Data Exploration & Visualization
Unit-I

Dr. SELVA KUMAR S
ASSISTANT PROFESSOR
DEPT. OF CSE

Prescribed Text Book					
Sl. No.	Book Title	Authors	Edition	Publisher	Year
1	Hands-On Exploratory Data Analysis with Python	Suresh Kumar Mukhiya, Usman Ahmed	1st Edition	Packt	2020
2	Fundamental of Data Visualization	Claus O. Wilke	1st Edition	O'Reilly	2019

At the end of the course the student will be able to

CO1	**Apply the computational approaches to perform Data Exploration and Visualization.**
CO2	Analyse the different techniques to perform Data Exploration and Visualization for a given application.
CO3	Demonstrate exploratory data analysis to real data sets and provide interpretations through relevant visualization tools.

Instructions

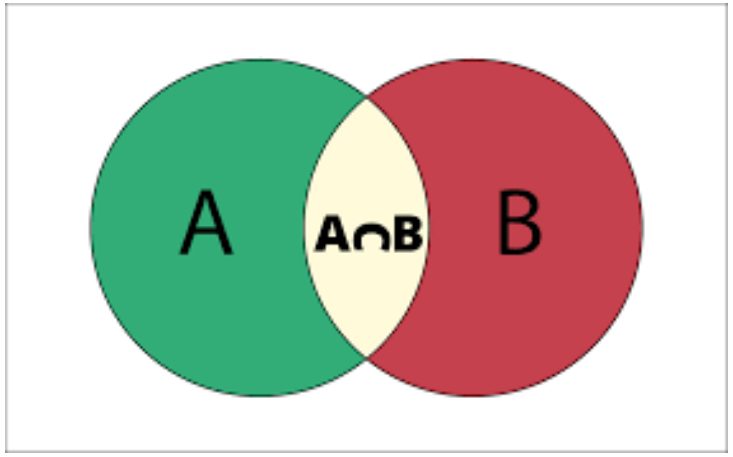
Google Classroom code: mnjiy3o

Sl. No	Week	Activity
1	1st	Formation of groups. Note: Student groups of size 2 to 4
2	2nd and 3rd	Project topic selection by each group
3	4th	Presentation-1: Student and Project topic introduction by each group
4	5th	Data Acquisition and Data Preparation
5	6th and 7th	Presentation-2: Exploratory tools demonstration
6	8th and 9th	Presentation-3: Techniques applied on EDA
7	10th	Presentation-4: Visualization tools demonstration
8	11th	Complete Project Work Demonstration by each group
9	12th	Project Report Submission

Guess the topic?

Guess topic?

Guess the topic?



Guess the topic?

Guess the topic?

Guess the topic?

Guess the topic?

Agenda

Exploratory Data Analysis

- **Introduction**
- **Steps in EDA**
- **Data Types**
  - Numerical Data
  - Categorical Data
- **Measurement Scales**
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- **Comparing EDA with classical & Bayesian Analysis**
- **Software tools for EDA**

WHAT IS EDA?

- The analysis of datasets based on various numerical methods and graphical tools.
- Exploring data for patterns, trends, underlying structure, deviations from the trend, anomalies and strange structures.
- It facilitates discovering unexpected as well as conforming the expected.
- Another definition: An approach/philosophy for data analysis that employs a variety of techniques (mostly graphical).

WHAT IS EDA?

Primary Aim is to performing quantitative and qualitative evaluation of the data to draw meaningful insights from it.

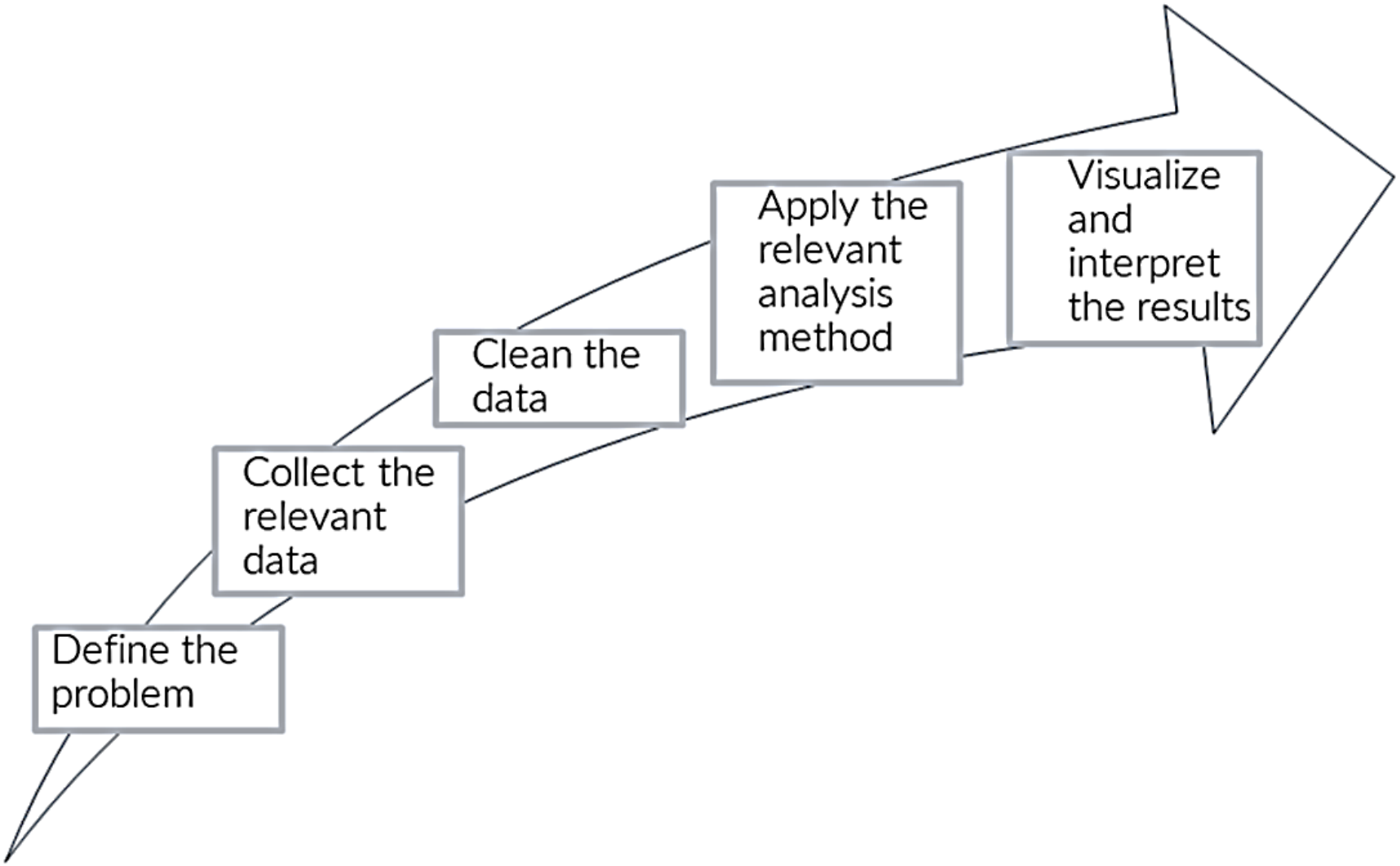
AIM OF THE EDA

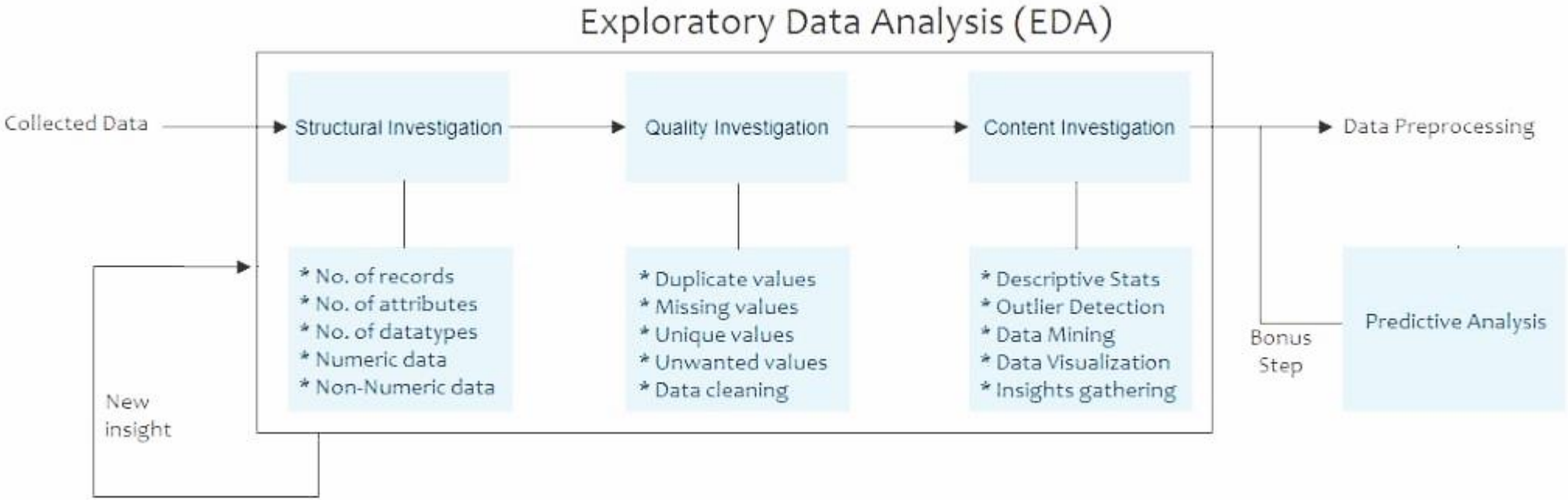
- Maximize insight into a dataset
- Uncover underlying structure
- Extract important variables
- Detect outliers and anomalies
- Test underlying assumptions
- Develop valid models
- Determine optimal factor settings (Xs)

Exploratory vs Confirmatory Data Analysis

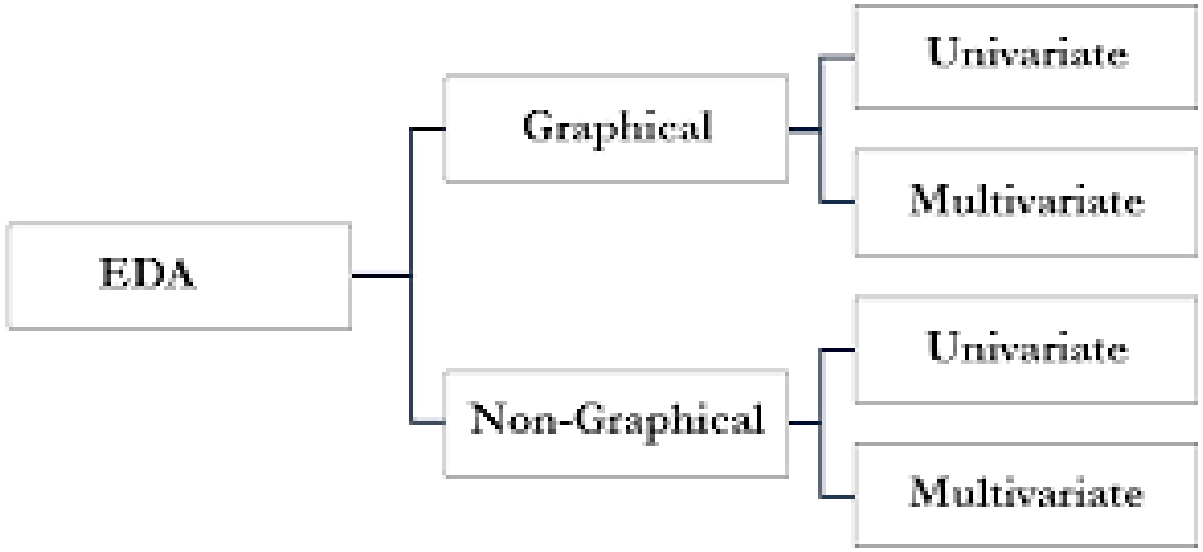
EDA	CDA
• No hypothesis at first • Generate hypothesis • Uses graphical methods (mostly)	• Start with hypothesis • Test the null hypothesis • Uses statistical models

STEPS OF EDA


Define the problem
Collect the relevant data
Clean the data
Apply the relevant analysis method
Visualize and interpret the results


Exploratory Data Analysis (EDA)
Collected Data
Structural Investigation
Quality Investigation
Content Investigation
Data Preprocessing
* No. of records
* No. of attributes
* No. of datatypes
* Numeric data
* Non-Numeric data
* Duplicate values
* Missing values
* Unique values
* Unwanted values
* Data cleaning
* Descriptive Stats
* Outlier Detection
* Data Mining
* Data Visualization
* Insights gathering
Bonus Step
Predictive Analysis
New insight

Classification of EDA



EXAMPLE 1 – Professional Sports

Guess??? how EDA can be used in Professional Sports

EXAMPLE 1 - Professional Sports

1. **Player Performance Analysis**:
    1. EDA can be used to analyze player statistics, such as scoring, rebounds, assists, steals, and turnovers, to identify strengths and weaknesses.
    2. Visualizations like heat maps and shot charts can help understand a player's shooting patterns and effectiveness from different areas of the court.
2. **Team Performance Analysis**:
    1. EDA can provide insights into team performance by examining key metrics like points scored, turnovers, field goal percentages, and more.
    2. Time-series analysis can help identify trends in team performance throughout a season or across multiple seasons.
3. **Opponent Analysis**:
    1. EDA can be used to study opponent teams' historical data, helping teams prepare game strategies.
    2. It can reveal opponent strengths and weaknesses, player tendencies, and optimal defensive strategies.

EXAMPLE 1 – Professional Sports

4. **Player Health and Injury Prevention**:
    1. By analyzing player health data (e.g., player load, heart rate, and injury history), teams can identify patterns that may lead to injuries.
    2. EDA can help in developing training and recovery strategies to reduce injury risks.
5. **Player Recruitment and Drafting**:
    1. EDA can assist in scouting potential new players for recruitment or drafting by comparing their performance statistics to existing team needs.
    2. It can help identify undervalued or underappreciated talent.
6. **Game Strategy Optimization**:
    1. By analyzing historical game data, teams can identify effective offensive and defensive strategies against different opponents.
    2. EDA can reveal optimal lineup combinations, substitutions, and in-game decision-making.
7. **Fan Engagement and Marketing**:
    1. EDA can help sports organizations understand fan demographics, preferences, and behaviors.
    2. It can be used to tailor marketing strategies, ticket pricing, and fan engagement activities.

EXAMPLE 1 – Professional Sports

8. **In-Game Analytics**:
    1. Real-time EDA during games can provide coaches and analysts with immediate insights.
    2. Data on player performance, shot selection, and opponent behavior can be analyzed to make in-game adjustments.
9. **Performance Tracking Wearables**:
    1. Many athletes wear devices that track their movements and physiological data. EDA can be used to interpret this data and make real-time performance assessments.
10. **Revenue Optimization**:
    1. EDA can be used to analyze revenue sources, such as ticket sales, merchandise, and broadcasting deals, to optimize revenue streams.
11. **Player Development**:
    1. Coaches and trainers can use EDA to track player development over time and make adjustments to training and practice routines.
12. **Fantasy Sports and Betting**:
    1. EDA is also used by analysts, enthusiasts, and sports gamblers to gain insights for fantasy sports and betting purposes.

EXAMPLE 2 - Healthcare

Guess?? How EDA can be used on Healthcare domain

EXAMPLE 2 - Healthcare

1. **Patient Data Analysis**:
    1. EDA can be used to analyze patient medical records and electronic health records (EHRs) to identify trends and patterns in patient health.
    2. It can help identify risk factors, comorbidities, and treatment effectiveness.
2. **Disease Surveillance**:
    1. EDA is valuable for monitoring the spread of diseases and outbreaks by analyzing data such as the number of cases, geographic distribution, and demographic information.
    2. It aids in early detection and timely response to public health emergencies.
3. **Clinical Trials**:
    1. EDA can be used to examine clinical trial data to assess the safety and efficacy of new treatments.
    2. It helps identify patient subgroups that may benefit more from specific treatments.
4. **Drug Safety Analysis**:
    1. EDA can be applied to pharmacovigilance data to detect adverse drug reactions and ensure medication safety.
    2. It helps in making informed decisions about drug approvals and withdrawals.

EXAMPLE 2 - Healthcare

5. **Quality of Care Assessment**:
    1. Healthcare facilities can use EDA to assess the quality of care by analyzing patient outcomes, readmission rates, and adherence to clinical guidelines.
    2. It aids in identifying areas for improvement.
6. **Resource Allocation**:
    1. EDA can help healthcare organizations optimize resource allocation, such as the allocation of healthcare staff, medical equipment, and hospital beds.
    2. It ensures efficient use of resources and cost savings.
7. **Patient Flow and Wait Times**:
    1. EDA can be used to analyze patient flow within hospitals and clinics, helping to reduce wait times and improve patient satisfaction.
8. **Predictive Modeling**:
    1. EDA is often a precursor to building predictive models, which can forecast disease trends, patient readmissions, and resource demands.
9. **Chronic Disease Management**:
    1. EDA helps identify at-risk patient populations for chronic diseases and develop personalized care plans.
    2. It enables the early detection of complications and proactive intervention.

EXAMPLE 2 - Healthcare

10. **Telemedicine and Remote Monitoring**:
    1. EDA can be applied to data collected from remote patient monitoring devices, supporting telemedicine and remote care initiatives.
11. **Patient Engagement**:
    1. By analyzing patient feedback and satisfaction surveys, healthcare organizations can improve patient engagement and the overall healthcare experience.
12. **Public Health Policy**:
    1. EDA assists policymakers in making data-informed decisions on issues like vaccination campaigns, public health initiatives, and health regulations.
13. **Health Insurance**:
    1. Health insurance providers use EDA to assess risks, set premiums, and design healthcare plans.

EXAMPLE 3 - Marketing

Any Guess??

EXAMPLE 3 - Marketing

1. **Customer Segmentation**:
    1. EDA can identify customer segments based on demographics, behavior, and purchase history.
    2. It helps in tailoring marketing strategies to specific customer groups.
2. **Product Analysis**:
    1. EDA can help analyze product performance, identifying top-selling products, underperforming items, and opportunities for product development.
3. **Pricing Strategies**:
    1. Analyzing price elasticity and consumer demand through EDA can help optimize pricing strategies.
4. **Market Basket Analysis**:
    1. EDA can uncover patterns of products that are frequently purchased together, aiding in cross-selling and recommendation systems.
5. **Customer Churn Analysis**:
    1. EDA can identify factors contributing to customer churn and assist in designing retention strategies.
6. **Campaign Effectiveness**:
    1. Analyzing marketing campaign data helps determine the effectiveness of various channels, messages, and timing.
    2. EDA can reveal which campaigns generate the highest ROI.

EXAMPLE 4*

New cancer cases in the U.S. based on a cancer registry

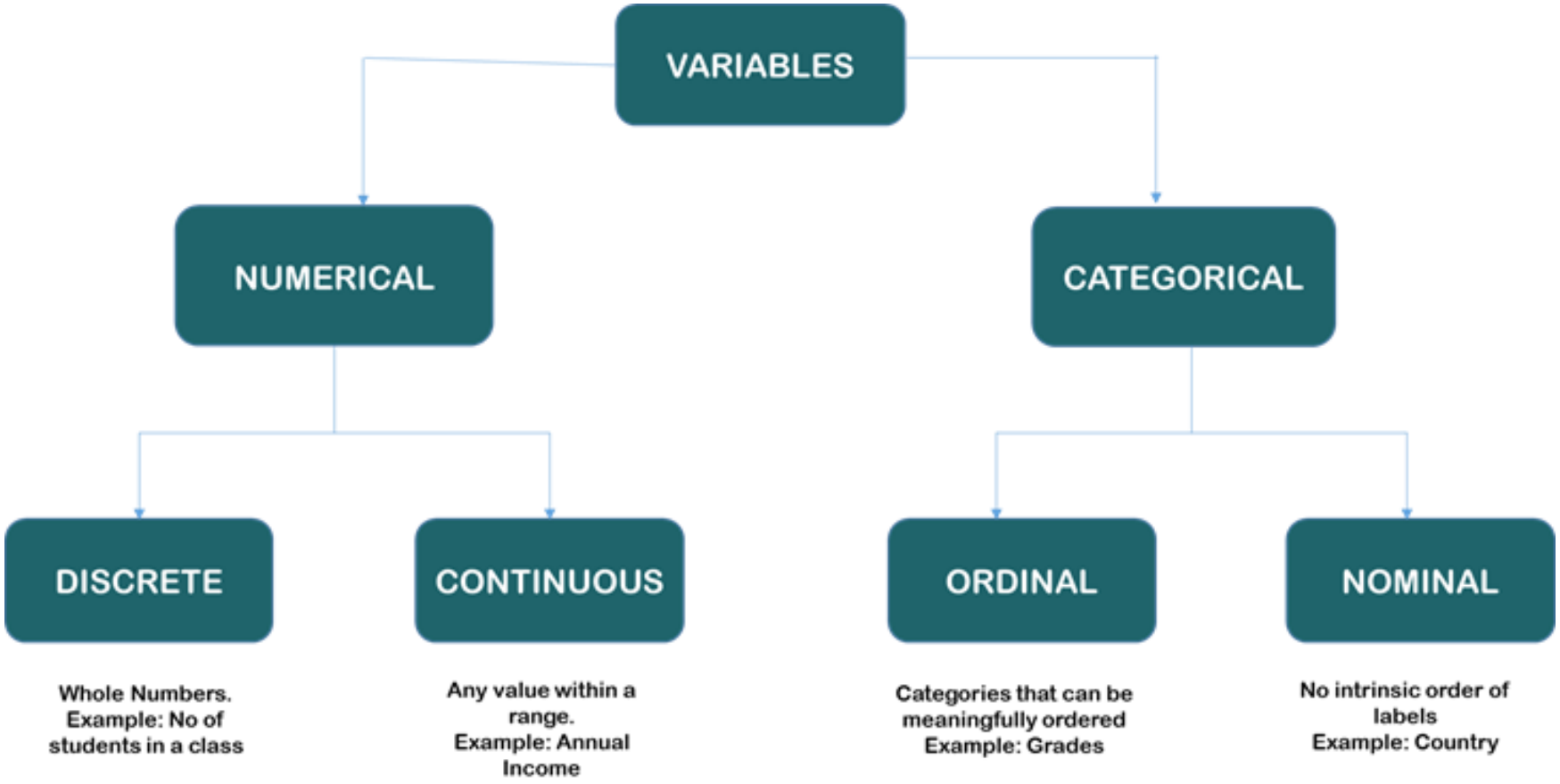
- The rows in the registry are called observations they correspond to individuals
- The columns are variables or data fields they correspond to attributes of the individuals

Cancer Registry, 2001

ID	Sex	Age	Date	Site (ICD-9)
⋮	⋮	⋮	⋮	⋮
1023	M	58	1/1/01	1530
1024	F	69	1/2/01	1740
1025	F	29	1/2/01	1610
1026	M	46	1/4/01	1410
⋮	⋮	⋮	⋮	⋮

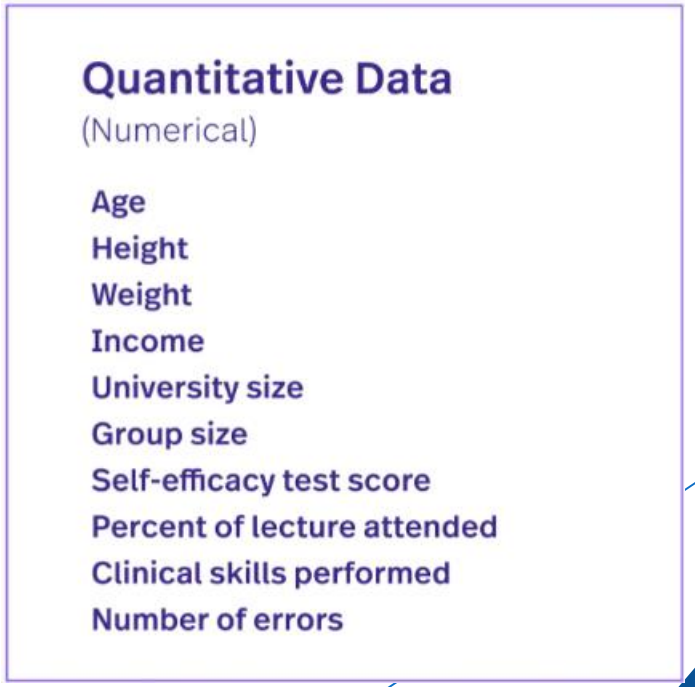
https://www.biostat.wisc.edu/~lindstro/2.EDA.9.10.pdf

Making Sense of Data



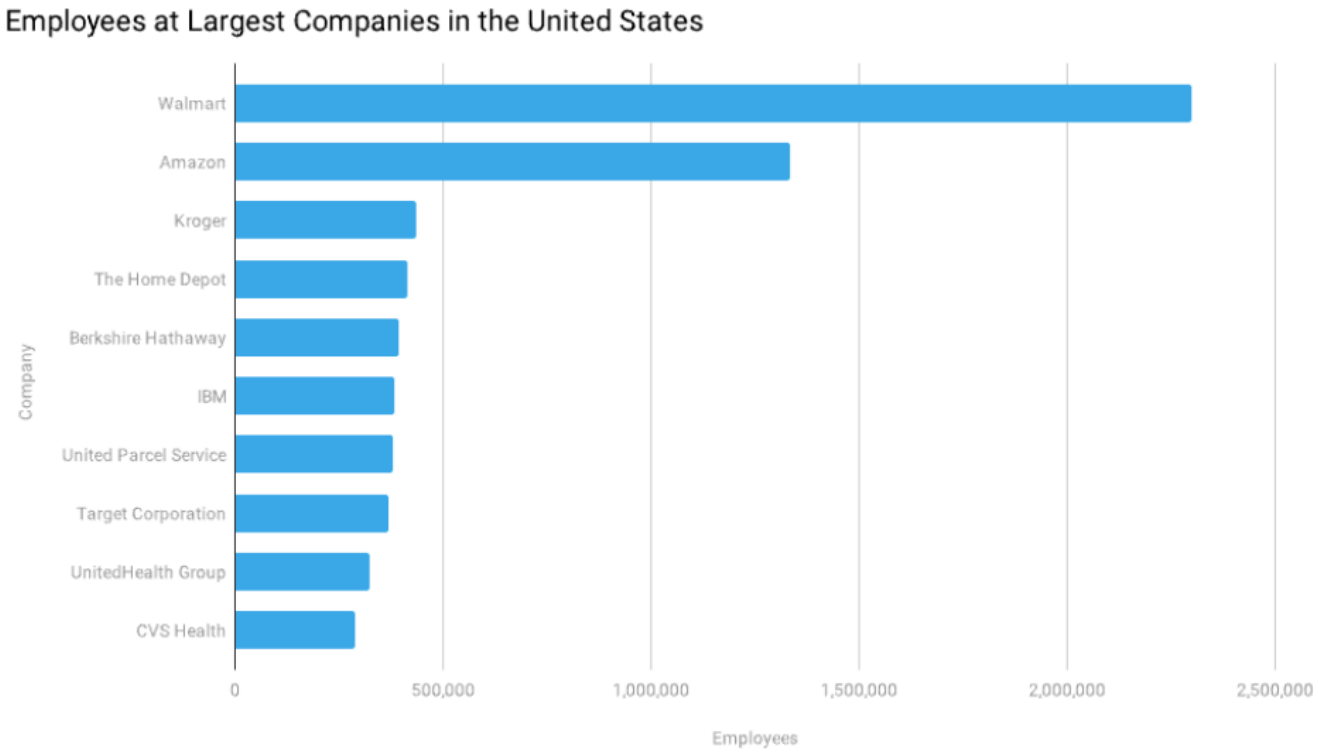
Quantitative/Numerical Data

- Quantitative data is data that can be counted or measured in numerical values.
- The two main types of quantitative data are discrete data and continuous data.



Discrete Data

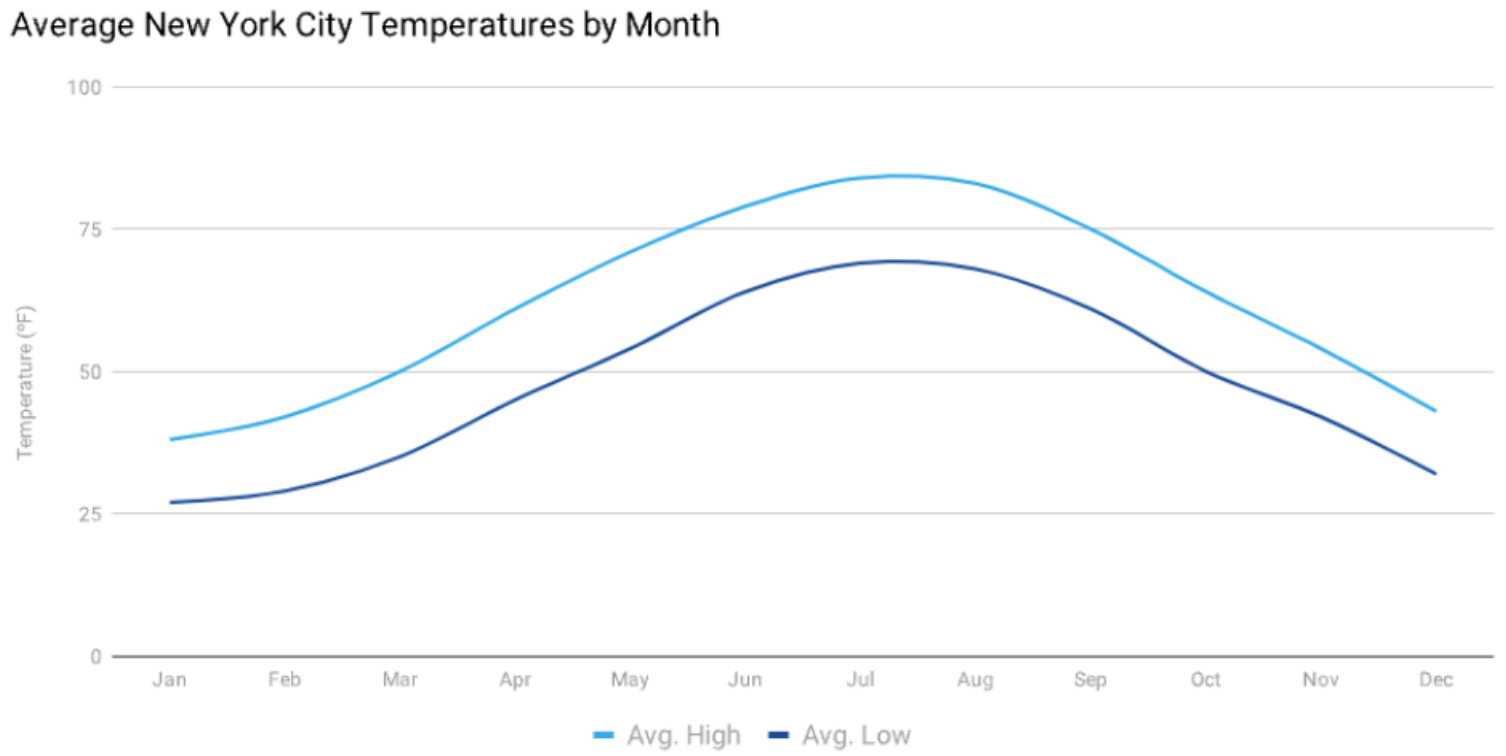
- Discrete data is information that can only take certain fixed values.
- Data that is countable and its values can be listed out.
- Example: The number of players in a team, No. of employees, etc.



An example of a bar graph representing quantitative data

Continuous Data

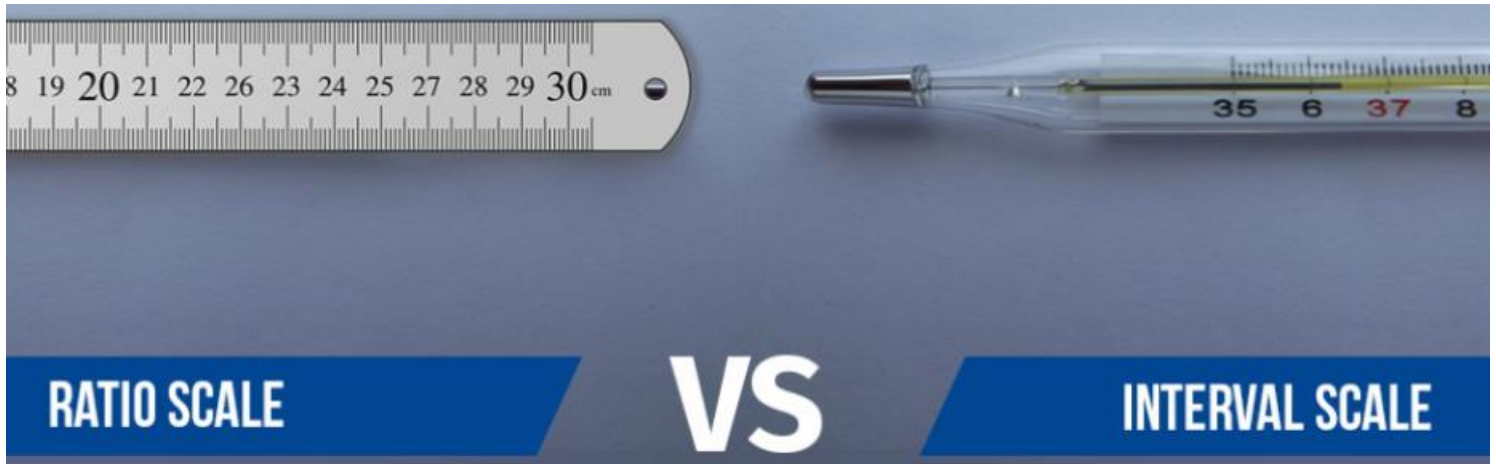
- Variable that can have an infinite number of numerical values within a specific range.
- Example: Website traffic, Water Temperature, Wind Speed, etc.



An example of a line graph representing quantitative data

Continuous Data

- Continuous data can be further broken down into two categories: interval data and ratio data.
- Interval data can include numerical data that does not use zero as a reference.
- Ratio data uses absolute zero as a reference point for measurement.



Qualitative or Categorical Data

- Categorical data is divided into groups or categories.
- The categories are based on qualitative characteristics.
- There is no order to categorical values and variables.
- Categorical data can take numerical values, but those numbers don’t have any mathematical meaning.
- Categorical data is displayed graphically by bar charts and pie charts.

Categorical data Example

Qualitative Data
(Categorical)

- Gender
- Religion
- Marital status
- Native language
- Social class
- Qualifications
- Type of instruction
- Method of treatment
- Type of teaching approach
- Problem-solving strategy used

Two-way Table

	Eye Color				
Hair Color	Green	Blue	Brown	Black	Total
Blonde	4	7	2	1	14
Brown	2	4	18	2	26
Black	1	2	5	2	10
Total	7	13	25	5	50

Nominal Data & Ordinal Data

- Unordered categorical data (nominal)
  2 possible values (binary)
    Examples: gender, alive/dead, yes/no.
  Greater than 2 possible values - No order to categories
    Examples: marital status, religion, country of birth, race.

- Ordered categorical data (ordinal)
  Ratings or preferences
  Cancer stage
  Quality of life scales,
  Severity of a software bug (critical, high, medium, low)
  Experience working with us (very great, great, bad)

Problem-1

Given the following questions from the survey, state what type of variable each question deals with.

Question

1. How old are you?
2. Where do you live? Give the name of your city
3. How many siblings do you have?
4. What is your height?
5. What is your birth date?
6. Do you have a pet?
7. What grade level are you in?

Problem-1 : Solution

Question Variable Type

1. How old are you? **Quantitative**
2. Where do you live? Give the name of your city **Categorical**
3. How many siblings do you have? **Quantitative**

It can also be categorical if put into groups (group people who have 1 sibling, then those that have 2 and so on or those who have less than 3, those who have more than 3, etc).

4. What is your height? **Quantitative**
5. What is your birth date? **Categorical**

6. Do you have a pet? **Categorical**
7. What grade level are you in? **Categorical**

Problem-2

Consider a dataset containing information about a group of students. Determine whether each of the following attributes is numerical or categorical:

1. Age
2. Gender
3. Student ID
4. Test Scores
5. Favorite Color
6. ZIP Code
7. Number of Siblings
8. Country of Birth
9. Student's Email Address
10. Height (in inches)

Problem-2 - Solution

1.Age :**Numerical**: can be measured on a continuous scale.
2.Gender :**Categorical**: Gender typically has two or more categories.
3.Student ID : **Categorical**: Student ID is typically an identifier or label, and it doesn't have inherent mathematical meaning.
4.Test Scores :**Numerical**: Test scores are quantitative data that can be measured and analyzed mathematically.
5.Favorite Color : **Categorical**.
6.ZIP Code : **Categorical**: ZIP codes are typically used to represent geographic regions.
7.Number of Siblings : **Numerical**: The number of siblings is a numerical variable that represents a count of discrete entities.
8.Country of Birth : **Categorical**: Country of birth is a categorical attribute with different countries as categories.
9.Student's Email Address : **Categorical**: Email addresses are text-based identifiers and do not have numerical values.
10.Height (in inches) :**Numerical**: Height is a quantitative attribute, and when expressed in inches, it can be measured as a numerical value.

Excercise-1

Customer ID	Customer Name	Age	Gender	Email Address	Purchase Amount	Payment Method
1	John Smith	35	Male	john@email.com	$120.50	Credit Card
2	Sarah Johnson	28	Female	sarah@email.com	$85.00	PayPal
3	Michael Brown	42	Male	michael@email.com	$220.75	Cash
4	Emily Davis	31	Female	emily@email.com	$75.20	Credit Card
5	David Wilson	29	Male	david@email.com	$112.60	PayPal
6	Laura Lee	35	Female	laura@email.com	$95.30	Credit Card

Excercise-2

	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
0	15.0	8	350.0	165.0	3693.0	11.5	70	1	buick skylark 320
1	18.0	8	318.0	150.0	3436.0	11.0	70	1	plymouth satellite
2	16.0	8	304.0	150.0	3433.0	12.0	70	1	amc rebel sst
3	17.0	8	302.0	140.0	3449.0	10.5	70	1	ford torino
4	15.0	8	429.0	198.0	4341.0	10.0	70	1	ford galaxie 500

Measurement Scales

Measurement Scales

1.Nominal Scale:

1. Represents categorical data without any inherent order or ranking.
2. Examples: Gender, colors, categories.
3. Differences: Different categories are distinct but not ordered. No quantitative relationship exists between categories.

2.Ordinal Scale:

1. Represents data with ordered categories or ranks but without precise differences between them.
2. Examples: Rankings (1st, 2nd, 3rd), Likert scales (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree).
3. Differences: Ordered categories, but the exact difference between ranks may not be defined.

Measurement Scales

3.Interval Scale:

1. Represents data with ordered categories and precise, equal intervals between them.
2. Examples: Temperature in Celsius or Fahrenheit, calendar dates.
3. Differences: Equal intervals between points on the scale, but there's no true zero point (zero doesn't indicate the absence of the attribute being measured).

4.Ratio Scale:

1. Represents data with ordered categories, equal intervals, and a true zero point.
2. Examples: Height, weight, time, income.
3. Differences: Possesses all the properties of interval scale but also has a true zero, enabling meaningful ratios and arithmetic operations.

Example: Identify the measurement scale?

1.What is your favorite color?

2.Which country are you from?

3.Rate your satisfaction with our service on a scale from 1 to 5.

4.Rank the following in order of preference: Action, Comedy, Drama.

5.What is the temperature today in Celsius?

6.On a scale of 1 to 10, how happy are you today?

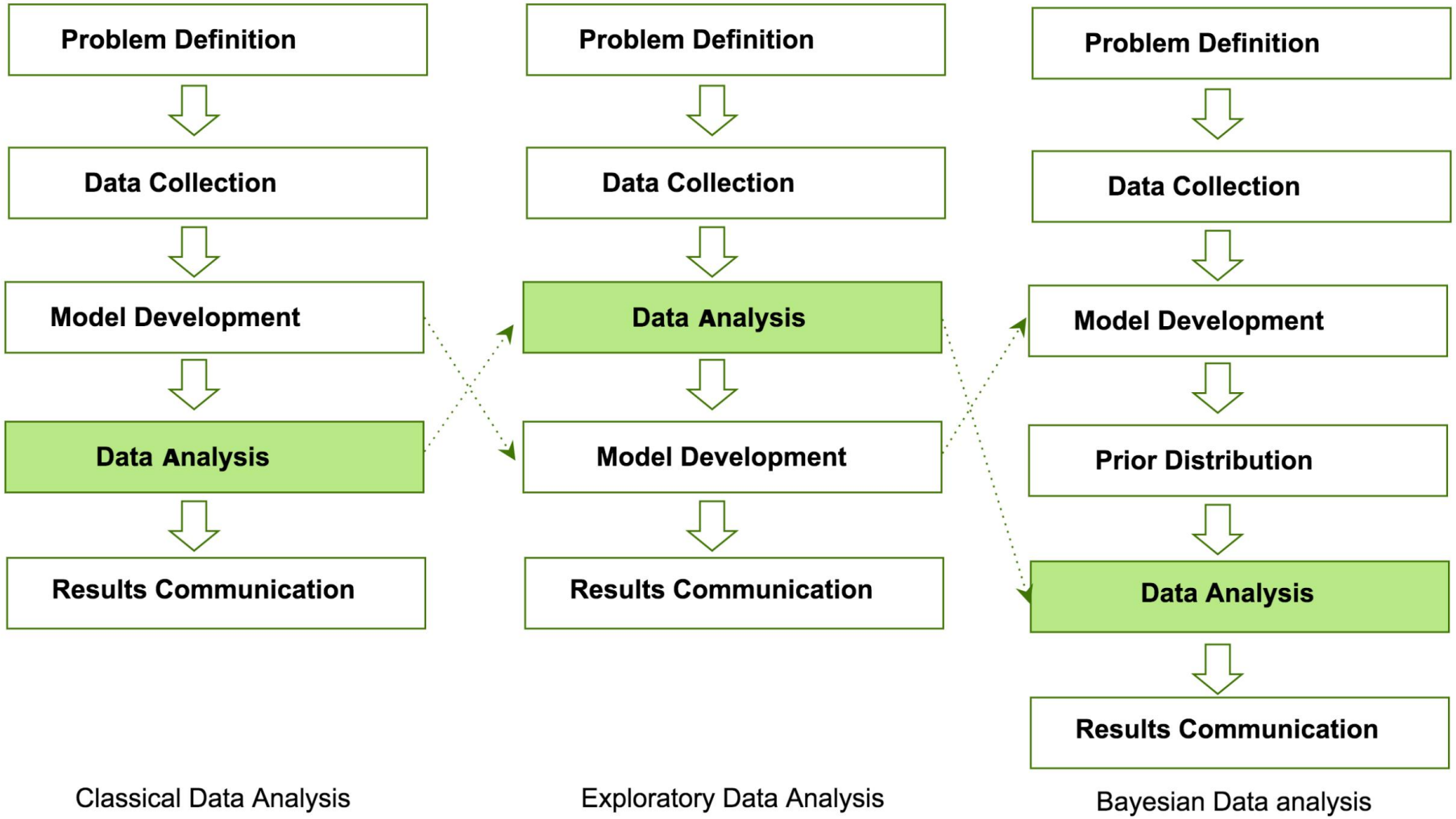
7.How many hours did you study for the exam?

8.What is your annual income?

Example

1. **Question:** What is your favorite color?
   Solution: The answers (e.g., "Red," "Blue," "Green") represent categories without an inherent order, indicating a nominal scale.
2. **Question:** Which country are you from?
   Solution: Responses like "USA," "France," "Japan" represent categories without a specific order, indicating a nominal scale.
3. **Question:** Rate your satisfaction with our service on a scale from 1 to 5.
   Solution: Responses with ordered categories (e.g., "1 - Very Dissatisfied," "5 - Very Satisfied") without precise intervals represent an ordinal scale.
4. **Question:** Rank the following in order of preference: Action, Comedy, Drama.
   Solution: Responses like "1st - Comedy," "2nd - Action," "3rd - Drama" represent ordered categories without quantifiable differences, indicating an ordinal scale.
5. **Question:** What is the temperature today in Celsius?
   Solution: Answers like "20°C," "25°C," represent ordered categories with equal intervals but lack a true zero, indicating an interval scale.
6. **Question:** On a scale of 1 to 10, how happy are you today?
   Solution: Responses on a numerical scale from 1 to 10 represent ordered categories with equal intervals but without a true zero, indicating an interval scale.
7. **Question:** How many hours did you study for the exam?
   Solution: Answers like "0 hours," "2 hours," "5 hours" involve ordered categories with equal intervals and a true zero, indicating a ratio scale.
8. **Question:** What is your annual income?
   Solution: Responses involving numerical values (e.g., $30,000, $50,000) with a true zero point represent ordered categories with equal intervals and a true zero, indicating a ratio scale.

Comparing EDA with classical and Bayesian analysis



Additional Resources

Classification of Digital Data

Digital data is classified into the following categories:

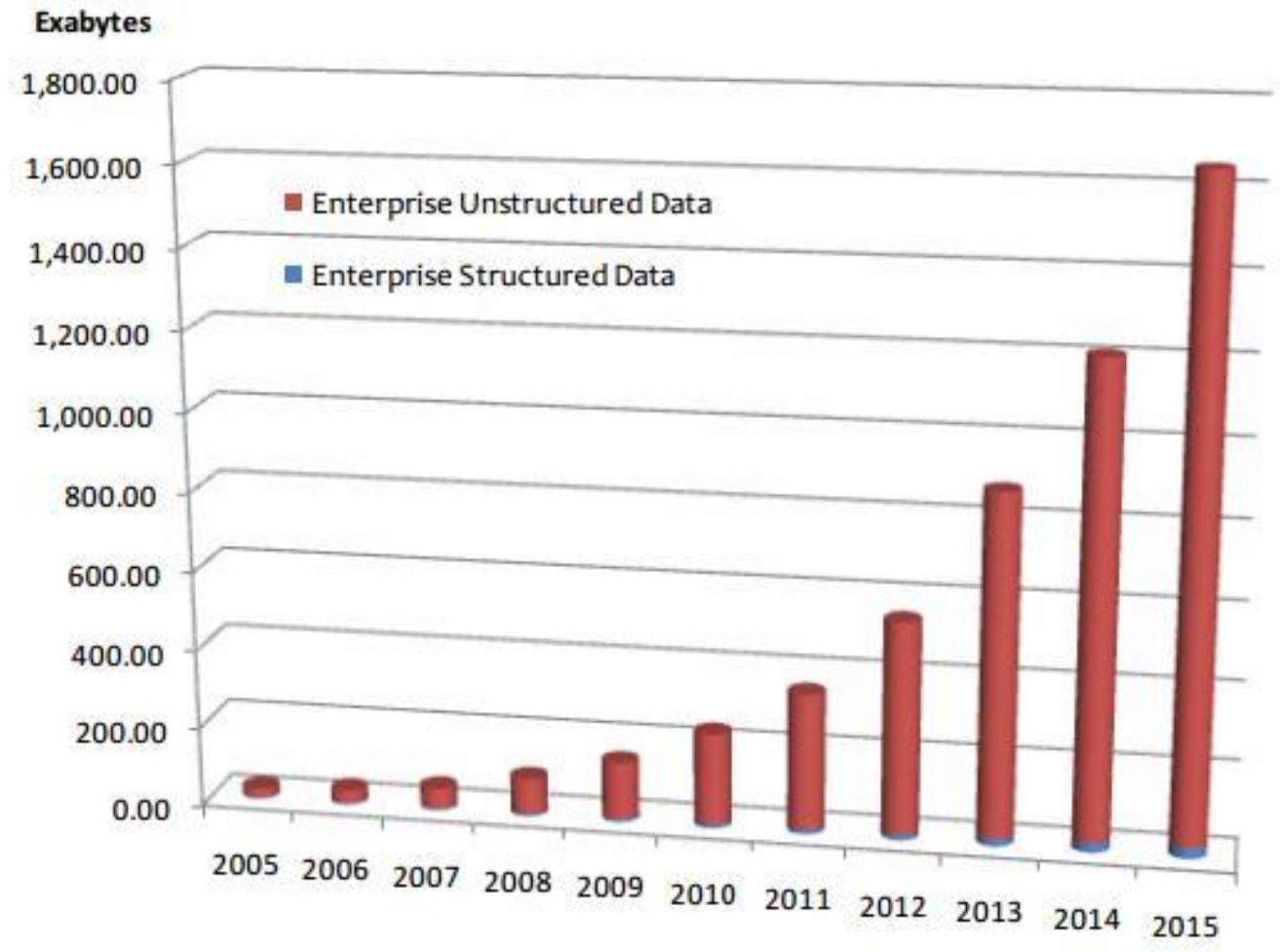
- Structured data- This is the data which is in an organized form(e.g, rows and columns) and can be easily used by a computer program. Relationships exist between entities of data, such as classes and their objects. Data stored in databases is an example of structured data.

- Semi-structured data- This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer program, for example, emails, XML, markup languages like HTML etc.,

- Unstructured data- -This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program. About 80%-90% data of an organization is in this format for example, memos, chat rooms, powerpoint presentations, images, videos, letters etc,.

Approximate Distribution of Digital Data

Approximate percentage distribution of digital data

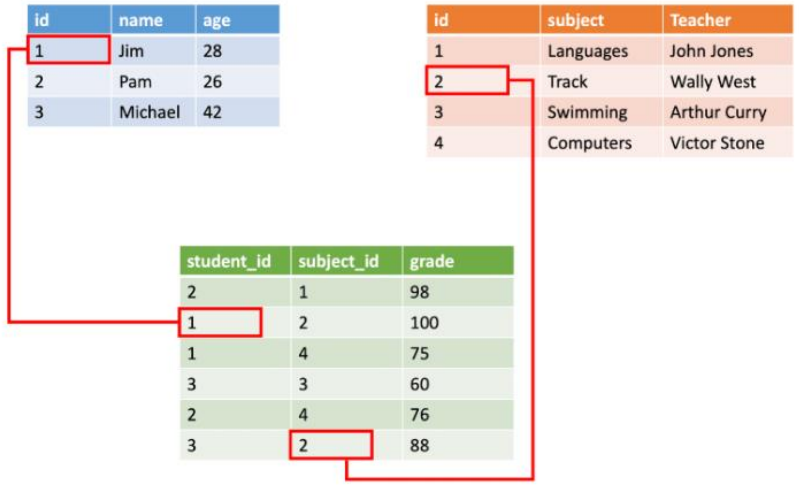


Structured Data

Structured Data

- This is the data which is in an organized form (e.g., in rows and columns) and can be easily used by a computer program.
- In structured data, all row in a table has the same set of columns.

- Data stored in databases is an example of structured data.

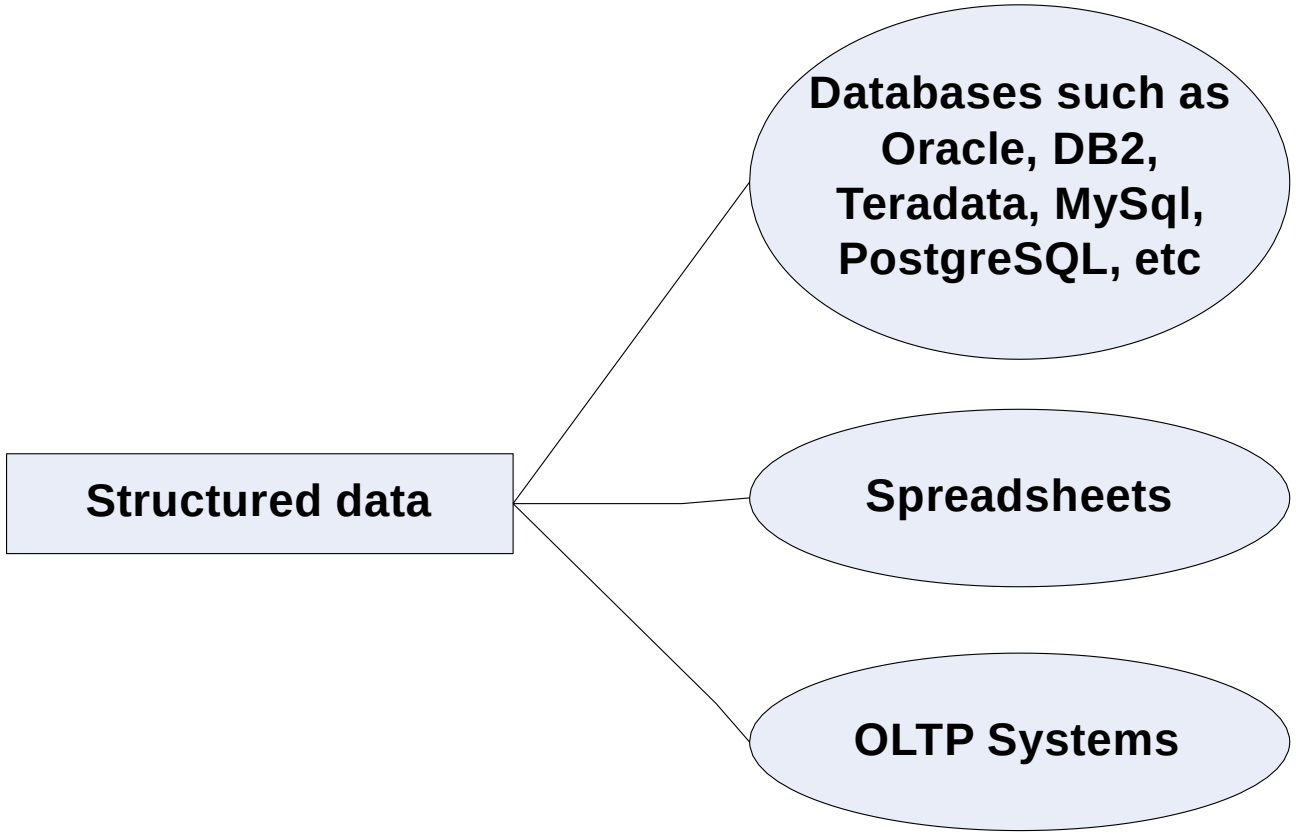


id	name	age
1	Jim	28
2	Pam	26
3	Michael	42

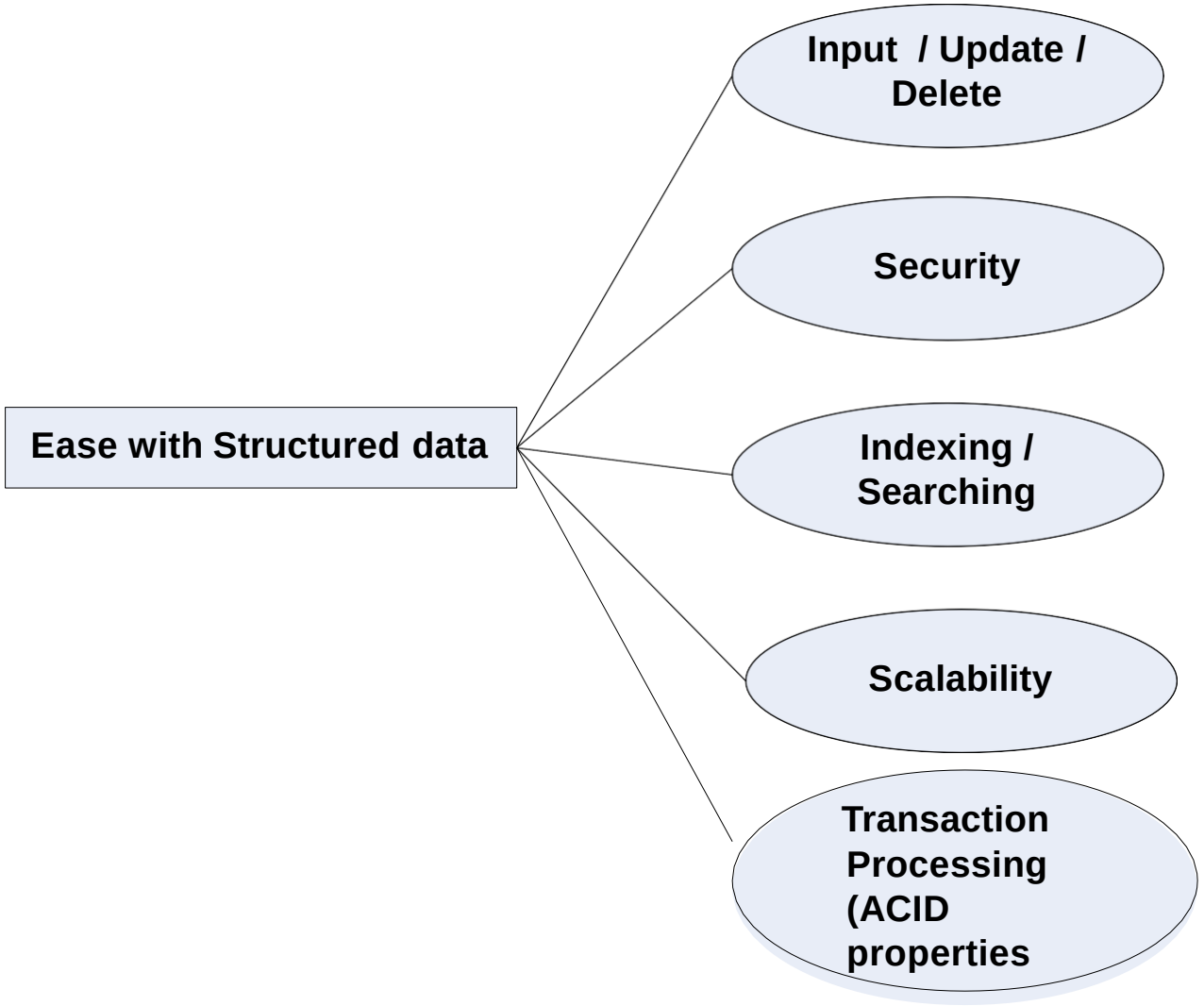
id	subject	Teacher
1	Languages	John Jones
2	Track	Wally West
3	Swimming	Arthur Curry
4	Computers	Victor Stone

student_id	subject_id	grade
2	1	98
1	2	100
1	4	75
3	3	60
2	4	76
3	2	88

Sources of Structured Data



Ease with Structured Data



Semi-structured Data

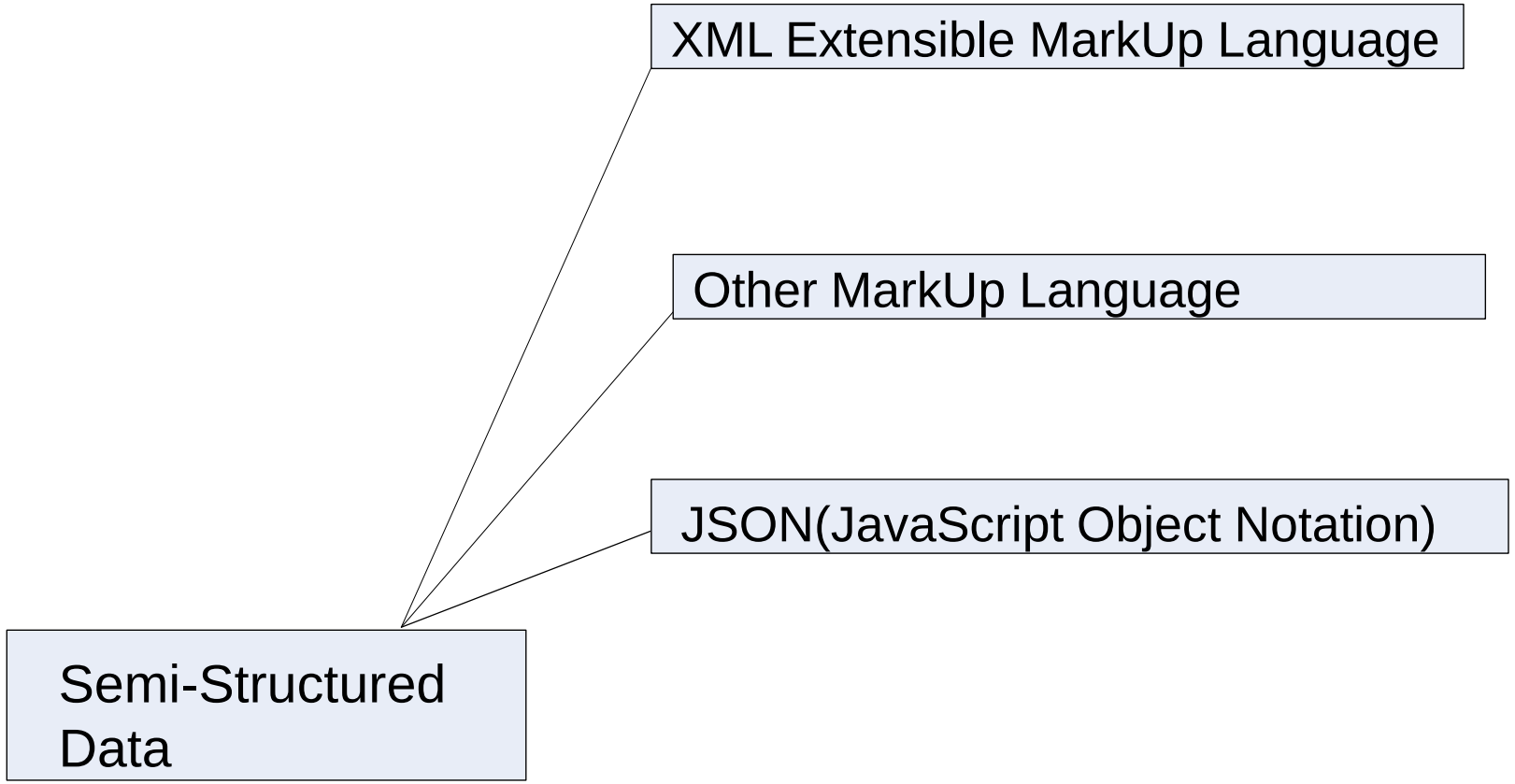
Semi-structured Data

- This is the data which does not conform to a data model but has some structure. However, it is not in a form which can be used easily by a computer program.

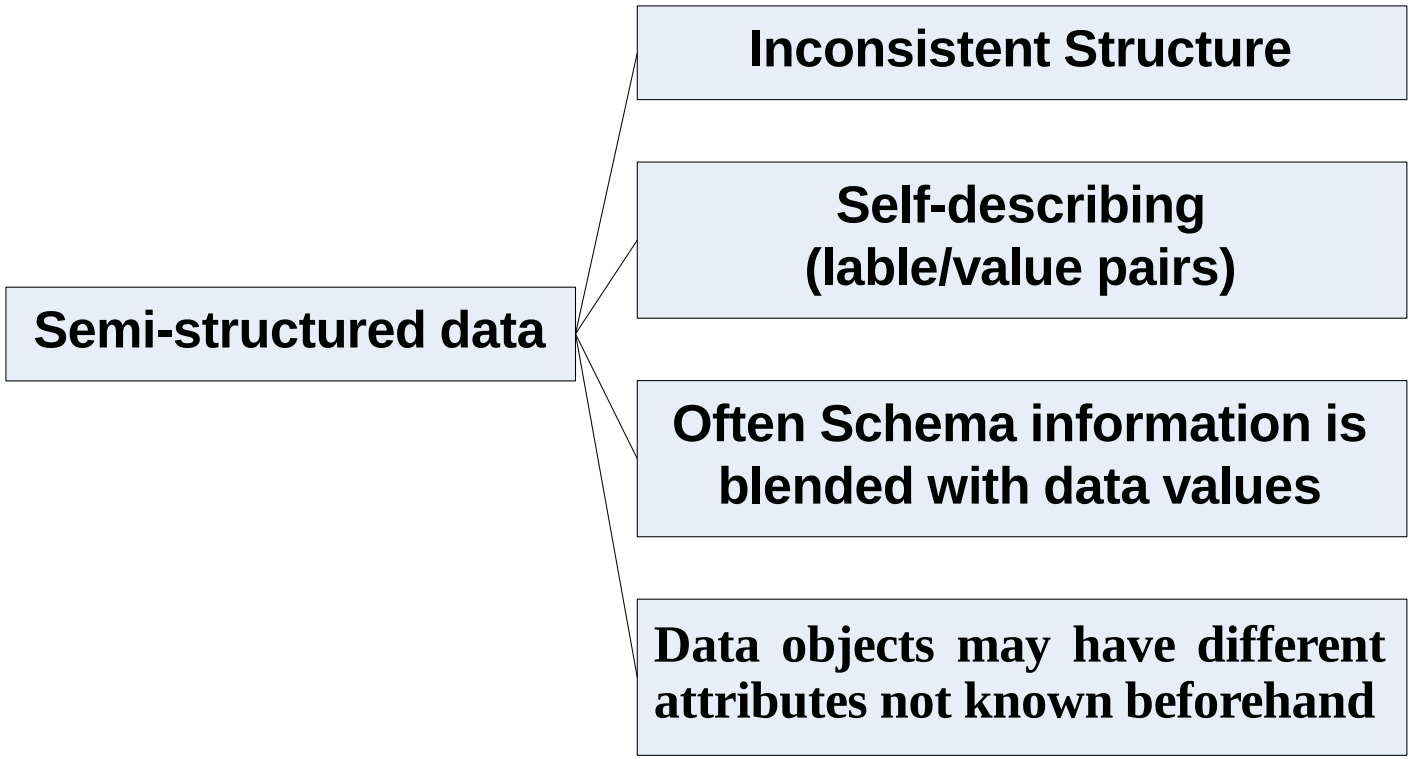
- Example, emails, XML, markup languages like HTML, etc. Metadata for this data is available but is not sufficient.

```
## Document 1 ##
{
  "customerID": "103248",
  "name":
  {
    "first": "AAA",
    "last": "BBB"
  },
  "address":
  {
    "street": "Main Street",
    "number": "101",
    "city": "Acity",
    "state": "NY"
  },
  "ccOnFile": "yes",
  "firstOrder": "02/28/2003"
}
```

Sources of Semi-structured Data



Characteristics of Semi-structured Data

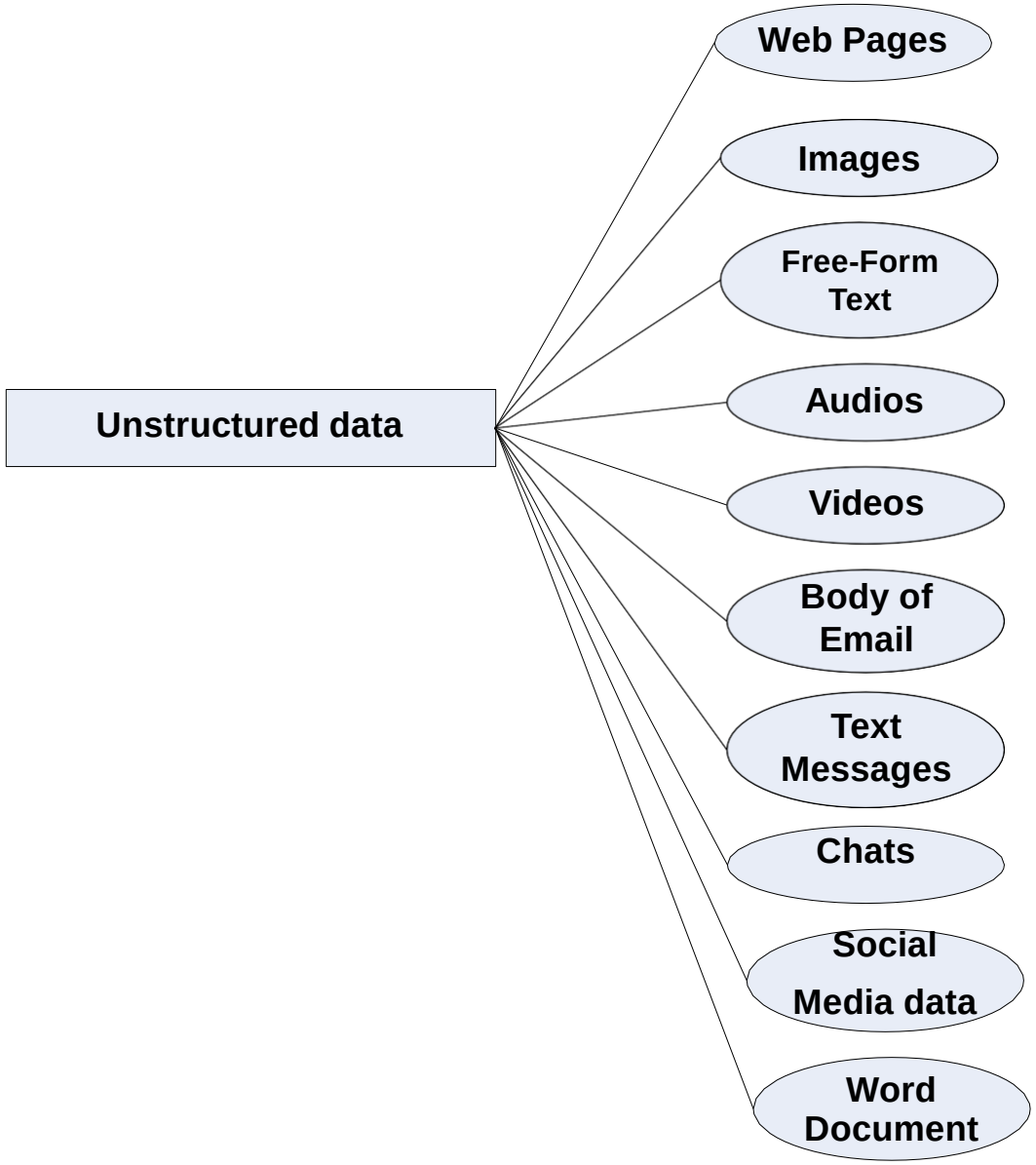


Unstructured Data

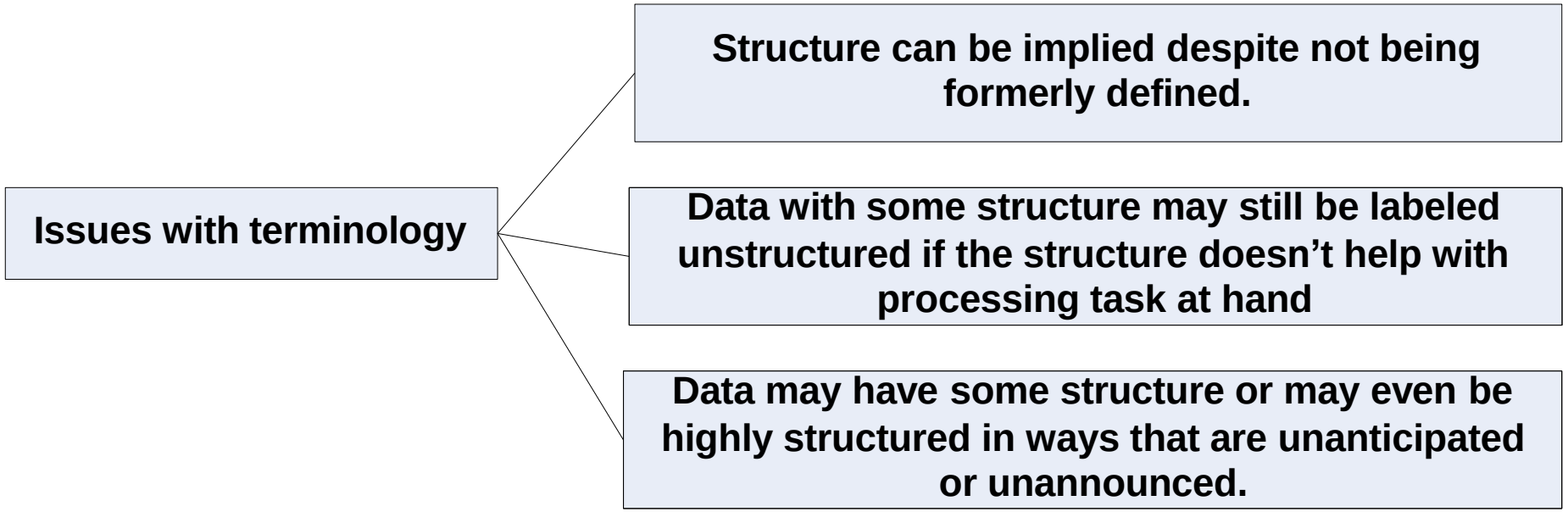
Unstructured Data

- This is the data which does not conform to a data model or is not in a form which can be used easily by a computer program.
- About 80–90% data of an organization is in this format.
- Example: memos, chat rooms, PowerPoint presentations, images, videos, letters, researches, white papers, body of an email, etc.

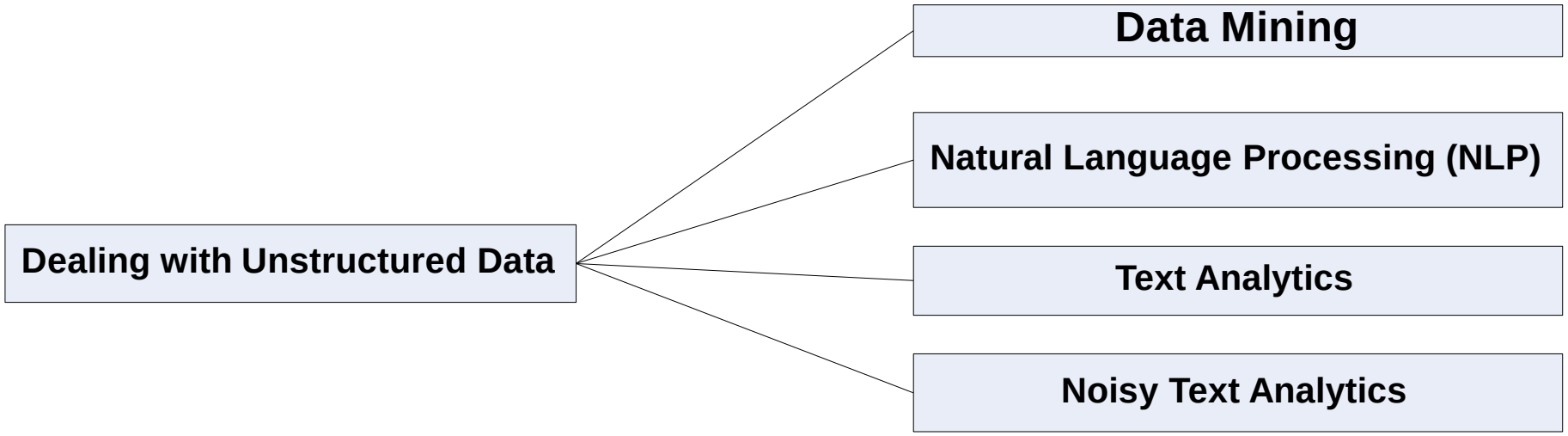
Sources of Unstructured Data



Issues with terminology - Unstructured Data



Dealing with Unstructured Data



Dealing with Unstructured Data

- **Data Mining**
  - Association Rule Mining
  - Regression Analysis
  - Collaborative Filtering

- **Text analysis and Text Mining**

- **Natural Language Processing(NLP)**

- **Noisy text Analysis**

- **Manual tagging with metadata**

- **Part-of-speech tagging**

- **Unstructured Information Management Architecture(UIMA)**

Answer a few quick questions …

Answer Me

- Which category (structured, semi-structured, or unstructured) will you place a Web Page in?

- Which category (structured, semi-structured, or unstructured) will you place Word Document in?

- State a few examples of human generated and machine-generated data.

Place Me in the Basket

Structured	Unstructured	Semi-Structured

Following words are to be placed in the relevant basket:

Email
MS Access
Images
Database
Chat conversations
Relations/Tables
Facebook
Videos
MS Excel
XML

Answer:

Structured	Unstructured	Semi-Structured
MS Access	Email	XML
Database	Images	
Relations/Tables	Chat conversations	
MS Excel	Facebook	
	Videos	

References ...

Further Readings

Exploratory Data Analysis (EDA) | Introduction to EDA (analyticsvidhya.com)

https://www.simplilearn.com/tutorials/data-analytics-tutorial/exploratory-data-analysis

https://www.analyticsvidhya.com/blog/2021/05/exploratory-data-analysis-eda-a-step-by-step-guide/

https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/

https://intellipaat.com/blog/what-is-eda-in-data-science/

https://www.knowledgehut.com/blog/data-science/eda-data-science#frequently-asked-questions

https://www.powershow.com/view/aca5d-OTEwN/Exploratory_Data_Analysis_powerpoint_ppt_presentation