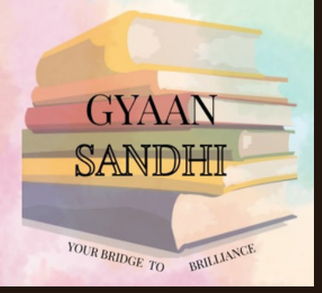


UNIT 1

Data science & big data analytics

Introduction to data science and big data

Data

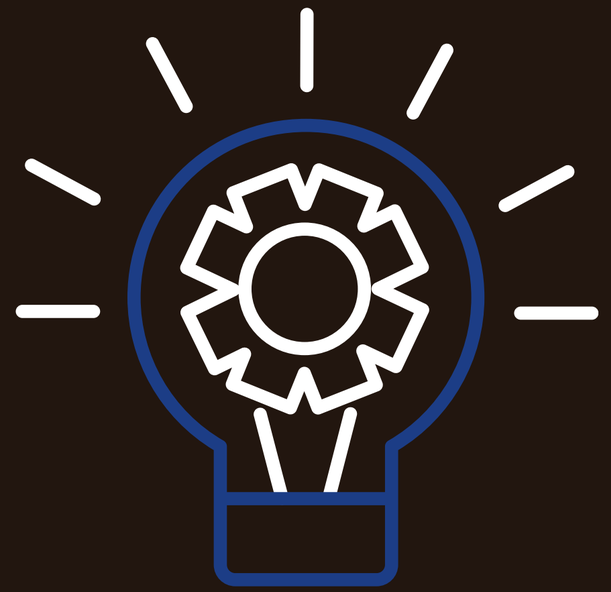
- Data is a collection of raw facts and figures
- It represents something specific, but is not organized

Data can be in the form of:

- Numbers
- Words
- Observations
- Measurements

- Data by itself has no meaning
- It is the raw material used for producing information

Data Science & Big Data



Data Science

- Data Science is an interdisciplinary field
- It uses scientific methods, processes, algorithms, and systems
- It is used to extract knowledge and insights
- Data can be structured or unstructured
- It uses a variety of tools and techniques to extract information from data

Purpose of Data Science

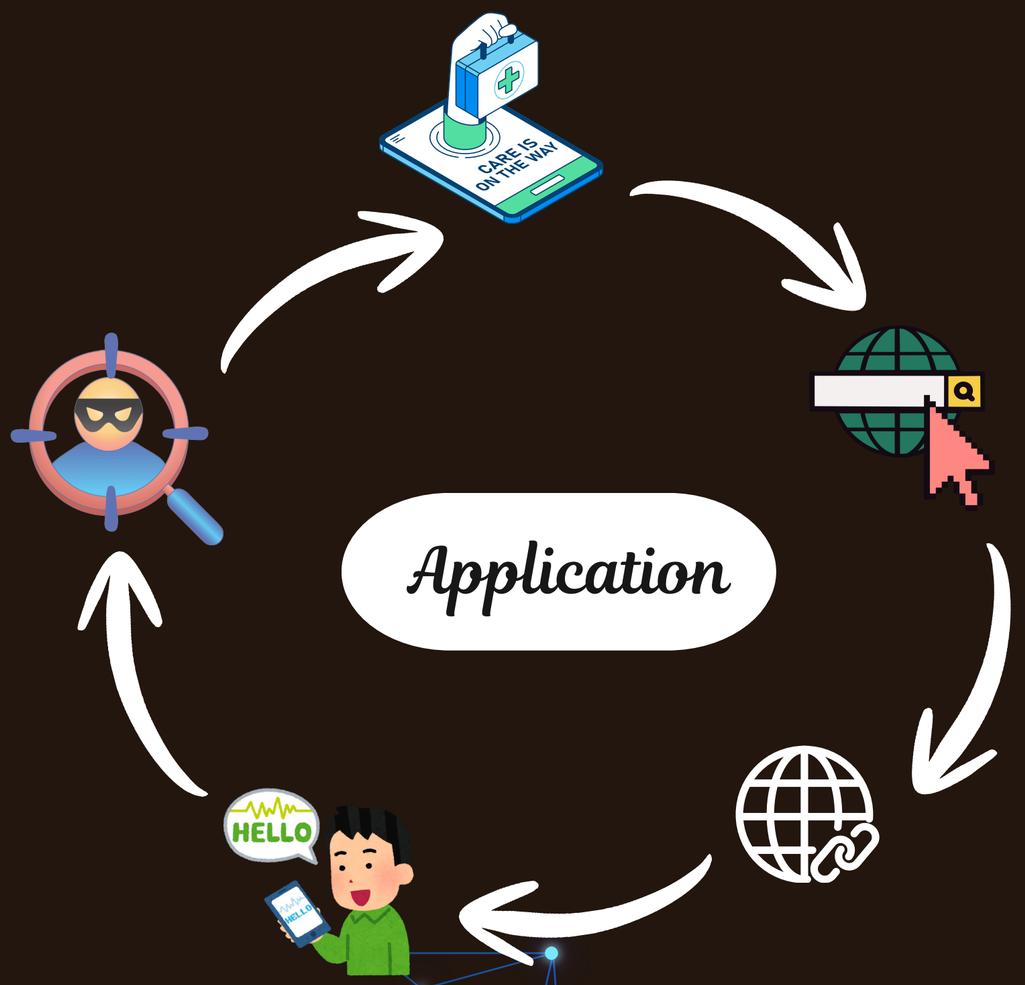
- Main purpose is to find patterns in data
- Helps in data analysis and decision making

Data Science & Big Data

- Uses statistical techniques to analyze data
- Helps to draw useful insights from data

Applications of Data Science

- Fraud and Risk Detection
- Healthcare
- Internet Search
- Website Recommendations
- Speech Recognition



Data Science & Big Data

Big Data

- Big Data is a field that treats and analyzes huge datasets
- It systematically extracts information from very large or complex data
- These datasets are too large or complex to be handled by traditional software
- Big Data helps in better decision making and strategic business planning

Data Science & Big Data

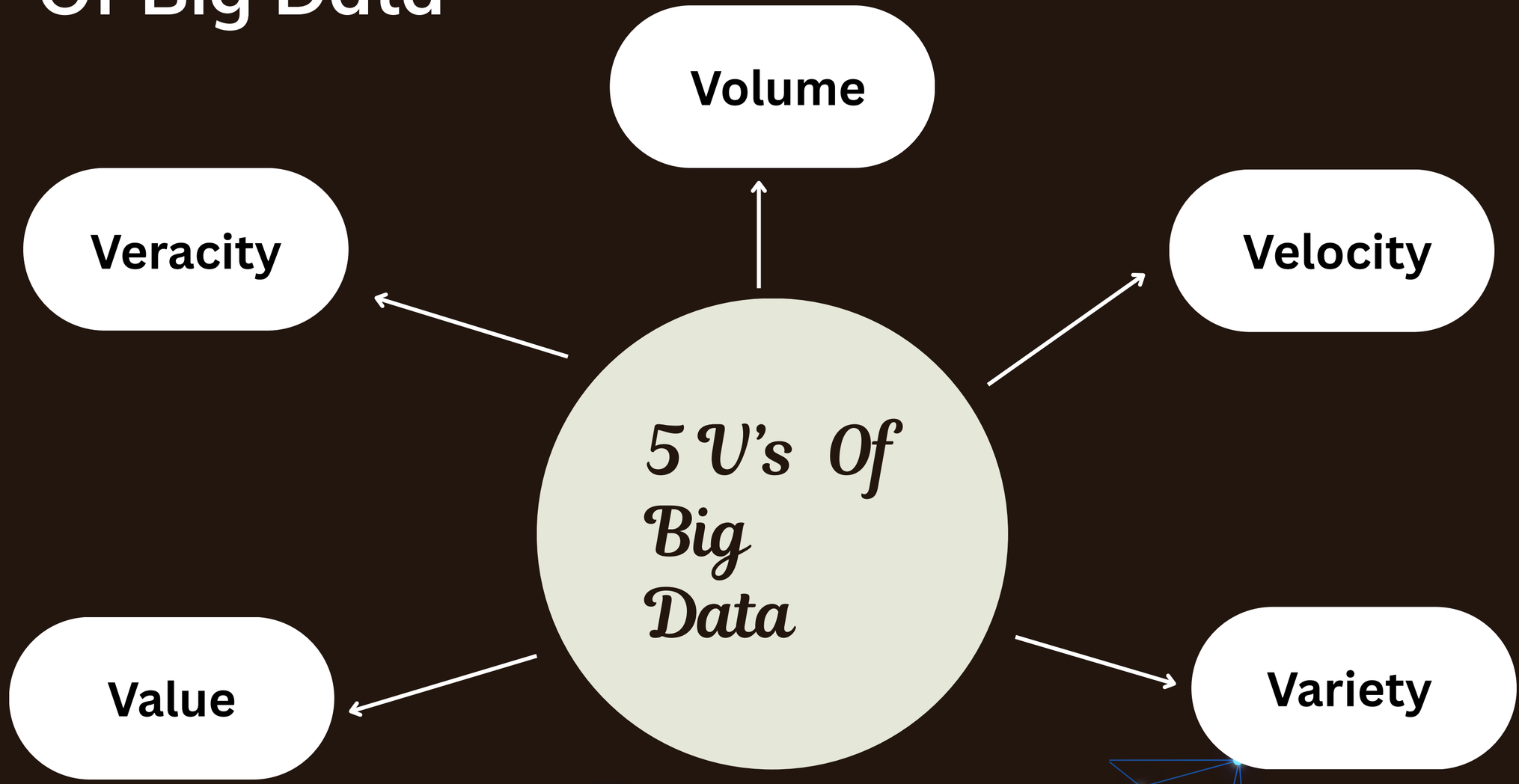
Advantages of Big Data

- Product Development
- Manufacturing
- Marketing
- Price Management

Big Data Processing

- Big Data processing starts with raw data
- Raw data is not aggregated or organized
- Data size is often too large to store in memory
- Cannot be processed using a single computer

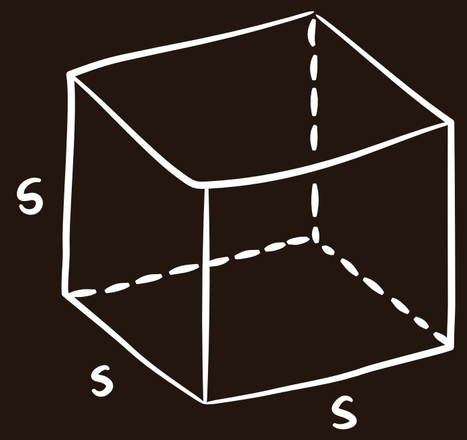
5 V’s Of Big Data



5 V’s Of Big Data

1. Volume

- Refers to the sheer size or quantity of data
- Data is generated or collected in huge amounts
- Measured in Terabytes, Petabytes, Exabytes
- Example: Social media data, transaction records



2. Velocity

- Refers to the speed at which data is generated, processed, and updated
- Data needs to be handled quickly
- Example: Real-time data processing, live streaming data

5 V’s Of Big Data

3. Variety

- Refers to different types and formats of data
- Includes both structured and unstructured data
- Examples: Text, images, videos, audio, emails

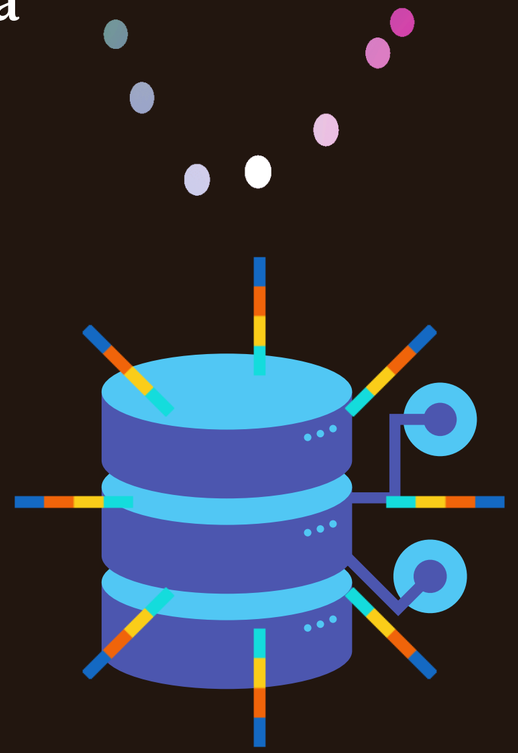
4. Value

- Refers to the importance of extracting meaningful insights
- Focuses on converting data into useful and actionable information
- Example: Business decisions based on data analysis

5 V’s Of Big Data

5. Veracity

- Refers to the accuracy, reliability, and trustworthiness of data
- Deals with noise, inconsistencies, and uncertainties in data
- Example: Cleaning inaccurate or incomplete data



Difference Between Data Science & BI

Data Science	Business Intelligence (BI)
Focuses on **predictive & prescriptive analysis**	Focuses on **descriptive & diagnostic analysis**
Uses **advanced algorithms & machine learning**	Uses reports, dashboards, and queries
Works with **structured & unstructured data**	Mainly works with **structured data**
Helps in **future predictions & pattern discovery**	Helps in **understanding past & present data**
Uses tools like **Python, R, ML models**	Uses tools like Power BI, Tableau, Excel

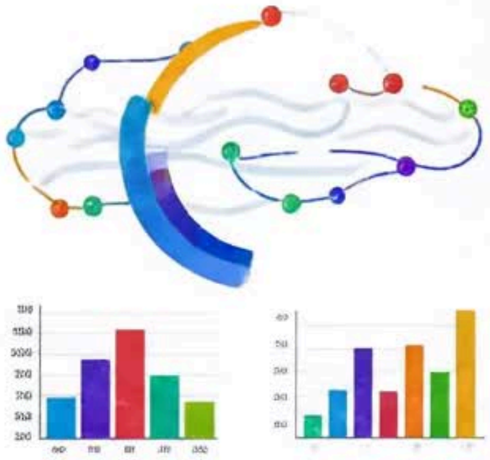
Difference Between Data Science & BI

Data Science	Machine Learning	Artificial Intelligence (AI)
Focuses on extracting insights from data	Focuses on **learning from data automatically**	Focuses on making machines intelligent
Involves data collection, cleaning, analysis	Involves **training models using algorithms**	Involves decision making & problem solving
Uses statistics, visualization, ML	Uses **statistical & mathematical models**	Uses ML, logic, reasoning & rules

Difference Between Data Science & BI

Data Science	Machine Learning	Artificial Intelligence (AI)
Works with structured & unstructured data	Works mainly with training datasets	Can work with or without data
Goal is insight & prediction	Goal is improving performance with experience	Goal is human-like intelligence
Example: Data analysis for business decisions	Example: Spam detection, recommendation systems	Example: Self-driving cars, chatbots

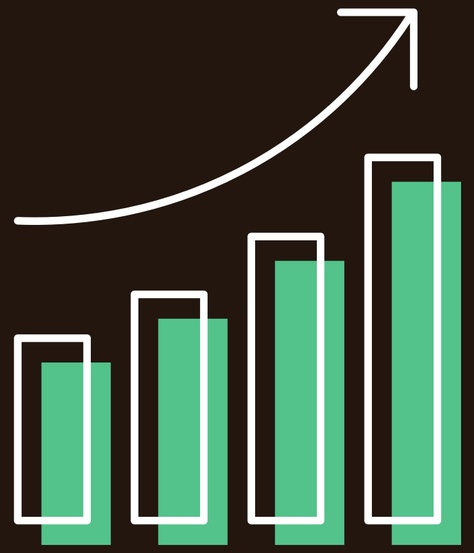
Data Science



Machine Learning

Artificial Intelligence

Data Types



- **Data can be divided into two main types:**
  a. **Qualitative Data**
  b. **Quantitative Data**

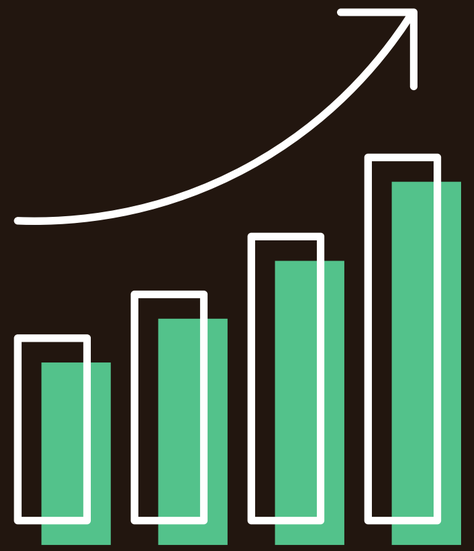
1. Qualitative Data

- Provides information about the quality or characteristics of an object
- Cannot be measured numerically
- Describes categories or labels

a) Nominal Data

- Deals with non-numeric variables

Data Types

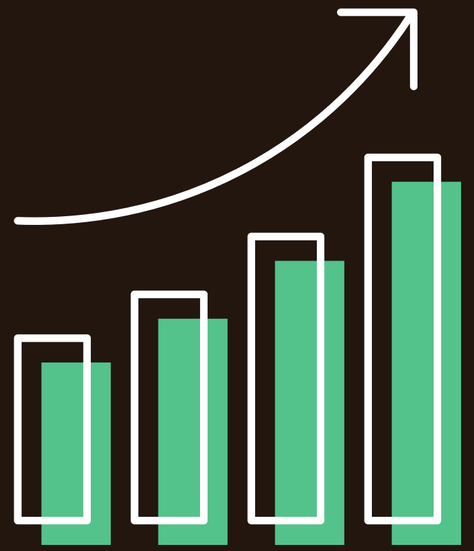


- Data has no specific order
- Example:
- Gender: Male, Female, Other
- Blood group, Color

b) Ordinal Data

- Data values are taken from an ordered set
- Order matters, but difference is not measurable
- Example:
- University ranking (1st, 2nd, 3rd)
- Class grades (A, B, C)

Data Types



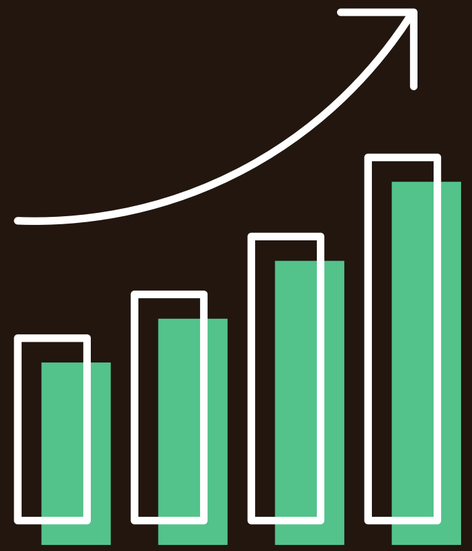
2. Quantitative Data

- Data based on numbers
- Mathematical calculations can be performed
- Measures quantity

a) Interval Data

- Values are chosen from an interval scale
- No true zero point
- Difference between values is meaningful

Data Types

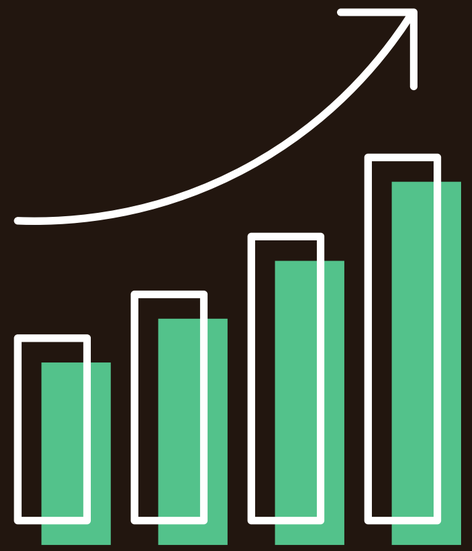


- Example:
  - Temperature in Celsius
  - Dates on a calendar

b) Ratio Data

- Data where ratios can be calculated
- Has a true zero point
- Most meaningful type of data
- Example:
  - Age
  - Weight
  - Height

Data Types

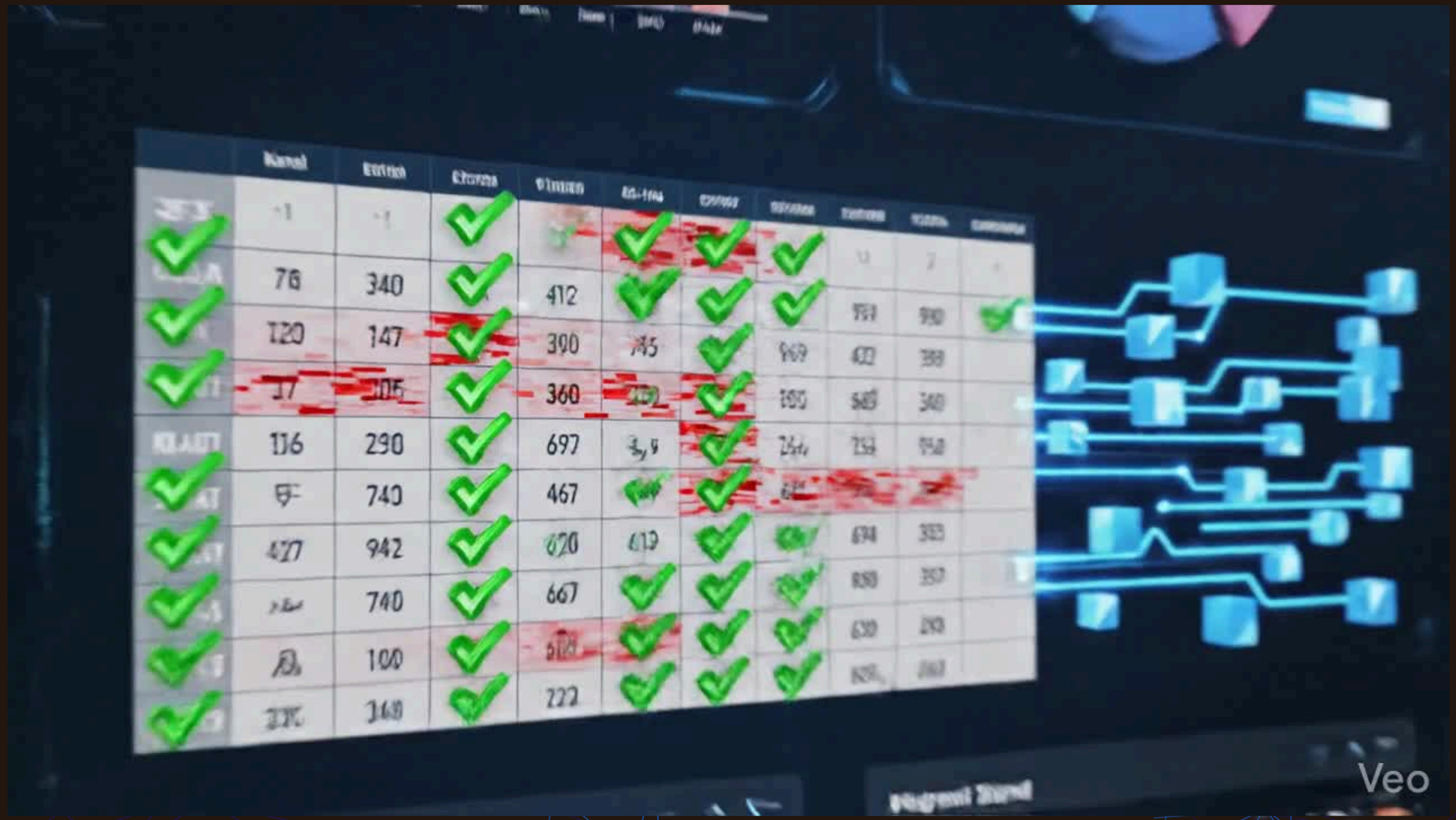


- Example:
  - Temperature in Celsius
  - Dates on a calendar

b) Ratio Data

- Data where ratios can be calculated
- Has a true zero point
- Most meaningful type of data
- Example:
  - Age
  - Weight
  - Height

Data Wrangling



Data Wrangling

Data Wrangling

- Data Wrangling is the process of converting raw data into a format suitable for analysis
- It prepares data for better and more accurate analytics
- The main goal is to ensure data quality and usability

Steps in Data Wrangling

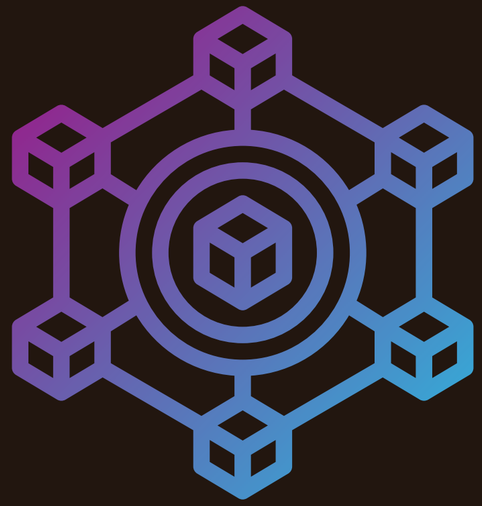
1. Discovering

- First step of data wrangling
- Helps in understanding the data
- Identifies data structure, issues, and patterns

Data Wrangling

2. Structuring

- Data is organized into a proper format
- Converts data into rows, columns, or tables
- Makes data ready for processing



3. Cleaning

- Handles null or missing values
- Removes errors and inconsistencies
- Applies standard formatting

Data Wrangling

4. Enriching

- Determines whether additional data is required
- Extra data is added to improve dataset quality
- Enhances usefulness of data

5. Validating

- Data is checked using validation rules
- Ensures consistency, quality

Data Integration

- Definition: Combines data from multiple sources to form a coherent dataset
- Helps in metadata conflict detection and correlation analysis
- Provides a unified view of collected data and maintains accuracy

Issues in Data Integration

1. Entity Identification Problem – recognizing same entities across datasets
2. Redundancy – duplicate data from multiple sources

Data Transformation

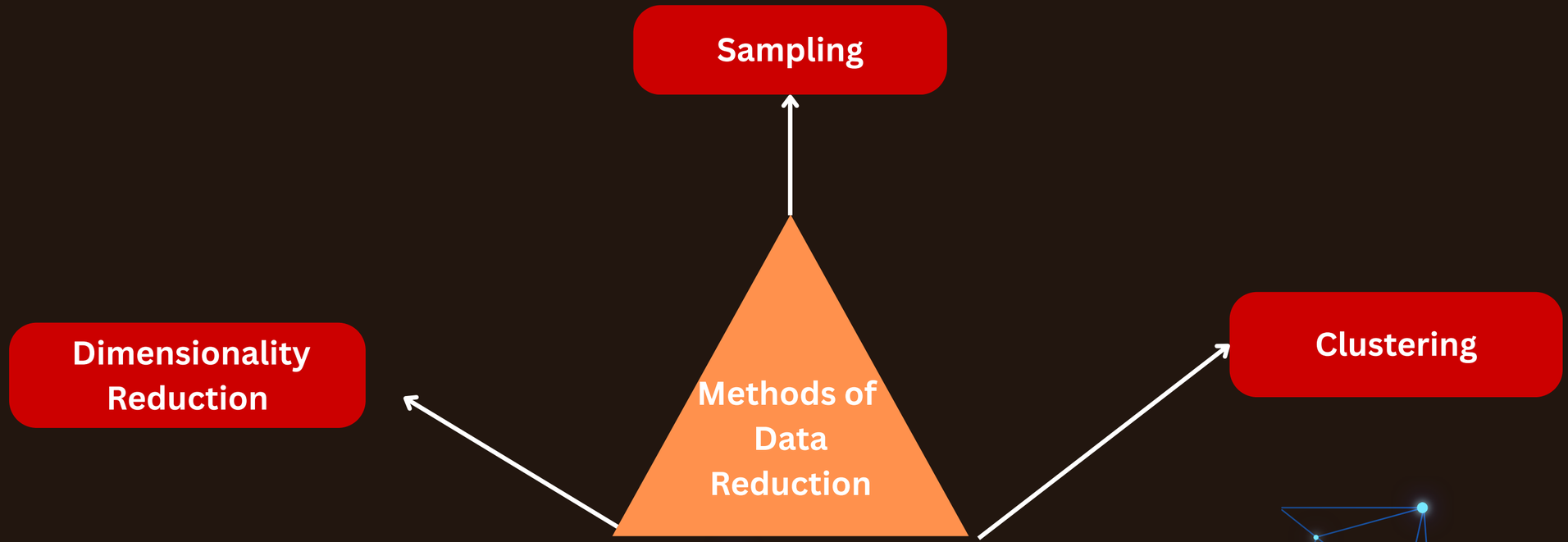
- Converts or consolidates data into a format suitable for mining or analysis

Steps involved:

- Smoothing – Removes noise using binning, regression, or clustering
- Aggregation – Applies summary operations to data
- Generalization – Replaces low-level data with higher-level concepts
- Normalization – Scales attribute data within a specified range
- Attribute Construction – Creates new attributes from existing ones for better analysis

Data Reduction

- Definition: Process of obtaining a reduced representation of a dataset
- Reduces the volume of original data while producing the same analytical results
- Ensures integrity of data even after reduction



Data Reduction

1. Dimensionality Reduction
   - Reduces the number of random variables under consideration
   - Obtains a set of principal variables
   - Helps simplify complex datasets

2. Clustering
   - Groups similar numerical attributes together
   - Uses clustering algorithms
   - Partitions data into clusters or groups for analysis

Data Reduction

3. Sampling

- Selects a smaller subset of data representing the whole dataset
- Reduces data volume while retaining accuracy
- Types of Sampling:
  - Simple Random Sampling (SRS)
  - Cluster Sampling
  - Systematic Sampling

Data Discretization

- Definition: Process of dividing the range of a continuous attribute into intervals
- Reduces the number of values for a given continuous attribute
- Helps in concise, easy-to-use, knowledge-level representation of mining results

Types of Discretization Techniques

1. Supervised Discretization
   - Uses class information while discretizing
   - Intervals are created based on class distinctions

2. Unsupervised Discretization
   - Does not use class information
   - Intervals are created based on data distribution alone

Data Discretization

Approaches in Discretization

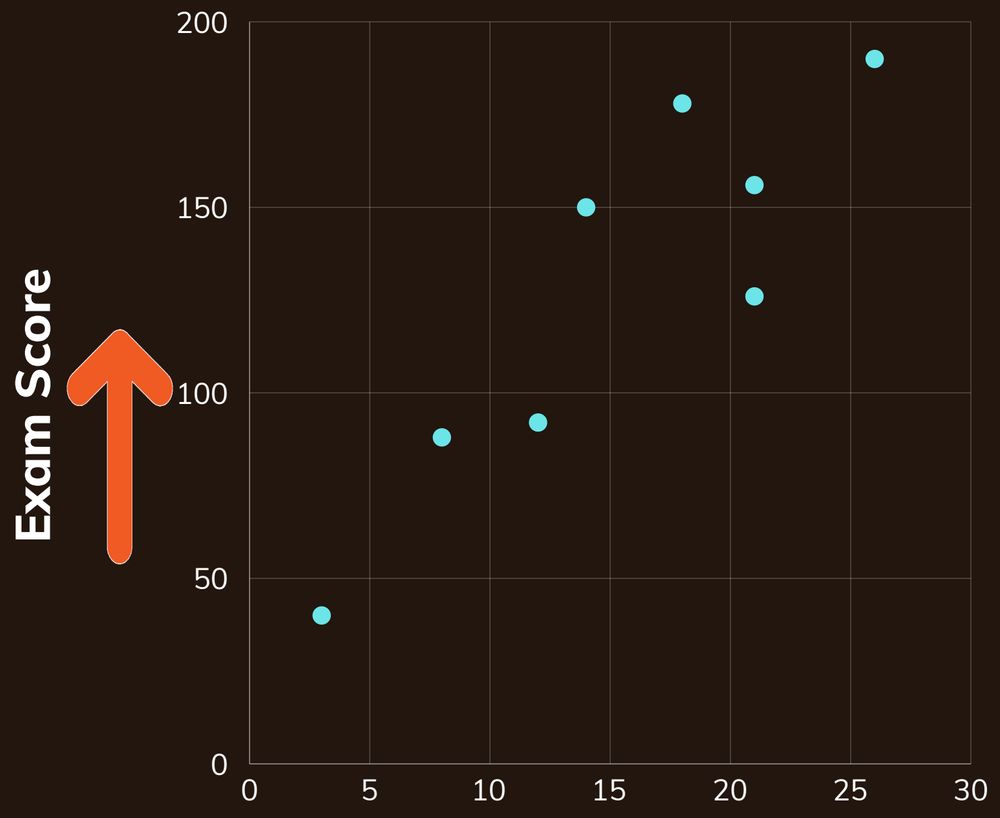
Top-Down Approach (Splitting)

- Start by finding one or two points to split the entire attribute range
- Recursively splits intervals further
- Called Splitting method

Bottom-Up Approach (Merging)

- Start by considering all continuous values separately
- Merge neighboring values to form intervals
- Recursively merges until desired intervals are formed
- Called Merging method

Regression


Exam Score
200
150
100
50
0
0
5
10
15
20
25
30
Study hours


regression line
40
30
20
10
0
18
26
23
35
36

Types of Regression

1. Linear Regression

- Simplest and most widely used regression technique
- Predicts dependent variable (Y) using independent variable(s) (X)
- Equation (Simple Linear Regression):
- $Y=mX+bY = mX + bY=mX+b$
- Example: Predicting house prices based on features like size, location, etc.

2. Polynomial Regression

- Extension of linear regression

Types of Regression

- Relationship between dependent and independent variables is modeled as nth-degree polynomial
- Example: Modeling relationship between temperature and humidity

3. Ridge Regression

- Regularization technique to prevent overfitting in linear regression models
- Adds a penalty term to the loss function
- Example: Predicting student's GPA based on study hours, attendance, etc.

Thank You