Introduction to Data Mining

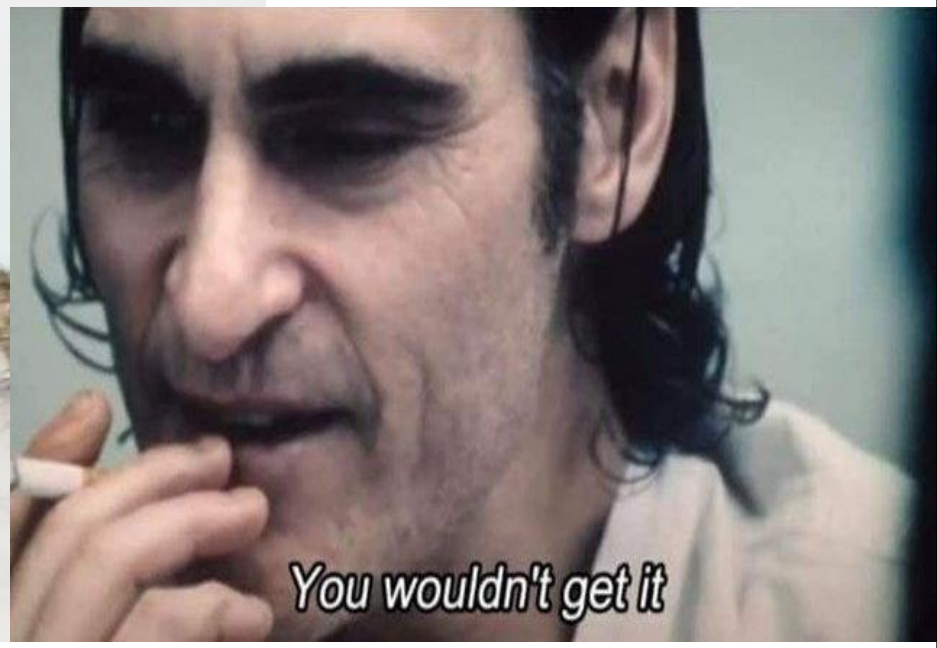
acknowledgements

- These slides has been taken from (as it is or with some modifications) for lots of resources such as -
  - Data Mining book by Han and Kamber
  - Data Mining Slides of Prof. Panayiotis Tsaparas, University of Ioannina

grading policy

- **Quizzes:** 20 **marks**
  - 4 Quizzes will be conducted, 5 marks each
  - No make-up quiz, If someone misses a quiz, he/she has to give viva
- **Case Study/ Programming Assignment: 10 marks**
  - Can be done in group of at max 3 people.
- **Midterm: 20 marks**
- **Endterm: 50 marks**

When they asked you to clean the data and you cleaned all of it.



What is data mining?

- After years of data mining there is still no unique answer to this question.

- One definition:

> Data mining is the process of discovery interesting patterns, models, and other kinds of knowledge in large data sets.

- Also known as *Knowledge Discovery from Data (KDD), Pattern Discovery, Knowledge Extraction, etc.*

Why do we need data mining?

- Really, really huge amounts of raw data!!
  - In the digital age, TeraBytes (10^{12} bytes) of data is generated every second
    - Mobile devices, digital photographs, web documents.
    - Facebook updates, Tweets, Blogs, User-generated content
    - Transactions, sensor data, surveillance data
    - Queries, clicks, browsing
  - Cheap storage has made possible to maintain this data
- Need to analyze the raw data to extract knowledge.

Why do we need data mining?

- Large amounts of data can be more powerful than complex algorithms and models
  - Google has solved many Natural Language Processing problems, simply by looking at the data
  - Example: misspellings, synonyms
- Data is power!
  - Today, the collected data is one of the biggest assets of an online company
    - Query logs of Google
    - The friendship and updates of Facebook
    - Tweets and follows of Twitter
    - Amazon transactions
- We need a way to harness the collective intelligence

Example: transactions data

- UPI transactions in 2024 in India -

Month	No. of Banks live on UPI	Volume (in Mn)	Value (in Cr.)
Jun-24	602	13,885.14	20,07,081.16
May-24	598	14,035.84	20,44,937.05
Apr-24	581	13,303.99	19,64,464.52
Mar-24	572	13,440.00	19,78,353.23
Feb-24	560	12,102.67	18,27,869.35
Jan-24	550	12,203.02	18,41,083.97

Example: Social Media Data

- As of April 2024, there are an estimated active **5.07 billion global social media users**, that's more than 60% of the world population.
- Instagram has over **2 billion monthly active users**. With India having the most **Instagram users with 363 million**.
- AI-recommended reels make people spend **24% more time on Instagram**.
- On average, a person spends 35 minutes a day on Facebook.
- Users worldwide spend approximately **28 hours each month using the YouTube mobile app**.
- Over 70% of Youtube users say that **YouTube makes them more aware of new brands**.

Example: document data

- Web as a document repository: estimated 50 billions of web pages
- Wikipedia: 4 million articles (and counting)
- Online news portals: steady stream of 100’s of new articles every day
- Twitter: ~300 million tweets every day

Example: network data

- Web: 50 billion pages linked via hyperlinks
- Facebook: 500 million users
- Twitter: 300 million users
- Instant messenger: ~1billion users
- Blogs: 250 million blogs worldwide, presidential candidates run blogs

Example: genomic sequences

- http://www.1000genomes.org/page.php
- Full sequence of 1000 individuals
- $3*10^9$ nucleotides per person → $3*10^{12}$ nucleotides
- Lots more data in fact: medical history of the persons, gene expression data

Example: environmental data

- Climate data (just an example)

http://www.ncdc.gov/oa/climate/ghcn-monthly/index.php

- “a database of temperature, precipitation and pressure records managed by the National Climatic Data Center, Arizona State University and the Carbon Dioxide Information Analysis Center”

- “6000 temperature stations, 7500 precipitation stations, 2000 pressure stations”
  - Spatiotemporal data

Behavioral data

- Mobile phones today record a large amount of information about the user behavior
  - GPS records position
  - Camera produces images
  - Communication via phone and SMS
  - Text via facebook updates
  - Association with entities via check-ins
- Amazon collects all the items that you browsed, placed into your basket, read reviews about, purchased.
- Google and Bing record all your browsing activity via toolbar plugins. They also record the queries you asked, the pages you saw and the clicks you did.
- Data collected for millions of users on a daily basis

What can you do with the data?

- Suppose that you are the owner of a supermarket and you have collected billions of market basket data. What information would you extract from it and how would you use it?

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Product placement

Catalog creation

Recommendations

- What if this was an online store?

What can you do with the data?

- Suppose you are a search engine and you have a toolbar logs consisting of
  - pages browsed,
  - queries,
  - pages clicked,
  - ads clicked

Ad click prediction

Query reformulations

each with a user id and a timestamp. What information would you like to get out of the data?

What can you do with the data?

- Suppose you are a stock broker and you observe the fluctuations of multiple stocks over time. What information would you like to get out of your data?

What can you do with the data?

- You are the owner of a social network, and you have full access to the social graph, what kind of information do you want to get out of your graph?

- Who is the most important node in the graph?
- What is the shortest path between two nodes?
- How many friends two nodes have in common?
- How does information spread on the network?

different types of data

- As a general technology, data mining can be applied to any kind of data as long as the data is meaningful for our purpose.
- Data can be available in many forms such as -
  - **Structured vs. Semi-Structured vs. Unstructured data**
  - **Stored vs. Streaming data**

structured data

- Structured data is highly organized and easily searchable.
- It follows a strict schema, meaning each data point is consistent in format and can be easily queried.
- Structured data is table-like structure with a fixed set of attributes (or fields, columns), each with a fixed set of value ranges and semantic meaning.
- Structured data is stored in relational databases (SQL Databases), Spreadsheets (Excel or Google Sheets), CSV Files, etc.

structured data

id [PK] integer	country character varying (45)	rice double precision	wheat double precision
1	Australia	0.42	31.9
2	Brazil	13.98	7.9
3	China	212.84	136.9
4	Ethiopia	0.2	5.2
5	India	195.43	109.6

semi-structured data

- Semi-structured data does not have a strict structure but still contains tags or markers to separate semantic elements.
- This data type is more flexible than structured data but still provides a degree of organization.
- Semi-structured is stored in XML Files, JSON Files, NoSQL Databases like MongoDB, etc.

semi-structured data

JSON Format

```
{
    "userid": 1,
    "Name": "Bob",
    "Age": 24,
    "Email": "bob@example.com"
}
```

XML Format

```
<user>
    <UserID>1</UserID>
    <Name>Bob</Name>
    <Age>24</Age>
    <Email>bob@example.com</Email>
</user>
```

unstructured data

- Unstructured data lacks a predefined data model or structure, making it more difficult to process, and analyze.
- It encompasses a wide variety of data types that do not fit neatly into structured or semi-structured categories.
- For example -
  - text data like emails, messages
  - media files like images, videos, audios

stored vs. streaming data

- Usually, we perform data mining on stored data sets which are stored in large data repositories or databases.
- But for some applications such as video surveillance data may come in streaming manner.
- Mining stream data may require different set of methods compared to the stored data.

Mining various kinds of knowledge

- Different kinds of patterns and knowledge can be uncovered via data mining. In general, data mining tasks can be put into two categories - *descriptive data mining* and *predictive data mining*
- **Descriptive Data Mining -** In the descriptive data mining, the goal is to derive patterns (like clusters, trends, anomalies) that summarize the underlying relationships in data.
- **Predictive Data Mining -** In the predictive data mining, the goal is to predict the value of a particular attribute based on the values of other attributes of the data. The attribute to be predicted is commonly known as the *target or dependent variable*, while the attributes used for making the prediction are known as the *independent variables*.

mining various kinds of knowledge

- **Descriptive Data Mining Tasks -**
  - Data Summarization
  - Mining Frequent Patterns and Association Rules
  - Clustering
  - Anomaly Detection
- **Predictive Data Mining Tasks -**
  - Regression
  - Classification

data summarization

- As the name suggests, Data Summarization is task of summarizing a large dataset.
- Data summarization includes tasks like -
  - Generating summary statistics (mean, median, mode) of data.
  - Creating data visualizations like histograms and pie charts.
  - Producing reports that summarize key insights from data.

Frequent Itemsets and Association Rules

- Given a set of records each of which contain some number of items from a given collection;
  - Identify sets of items (itemsets) occurring frequently together
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Itemsets Discovered:
{Milk,Coke}
{Diaper, Milk}

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Frequent Itemsets: Applications

- Product recommendations

Frequently bought together

This item: The Kalki Trilogy (Set of 3 Books) - Avatar of Vishnu; Eye of Brahma; Sword of Shiva

₹729.00

The Shiva Triology Boxset of 3 Books (Perfect Gift for this Festive Season) : The Immortals…

₹777.00

The Ram Chandra Series Boxset: Boxset of 4 Books (Ram - Scion of Ikshvaku, Sita : Warrior of Mithi…

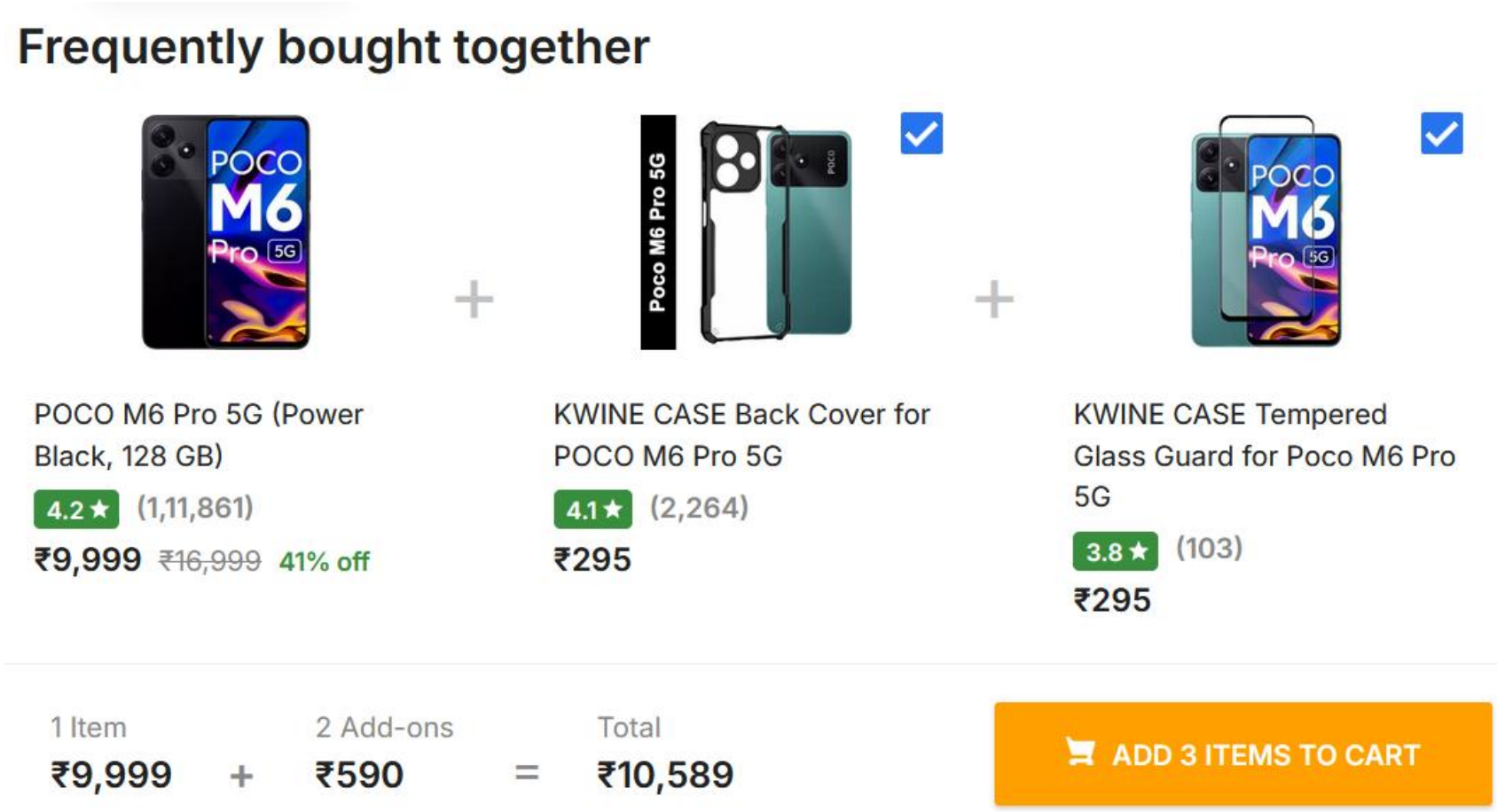
₹966.00

Total price: ₹2,472.00

Add all 3 to Cart

Frequent Itemsets: Applications

- Product recommendations

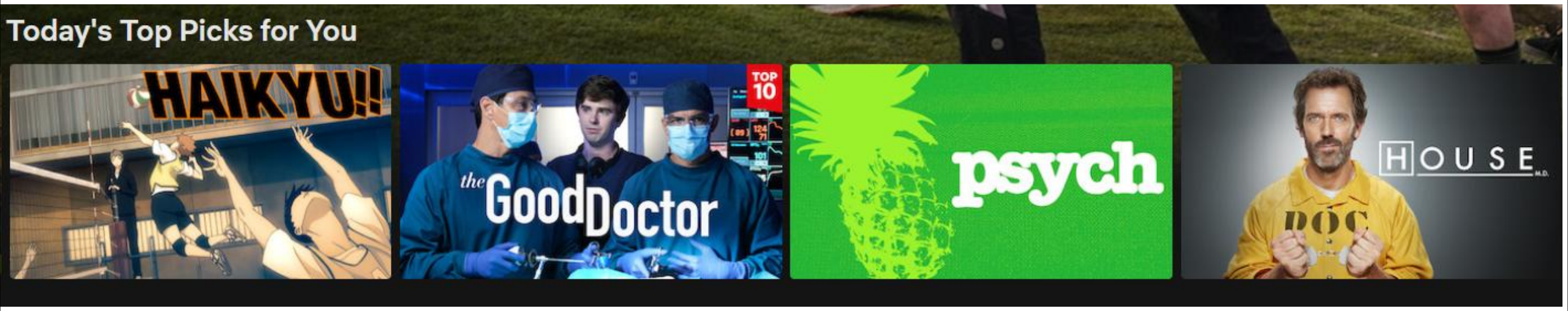


Association Rule Discovery: Application

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don’t be surprised if you find six-packs stacked next to diapers!

Association Rule Discovery: Application

- Movies recommendation



Clustering Definition

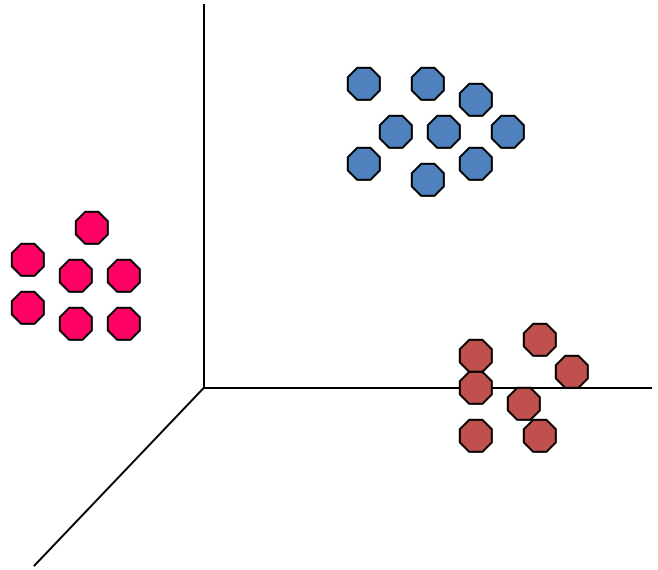
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures?
  - Euclidean Distance if attributes are continuous.
  - Other Problem-specific Measures.

Illustrating Clustering

Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized

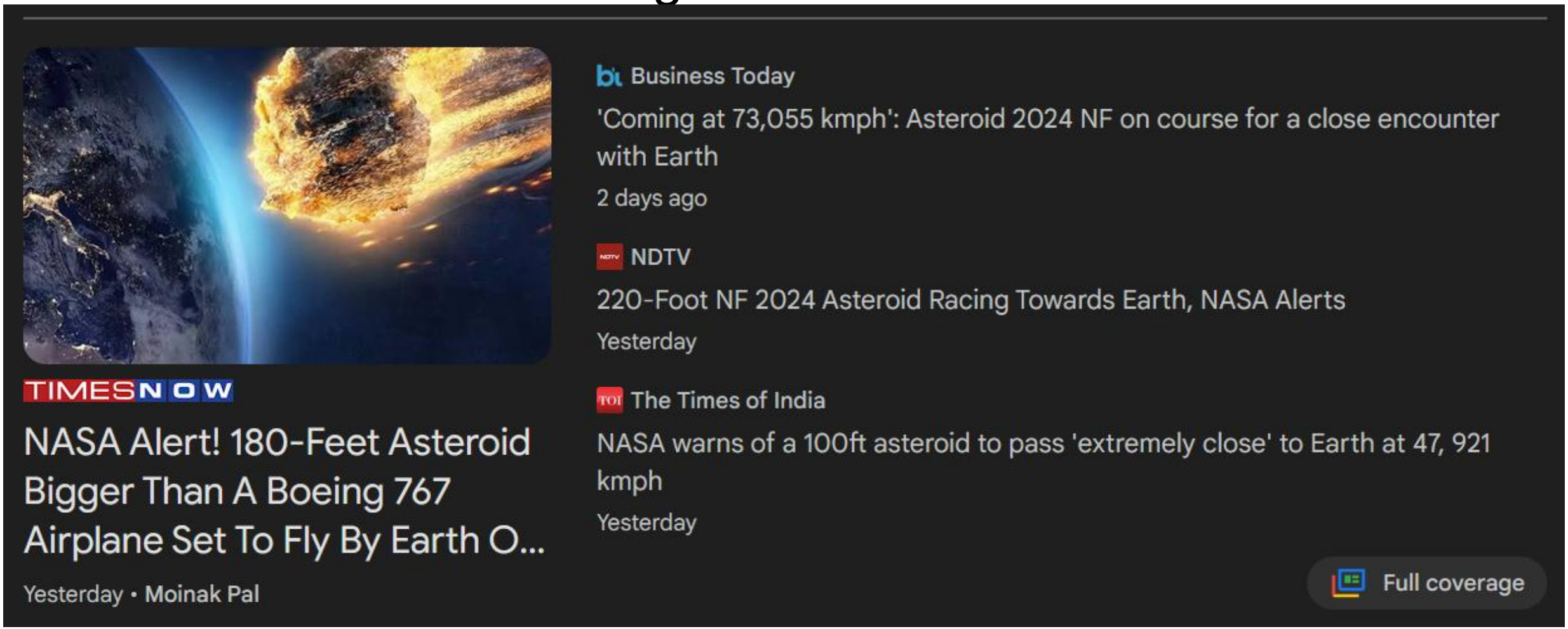


Clustering: Application

- Document Clustering:
  - Goal: To find **groups of documents that are similar to each** other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

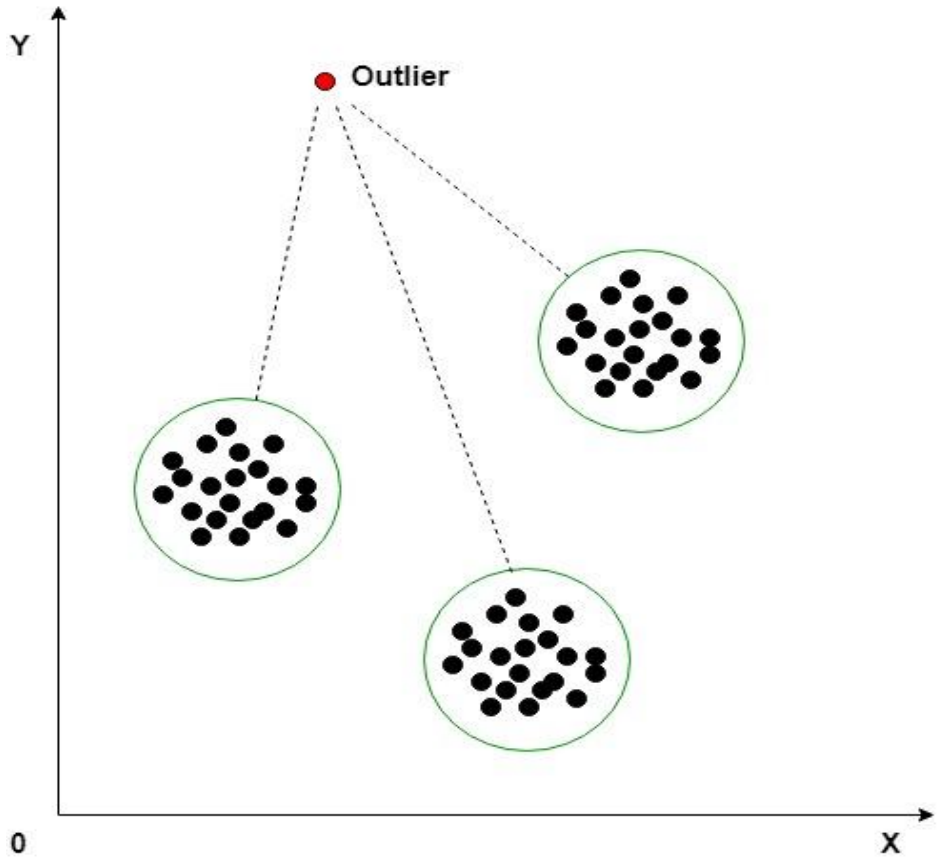
Clustering: Application

- News articles clustering



Outlier Analysis

- A dataset may contain certain objects that do not comply with the general behaviour or model of the data. These data objects are called **outliers.**
- Many data mining methods discard outliers as noise or exceptions. However, in some applications (e.g., fraud detection) the rare events can be more interesting than the more regularly occurring ones.



Classification: Definition

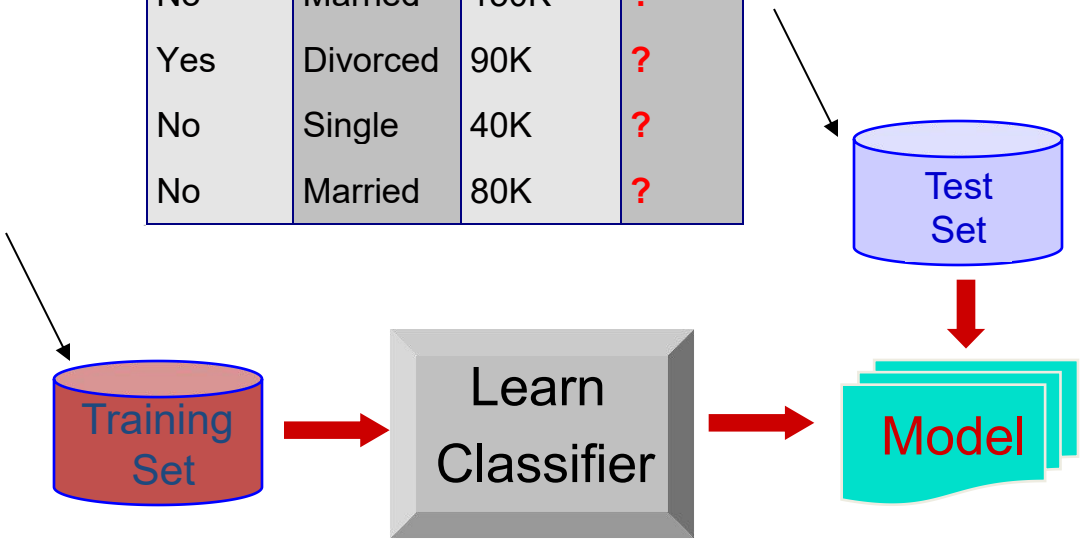
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.

- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

Classification Example

	categorical	categorical	continuous	class
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**

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Application 1

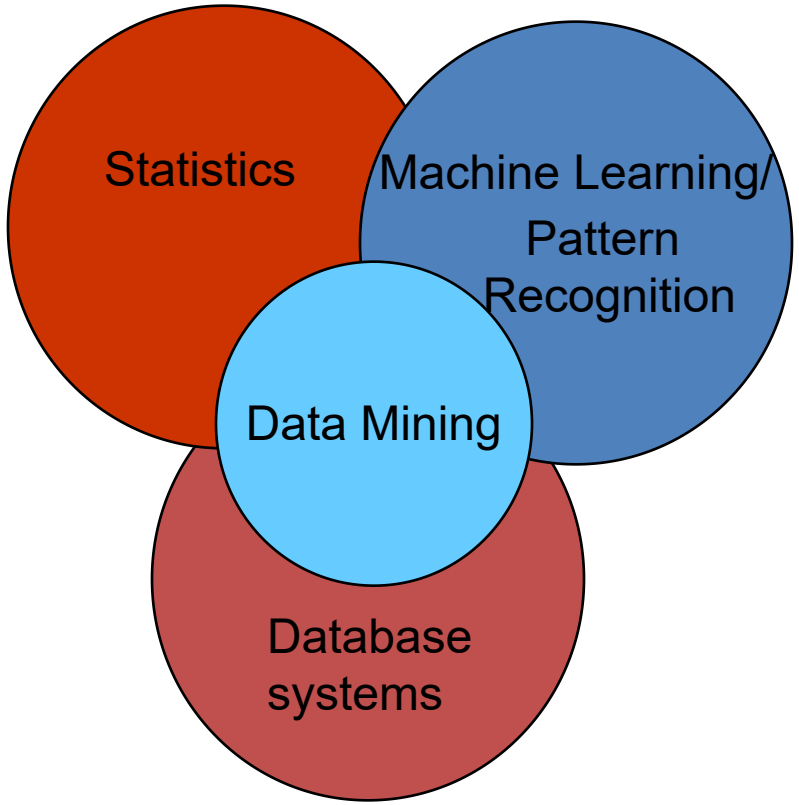
- Ad Click Prediction
  - Goal: Predict if a user that visits a web page will click on a displayed ad. Use it to target users with high click probability.
  - Approach:
    - Collect data for users over a period of time and record who clicks and who does not. The {click, no click} information forms the class attribute.
    - Use the history of the user (web pages browsed, queries issued) as the features.
    - Learn a classifier model and test on new users.

Classification: Application 2

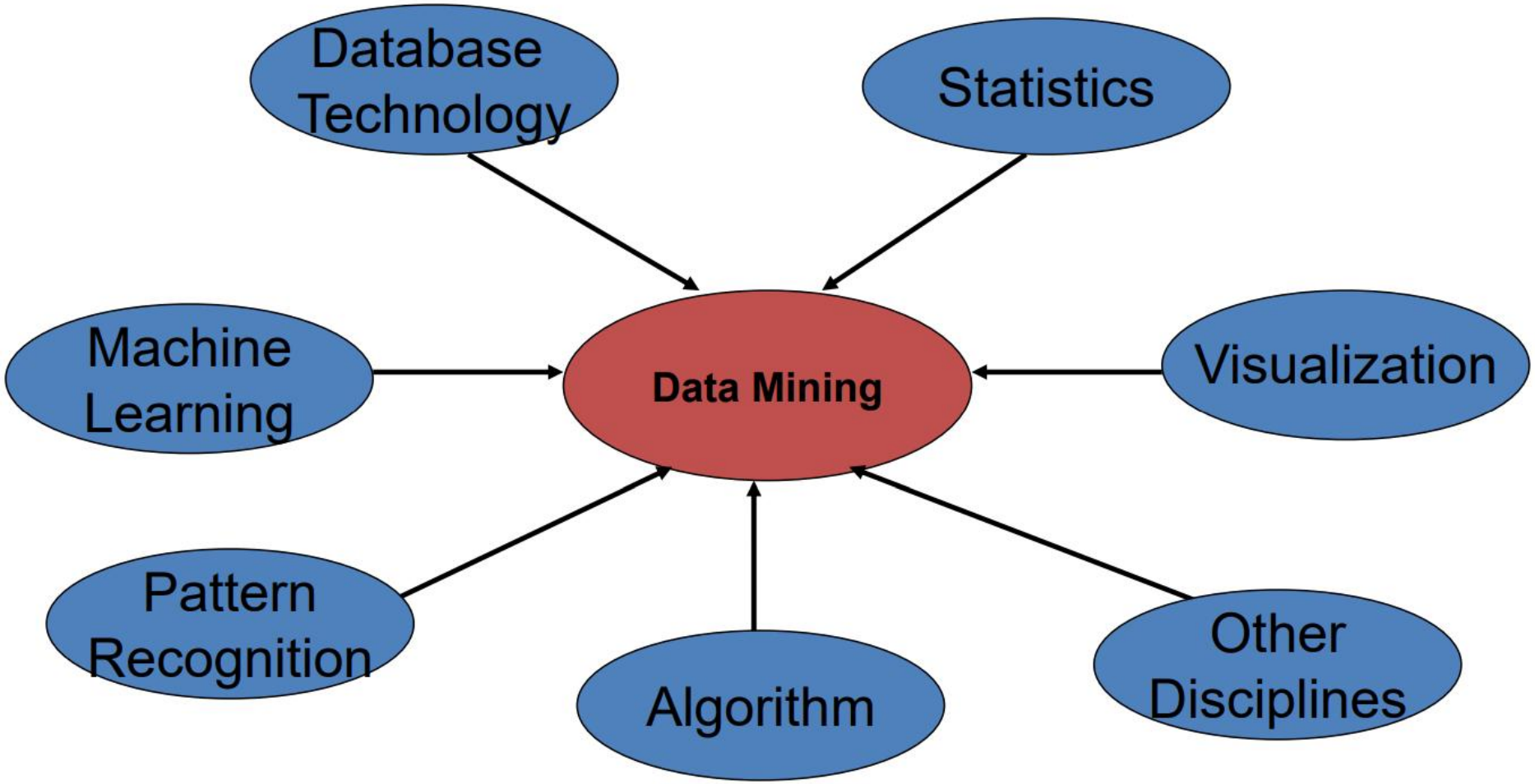
- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
    - When does a customer buy, what does he buy, how often he pays on time, etc
    - Label past transactions as fraud or fair transactions. This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

Connections of Data Mining with other areas

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems



Connections of Data Mining with other areas



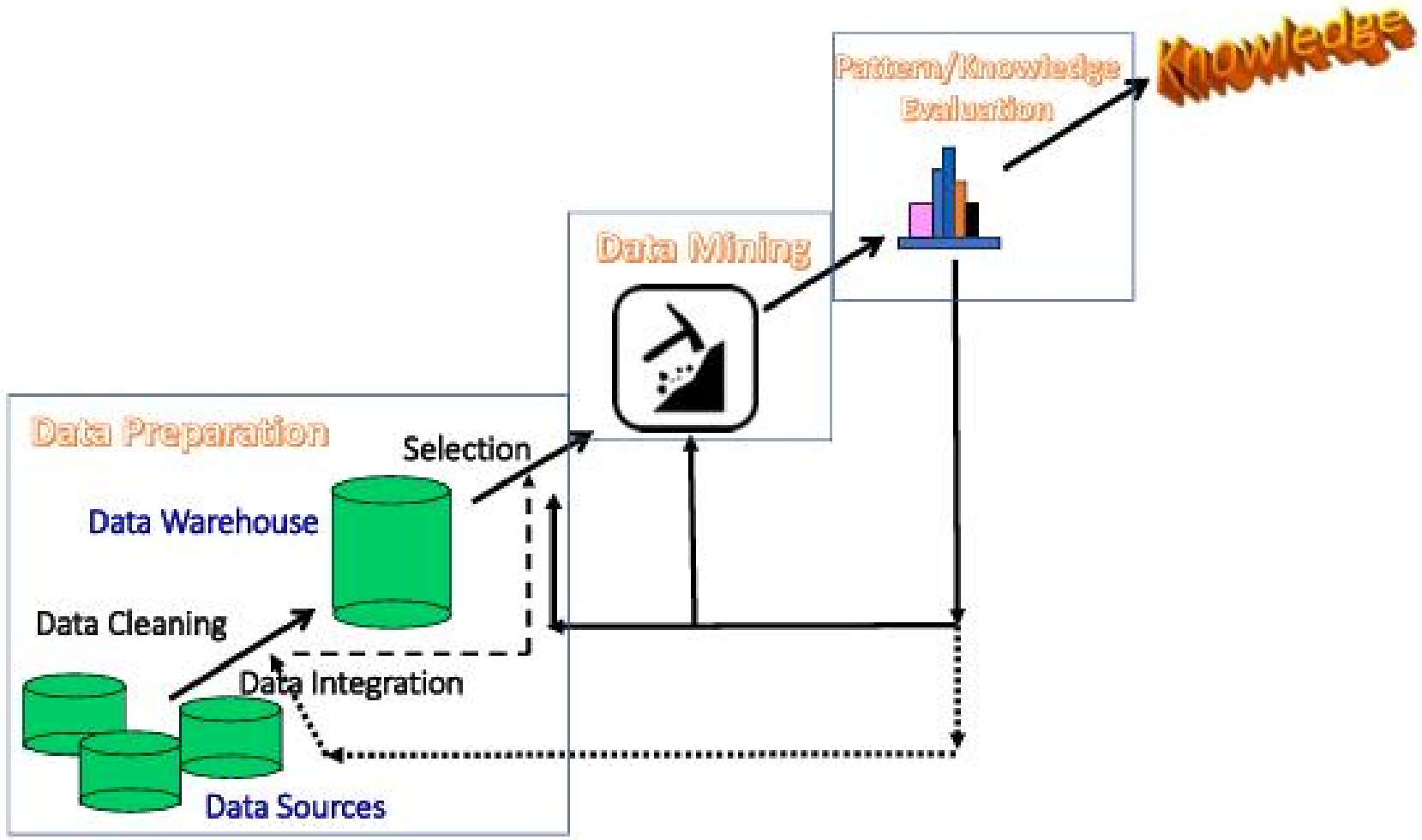
Major Challenges in Data Mining

- **Data Quality Issues -** In most scenarios, data available is not perfect. Available data might not be complete, it may have some missing values, or data might contain some garbage or incosistent values also.
- **Data Complexity Issues -** Data is geenrated from various sources like IoT Devices, social media platforms, etc and each source has their own way of storing and handling data.
- **Scalability and Efficiency -** Data mining techniques should be able to handle large amount of data efficiently.

Steps Involved in Knowledge Discovery

- The process of finding some patterns from data involves a sequence of steps -
  - **Data Preparation**
    - ***Data Cleaning*** - Removing noise and inconsistent data
    - ***Data Integration*** - Combining data from multiple sources
    - ***Data Transformation*** - Transforming data into forms appropriate for mining by performing summary or aggregation operations
    - ***Data Selection*** - Selecting relevant data for the task
  - **Data Mining** - Extracting patterns from the data
  - **Pattern / Model Evaluation** - Comparing extracted patterns / models based on different evaluation metrices
  - **Knowledge Presentation** - Using visualization and knowledge representation techniques to present mined knowledge to users.

Steps Involved in Knowledge Discovery



Data Measurements and Data Preprocessing

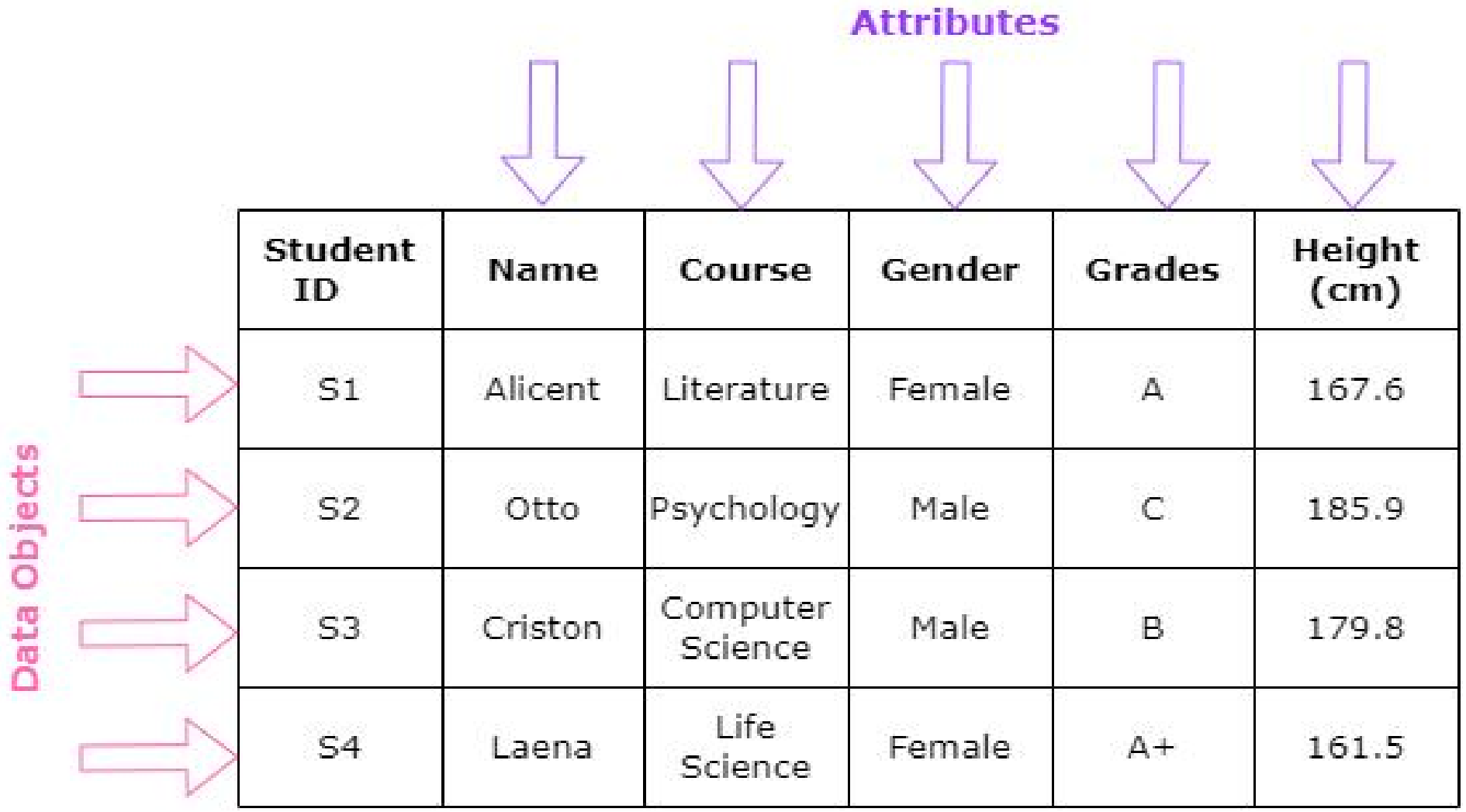
Data Related Issues for Successful Data Mining

- **Types of Data:**
  - Data sets differ in a number of ways
  - Type of data determines which techniques can be used to analyze the data
- **Quality of Data:**
  - Data is often far from perfect.
  - Improving data quality improves the quality of the resulting analysis.
- **Preprocessing of data:**
  - Raw data must be processed in order to make it suitable for analysis.
    - Improve data quality
    - Modify data so that is better fits a specified data mining technique.

What is Data?

- Data sets are made up of **data objects**. Also known as *samples, examples, instances, or data points etc.*
- **A data object represents an entity -**
  - In a sales database, the objects may be customers, store items or sales etc.
  - In a medical database, the objects may be patients, doctors, etc.
  - In a university database, the objects may be students, professors, or courses etc.
- An ***attribute*** is a data field, representing a characteristic or feature of a data object. Attributes are also known as *dimensions, features, variables, etc.*
  - Attributes describing a customer data object can be *customerID, name* or *address, etc.*

What is Data?



Student ID	Name	Course	Gender	Grades	Height (cm)
S1	Alicent	Literature	Female	A	167.6
S2	Otto	Psychology	Male	C	185.9
S3	Criston	Computer Science	Male	B	179.8
S4	Laena	Life Science	Female	A+	161.5

Attributes

- **Attribute** (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
  - E.g., customer _ID, name, address
- **Attribute values** are numbers or symbols assigned to an attribute

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters
  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different; ID has no limit but age has a maximum and minimum value

Attribute Types

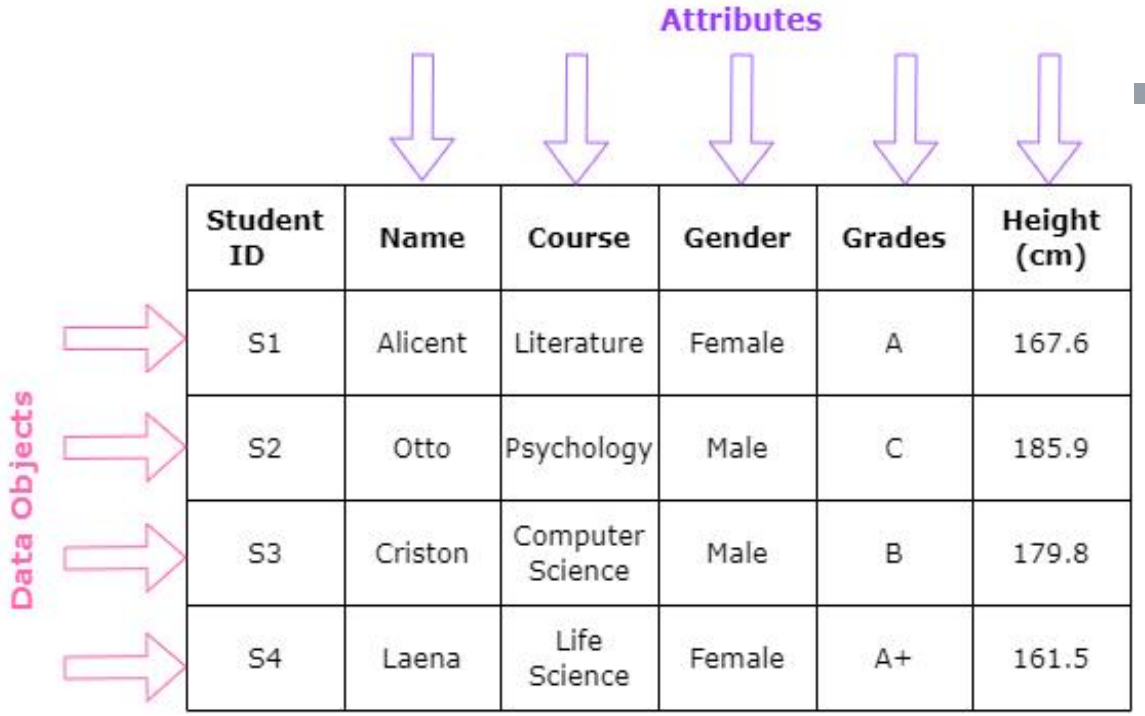
Nominal: Categorical (Qualitative)
- categories, states, or "names of things"
  - Hair color, marital status, occupation, ID numbers, zip codes
- An important nominal attribute: **Binary**
  - Nominal attribute with only 2 states (0 and 1)

Ordinal: Categorical (Qualitative)
- Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = {small, medium, large}, grades, army rankings

Interval: Numeric (Quantitative)
- Measured on a scale of equal-sized units
- Values have order:
  - temperature in C° or F°, calendar dates
- No true zero-point: ratios are not meaningful

Ratio: Numeric (Quantitative)
- Inherent zero-point: ratios are meaningful
  - temperature in Kelvin, length, counts, monetary quantities

Attribute Types

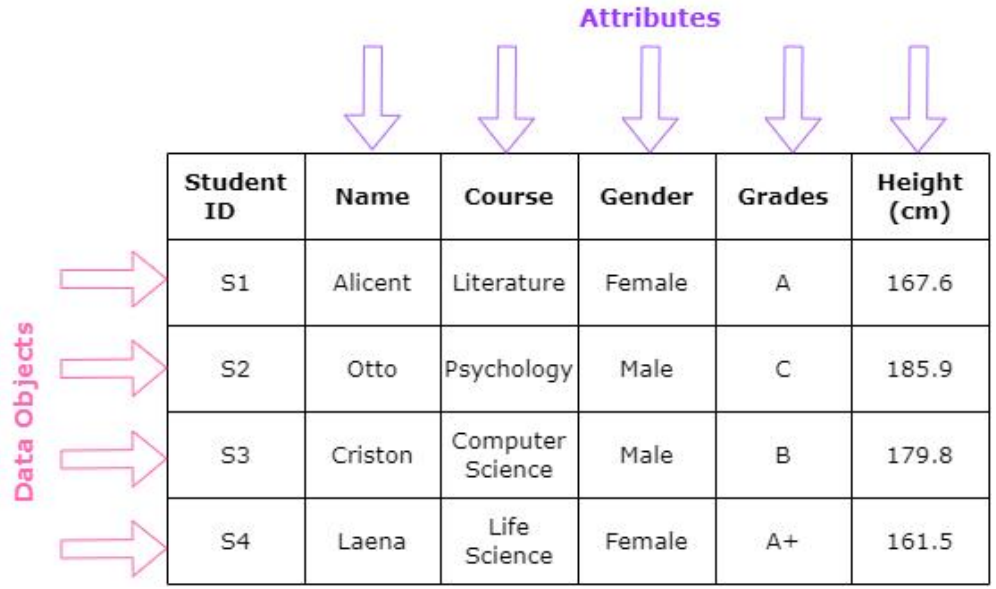


Student ID	Name	Course	Gender	Grades	Height (cm)
S1	Alicent	Literature	Female	A	167.6
S2	Otto	Psychology	Male	C	185.9
S3	Criston	Computer Science	Male	B	179.8
S4	Laena	Life Science	Female	A+	161.5

- A set of attributes used to describe a given object is called an *attribute vector* (or *feature vector*).
- Depending upon what kind of values a attribute take, attributes can be divided into -
  - Nominal attributes
    - Binary attributes
  - Ordinal attributes
  - Numeric attributes
    - Interval based attributes
    - Ratio based attributes

Nominal Attributes

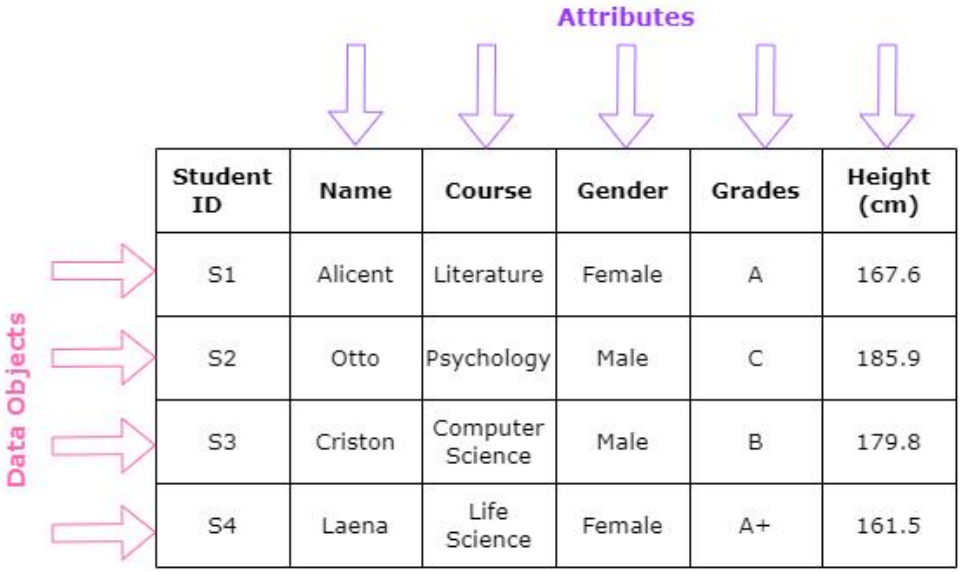
- Nominal attributes are attributes that are symbols or names of things.
- Each value represents some kind of category, code, state, etc., and hence can also be referred to as categorical.
- The attribute values do not have any meaningful order between them.
- In the example, Attributes Name, Course, Gender are nominal attributes.



Student ID	Name	Course	Gender	Grades	Height (cm)
S1	Alicent	Literature	Female	A	167.6
S2	Otto	Psychology	Male	C	185.9
S3	Criston	Computer Science	Male	B	179.8
S4	Laena	Life Science	Female	A+	161.5

Binary Attributes

- A Binary Attribute is the same as a Nominal attribute but with only two categories, for example, 0/1, True/False (Boolean), Pass/Fail, Male/Female, Yes/No, positive/negative, etc.
- In our example, the attribute Gender is a binary attribute as it only consists of two values: Male and Female.
- A binary attribute is ***symmetric*** if both of its states are equally valuable and carry the same weights. For example, in our case, the attribute Gender is symmetric as both of its values (Male and Female) carry equal weight (meaning none of the values is higher or lower importance than the other).
- A binary attribute is ***asymmetric*** if both of its states are not equally important and do not carry the same weight. For example, the outcome of an HIV test (positive/negative) as a positive value is more concerning than a negative value.



Student ID	Name	Course	Gender	Grades	Height (cm)
S1	Alicent	Literature	Female	A	167.6
S2	Otto	Psychology	Male	C	185.9
S3	Criston	Computer Science	Male	B	179.8
S4	Laena	Life Science	Female	A+	161.5

Ordinal Attributes

- An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them.
- Ordinal attributes are qualitative. They describe a feature of an object without giving an actual size or quantity.
- In our example, the attribute Grades is an ordinal attribute as its values A+, A, B, and C have an order.
- Ordinal values can also be obtained by discretization of numeric quantities by splitting the value range into a finite number of ordered categories. Let us understand this with an example of the age of people which say is ranging from 13 to 100. This age attribute contains quantitative values which can be converted to ordinal attribute as Teen (13–19), Adult (20–49), Elder (50–79), and Old (80+). The values Teen, Adult, Elder, and Old have got an order among them.

Attributes

Data Objects

Student ID	Name	Course	Gender	Grades	Height (cm)
S1	Alicent	Literature	Female	A	167.6
S2	Otto	Psychology	Male	C	185.9
S3	Criston	Computer Science	Male	B	179.8
S4	Laena	Life Science	Female	A+	161.5

Numeric Attributes

- Numeric attributes are measurable quantities represented using integers or real values.
- **Interval-scaled:** Interval-scaled attributes are measured on a scale of equal-size units. The values of interval-scaled attributes have order and can be positive, 0, or negative. Thus, in addition to providing a ranking of values, such attributes allow us to compare and quantify the difference between values.
- **Ratio-scaled:** A ratio-scaled attribute is a numeric attribute with an inherent zero-point. That is, if a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value. In addition, the values are ordered, and we can also compute the difference between values, as well as the mean, median, and mode.

Attributes

Data Objects

Student ID	Name	Course	Gender	Grades	Height (cm)
S1	Alicent	Literature	Female	A	167.6
S2	Otto	Psychology	Male	C	185.9
S3	Criston	Computer Science	Male	B	179.8
S4	Laena	Life Science	Female	A+	161.5

Interval Attributes

- **Interval attributes** are measured on a ***scale of equal-size units***.
  - We can compare and quantify the difference between values of interval attributes.
- Example: A ***temperature*** attribute is an interval attribute.
  - We can quantify the difference between values. For example, a temperature of 20°C is five degrees higher than a temperature of 15°C.
  - Temperatures in Celsius do not have a **true zero-point**, that is, 0°C does not indicate "no temperature."
  - Although we can compute the difference between temperature values, we cannot talk of one temperature value as being a multiple of another.
    - Without a true zero, we cannot say, for instance, that 10°C is twice as warm as 5°C . That is, we cannot speak of the values in terms of ratios.
- The central tendency of an interval attribute can be represented by its ***mode***, its ***median*** (middle value in an ordered sequence), and its ***mean***.

Ratio Attributes

- A **ratio attribute** is a numeric attribute with an *inherent zero-point*.
- Example: A ***number_of_words*** attribute is a ratio attribute.
  - If a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value.
- The central tendency of an ratio attribute can be represented by its ***mode***, its ***median*** (middle value in an ordered sequence), and its ***mean***.

Numerical Attributes

Feature	Interval Attributes	Ratio Attributes
Zero Point	No true zero point; zero is arbitrary	True zero point; zero indicates the absence of the attribute
Arithmetic	Addition and subtraction are meaningful	All arithmetic operations (addition, subtraction, multiplication, division) are meaningful
Examples	Temperature (Celsius/Fahrenheit), Calendar dates	Height, Weight, Age, Salary

Properties of Attribute Values

- The type of an attribute depends on which of the following properties it possesses:
  - Distinctness: $= \neq$
  - Order: $< >$
  - Addition: $+ -$
  - Multiplication: $* /$

- Nominal attribute: **distinctness**
- Ordinal attribute: **distinctness & order**
- Interval attribute: **distinctness, order & addition**
- Ratio attribute: **all 4 properties**

Properties of Attribute Values

Attribute Type	Description	Examples
Nominal	The values of a nominal attribute are just different names, i.e., nominal attributes provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}
Ordinal	The values of an ordinal attribute provide enough information to order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers
Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	calendar dates, temperature in Celsius or Fahrenheit
Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length,

Attribute Types: Categorical and Numeric

- **Nominal** and **Ordinal** attributes are collectively referred to as ***categorical or qualitative attributes***.
  - qualitative attributes, such as employee ID, lack most of the properties of numbers.
  - Even if they are represented by numbers, i.e. , integers, they should be treated more like symbols .
  - *Mean* of values does not have any meaning.

- **Interval** and **Ratio** are collectively referred to as ***quantitative or numeric attributes.***
  - Quantitative attributes are represented by numbers and have most of the properties of numbers .
  - Note that quantitative attributes can be integer-valued or continuous.
  - Numeric operations such as *mean*, *standard deviation* are meaningful

Discrete vs. Continous Attributes

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes
  - Binary attributes where only non-zero values are important are called **asymmetric binary attributes.**

- **Continuous Attribute**
  - Has real numbers as attribute values
    - temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

	Discrete Nominal	Discrete Nominal	Binary Discrete Nominal	Ordinal Discrete Nominal	Continuous Ratio-scaled
Student ID	**Name**	**Course**	**Gender**	**Grades**	**Height (cm)**
S1	Alicent	Literature	Female	A	167.6
S2	Otto	Psychology	Male	C	185.9
S3	Criston	Computer Science	Male	B	179.8
S4	Laena	Life Science	Female	A+	161.5

Basic Statistical Descriptions of Data

- Basic statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.
- For data preprocessing tasks, we want to learn about data characteristics regarding both ***central tendency*** and ***dispersion of the data***.
- Measures of central tendency include ***mean, median, mode*** and **midrange**.
- Measure of data dispersion include ***quartiles, interquartile range (IQR)*** and ***variance***.
- These descriptive statistics are of great help in understanding the distribution of the data.

Measuring Central Tendency: Mean

- The most common and most effective numerical measure of the "center" of a set of data is the arithmetic mean. Let $x_1, x_2, \ldots, x_N$ be a set of N values or observations, such as for some numeric attribute X. The mean of this set of values is -

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

- Sometimes, each value x_i, in a set may be associated with a weight w_i for $i = 1, 2, \ldots, N$. The weights reflect the significance, importance or occurrence frequency attched to their respective values. In this case, we can compute weighted mean as following -

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

Measuring Central Tendency: Mean

- Although the mean is the single most useful quantity for describing a data set, it is not always the best way of measuring the center of the data.
  - A major problem with the mean is its sensitivity to extreme (e.g., outlier) values. Even a small number of extreme values can corrupt the mean.
  - For example, the mean salary at a company may be substantially pushed up by that of a few highly paid managers.
- To offset the effect caused by a small number of extreme values, we can instead use the trimmed mean.
- **Trimmed mean** can be obtained after chopping off values at the high and low extremes and then by taking the mean of remaining values.

Measuring Central Tendency: Median

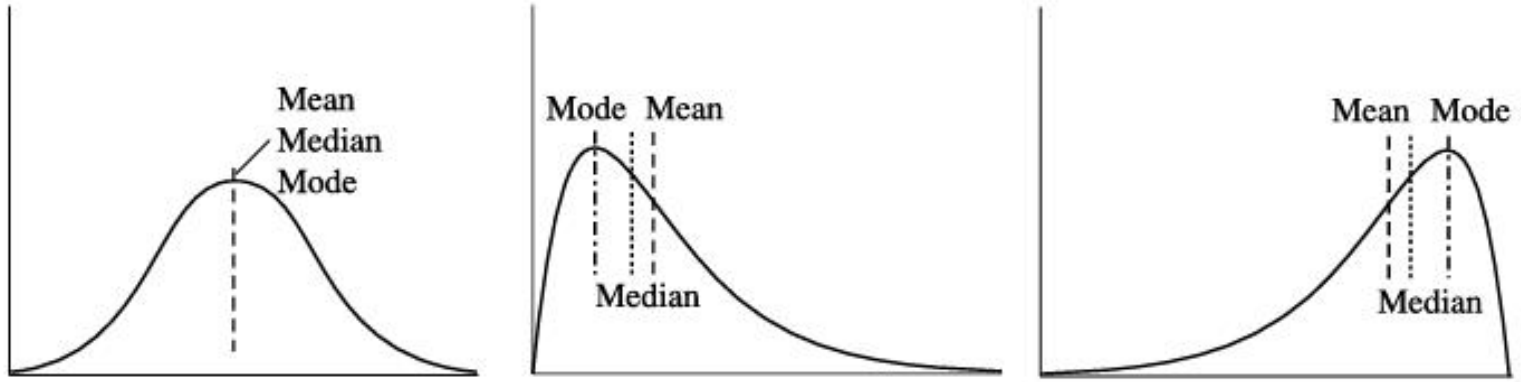
- For skewed (asymmetric) data, a better measure of the center of data is the median, which is the middle value in a set of ordered data values.
- It is the value that separates the higher half of a data set from the lower half.
- Suppose that a given data set of N distinct values is sorted in numerical order.
  - *If N is odd; the median is the middle value of the ordered set.*
  - *If N is even; the median in the average of the middle two values (N/2, N/2 + 1 values).*
- In probability and statistics, the median generally applies to numeric data; however, we may extend the concept to ordinal data.

Measuring Central Tendency: Mode

- Another measure of central tendency is the mode.
- The mode for a set of data is the value that occurs most frequently in the set.
  - It is possible for the greatest frequency to correspond to several different values, which results in more than one mode.
  - Data sets with one, two, or three modes are called *unimodal, bimodal*, and *trimodal* respectively.
  - At the other extreme, if each data value occurs only once, then there is no mode.
  - In general, a data set with two or more modes is *multimodal*.
- For unimodal numeric data that are moderately skewed (asymmetrical), we have the following empirical relation: $mean - mode \approx 3 * (mean - median)$
- This implies that the mode for unimodal frequency curves that are moderately skewed can easily be approximated if the mean and median values are known.

Measuring Central Tendency: Mean, Median, Mode

- The **midrange** can also be used to assess the central tendency of a numeric data set. It is the *average of the largest and smallest values* in the set.
- Central Tendency Measures for Numerical Attributes: *Mean, Median, Mode*
- Central Tendency Measures for Categorical Attributes: *Mode*
  - Central Tendency Measures for Nominal Attributes: *Mode*
  - Central Tendency Measures for Ordinal Attributes: *Mode, Median*



(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

Measuring Central Tendency: Example

What are central tendency measures (mean, median, mode)for the following attributes?

attr1 = {2,4,4,6,8,24}

attr2 = {2,4,7,10,12}

attr3 = {xs,s,s,s,m,m,l}

Measuring Central Tendency: Example

What are central tendency measures (mean, median, mode)for the following attributes?

attr1 = {2,4,4,6,8,24}

mean = (2+4+4+6+8+24)/6 = 8	average of all values
median = (4+6)/2 = 5	avg. of two middle values
mode = 4	most frequent item

attr2 = {2,4,7,10,12}

mean = (2+4+7+10+12)/5 = 7	average of all values
median = 7	middle value
mode = any of them (no mode)	all of them has same freq.

attr3 = {xs,s,s,s,m,m,l}

mean is meaningless for categorical attributes.

median = s	middle value
mode = s	most frequent item

Example

attr1 : {10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30}

attr2: {1, 2, 2, 3, 3, 3, 4, 4, 5, 5, 6, 7, 8, 9, 10, 20}

attr3: {1, 5, 5, 6, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10, 11, 15}

- Calculate the *mean, median* and mode of **attr1**?
- Calculate *the mean,* median and mode of **attr2**?
- Calculate the mean, median and mode of **attr3**?
- Which of these attributes are *symmetric, positively skewed* and *negatively skewed*?

Example

attr1 : {10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30}

attr2: {1, 2, 2, 3, 3, 3, 4, 4, 5, 5, 6, 7, 8, 9, 10, 20}

attr3: {1, 5, 5, 6, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10, 11, 15}

- Calculate the *mean, median* and mode of **attr1**?
  - Mean = 20, Median = 20, Mode = *no mode exists as all values are unique*
- Calculate *the mean,* median and mode of **attr2**?
  - Mean = 5.75, Median = 4.5, Mode = 3
- Calculate the mean, median and mode of **attr3**?
  - Mean = 7.69, Median = 7.5, Mode = 6
- Which of these attributes are *symmetric, positevely skewed* and *negatively skewed*?
  - attr1 is symmetric, attr2 and attr3 are positively skewed.

Mean, Median and Mode from Grouped Frequencies

The Race.....

This starts with some raw data (not a grouped frequency yet) ...

Alex timed 21 people in the sprint race, to the nearest second:

59, 65, 61, 62, 53, 55, 60, 70, 64, 56, 58, 58, 62, 62, 68, 65, 56, 59, 68, 61, 67

After sorting of time points in dataset;

53, 55, 56, 56, 58, 58, 59, 59, 60, 61, 61, 62, 62, 62, 64, 65, 65, 67, 68, 68, 70

To find the Mean, Median & Mode Alex follows the normal methods and return as,

Mean = **61.38095...**

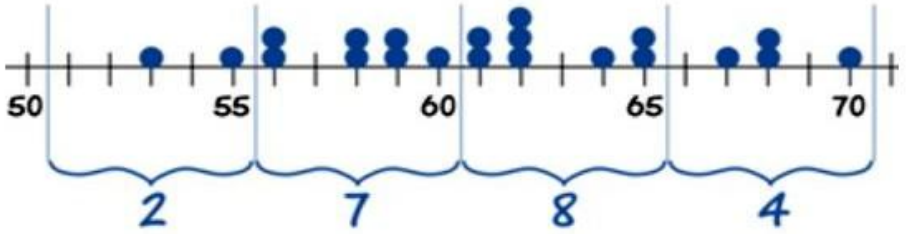
Median = 61

Mode = 62

Mean, Median and Mode from Grouped Frequencies

- Grouped Frequency Table
- Alex then makes a Grouped Frequency Table:
- *So 2 runners took between 51 and 55 seconds, 7 took between 56 and 60 seconds, etc*

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4



Oh No!

Suddenly all the original data gets lost (naughty pup!)

Only the Grouped Frequency Table survived ...

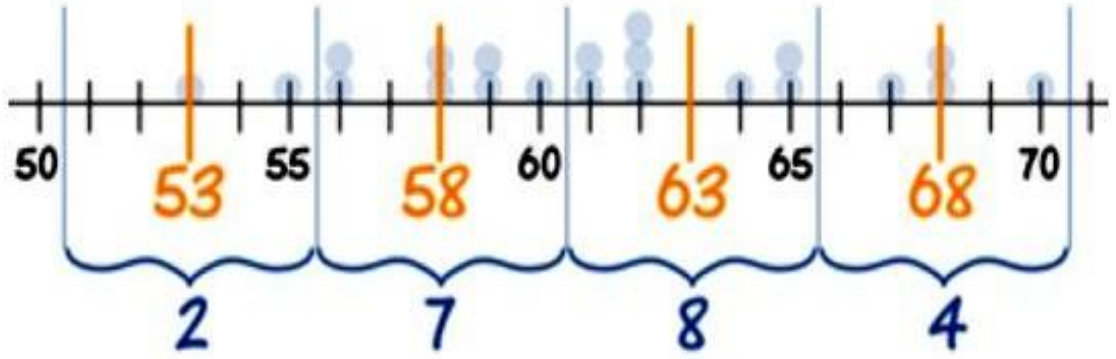
... can we help Alex calculate the Mean, Median and Mode from just that table?

The answer is ... no we can't. Not accurately anyway. But, we can make **estimates**.

Estimating the Mean from Grouped Data

The groups (51-55, 56-60, etc), also called **class intervals**, are of **width** 5

The **midpoints** are in the middle of each class: 53, 58, 63 and 68



Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

We can estimate the Mean by using the ***midpoints.***

Let's now make the table using midpoints:

Midpoint	Frequency
53	2
58	7
63	8
68	4

Estimating the Mean from Grouped Data

Midpoint	Frequency
53	2
58	7
63	8
68	4

- Our thinking is: "2 people took 53 sec, 7 people took 58 sec, 8 people took 63 sec and 4 took 68 sec". In other words we imagine the data looks like this:

53, 53, 58, 58, 58, 58, 58, 58, 58, 63, 63, 63, 63, 63, 63, 63, 63, 68, 68, 68, 68

- Then we add them all up and divide by 21. The quick way to do it is to multiply each midpoint by each frequency:

Midpoint x	Frequency f	Midpoint × Frequency fx
53	2	106
58	7	406
63	8	504
68	4	272
Totals:	**21**	**1288**

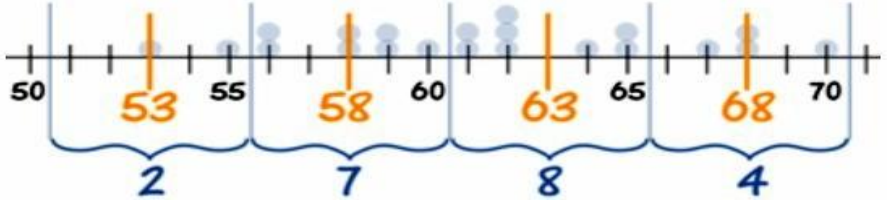
And then our estimate of the mean time to complete the race is:

$$\textbf{Estimated Mean} = \frac{1288}{21} = \textbf{61.333...}$$

Estimating the Median from Grouped Data

- Let's look at our data again:
- The median is the middle value, which in our case is the 11th one, which is in the 61 - 65 group:
- We can say "the **median group** is 61 - 65"
- But if we want an estimated **Median value** we need to look more closely at the 61 - 65 group.

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4



We call it "61 - 65", but it really includes values from 60.5 up to (but not including) 65.5.

Why? Well, the values are in whole seconds, so a real time of 60.5 is measured as 61. Likewise 65.4 is measured as 65.

Estimating the Median from Grouped Data

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

- At 60.5 we already have 9 runners, and by the next boundary at 65.5 we have 17 runners.
- By drawing a straight line in between we can pick out where the median frequency of n/2 runners is:

And this handy formula does the calculation:

$$\textbf{Estimated Median} = L + \frac{(n/2) - B}{G} \times w$$

where:

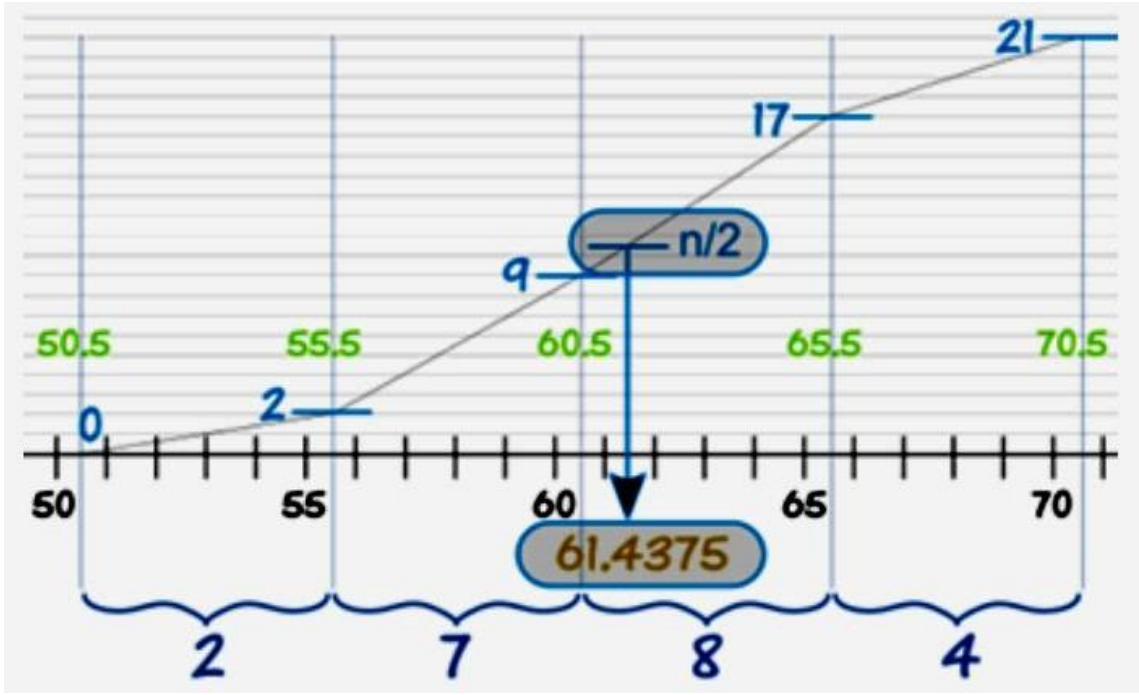
L is the lower class boundary of the group containing the median

n is the total number of values

B is the cumulative frequency of the groups before the median group

G is the frequency of the median group

w is the group width



For our example:

Seconds	Frequency
51 - 55	2
56 - 60	7
61 - 65	8
66 - 70	4

L = 60.5

n = 21

B = 2 + 7 = 9

G = 8

w = 5

$$median = L_1 + (\frac{n/2 - (\sum freq)l}{freq_{median}})width$$

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

Estimated Median $= 60.5 + \frac{(21/2) - 9}{8} \times 5$

$= 60.5 + 0.9375$

$= \mathbf{61.4375}$

Measuring Dispersion of Data

- The degree to which numerical data tend to spread is called the dispersion, or variance of the data.
- The most common measures of data dispersion are -
  - **Range**
  - **quantiles**
  - **quartiles**
  - **percentiles**
  - **interquartile range**
  - **five-number summary**
  - **variance and standard deviation**

Variance and Standard Deviation

- Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is.
- A low standard deviation means that the data observations tend to be very close to the mean, whereas a high standard deviation indicates that the data are spread out over a large range of values.

The **variance** of N observations, $x_1, x_2, \ldots, x_N$ (when N is large), for a numeric attribute X is

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2,$$

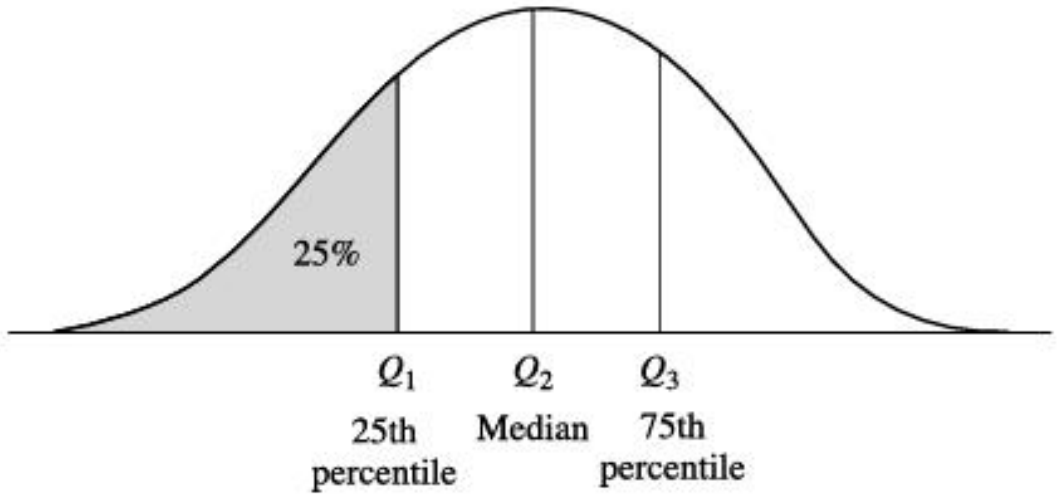
The **Standard Deviation** is the square root of variance.

Range, Quartiles, and Interquartile Range

- Let $x_1, x_2, \ldots, x_N$ be a set of observations for some numeric attribute X.
- The **range** of the set is the difference between the largest and smallest values.
- Suppose that the data for attribute X are sorted in ascending numeric order.
  - **Quantiles** are points taken at regular intervals of a data distribution, dividing it into essentially equal-size consecutive sets.
  - The kth q-quantile for a given data distribution is the value x such that at most k/q of the data values are less than x and at most (q –k)/q of the data values are more than x, where k is an integer such that 0 <k<q.
  - The 2-quantile is the data point dividing the lower and upper halves of the data distribution. It corresponds to the median.
  - The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as **quartiles** represented as **Q1, Q2,** and **Q3.**

Range, Quartiles, and Interquartile Range

- The 100-quantiles are more commonly referred to as percentiles; they divide the data distribution into 100 equal-size consecutive sets. The median, quartiles, and percentiles are the most widely used forms of quantiles.
- **Interquartile range (IQR) = Q3 - Q1**



Example

- Suppose wehavethe following values for salary (in thousands of dollars), shown in ascending order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.
- Find out Q1, Q2 and Q3?
- Find out IQR of the data-set?

Example

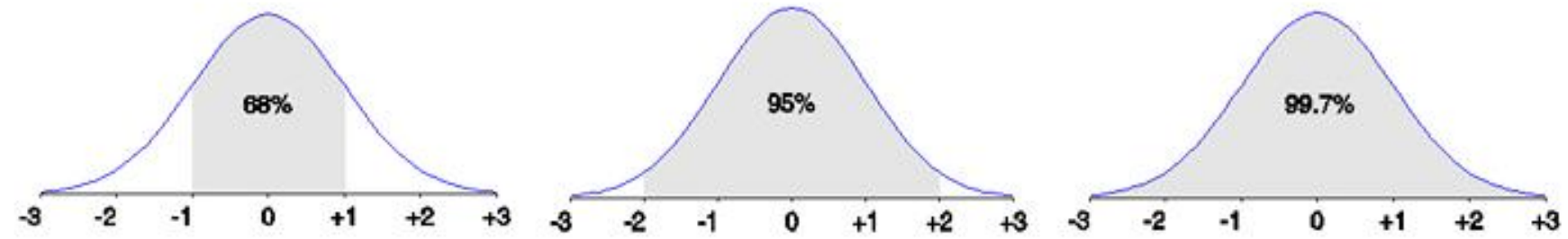
- Suppose wehavethe following values for salary (in thousands of dollars), shown in ascending order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.
- Find out Q1, Q2 and Q3?
  - Q1 would be mean of 3rd and 4th values.
  - Q2 would be mean of 6th and 7th values.
  - Q3 would be mean of 9th and 10th values.
- Find out IQR of the data-set?
  - Q3 - Q1

Five-number summary

- Because Q1, the median, and Q3 together contain no information about the endpoints (e.g., tails) of the data, a fuller summary of the shape of a distribution can be obtained by providing the lowest and highest data values as well. This is known as the five-number summary.
- The **five-number summary** of a distribution consists of the median (Q2), the quartiles Q1 and Q3, and the smallest and largest individual observations, written in the order of ***Minimum, Q1, Median, Q3, Maximum***.

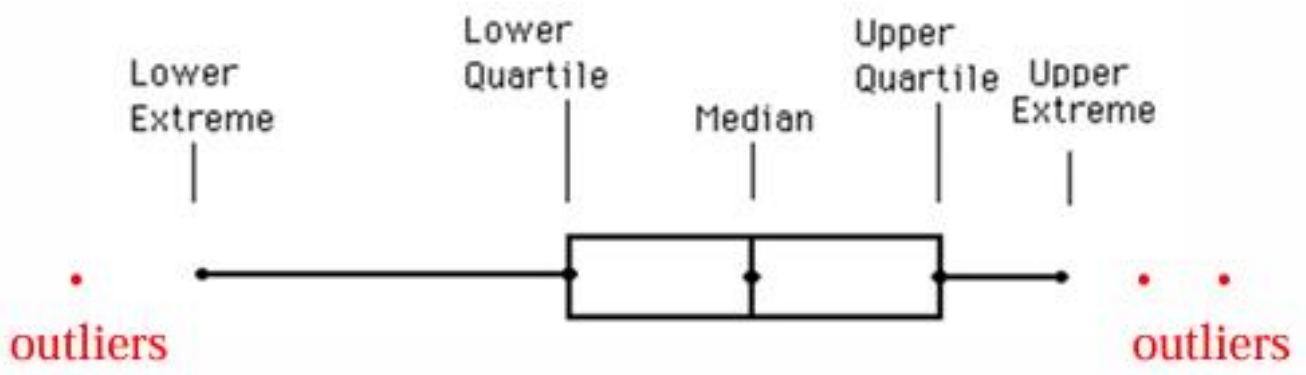
Outliers

- **Outliers** can be identified by the help of ***interquartile range*** or ***standard deviation*** measures.
  - Suspected outliers are values falling at least 1.5xIQR above the third quartile or below the first quartile.
  - Suspected outliers are values that fall outside of the range of $\mu - N\sigma$ and $\mu + N\sigma$ where μ is mean and σ is standard deviation. N can be chosen as 2.5.
- The normal distribution curve: (μ: mean, σ: standard deviation)
  - From $\mu - \sigma$ to $\mu + \sigma$: contains about 68% of the measurements
  - From $\mu - 2\sigma$ to $\mu + 2\sigma$: contains about 95% of it
  - From $\mu - 3\sigma$ to $\mu + 3\sigma$: contains about 99.7% of it



Boxplots, and Outliers

- **Boxplots** are a popular way of visualizing a distribution and a boxplot incorporates *five-number summary*:
  - The ends of the box are at the **quartiles Q1 and Q3**, so that the box length is the interquartile range, IQR.
  - The **median** is marked by a line within the box. (**median** of values in IQR)
  - Two lines outside the box extend to the **smallest and largest observations** (outliers are excluded). Outliers are marked separately.
    - If there are no outliers, lower extreme line is the smallest observation (Minimum) and upper extreme line is the largest observation (Maximum).



Example

Consider following two attribute values:

attr1: {2,3,4,5,6,7,8,9} **attr2: {1,5,9,10,11,12,18,30}**

Which attribute has biggest standard deviation? Do not compute standard deviations.

Give interquartile ranges of attribute values?

Are there any outliers (wrt IQR) in these datasets?

Give a 4 element dataset whose standard deviation is zero?

Example

Consider following two attribute values:

attr1: {2,3,4,5,6,7,8,9} **attr2: {1,5,9,10,11,12,18,30}**

Which attribute has biggest standard deviation? Do not compute standard deviations.

attr2

Give interquartile ranges of attribute values?

attr1: Q1: (3+4)/2=3.5 Q3:(7+8)/2=7.5 IQR:3.5-7.5 = 4

attr2: Q1: (5+9)/2=7 Q3:(12+18)/2=15 IQR:7-15 = 8

Are there any outliers (wrt IQR) in these datasets?

Yes. 30 in attr2. 30 > 15+1.5*IQR

Give a 4 element dataset whose standard deviation is zero? **{1,1,1,1}**

Covariance

- Two attributes may change in relation to each other. Covariance is a measure of association of two variables.
- If covariance is positive, then both attributes increase or decrease together.
- If covariance is negative, then if one attribute values increases then other will decrease or vice versa.

Covariance

- Consider two numeric attributes A and B and a set of n real valued observations **$\{(a_1, b_1), \ldots, (a_n, b_n)\}$.**

The mean or expected values of A and B would be -

$$E(A) = \bar{A} = \frac{\sum_{i=1}^{n} a_i}{n}$$

$$E(B) = \bar{B} = \frac{\sum_{i=1}^{n} b_i}{n}$$

The **covariance** between A and B is defined as -

$$Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n}$$

Mathematically, it can also be shown that -

$$Cov(A, B) = E(A.B) - \bar{A}\,\bar{B}$$

Covariance

- For two attributes A and B that tend to change together, if a value a_i of A is larger than expected value of A, then the corresponding value of b_i of attribute B is likely to be larger than the expected value of B. Therefore the covariance between A and B is positive.
- On the other hand, if one of the attributes tends to be above its expected value when the other attribute is below its expected value, then the covariance of A and B is negative.
- If A and B are independent, then Cov(A, B) = 0 however converse in not always true i.e. some pair of attributes might have a covariance of 0 but are not independent.

Example

Table 2.1 Stock prices for *AllElectronics* and *HighTech*.

Time point	AllElectronics	HighTech
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

Check if both stocks follow same trends?

Correlation Coefficient

- For numeric attributes, we can evaluate the correlation between two attributes, A and B, by computing the correlation coefficient (also known as **Pearson's product moment coefficient**) -

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}$$

- Note that -1 <= $r_{A,B}$ <= +1.
- If $r_{A,B}$ is greater than 0, then A and B are positively correlated, meaning if values A increases then values of B will also increase.
- If $r_{A,B}$ is 0 then A and B are independent.
- If $r_{A,B}$ is less than 0, then A and B are negatively correlated, meaning if values of A increases then values of B will decrease.

Chi-square test for Independence

- A chi-square test of independence is to test whether two categorical (nominal) variables (attributes) are related to each other or not.
- **Example 1:** we have a list of movie genres; this is the first variable. The second variable is whether or not the viewers of those genres bought snacks at the theater. **The idea (or null hypothesis) is that the type of movie and whether or not people bought snacks are unrelated.** The owner of the movie theater wants to estimate how many snacks to buy. If movie type and snack purchases are unrelated, estimating will be simpler than if the movie types impact snack sales.
- **Example 2:** a veterinary clinic has a list of dog breeds they see as patients. The second variable is whether owners feed dry food, canned food or a mixture. The idea (or null hypothesis) is that the dog breed and types of food are unrelated. If this is true, then the clinic can order food based only on the total number of dogs, without consideration for the breeds.

Chi-square Test for Independence Example

- Let's take a closer look at the movie snacks example. Suppose we collect data for 600 people at our theater. For each person, we know the type of movie they saw and whether or not they bought snacks.
- For the valid Chi-square test, the following conditions to be satisfied:
  - Data values that are a simple random sample from the population of interest.
  - Two categorical or nominal variables.
  - For each combination of the levels of the two variables, we need at least five expected values. When we have fewer than five for any one combination, the test results are not reliable. To confirm this, we need to know the total counts for each type of movie and the total counts for whether snacks were bought or not. For now, we assume we meet this requirement and will check it later.

Statistical details

- The null hypothesis is that the type of movie and snack purchases are independent. It is written as: **H_0:Movie Type and Snack purchases are independent**
- The alternative hypothesis is the opposite i.e., **H_a: Movie Type and Snack purchases are not independent.**

Chi-square Test for Independence Example cont…

- The data summarized in a **contingency table** is as follows:

Type of movie	Snacks	No snacks
Action	50	75
Comedy	125	175
Family	90	30
Horror	45	10

- Before we go any further, let's check the assumption of five expected values in each category. The data has more than five counts in each combination of Movie Type and Snacks.

- To find expected counts for each Movie-Snack combination, we first need the row and column totals, which are shown below:

Type of movie	Snacks	No snacks	Row Totals
Action	50	75	50 + 75 = 125
Comedy	125	175	125 + 175 = 300
Family	90	30	90 + 30 = 120
Horror	45	10	45 + 10 = 55
Column Totals	50+125+90+45 = 310	75+175+30+10 = 290	**Grand Total** = 600

Chi-square Test for Independence Example cont…

- The expected counts for each Movie-Snack combination are based on the row and column totals. We multiply the row total by the column total and then divide by the grand total. This gives us the expected count for each cell in the table.
- For example, for the Action-Snacks cell: (125 * 310) / 600 = 65. If there is not a relationship between movie type and snack purchasing we would expect 65 people to have watched an action film with snacks.

- For the Action-No Snacks cell: (125 * 290) / 600 = 60. Similarly, it can be counted for others…
- The expected count appears (in bold) beneath the actual count.

Type of movie	Snacks	No snacks	Row Totals
Action	50 125*310/600 = **65**	75 125*290/600 = **60**	125
Comedy	125 300*310/600 = **155**	175 300*290/600 = **145**	300
Family	90 120*310/600 = **62**	30 120*290/600 = **58**	120
Horror	45 55*310/600 = **28**	10 55*290/600 = **27**	55
Column Totals	310	290	Grand Total = 600

Chi-square Test for Independence Example cont…

- All of the expected counts for our data are larger than five, so we meet the requirement for applying the independence test.
- If we look at each of the cells, we can see that some expected counts are close to the actual counts but most are not.
- If there is no relationship between the movie type and snack purchases, the actual and expected counts will be similar. If there is a relationship, the actual and expected counts will be different.

Performing the Chi-square Test

- The basic idea in calculating the test statistic is to compare actual and expected values, given the row and column totals that we have in the data.
- ***First, we calculate the difference from actual and expected for each Movie-Snacks combination.***
- ***Next, we square that difference. Squaring gives the same importance to combinations with fewer actual values than expected and combinations with more actual values than expected.***
- ***Next, we divide by the expected value for the combination. We add up these values for each Movie-Snacks combination. This gives the test statistic.***

Chi-square Test for Independence Example cont…

Type of movie	Snacks	No snacks	Row Totals
Action	Actual: 50 Expected: **65** Difference: 50 – 65 = -15 Squared Difference = 225 Divide by Expected: 225/65 = 3.46	Actual: 75 Expected: **60** Difference: 75 – 60 = 15 Squared Difference = 225 Divide by Expected: 225/60 = 3.75	125
Comedy	Actual: 125 Expected:**155** Difference: 125– 155 = -30 Squared Difference = 900 Divide by Expected: 900/155 = 5.81	Actual: 175 Expected: **145** Difference: 175– 145 = 30 Squared Difference = 900 Divide by Expected: 900/145 = 6.21	300
Family	Actual: 90 Expected:**62** Difference: 90 – 62 = 28 Squared Difference = 784 Divide by Expected: 784/62 = 12.65	Actual: 30 Expected: **58** Difference: 30 – 58 = -28 Squared Difference = 784 Divide by Expected: 784/58 = 13.52	120
Horror	Actual: 45 Expected:**28** Difference: 45 – 28 = -16 Squared Difference = 256 Divide by Expected: 256/28 = 9.14	Actual: 10 Expected: **27** Difference: 10 – 27 = -17 Squared Difference = 289 Divide by Expected: 289/27 = 10.70	55
Column Totals	310	290	Grand Total = 600

Chi-square Test for Independence Example cont…

- Lastly, to get our test statistic, we add the numbers in the final row for each cell: ***3.46 + 3.75 + 5.81 + 6.21 + 12.65 + 13.52 + 9.14 + 10.70 = 65.24***
- Now, we need to find the critical value from the Chi-square distribution based on degrees of freedom and significance level. This is the value to expect if the two variables are independent.
- The degrees of freedom depend on how many rows and how many columns we have in contigency table. The degrees of freedom (df) are calculated as $df=(r-1)\times(c-1)$ where r is the number of rows, and c is the number of columns in the contingency table. From the example, r is 4 and c is 2. Hence, $df = (4-1)\times(2-1)=3\times1=3$.

- The Chi-square value with $\alpha = 0.05$ (it is given and represents the probability of rejecting the null hypothesis when it is true) and three degrees of freedom is 7.815. **Note:** This value of 7.815 to be infer from the Chi-square distribution table.

- We compare the value of our test statistic (65.24) to the Chi-square value. Since $65.24 > 7.815$, we reject the idea that movie type and snack purchases are independent.

Chi-Square distribution table

Degrees of freedom (*df*)	Significance level (α)							
	.99	.975	.95	.9	.1	.05	.025	.01
1	-------	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578

- Therefore, we conclude that there is some relationship between movie type and snack purchases.
- However, the owner of the movie theater cannot estimate how many snacks to buy regardless of the type of movies being shown. Instead, the owner must think about the type of movies being shown when estimating snack purchases.
- It's important to note that we cannot conclude that the type of movie causes a snack purchase. The independence test tells us only whether there is a relationship or not; it does not tell that one variable causes the other.

Example

Degree	Male	Female
High School	185	251
Bachelor's	1824	1199
Master's	791	1068
PhD	873	496

Graphic Display of Basic Statistics of Data: Quantile Plot

- A quantile plot is a simple and effective way to have a first look at a univariate data distribution. First, it displays all of the data for the given attribute (allowing a user to assess both the overall behavior and unusual occurrences). Second, it plots quantile information.
- Let x_i, for i = 1 to N, be the data sorted in ascending order so that x_1 is the smallest observation and x_N is the largest for some ordinal or numeric attribute X.
- Each observation, x_i, is paired with a percentage, f_i, which indicates that approximately $f_i \times 100\%$ of the data are below the value, x_i.

$$f_i = \frac{i - 0.5}{N}.$$

Quantile Plot

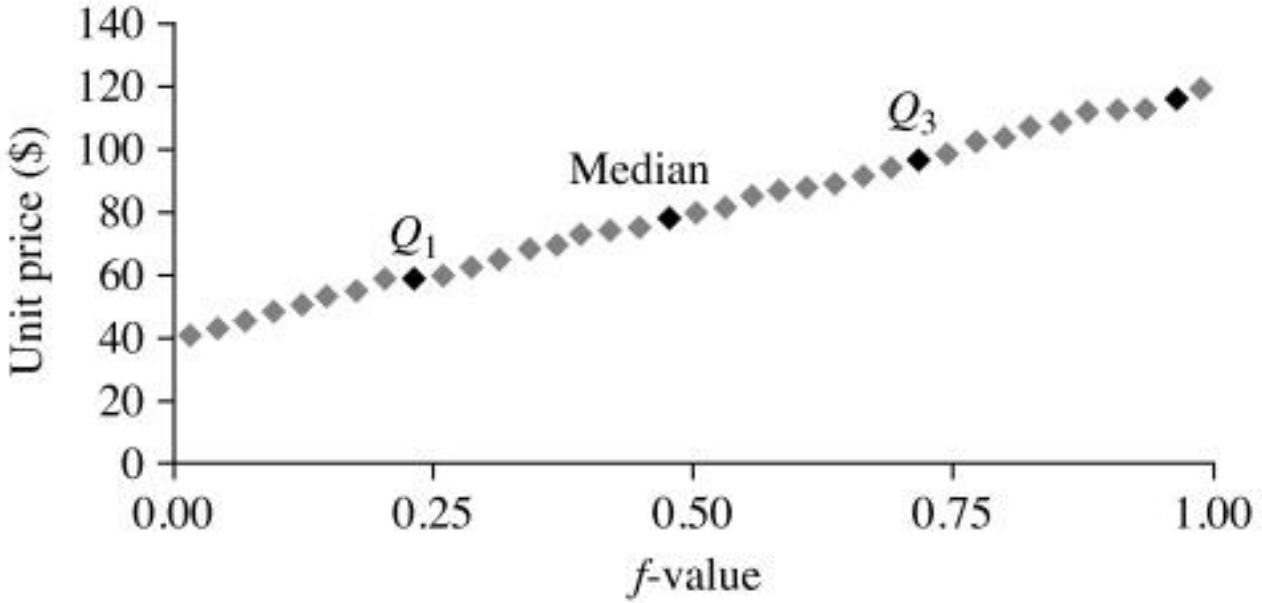


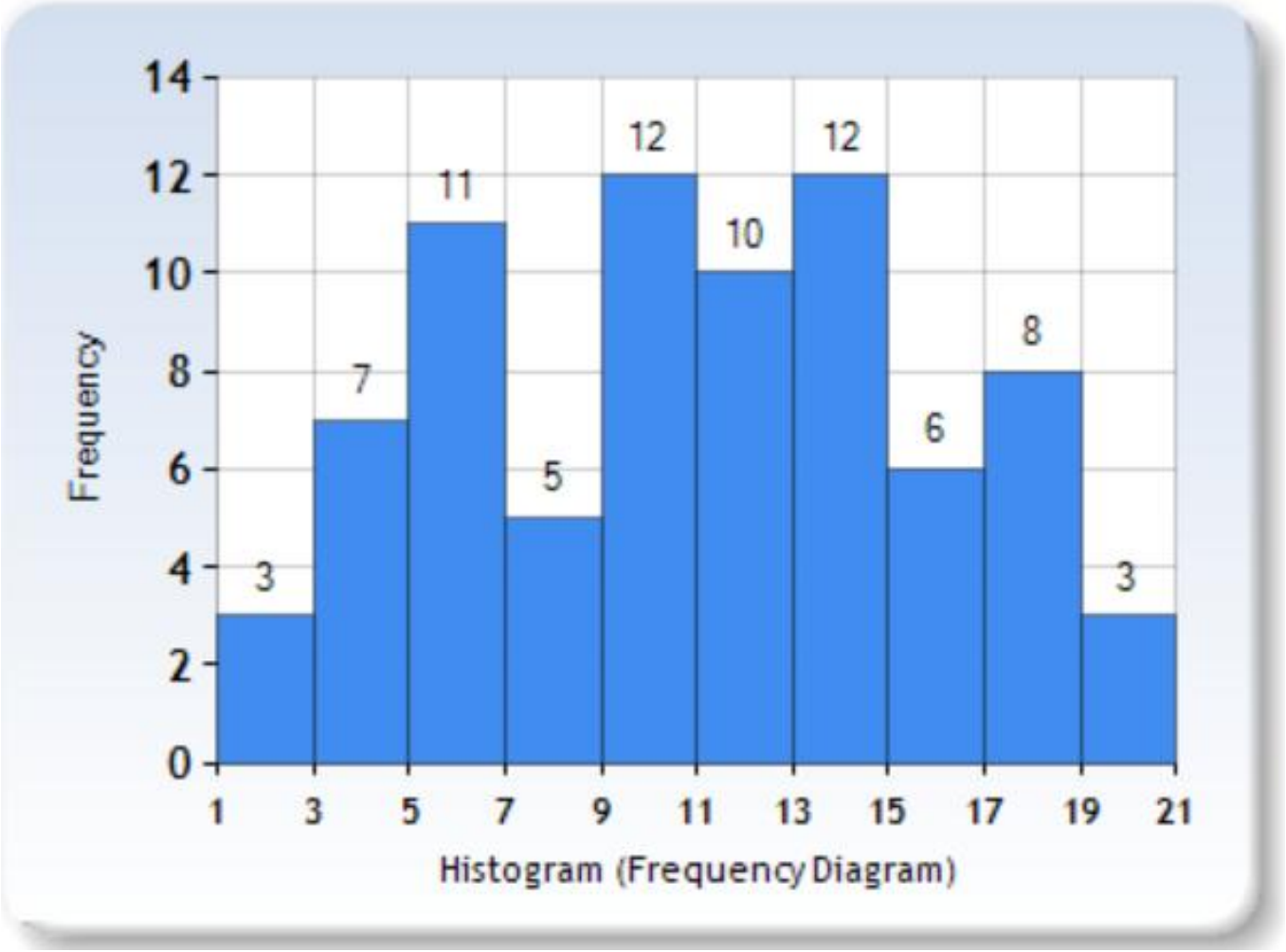
Table 2.3 A set of unit price data for items sold at a branch of the online store.

Unit price ($)	Count of items sold
40	275
43	300
47	250
⋮	⋮
74	360
75	515
78	540
⋮	⋮
115	320
117	270
120	350

Histograms or Frequency Histograms

- To create a histogram plot of attribute X, first particition range of values of X into a set of disjoint consecutive subranges called bins or buckets.

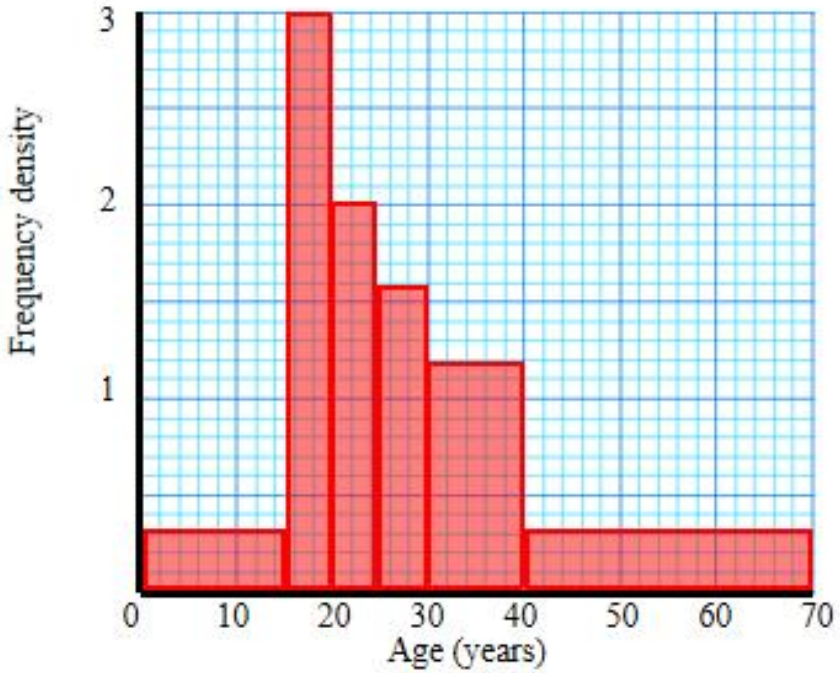
Histogram



Range	Frequency
0 <= x <= 2	3
2 < x <= 4	7
4 < x <= 6	11
6 < x <= 8	5
8 < x <= 10	12
10 < x <= 12	10
12 < x <= 14	12
14 < x <= 16	6
16 < x <= 18	8
18 < x <= 20	3

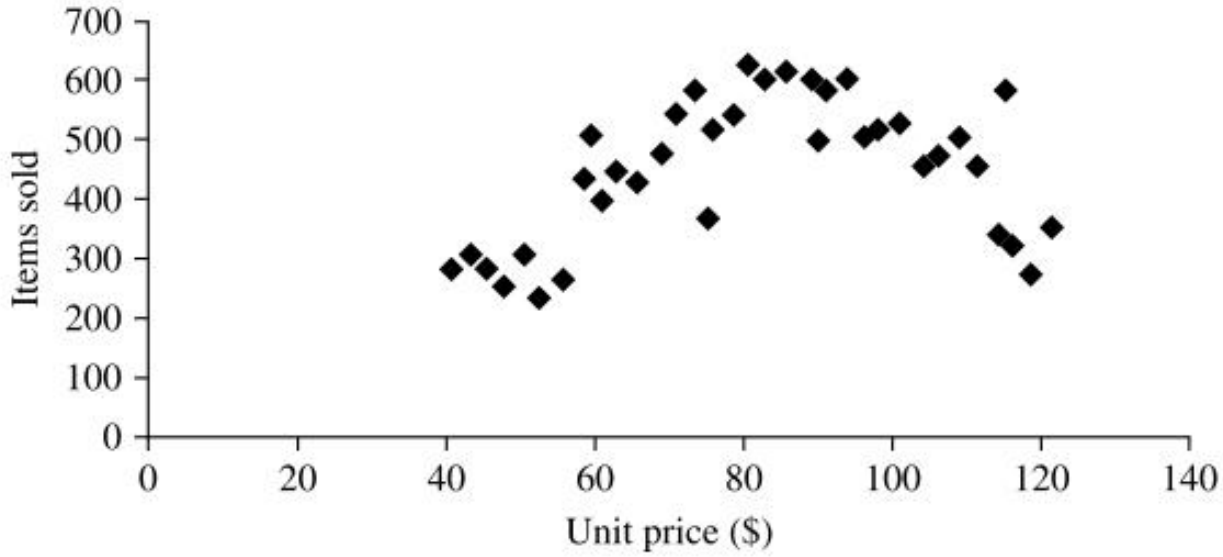
Draw a histogram to represent this data

Age	Frequency	Class Width	Frequency Density
$0 < x \leq 15$	5	15	$\frac{1}{3}$
$15 < x \leq 20$	15	5	3
$20 < x \leq 25$	10	5	2
$25 < x \leq 30$	8	5	1.6
$30 < x \leq 40$	12	10	1.2
$40 < x \leq 70$	10	30	$\frac{1}{3}$



Scatter Plot

- A scatter plot is one of the most effective graphical methods for determining whether there appears to be a relationship, pattern, or trend between two numeric attributes.
- To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane.



Scatter Plots

- The scatter plot is a useful method for providing a first look at bivariate data to see clusters of points and outliers, or to explore the possibility of correlation relationships.
- Two attributes, X and Y, are correlated if the knowledge of one attribute enables to predict the other with some accuracy.
- Correlations can be positive, negative, or null (uncorrelated).

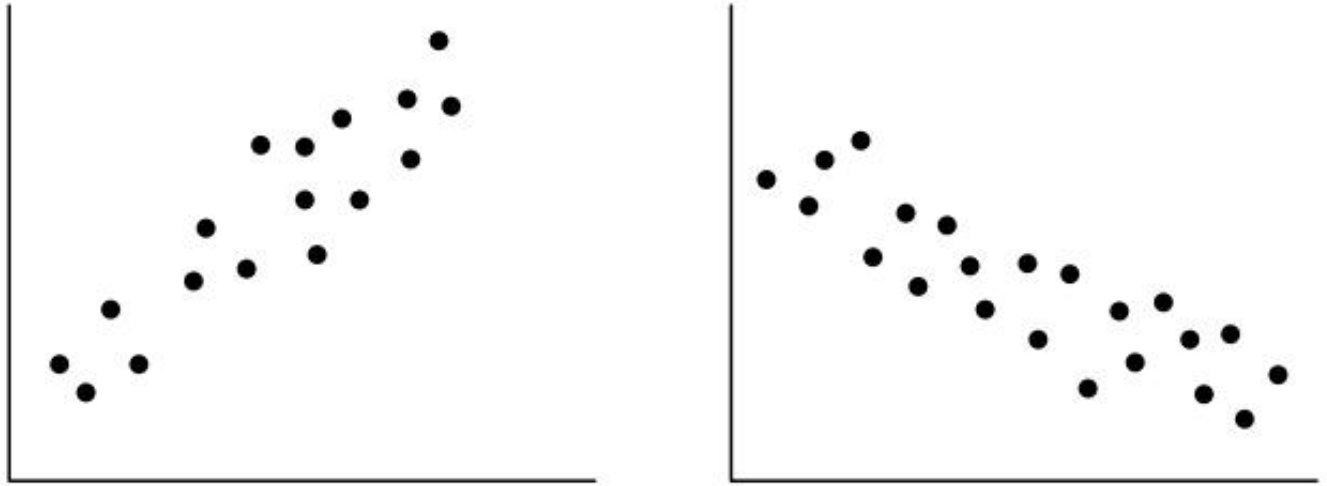


FIGURE 2.8

Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.

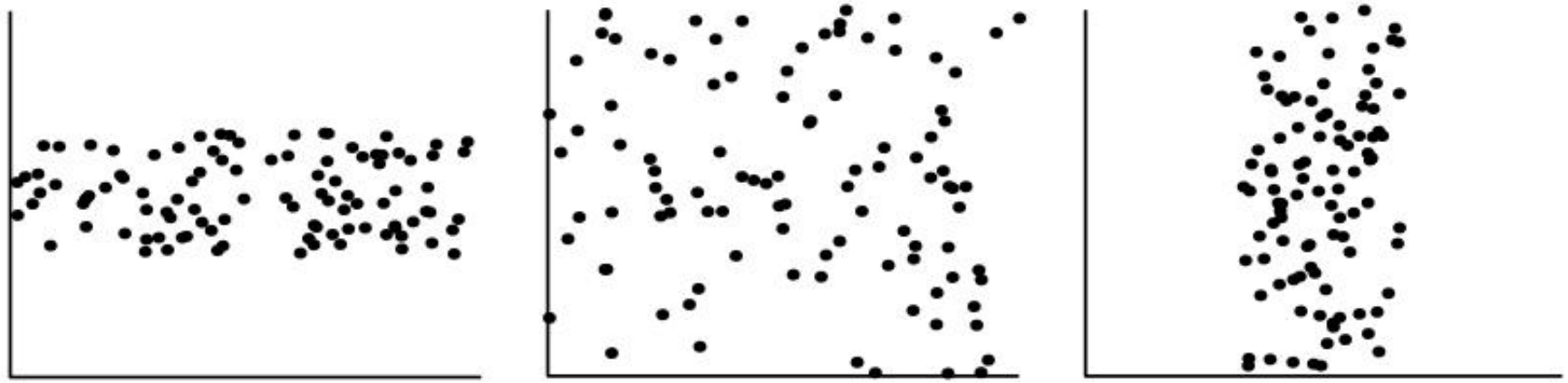


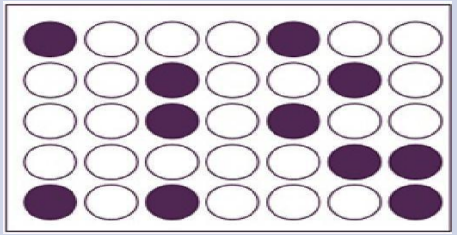
FIGURE 2.9

Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.

Data Preprocessing

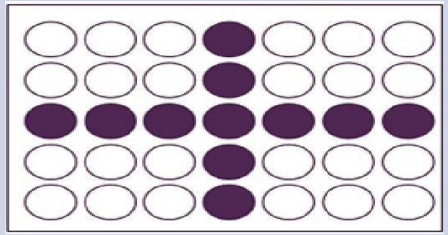
Sampling Methods

P robability Sampling



- Any element can be chosen **randomly** from the population. It deals with choosing the sample randomly.
- The most critical requirement of probability sampling is that everyone in your population has a known and equal chance of getting selected.
- Ex. When an unbiased coin is thrown (randomly), the probability of getting the head is ½.
- Ex. Probability of getting a number i.e 6 when a dice will be thrown.

Non-Probability Sampling



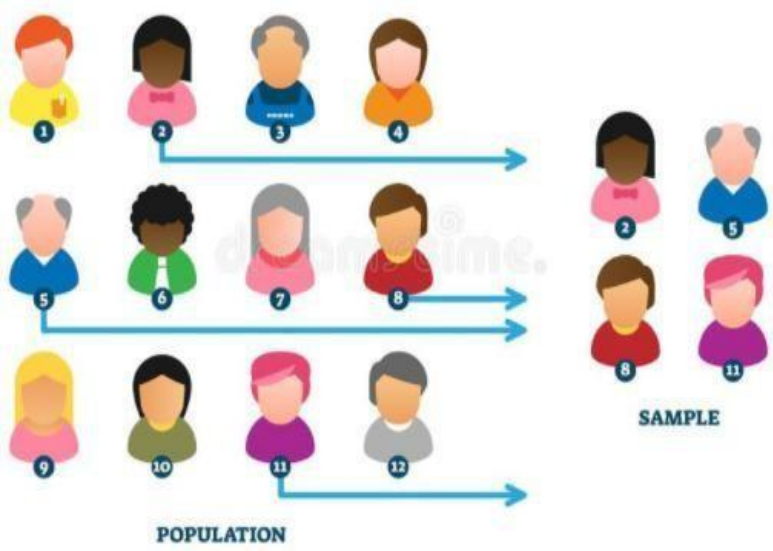
- Every element will be chosen on the subjective judgment (purposefully /intentionally) from the population on the basis of certain past experience & knowledge rather than random selection.
- A sampling process where every single individual elements in the population may not have an opportunity to be choosen as a sample.
- For example, one person could have a 10% chance of being selected and another person could have a 50% chance of being selected.

Probability Sampling

- Simple Random Sampling
- Systematic Sampling
- Stratified Sampling
- Cluster sampling
- Multi Stage Sampling

Simple Random Sampling

- Randomly any element can be chosen
- Chance of selection is totally in a randomized fashion.
- No previous knowledge, criteria and procedure is followed at the time of selection of the sample from the population.

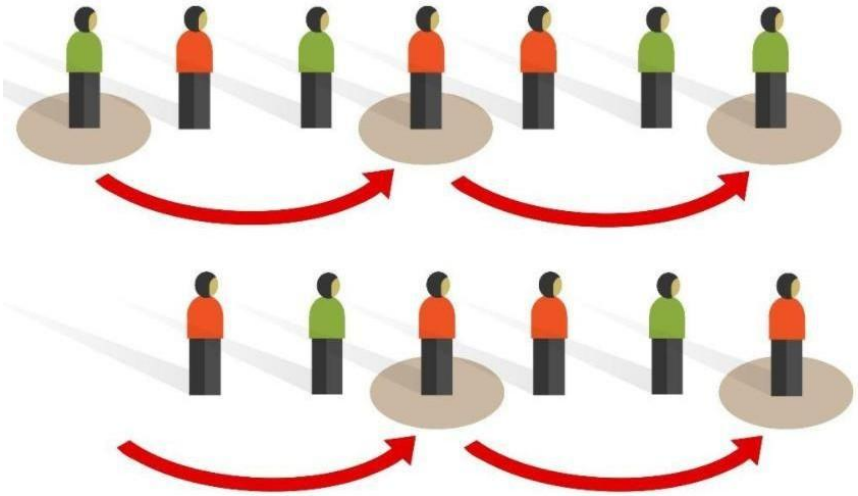


Example: Suppose we would like to select 10 students from any class consists of 75 students. Write the roll numbers of each students in separate cheats and put it in a container and 10 cheats from the container one by one randomly. Here probability of selection is 1/75

Advantage: Every element has an equal chance of getting selected to be the part sample.

Systematic Sampling

- Each member of the sample comes after an equal interval from its previous member.
- All the elements are put together in a sequence first where each element has the equal chance of being selected.
- Select a random starting point and then select the individual at regular intervals



Example: Suppose we would like to select 10 students from any class consists of 75 students. Choosing a random stating roll choose every 5th student.

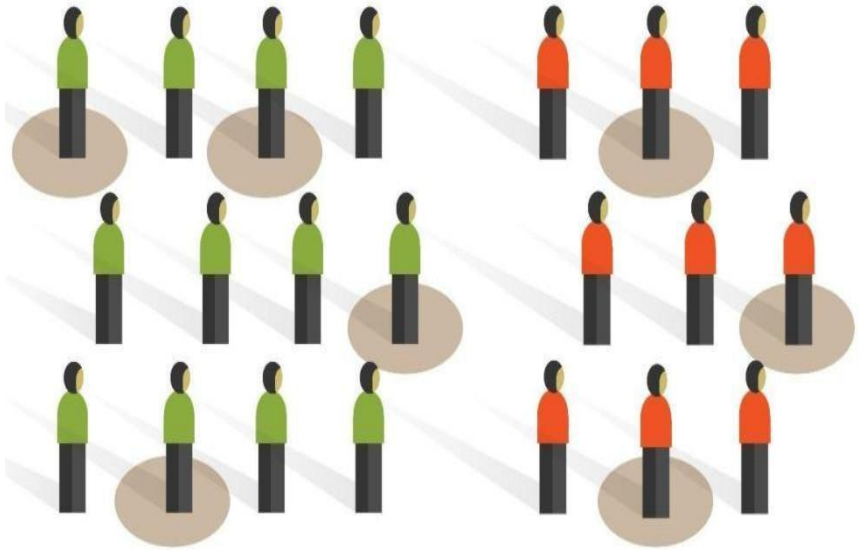
Advantage: As each student has a chance of getting selected there is no biasness in selection.

Systematic Clustering (cont..)

- For a sample of size n, we divide our population of size N into subgroups of k elements.
- We select our first element randomly from the first subgroup of k elements.
- To select other elements of sample, perform following:
  - We know number of elements in each group is k i.e N/n
  - So if our first element is n1 then Second element is n1+k i.e n2
  - Third element n2+k i.e n3 and so on.. Taking an example of N=20, n=5
  - No of elements in each of the subgroups is N/n i.e 20/5 =4= k
  - Now, randomly select first element from the first subgroup. If we select n1= 3, n2 = n1+k = 3+4 = 7, n3 = n2+k = 7+4 = 11

Stratified Sampling

- The population is divided into smaller homogeneous groups or strata by some characteristics.
- i.e the elements within the group are homogeneous and heterogeneous among the other subgroups formed.
- The samples are selected randomly from these strata.
- We need to have prior information about the population to create subgroups

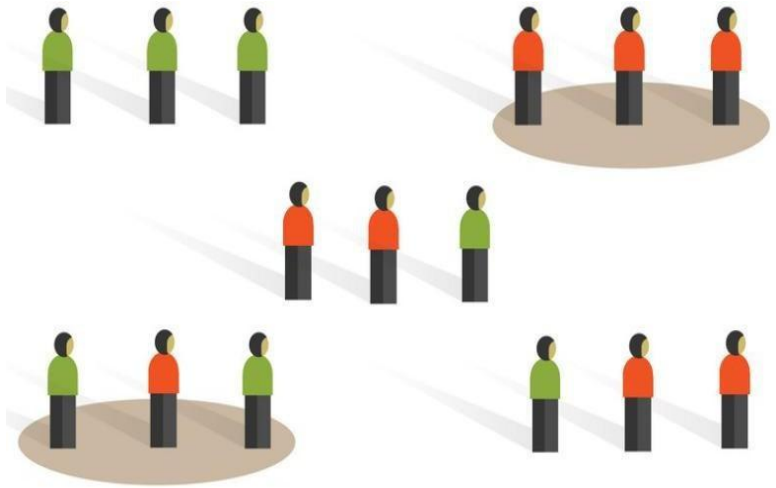


Example: Suppose we would like to select some students from any class consists of 75 students. The students will be divided into groups of boys and girls. Then some students will be chosen from boys and some from the girls.

Advantage: Members of each category or group will be chosen without any bias.

Cluster Sampling

- From the big population, choose a small group by diving it into clusters/sections i.e area wise.
- The clusters are randomly selected.
- All the elements of the cluster are used for sampling.



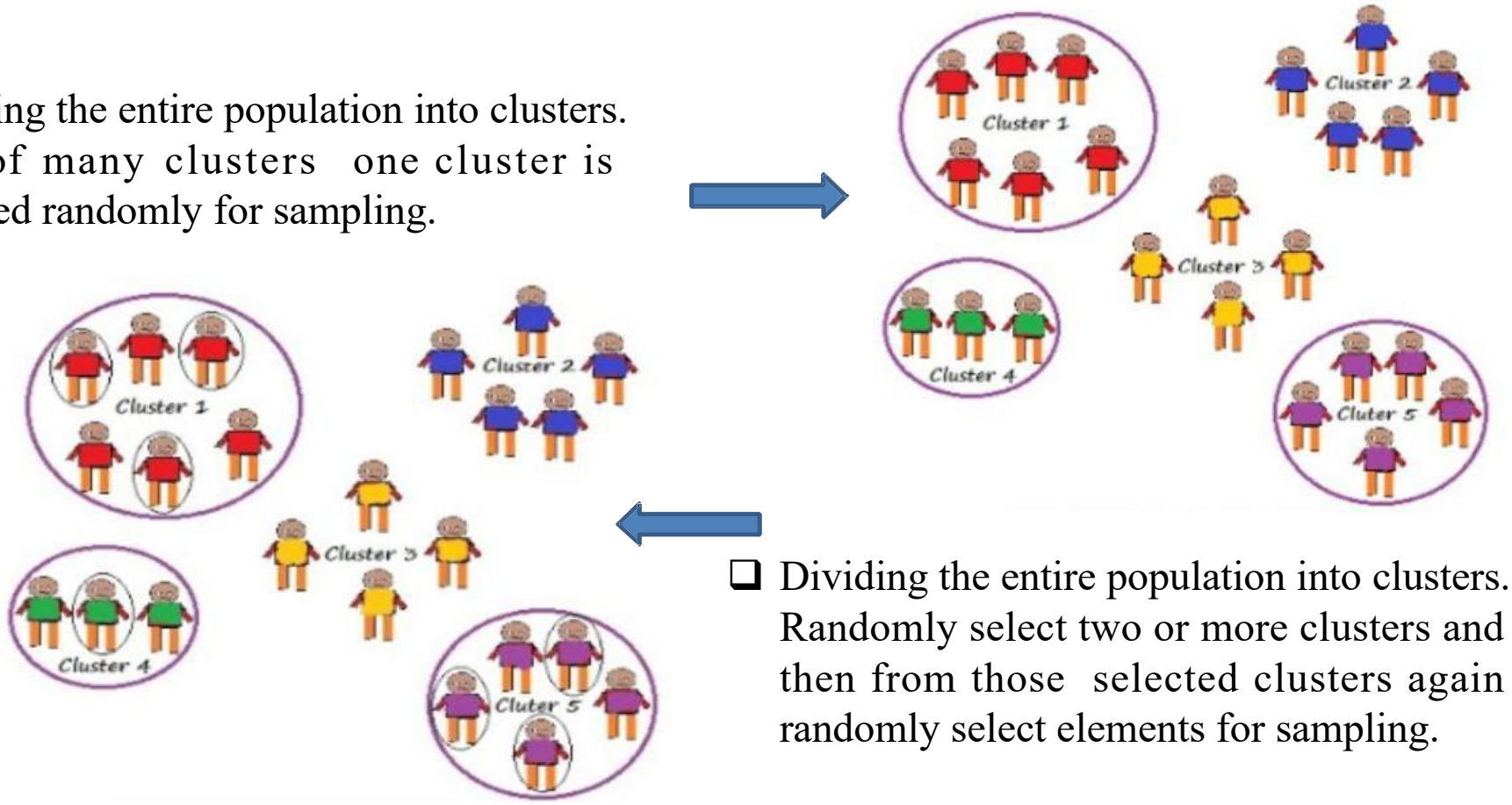
Example: Suppose we would like to know the awareness about COVID in a city. Instead of going the details survey of the entire city one can divide the city into clusters and randomly choose a cluster from that. All the members of the cluster will be considered.

Cluster sampling can be done in following ways:

- **Single Stage Cluster Sampling**
- **Two Stage Cluster Sampling**

Single and Two stage Cluster Sampling

- Dividing the entire population into clusters. Out of many clusters one cluster is selected randomly for sampling.

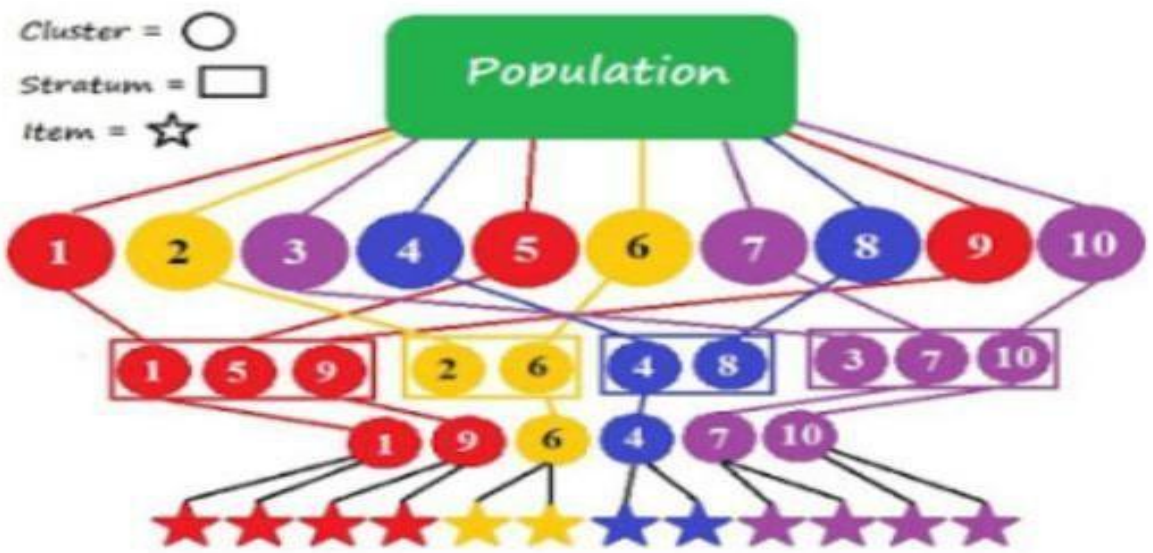


- Dividing the entire population into clusters. Randomly select two or more clusters and then from those selected clusters again randomly select elements for sampling.

Example: An airline company wants to survey its customers one day, so they randomly select 55 flights that day and survey every passenger on those flights.

Multi Stage Sampling

- Population is divided into multiple clusters and then these clusters are further divided and grouped into various sub groups (strata) based on similarity.
- One or more clusters can be randomly selected from each stratum.
- This process continues until the cluster can’t be divided anymore.
- Example : A country can be divided into states, cities, urban and rural and all the areas with similar characteristics can be merged together to form a strata.



Non-Probability Sampling

- Every element will be chosen purposefully/intentionally from the population on the basis of certain past experience and knowledge.
- It is a less stringent method.
- This sampling method depends heavily on the expertise of the researchers.
- It is carried out by observation, and researchers use it widely for qualitative research.
- Mainly classified into
  - **Quota Sampling**
  - **Purpose Sampling/Judgemental Sampling**
  - **Convenience Sampling**
  - **Referral / Snowball Sampling**

Quota Sampling

- Quota sampling works by first dividing the selected population into exclusive subgroups.
- The proportions of each subgroup are measured, and the ratio of selected subgroups are then used in the final sampling process.
- The proportions of the selected subgroups are used as boundaries for selecting a sample population of proportionally represented subgroups.
- There are two types of quota sampling:
  - ✓ **proportional**
  - ✓ **non proportional**

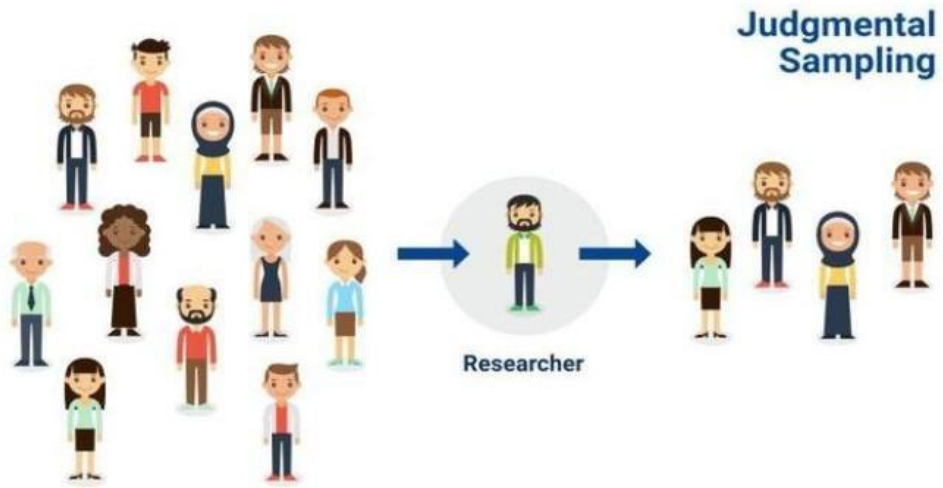
Proportional Quota Sampling

- In proportional quota sampling you want to represent the major characteristics of the population by sampling a proportional amount of each.
- The problem here is that you have to decide the specific characteristics on which you will base the quota. Will it be by gender, age, education race, religion, etc.?
- ***For example, if you know the population has 40% women and 60% men, and that you want a total sample size of 100, you will continue sampling until you get those percentages and then you will stop. So, if you've already got the 40 women for your sample, but not the sixty men, you will continue to sample men but even if legitimate women respondents come along, you will not sample them because you have already "met your quota."***

Non-Proportional Quota Sampling

- Use when it is important to ensure that a number of sub-groups in the field of study are well-covered.
- Use when you want to compare results across sub-groups.
- Use when there is likely to a wide variation in the studied characteristic within minority groups.
  - Identify sub-groups from which you want to ensure sufficient coverage. Specify a minimum sample size from each sub-group.
  - Here, you're not concerned with having numbers that match the proportions in the population. Instead, you simply want to have enough to assure that you will be able to talk about even small groups in the population.
- *Example:A study of the prosperity of ethnic groups across a city, specifies that a minimum of 50 people in ten named groups must be included in the study. The distribution of incomes across each ethnic group is then compared against one another.*

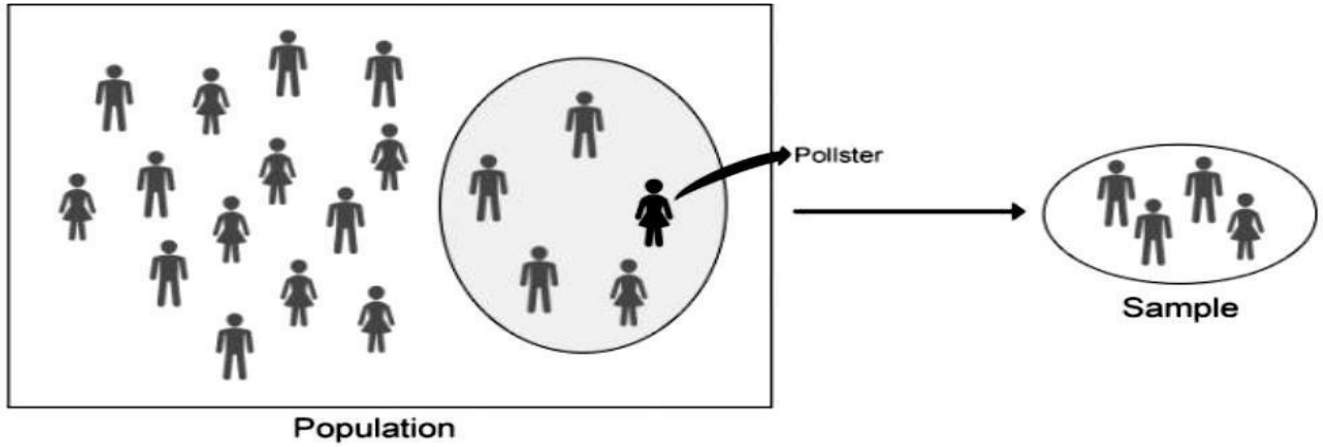
Purpose Sampling/Judgemental Sampling



- Samples are chosen only on the basis of the researcher's knowledge and judgement.
- It enables the researcher to select cases that will best enable him to answer his research questions that meet the objective.
- Choosing a sample because of represent the certain purpose.
- ***Example-1: In online live voting for selecting a GOOD Singer from a competition, the people who have interest in singing can be selected in the sample .***
- ***Example-2: If we want to understand the thought process of the people who are interested in pursuing master's degree then the selection criteria would be "Are you interested for Masters in..?"***
- ***All the people who respond with a "No" will be excluded from our sample.***

Convenience Sampling

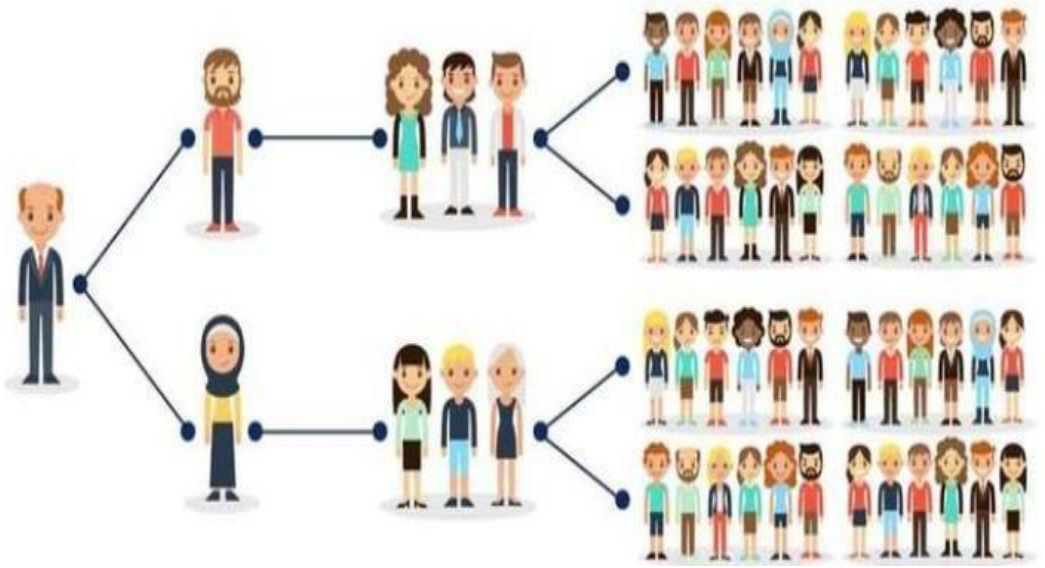
- Convenience sampling (also called accidental sampling or grab sampling) is where you include people who are easy to reach.
- Sample are taken mainly on basis of the readily available.
- Sample which is convenient to the researcher or the data analyst can be chosen. The task is done without any principles or theories.
- For example, you could survey people from:
  - ✓ Your workplace,
  - ✓ Your school,
  - ✓ A club you belong to,
  - ✓ The local mall.



- ***Example: Suppose I would like to select 5 students from any class consists of 75 students. Choosing the 5 students who sits near by me without any principle of selection.***

Referral / Snowball Sampling

- Snowball sampling method is purely based on referrals and that is how a researcher is able to generate a sample.
- So the researcher will take the help from the first element which he select for the population and ask him to recommend others who will fit for the description of the sample needed.
- So this referral technique goes on, increasing the size of population like a snowball.



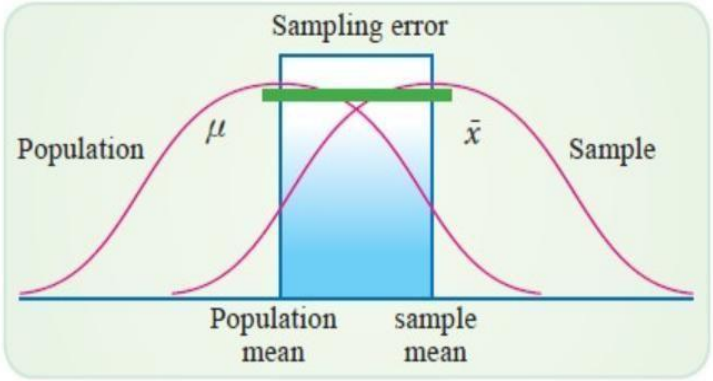
Example: ***If you are studying the level of customer satisfaction among the members of an elite country club, you will find it extremely difficult to collect primary data sources unless a member of the club agrees to have a direct conversation with you and provides the contact details of the other members of the club.***

Sampling Errors

- Sampling error is a statistical error that occurs when an analyst does not select a sample that represents the entire population of data.
- The results found in the sample thus do not represent the results that would be obtained from the entire population.
- Sampling error can be reduced by randomizing sample selection increasing the number of observations.
- It mainly happens when the sample size is very small (10 to 100).

For example, if you wanted to figure out how many people out of a thousand were under 18, and you came up with the figure 19.357%. If the actual percentage equals 19.300%, the difference (19.357 – 19.300) of 0.57 or 3% = the margin of error. If you continued to take samples of 1,000 people, you'd probably get slightly different statistics, 19.1%, 18.9%, 19.5% etc, but they would all be around the same figure. This is one of the reasons that you'll often see sample sizes of 1,000 or 1,500 in surveys: they produce a very acceptable margin of error of about 3%.

Formula: the formula for the margin of error is $1/\sqrt{n}$, where n is the size of the sample. For example, a random sample of 1,000 has about a $1/\sqrt{n}$; = 3.2% error.



Five Common Types of Sampling Errors

- **Population Specification Error**—This error occurs when the researcher does **not understand who they should survey**.
- **Sample Frame Error**—A frame error occurs when **the wrong sub-population is used to select a sample**.
- **Selection Error**—This occurs when **respondents self-select their participation in the study – only those that are interested respond**. Selection error can be controlled by going extra lengths to get participation.
- **Non-Response**—**Non-response errors occur when respondents are different than those who do not respond**. This may occur because either the potential respondent was not contacted or they refused to respond.
- **Sampling Errors**—These errors occur because of **variation in the number or representativeness of the sample that responds**. Sampling errors can be controlled by (1) careful sample designs, (2) large samples, and (3) multiple contacts to assure representative response.

- **Data Quality and Major Tasks in Data Preprocessing**
- Data Cleaning
- Data Integration
- Data Transformation and Data Discretization
- Data Reduction

Data Preprocessing

- Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogenous sources.
- **Low-quality data** will lead to low-quality mining results.
- "How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the mining results?
- How can the data be preprocessed so as to improve the efficiency and ease of the mining process?"
- Data preprocessing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining.

Data Quality

- What kinds of data quality problems?
- How can we detect problems with the data?
- What can we do about these problems?

- Examples of data quality problems:
  - **Noise and outliers**
    - **Noise**: random error or variance in a measured variable
    - **Outliers** are data objects with characteristics that are considerably different than most of the other data objects in the data set
  - **Missing values**
  - **Duplicate data**

Data Quality
Missing Values and Duplicate Data

- **Reasons for missing values**
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)

- **Handling missing values**
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

- Data set may include data objects that are **duplicates**, or almost duplicates of one another
  - Major issue when merging data from heterogenous sources

Data Quality: How do we know data have quality?

- **Data have quality if they satisfy the requirements of the intended use.**

- Measures for data quality: A multidimensional view
  - **Accuracy**: correct or wrong, accurate or not
  - **Completeness**: not recorded, unavailable
  - **Consistency**: some modified but some not, different values for same data object
  - **Timeliness**: timely update?
  - **Believability**: how trustable the data are correct?
  - **Interpretability**: how easily the data can be understood?

Data Quality: How do we know data have quality?

Accuracy: correct or wrong, accurate or not

- There are many possible reasons for inaccurate data.
  - Human or computer errors occurring at data entry.
  - Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information such as choosing the default value "January 1" displayed for birthday.
  - Incorrect data may also result from inconsistencies in naming conventions or data codes, or inconsistent formats for input fields (e.g., date).

Completeness: not recorded, unavailable

- Attributes of interest may not always be available
- Data may not be included simply because they were not considered important at the time of entry.
- Relevant data may not be recorded due to a misunderstanding
- Missing data, tuples with missing values for some attributes, may need to be inferred.

Data Quality: How do we know data have quality?

Consistency: some modified but some not, inconsistent data, ...

- Containing discrepancies in the department codes used to categorize items.
- Inconsistencies in data codes, or inconsistent formats for input fields (e.g., *date*).

Timeliness: timely update?

- Is the data updated regularly?
- For example, data of sales of few days or months might be missing.

Believability: how trustable the data are correct?

- How much the data are trusted by users?
- The past errors can effect the trustability of the data.

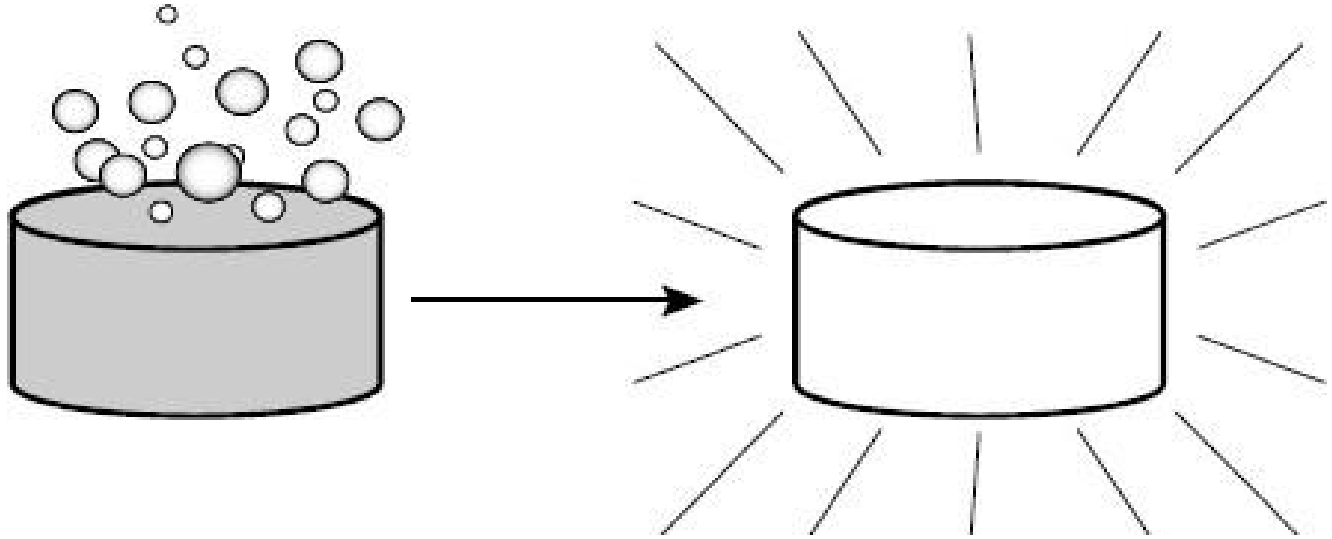
Interpretability: how easily the data can be understood?

Major Tasks in Data Preprocessing

- **Data cleaning** can be applied to remove noise and correct inconsistencies in the data.
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration** merges data from multiple sources into a coherent data store, such as a data warehouse.
  - Integration of multiple databases, data cubes, or files
- **Data reduction** can reduce the data size by aggregating, eliminating redundant features, or clustering.
  - Dimensionality reduction, Numerosity reduction, Data compression
- **Data transformations and Data Discretization**, such as normalization, may be applied.
  - For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements.
  - Concept hierarchy generation

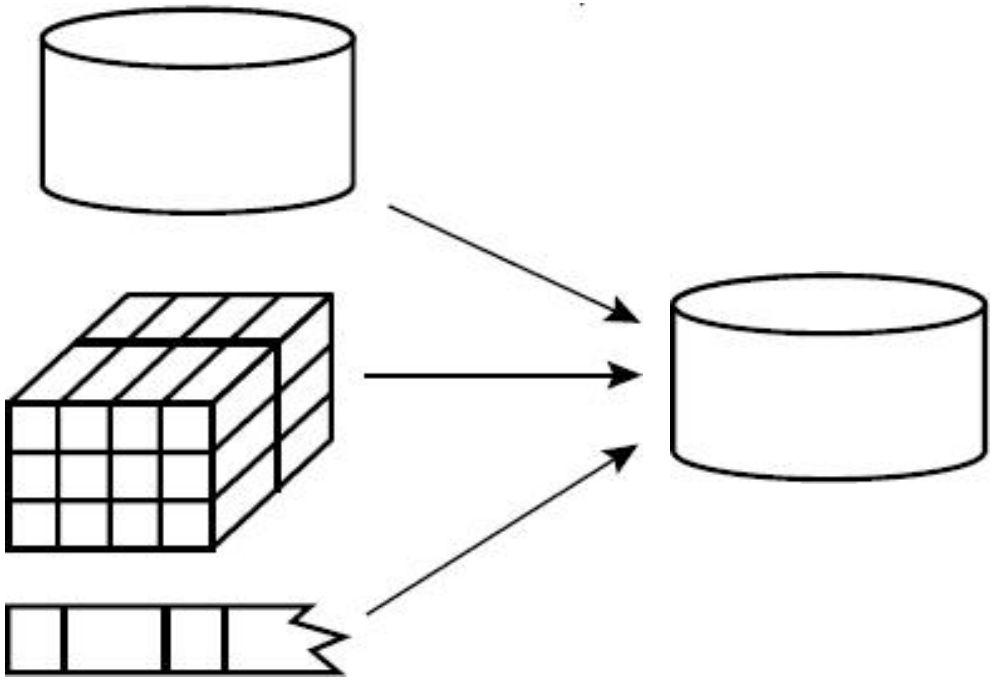
Major Tasks in Data Preprocessing
Data Cleaning

- **Data cleaning** routines work to "clean" the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.
  - If users believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied to it.
  - Dirty data can cause confusion for the mining procedure, resulting in unreliable output



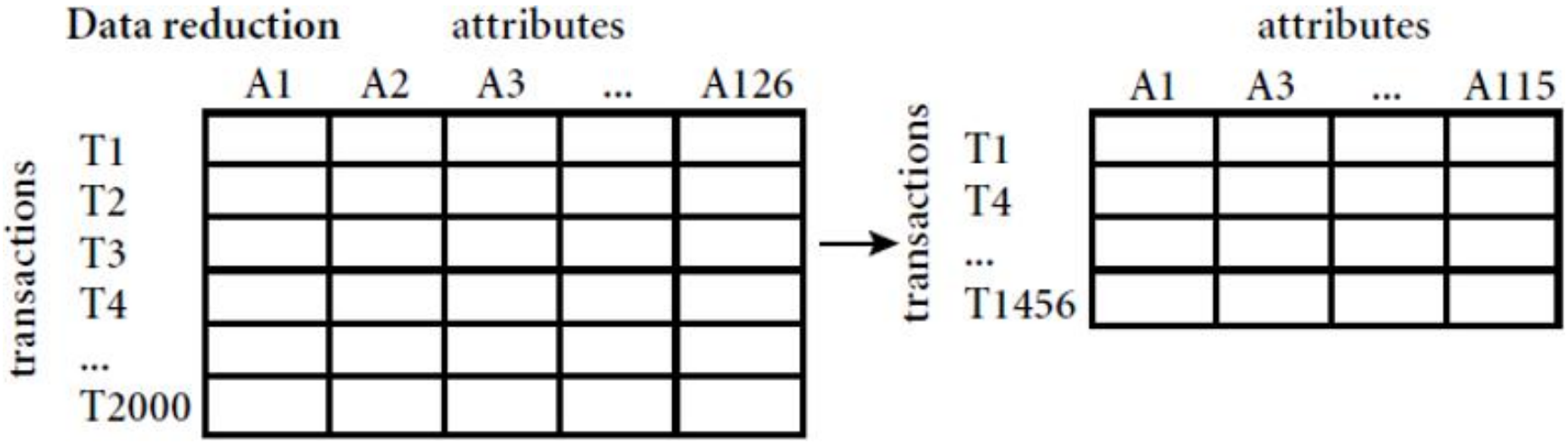
Major Tasks in Data Preprocessing
Data Integration

- **Data integration** merges data from multiple sources into a coherent data store, such as a data warehouse.



Major Tasks in Data Preprocessing
Data Reduction

- **Data reduction** obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results.
- Data reduction strategies include **dimensionality reduction** and **numerosity reduction**.



Major Tasks in Data Preprocessing
Data transformations and Data Discretization

- The data are transformed or consolidated so that the resulting mining process may be more efficient, and the patterns found may be easier to understand.

- **Data discretization** is a form of **data transformation**.
  - Data discretization transforms numeric data by mapping values to interval or concept labels.

- Data Transformation: Normalization

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

- Data Quality and Major Tasks in Data Preprocessing
- **Data Cleaning**
- Data Integration
- Data Transformation and Data Discretization
- Data Reduction

Data Cleaning

- Data in the real world is dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
  - **incomplete:** lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., *Occupation* = " " (missing data)
  - **noisy:** containing noise, errors, or outliers
    - e.g., *Salary* = "−10" (an error)
  - **inconsistent:** containing discrepancies in codes or names, e.g.,
    - *Age* = "42", *Birthday* = "03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - **intentional:** (e.g., *disguised missing* data)
    - Jan. 1 as everyone's birthday?

Incomplete (Missing) Data

- **Data is not always available**
  - E.g., many tuples have no recorded value for several attributes, such as customer income in sales data.
- **Missing data may be due to**
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data
- Missing data may need to be inferred.

How to Handle Missing Data?

- **Ignore the tuple**: usually done when class label is missing (when doing classification)
  - can be used if many attributes have missing values of a data object.
  - not effective when the % of missing values per attribute varies considerably.
- **Fill in the missing value manually**: tedious and time consuming approach
  - may not be feasible for large datasets.
- **Use a global constant value to fill in the missing values**
  - Replace all missing attribute values by the same constant such as a label like "Unknown" or $-\infty$
- **Use mean or median of all attribute values to fill in the missing values.**
- **Use the attribute mean or median for all samples belonging to the same class as the given data object.**
- **Use the most probable value to fill in the missing value -** Use data mining techniques or models to predict missing values.

Noisy Data and How to Handle Noisy Data?

- **Noise**: random error or variance in a measured variable
- Outliers may represent noise.
- Given a numeric attribute such as, say, *price*, how can we "**smooth**" out the data to remove the noise?

Data Smoothing Techniques:

- **Binning**
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- **Regression**
  - smooth by fitting the data into regression functions
- **Clustering**
  - detect and remove outliers
- **Combined computer and human inspection**
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Binning Methods for Data Smoothing

Binning methods smooth a sorted data by distributing them into bins (buckets).

Smoothing by bin means:

- Each value in a bin is replaced by the mean value of the bin.

Smoothing by bin medians:

- Each bin value is replaced by the bin median.

Smoothing by bin boundaries:

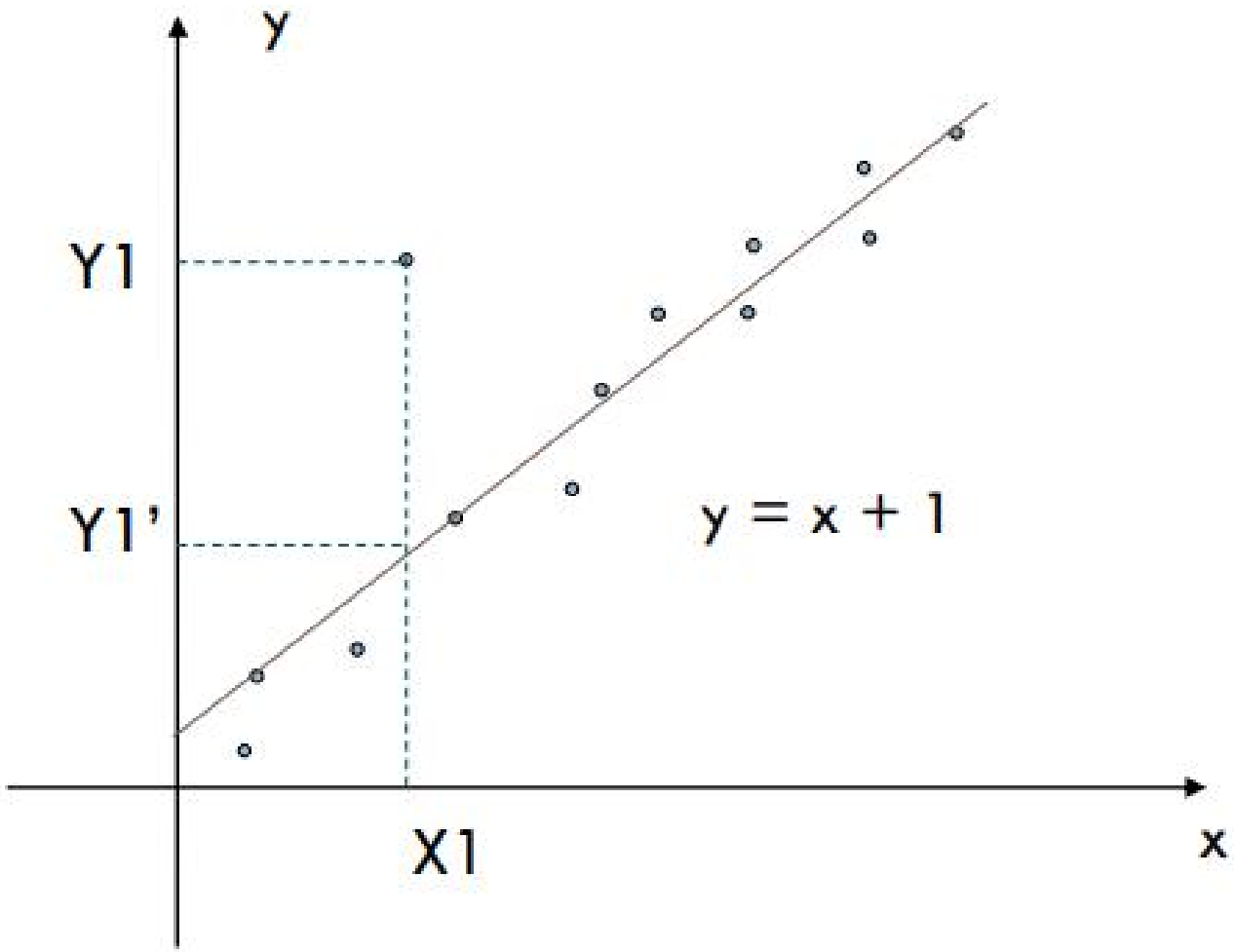
- The minimum and maximum values in a given bin are identified as the bin boundaries.
- Each bin value is then replaced by the closest boundary value.

Binning Methods for Data Smoothing: Example

- Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34
- **Partition into (equal-frequency) bins**:
  - Bin 1: 4, 8, 15
  - Bin 2: 21, 21, 24
  - Bin 3: 25, 28, 34
- **Smoothing by bin means**:
  - Bin 1: 9, 9, 9
  - Bin 2: 22, 22, 22
  - Bin 3: 29, 29, 29
- **Smoothing by bin medians**:
  - Bin 1: 8, 8, 8
  - Bin 2: 21, 21, 21
  - Bin 3: 28, 28, 28
- **Smoothing by bin boundaries**:
  - Bin 1: 4, 4, 15
  - Bin 2: 21, 21, 24
  - Bin 3: 25, 25, 34

Handling Noise: Regression

- Replace noisy or missing values by predicted values
- Requires model of attribute dependencies (maybe wrong!)
- Can be used for data smoothing or for handling missing data



Data Smoothing

- *Many methods for data smoothing* are also methods for *data reduction involving discretization*.
  - For example, the binning techniques reduce the number of distinct values per attribute.
    - This acts as a form of data reduction for logic-based data mining methods, such as decision tree induction, which repeatedly make value comparisons on sorted data.

- Concept hierarchies are a form of data discretization that can also be used for data smoothing.
  - A concept hierarchy for price, for example, may map real price values into inexpensive, moderately priced, and expensive, thereby reducing the number of data values to be handled by the mining process.

Data Cleaning as a Process

- **Data discrepancy detection**
  - Use metadata (e.g., domain, range, dependency, distribution)
  - Check uniqueness rule, consecutive rule and null rule
  - For example, values that are more than two standard deviations away from the mean for a given attribute may be flagged as potential outliers.
  - Use commercial tools
    - Data scrubbing: use simple domain knowledge (e.g., postal code, spell-check) to detect errors and make corrections
    - Data auditing: by analyzing data to discover rules and relationship to detect violators (e.g., correlation and clustering to find outliers)
- **Data migration and integration**
  - Data migration tools: allow transformations to be specified
  - ETL (Extraction/Transformation/Loading) tools: allow users to specify transformations through a graphical user interface
- **Integration of the two processes**
  - Iterative and interactive (e.g., Potter's Wheels is a data cleaning tool)

- Data Quality and Major Tasks in Data Preprocessing
- Data Cleaning
- **Data Integration**
- Data Transformation and Data Discretization
- Data Reduction

Data Integration

- Data mining often requires data integration—the merging of data from multiple data sources.
- Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set. This can help improve the accuracy and speed of the subsequent data mining process.
- The semantic heterogeneity and structure of data pose great challenges in data integration.

Entity Identification Problem

- One problem that arise in data integration is that, How can equivalent real-world entities from multiple data sources be matched up?
- For example, how can a data analyst or a computer be sure that *customer_id* in one database and *cust_number* in another refer to the same attribute?
- This is known as ***entity identification problem.***
- To solve this problem, we can use metadata for entity identification. Metadata is extra information regarding the data.
- Examples of metadata for each attribute include the name, meaning, data type, range of values permitted for the attribute, and null rules for handling blank, zero, or null values.
- Such metadata can be used to help avoid errors in *schema integration*.

Schema and Metadata

```sql
CREATE TABLE Students (
    StudentID INT PRIMARY KEY,
    FirstName VARCHAR(50),
    LastName VARCHAR(50),
    DateOfBirth DATE,
    Gender CHAR(1)
);
```

- **Table Name:** Students
- **Columns:**
  - `StudentID`: INT, Primary Key, Unique Identifier for each student.
  - `FirstName`: VARCHAR(50), Student's first name.
  - `LastName`: VARCHAR(50), Student's last name.
  - `DateOfBirth`: DATE, Student's date of birth.
  - `Gender`: CHAR(1), Student's gender (M/F).

Issues in Data Integration

- **Redundancy**
  - An attribute may be redundant if it can be “derived” from another attribute or set of attributes. For example Age can be derived from birthdate.
- **Duplication**
  - While merging data from multiple sources, some data objects can be added multiple times resulting in duplicates in the dataset.
- **Data value conflicts**
  - For the same real-world entity, attribute values from different sources may differ. This may be due to differences in representation, scaling, or encoding.
  - For instance, a weight attribute may be stored in metric units (in Kg) in one system and British imperial units (in pounds) in another.

Data Transformation

- In data transformation, the data are transformed or consolidated into forms appropriate for mining.
- Through appropriate data transformation, the resulting mining process may be more efficient, and the patterns found may be easier to understand.
- Data Transformation involves techniques like -
  - Normalization
  - Discretization
  - Data Compression

Data Transformation: Normalization

- Representing data into different units, may result in different range of values.
- To help avoid dependence on the choice of measurement units, the data should be normalized or standardized. This involves transforming the data to fall within a smaller or common range such as [−1.0, 1.0] or [0.0, 1.0].
- Let A be a numeric attribute with n values $v_1, v_2, \ldots, v_n$. We can normalize this data in different ways.

Data Transformation: Normalization

- min-max normalization

$$v' = \frac{v - min}{max - min}(new\_max - new\_min) + new\_min$$

- z-score normalization

$$v' = \frac{v - mean}{stand\_dev}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that $\text{Max}(|v'|) < 1$

Normalization Example

- Suppose you are given 12 values for an attribute

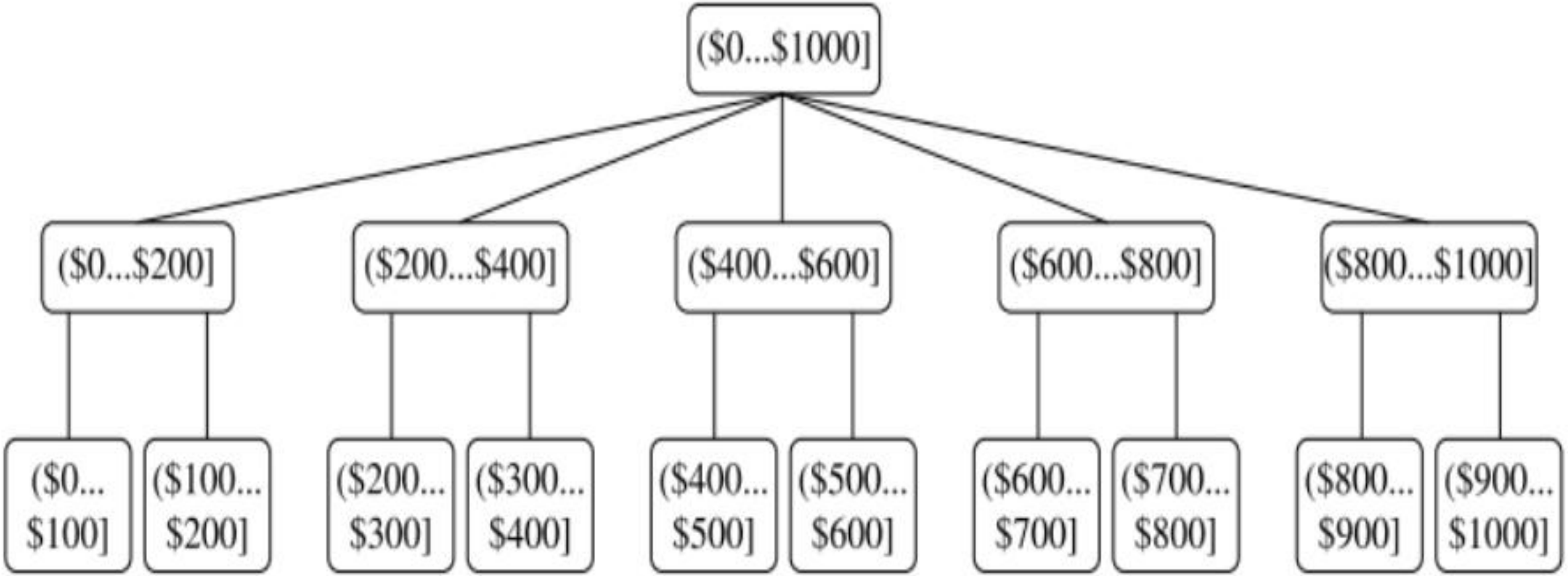
$$X = \{-100, -150, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55\}$$

- Normalize above data so that each value belongs to range [5, 10]
- Normalize above data using *z-normalization*.
- Normalize above data using *decimal scaling*

Data Discretization

- Data discretization is a common data transformation technique, where the raw values of a numeric attribute (e.g., age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior).
- The labels, in turn, can be recursively organized into higher-level concepts, resulting in a concept hierarchy for the numeric attribute.

Data Discretization: Concept Hierarchy



Data Discretization

- Discretization techniques can be **categorized based on how the discretization is performed**, such as whether it uses class information or which direction it proceeds (i.e., top-down vs. bottom-up).
- If the discretization process **uses class information, then we say it is supervised discretization. Otherwise, it is unsupervised**.
- If the process starts by **first finding one or a few points (called split points or cut points) to split the entire attribute range** and then repeats this recursively on the resulting intervals, it is called top-down discretization or splitting.
- This contrasts with **bottom-up discretization or merging, which starts by considering all of the continuous values as potential split-points, removes some by merging neighborhood values** to form intervals, and then recursively applies this process to the resulting intervals.

Discretization by binning

- Binning is a top-down splitting technique based on a specified number of bins.
- Attribute values can be discretized by applying *equal-width or equal-frequency* binning and then replacing each bin value by the bin mean or median, as in smoothing by bin means or smoothing by bin medians, respectively.
- Binning does not use class information and is therefore an unsupervised discretization technique. It is sensitive to the user-specified number of bins, as well as the presence of outliers.

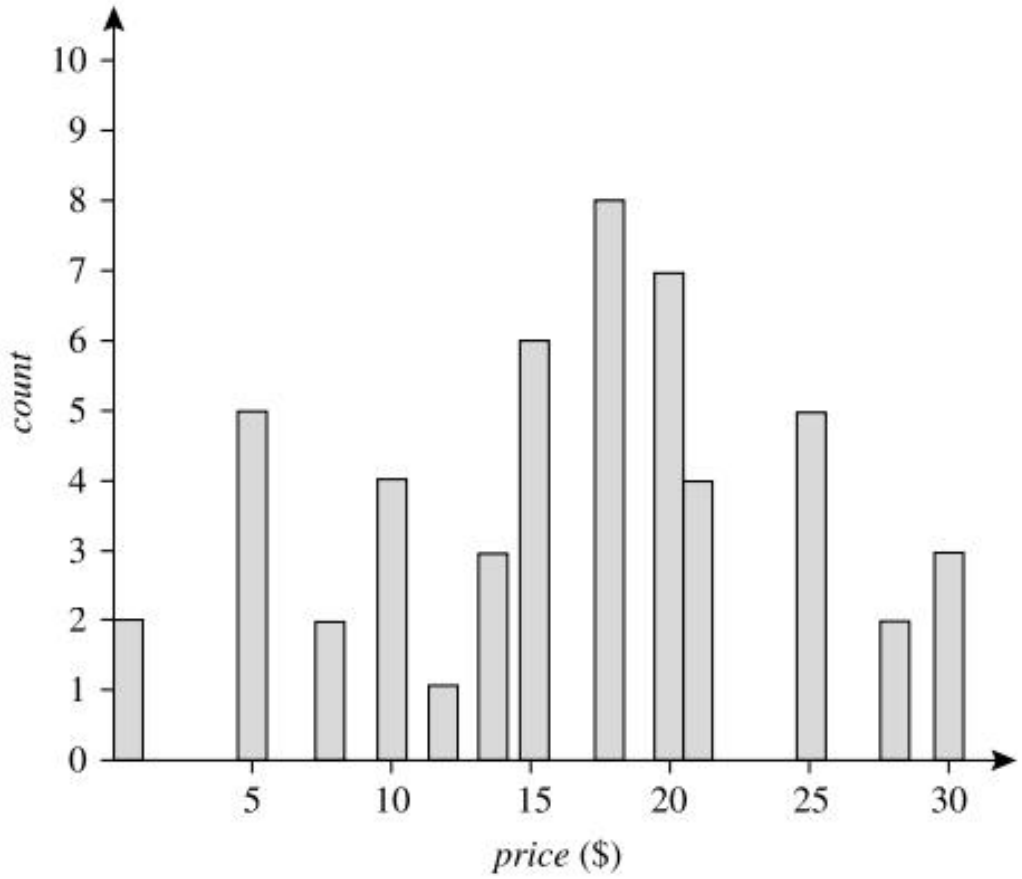
Discretization by Histogram Analysis

- Histogram analysis is an unsupervised discretization technique because it does not use class information.
- The following data are a list of prices for commonly sold items in the company (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, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.
- Can you create a histogram for this data?
- How would you decide ranges?

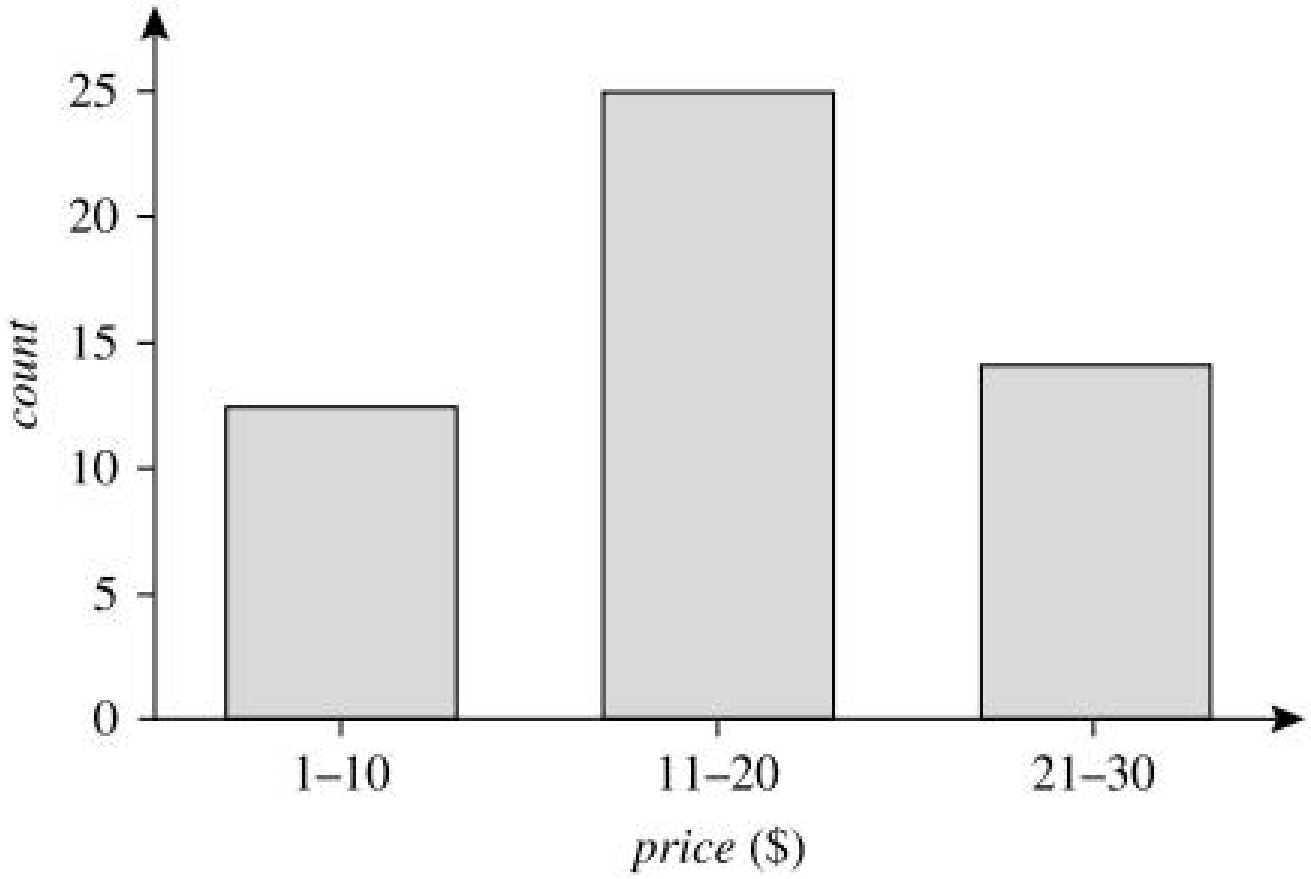
Value	Frequency
1	2
5	5
8	2
10	4
12	1
14	3
15	6
18	8
20	7
21	4
25	5
28	2
30	3

Discretization by Histogram Analysis

- Ranges can be decided in two ways-
  - Equal Width
  - Equal Frequency



Discretization by Histogram Analysis



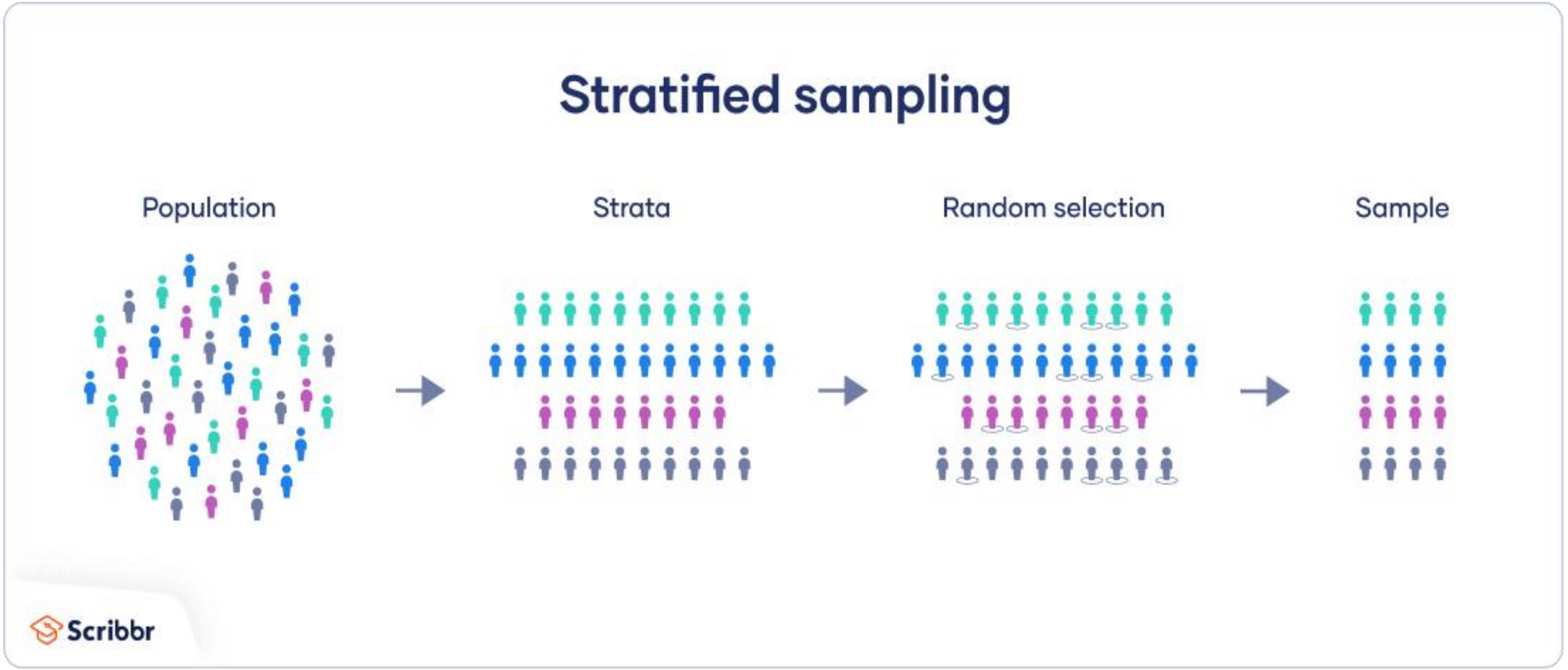
Sampling

- Data sampling is the process of selecting a subset of data from a larger dataset. This subset, known as a **sample**, is used to make inferences or draw conclusions about the **entire dataset (population)**.
- Sampling is crucial in various fields such as statistics, data science, and machine learning, especially when working with large datasets where processing the entire population would be impractical or impossible.
- Suppose we have a large dataset, ***D*** which contains ***N*** data objects. How can I create a sample which contains ***s*** data object?

Sampling Techniques

- **Simple random sample without replacement of size *s*:** This is created by drawing *s* samples from *D*, and every time a sample is drawn, it is not to be placed back to the data set *D*.
- **Simple random sample with replacement of size *s*:** Each time a sample is drawn from *D*, it is recorded and then placed back in the data set.
- **Cluster Sample:** Group objects of *dataset D* of *into M mutually disjoint "clusters", then a sample of s clusters can be obtained where s < M*.
- **Stratified Sample:** Divide *D* into mutually disjoint parts called *strata*. Then create a stratified sample by obtaining a sample at each strata and combining those samples.
  - This helps ensure a representative sample, especially when the data are skewed. For example, a stratified sample may be obtained from customer data, where a stratum is created for each customer age group. In this way, the age group having the smallest number of customers will be sure to be represented.

Sampling Techniques



Sampling

- An advantage of sampling for data reduction is that the cost of obtaining a sample is proportional to the size of the sample, s, as opposed to N, the data set size. Hence, sampling complexity is potentially sublinear to the size of the data.

Statistical Measures for Sample

Suppose we have a large dataset **D** (Population) which contains **N** data objects. Out of these, we have selected a sample **X** which only contains **s** data objects.

So the basic statistics for both population and sample are -

1. **Mean**

   a. Population Mean

$$\mu = \frac{\sum_{i=1}^{N} v_i}{N}$$

   b. Sample Mean

$$\bar{X} = \frac{\sum_{i=1}^{s} v_i}{s}$$

Statistical Measures for Sample

2. **Variance** and **Standard** Deviation

a. For population

$$\sigma^2 = \frac{\sum_{i=1}^{N}(v_i - \mu)^2}{N}$$

b. For sample

$$S^2 = \frac{\sum_{i=1}^{s}(v_i - \mu)^2}{s-1}$$

3. **Covariance**

a. For population

$$COV(X,Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N}$$

b. For sample

$$COV(X,Y) = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{s-1}$$

Exercise

Let the minimum, maximum, mean and standard deviation values for the attribute income of data scientists be ₹46000, ₹170000, ₹96000, and ₹21000, respectively.

The z-score normalized income value of ₹106000 is closest to which ONE of the following options?

a. 0.217

b. 0.476

c. 623

d. 2.304

Exercise

Two fair coins are tossed independently. X is a random variable that takes a value of 1 if both tosses are heads and 0 otherwise. Y is a random variable that takes a value of 1 if at least one of the tosses is heads and 0 otherwise.

The value of the covariance of X and Y is _____ (rounded off to **three** decimal places).

Exercise

Let X be a random variable exponentially distributed with parameter $\lambda > 0$. The probability density function of X is given by:

$$f_X(x) = \begin{cases} \lambda\, e^{-\lambda x}, & x \geq 0 \\ 0, & otherwise \end{cases}$$

If $5E(X) = Var(X)$, where $E(X)$ and $Var(X)$ indicate the expectation and variance of X, respectively, the value of λ is _____ (rounded off to **one** decimal place).