




Awarding
Great British
Qualifications



Introduction to Data Science and Big Data
Topic 2: Lecture 1
Introduction to Data

Unit Syllabus


- Data Science and Big Data Fundamentals
- **Introduction to Data**
- Understanding Data & Exploration
- Data Pre-Processing
- Data Processing
- Model Selection and Evaluation
- Data Visualization
- Business Intelligence and Tools
- Data Science Ethical and Privacy Issues
- Unit Summary



Scope and Coverage

This topic will cover:


- What is Data?
- How big is Big Data? Some examples
- Sources of Data
- Big Data Challenges
- Data quality and issues
- 5V's of Big Data
- Types of Data: Structured, Unstructured, Semi Structured
- Categories of Data Types: Continuous, Categorical, Text Data, Time Series, Binary



Learning Outcomes

By the end of this topic students will be able to:

- Understand the need and applications of Data Science and Big Data.
- Understand and classify different types of data.
- Understand the data quality and issues.
- Demonstrate a systematic awareness of the theoretical foundations of data processing.



Review Quiz

Introduction to Data Topic 2 - 2.5

- 1. What best describes Data Science?
 - a. Managing large datasets efficiently
 - b. Extracting insights from data using statistical methods **Correct Answer**
 - c. Handling and analyzing extremely large datasets
 - d. Developing algorithms for data storage



Review Quiz

Introduction to Data Topic 2 - 2.6

- 2. What role does statistics play in Data Science?
 - a. It is not relevant to Data Science
 - b. Statistics helps in making informed decisions based on data **Correct Answer**
 - c. Statistics is only useful in Big Data
 - d. It is used for data encryption



Review Quiz

Introduction to Data Topic 2 - 2.7

- 3. Which of the following is a key phase in the Data Science Lifecycle?
 - a. Data Storage
 - b. Data Manipulation
 - c. Model Deployment **Correct Answer**
 - d. Data Encryption



Review Quiz

Introduction to Data Topic 2 - 2.8

- 4. Which phase in the Data Science Process is focused on fine-tuning and optimizing machine learning models?
 - a. Feature Engineering
 - b. Model Deployment
 - c. Model Evaluation **Correct Answer**
 - d. Model Selection

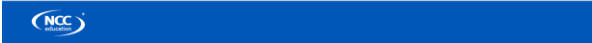


Introduction to Data Topic 2 - 2.9

Review Topic 1

- Data Science and Big Data Fundamentals:
- Data Science is not same as Big Data.
 - Data science is a multidisciplinary field that aims to produce broader insights from the data.
 - Data science is a superset of Data Analytics and Data mining.
 - Data Science Process includes understanding business and data, data exploration, data pre-Processing, data analysis, data modelling, model evaluation and deployment, knowledge &action.

Today, we will go through the first step of data science process and understand **data and Big Data** in more detail

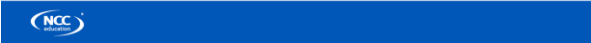


Introduction to Data Topic 2 - 2.10

What is Data?

- In science: Factual information, especially information organised for analysis or used to reason or make a decision
 - ✓ e.g., values derived from scientific experiments
- In computer science: Numerical or other information represented in a form suitable for processing by computers
 - ✓ e.g., values captured via various sensors/devices
- From Latin: Plural of datum, 'a given'

The purpose of computing is insight, not numbers
— Richard Hamming (1962)



Introduction to Data Topic 2 - 2.11

Data Types

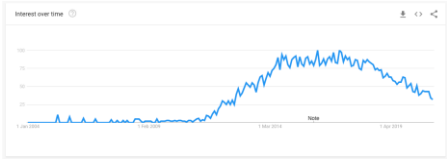
- Physical type (model)
 - ✓ Characterised by storage format
 - ✓ Characterised by machine operations
 - ✓ Examples: Boolean, int, float, double, string, ...
- Abstract type
 - ✓ Provide (conceptual) descriptions of the data
 - ✓ May be characterised by methods/attributes
 - ✓ May be organised into a hierarchy
 - ✓ Examples: qualitative, quantitative, categorical, ordered, plants, animals, fungi, bacteria, ...



Introduction to Data Topic 2 - 2.12

Introduction to Big Data

- The term "Big Data" gained momentum around 2005.
- As the name suggests, the term "Big Data" indicate larger, more complex data sets, especially from new data sources.
- These data sets are so voluminous that traditional data processing software just can't manage them.



Introduction to Data Topic 2 - 2.13

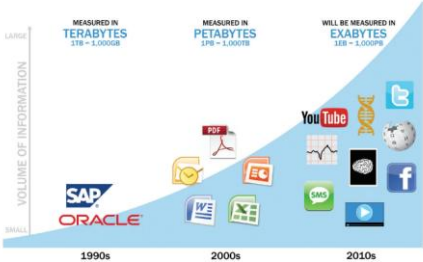
How Big is the Big Data

- Data is expected to double every two years for the next decades
- The amount of data produce every day is mind-blowing:
 - ✓ There are 2.5 quintillion bytes of data created each day at our current pace.
 - ✓ A **single Jet engine** can generate 10+ terabytes of data in 30 minutes of flight time. With many thousand flights per day, generation of data reaches to Petabytes.
 - ✓ More than **3.7 billion humans** use the internet (that's a growth rate of 7.5 percent over 2016).
 - ✓ On average, Google now processes more than **40,000 searches** *every* second (3.5 billion searches per day)!



Introduction to Data Topic 2 - 2.14

How Big is the Big Data



<https://catalogimages.wiley.com/images/db/pdf/9781118876138.excerpt.pdf>



Introduction to Data Topic 2 - 2.15

Big Data Ecosystem



<http://what-when-how.com/Tutorial/topic-7157b5ut/Data-Science-and-Big-Data-Analytics-44.html>



Introduction to Data Topic 2 - 2.16

Big Data Ecosystem

- **Data Devices** and the “Sensor network” gather data from multiple locations and continuously generate new data about this data. For each gigabyte of new data created, an additional petabyte of data is created about that data.
- **Data Collectors** include sample entities that collect data from the device and users.
- **Data Aggregators** make sense of the data collected from the various entities from the “Sensor Network” or the “Internet of Things.” These organisations compile data from the devices and usage patterns collected by government agencies, retail stores and websites.
- **Data Users and Buyers** are the groups directly benefit from the data collected and aggregated by others within the data value chain.



Introduction to Data Topic 2 - 2.17



Introduction to Data Topic 2 - 2.18

-



Introduction to Data Topic 2 - 2.19

- 



Introduction to Data Topic 2 - 2.20


-



Big Data Applications Are Everywhere

Introduction to Data Topic 2 - 2.21

- Agriculture
- Retail
- Cyber security and intelligence
- Crime prediction and prevention
- E-commerce
- Fake news detection
- Fraud detection
- Education
- Weather forecasting
- Tax compliance



NCC

How Big Data Works

Introduction to Data Topic 2 - 2.22

INTEGRATE

Big data brings together data from many disparate sources and applications.
Collate and process in a way it's formatted in useable form.

MANAGE

Big data requires storage which can be cloud, on premises, or both.

VISUALIZE

Get new clarity with a visual analysis of your varied data sets. Obtain patterns and insight through visualization.

ANALYSE

Your investment pays off when analyse and act on data. Build data models with data analytics, ML and AL. Put your data to work.

NCC

Big Data Challenges - Storage

Introduction to Data Topic 2 - 2.23

- Traditional systems are not capable of storing and processing the amount of data that is generated in such speed.
- Example:
 - ✓ Suppose that we need to store/read 100 TB of data.
 - ✓ Even if we do not have space constraint because we have access to large servers...
 - ✓ Assume that the Hard Drive of your server is able to read/write files at 100 MB/sec.
 - ✓ What's the amount of time to read/write 100 TB of data?
It's around 1,000,000 sec → more than 11 days!
- Traditional ways do not scale!
- There should be a better way to do so!

NCC

Big Data Challenges - Storage

Introduction to Data Topic 2 - 2.24

- Organisations still struggle to keep pace with their data and find ways to effectively store it

But it's not enough to just store the data.

- Data must be used to be valuable and that depends on curation
- Data cleaning and curation requires a lot of work
- Data scientists spend 50 to 80 percent of their time curating and preparing data before it can be used
- Data security management
- Keeping up with big data handling tool/technology is an ongoing challenge

NCC

Big Data Challenges –
Data Quality and Issues

Introduction to Data Topic 2 - 2.25

- Data quality refers to the reliability, accuracy, completeness, consistency, and relevance of data.
- It is crucial to have high-quality data for effective decision-making, analysis, and deriving meaningful insights. However, several issues can affect data quality, including:
 - **Inaccurate Data:** Inaccurate data occurs when incorrect or erroneous information is recorded or entered into a dataset.
 - **Incomplete Data:** Incomplete data refers to missing or partial information within a dataset. It can occur if certain data points were not collected or were not recorded properly.



Big Data Challenges –
Data Quality and Issues

Introduction to Data Topic 2 - 2.26

- **Inconsistent Data:** Inconsistent data occurs when the same attribute or information is recorded differently across different data sources or within the same dataset.
- **Duplicate Data:** Duplicate data refers to the presence of identical or similar records within a dataset.
- **Outdated Data:** Outdated data refers to information that is no longer current or relevant. Data can become outdated due to changes in business processes, system upgrades, or the passage of time.



Big Data Challenges –
Data Quality and Issues

Introduction to Data Topic 2 - 2.27

- **Biased Data:** Biased data occurs when the data collected is skewed or not representative of the entire population or group.
- **Lack of Data Governance:** Data governance refers to the overall management, policies, and processes for ensuring data quality and integrity.

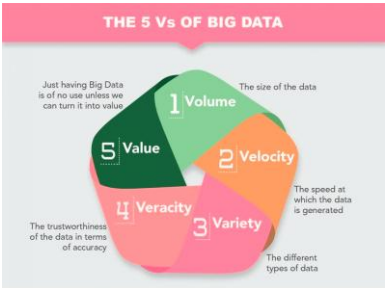
Addressing data quality issues requires implementing robust data management practices, establishing data quality frameworks, ensuring data validation and verification, and regularly monitoring data integrity. Additionally, organisations should invest in data cleansing, deduplication, and validation techniques to improve data quality and reliability.



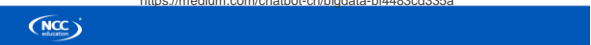
5V's of Big Data

Introduction to Data Topic 2 - 2.28

- Volume
- Velocity
- Variety
- Veracity
- Value



<https://medium.com/chatbot-ch/bigdata-bf4483cd335a>



Types of Big Data

Introduction to Data Topic 2 - 2.29

Big Data can be found in three forms:

- 1. Structured
- 2. Unstructured
- 3. Semi-structured



Structured Data

Introduction to Data Topic 2 - 2.30

- Data that can be stored, accessed and processed in the form of fixed format is termed as a 'structured' data.
- It is typically stored in relational databases or spreadsheets.
- Structured data is highly organised and can be easily categorised, queried, and analysed.
- Examples include transaction records, customer profiles, financial data.

SUMMER SCHOOL STUDENT PARTICIPATION BY STATE				
Data as of August 15, 2016				
State	Number of Students	Peak School Participation	Number of Schools	Total Enrollment
Alabama	1,234	1,234	1	1,234
Alaska	567	567	1	567
Arizona	2,345	2,345	1	2,345
Arkansas	876	876	1	876
California	12,345	12,345	1	12,345
Colorado	3,456	3,456	1	3,456
Connecticut	1,567	1,567	1	1,567
Delaware	987	987	1	987
District of Columbia	2,109	2,109	1	2,109
Florida	15,678	15,678	1	15,678
Georgia	4,567	4,567	1	4,567
Hawaii	1,098	1,098	1	1,098
Idaho	765	765	1	765
Illinois	18,765	18,765	1	18,765
Indiana	2,876	2,876	1	2,876
Iowa	1,432	1,432	1	1,432
Kansas	1,210	1,210	1	1,210
Kentucky	1,321	1,321	1	1,321
Louisiana	1,654	1,654	1	1,654
Maine	876	876	1	876
Massachusetts	2,345	2,345	1	2,345
Michigan	3,210	3,210	1	3,210
Minnesota	2,109	2,109	1	2,109
Mississippi	1,098	1,098	1	1,098
Missouri	2,987	2,987	1	2,987
Montana	765	765	1	765
Nebraska	1,234	1,234	1	1,234
Nevada	1,567	1,567	1	1,567
New Hampshire	876	876	1	876
New Jersey	3,456	3,456	1	3,456
New Mexico	1,098	1,098	1	1,098
New York	20,123	20,123	1	20,123
North Carolina	4,567	4,567	1	4,567
North Dakota	765	765	1	765
Ohio	3,210	3,210	1	3,210
Oklahoma	1,234	1,234	1	1,234
Oregon	1,567	1,567	1	1,567
Pennsylvania	5,678	5,678	1	5,678
Rhode Island	876	876	1	876
South Carolina	1,234	1,234	1	1,234
South Dakota	765	765	1	765
Tennessee	2,345	2,345	1	2,345
Texas	19,876	19,876	1	19,876
Utah	1,098	1,098	1	1,098
Vermont	876	876	1	876
Virginia	2,109	2,109	1	2,109
Washington	3,456	3,456	1	3,456
West Virginia	765	765	1	765
Wisconsin	2,345	2,345	1	2,345
Wyoming	765	765	1	765



Unstructured Data

Introduction to Data Topic 2 - 2.31

- Unstructured Data refers to data that does not have a predefined format.
- It includes textual data, social media posts, emails, videos, images, audio files, and web logs.
- Unstructured data is more complex to analyse.
- Unstructured data poses challenges in terms of storage, processing, and analysis.



Semi-structured Data

Introduction to Data Topic 2 - 2.32

- Semi-structured lies between structured and unstructured data.
- It has some organisational structure.
- Semi-structured data is typically represented using formats like XML (eXtensible Markup Language) or JSON (JavaScript Object Notation).
- Examples of semi-structured data include log files, sensor data with varying attributes, and data from web APIs.



Introduction to Data Topic 2 - 2.33

Class Activity 1

- Students should form small groups (3-4 students per group).
- Each group take one Big Data challenge (e.g., volume, velocity, variety, veracity, or value).

Each group to discuss and brainstorm examples or real-life scenarios that illustrate their assigned challenge.



Introduction to Data Topic 2 - 2.34

Class Activity 1: Real-life Examples

Volume

- ✓ **Social Media Data:** The massive volume of user-generated content on social media platforms like Facebook and Twitter creates a challenge in storing, processing, and analysing this vast amount of data.
- ✓ **Sensor Data:** IoT devices and sensors generate a tremendous volume of data in industries such as manufacturing, healthcare, and smart cities, requiring efficient handling and processing.



Introduction to Data Topic 2 - 2.35

Class Activity 1: Real-life Examples

Velocity

- ✓ **Financial Transactions:** Financial institutions process a high velocity of transactions per second, requiring real-time data processing and analysis to detect fraudulent activities promptly.
- ✓ **Streaming Data:** Streaming services like Netflix or Spotify continuously collect data on user interactions and preferences, necessitating real-time analysis to provide personalised recommendations.



Introduction to Data Topic 2 - 2.36

Class Activity 1: Real-life Examples

Variety

- ✓ **Multimedia Data:** Platforms like YouTube or Instagram handle diverse types of data, including images, videos, text, and audio. Analysing such data requires techniques capable of handling multiple formats.
- ✓ **Customer Interactions:** Businesses receive data from various sources such as emails, customer support chats, and social media comments, which can be unstructured text data. Extracting insights from these sources requires handling the variety of data formats and structures.



Introduction to Data Topic 2 - 2.37

Class Activity 1: Real-life Examples

Veracity

- ✓ **Social Media Sentiment Analysis:** Analysing social media data for sentiment analysis can be challenging due to the presence of noise, sarcasm, slang, and language variations. Ensuring data accuracy and reliability becomes crucial in drawing meaningful insights.
- ✓ **Sensor Data in Industrial Settings:** Sensor data collected in industrial environments may suffer from inaccuracies, measurement errors, or data corruption, which can impact the validity and trustworthiness of the data.



Introduction to Data Topic 2 - 2.38

Class Activity 1: Real-life Examples

Value

- ✓ **Personalised Marketing:** Extracting valuable insights from customer data enables businesses to deliver personalised marketing campaigns tailored to individual preferences and behaviors.
- ✓ **Predictive Maintenance:** Analysing data from machinery, equipment, and sensors can help identify patterns and predict maintenance requirements, allowing companies to optimise maintenance schedules and reduce downtime.



Introduction to Data Topic 2 - 2.39

Checkpoint Summary

- Big data refers to any large and complex collection of data.
- Data can be classified as structured, unstructured and semi-structured.
- Normal architectures fail to work with a such large volume of data.
- Problem in:
 - ✓ Storing data.
 - ✓ Processing data
 - ✓ Data Quality





Awarding
Great British
Qualifications



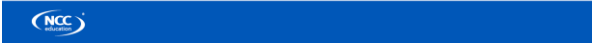
Introduction to Data Science and Big Data
Topic 2: Lecture 2
Categories of Data Types

Introduction to Data Topic 2 - 2.41

Categorising the Data Types

Data types can be categorised into several broad categories based on their characteristics and usage. Here are some common ways to categorise data types:

- Numerical Data
- Categorical Data
- Textual Data
- Temporal Data
- Multimedia Data
- Derived Data

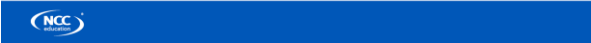


Introduction to Data Topic 2 - 2.42

Numerical Data

Discrete: Integers that represent whole numbers, such as counts or identifiers.

Continuous: Real numbers that represent measurements or quantities with a continuous range.



Introduction to Data Topic 2 - 2.43

Numerical Data - Discrete

- Discrete data refers to a type of numerical data that represents values that are distinct and separate, with no intermediate values possible within the given range.
- It consists of individual, separate, and countable data points.
- Discrete data is typically based on a finite or countable set of values and is often associated with categorical or qualitative measurements.



Introduction to Data Topic 2 - 2.44

Numerical Data - Discrete

- **Characteristics of Discrete Data:**
 - ✓ Countable and distinct values: Discrete data can only take on specific, separate values, and there are no intermediate values.
 - ✓ Non-continuous: There are no fractional or decimal values between points.
 - ✓ Often represented as whole numbers: Discrete data values are often represented as integers or whole numbers.
- **Examples of Discrete Data:**
 - ✓ Number of students in a class: The count of students is discrete (e.g., 25 students, not 25.5 students).
 - ✓ Number of cars in a parking lot: The count of cars is discrete since you cannot have a fractional number of cars (e.g., 10 cars).
 - ✓ Number of books on a shelf: The count of books is discrete since it involves individual, separate objects (e.g., 50 books, not 50.2 books).



Introduction to Data Topic 2 - 2.45

Numerical Data - Continuous

- Continuous data refers to a type of numerical data that can take on any value within a specified range.
- Unlike discrete data, continuous data is not restricted to specific, separate values, but instead can have infinite possibilities within the given range.
- It represents measurements that can be made with precision and can include fractional or decimal values.



Introduction to Data Topic 2 - 2.46

Numerical Data - Continuous

- **Characteristics of Continuous Data:**
 - ✓ Infinite possibilities: Continuous data can take on an infinite number of values within a given range, including fractional or decimal values.
 - ✓ Continuous and unbroken range: There are no gaps or interruptions in the range of possible values.
 - ✓ Measurable and precise: Continuous data can be measured with precision and can have varying levels of accuracy and precision.
- **Examples of Continuous Data:**
 - ✓ Height: Height is a continuous variable as it can take on any value within a range (e.g., 165.3 cm, 173.8 cm, etc.).
 - ✓ Weight: Weight is continuous data as it can have any value within a range (e.g., 68.5 kg, 75.2 kg, etc.).
 - ✓ Temperature: Temperature is continuous data as it can vary continuously within a range (e.g., 25.7° C, 30.2° C, etc.).



Introduction to Data Topic 2 - 2.47

Categorical Data

- Nominal: Categories without any inherent order or ranking, such as colors or categories.
- Ordinal: Categories with a specific order or ranking, such as ratings or survey responses.



Introduction to Data Topic 2 - 2.48

Categorical Data- Nominal

- Nominal data, also known as categorical data, is a type of data that represents discrete, qualitative, or non-numeric values.
- Nominal data consists of categories or labels that do not have any inherent order or ranking.
- Each category in nominal data is distinct and unrelated to the others, and there is no numerical meaning or relationship between the categories.



Introduction to Data Topic 2 - 2.49

Categorical Data- Nominal

- **Characteristics of Nominal Data:**
 - ✓ Categories or labels: Nominal data consists of distinct categories or labels that represent different groups or classes.
 - ✓ No inherent order: The categories in nominal data have no natural or inherent order. They are merely different and unrelated labels.
 - ✓ Non-numeric values: Nominal data does not involve numerical values or measurements.
- **Examples of Nominal Data:**
 - ✓ Gender: The categories "male" and "female" represent nominal data as they are distinct labels without any inherent order.
 - ✓ Marital status: The categories "married," "single," "divorced," and "widowed" represent nominal data.
 - ✓ Eye color: The categories "blue," "brown," "green," represent nominal data as they represent eye colors without any inherent order.



Introduction to Data Topic 2 - 2.50

Categorical Data- Ordinal

- Ordinal data is a type of categorical data that represents values with a natural order or ranking.
- Unlike nominal data, ordinal data categories have a meaningful relationship or hierarchy between them.
- The categories in ordinal data can be ranked or ordered based on some criteria, but the intervals between the categories may not be equal or precisely quantifiable.



Introduction to Data Topic 2 - 2.51

Categorical Data- Ordinal

- **Characteristics of Ordinal Data:**
 - ✓ Ordered categories: Ordinal data consists of categories or labels that can be arranged in a specific order or ranking.
 - ✓ Meaningful ranking: The order of the categories represents a meaningful relationship, indicating a higher or lower value.
 - ✓ Non-uniform intervals: The intervals between the categories may not be equal or precisely measurable.
- **Examples of Ordinal Data:**
 - ✓ Educational attainment: Categories such as "diploma," "certificate" "bachelor's degree," represent ordinal data. They rank the level of education, but difference between each category may not be uniform.
 - ✓ Socioeconomic status: Categories like "lower class," "middle class," and "upper class" represent ordinal data. They indicate a status, but the intervals between them is not quantifiable.



Introduction to Data Topic 2 - 2.52

Textual Data

- Textual data refers to any form of data that is represented as text, including written or typed words, sentences, paragraphs, documents, articles, social media posts, emails, chat conversations, and more.
- Textual data plays a crucial role in various fields, such as natural language processing, sentiment analysis, information retrieval, text mining, and content analysis.



Introduction to Data Topic 2 - 2.53

Textual Data

- **Characteristics of Textual Data:**
 - ✓ Unstructured format: Textual data often lacks a predefined structure, making it more challenging to process and analyse.
 - ✓ Linguistic complexity: Textual data contains linguistic elements such as grammar, syntax, semantics, and context, which adds complexity. Contextual nuances: Textual data can include ambiguities, sarcasm, slang, abbreviations, misspellings, cultural references, and other contextual nuances, requiring specialised techniques for accurate interpretation.
 - ✓ Large volumes: Textual data can accumulate in massive volumes, especially in sources like social media, blogs, news articles, and customer reviews, making it difficult to analyse manually without automated tools.



Introduction to Data Topic 2 - 2.54

Textual Data

- **Examples of Textual Data:**

I recently watched the new movie that came out last week. The plot was intriguing, and the acting was phenomenal. The special effects were mind-blowing. Overall, it was a fantastic cinematic experience that I would highly recommend to others.

In this example, the textual data consists of a snippet of text expressing someone's opinion. It contains sentences and descriptive language conveying the person's thoughts and impressions of the film.



Introduction to Data Topic 2 - 2.55

Temporal/Time Series Data

- Temporal data, also known as time-series data, refers to any type of data that is associated with a specific time or temporal dimension.
- Date: Represents specific dates without any associated time.
- Time: Represents specific times of the day without any associated date.
- Timestamp: Represents specific points in time, including both date and time.



Introduction to Data Topic 2 - 2.56

Temporal/Time Series Data

- **Characteristics of Temporal Data:**
 - ✓ Time dimension: Temporal data includes a timestamp that represents when each observation or measurement was recorded.
 - ✓ Sequential nature: Temporal data points are ordered chronologically, forming a sequence or time series.
 - ✓ Granularity: Temporal data captured at different time resolutions, such as seconds, minutes, hours, days, months, or longer intervals.
 - ✓ Trends and patterns: Temporal data often exhibits trends, seasonality, cyclical, or other patterns
 - ✓ Irregular intervals: Temporal data may not always be recorded at fixed time intervals, as some data points might be missing or have varying time gaps between them.
 - ✓ Time-dependent relationships: Temporal data may demonstrate dependencies between variables that change over time.



Introduction to Data Topic 2 - 2.57

Temporal/Time Series Data

- **Examples of Temporal Data:** Timestamp | Temperature (° C)

2022-01-01 00:00:00 | 10
2022-01-01 01:00:00 | 9



Introduction to Data Topic 2 - 2.58

Multimedia Data

- Image: Represents visual data in the form of pixels, such as photographs or diagrams.
- Audio: Represents sound data, such as music or speech recordings.
- Video: Represents a sequence of images with accompanying audio, forming a moving picture.



Introduction to Data Topic 2 - 2.59

Multimedia Data

- **Characteristics of Multimedia Data:**
 - ✓ Multiple media formats: Multimedia data such as images, videos, audio, text, 3D models, animations, and interactive elements.
 - ✓ Rich content: Multimedia data provides a rich and immersive experience by combining media elements to convey information.
 - ✓ Large file sizes: Multimedia data tends to have larger file sizes, requiring appropriate storage and processing capabilities.
 - ✓ Synchronisation: In multimedia presentations, the different media components need to be synchronized and coordinated.
 - ✓ Interactivity: Multimedia data often incorporates interactive elements.
- **Examples of Multimedia Data:**
 - ✓ Photos or videos, Surveillance footage, etc.



Introduction to Data Topic 2 - 2.60

Derived Data


- Aggregated: Represents summarised or aggregated data from individual records or observations.
- Calculated: Represents data that is derived or computed based on other data using mathematical or logical operations.
- **Examples of Derived Data:**
 - ✓ Moving Averages: In financial analysis, a 30-day moving average of a stock's closing prices is derived data. It helps smooth out fluctuations to identify trends more easily.
 - ✓ Sentiment Scores: Analyzing customer reviews using NLP to derive sentiment scores (e.g., positive, negative, neutral) for products or services.



Introduction to Data Topic 2 - 2.61

Checkpoint Summary


- Discrete Data: Separate values with no intermediates (e.g., categorical variables, counts).
- Continuous Data: Measurements that can take any value within a range (e.g., real numbers, temperature).
- Nominal Data: Categories with no inherent order (e.g., gender, eye color).
- Ordinal Data: Categories with a predefined order or ranking (e.g., rating scales, educational levels).
- Textual Data: Written or textual information (e.g., documents, social media posts).
- Temporal Data: Observations recorded over time (e.g., stock prices).
- Multimedia Data: Combines different media elements (e.g., images, videos, audio).
- Derived data: Generated or calculated from existing data through transformation or aggregation.



Introduction to Data Topic 2 - 2.62

Discussion Session

Understanding the Importance of Data Types in Data Science Process? How does the choice of data type impact the selection of analytical techniques or methods?




Introduction to Data Topic 2 - 2.63

Discussion Session

The choice of data type impacts the selection of analytical techniques or methods in several ways:

- Statistical Techniques:** Different data types require specific statistical techniques, such as regression for continuous data or chi-square tests for categorical data.
- Visualisation Techniques:** Data types determine the appropriate visualisation methods, such as scatter plots for numerical data or bar charts for categorical data.
- Machine Learning Algorithms:** Data types influence the selection of machine learning algorithms, with different algorithms suited for numerical, categorical, or textual data.




Introduction to Data Topic 2 - 2.64

Discussion Session

- Data Preprocessing:** Data type affects preprocessing steps like missing value handling or data transformation, which vary based on the data type.
- Result Interpretation:** Data type impacts how results are interpreted, with different implications for continuous and categorical variables.
- Data Integration:** When combining datasets with different data types, integration techniques must consider the compatibility of the data types.

Understanding the data type is crucial for selecting the right analytical techniques, interpreting results accurately, and effectively deriving insights from the data.



Topic Summary

Introduction to Data Topic 2 - 2.65

- Big data refers to any large and complex collection of data.
- Normal architectures fail to work with a such large volume of data.
- There are different types of data types.
- Understanding the data type is crucial for selecting the right analytical techniques, interpreting results accurately, and effectively deriving insights from the data.



Next Topic 3:
Understanding Data & Exploration

Introduction to Data Topic 2 - 2.66

- What is Descriptive Statistic?
- What is Inferential Statistic?
- What is EDA? Numerical or Graphical methods
- Aims of EDA?
- Exploratory vs Confirmatory Data Analysis?
- Numerical Methods of EDA: Central Tendency, Measurement of Variability
- Graphical Methods of EDA: histogram, box plot, scatter plot, stem, and leaf plot.



References

Introduction to Data Topic 2 - 2.67

- Grus, J. (2019). [Data Science from Scratch](#). O'Reilly Media.
- Holmes, D. E. (2013). [Big Data: A Very Short Introduction](#). Oxford University Press.
- Kotu, V., & Deshpande, B. (2014). [Predictive analytics and data mining: concepts and practice with RapidMiner](#). Morgan Kaufmann.
- Marz, N., & Warren, J. (2015). [Big Data: Principles and best practices of scalable real-time data systems](#). Manning Publications.
- Provost, F., & Fawcett, T. (2013). [Data Science for Business](#). O'Reilly Media.



Topic 2 – Introduction of Data

Any Questions?