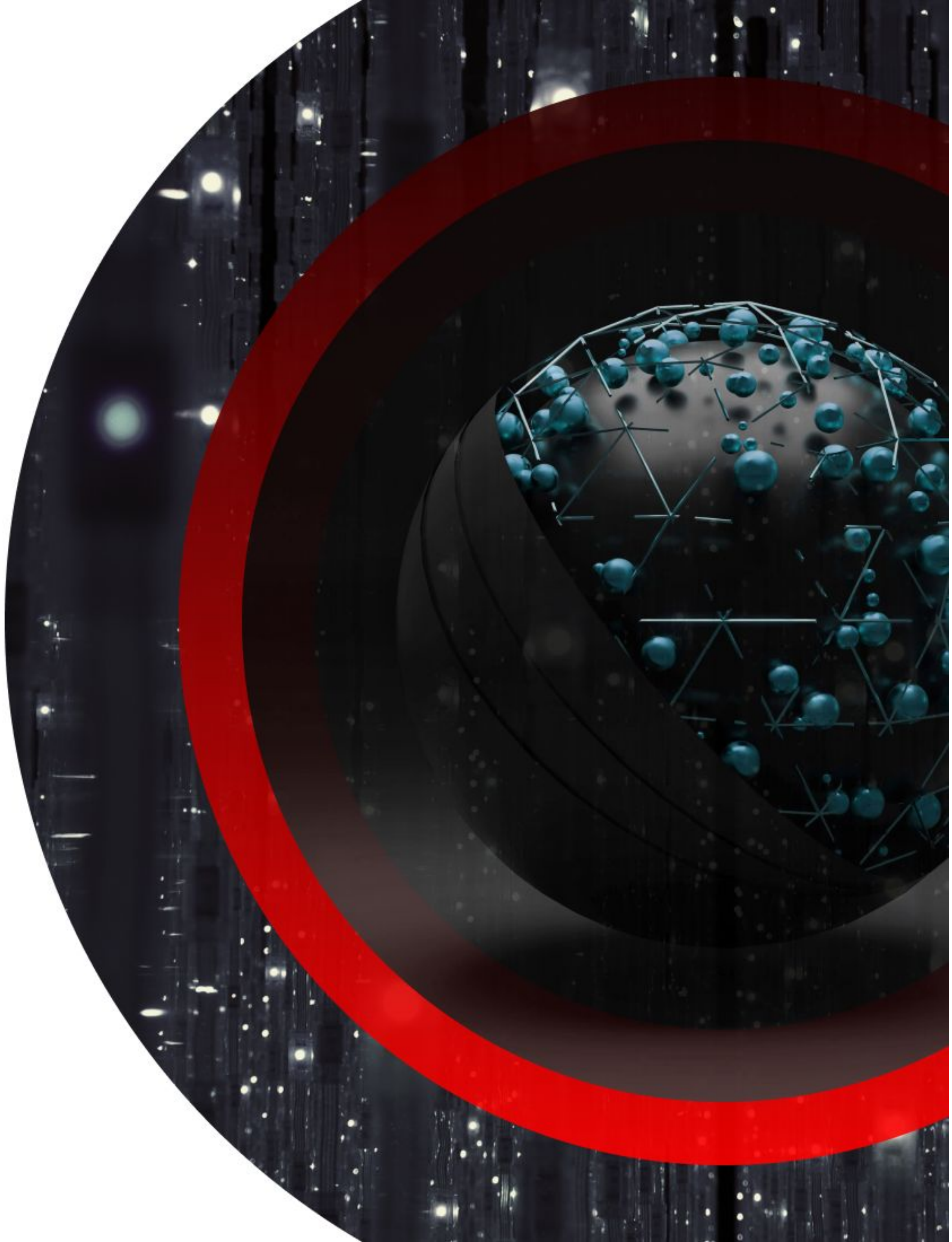
DATA SCIENCE AND ANALYTICS

Introductory Course



CLASS NORMS

COURTESY IN CLASS

Remaining on mute unless called on, exercising courtesy during breakout rooms, using the chat box for questions only

ATTENDANCE

100% attendance is expected and contributes to success in passing the course and the program

PARTICIPATION

Keeping an open mind in discussions and sharing experiences, making contributions during team assignments, submitting assignments in Canvas, and participating in discussion boards

USE OF CLASS RESOURCES

Follow along during the lecture with the lesson companion and download any in-class documents prior to class.

INSTRUCTIONAL TEAM

INSTRUCTIONAL TEAM

Name	Contact information
Name	Contact information
Name	Contact information

PROGRAM PATH

1 Introductory Course

2 SQL and Databases

3 Statistics and Probability

4 Data Storytelling

Milestone 1: Building and Presenting Data Stories

5 Python Programming

6 Data Wrangling

7 Visual Communications

8 Advanced SQL Programming

Milestone 2: Data Integration, Preparation, Reporting, and Presentation

9 Business Intelligence

10 Big Data

11 Machine Learning

12 Applied AI

Milestone 3: Capstone Project: Delivering Insights and Presentations

INTRODUCTORY COURSE PATH

1

INTRODUCTION TO DATA SCIENCE AND ANALYTICS

2

COMPUTING PRIMER

3

PROGRAMMING CONCEPTS

4

DISCOVERING AND CURATING DATA

5

STRUCTURING AND ANALYZING DATA

6

CLEANING AND ENRICHING DATA

7

VALIDATING AND PRESENTING DATA

8

INTRODUCTION TO DATA SCIENCE PROJECTS

9

ASSESSMENT NIGHT

MODULE OUTLINE

Lesson 1: Data Structure Types

Lesson 2: Data Storage Tools

Lesson 3: Data Analysis Process

LESSON 1: DATA STRUCTURE TYPES

LESSON OUTLINE

- What Is Data?
- Structured Data
- Semi-Structured Data
- Unstructured Data

BRAINSTORMING

Define *data* in your own words and give an example.

What are some sources to find data?

LESSON GOALS

WHAT ARE THE GOALS?

- Identify three basic categories of data.
- Explain the pros and cons of each data category.
- Provide examples of each data structure type.

WHY ARE THEY IMPORTANT?

Data professionals are expected to identify the type of data they are handling and use the right tools for different categories.

YOUR TAKE

What do you wish to accomplish by completing this module?

REVIEW

In the last module, we discussed:

- How the velocity and volume of data affects collection
- What APIs are and how they are used

WHAT IS DATA?

Data is information that is stored on a computer or server and used as a basis for analysis.

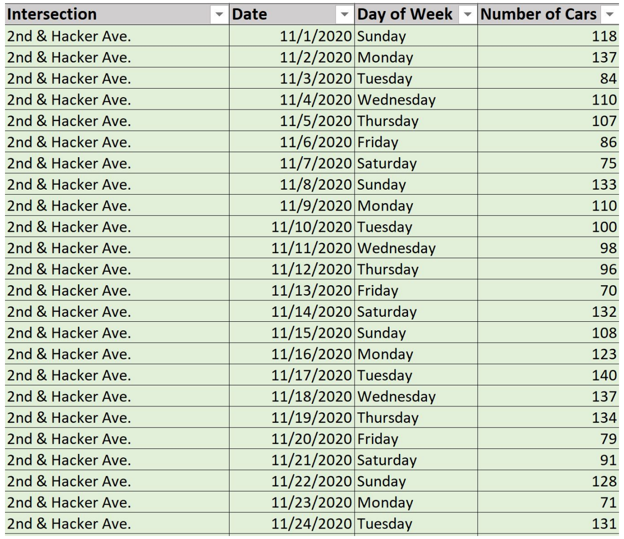
Data can be separated into three categories:

- Structured data
- Semi-structured data
- Unstructured data

STRUCTURED DATA

Structured data is information that is well-organized, such as in rows and columns.

Structured data makes up 20% of all corporate data.



Intersection	Date	Day of Week	Number of Cars
2nd & Hacker Ave.	11/1/2020	Sunday	118
2nd & Hacker Ave.	11/2/2020	Monday	137
2nd & Hacker Ave.	11/3/2020	Tuesday	84
2nd & Hacker Ave.	11/4/2020	Wednesday	110
2nd & Hacker Ave.	11/5/2020	Thursday	107
2nd & Hacker Ave.	11/6/2020	Friday	86
2nd & Hacker Ave.	11/7/2020	Saturday	75
2nd & Hacker Ave.	11/8/2020	Sunday	133
2nd & Hacker Ave.	11/9/2020	Monday	110
2nd & Hacker Ave.	11/10/2020	Tuesday	100
2nd & Hacker Ave.	11/11/2020	Wednesday	98
2nd & Hacker Ave.	11/12/2020	Thursday	96
2nd & Hacker Ave.	11/13/2020	Friday	70
2nd & Hacker Ave.	11/14/2020	Saturday	132
2nd & Hacker Ave.	11/15/2020	Sunday	108
2nd & Hacker Ave.	11/16/2020	Monday	123
2nd & Hacker Ave.	11/17/2020	Tuesday	140
2nd & Hacker Ave.	11/18/2020	Wednesday	137
2nd & Hacker Ave.	11/19/2020	Thursday	134
2nd & Hacker Ave.	11/20/2020	Friday	79
2nd & Hacker Ave.	11/21/2020	Saturday	91
2nd & Hacker Ave.	11/22/2020	Sunday	128
2nd & Hacker Ave.	11/23/2020	Monday	71
2nd & Hacker Ave.	11/24/2020	Tuesday	131

STRUCTURED DATA: PROS AND CONS

Pros:

- Easy to store and access
- Easy to understand
- Abundance of existing tools to work with it
- Easy for machine algorithms to use

Cons:

- Lack of flexibility
- Limited storage options
- Most uncommon type of data
- Predefined data confined to a specific use

SEMI-STRUCTURED DATA

Semi-structured data is information that has an organizational structure but does not fit into a traditional tabular format.

```
{"student_details":  [
        {
"student_first_name": "Aparna“,
"student_last_name": "Singh“,
"graduation_year": "2024“,
"student_id_number: "00102"
         },
         {
"student_first_name": "Ava“,
"student_last_name": "Del Rio“,
"graduation_year": "2026“,
"student_id_number: "00103"
          }
      ]
}
```

SEMI-STRUCTURED DATA: PROS AND CONS

Pros:

- Flexible, no rigid schema
- Structured data can be viewed as semi-structured data.
- Many different data sources can be analyzed.

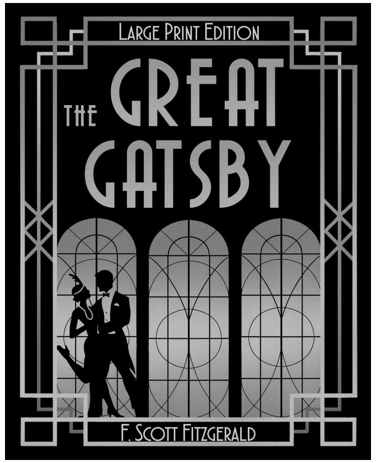
Cons:

- Difficult to determine relationships
- Hard to query information
- Lack of storage options

UNSTRUCTURED DATA

Unstructured data has no organizational structure. It describes the data within most companies.

Free-form text from a book is an example of unstructured data.



From: Simon & Schuster

UNSTRUCTURED DATA: PROS AND CONS

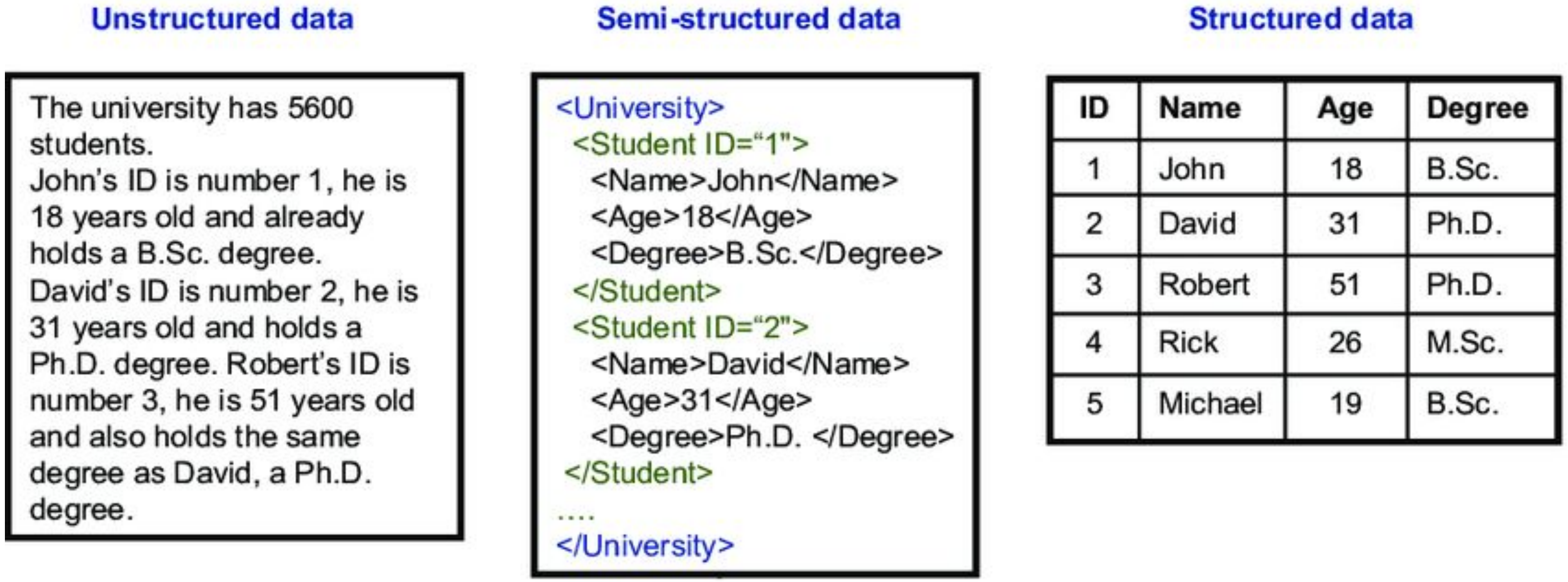
Pros:

- Maximum flexibility
- Scalable
- Not constrained by formatting

Cons:

- Difficult to manage
- Organization issues make it difficult to access data.
- Requires the most storage space

COMPARING DATA STRUCTURES



Unstructured data

The university has 5600 students. John's ID is number 1, he is 18 years old and already holds a B.Sc. degree. David's ID is number 2, he is 31 years old and holds a Ph.D. degree. Robert's ID is number 3, he is 51 years old and also holds the same degree as David, a Ph.D. degree.

Semi-structured data

```
<University>
 <Student ID="1">
   <Name>John</Name>
   <Age>18</Age>
   <Degree>B.Sc.</Degree>
 </Student>
 <Student ID="2">
   <Name>David</Name>
   <Age>31</Age>
   <Degree>Ph.D. </Degree>
</Student>
....
</University>
```

Structured data

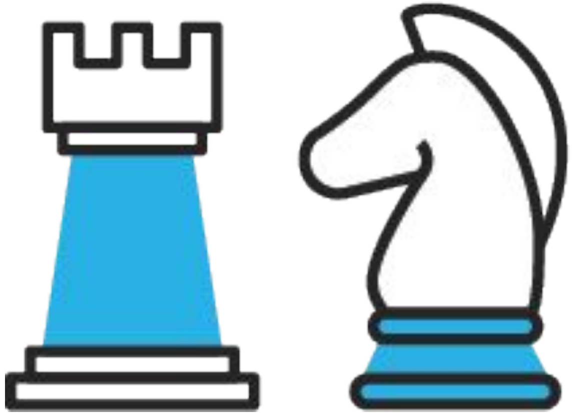
ID	Name	Age	Degree
1	John	18	B.Sc.
2	David	31	Ph.D.
3	Robert	51	Ph.D.
4	Rick	26	M.Sc.
5	Michael	19	B.Sc.

From: Cardoso, Jorge

ACTIVITY: SONYA'S COFFEE SHOP

In this activity, you will explore an unorganized data set.

You will need the 1.5.1 Activity document and Excel (1.5.1_Activity_Data.xlsx) to get started.



BREAKOUT ROOMS

QUESTIONS?

BREAK TIME

LESSON 2: DATA STORAGE TOOLS

LESSON OUTLINE

Databases

- Hierarchical
- Network
- Relational
- Non-relational
- Object-oriented

BRAINSTORMING

Define *database* in your own words.

What is the function and purpose of a database?

LESSON GOALS

WHAT ARE THE GOALS?

- Explain the functions of a database.
- Identify the best database management system for a particular data structure.

WHY ARE THEY IMPORTANT?

It is important to know the limitations of certain database management systems to store your data. Knowledge of how data is stored will give you an idea of how you can interact with it.

REVIEW

In the last lesson, we discussed:

- Three different data structures: structured, unstructured, and semi-structured
- The pros and cons of each type of data structure

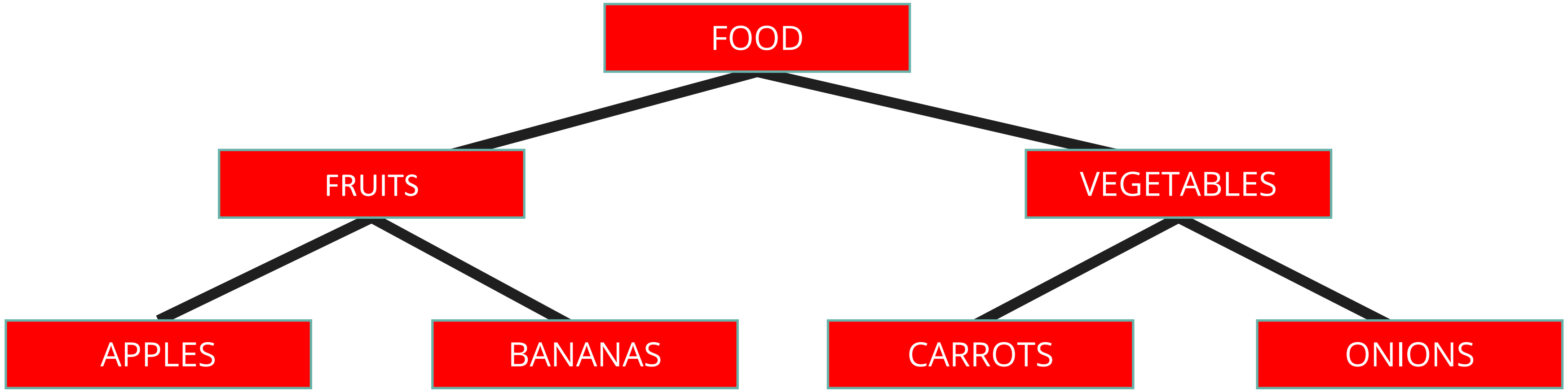
DATABASES

A **database** is an organized structure where data is stored.

You can access data in a DataBase Management System **(DBMS)** to:

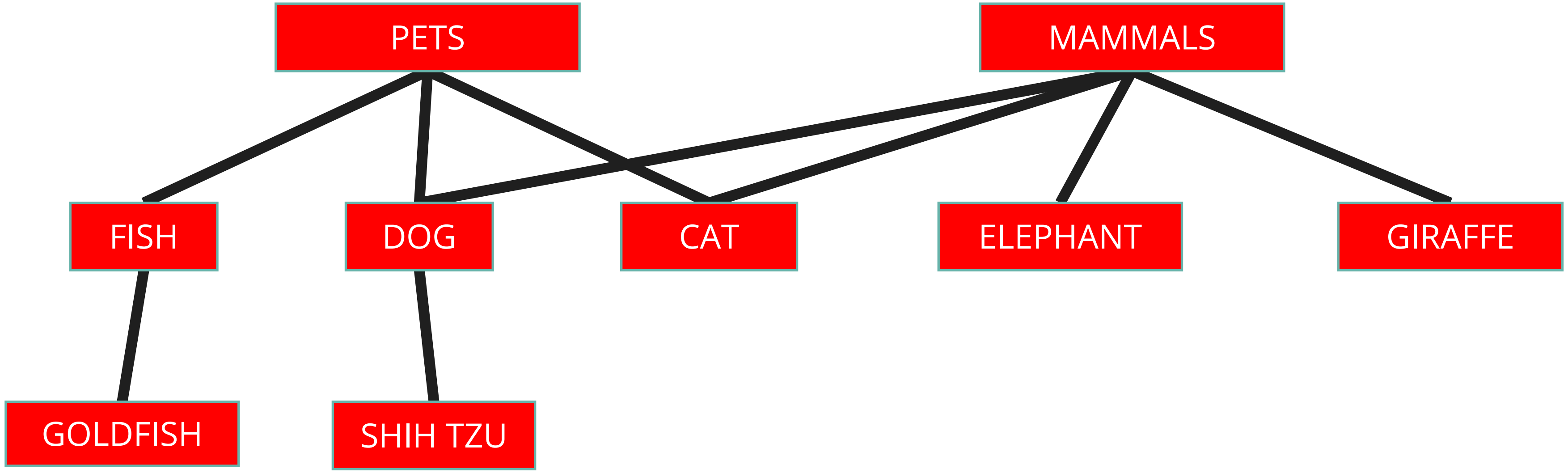
- Define the organizational structure of data.
- Add/delete/modify data.
- Access existing data.

HIERARCHICAL DATABASE



Hierarchical Database: Data organized by a series of parent/child relationships

NETWORK DATABASE



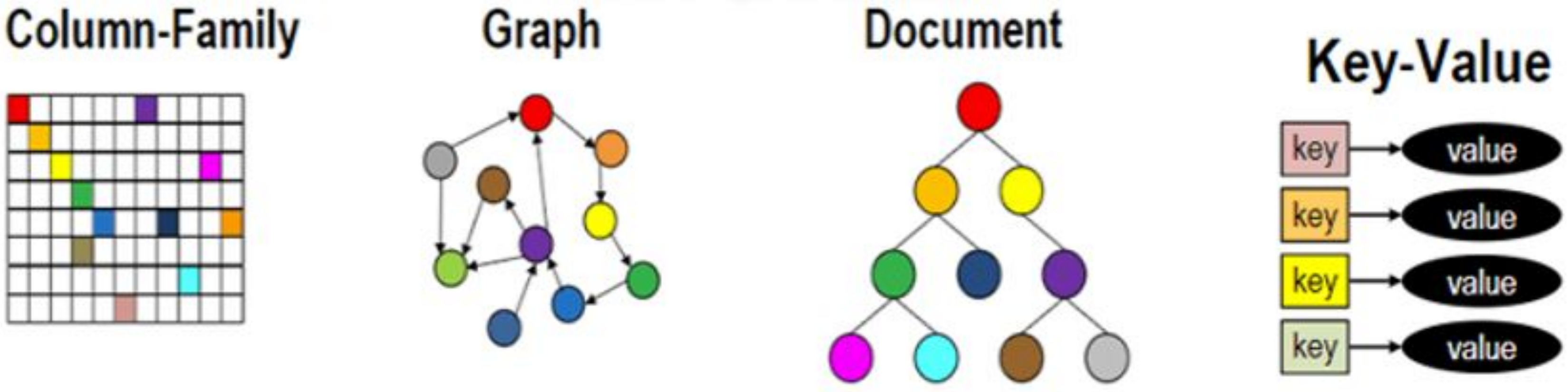
Network Database: A series of data relationships where child nodes may have *multiple* parent nodes

RELATIONAL DATABASE

MOVIE	YEAR	STUDIO	ROTTEN TOMATOES RATING
Toy Story 4	2019	Disney-Pixar	97%
Spider-Man: Into the Spider-Verse	2018	Sony	97%
Inside Out	2015	Disney-Pixar	98%
Coco	2017	Disney-Pixar	97%
Snow White	1937	Disney	98%

Relational Database: Organized into rows and columns within tables

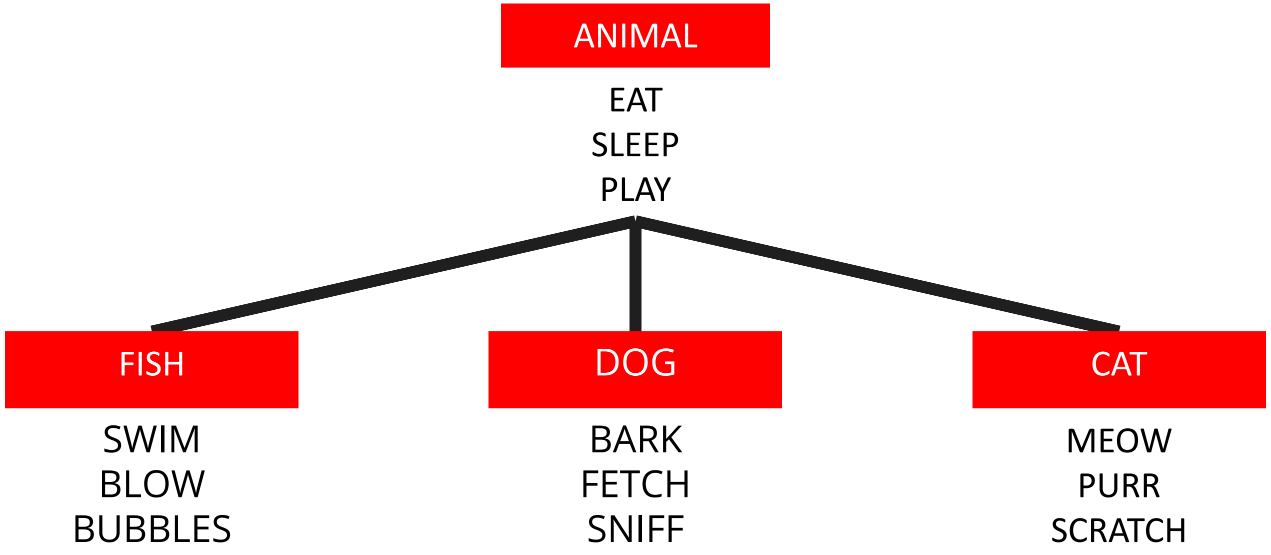
NON-RELATIONAL DATABASE



From: Guru99

A **non-relational database** does not rely on a tabular structure.

OBJECT-ORIENTED DATABASE



ANIMAL

EAT
SLEEP
PLAY

FISH

SWIM
BLOW
BUBBLES

DOG

BARK
FETCH
SNIFF

CAT

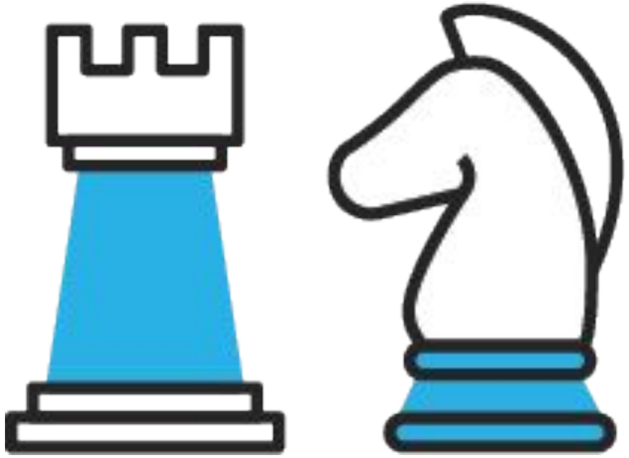
MEOW
PURR
SCRATCH

An **object-oriented database** is stored as objects that exist within classes.

ACTIVITY: BRICKBUSTER

In this activity, you will assist AJ in choosing the best database management system for his movie rental business.

You will need the 1.5.2 Activity document to get started.



BREAKOUT ROOMS

QUESTIONS?

BREAK TIME

LESSON 3: DATA ANALYSIS PROCESS

LESSON OUTLINE

- Data Analysis
- Define the Data Goal
- Collect Data
- Clean and Process Data
- Analyze the Data
- Initial Data Analysis
- Primary Data Analysis
- Confirmatory Data Analysis
- Exploratory Data Analysis
- Draw Conclusions

BRAINSTORMING

What do you think the analytic process is?

What are some things we might need to consider?

LESSON GOALS

WHAT ARE THE GOALS?

- List the five steps of the analytic process.
- Identify when to collect, clean, and analyze data.
- Draw conclusions based on the data goals.

WHY ARE THEY IMPORTANT?

Understanding the analytic process is key for any data professional. This lesson acts as a foundation for the data analysis process.

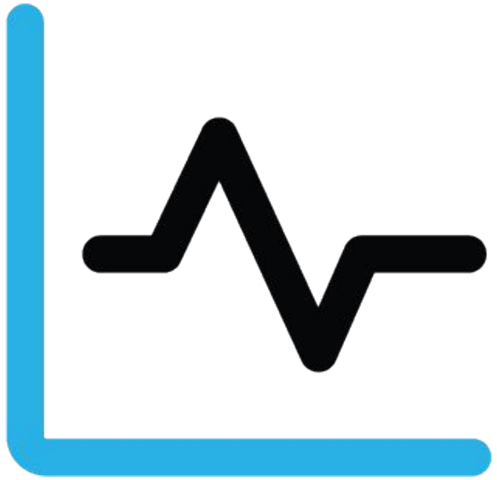
REVIEW

In the last lesson, we discussed that a database is an organized structure where data is stored.

There are many different databases, such as hierarchical, network, relational, non-relational, and object-oriented databases.

DATA ANALYSIS

Data analysis is the journey of collecting, modeling, and analyzing data to gain insights that support decision-making or change.



DATA ANALYSIS STEPS

1. Define the goal.
2. Collect the data.
3. Clean and process the data.
4. Analyze the data.
5. Draw conclusions.

DEFINE THE GOAL

The first step in data analysis is to **determine the goals of the analysis.**

Consider:

- What problem are we trying to solve?
- What are the business's expectations?

COLLECT DATA

The second step of data analysis is **collecting the data needed to meet the goals of the project.**

You may need to ask:

- Where is the data located?
- How will the data be transferred or received?

CLEAN AND PROCESS DATA

A large portion of the analytic process is cleaning and processing data.

Proper preprocessing is essential for accurate results.

ANALYZE THE DATA

Data professionals analyze data to try to answer difficult business questions.

Data analysis can be further broken down into two categories:

1. Initial data analysis
2. Primary data analysis

INITIAL DATA ANALYSIS

Initial data analysis **does not aim to answer a specific question.**

Initial data analysis sets out to explore:

- Quality of the data
- Quality of the measurements

PRIMARY DATA ANALYSIS

Primary data analysis is **aimed at answering the predefined business questions.**

The two categories of primary data analysis are:

- Confirmatory
- Exploratory

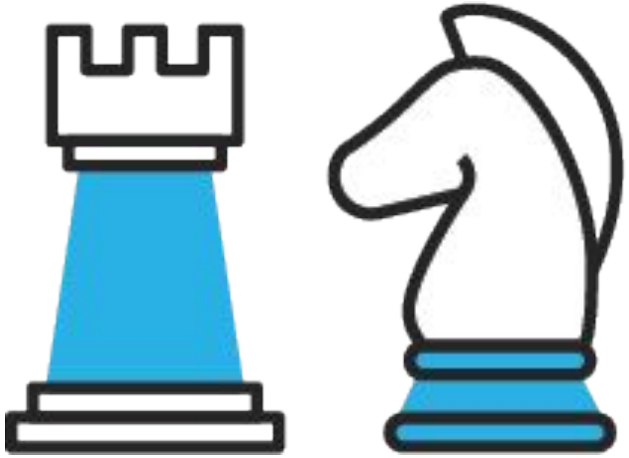
DRAW CONCLUSIONS

- Conclusions come from important data points and insights.
- Data professionals act as the bridge between the data and other business professionals.
- Revisit the initial goals of the analysis.

ACTIVITY: MIGUEL'S LIBRARY

In this activity, you will help Miguel decide how to spend his grant money to buy new books and update his library's book collection.

You will need the 1.5.3 Activity document and Excel (1.5.3 Activity Data.csv) to get started.



BREAKOUT ROOMS

REVIEW AND WRAP-UP

Today you learned about:

- The characteristics of structured, semi-structured, and unstructured data
- The functions of a database and how a database management system works
- The five steps of the analytic process

YOUR TAKE

- Reflect on what you have learned so far.
- Share key takeaways.

QUESTIONS?

NEXT STEPS

☑ Assigned Activities

☑

☑