Statistics for Data Science

-MUKESH KUMAR

What is Statistics?

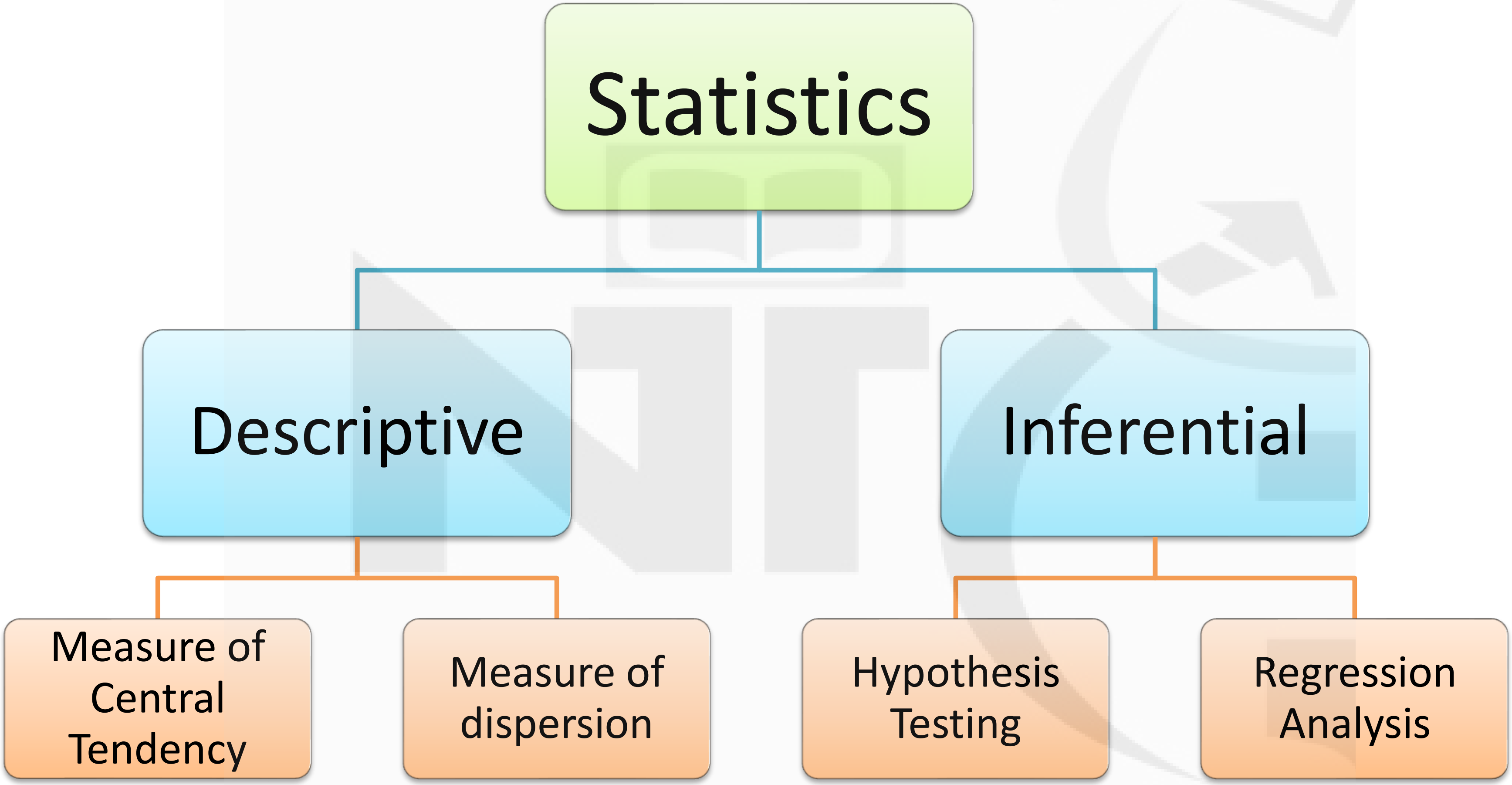
- **Statistics** (from German: *Statistik*, orig. "description of a state, a country")is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.

- In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model to be studied.

- Populations can be diverse groups of people or objects such as "all people living in a country" or "every atom composing a crystal".

- Statistics deals with every aspect of data, including the planning of data collection in terms of the design of surveys and experiments.[6]

https://en.wikipedia.org/wiki/Statistics



Statistics

Descriptive

Inferential

Measure of Central Tendency

Measure of dispersion

Hypothesis Testing

Regression Analysis

Descriptive Statistics

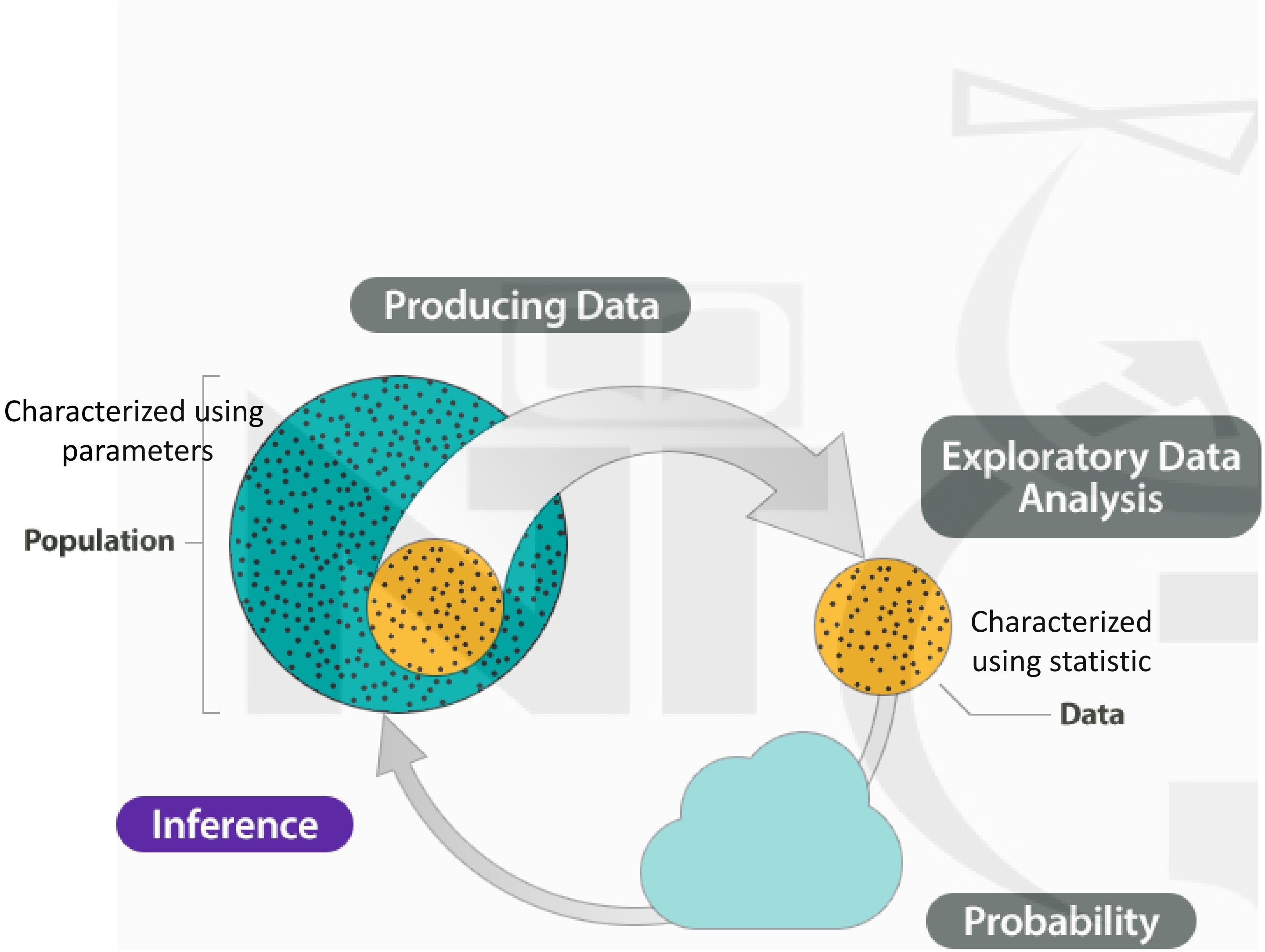
- This branch focuses on summarizing and describing the main features of a dataset.
- It includes:
  - measures of central tendency (mean, median, mode)
  - measures of dispersion (range, variance, standard deviation)
  - and visual representations like graphs and charts.

Inferential Statistics

- This branch involves using sample data to make inferences about a larger population.

- It includes techniques like
  - hypothesis testing
  - confidence intervals
  - regression analysis.

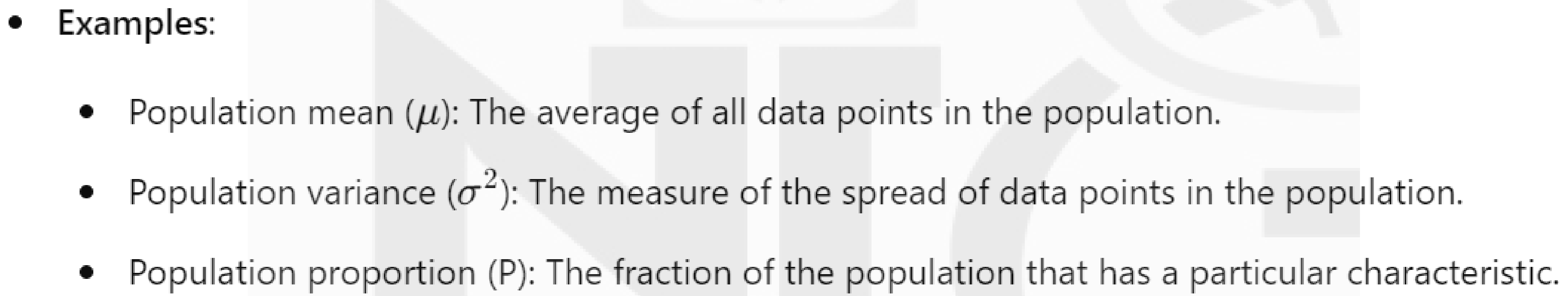
- Inferential statistics allows researchers to draw conclusions and make predictions based on the available data.


Producing Data
Characterized using parameters
Population
Exploratory Data Analysis
Characterized using statistic
Data
Inference
Probability

Prameters Vs Statistcs

- **Parameters**
  - **Definition:** Parameters are numerical characteristics that describe an entire population. A population is the complete set of items or individuals of interest in a study.
  - **Nature:** Parameters are fixed values, though in practice, they are often unknown because it is usually impractical to collect data from an entire population.

Parameters Example



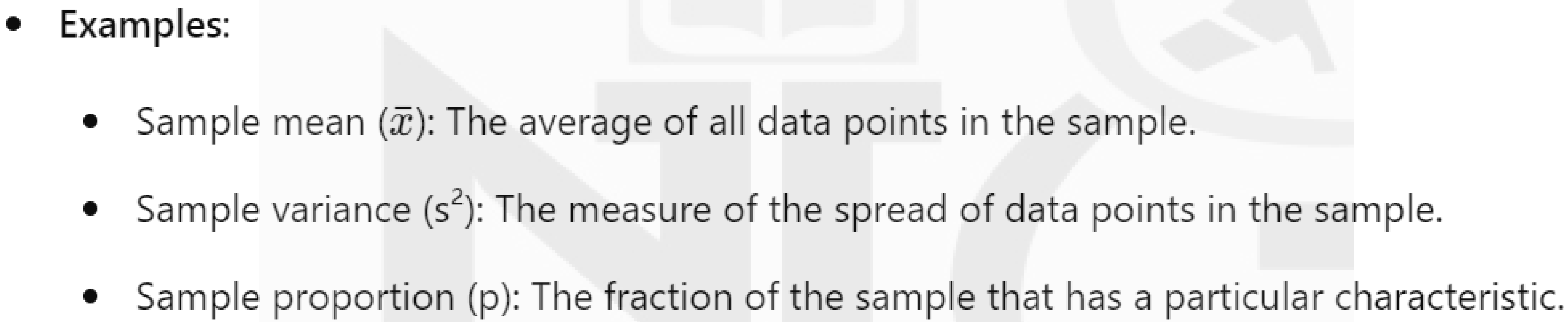
- **Examples:**
  - Population mean (μ): The average of all data points in the population.
  - Population variance (σ^2): The measure of the spread of data points in the population.
  - Population proportion (P): The fraction of the population that has a particular characteristic.

Statistics

- **Definition**: Statistics are numerical characteristics that describe a sample, which is a subset of the population. A sample is a smaller, manageable version of the population used to infer conclusions about the population.

- **Nature**: Statistics are variable because they can change depending on the sample chosen from the population.

Statistics Examples



- **Examples:**
  - Sample mean ($\bar{x}$): The average of all data points in the sample.
  - Sample variance (s^2): The measure of the spread of data points in the sample.
  - Sample proportion (p): The fraction of the sample that has a particular characteristic.

Key Differences

- **Scope:**
  - Parameters describe the entire population.
  - Statistics describe a sample taken from the population.

- **Representation:**
  - Parameters are typically denoted using Greek letters

    (e.g., μ, σ, P)
  - Statistics are usually denoted using Roman letters

    (e.g., $\bar{x}$, s, p)

- **Variability:**
  - Parameters are constant for a given population.
  - Statistics can vary from sample to sample.

- **Purpose:**
  - Parameters are the actual values we aim to understand or estimate.
  - Statistics are used to estimate parameters.

Example Scenario

- Imagine a company wants to know the average height of all employees (a population parameter).
  - Measuring the height of every employee might be impractical, so they take a sample of 100 employees
  - and calculate the average height of this sample (a sample statistic).
  - The sample mean (x¯) is used to estimate the population mean (μ).
- In summary,
  - while parameters pertain to entire populations and are often unknown constants, statistics are derived from samples and used to estimate these unknown parameters.
  - Understanding the distinction between these two concepts is crucial for accurate data analysis and making reliable inferences about populations from samples.

Applications

- Statistics is widely used in various fields, such as:
  - Business: For market analysis, forecasting, and decision-making.
  - Science: To design experiments, test hypotheses, and draw conclusions.
  - Social Sciences: For survey analysis, demographic studies, and policy evaluation.
  - Medicine: For clinical trials, epidemiological studies, and drug development.

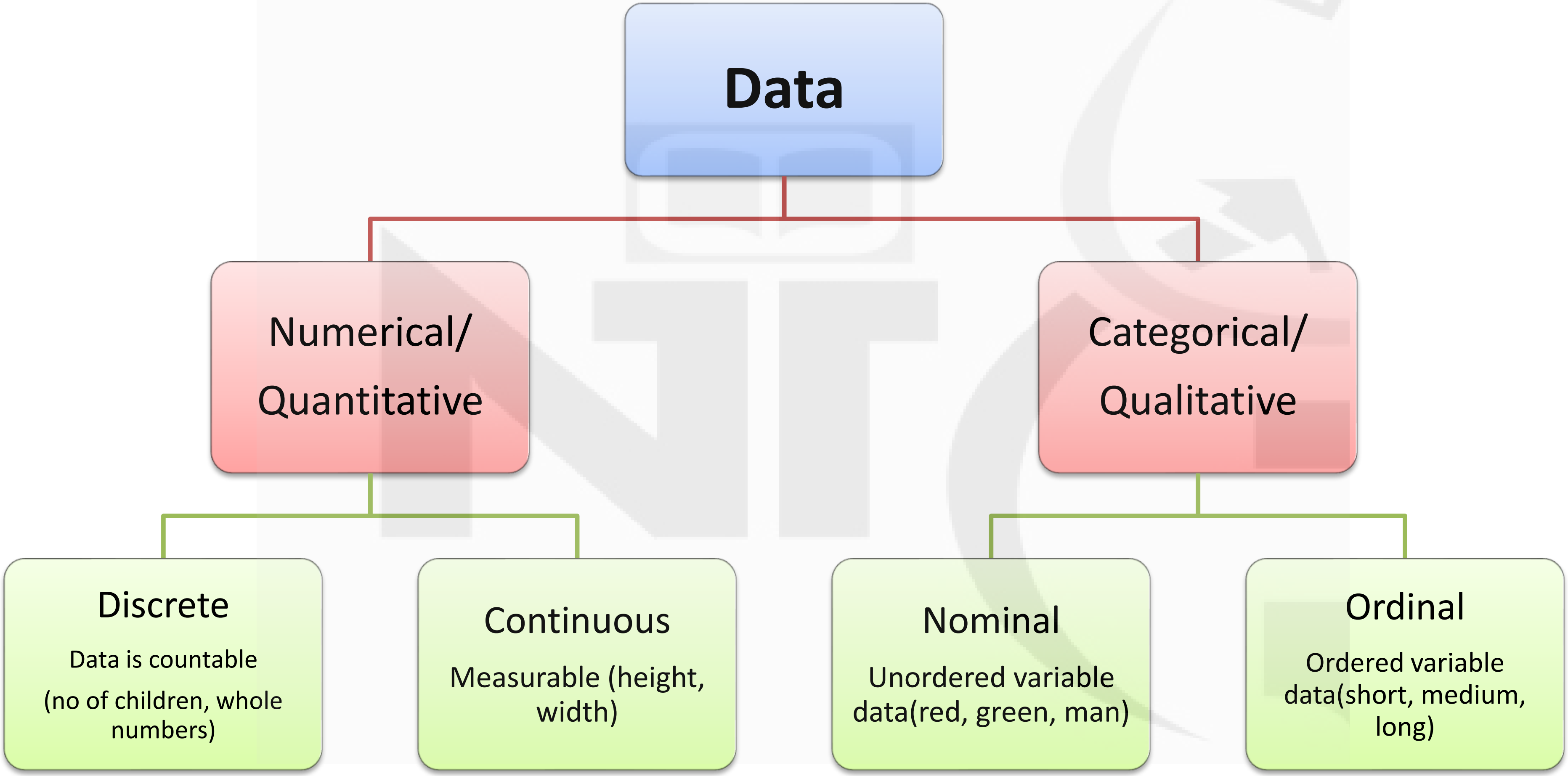
TYPES OF DATA

Introduction to Data Types

- Having a good understanding of the different data types, also called measurement scales, is a crucial prerequisite for doing Exploratory Data Analysis (EDA), since you can use certain statistical measurements only for specific data types.

- You also need to know which data type you are dealing with to choose the right visualization method.

- Think of data types as a way to categorize different types of variables. We will sometimes refer to them as measurement scales.



Data

Numerical/ Quantitative

Categorical/ Qualitative

Discrete

Data is countable

(no of children, whole numbers)

Continuous

Measurable (height, width)

Nominal

Unordered variable data(red, green, man)

Ordinal

Ordered variable data(short, medium, long)

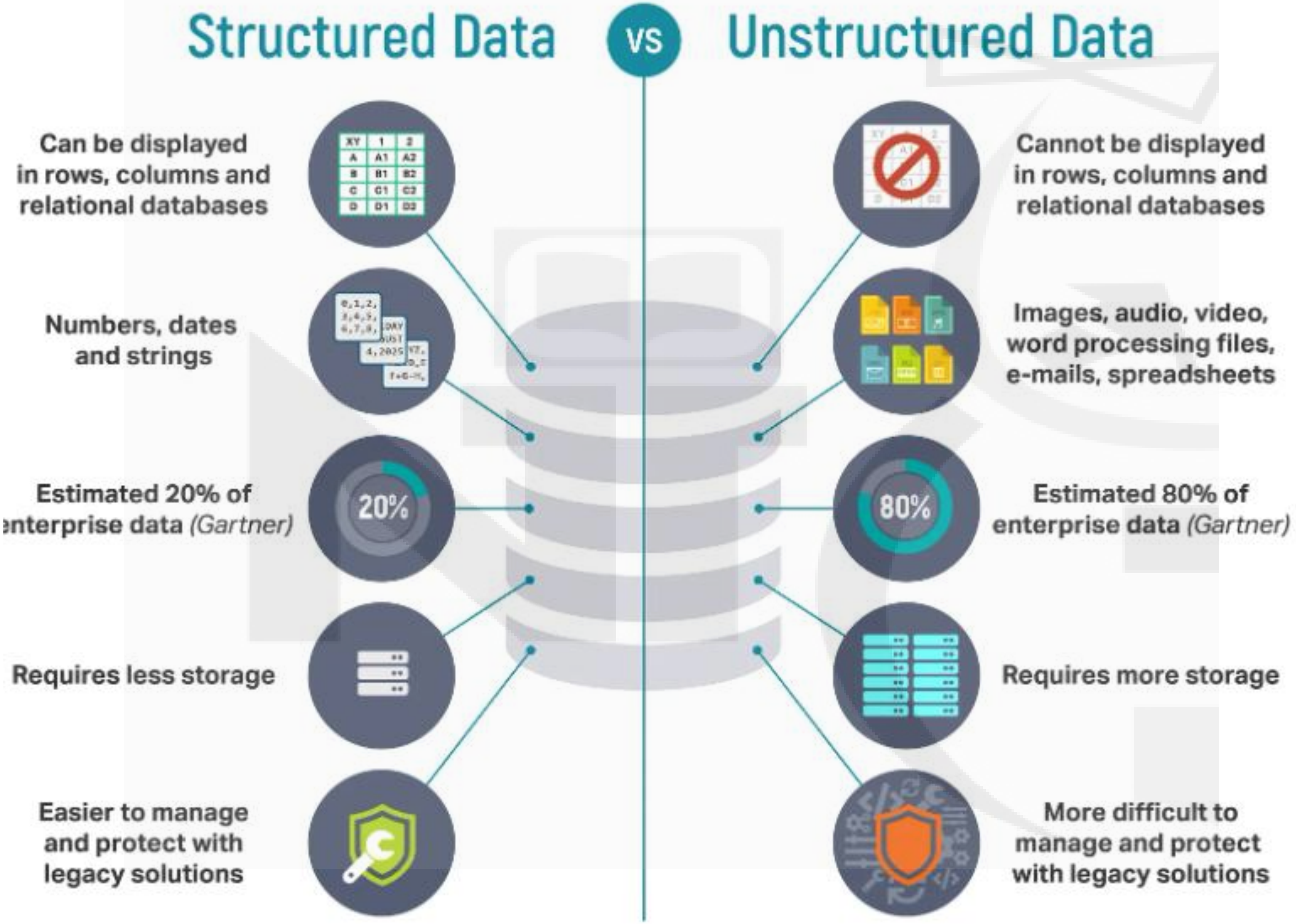
Examples

- **Quantitative Data**
  - **Discrete Data**:
    - Number of students in a classroom.
    - Number of cars in a parking lot.
    - Number of books on a shelf.
    - Number of goals scored in a soccer match.
  - **Continuous Data**:
    - Height of individuals.
    - Weight of fruits.
    - Temperature over a day.
    - Time taken to run a marathon.

Examples

- **Qualitative Data**
  - **Nominal Data**:
    - Types of fruits (apple, banana, cherry).
    - Colors of cars (red, blue, green).
    - Types of animals (mammals, birds, reptiles).
    - Brands of shoes (Nike, Adidas, Puma).
  - **Ordinal Data**:
    - Education levels (high school, bachelor's, master's, PhD).
    - Satisfaction ratings (very dissatisfied, dissatisfied, neutral, satisfied, very satisfied).
    - Movie ratings (one star, two stars, three stars, four stars, five stars).
    - Class ranks (first, second, third).

Structured Data VS Unstructured Data



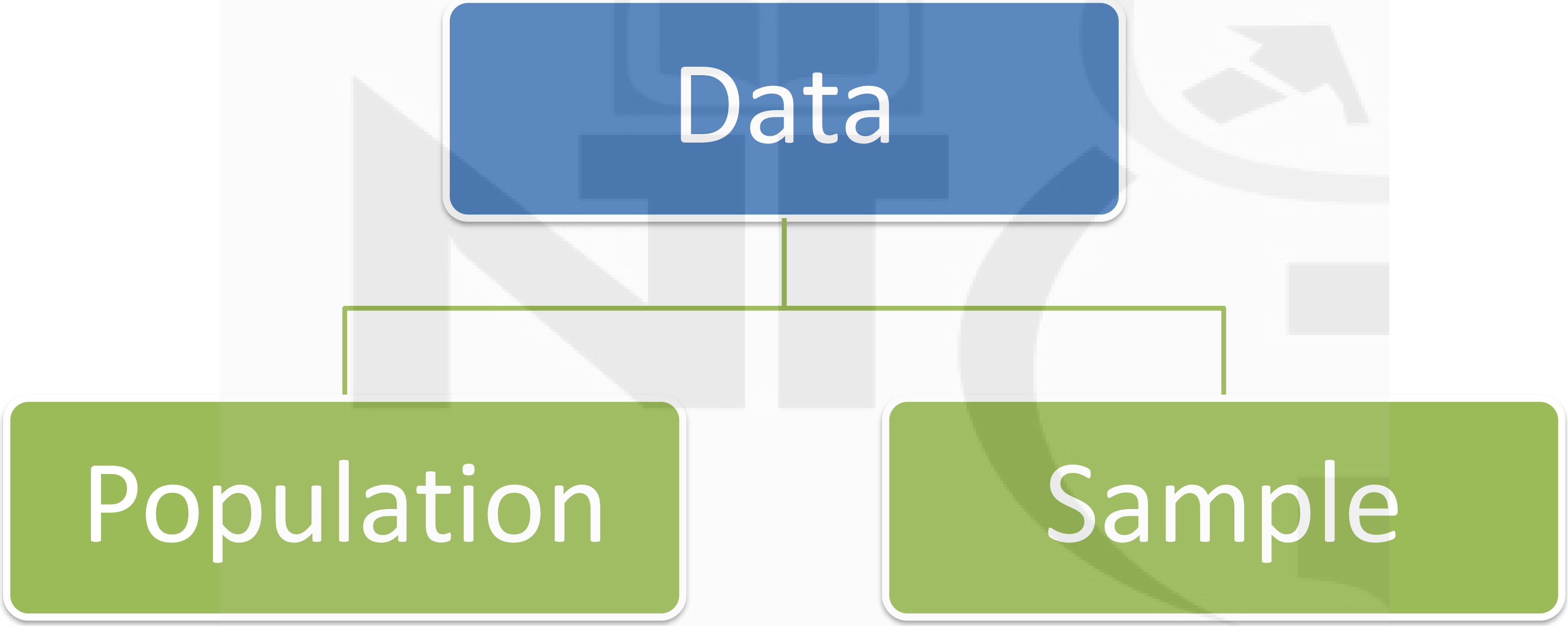
Structured Data

- Can be displayed in rows, columns and relational databases
- Numbers, dates and strings
- Estimated 20% of enterprise data *(Gartner)*
- Requires less storage
- Easier to manage and protect with legacy solutions

Unstructured Data

- Cannot be displayed in rows, columns and relational databases
- Images, audio, video, word processing files, e-mails, spreadsheets
- Estimated 80% of enterprise data *(Gartner)*
- Requires more storage
- More difficult to manage and protect with legacy solutions

- https://www.youtube.com/watch?v=u_7G8Xy61zs&t=11s



Data

Population

Sample

- **Population Data:** Population data is the collection of all items of interest which is denoted by ‘N’ and the numbers we obtained when using population are called parameters.

- **Sample Data:** Sample data is a subset of the population which is denoted by ‘n’ and the numbers we obtained when using sample are called statistics.