

BIT34503 Data Science

CHAPTER 3: DEALING WITH DATABASES

3.0 DEALING WITH DATABASES

3.1 Types of Databases

3.2 Relational Databases

3.3 Structured Query language (SQL)


3.3.1 Performing CRUD (Create, Retrieve, Update, Delete)

3.3.2 Designing a real-world database

3.3.3 Normalizing a table

3.4 NoSQL

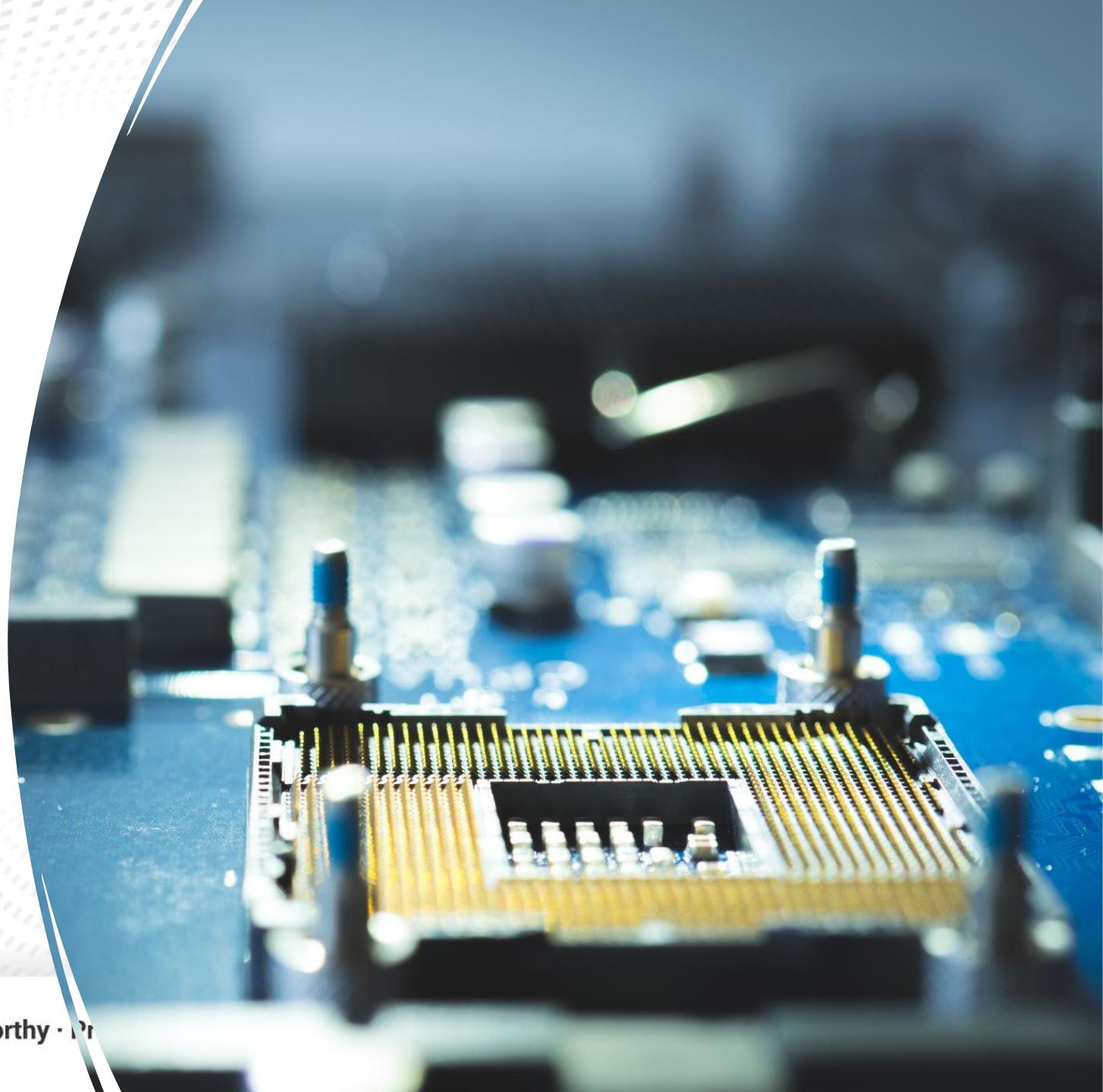
3.5 Hybrid database

A faint, light gray background image showing a network of interconnected nodes and lines, resembling a molecular structure or a data network.

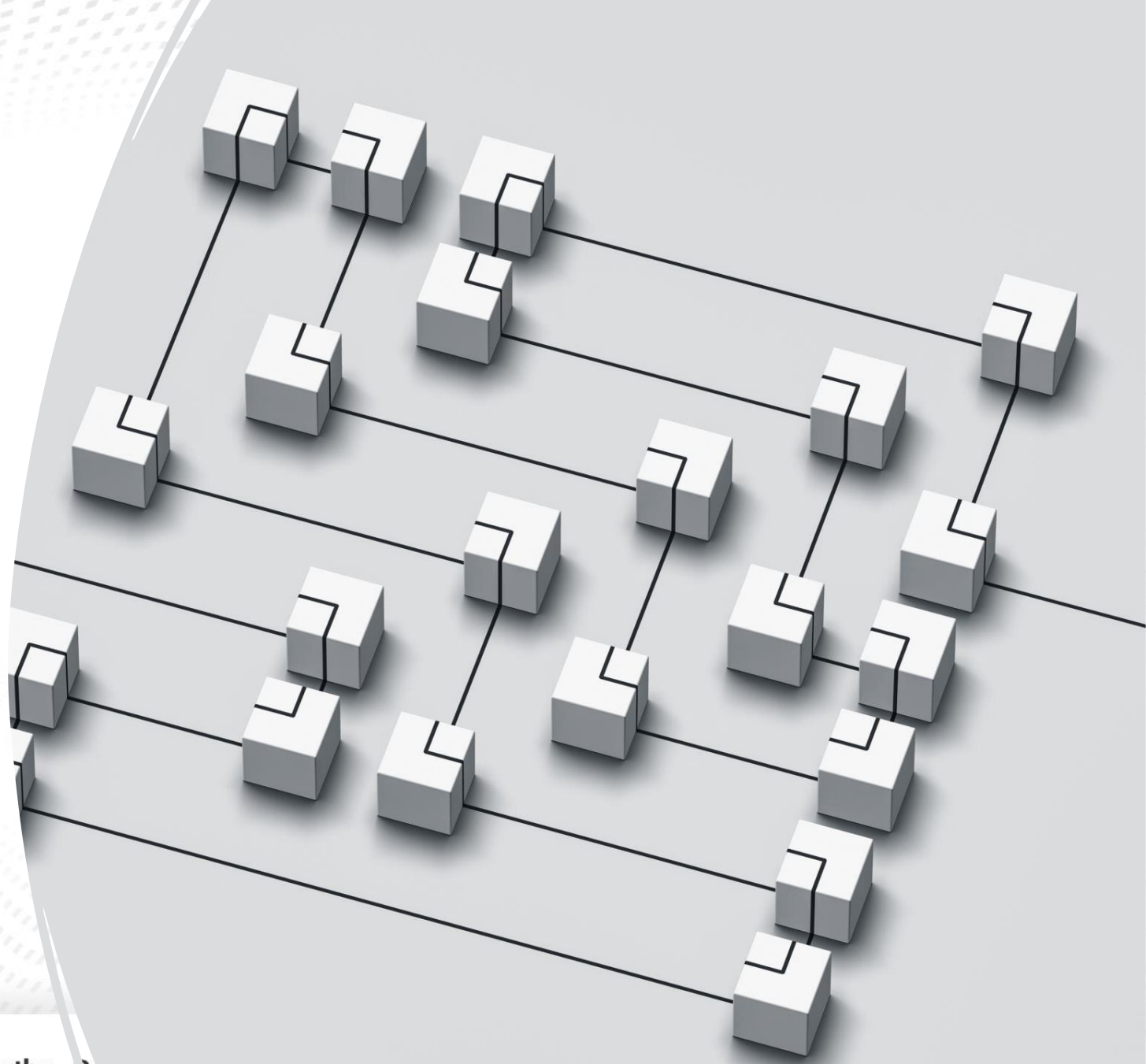
Data science is one of the fast-growing fields

Data science is all about data, collecting it, cleaning it, analyzing it, visualizing it, and using it to make our life better.

- Handling large amounts of data can be a challenging task for data scientists. Most of the time, that data we need to process and analyze is much larger than the capacity of our devices (the size of the RAM). Storing the information on the hard-drive might cause our code to be much slower.



- A database is defined as a structured set of data held in a computer's memory or on the cloud that is accessible in various ways.
- As a data scientist, you will need to design, create, and interact with databases on most of the projects you will work on. Sometimes you will need to create everything from scratch, while at other times, you will just need to know how to communicate with an already existing database.



Why use databases?

Data surround us; everything we use in our daily life is based on massive amounts of data. You turn on Netflix, it suggests what you should watch next, based on your previous selections. You open the Spotify app; it tells you to want songs you might like based on your preferences.

Collecting and analyzing data is one of the ways to personalize the experience of every one of us. It's a way to build one product that can fit everyone.



- Databases make structured storage secure, efficient, and fast. They provide a framework for how the data should be stored, structured, and retrieved. Having databases saves you the hassle of needing to figure out what to do with your data in every new project.

TYPES OF DATABASES

Relational
Databases

PostgreSQL
MySQL
Db2

NoSQL
Databases

RabbitMQ
MongoDB
JanusGraph

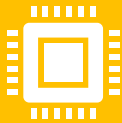
3.2 Relational databases



In a relational database, the data is organized and stored into tables that can be linked to each other use some relation. For example, an airline company can have a table of passengers for all flights, and another for passengers on a specific flight. A flight code can connect these two tables.



This ability to have connected tables allows us — as developers and data scientists — to understand better the relation between the different elements of the table. Understanding the relationship can give us hints and insight that will make the process of analyzing and visualizing the data an easier task.

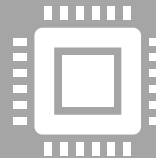


The way to communicate and interact with relational databases is through using the SQL language.

3.3 Structured Query Language (SQL)

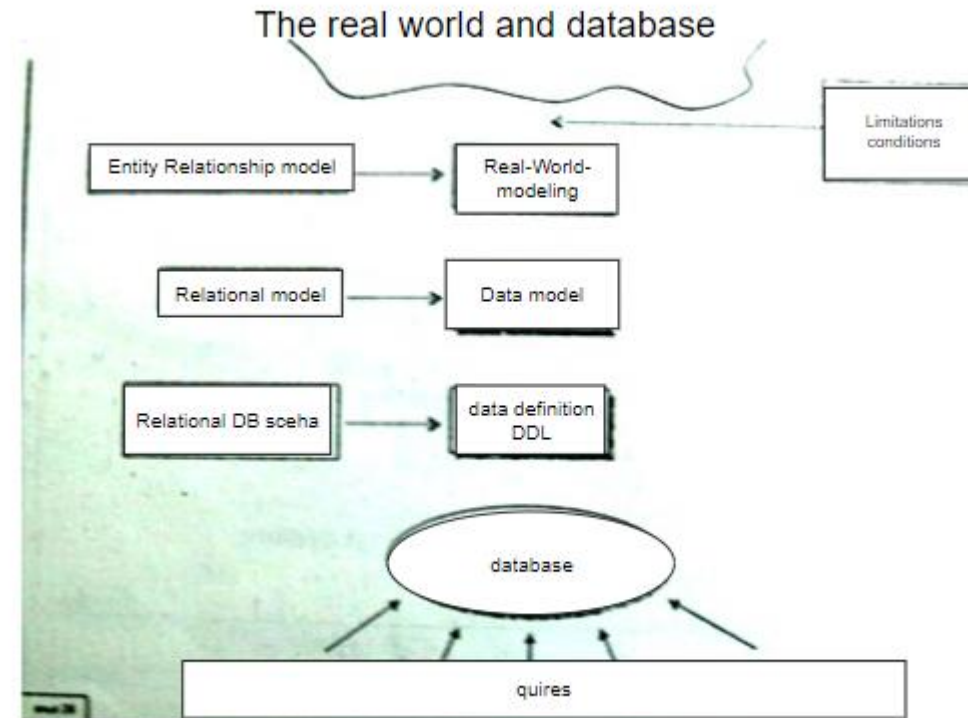


SQL is a powerful programming language used to manipulate data in a relational database management system (RDBMS). SQL is relatively easy, yet so powerful and efficient. Developers and data scientists use SQL to add, delete, update, or perform specific operation on a relational database.



SQL is not just for performing simple operations on databases; it can also be used to design databases or perform some analytics of the data stored.

CRUD (create, retrieve, update, delete) is an acronym that represents the four functions used to handle stored data in database applications. Create allows new database records to be made. Retrieve is for searching and reading data. Update permits users to change existing records.



- **Describing real-world data**
 - How can a user describe a real-world (e.g. university) in terms of data to be stored?
 - Understanding what data to be stored
 - What applications to be built on top of it?
 - What operations?
 - Study of the current operating environment and how it is expected to change?
 - What factors must be considered in deciding how to organize the stored data
 - Analysis of any available documentation on existing applications that are expected to be replaced or complemented by the DB.



Why we use normalization in data science?



Normalization avoids raw data and various problems of datasets by creating new values and maintaining general distribution as well as a ratio in data. Further, it also improves the performance and accuracy of machine learning models using various techniques and algorithms.

Non-relational databases, also known as NoSQL databases. These databases are those that connect the information stored in them by categories rather than relations.

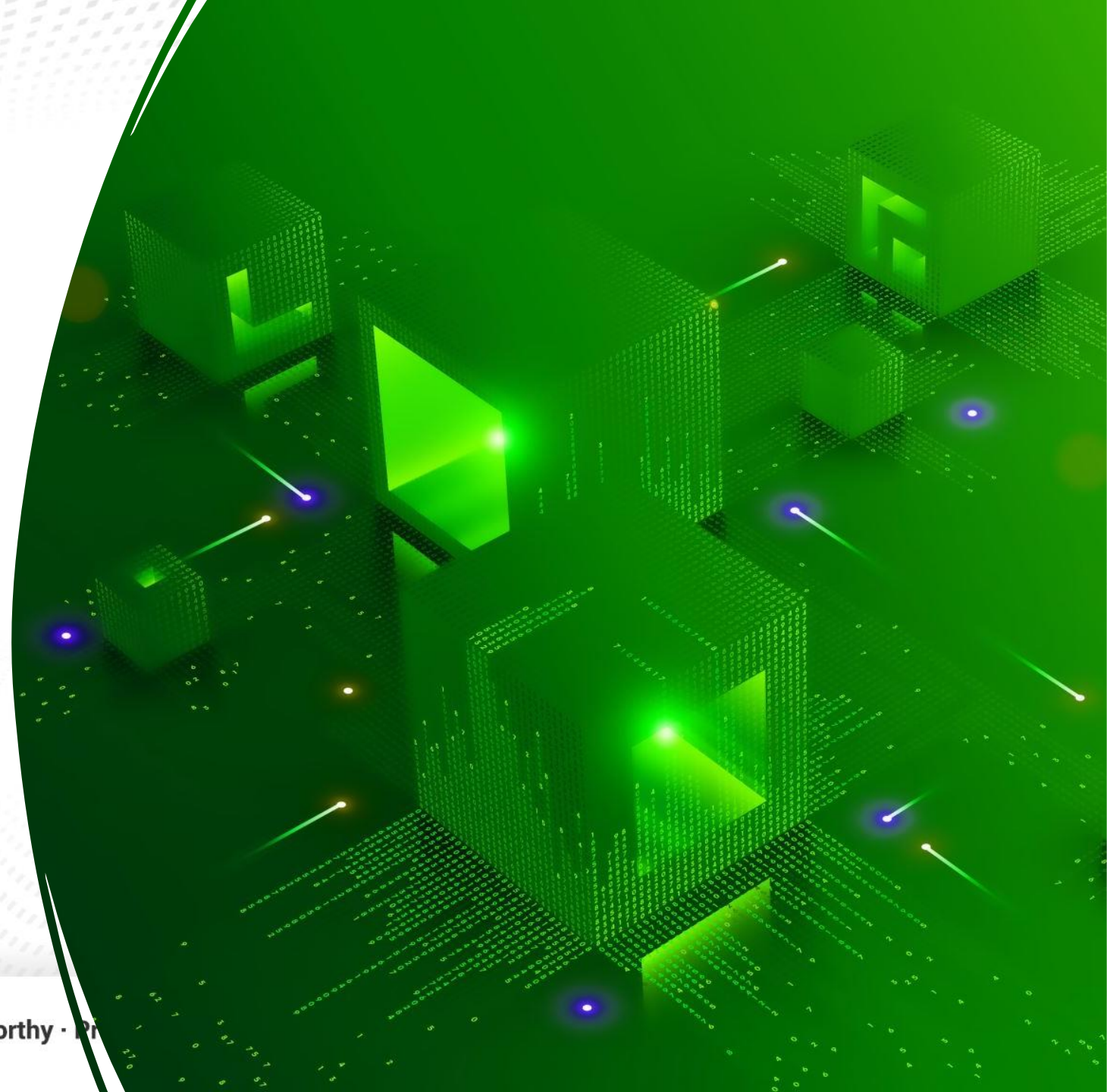
The most popular form of the NoSQL database is key-value pairs, which you can think about the same you do a Python dictionary. Keys have to be unique, as long as they are, a key-value pair can store all the relations in one document.

Relational databases use tables as their core storing unit. A table in a database consists of a collection of rows and columns, and you can connect several tables using relations. In NoSQL, however, the data is stored on document-like storage. You can still perform all everyday tasks, such as add, delete, update your data as long as you know how the document is structured.

3.5 Hybrid database

Conclusion

- Data is the most crucial part of data science; you can't have data science without data. Designing, creating, and communicating with databases is essential for any data scientist to grow her/his career and enrich their knowledge-base.
- Databases are a vast and broad field



References

- <https://towardsdatascience.com/databases-101-introduction-to-databases-for-data-scientists-ee18c9f0785d>
- <https://www.w3schools.com/python/>
- <https://www.analyticsvidhya.com/blog/2021/06/sql-for-data-science-a-beginners-guide/>





Thank you



Global Technopreneur
University 2030

Trustworthy · Professional · Innovative

