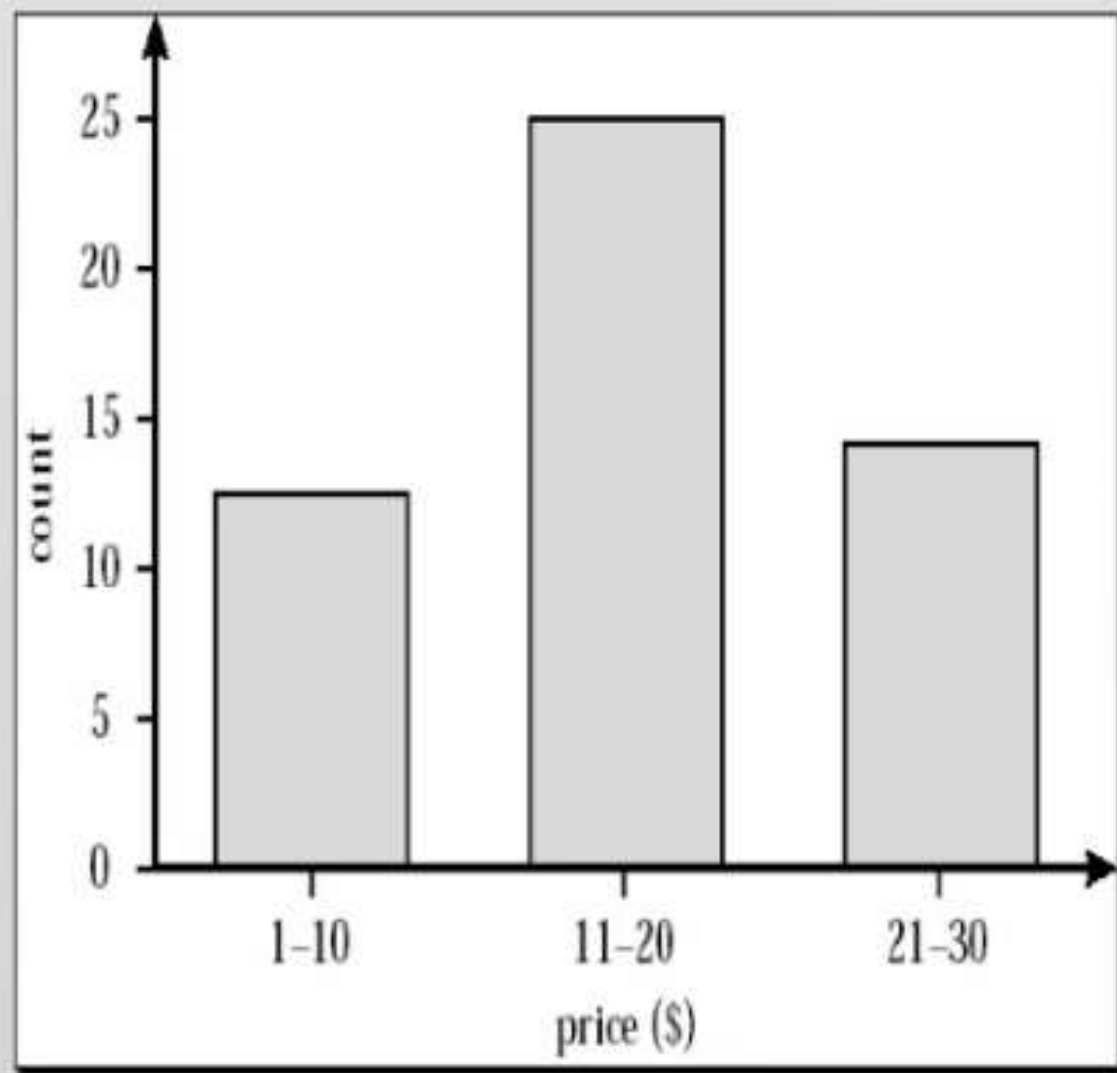
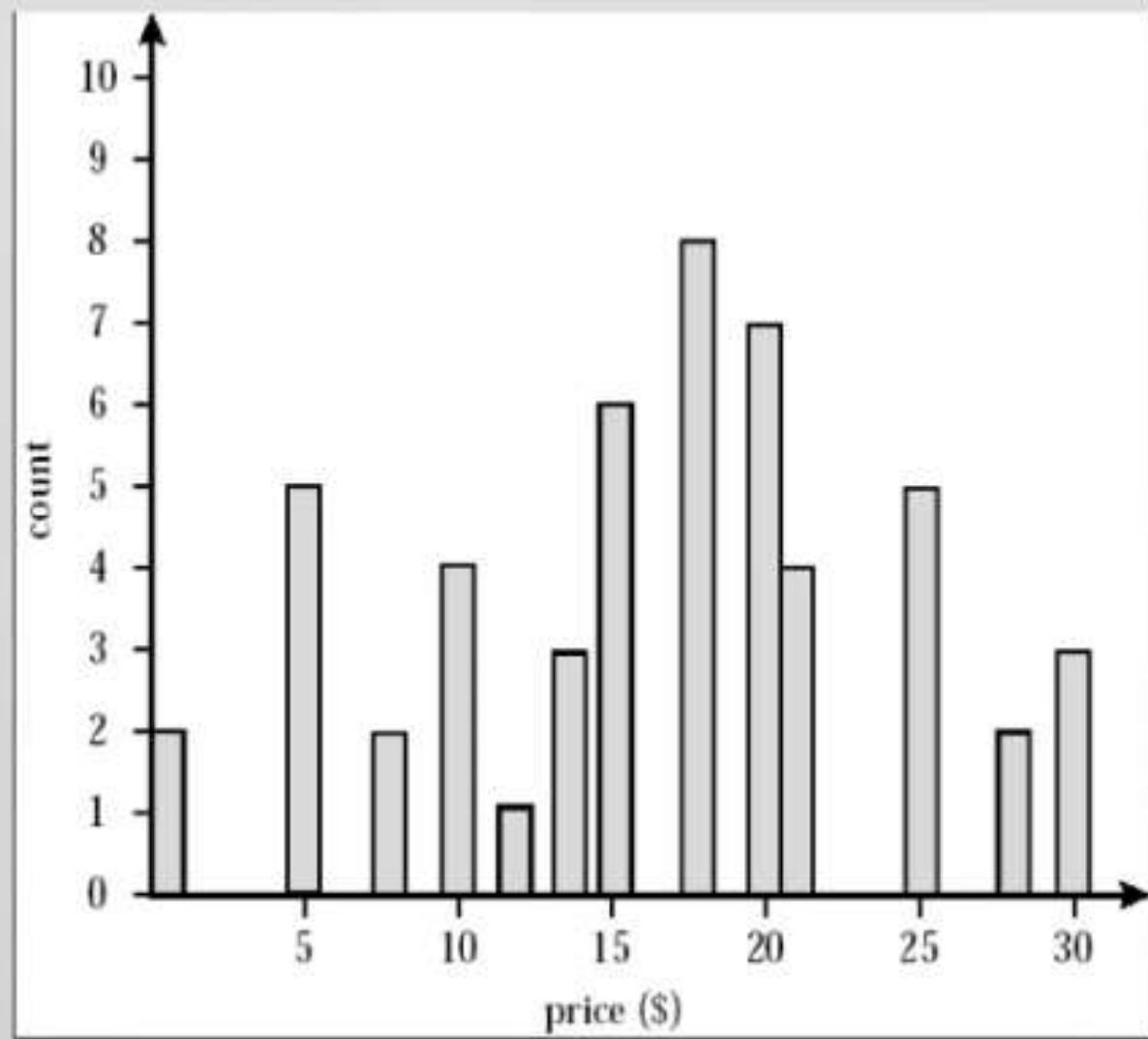
► Data Cube Aggregation

- ↪ Data cube aggregation involves moving the data from detailed level to a fewer number of dimensions.
- ↪ The resulting data set is smaller in volume, without loss of information necessary for the analysis task.

Histograms. The following data are a list of *AllElectronics* prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

Figure 3.7 shows a histogram for the data using singleton buckets. To further reduce the data, it is common to have each bucket denote a continuous value range for the given attribute. In Figure 3.8, each bucket represents a different \$10 range for *price*. ■

Histograms

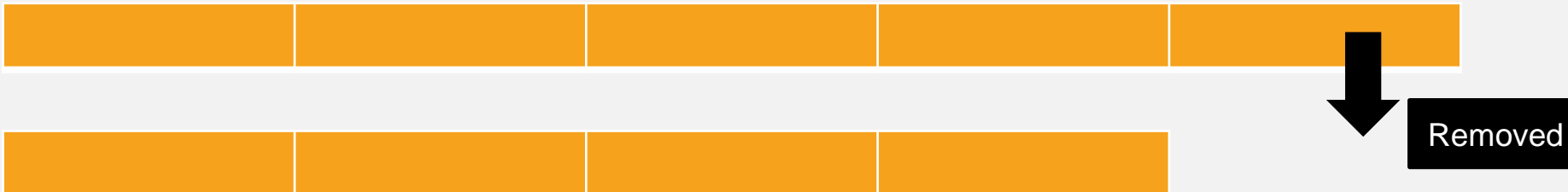


► 1) SRS (Simple Random Sample)

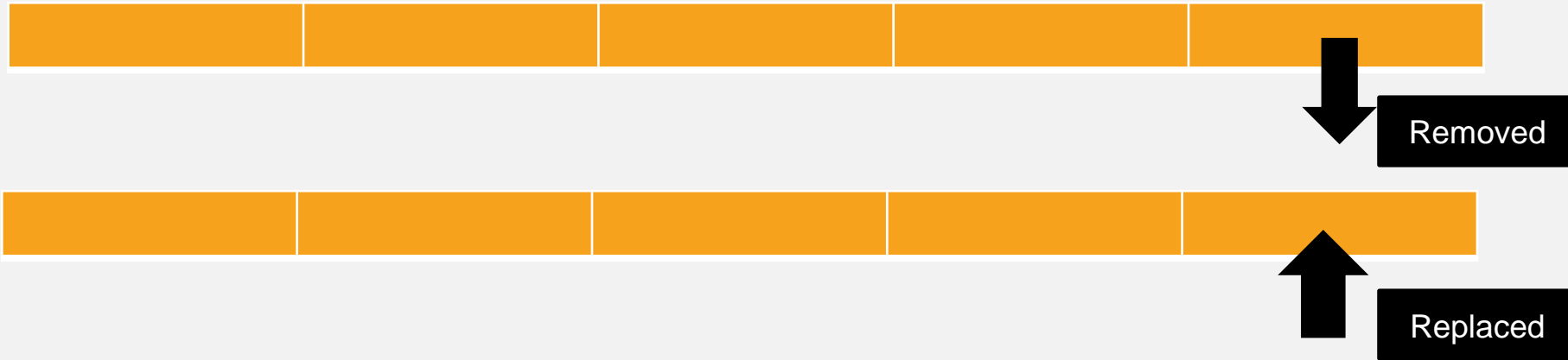


Large DataSet Divided in Equal Partition for taking Sample

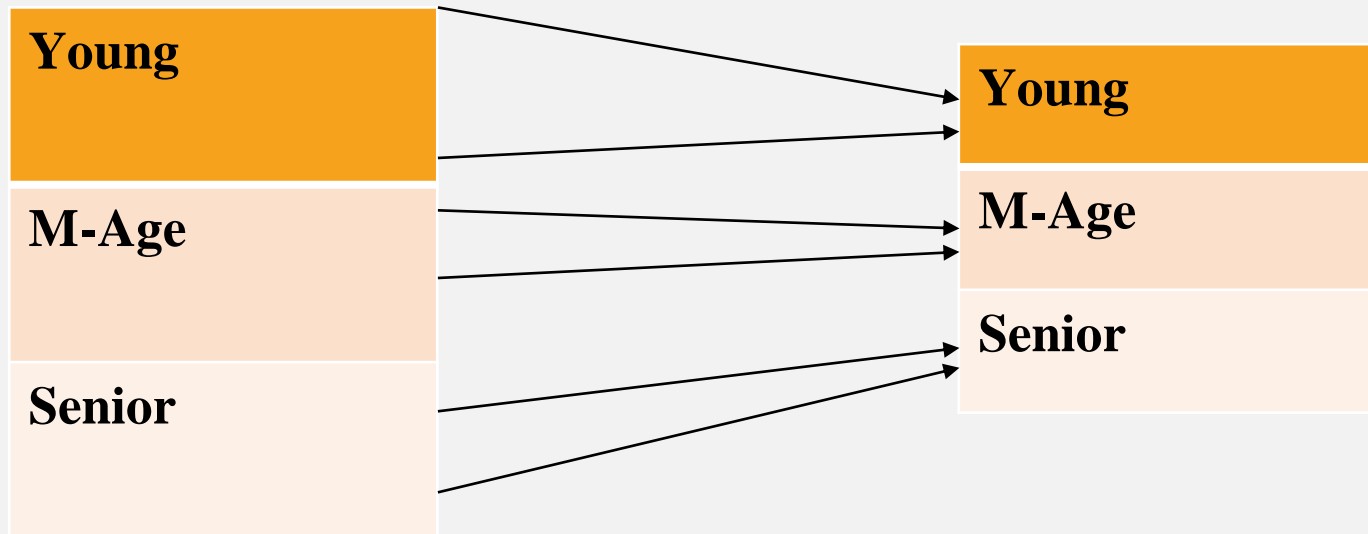
► 2) SRSWOR (Simple Random Sampling Without Replacement)



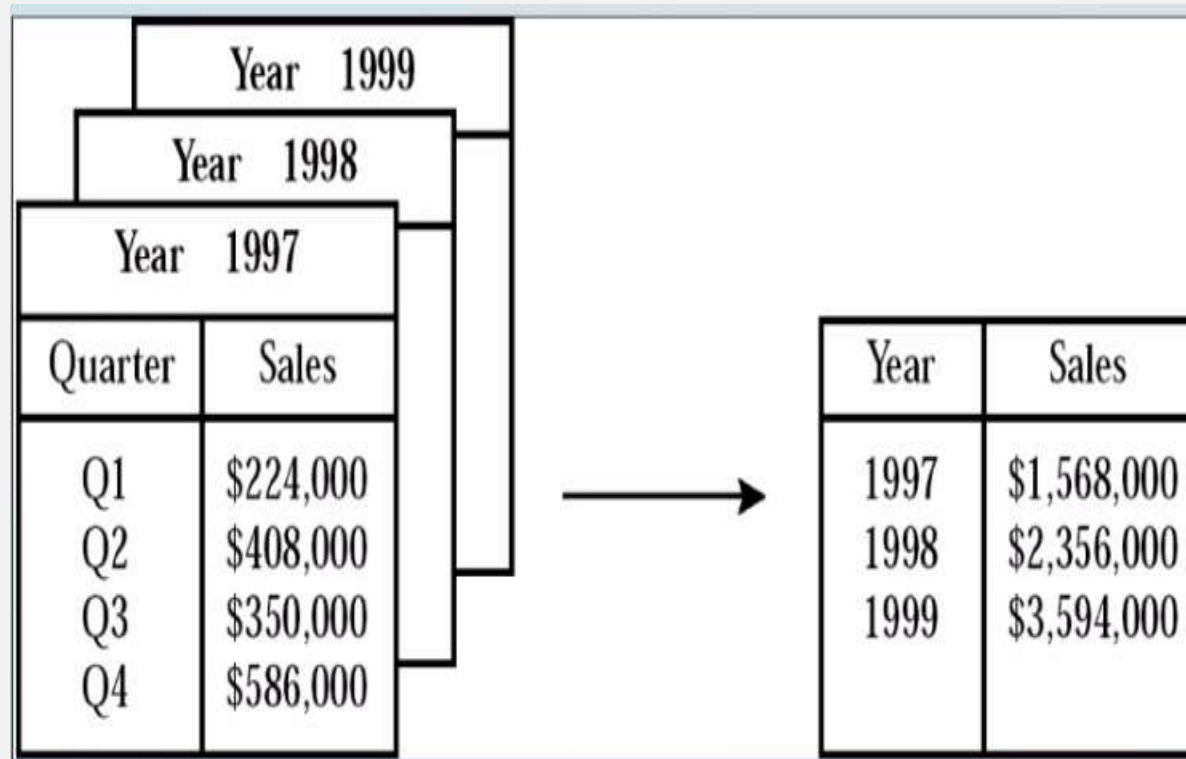
► 3) SRSWOR (Simple Random Sampling With Replacement)



4) Stratified Sampling



- Data cube aggregation involves moving the data from detailed level to a fewer number of dimensions.
- The resulting data set is smaller in volume, without loss of information necessary for the analysis task.



Year 1999	
Year 1998	
Year 1997	
Quarter	Sales
Q1	\$224,000
Q2	\$408,000
Q3	\$350,000
Q4	\$586,000

Year	Sales
1997	\$1,568,000
1998	\$2,356,000
1999	\$3,594,000

Thank You