Algorithms for Data Science

Data Types and Representations

Introduction to Data Types

- **Data** is information in a raw or unorganized form that can be processed to obtain meaning.
- It is crucial to understand the meaning/usefulness of various **Data Types**:
  - Optimal data type selection is crucial for efficient storage and manipulation.
  - Data types impact analysis, computations, and the accuracy of results.
  - Example:

```
print("The sum of two strings 'a' and 'b': ", "a" + "b")
print("The sum of two integers 1 and 2: ", 1 + 2)
```

✓ 0.0s

```
The sum of two strings 'a' and 'b':  ab
The sum of two integers 1 and 2:  3
```

What are Data Types?

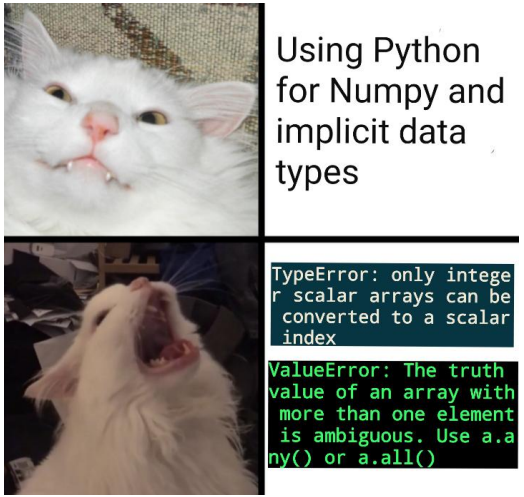
Data Types categorize the kind of value that can be stored and manipulated in a programming language or system.

Examples of Data Types:

- Integer: Whole numbers without decimals (e.g. 5, 10, -1)
- Float: Numbers with decimals (e.g. 3.14, -0.1)
- String: Sequence of characters (e.g. "Hello", "World")
- Boolean: Represents truth values (e.g. True, False)

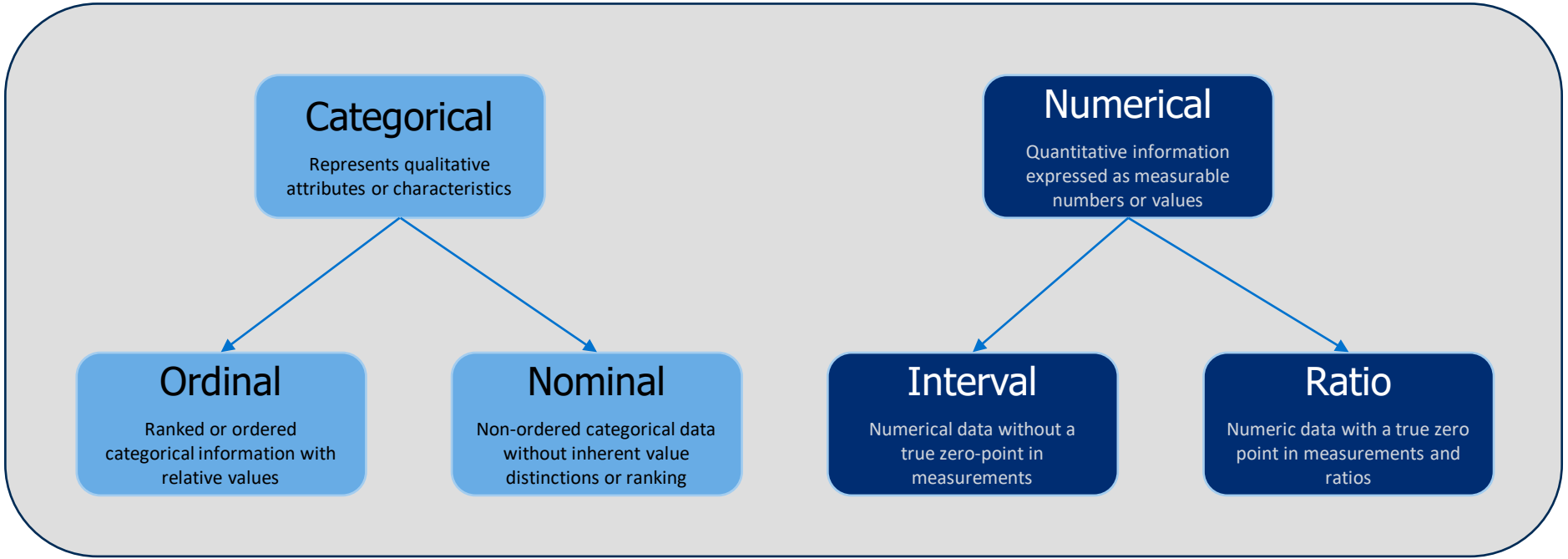
Purpose:

- To define the type of data a variable can hold
- To ensure proper storage allocation and manipulation



Importance: Determines how data is interpreted and operated upon by the system

Data Categories



Data Categories Examples

Ordinal

- Education Levels (high school, undergraduate, graduate, doctoral)
- Satisfaction Levels (unsatisfied, neutral, satisfied)

Interval

- Temperature (in degrees Celsius or Fahrenheit)
- Calendar Years (2000, 2010, 2020)

Nominal

- Fruit Types (apple, banana, orange)
- Colors (red, blue, green)

Ratio

- Height/Weight (in cm or in, lbs or kgs)
- Money in various currencies ($20, €50, £75.50

Structured vs. Unstructured Data

Structured

- Highly organized and easily searchable, often stored in relational databases
- Follows a predefined schema with clear data types and relationships
- E.g. An SQL database containing employee records with fields for ID, name, department, and salary

Unstructured

- Lacks a predefined format, making it more difficult to analyze and search
- Includes a wide variety of data types, often text-heavy or multimedia
- Emails, social media posts, multimedia files with no specific structure

Classes of Data

Tabular Data

- Structured data organized in rows and columns
- CSV, Excel, TSV

Image Data

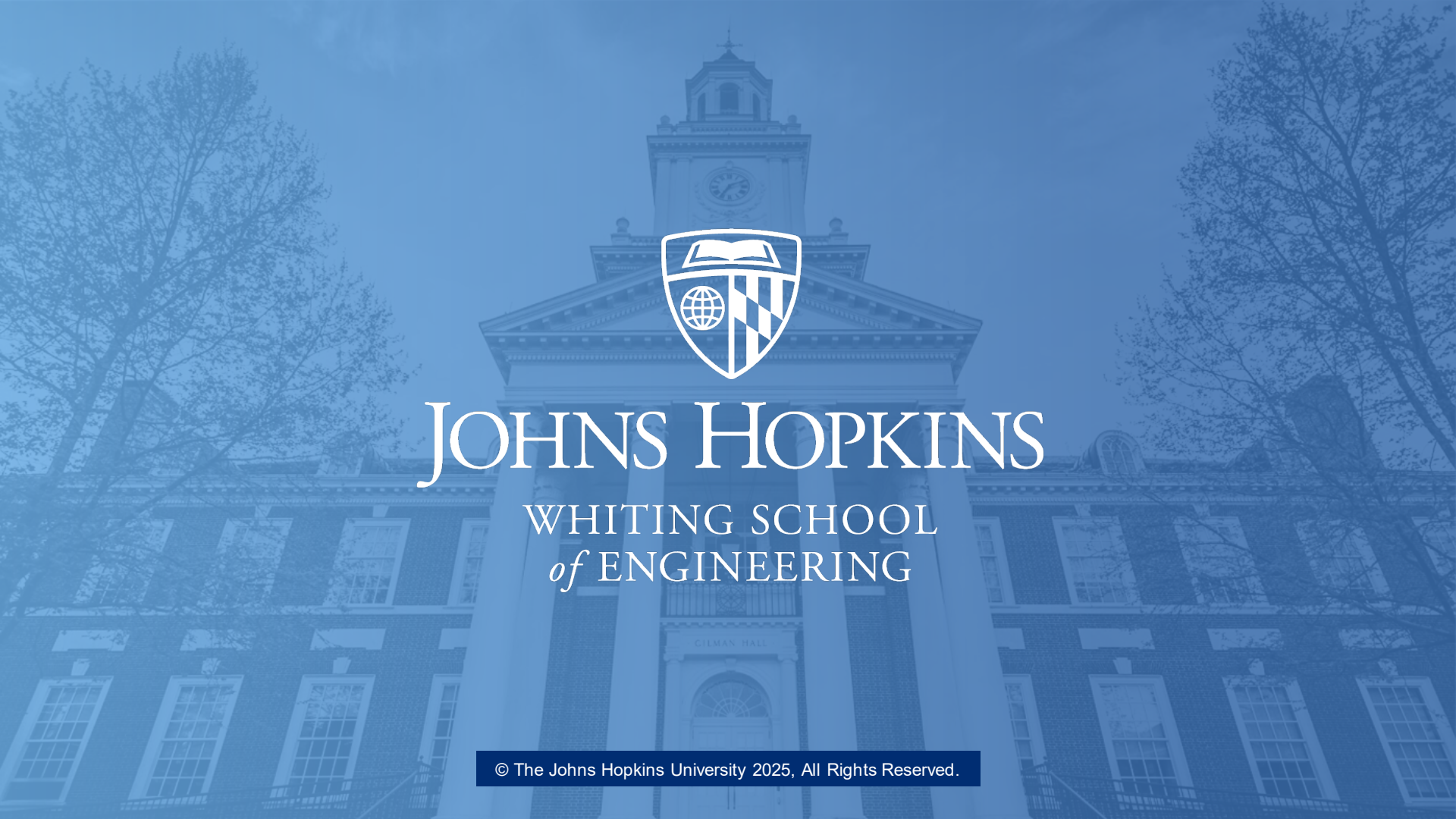
- Visual data captured as pixel arrays with intensities
- JPEG, PNG, Xray

Video Data

- Sequence of images displayed in succession
- MP4

Natural Language

- Data in the form of human language, written and spoken
- Text, Audio



© The Johns Hopkins University 2025, All Rights Reserved.