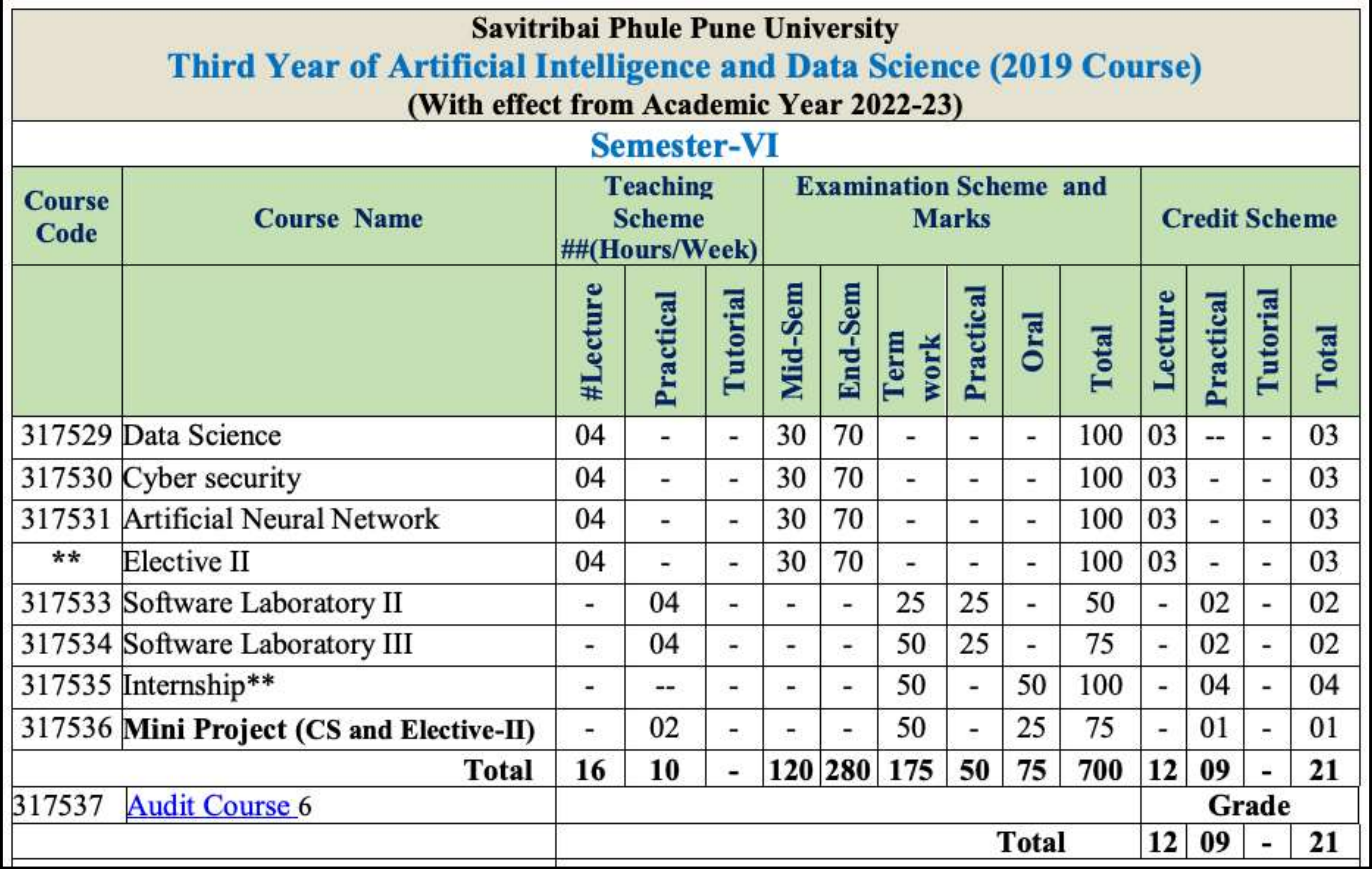
DATA SCIENCE

UNIT - I



Savitribai Phule Pune University

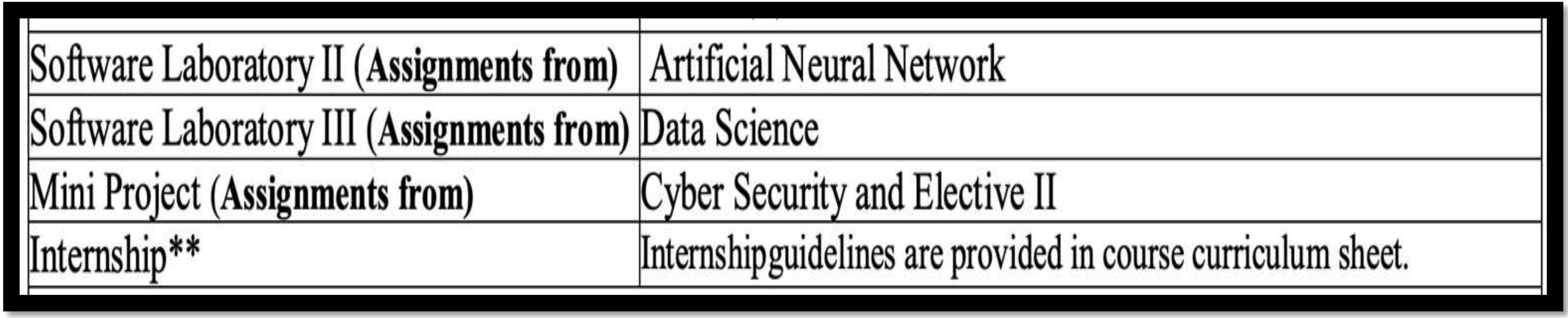
Third Year of Artificial Intelligence and Data Science (2019 Course)

(With effect from Academic Year 2022-23)

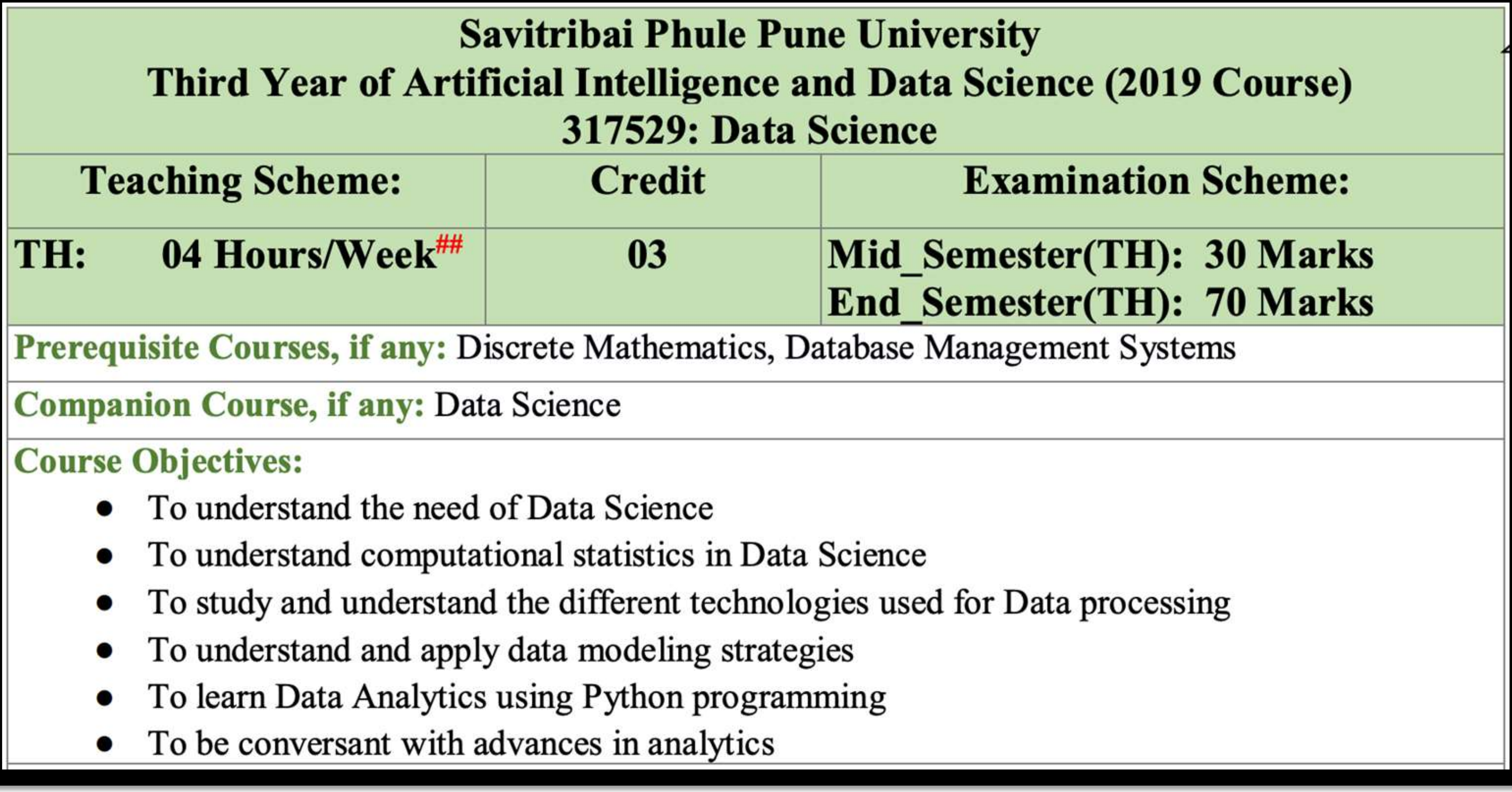
Semester-VI

Course Code	Course Name	Teaching Scheme ##(Hours/Week)			Examination Scheme and Marks						Credit Scheme			
		#Lecture	Practical	Tutorial	Mid-Sem	End-Sem	Term work	Practical	Oral	Total	Lecture	Practical	Tutorial	Total
317529	Data Science	04	-	-	30	70	-	-	-	100	03	--	-	03
317530	Cyber security	04	-	-	30	70	-	-	-	100	03	-	-	03
317531	Artificial Neural Network	04	-	-	30	70	-	-	-	100	03	-	-	03
**	Elective II	04	-	-	30	70	-	-	-	100	03	-	-	03
317533	Software Laboratory II	-	04	-	-	-	25	25	-	50	-	02	-	02
317534	Software Laboratory III	-	04	-	-	-	50	25	-	75	-	02	-	02
317535	Internship**	-	--	-	-	-	50	-	50	100	-	04	-	04
317536	**Mini Project (CS and Elective-II)**	-	02	-	-	-	50	-	25	75	-	01	-	01
	Total	**16**	**10**	**-**	**120**	**280**	**175**	**50**	**75**	**700**	**12**	**09**	**-**	**21**
317537	Audit Course 6										**Grade**			
									Total		**12**	**09**	**-**	**21**

PRACTICAL



Software Laboratory II (**Assignments from**)	Artificial Neural Network
Software Laboratory III (**Assignments from**)	Data Science
Mini Project (**Assignments from**)	Cyber Security and Elective II
Internship**	Internshipguidelines are provided in course curriculum sheet.



Savitribai Phule Pune University
Third Year of Artificial Intelligence and Data Science (2019 Course)
317529: Data Science

Teaching Scheme:	Credit	Examination Scheme:
TH: 04 Hours/Week##	03	Mid_Semester(TH): 30 Marks End_Semester(TH): 70 Marks

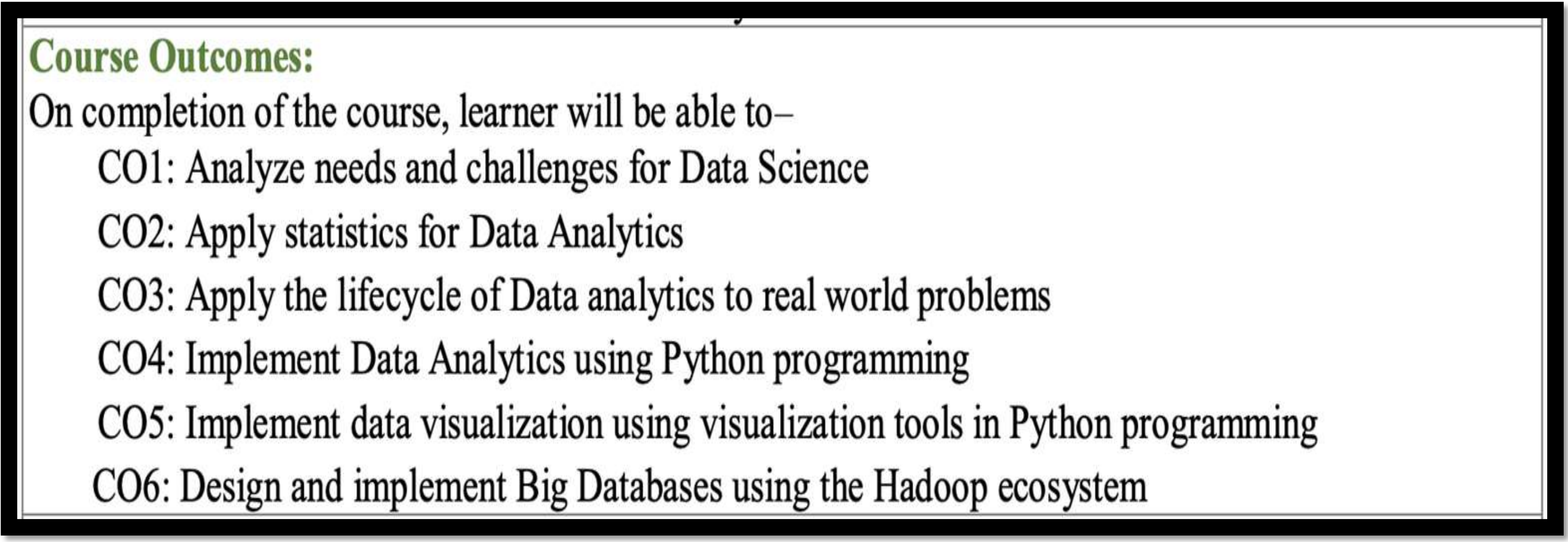
Prerequisite Courses, if any: Discrete Mathematics, Database Management Systems

Companion Course, if any: Data Science

Course Objectives:

- To understand the need of Data Science
- To understand computational statistics in Data Science
- To study and understand the different technologies used for Data processing
- To understand and apply data modeling strategies
- To learn Data Analytics using Python programming
- To be conversant with advances in analytics

COURSE OUTCOMES



Course Outcomes:

On completion of the course, learner will be able to–

CO1: Analyze needs and challenges for Data Science

CO2: Apply statistics for Data Analytics

CO3: Apply the lifecycle of Data analytics to real world problems

CO4: Implement Data Analytics using Python programming

CO5: Implement data visualization using visualization tools in Python programming

CO6: Design and implement Big Databases using the Hadoop ecosystem

Types of Data Science Job

If you learn data science, then you get the opportunity to find the various exciting job roles in this domain. The main job roles are given below:

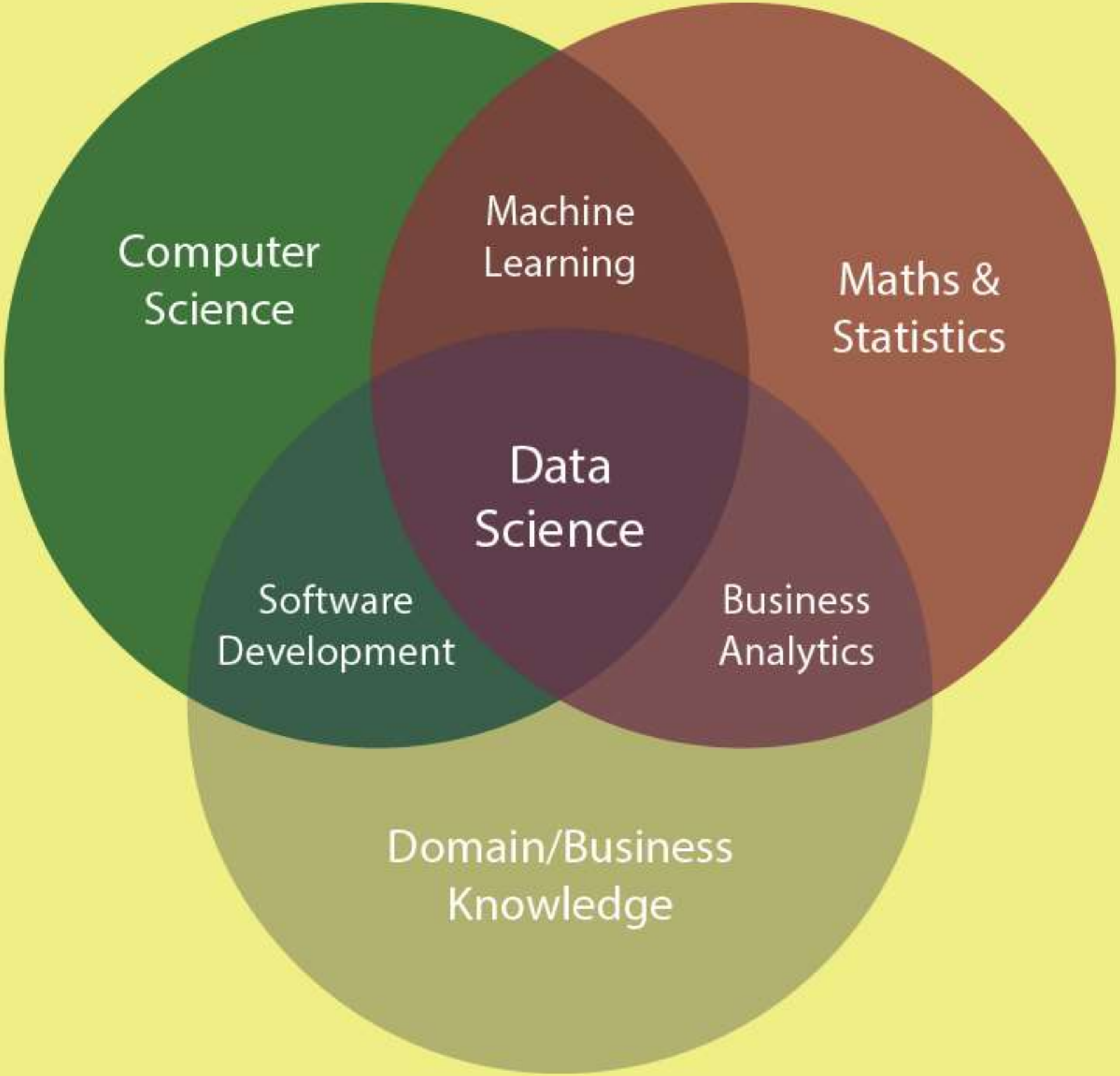
1. Data Scientist
2. Data Analyst
3. Data engineer
4. Data Architect
5. Data Administrator
6. Business Analyst
7. Business Intelligence Manager

UNIT 1: INTRODUCTION TO DATA SCIENCE

- 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, and Data Discretization.

DATA SCIENCE

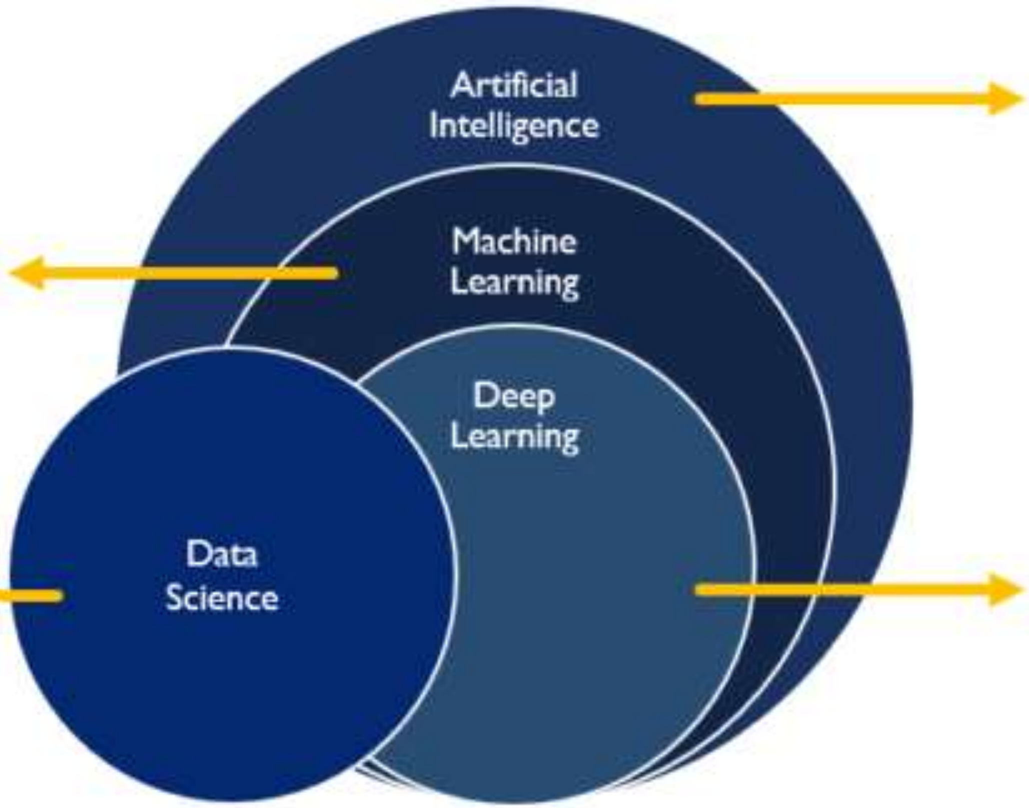
- Data Science is the area of study which involves extracting insights from vast amounts of data using various scientific methods, algorithms, and processes.
- It helps you to discover hidden patterns from the raw data.
- The term Data Science has emerged because of the evolution of mathematical statistics, data analysis, and big data.
- Data Science is a field that allows you to extract knowledge from structured or unstructured data.


Computer Science
Machine Learning
Maths & Statistics
Data Science
Software Development
Business Analytics
Domain/Business Knowledge

DATA SCIENCE

Machine Learning is a subset of AI, and builds a model based on training data to make predictions

Data Science is a subset of AI. It is an area of statistics, scientific methods, etc. to extract meaning and insights from data..



Artificial Intelligence means creating smart machines to mimic human behavior

Deep learning is a subset of ML, a class of ML algorithms to solve complex problems.

Need of Data Science

1. Informed Decision Making:
   - □ Enables organizations to make data-driven decisions for better outcomes.

1. Big Data Handling:
   - □ Extracts insights from vast and complex datasets, which traditional methods struggle to manage.

1. Business Innovation:
   - □ Drives innovation, leading to new opportunities and maintaining a competitive advantage.

Need of Data Science

4. Operational Efficiency:
   - □ Enhances operational efficiency by optimizing processes, reducing costs, and improving productivity.

4. Personalization and Customer Experience:
   - □ Utilizes data for personalized products, services, and marketing, improving customer experience.

4. Risk Management and Fraud Detection:
   - □ Plays a vital role in risk assessment, fraud detection, and credit scoring.

Need of Data Science

7. Healthcare Advancements:
   - Contributes to advancements in healthcare through data analysis, personalized medicine, and disease prediction.
8. Scientific Research:
   - Accelerates scientific discoveries by analyzing large datasets in various research fields.
9. Supply Chain Optimization:
   - Optimizes supply chain processes, leading to inventory efficiency and cost savings.
10. Smart Cities and Urban Planning:
    - Facilitates the development of smart cities through data analysis for better urban planning and resource management.

APPLICATIONS

1. Healthcare:

   □ Predictive Analytics: Predicting patient outcomes and identifying potential health risks.

   □ Disease Surveillance: Monitoring and predicting the spread of diseases.

   □ Personalized Medicine: Tailoring treatment plans based on individual patient data.

2. Finance:

   □ Fraud Detection: Identifying unusual patterns to detect and prevent fraudulent activities.

   □ Credit Scoring: Assessing creditworthiness of individuals and businesses.

   □ Algorithmic Trading: Using algorithms to make investment decisions.

APPLICATIONS

3. Retail:
   - Recommendation Systems: Providing personalized product recommendations to customers.
   - Inventory Management: Optimizing stock levels based on demand forecasts.
   - Customer Segmentation: Dividing customers into groups for targeted marketing.
4. Manufacturing:
   - Predictive Maintenance: Anticipating equipment failures to reduce downtime.
   - Quality Control: Analyzing production data to ensure product quality.
   - Supply Chain Optimization: Streamlining logistics and inventory management.

APPLICATIONS

5. Telecommunications:
   - Churn Prediction: Identifying customers at risk of leaving the service.
   - Network Optimization: Enhancing network performance and efficiency.
   - Customer Experience Analysis: Improving services based on customer feedback and behavior.
6. Marketing:
   - Customer Segmentation: Dividing customers into groups for targeted campaigns.
   - Social Media Analytics: Analyzing social media data to understand customer sentiment.

APPLICATIONS

7. Education:

□ Adaptive Learning Platforms: Personalizing educational content based on student performance.

□ Student Retention: Identifying factors influencing student dropout rates.

□ Performance Analysis: Analyzing data to improve teaching methodologies.

8. Energy:

□ Predictive Maintenance for Infrastructure: Monitoring and maintaining energy infrastructure.

□ Smart Grids: Optimizing energy distribution for efficiency.

□ Energy Consumption Forecasting: Predicting future energy demand.

APPLICATIONS

9. Transportation:
   - Route Optimization: Finding the most efficient routes for vehicles.
   - Demand Forecasting: Predicting transportation demand for better planning.
   - Traffic Management: Analyzing traffic patterns for improved city planning.
10. Government:
    - Crime Prediction: Predicting areas with high likelihood of criminal activities.
    - Public Health Monitoring: Tracking and managing public health crises.
    - Policy Planning: Analyzing data for evidence-based policy decision-making.

Relationship between Data Science and Information Science

- Data Science and Information Science are closely related fields but focus on different aspects of handling and utilizing information.

- Data Science:
  - Primarily deals with extracting insights and knowledge from structured and unstructured data. It involves a combination of statistics, machine learning, and domain expertise to analyze and interpret data.

- Information Science:
  - Focuses on the organization, classification, retrieval, and dissemination of information. It encompasses a broader view, including the study of information systems, knowledge management, and the design of information structures.

Scope

- Data Science: Primarily focuses on the extraction of knowledge and insights from data to support decision-making and predictions.

- Information Science: Encompasses a broader spectrum, including the study of information processes, systems, and the effective use of information in various domains.

Methods and Techniques

- Data Science: Utilizes statistical analysis, machine learning algorithms, data modeling, and programming to extract patterns and insights from data.

- Information Science: Emphasizes the organization and retrieval of information, often involving the design and management of databases, information systems, and knowledge repositories.

Data vs Information

- Data is something raw, meaningless, an object that, when analyzed or converted to a useful form, becomes information.

- Information is also defined as “data that are endowed with meaning and purpose.

- For example, the number “480,000” is a data point. But when we add an explanation that it represents the number of deaths per year in the USA from cigarette smoking, it becomes information.

Applications: Information Science

1. Library Science: Organization, cataloging, and classification.
2. Information Retrieval Systems: Search engine development, algorithm design.
3. Database Management: Creation, maintenance, and optimization.
4. Knowledge Management: Capture, organization, and distribution.
5. Information Architecture: User-friendly structure design.
6. Digital Asset Management: Organization of digital media assets.
7. Document Management: Tracking, access control, versioning.
8. Health Information Management: Patient record organization.
9. Records Management: Lifecycle management and compliance.
10. Digital Preservation: Strategies for preserving digital content.

Overlap

- There is an overlap between Data Science and Information Science, especially in areas such as information retrieval, data management, and the development of information systems.

Business Intelligence vs Data Science

- Business intelligence (BI) is a set of strategies and technologies enterprises use to analyze business information and transform it into actionable insights that inform strategic and tactical business decisions.

- BI tools access and analyze data sets and present analytical findings in reports, summaries, dashboards, graphs, charts, and maps to provide users with detailed intelligence about the state of the business.

Factors	Business Intelligence	Data Science
Concept	It is a collection of processes, tools, and technologies that help a business with data analysis.	It consists of mathematical and statistical models used for processing the data, discovering hidden patterns, and predicting future actions based on those patterns.
Data	It deals mainly with structured data.	It accepts both structured and unstructured data.
Flexibility	Data sources should be planned before the visualization.	Data Sources can be added anytime based on the requirements.
Approach	It has both statistical and visual approaches toward data analysis.	Graph analysis, NLP, machine learning, neural networks, and other methods can be used to process the data.
Expertise	It is made for business users to visualize raw business information without any technical knowledge.	It requires sound knowledge of data analysis and programming.
Complexity	For a single user, compared to data science, business intelligence is much simpler to use and visualize data.	Data science is much more complex when compared to business intelligence.
Tools	Business intelligence tools include MS Excel, Power BI, SAS BI, MicroStrategy, IBM Cognos, Throughput, and more.	Some of the most popular Data Science tools are Python, Hadoop, Spark, R, TensorFlow, and more.

Data

1. Data Types
   a. Structured data
   b. Unstructured data

1. Data Collection
   a. Open Data
   b. Social Media Data
   c. Multimodal Data
   d. Data Storage and Presentation

Data Types: Structured data

1. Structured data is the most important data type.

1. Highly organized information that can be seamlessly included in a database and readily searched via simple search operations.

1. Someone would have to collect, store, and present the data in such a format.

Example

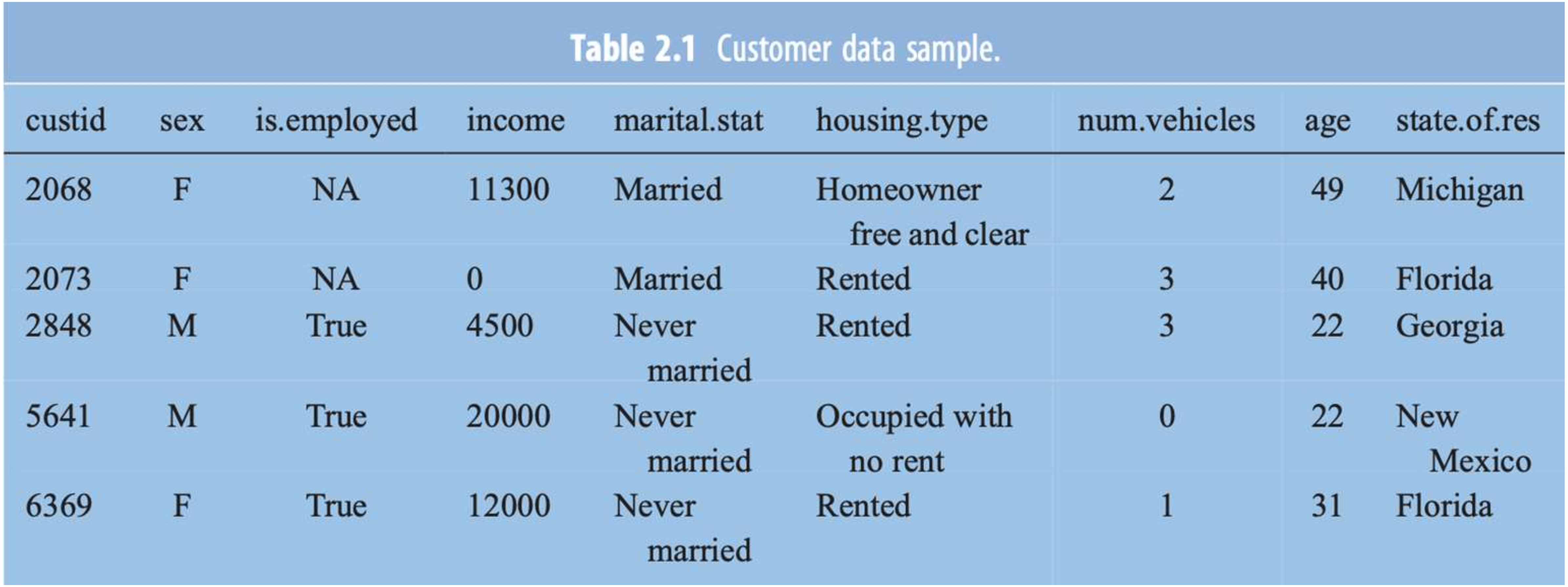


Table 2.1 Customer data sample.

custid	sex	is.employed	income	marital.stat	housing.type	num.vehicles	age	state.of.res
2068	F	NA	11300	Married	Homeowner free and clear	2	49	Michigan
2073	F	NA	0	Married	Rented	3	40	Florida
2848	M	True	4500	Never married	Rented	3	22	Georgia
5641	M	True	20000	Never married	Occupied with no rent	0	22	New Mexico
6369	F	True	12000	Never married	Rented	1	31	Florida

Data Types: Unstructured Data

- Unstructured data is data that does not have a pre-defined data model or format, making it less organized and more challenging to analyze using traditional methods.
- Examples of unstructured data include text documents, emails, social media posts, images, videos, audio recordings, etc.
- Challenges with Unstructured Data:
  - The lack of structure makes compilation and organizing unstructured data a time and energy-consuming task.
  - Structured data is akin to machine language, in that it makes information much easier to be parsed by computers.

Data Collection: Open Data

a. Open Data

- □ Data should be freely available in a public domain
- □ Can be used by anyone as they wish, without restrictions from copyright, patents, or other mechanisms of control.
- □ List of principles associated with open data
  1. Public
  2. Accessible
  3. Described
  4. Reusable
  5. Complete
  6. Timely

Data Collection: Social Media Data & Multimodal Data

b. Social Media Data
   - Social media has become a gold mine for collecting data to analyze for research or marketing purposes.
   - This is facilitated by the Application Programming Interface (API) that social media companies provide to researchers and developers.

b. Multimodal Data
   - Multimodal data refers to data that involves multiple modes or types of information.
   - Each mode represents a distinct form of data, such as text, images, audio, video, or other types, and these modes are often interconnected.

Data Collection: Multimodal Data

1. **Text**: Involves written or spoken language, such as documents, articles, transcripts, and textual information.
2. **Image**: Represents visual content, including photographs, graphics, and other visual elements.
3. **Audio**: Involves sound or spoken words, captured in audio files, recordings, or other formats.
4. **Video**: Integrates moving images and audio, often depicting dynamic scenes or events.
5. **Sensor Data**: Captures information from various sensors, such as temperature sensors, accelerometers, or environmental sensors.
6. **Geospatial Data**: Involves location-based information, including maps, GPS coordinates, and spatial data.

Data Storage and Presentation

- Depending on its nature, data is stored in various formats.
- If data is structured, it is common to store and present it in some kind of delimited way.
- That means various fields and values of the data are separated using delimiters, such as commas or tabs.
- Data Formats:
  1. CSV (Comma-Separated Values)
  2. TSV (Tab-Separated Values)
  3. XML (eXtensibleMarkupLanguage)
  4. RSS (Really Simple Syndication)
  5. JSON (JavaScript Object Notation)

CSV (Comma-Separated Values)

- CSV (Comma-Separated Values) format is the most common import and export format for spreadsheets and databases.
- For example, Depression.csv is a dataset that is available at UF Health, UF Biostatistics for downloading.
- The dataset represents the effectiveness of different treatment procedures on separate individuals with clinical depression.
- Any spreadsheet program such as Microsoft Excel or Google Sheets can readily open a CSV file and display it correctly most of the time.

CSV File Format

```
treat,before,after,diff
No Treatment,13,16,3
No Treatment,10,18,8
No Treatment,16,16,0
Placebo,16,13,-3
Placebo,14,12,-2
Placebo,19,12,-7
Seroxat (Paxil),17,15,-2
Seroxat (Paxil),14,19,5
Seroxat (Paxil),20,14,-6
Effexor,17,19,2
Effexor,20,12,-8
Effexor,13,10,-3
```

TSV (Tab-Separated Values)

- TSV (Tab-Separated Values) files are used for raw data and can be imported into and exported from spreadsheet software.
- Tab-separated values files are essentially text files, and the raw data can be viewed by text editors, though such files are often used when moving raw data between spreadsheets.

Name<TAB>Age<TAB>Address

Ryan<TAB>33<TAB>1115 W

Franklin Paul<TAB>25<TAB>Big Farm

Way Jim<TAB>45<TAB>W Main St

Samantha<TAB>32<TAB>28 George St

XML (eXtensible Markup Language)

- □ XML (eXtensible Markup Language) was designed to be both human and machine readable, and can thus be used to store and transport data.
- □ In the real world, computer systems and databases contain data in incompatible formats.
- □ As the XML data is stored in plain text format, it provides a software and hardware independent way of storing data.
- □ This makes it much easier to create data that can be shared by different applications.

XML File Format

```
<?xml version=“1.0” encoding=“UTF-8”?> <bookstore>
<book category=“information science” cover=“hardcover”> <title lang=“en”>Social
Information Seeking</title> <author>Chirag Shah</author>
<year>2017</year>
          <price>62.58</price>
      </book>
<book category=“data science” cover=“paperback”> <title lang=“en”>Hands-On
Introduction to Data
Science</title> <author>Chirag Shah</author> <year>2019</year> <price>50.00</price>
      </book>
  </bookstore>
```

RSS & JSON

- **RSS (Really Simple Syndication)**
  - It is a format used to share data between services, and which was defined in the 1.0 version of XML.

- **JSON (JavaScript Object Notation)**
  - It is a lightweight data-interchange format.
  - It is not only easy for humans to read and write, but also easy for machines to parse and generate. It is based on a subset of the JavaScript Programming Language.

Data Pre-processing

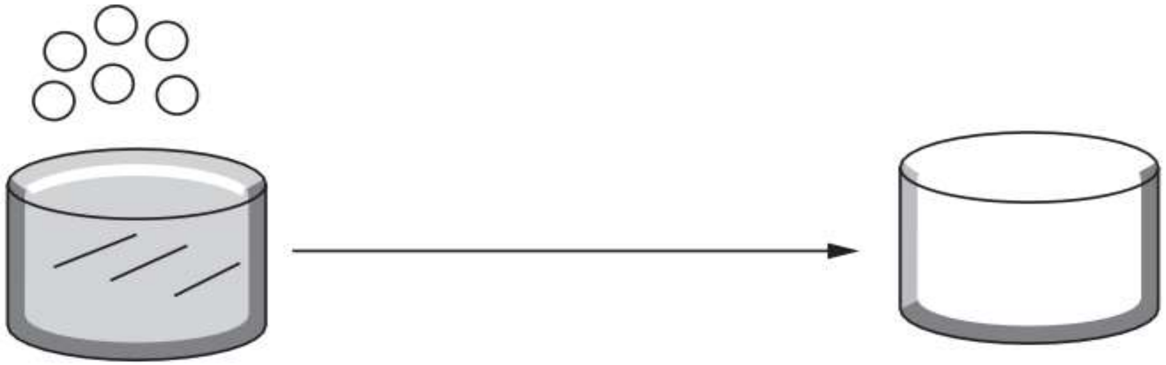
- Data in the real world is often dirty; that is, it is in need of being cleaned up before it can be used for a desired purpose. This is often called data pre-processing.
- Here are some of the factors that indicate that data is not clean or ready to process:
  - Incomplete:
    - When some of the attribute values are lacking, certain attributes of interest are lacking, or attributes contain only aggregate data.
  - Noisy:
    - When data contains errors or outliers. For example, some of the data points in a dataset may contain extreme values that can severely affect the dataset's range.

Data Pre-processing

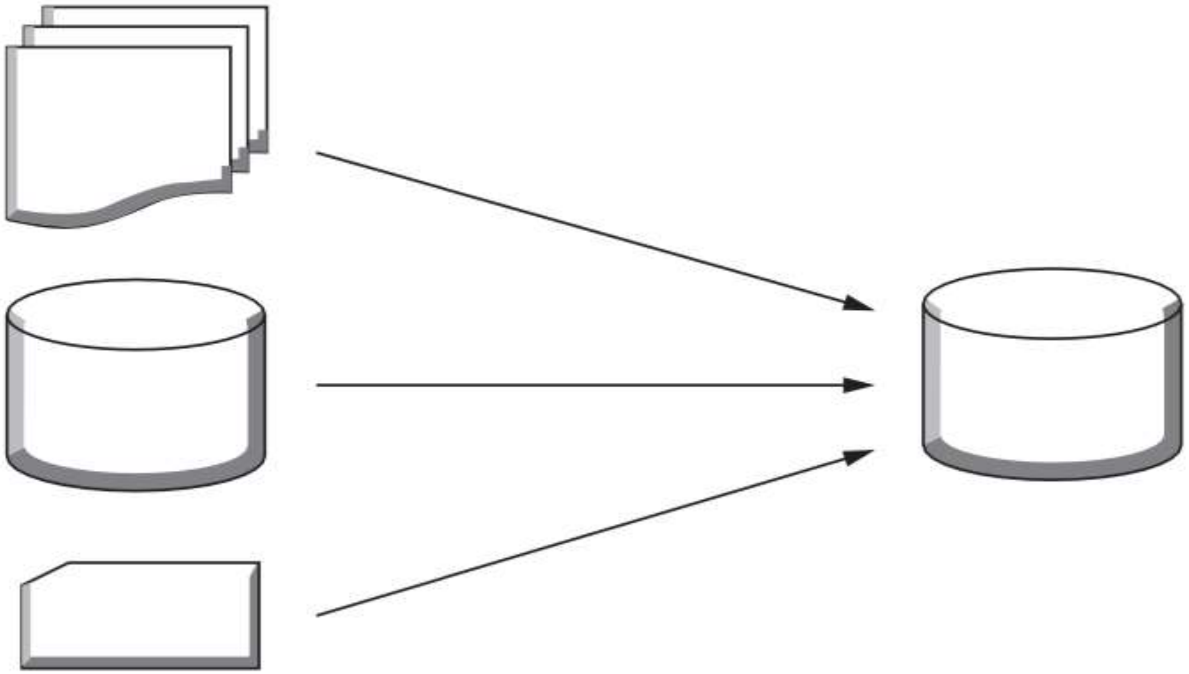
- Inconsistent:
  - Data contains discrepancies in codes or names.
  - For example, if the “Name” column for registration records of employees contains values other than alphabetical letters, or if records do not start with a capital letter, discrepancies are present.

- The most important tasks involved in data pre-processing are:
  - Data Cleaning
  - Data Integration
  - Data Transformation
  - Data Reduction
  - 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

- Since there are several reasons why data could be "dirty," there are just as many ways to "clean" it.
- There are three key methods that describe ways in which data may be "cleaned," or better organized, or scrubbed of potentially incorrect, incomplete, or duplicated information.
  1. Data Munging
  2. Handling Missing Data
  3. Smooth Noisy Data

Data Munging

- Often, the data is not in a format that is easy to work with.
- For example, it may be stored or presented in a way that is hard to process.
- Thus, we need to convert it to something more suitable for a computer to understand.
- To accomplish this, there is no specific scientific method.
- The approaches to take are all about manipulating or wrangling (or munging) the data to turn it into something that is more convenient or desirable.
- This can be done manually, automatically, or, in many cases, semi-automatically.
- Consider the following text recipe.
- "Add two diced tomatoes, three cloves of garlic, and a pinch of salt in the mix."
- This can be turned into a table

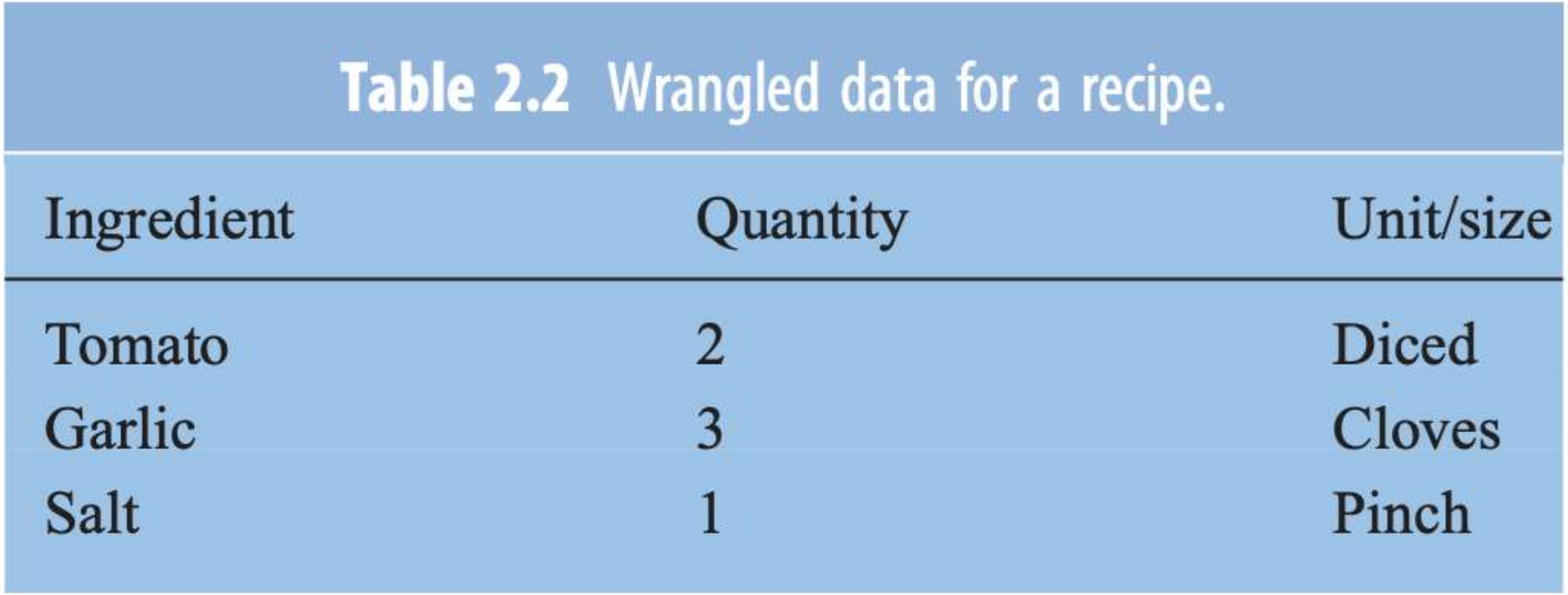


Table 2.2 Wrangled data for a recipe.

Ingredient	Quantity	Unit/size
Tomato	2	Diced
Garlic	3	Cloves
Salt	1	Pinch

Handling Missing Data

- Sometimes data may be in the right format, but some of the values are missing.
- Consider a table containing customer data in which some of the home phone numbers are absent. This could be due to the fact that some people do not have home phones, instead they use their mobile phones as their primary or only phone.
- Other times data may be missing due to problems with the process of collecting data, or an equipment malfunction. Or, comprehensiveness may not have been considered important at the time of collection.
- Furthermore, some data may get lost due to system or human error while storing or transferring the data.
- Strategies to combat missing data include ignoring that record, using a global constant to fill in all missing values, imputation, inference-based solutions (Bayesian formula or a decision tree), etc.

Smooth Noisy Data

- There are times when the data is not missing, but it is corrupted for some reason.
- This is, in some ways, a bigger problem than missing data.
- Data corruption may be a result of faulty data collection instruments, data entry problems, or technology limitations.
- For example, a digital thermometer measures temperature to one decimal point (e.g., 70.1°F), but the storage system ignores the decimal points.
- So, now we have 70.1°F and 70.9°F both stored as 70°F. This may not seem like a big deal, but for humans a 99.4°F temperature means you are fine, and 99.8°F means you have a fever, and if our storage system represents both of them as 99°F, then it fails to differentiate between healthy and sick persons!

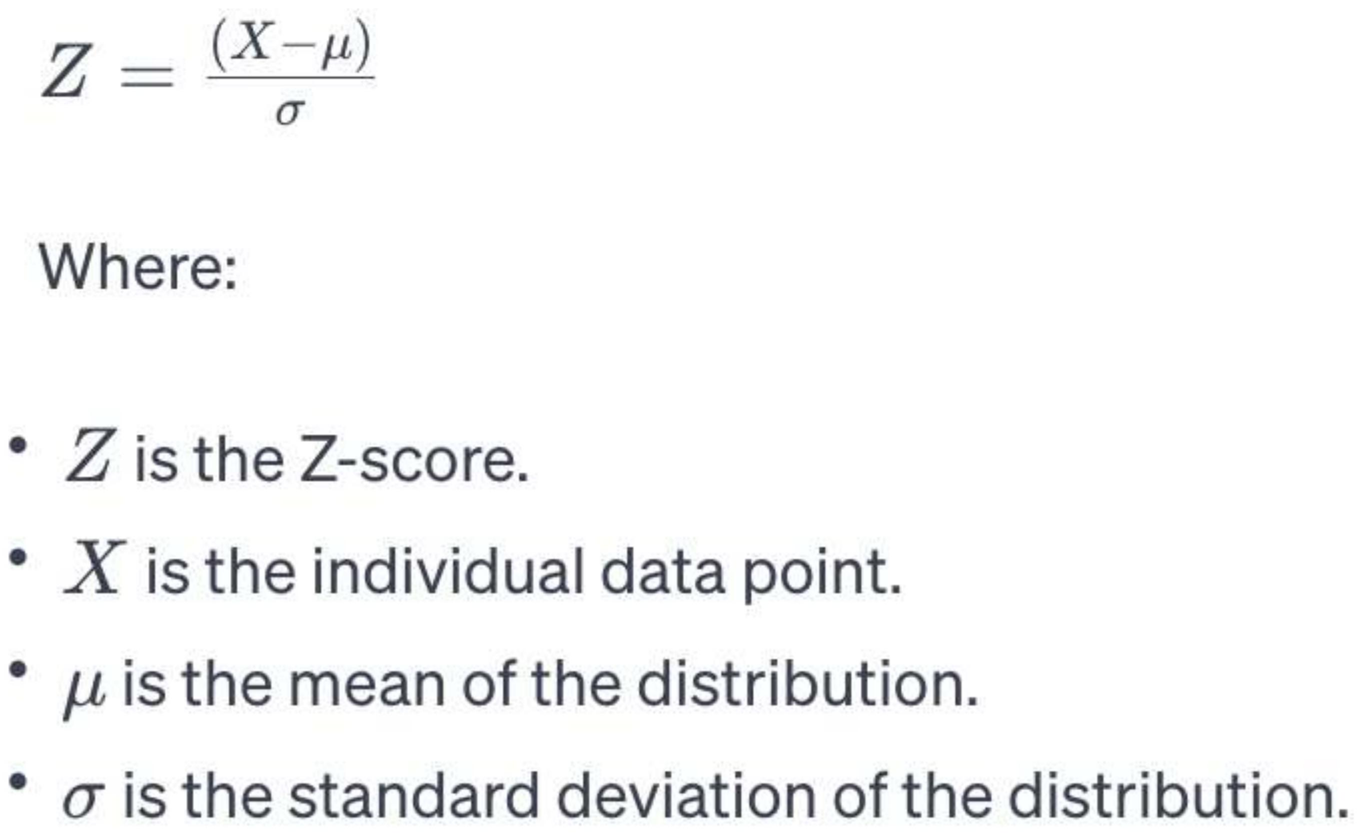
Smooth Noisy Data

- There is no one way to remove noise, or smooth out the noisiness in the data.
- However, there are some steps to try. First, you should identify or remove outliers.
- For example, records of previous students who sat for a data science examination show all students scored between 70 and 90 points, barring one student who received just 12 points.
- It is safe to assume that the last student's record is an outlier (unless we have a reason to believe that this anomaly is really an unfortunate case for a student!).
- Second, you could try to resolve inconsistencies in the data.
- For example, all entries of customer names in the sales data should follow the convention of capitalizing all letters, and you could easily correct them if they are not.

Smooth Noisy Data

- **Simple Moving Average:** Averages a set of data points within a specified window, providing a smoothed representation of the underlying trend.
- **Exponential Moving Average:** Gives more weight to recent data points, smoothing a time series while emphasizing the most recent trends.
- **Z-Score:** Measures how many standard deviations a data point is from the mean, identifying outliers based on their deviation from the average.
- **Inter Quartile Range (IQR):** Defines a range between the first and third quartiles, detecting outliers based on values outside this range.

Z-Score

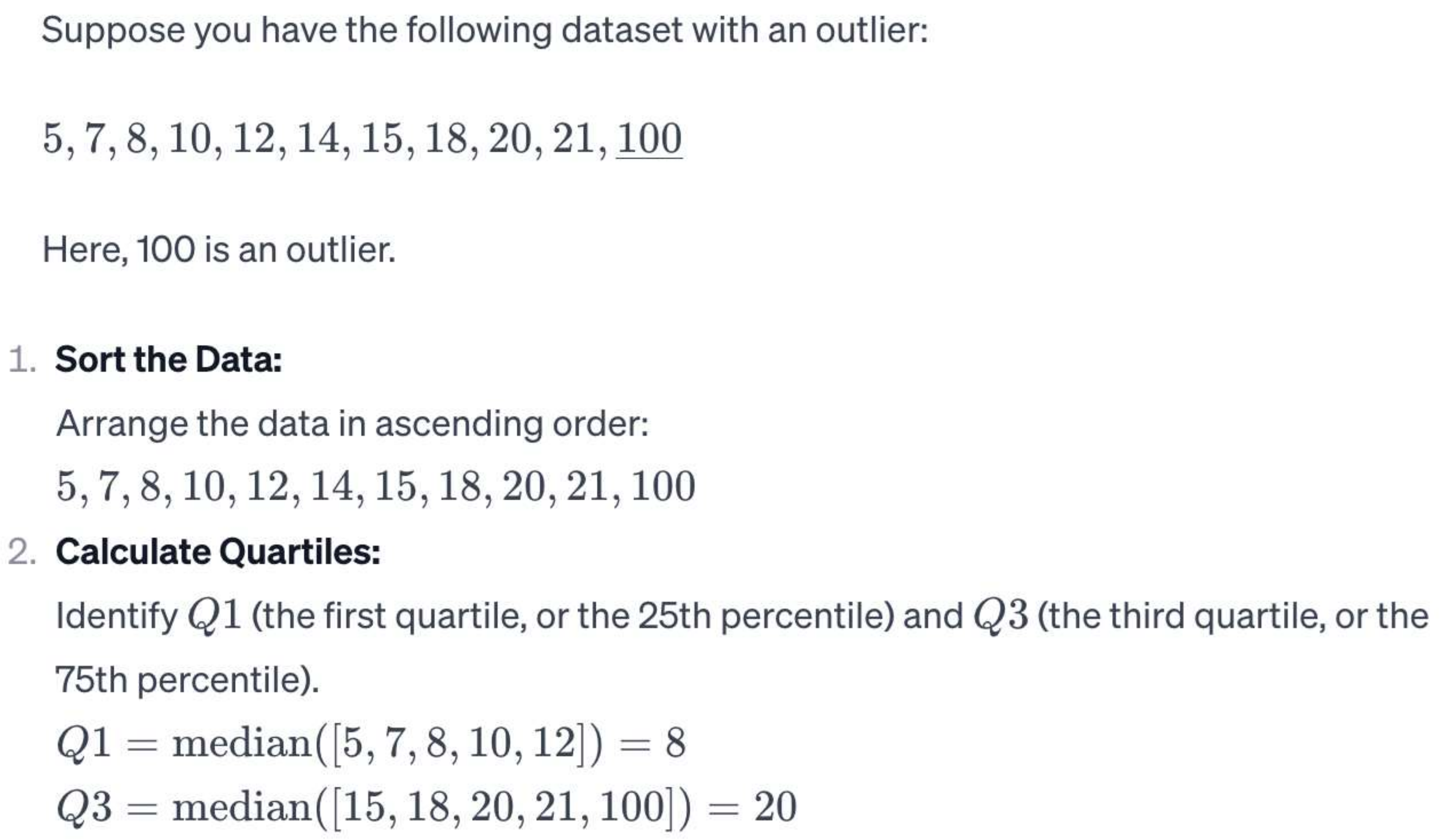


$$Z = \frac{(X - \mu)}{\sigma}$$

Where:

- Z is the Z-score.
- X is the individual data point.
- μ is the mean of the distribution.
- σ is the standard deviation of the distribution.

IQR



Suppose you have the following dataset with an outlier:

$$5, 7, 8, 10, 12, 14, 15, 18, 20, 21, \underline{100}$$

Here, 100 is an outlier.

1. **Sort the Data:**

   Arrange the data in ascending order:

   $$5, 7, 8, 10, 12, 14, 15, 18, 20, 21, 100$$

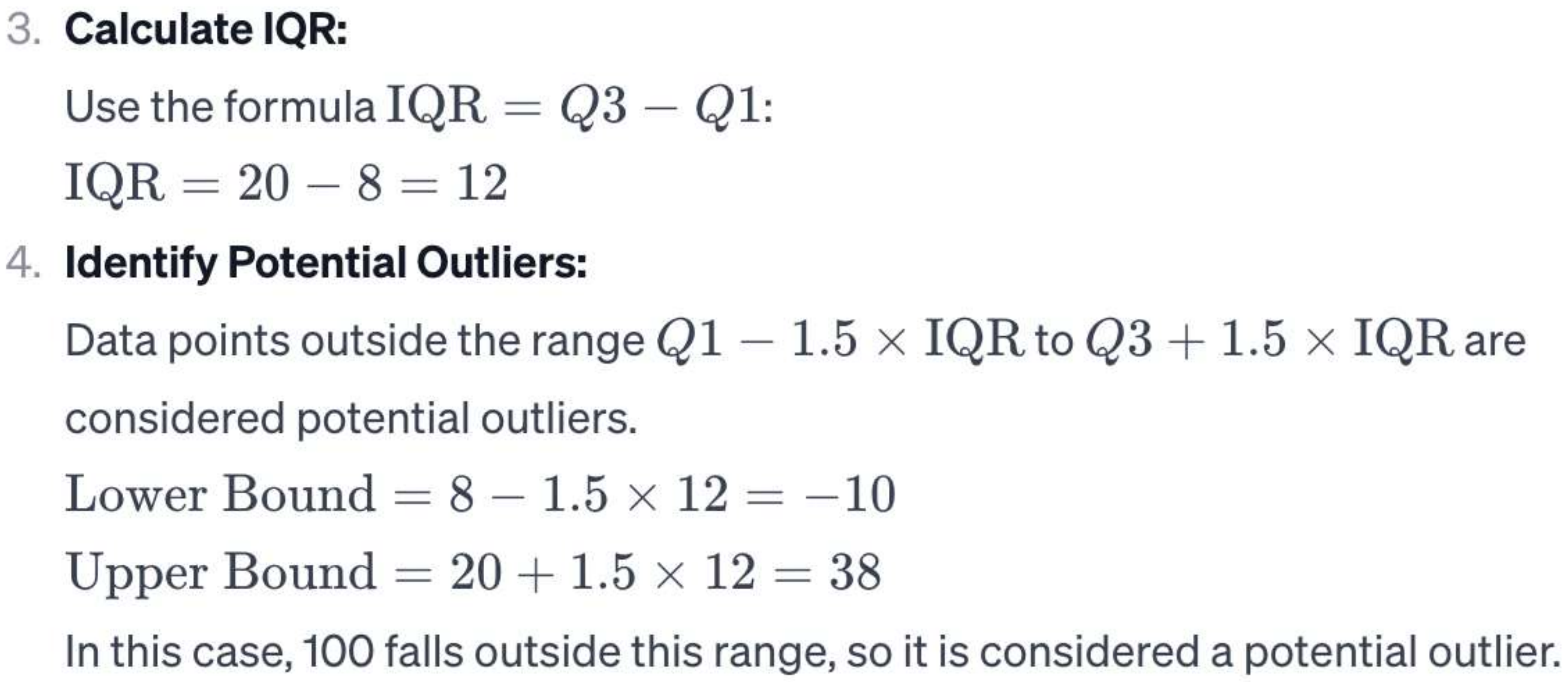
2. **Calculate Quartiles:**

   Identify $Q1$ (the first quartile, or the 25th percentile) and $Q3$ (the third quartile, or the 75th percentile).

   $$Q1 = \text{median}([5, 7, 8, 10, 12]) = 8$$

   $$Q3 = \text{median}([15, 18, 20, 21, 100]) = 20$$

IQR



3. **Calculate IQR:**

   Use the formula $\text{IQR} = Q3 - Q1$:

   $\text{IQR} = 20 - 8 = 12$

4. **Identify Potential Outliers:**

   Data points outside the range $Q1 - 1.5 \times \text{IQR}$ to $Q3 + 1.5 \times \text{IQR}$ are considered potential outliers.

   $\text{Lower Bound} = 8 - 1.5 \times 12 = -10$

   $\text{Upper Bound} = 20 + 1.5 \times 12 = 38$

   In this case, 100 falls outside this range, so it is considered a potential outlier.

Data Integration

- To be as efficient and effective for various data analyses as possible, data from various sources commonly needs to be integrated.
- The following steps describe how to integrate multiple databases or files.
  1. Combine data from multiple sources into a coherent storage place (e.g., a single file or a database).
  2. Engage in schema integration, or the combining of metadata from different sources.

Data Integration

3. Detect and resolve data value conflicts. For example:
   a. A conflict may arise; for instance, such as the presence of different attributes and values from various sources for the same real-world entity.
   b. Reasons for this conflict could be different representations or different scales; for example, metric vs. British units.

Data Integration

4. Address redundant data in data integration. Redundant data is commonly generated in the process of integrating multiple databases. For example:
   a. The same attribute may have different names in different databases.
   b. One attribute maybe a“derived” attribute in another table; for example, annual revenue.
   c. Correlation analysis may detect instances of redundant data.

Data Transformation

□ Data must be transformed so it is consistent and readable (by a system). The following five processes may be used for data transformation.

1. Smoothing: Remove noise from data.
2. Aggregation: Summarization, data cube construction.
3. Generalization: Concept hierarchy climbing.

Data Transformation

4. Normalization: Scaled to fall within a small, specified range and aggregation.

   Some of the techniques that are used for accomplishing normalization are:

   a. Min–max normalization.

   b. Z-score normalization.

   c. Normalization by decimal scaling.

5. Attribute or feature construction: New attributes constructed from the given ones.

Aggregation

- Aggregation is a process in data analysis that involves combining and summarizing data from multiple sources or rows into a single value.
- It is often used for creating summary statistics, constructing data cubes, or deriving insights from large datasets.
- There are various aggregation functions that can be applied depending on the nature of the data and the desired summary.
- Two common techniques involving aggregation are summarization and data cube construction.

Aggregation: Summarization

- Summarization is a specific form of aggregation where the goal is to provide a condensed overview of key characteristics in a dataset.
- Summary statistics, such as mean, median, mode, range, and standard deviation, are often used to capture essential features of the data.
- Summarization is crucial for understanding the central tendencies and variability within a dataset.

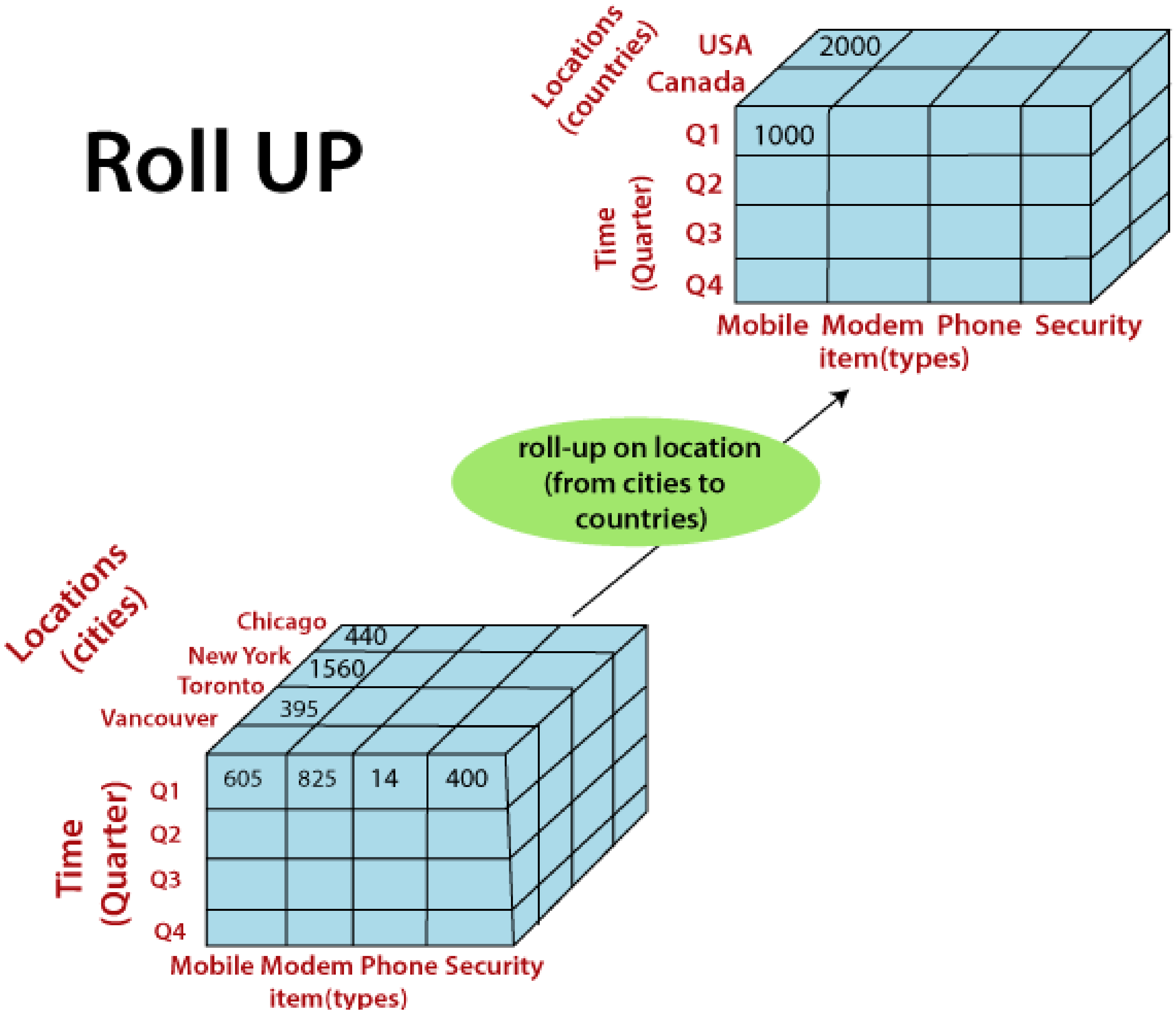
Aggregation: Data Cube Construction

- □ **Definition**: A data cube is a multi-dimensional representation of data that allows for the analysis of information along multiple dimensions.
- □ **Dimensions**: Data cubes have dimensions, which are the categorical variables along which data is analyzed (e.g., time, geography, product).
- □ **Measures**: Measures are the numeric values being analyzed (e.g., sales, revenue).
- □ **Aggregation along Dimensions**: Data cubes involve aggregating measures along different dimensions to provide a comprehensive view of the data.
- □ **OLAP (Online Analytical Processing)**: Data cubes are often associated with OLAP systems, where users can interactively explore and analyze data in a multidimensional way.

Data Cube Operations: Roll-Up

- The roll-up operation (also known as drill-up or aggregation operation) performs aggregation on a data cube, by climbing down concept hierarchies, i.e., dimension reduction.
- Roll-up is like zooming-out on the data cubes.

Roll UP



roll-up on location (from cities to countries)

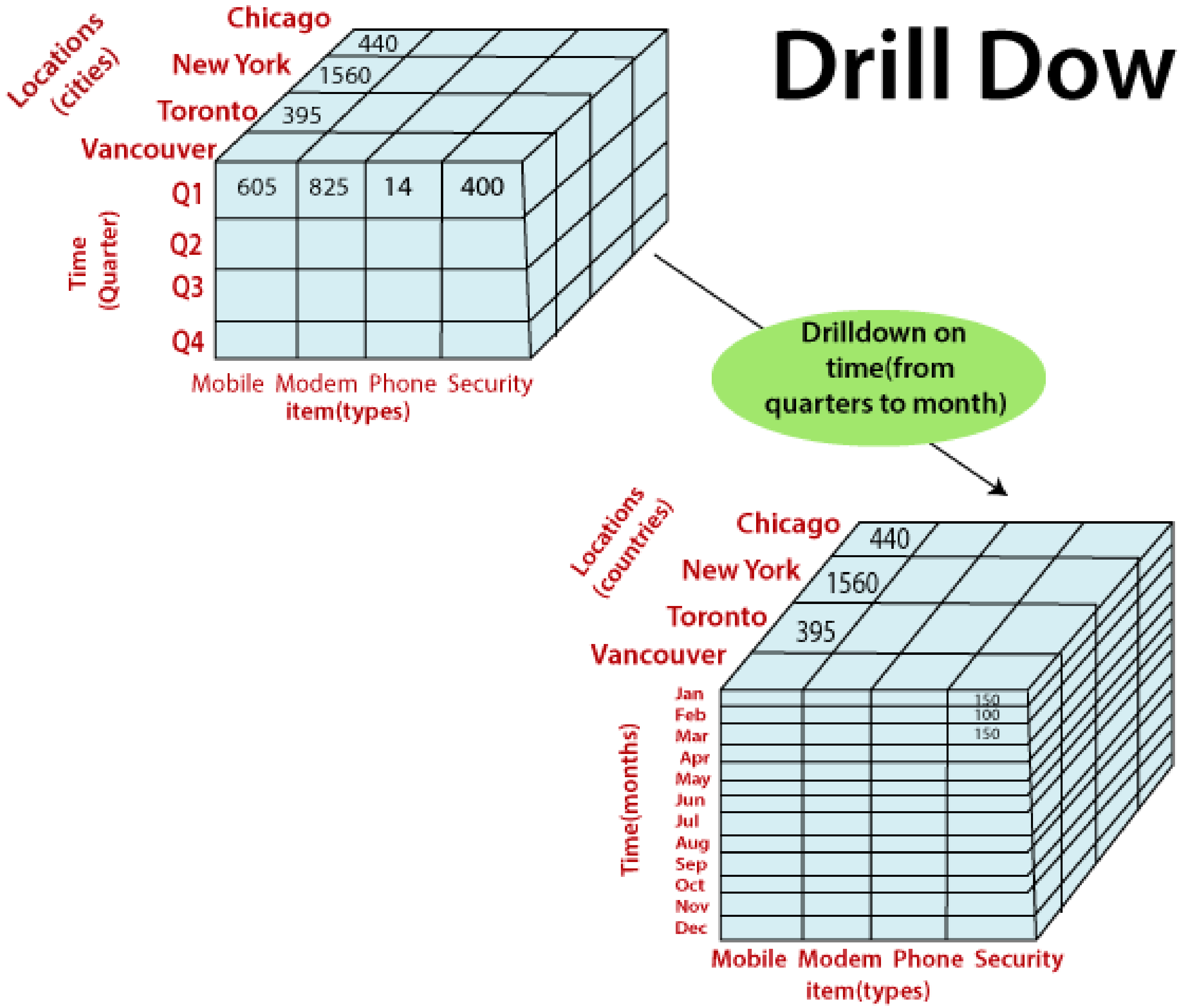
Data Cube Operations: Drill-Down

□ The drill-down operation (also called roll-down) is the reverse operation of roll-up.

□ Drill-down is like zooming-in on the data cube.

□ It navigates from less detailed record to more detailed data.

Drill Down



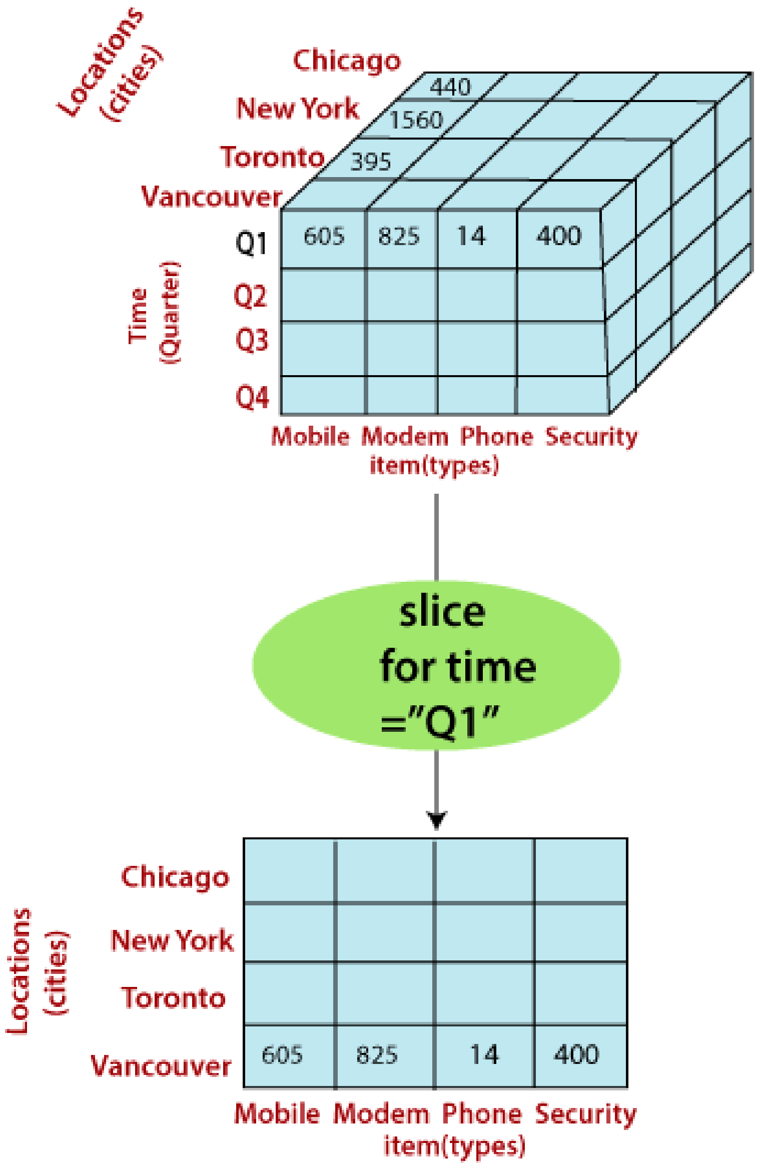
Drilldown on time(from quarters to month)

Data Cube Operations: Slice & Dice

- A **slice** is a subset of the cubes corresponding to a single value for one or more members of the dimension.
- For example, a slice operation is executed when the customer wants a selection on one dimension of a three-dimensional cube resulting in a two-dimensional site.

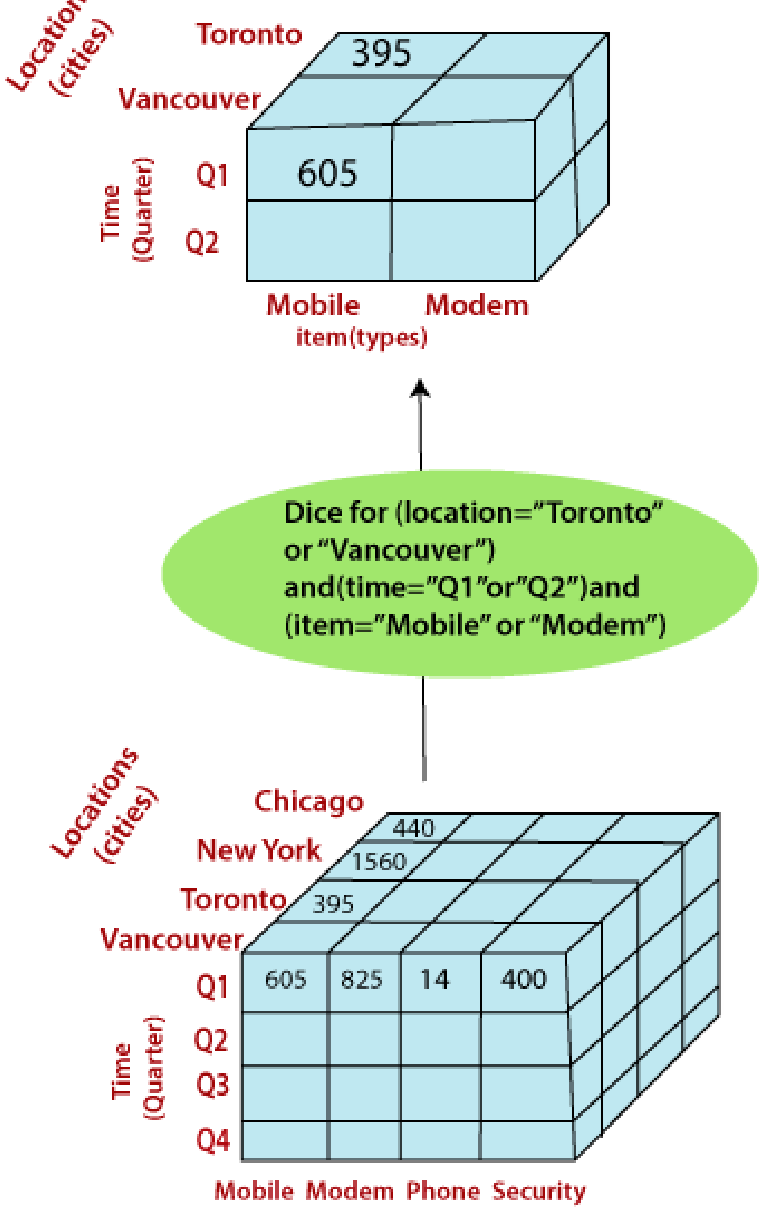
- The **dice** operation describes a subcube by operating a selection on two or more dimension.

Slice



	Mobile	Modem	Phone	Security
Chicago				
New York				
Toronto				
Vancouver	605	825	14	400

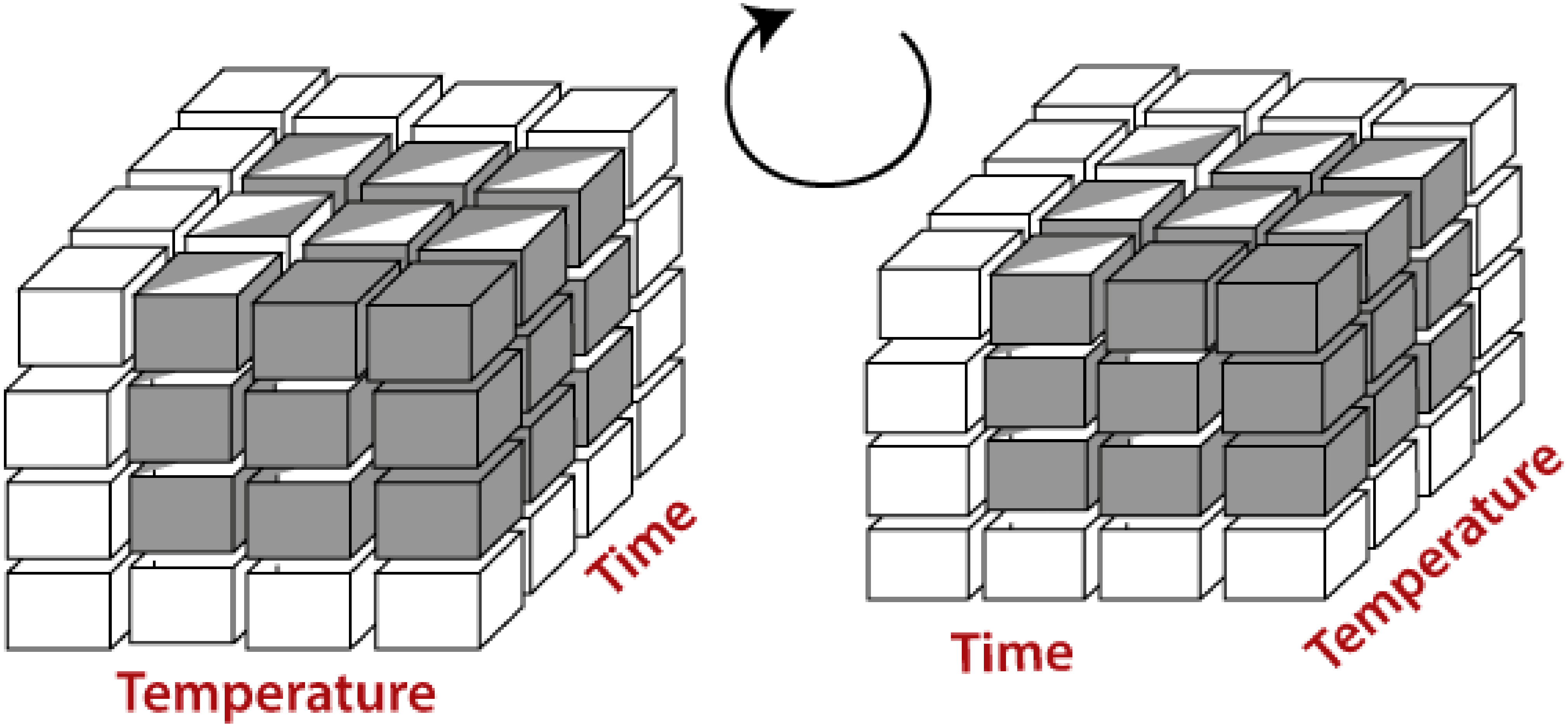
Dice



Data Cube Operations: Pivot

- The **pivot** operation is also called a rotation.
- Pivot is a visualization operations which rotates the data axes in view to provide an alternative presentation of the data.

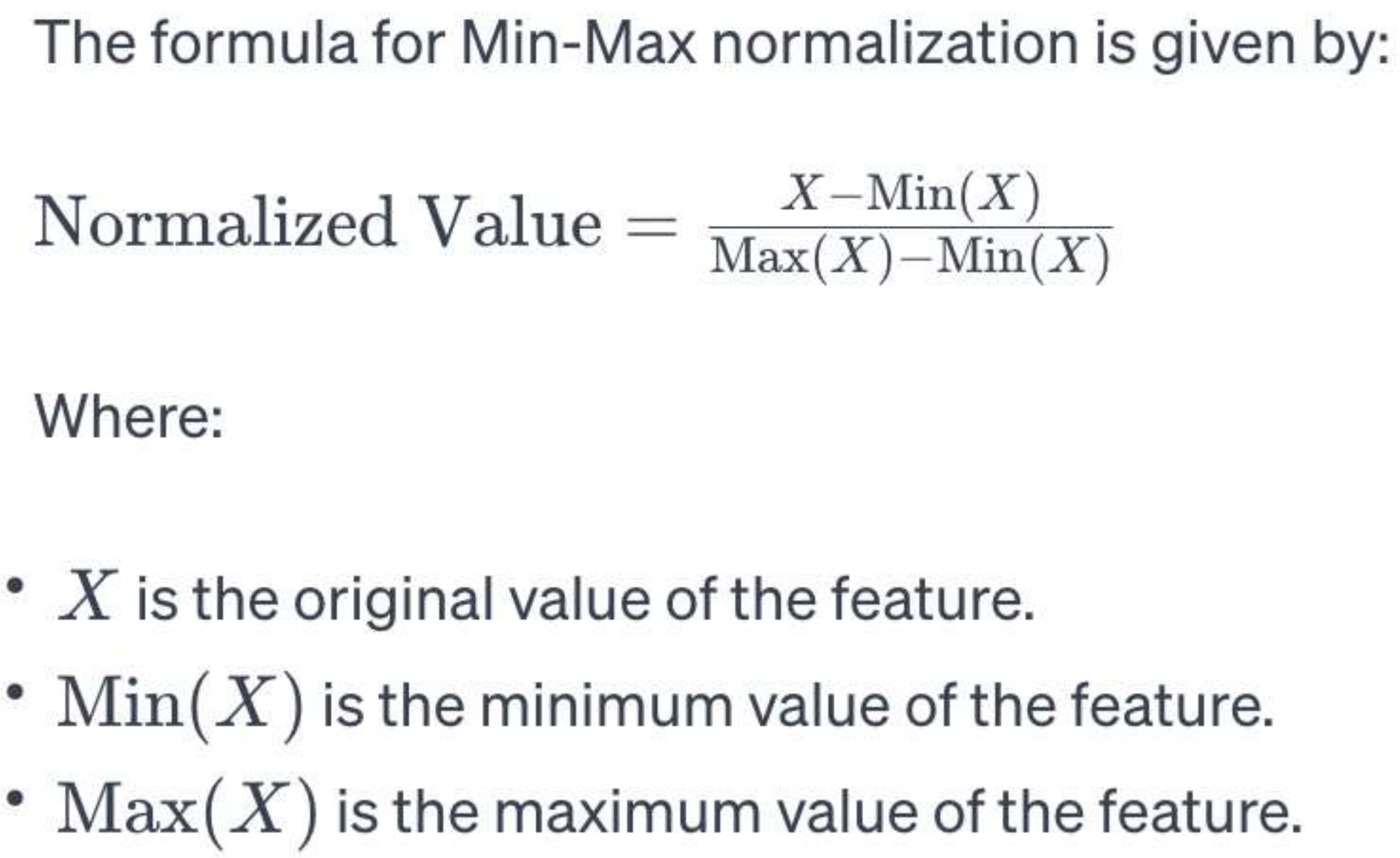
Pivot



Generalization: Concept Hierarchy Climbing.

- Generalization is the process of summarizing detailed and specific data into more abstract and generalized forms. It involves moving up in the hierarchy to a higher level of detail.
- Concept hierarchy climbing is the process of navigating up the levels of a concept hierarchy to access more general or summarized data.
- Example: Starting from daily sales data, climbing the concept hierarchy would involve aggregating to monthly, quarterly, and yearly levels.
- Allows users to view data at different levels of abstraction based on their analytical needs.

Min–max normalization



The formula for Min-Max normalization is given by:

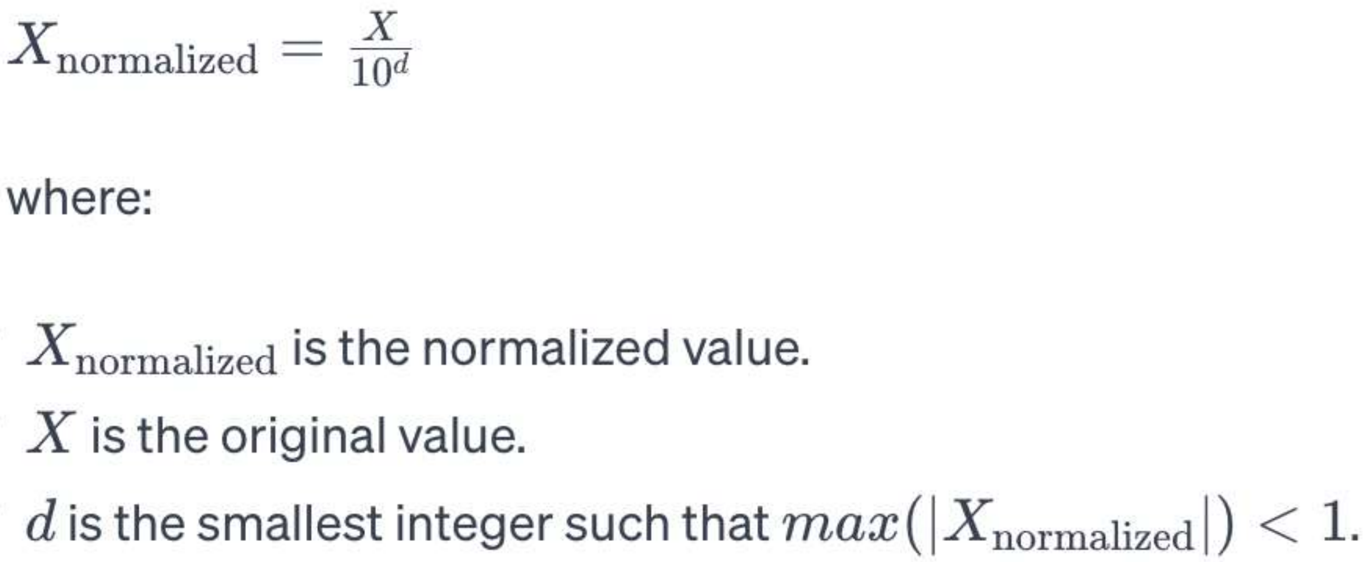
$$\text{Normalized Value} = \frac{X - \text{Min}(X)}{\text{Max}(X) - \text{Min}(X)}$$

Where:

- X is the original value of the feature.
- $\text{Min}(X)$ is the minimum value of the feature.
- $\text{Max}(X)$ is the maximum value of the feature.

Normalization by Decimal Scaling

- Decimal scaling is one method of normalization where the values are scaled by a power of 10.
- The goal is to bring all the values within a feature to a similar scale without changing their relative proportions.
- The formula for decimal scaling normalization is:

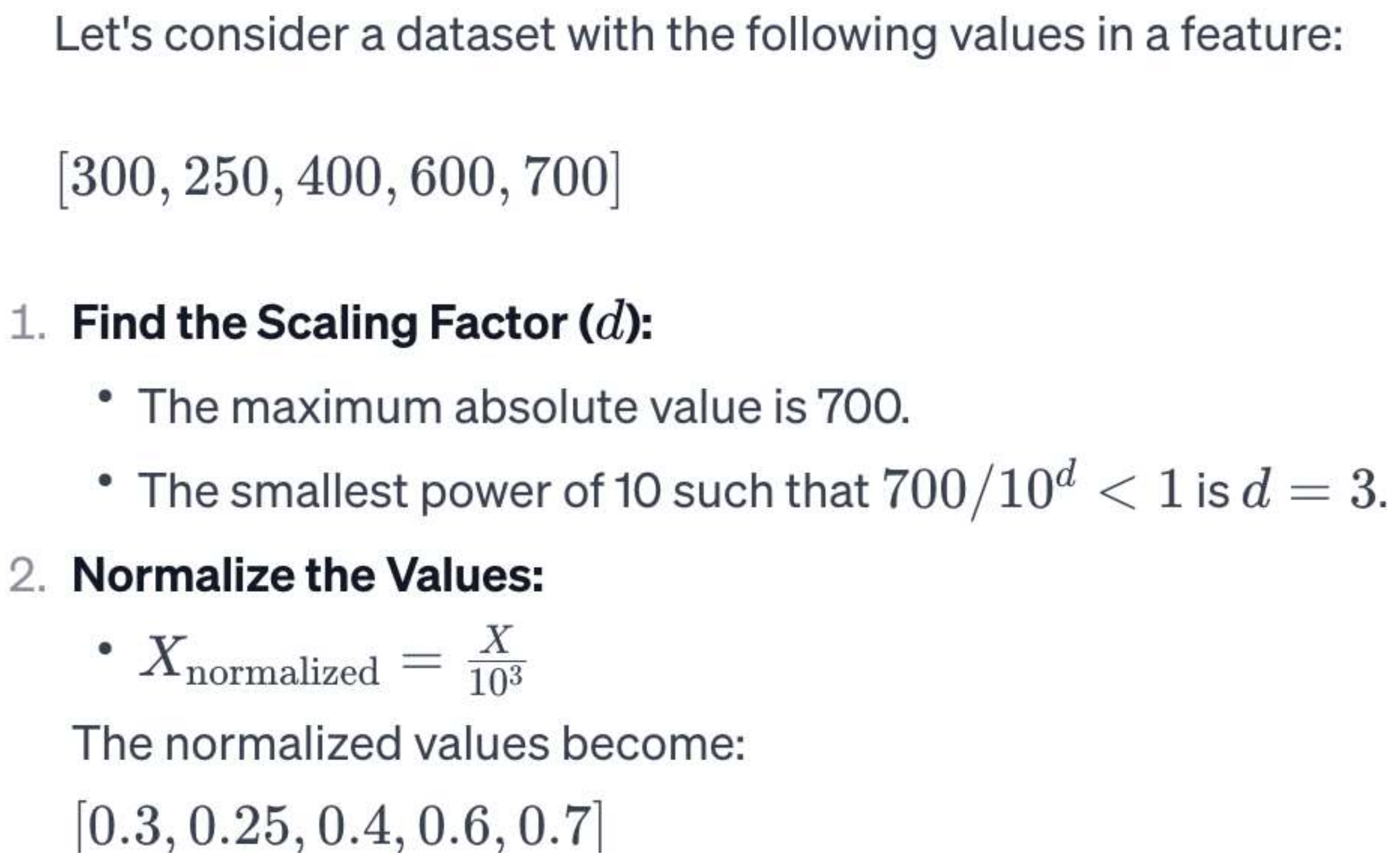


$$X_{\text{normalized}} = \frac{X}{10^d}$$

where:

- $X_{\text{normalized}}$ is the normalized value.
- X is the original value.
- d is the smallest integer such that $max(|X_{\text{normalized}}|) < 1$.

Example



Let's consider a dataset with the following values in a feature:

$$[300, 250, 400, 600, 700]$$

1. **Find the Scaling Factor (d):**
   - The maximum absolute value is 700.
   - The smallest power of 10 such that $700/10^d < 1$ is $d = 3$.
2. **Normalize the Values:**
   - $X_{\text{normalized}} = \frac{X}{10^3}$

   The normalized values become:

   $$[0.3, 0.25, 0.4, 0.6, 0.7]$$

Data Reduction

□ Data reduction is a key process in which a reduced representation of a dataset that produces the same or similar analytical results is obtained.

□ Two of the most common techniques used for data reduction.

1. Data Cube Aggregation
2. Dimensionality Reduction

Dimensionality Reduction

- Dimensionality reduction is a technique used in machine learning and data analysis to reduce the number of features or variables in a dataset.
- High-dimensional datasets, where the number of features is large, can pose challenges such as increased computational complexity, overfitting, and difficulty in visualization.
- Dimensionality reduction methods aim to capture the most important information in the data while reducing its dimensionality.
- Methods: Principal Component Analysis (PCA), Autoencoders, t-Distributed Stochastic Neighbor Embedding (t-SNE)

Overfitting

- Overfitting in machine learning is when a model learns the training data too well, capturing noise and specific details that don't generalize to new data.
- It leads to high accuracy on training but poor performance on unseen data.

Data Discretization

- We are often dealing with data that are collected from processes that are continuous, such as temperature, ambient light, and a company’s stock price.
- But sometimes we need to convert these continuous values into more manageable parts.
- This mapping is called discretization.

Data Discretization

- There are three types of attributes involved in discretization:
  1. Nominal: Values from an unordered set
  2. Ordinal: Values from an ordered set
  3. Continuous: Real numbers
- To achieve discretization, divide the range of continuous attributes into intervals.
- For instance, we could decide to split the range of temperature values into cold, moderate, and hot, or the price of company stock into above or below its market valuation.

Nominal Values

- In statistics and mathematics, nominal values are categorical data that represent different categories or groups, but the order among these categories is not meaningful.
- Nominal data can only be classified and cannot be ranked or ordered.
- Here's an example to illustrate nominal values:
- Example: Colors of Cars
- Consider a dataset that records the colors of cars in a parking lot. The colors are nominal because they represent different categories, but there is no inherent order among them.

Nominal Values

- The possible colors might include:
  - Red
  - Blue
  - Green
  - Black
  - White
- In this case, you can categorize the cars based on their colors, but you cannot say that one color is "greater" or "higher" than another in any meaningful way.
- The assignment of colors to the cars is arbitrary, and there is no inherent order or ranking associated with the colors.

Ordinal Values

- In contrast to nominal data, ordinal values represent categories with a meaningful order or ranking.
- However, the intervals between the values are not necessarily uniform or measurable.
- Here's an example to illustrate ordinal values:
- Imagine a survey that collects customer satisfaction ratings for a product or service. The ratings are on a scale from 1 to 5:
  1. Very Dissatisfied
  2. Dissatisfied
  3. Neutral
  4. Satisfied
  5. Very Satisfied

Ordinal Values

- In this case, the values have a clear order, with "Very Dissatisfied" being the lowest level of satisfaction and "Very Satisfied" being the highest.
- However, the intervals between the satisfaction levels are not necessarily uniform or quantifiable.
- The difference in satisfaction between "Dissatisfied" and "Neutral" may not be the same as the difference between "Satisfied" and "Very Satisfied."
- The data is ordinal because there is a meaningful order to the satisfaction levels, but the intervals between the categories are subjective and may not be consistently measurable.

Continuous Values

- Continuous data involves real numbers and represents measurements that can take any value within a given range.
- Unlike discrete data, which consists of distinct and separate values, continuous data can have an infinite number of possible values.
- Real numbers can include decimals and fractions, allowing for a continuous spectrum.
- Here's an example:
- Example: Height Measurement
- Consider measuring the height of individuals. Heights are continuous data because a person's height can take any value within a certain range.

Continuous Values

- You could measure someone's height as 165.2 cm, and it's conceivable that the next measurement might be 165.201 cm, with an infinite number of possible values between them.
- In this case, height is a continuous variable because it can be measured with a high level of precision, and there is no limit to the number of decimal places that could be considered.
- Continuous data is often associated with measurements in the physical world, where values can be as precise as the measuring instruments allow.