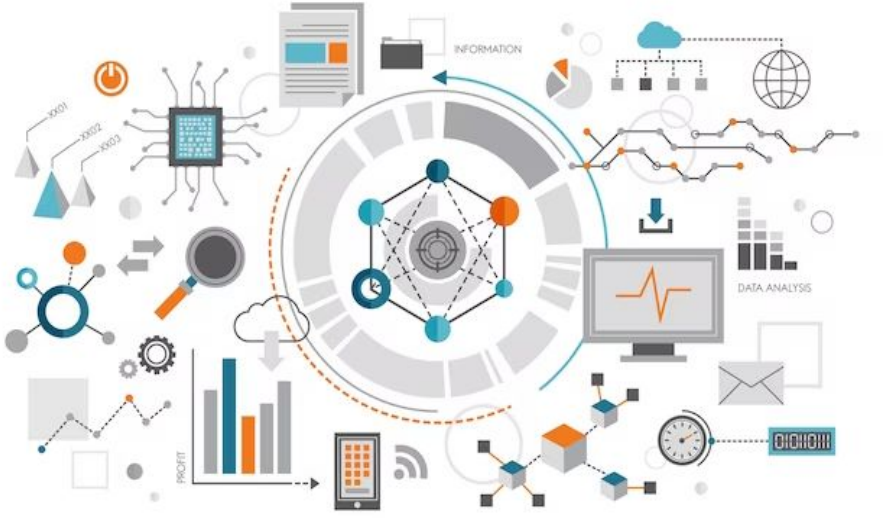
UNIT 1

Introduction to Data Science



Contents

Unit 1

Basics and need of Data Science

- Applications of Data Science
- Relationship between Data Science and Information Science
- Business intelligence versus Data Science

Data

- Data Types
- Data Collection
- Need of Data wrangling

Methods

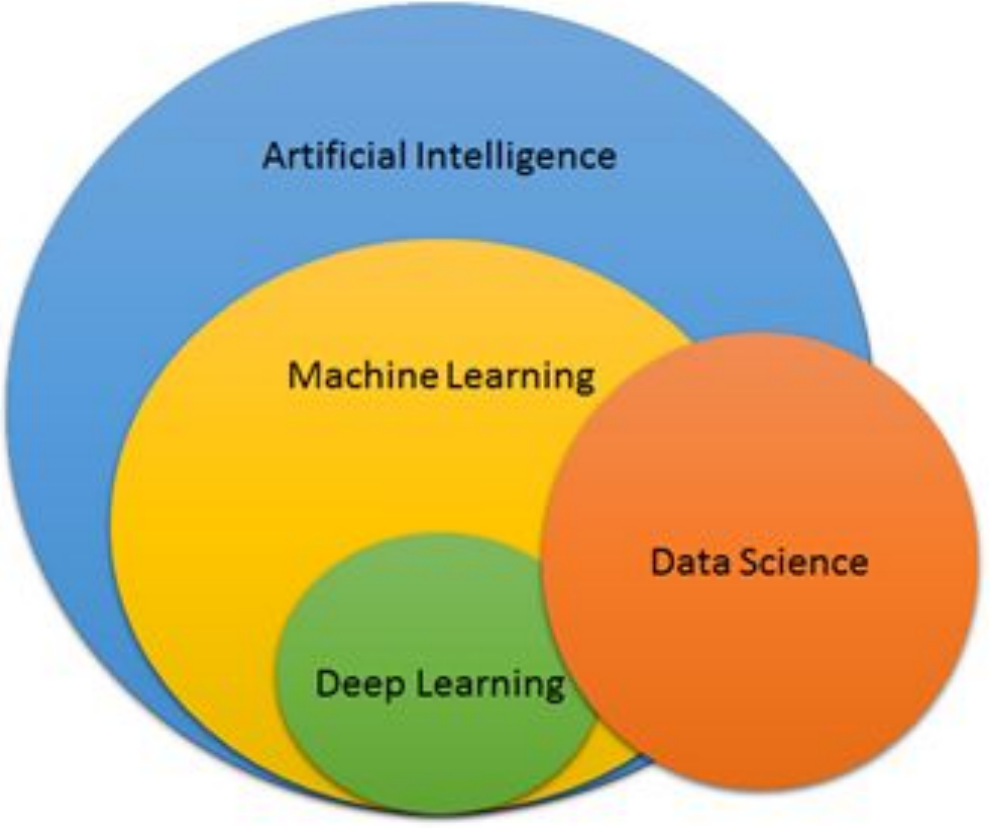
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation & Data Discretization

Course Outcome

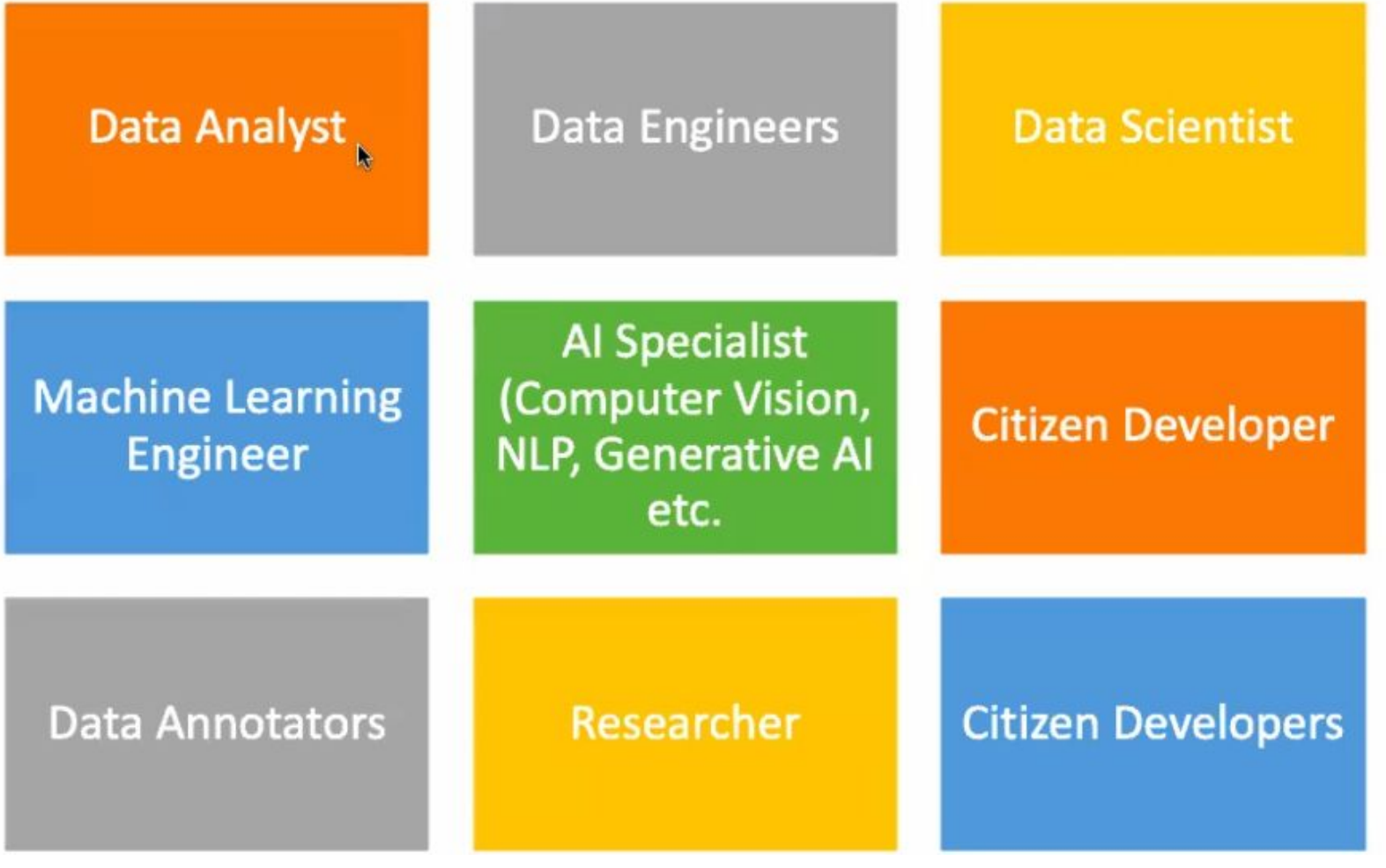
Unit 1

CO1: Analyze needs and challenges for Data Science

intersection of AI, ML, DL and DS



Data Science Career Opportunities



- **Citizen developers** are individuals who create or modify software applications for their own use or their team's needs, without having formal training as professional developers
- **Non-Technical Background**: They are not formally trained in programming but may have a basic understanding of technology.
- **Business-Oriented Solutions**: They focus on solving specific business problems or improving workflows within their departments.

Why Data Science?

- We live in a world where vast amounts of data are collected daily. Analyzing such data is an important need

- “We are living in the information age” is a popular saying; however, **we are actually living in the data age.**
- Terabytes or petabytes of data generated from our computer networks, the World Wide Web (WWW), and various data storage devices every day from business, society, science and engineering, medicine, and almost every other aspect of daily life

Why data Science?

Data Challenges

Data Imbalance: Unequal distribution of classes (e.g., fraud detection)

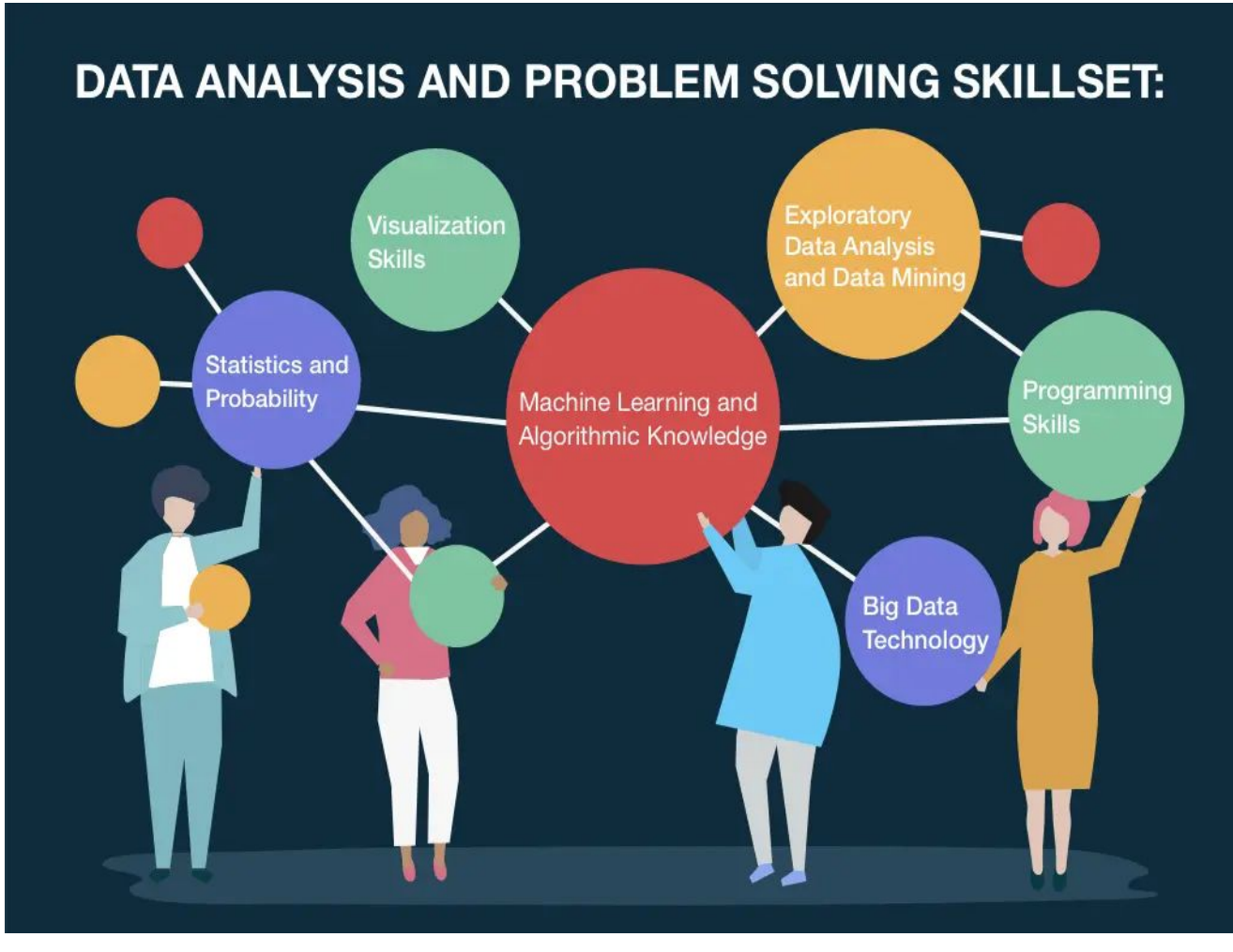
Bias in Data: Leads to biased models (e.g., gender or racial bias)

Data Drift: Changes in data over time, reducing model accuracy

Key Message: Addressing data challenges is crucial for reliable models.

Data Analytics

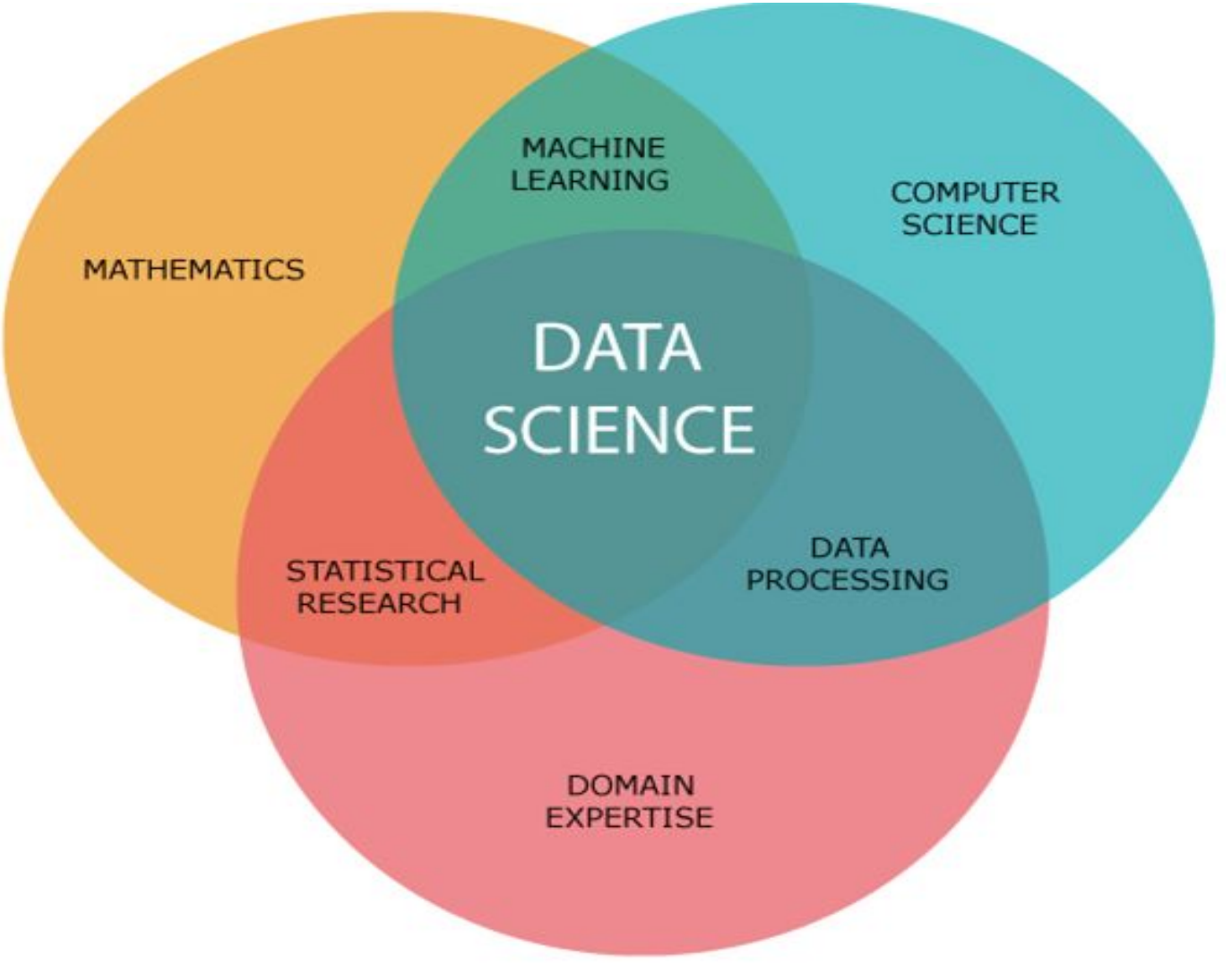
Analytics is utilizing data, machine learning, statistical analysis and computer-based models to get better insight and make better decisions from the data.



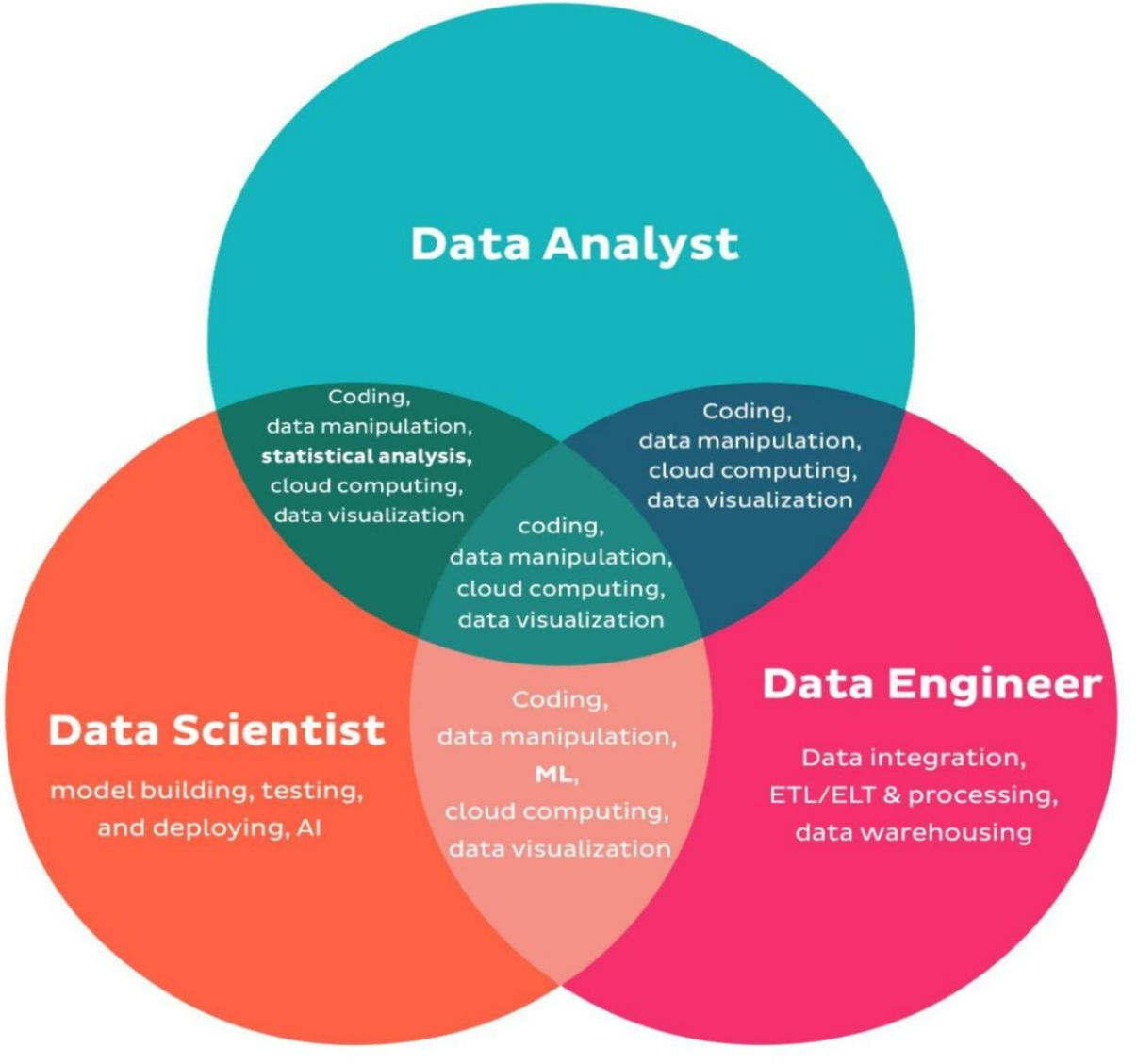
Data Science

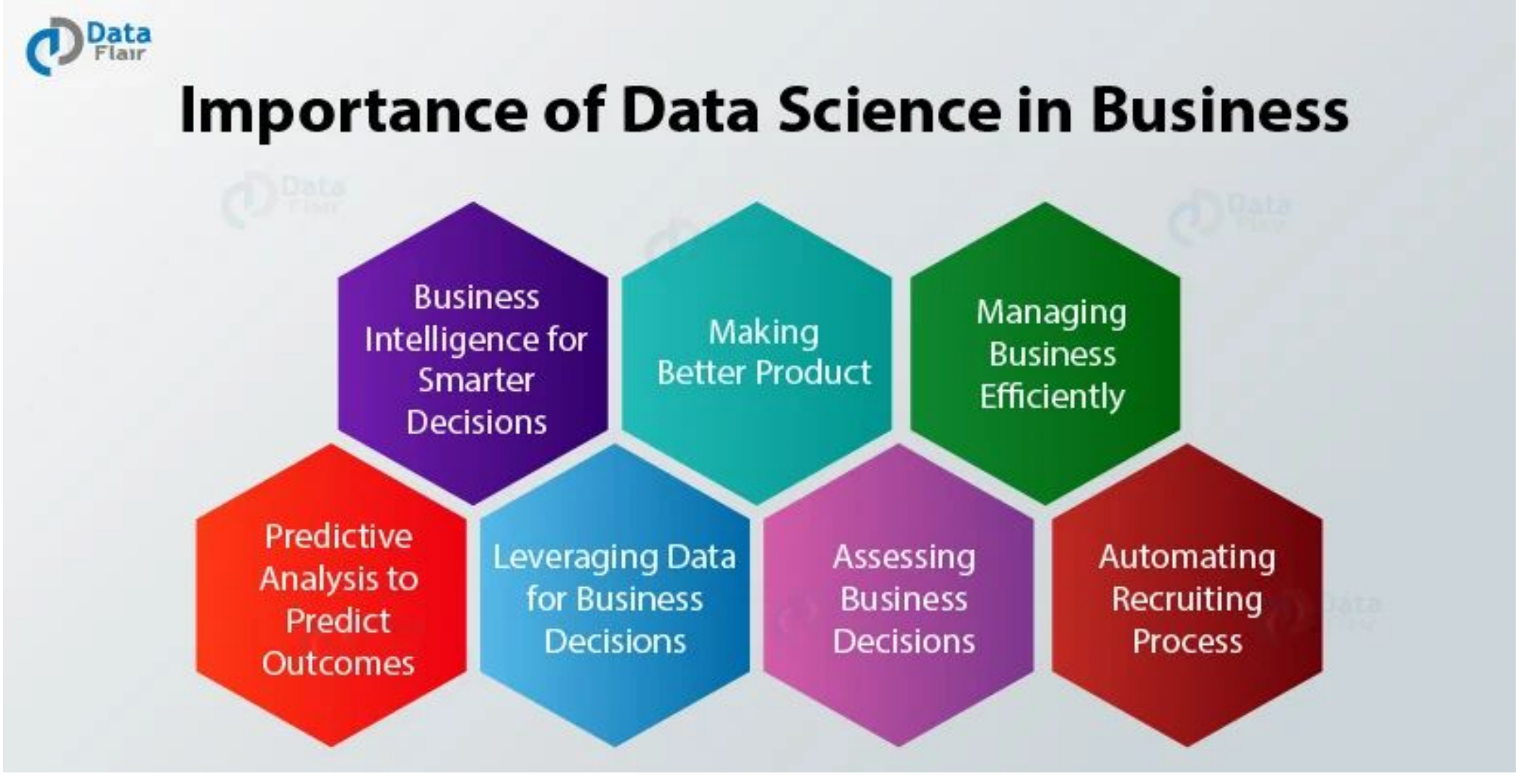
- A field focused on extracting insights and knowledge from structured and unstructured data using techniques like statistical analysis, machine learning, and data visualization.

Basics of Data Science



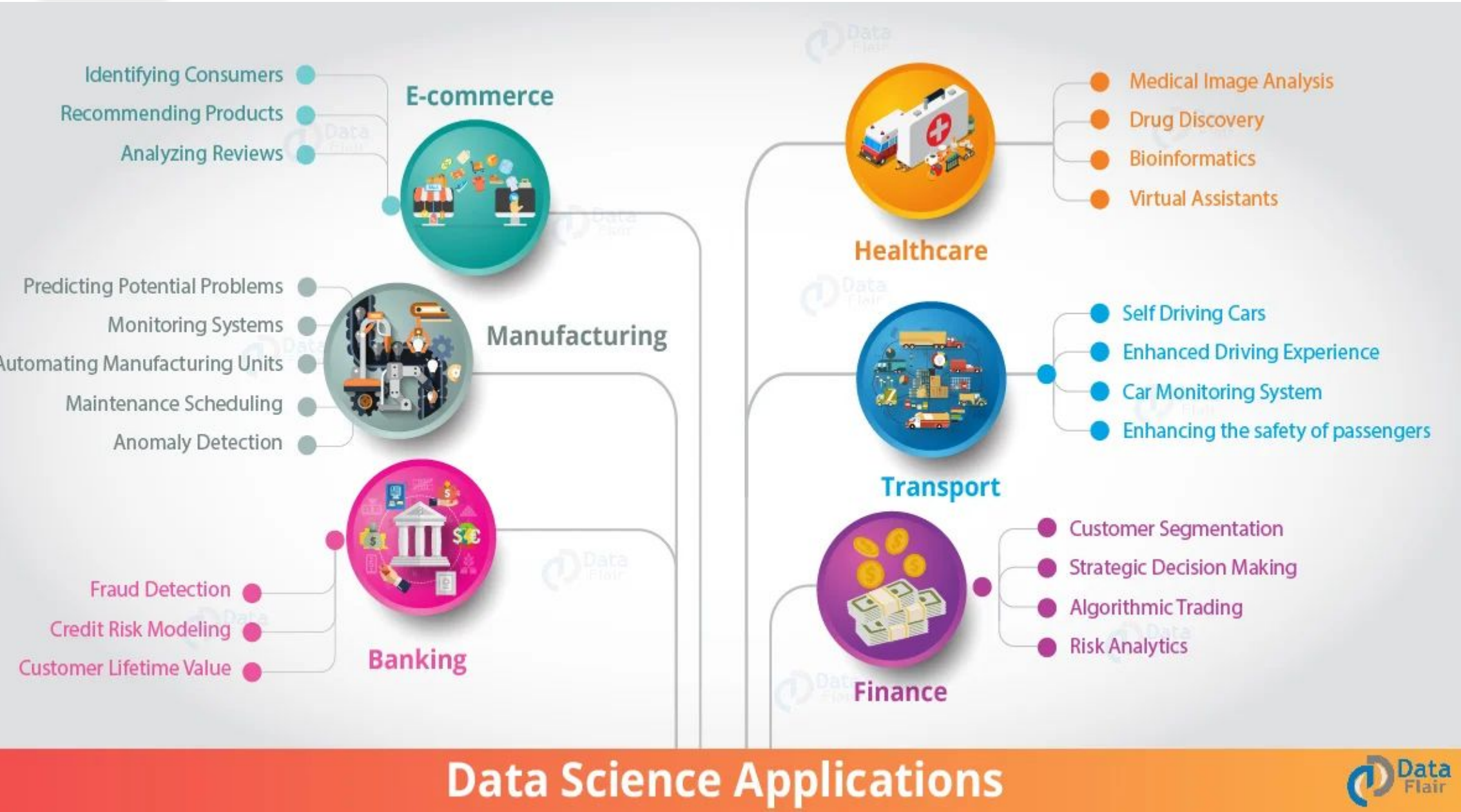
Basics of Data Science




Data Flair
Importance of Data Science in Business
Business Intelligence for Smarter Decisions
Making Better Product
Managing Business Efficiently
Predictive Analysis to Predict Outcomes
Leveraging Data for Business Decisions
Assessing Business Decisions
Automating Recruiting Process

Applications of Data Science

Unit 1



Data Science and Information Science

Information Science Definition:

- A broader interdisciplinary field that focuses on managing, storing, retrieving, and organizing information to improve its usability and accessibility for individuals and organizations.

Focus

Data Science:

- Analyzing and interpreting data to make predictions, discover trends, and support decision-making.
- Emphasizes quantitative and computational techniques.

Information Science:

- Organizing, classifying, and managing information resources to facilitate knowledge sharing and access.
- Focuses on the systems and tools that manage information (e.g., libraries, databases).

Tools and Techniques

Data Science:

- Programming languages: Python, R, SQL.
- Tools: Jupyter, TensorFlow, Pandas, scikit-learn.
- Methods: Machine learning, data mining, big data analytics, statistical modeling.

Information Science:

- Information systems: Database Management Systems (DBMS), Content Management Systems (CMS).
- Tools: Zotero, EndNote, and library cataloging systems.

Applications

Data Science:

- Fraud detection.
- Predictive maintenance in manufacturing.
- Recommendation systems (e.g., Netflix, Amazon).
- Natural language processing (e.g., chatbots, sentiment analysis).

Information Science:

- Library and knowledge management systems.
- Information architecture for websites and digital platforms.
- Developing search engines and indexing systems.

Skillset

Data Scientists:

- Mathematics, statistics, computer science, and domain expertise.
- Skills in data analysis, programming, and machine learning.

Information Scientists:

- Library science, information systems, and communication studies.
- Skills in information organization, metadata design, and user interaction.

Conclude

- **Data Science** uses data as a tool for discovery and decision-making.
- **Information Science** ensures information is well-organized, retrievable, and useful.

Business intelligence and data science

- BI involves collecting, analyzing, and visualizing **historical and current** data to help organizations make informed business decisions.
- It focuses on descriptive analytics—understanding **what happened and why**.

- **Business Intelligence (BI)** and **Data Science** are related fields that focus on data analysis and decision-making, but they differ in their approaches, goals, and methods

Focus

BI:

- Aggregating and reporting historical data.
- Providing dashboards and visualizations to track Key Performance Indicators (KPIs) i.e. measurable values.

Data Science:

- Predicting future outcomes based on data.
- Developing models to automate decision-making.

Objectives

BI:

- Monitor business performance.
- Provide executives with actionable summaries of business data.

Data Science:

- Build predictive models to forecast trends.
- Develop algorithms for recommendation systems.

Tools and Technologies

- **BI Tools:**
  - Power BI
  - Tableau
  - QlikView
  - SAP
  - Excel (for simpler BI tasks)
- **Data Science Tools:**
  - Python, R, SQL
  - Jupyter Notebook
  - TensorFlow, PyTorch (for AI/ML models)
  - Pandas, NumPy, scikit-learn
  - Hadoop, Spark (for big data processing)

Techniques

BI:

- Data Warehousing and ETL (Extract, Transform, Load).
- Reporting and dashboard creation.
- Descriptive analytics (e.g., year-over-year comparisons).

Data Science:

- Machine Learning and Deep Learning.
- Predictive and prescriptive analytics.
- Natural Language Processing, Computer Vision, and AI techniques.

Applications

BI Applications:

- Monitoring sales performance and customer trends.
- Budget and financial planning.
- Operational reporting and optimization.

Data Science Applications:

- Customer segmentation and personalization.
- Fraud detection and risk assessment.
- Forecasting demand or market trends.

Who Uses It?

BI Users:

- Business managers and executives.
- Operational teams for performance monitoring.

Data Science Users:

- Data scientists, analysts, and engineers.
- Researchers and developers building AI/ML models.

BI

- **Example:**
- A retail company uses BI tools like Power BI or Tableau to monitor sales performance across regions, products, and customer segments.
- Insights from the analysis help managers identify underperforming products or regions and implement targeted marketing campaigns.
- Increased sales and customer satisfaction.

Definition:

Data
raw facts, figures, symbols, or raw information that can be processed and converted into meaningful insights

Sources of Data: :

Data can be sourced from a variety of locations, depending on the context and the needs of the analysis or application.

Primary Data: Data collected directly by the researcher or organization for a specific purpose.

Secondary Data: Data collected by someone else but used for analysis or other purposes.

Primary Data Sources

- **Surveys**: Questionnaires or forms designed to collect specific information.
- **Interviews** Direct discussions with individuals or groups to gather insights.
- **Experiments**: Data generated through controlled scientific studies.
- **Observations**: Recording events, behaviors, or conditions in real-time.
- **Focus Groups**: Discussions with selected participants to explore their opinions or reactions.

Secondary Data Sources

- **Government Reports**: Census data, economic indicators, and public health statistics.
- **Industry Reports**: Data from trade associations or market research firms.
- **Academic Studies**: Published research papers or studies.
- **Open Data Portals**: Platforms offering free access to datasets (e.g., Data.gov, World Bank).

Machine-Generated Data

- Data generated automatically by systems, devices, or algorithms.
- **IoT Devices**: Sensors, smart devices, and wearables generating continuous data streams.
- **Web Logs**: Server logs tracking user activity on websites.
- **System Logs**: Data from software and hardware operations.
- **Automated Measurements**: Data from machines like GPS trackers, weather sensors, etc.

Transactional Data

- Data generated as a result of business transactions.
- **Online Transactions**: E-commerce purchases, payments, and interactions

Organizational Data

- Data generated or stored within organizations for internal use.
- **Customer Data**: CRM systems holding customer details and interactions.
- **Operational Data**: Inventory, logistics, and production data.
- **Employee Data**: HR records, payroll systems, and attendance logs.
- **Financial Data**: Budgets, income statements, and expense reports.

Social and Behavioral Data

- **Social Media Data**: Posts, likes, shares, and comments.
- **Mobile App Data**: User interactions with mobile applications.
- **Survey Responses**: Feedback from users, customers, or employees.

Data Types or attributes

- Nominal Attributes:
- Nominal means “relating to names.”
- The values of a nominal attribute are symbols or names of things.
- Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as categorical.
- The values do not have any meaningful order
- In computer science, called as enumeration
- Example, Student_name

Binary Attributes

- A binary attribute is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present.
- Binary attributes are referred to as Boolean if the two states correspond to true and false

- A binary attribute is symmetric if both of its states are equally valuable and carry the same weight; that is, there is no preference on which outcome should be coded as 0 or 1.
- One such example could be the attribute gender having the states male and female

- A binary attribute is asymmetric if the outcomes of the states are not equally important, such as the positive and negative

Ordinal Attributes

- An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them.
- Ordinal attributes. Suppose that cold drink size corresponds to the size of drinks available at a fast-food restaurant. This nominal attribute has three possible values: small, medium, and large

Numeric Attributes

- A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values

Discrete versus Continuous Attributes

- A discrete attribute has a finite or countably infinite set of values, which may or may not be represented as integers.

Example, **color_names**

- If an attribute is not discrete, it is continuous

Example, **House_price**

Need of Data Wrangling: Data Quality

- **Data Quality**: Why Preprocess the Data?
- Data have quality if they satisfy the requirements of the intended use.
- There are many factors comprising data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability

- incomplete (lacking attribute values or certain attributes of interest, or containing only aggregate data)
- inaccurate or noisy (containing errors, or values that deviate from the expected)
- inconsistent (e.g., containing discrepancies in the department codes used to categorize items).

Three of the elements defining data quality:

- accuracy
- completeness,
- and consistency.

Additional Qualities

- Believability reflects how much the data are trusted by users, while interpretability reflects how easy the data are understood.

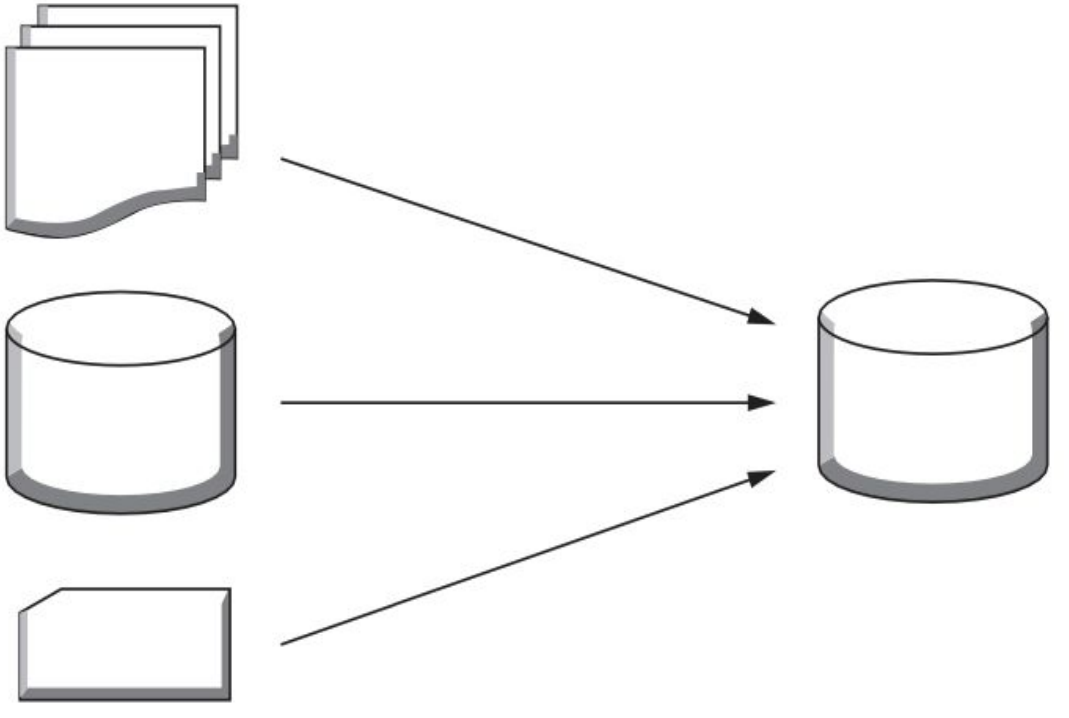
- **Welcome to the real world of Data!**

Data Wrangling Methods

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation
- Data Discretization

Data Cleaning

Data Integration



Data Transformation –17, 25, 39, 128, –39 → 0.17, 0.25, 0.39, 1.28, –0.39

Data Reduction

	A1	A2	A3	....	A200
T1					
T2					
T3					
....					
T200					

→

	A1	A2	A3	...	A120
T1					
T2					
T3					
....					
T150					

Data Cleaning

- Real-world data tend to be incomplete, noisy, and inconsistent
- Data cleaning routines work to “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.

Missing Values

- **Ignore the tuple**: This is usually done when the class label is missing .
- **Fill in the missing value manually- time consuming**
- Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value:
- Use the most probable value to fill in the missing value (Mode)

Noisy Data

- "What is noise?"

  Noise is a random error or variance in a measured variable

Given a numeric attribute such as, say, price, how can we "smooth" out the data to remove the noise? Let's look at the following data smoothing techniques.

Binning

- The sorted values are distributed into a number of “buckets,” or bins.

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 boundaries:

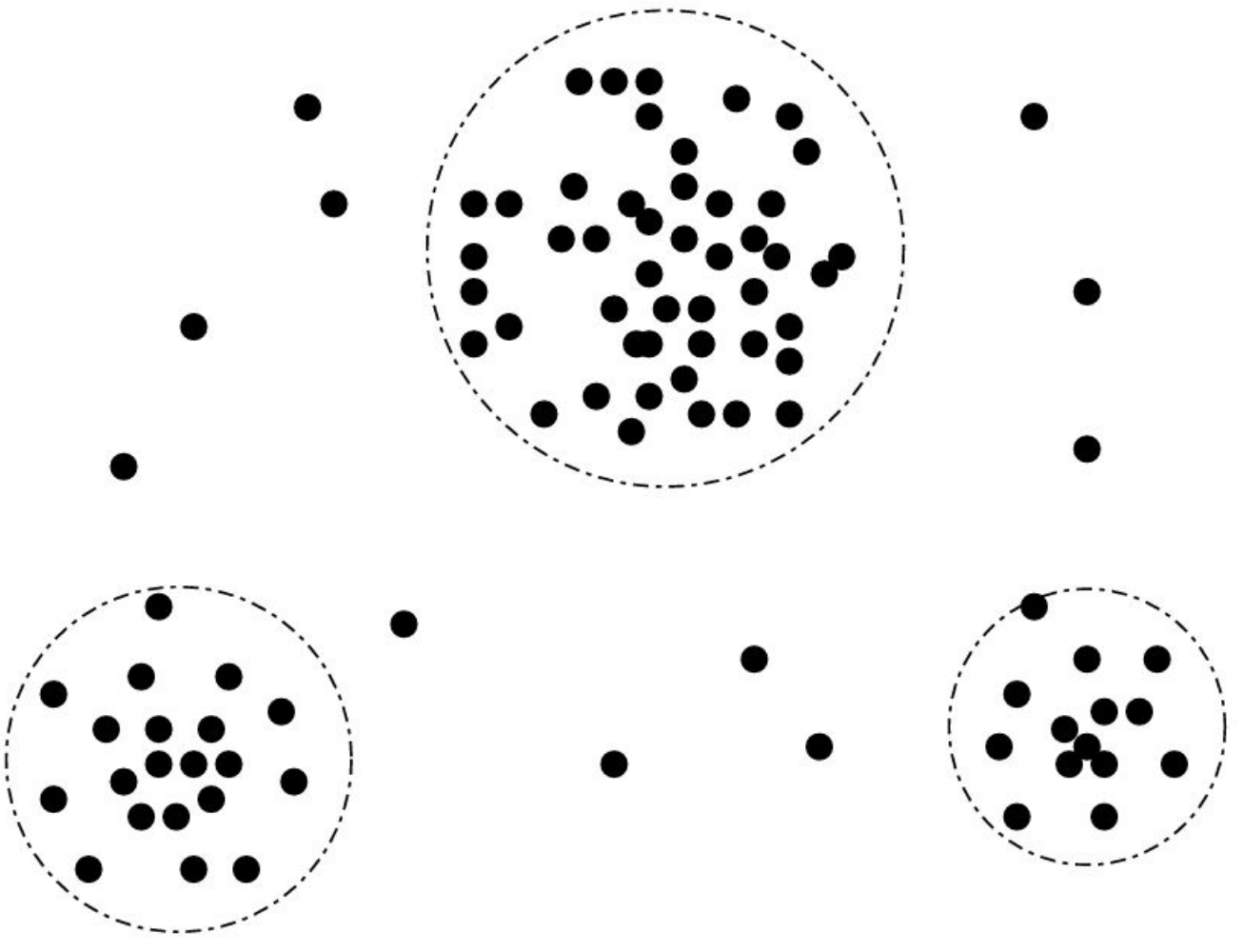
Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Regression

- Data smoothing can also be done by regression.
- Linear regression involves finding the “best” line to fit two attributes (or variables) so that one attribute can be used to predict the other.
- Multiple linear regression is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface

Outlier analysis

- Outliers may be detected by clustering, for example, where similar values are organized into groups, or “clusters.”
- Intuitively, values that fall outside of the set of clusters may be considered outliers



A 2-D customer data plot with respect to customer locations in a city, showing three data clusters. Outliers may be detected as values that fall outside of the cluster sets.

Pedagogy

- Data Driven Storytelling

Context and Background

- Before diving into data, it's crucial to provide context. Why are we looking at this data? What problem are we trying to solve, or what question are we trying to answer? Providing context helps the audience understand the relevance of the data being presented.

Example: In a dataset about climate change, it would be useful to explain the significance of global temperature trends, the impacts of rising temperatures, and the specific focus of the analysis (e.g., emissions by country, temperature anomalies over decades, etc.).

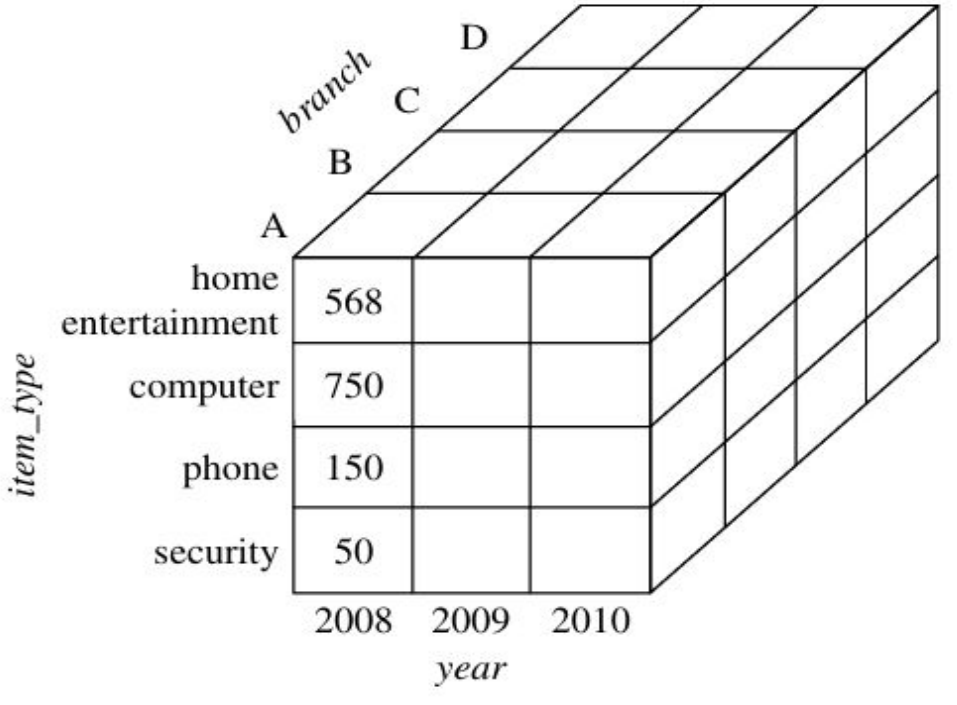
Data Integration

- Data mining often requires data integration—the merging of data from multiple data stores.
- 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 goal of data integration is to make the data more useful and meaningful for the purposes of analysis and decision making.
- Techniques used in data integration include data warehousing, ETL (extract, transform, load) processes

- Data Integration is a data preprocessing technique that combines data from multiple heterogeneous data sources into a coherent data store and provides a unified view of the data.
- These sources may include multiple data cubes, databases, or flat files.

Data cube example



A data cube for sales at *AllElectronics*.

Entity Identification Problem

- There are a number of issues to consider during data integration. Schema integration and object matching can be tricky

- How can equivalent real-world entities from multiple data sources be matched up?
- This is referred to as the entity identification problem.
- For example, how can the data analyst or the computer be sure that customer id in one database and cust number in another refer to the same attribute

Redundancy and Correlation Analysis

- Some redundancies can be detected by correlation analysis.

Data Integration

we've made sure to remove the impurities in data and make it clean. Now, the next step is to combine data from different sources to get a unified structure with more meaningful and valuable information

Say we have some data about an employee in a database. We can't expect all the data about the employee to reside in the same table.

- It's possible that the employee's personal data will be located in one table
- the employee's project history will be in a second table
- the employee's time-in and time-out details will be in another table, and so on

So, if we want to do some analysis about the employee, we need to get all the employee data in one common place. This process of bringing data together in one place is called data integration

Data_Integration.ipynb

File Edit View Insert Runtime Tools Help Last saved at January 27

Share

+ Code + Text

Connect Gemini

```
[ ] import pandas as pd
```

```
data=pd.read_excel("student.xlsx")
data
```

	student_id	marks	city
0	1	45	pune
1	2	25	satara
2	3	35	delhi
3	4	48	chennai
4	5	47	solapur
5	6	56	kolhapur
6	7	58	pune
7	8	54	mumbai
8	9	59	nagpur
9	10	60	nanded

Data_Integration.ipynb

File Edit View Insert Runtime Tools Help Last saved at January 27

Share

+ Code + Text

Connect Gemini

```
data1=pd.read_excel("stud_info.xlsx")
data1
```

	student_id	age	gender	grade	employed
0	1	14	M	First_class	Yes
1	2	15	F	Second_class	Yes
2	3	16	M	First_class	No
3	4	17	F	Second_class	Yes
4	5	18	M	First_class	Yes
5	6	19	F	Second_class	No
6	7	20	M	First_class	Yes
7	8	21	F	Second_class	Yes
8	9	22	M	First_class	No
9	10	23	F	Second_class	Yes

Merging of data

```
[ ] data2=pd.merge(data,data1,on='student_id')
    data2
```

	student_id	marks	city	age	gender	grade	employed
0	1	45	pune	14	M	First_class	Yes
1	2	25	satara	15	F	Second_class	Yes
2	3	35	delhi	16	M	First_class	No
3	4	48	chennai	17	F	Second_class	Yes
4	5	47	solapur	18	M	First_class	Yes
5	6	56	kolhapur	19	F	Second_class	No
6	7	58	pune	20	M	First_class	Yes
7	8	54	mumbai	21	F	Second_class	Yes
8	9	59	nagpur	22	M	First_class	No
9	10	60	nanded	23	F	Second_class	Yes

1. What is the primary goal of Data Science?

- a) Develop software applications
  b) Extract insights and knowledge from data
  c) Build websites
  d) Create multimedia content

- **Which of the following is NOT a step in the Data Science process?**
- a) Data collection
  b) Data cleaning
  c) Data visualization
  d) System reboot

- **Which programming language is most commonly used in Data Science?**
- a) Java
  b) Python
  c) HTML
  d) C++

What does the term "Big Data" refer to?

- a) Data stored in large physical drives
  b) Vast volumes of data that cannot be processed using traditional tools
  c) Data from big companies
  d) Data with long file names

Which of these is an application of Data Science?

- a) Fraud detection in banking
  b) Personalized recommendations in e-commerce
  c) Autonomous driving
  d) All of the above

What is the main purpose of data visualization?

- a) To store data in graphical formats
  b) To make data analysis visually accessible and interpretable
  c) To replace statistical analysis
  d) To design computer games

Which library is widely used for data manipulation in Python?

- a) TensorFlow
  b) NumPy
  c) Pandas
  d) OpenCV

Which library is most commonly used for data visualization in Python?

a) NumPy
b) Matplotlib
c) TensorFlow
d) OpenCV

What is Information Science primarily concerned with?

a) Data storage and retrieval
b) Human interaction with information
c) Organizing and managing information resources
d) All of the above

- **Which of the following is a common application of Information Science?**
  a) Recommendation Systems
  b) Library Catalog Systems
  c) Natural Language Processing
  d) All of the above
- **Which technology is commonly used in Information Retrieval systems?**
  a) Relational Databases
  b) Search Engines
  c) Machine Learning
  d) All of the above

Data Transformation

- In this preprocessing step, 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.

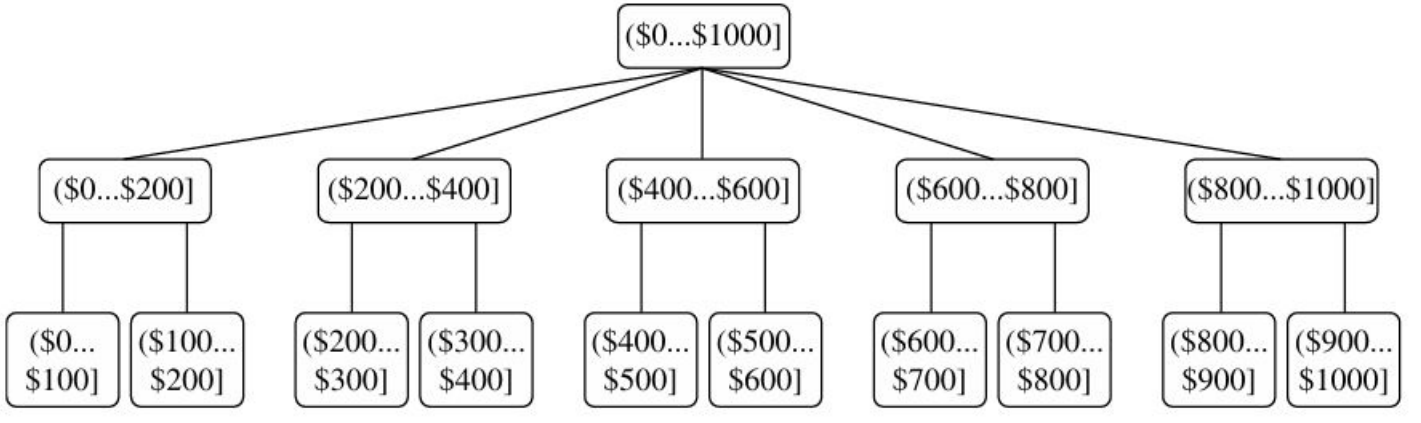
Data Transformation Strategies

- Smoothing: which works to remove noise from the data. Techniques include binning, regression, and clustering.
- Attribute construction (or feature construction), where new attributes are con structed and added from the given set of attributes to help the mining process

- Aggregation, where summary or aggregation operations are applied to the data.
- For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts.

- Normalization, where the attribute data are scaled so as to fall within a smaller range, such as 0.0 to 1.0.
- Discretization, 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)

Concept Hierarchy



A concept hierarchy for the attribute *price*, where an interval ($X . . . $Y] denotes the range from $X (exclusive) to $Y (inclusive).

Data Transformation by Normalization

- Changing measurement units from meters to inches for height,or from kilograms to pounds for weight, may lead to very different results.

- 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 [-1,1] or [0.0, 1.0].
- (The terms standardize and normalize are used interchangeably in data preprocessing

min-max normalization

Min-max normalization performs a linear transformation on the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, v_i, of A to v_i' in the range $[new\_min_A, new\_max_A]$ by computing

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A. \qquad (3.8)$$

Min-max normalization preserves the relationships among the original data values. It will encounter an "out-of-bounds" error if a future input case for normalization falls outside of the original data range for A.

Example

- Suppose that the minimum and maximum values for the attribute income are $12,000 and $98,000, respectively.
- We would like to map income to the range [0.0,1.0]. By min-max normalization, a value of $73,600 for income is transformed to ----

z-score normalization

In ***z*-score normalization** (or *zero-mean normalization*), the values for an attribute, A, are normalized based on the mean (i.e., average) and standard deviation of A. A value, v_i, of A is normalized to v_i' by computing

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}, \tag{3.9}$$

where $\bar{A}$ and σ_A are the mean and standard deviation, respectively, of attribute A. The

Example

- Suppose that the mean and standard deviation of the values for the attribute income are $54,000 and $16,000, respectively. With z-score normalization, a value of $73,600 for income is transformed to----

Normalization by decimal scaling

Normalization by decimal scaling normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value, v_i, of A is normalized to v_i' by computing

$$v_i' = \frac{v_i}{10^j}, \tag{3.12}$$

Example

- Suppose that the recorded values of A range from -986 to 917. The maximum absolute value of A is 986.
- Normalize -986 and 917 ?

Discretization by Histogram Analysis

- histogram analysis is an unsupervised discretization technique because it does not use class information.

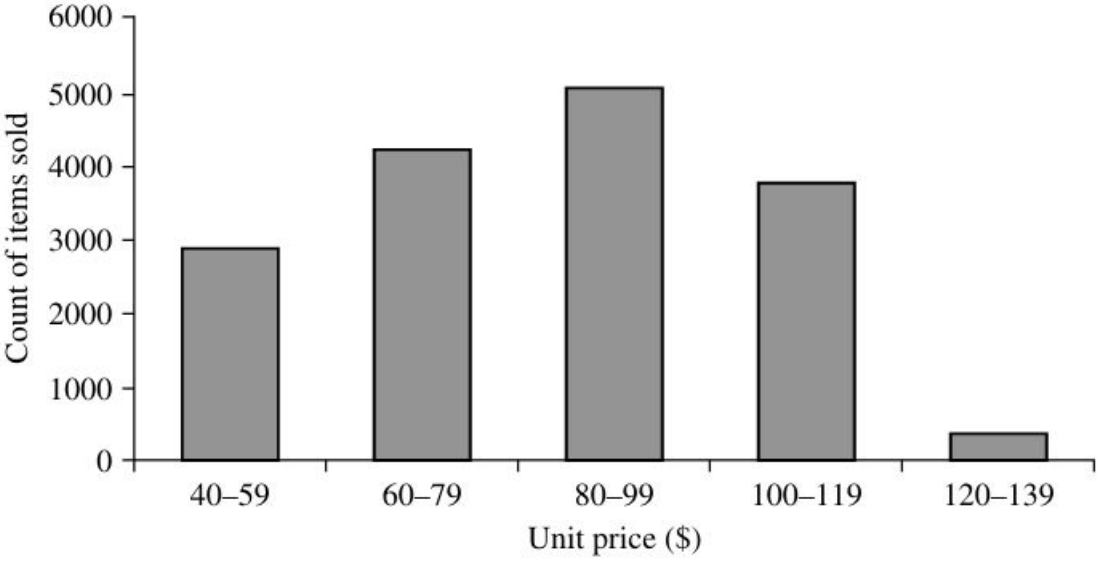


Figure 2.6 A histogram for the Table 2.1 data set.

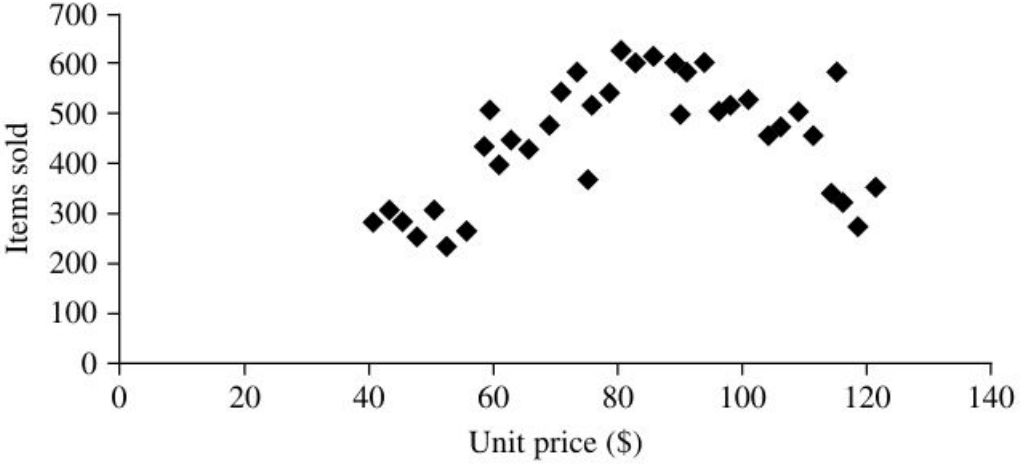


Figure 2.7 A scatter plot for the Table 2.1 data set.

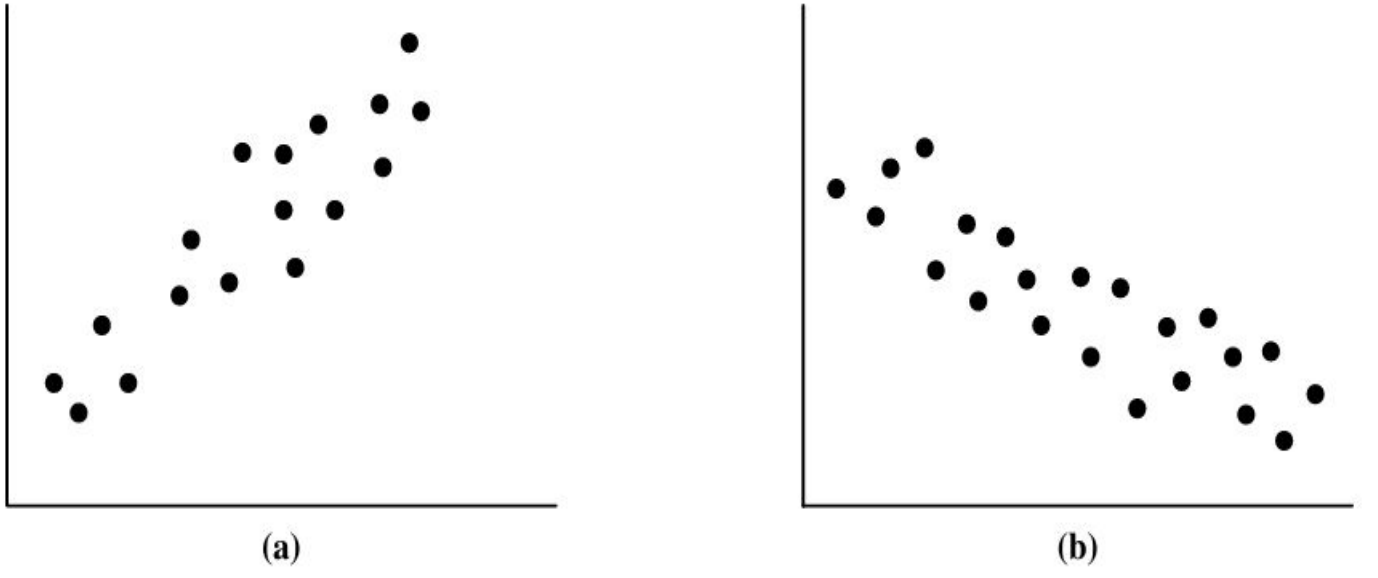


Figure 2.8 Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.