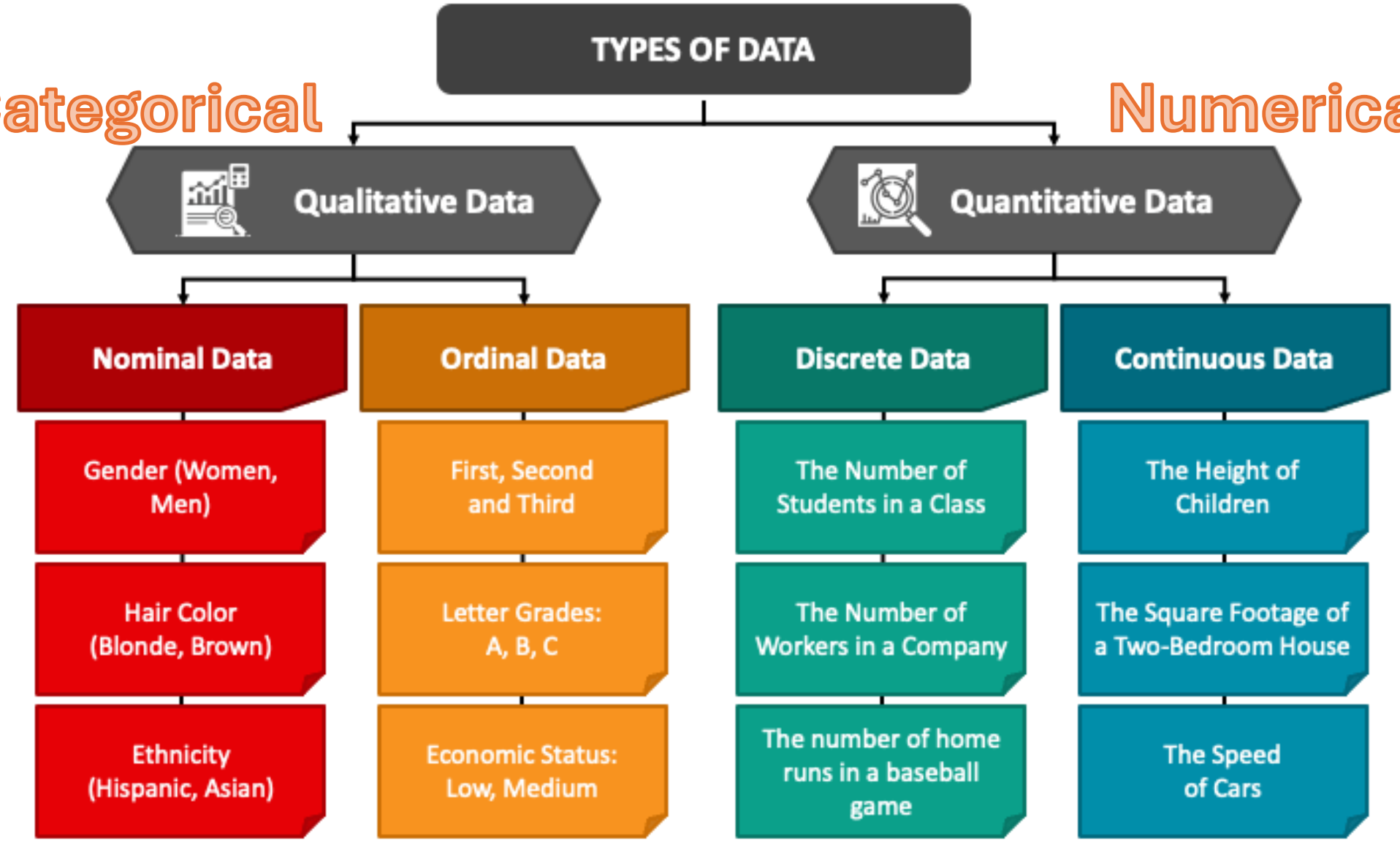
Data and Data Sources

Categorical

Numerical



Categorical Data

- The objects being studied are grouped into categories based on some **qualitative** trait.
- The resulting data are merely labels or categories.
- E.g. Hair color (blonde, brown, red, black, etc.), Opinion of students about riots (ticked off, neutral, happy), Smoking status (smoker, non-smoker), etc.
- **Nominal:** A type of categorical data in which objects fall into **unordered** categories.
  - E.g. Hair color (blonde, brown, red, black, etc.), Smoking status (smoker, non-smoker), etc.
- **Ordinal:** A type of categorical data in which **order** is important.
  - E.g. Class (fresh, sophomore, junior, senior, super senior), Opinion of students about riots (ticked off, neutral, happy), etc.
- **Binary:** A type of categorical data in which there are only two categories. Can be nominal or ordinal.
  - E.g. Attendance (present, absent), Smoking status (smoker, non-smoker), etc.

Numerical

- The objects being studied are “measured” based on some **quantitative** trait.
- The resulting data are set of numbers.
- E.g., Cholesterol level, Height, Age, SAT score, Number of students late for class, Time to complete a homework assignment, etc.
- **Discrete:** Only certain values are possible (there are gaps between the possible values).
  - E.g., SAT scores, Number of students late for class, etc.
- **Continuous:** Theoretically, any value within an interval is possible with a fine enough measuring device.
  - E.g., Cholesterol level, Height, Time to complete a homework assignment, etc.

Types of Data

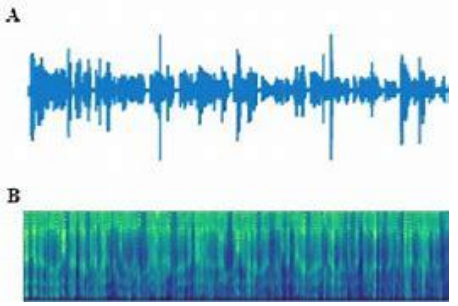
- The type(s) of data collected in a study determine the type of statistical analysis used. For instance:
- Categorical data are commonly summarized using "**percentages**" (or "**proportions**").
  - 11% of students have a tattoo
  - 2%, 33%, 39%, and 26% of the students in class are, respectively, freshmen, sophomores, juniors, and seniors
- Numerical data are typically summarized using "**averages**" (or "**means**").
  - Average number of siblings Fall 1998 Stat 250 students have is 1.9.
  - Average weight of male Fall 1998 Stat 250 students is 173 pounds.
  - Average weight of female Fall 1998 Stat 250 students is 138 pounds.

Types of Data

Based on Structure

- **Structured Data** – Organized, stored in relational databases (e.g., SQL tables).
  - Example: Customer records, transaction logs.
- **Unstructured Data** – No predefined format, difficult to analyze (e.g., text, images, videos).
  - Example: Social media posts, raw audio files.
- **Semi-Structured Data** – Some structure but not strictly formatted (e.g., JSON, XML).
  - Example: Emails, NoSQL databases.

	A	B	C	D	E
1	Cookie Sales by Region				
2	SalesRep	Region	# Orders	Total Sales	
3	Bill	West	217	$41,107	
4	Frank	West	268	$72,707	
5	Harry	North	224	$41,676	
6	Janet	North	286	$87,858	
7	Joe	South	226	$45,606	
8	Martha	East	228	$49,017	
9	Mary	West	234	$57,967	
10	Ralph	East	267	$70,702	
11	Sam	East	279	$77,738	
12	Tom	South	261	$69,496	
13					
14					
15					




```
{"widget": {
    "debug": "on",
    "window": {
        "title": "Sample Konfabulator Widget",
        "name": "main_window",
        "width": 500,
        "height": 500
    },
    "image": {
        "src": "Images/Sun.png",
        "name": "sun1",
        "hOffset": 250,
        "vOffset": 250,
        "alignment": "center"
    },
    "text": {
        "data": "Click Here",
        "size": 36,
        "style": "bold",
        "name": "text1",
        "hOffset": 250,
        "vOffset": 100,
        "alignment": "center",
        "onMouseUp": "sun1.opacity = (sun1.opacity / 100) * 90;"
    }
}}
```

Types of Data

Text Data

- **Definition:** Data in the form of natural language text.
- **Examples:** Social media posts, customer reviews, research papers.
- **Use Cases:** Sentiment analysis, language translation, text classification.

Image Data

- **Definition:** Data in the form of images.
- **Examples:** Photographs, medical scans, satellite images.
- **Use Cases:** Image classification, object detection, image segmentation.

Audio Data

- **Definition:** Data in the form of sound recordings.
- **Examples:** Speech recordings, music, environmental sounds.
- **Use Cases:** Speech recognition, audio classification, music generation.

Video Data

- **Definition:** Data in the form of moving images.
- **Examples:** Surveillance footage, video clips, movies.
- **Use Cases:** Action recognition, video summarization, video segmentation.

Types of Datasets

Datasets can be categorized based on their characteristics, structure, and the type of problem they are used to solve. Different types of datasets serve various purposes and are used in different applications. Understanding the nature of the data and choosing the right type of dataset for a specific problem is crucial for developing effective and accurate prediction models.
Here are the primary types of datasets:

1. Structured vs. Unstructured Dataset

Structured Data

•**Definition:** Data that is organized in a predefined manner, often in tabular format with rows and columns.
•**Examples:** Spreadsheets, SQL databases.
•**Use Cases:** Financial records, customer databases, sensor data.

Unstructured Data

•**Definition:** Data that does not have a predefined format or organization.
•**Examples:** Text documents, images, audio files, videos.
•**Use Cases:** Natural language processing (NLP), image recognition, speech-to-text conversion.

Types of Datasets

2. Labeled vs. Unlabeled Datasets

Labeled Data

- **Definition:** Data that has been tagged with one or more labels, providing explicit information about the target variable.
- **Examples:** Annotated images (with objects labeled), spam vs. non-spam emails.
- **Use Cases:** Supervised learning tasks such as classification and regression.

Unlabeled Data

- **Definition:** Data without any labels or target variables.
- **Examples:** Raw text, unlabeled images, customer behavior data.
- **Use Cases:** Unsupervised learning tasks such as clustering, anomaly detection.

3. Time Series Datasets

- **Definition:** Data points collected or recorded at specific time intervals.
- **Examples:** Stock prices, weather data, sensor readings.
- **Use Cases:** Forecasting, anomaly detection, trend analysis.

Types of Datasets

4. Training, Validation, and Test Sets

Training Set

- **Definition:** The portion of the dataset used to train the machine learning model.
- **Purpose:** To allow the model to learn patterns and relationships in the data.

Validation Set

- **Definition:** A subset of the dataset used to tune model parameters and make decisions about model architecture.
- **Purpose:** To provide an unbiased evaluation of a model fit on the training dataset while tuning hyperparameters.

Test Set

- **Definition:** The portion of the dataset used to evaluate the final model performance.
- **Purpose:** To provide an unbiased assessment of the model's performance on unseen data.

Popular sources for datasets

- https://www.kaggle.com/
- Hugging Face – The AI community building the future.
- Home - UCI Machine Learning Repository
- Dataset Storage and Dataset Search Platform | IEEE DataPort

Data Quality and Issues

Data quality is crucial for accurate analysis and decision-making. . Poor data quality leads to incorrect insights and business risks.

Key Dimensions of Data Quality

Accuracy – Data should be correct and free from errors.

Completeness – No missing values or gaps in data.

Consistency – Uniform format and values across different sources.

Timeliness – Data should be up to date and relevant.

Validity – Data should conform to predefined formats and rules.

Uniqueness – No duplicate or redundant records.

Data Quality and Issues

Common Data Issues

- **Missing Data** – Incomplete records causing bias in analysis.
- **Duplicate Data** – Multiple records for the same entity.
- **Inconsistencies** – Data mismatch across different systems.
- **Incorrect Data** – Human or system errors in data entry.
- **Data Drift** – Changes in data patterns over time affecting model accuracy.
- **Bias in Data** – Unrepresentative data leading to skewed results.
- **Scalability:** Handling large volumes of data requires scalable storage and processing solutions.

Data Quality and Issues

Improving Data Quality

- **Data Cleaning** – Removing errors, duplicates, and inconsistencies.
- **Data Standardization** – Ensuring a uniform format across datasets.
- **Validation Techniques** – Implementing rule-based validation and automated error detection.
- **Data Governance** – Setting policies for data integrity, security, and compliance.
- **Continuous Monitoring** – Regular data audits and updates to maintain quality.

Association Analysis and Prediction Analysis

Feature	Association Analysis	Prediction Analysis
Goal	To discover relationships and patterns between variables.	To build models that forecast future outcomes or unknown values.
Focus	Identifying "what goes with what."	Predicting "what will happen."
Purpose	Understanding the relationships and dependencies within data.	Maximizing the accuracy of predictions on new data.
Output	Rules, correlations, or patterns describing relationships.	A predictive model that produces forecasts.
Evaluation	Statistical significance of relationships (e.g., support, confidence, lift).	Accuracy metrics (e.g., precision, recall, accuracy, RMSE).
Interpretability	Often high; relationships are typically easier to understand.	Varies; complex models may be "black boxes" with low interpretability.
Example	Market basket analysis (finding which items are frequently bought together).	Predicting customer churn or forecasting sales.
Key Question	"What variables are related?"	"What outcome is most likely?"
Typical Methods	Association rule mining, correlation analysis.	Regression analysis, classification algorithms, time series analysis.