**Languages of Data Science**

For users who are just starting on their data science journey, the range of programming languages can be overwhelming. So, which language should you learn first? This module will bring awareness about the criteria that would determine which language you should learn. You will learn the benefits of Python, R, SQL, and other common languages such as Java, Scala, C++, JavaScript, and Julia. You will explore how you can use these languages in Data Science. You will also look at some sites to locate more information about the languages.

**Learning Objectives**

* Identify the criteria and roles for determining the language to learn.
* Identify the users and benefits of Python.
* Identify the users and benefits of R language.
* Define SQL elements and list its benefits.
* Review languages such as Java, Scala, C++, JavaScript, and Julia.
* List the global communities for connecting with other users.

# **Languages of Data Science**

choosing a programming language for data science.

Choosing the Right Programming Language for Data Science When starting your journey in data science, one of the first decisions you'll face is which programming language to learn. Think of it like picking the right tool for a job. Just as a carpenter might choose a hammer for nails and a saw for cutting wood, in data science, the language you choose depends on the specific problems you want to solve and the tasks you need to accomplish.

For beginners, popular choices include Python, R, and SQL. Python is like a Swiss Army knife—versatile and great for many tasks. R is fantastic for statistical analysis, while SQL is essential for managing and querying databases. Each language has its strengths, so consider what you want to do. For example, if you're interested in analyzing data trends, R might be your best bet. If you're looking to manipulate large datasets, SQL could be the way to go.

**Choosing a Programming Language for Data Science:**

* Selecting a programming language depends on your specific needs and the problems you want to solve.
* **Popular languages include:**
  + **Python:** Versatile and widely used for various tasks.
  + **R:** Excellent for statistical analysis.
  + **SQL:** Essential for managing and querying databases.
* The choice of language should align with your goals, the role you want to pursue, and the type of data you will work with.

# **Introduction to Python**

* **Users of Python**: Python is widely used by data professionals, with over 80% reporting regular use. It's suitable for both beginners and experienced programmers due to its clear syntax.
* **Benefits**: Python allows for writing less code to achieve the same results as other programming languages, making it efficient for various applications like data science, AI, web development, and more.
* **Community and Inclusion**: The Python community promotes diversity and inclusion, with initiatives like PyLadies supporting women in tech. The Python Software Foundation enforces a code of conduct to ensure a safe and welcoming environment.

# **Introduction to R**

**Open Source vs. Free Software** in simple terms.

Open source and free software both allow people to use software without paying for it, but they have different focuses. Open source software is more about making the source code available for anyone to see and modify, which encourages collaboration and innovation. Think of it like a community garden where everyone can plant, grow, and share their fruits and vegetables. On the other hand, free software emphasizes the freedom to use, modify, and share the software based on a set of values, like ensuring that everyone has access to the tools they need. Imagine it as a public library where everyone can borrow books, but they can also write their own stories and add them to the collection.

In summary:

* **Open Source**: Focuses on collaboration and making the code available.
* **Free Software**: Focuses on the freedom to use and share based on values.

Summary of video

* **Comparison of Open Source and Free Software**:
  + Both are free to use and support collaboration.
  + Open source is more business-focused, while free software emphasizes user freedoms and values.
* **Why Learn R?**:
  + R is free software, allowing for private, commercial, and public use.
  + It is widely used by statisticians, mathematicians, and data miners for statistical software, graphing, and data analysis.
  + R's array-oriented syntax makes it easier for beginners to translate math into code.
* **Popularity and Community**:
  + R is popular in academia and used by major companies like IBM, Google, and Facebook.
  + There are over 15,000 publicly released packages for complex data analysis.
  + R integrates well with other programming languages and has strong object-oriented programming features.
* **Global Communities**:
  + Communities like useR, WhyR, and R-ladies help connect R users worldwide.

This video highlights the importance of R in data science and its supportive community. If you have any specific questions or need further details, let me know!

# **Introduction to SQL**

concept of SQL and relational databases in simple terms.

**Understanding SQL and Relational Databases** SQL, which stands for Structured Query Language, is a special language used to communicate with databases. Think of it as a way to ask questions or give instructions to a library filled with books (the database). Each book represents a collection of information organized in tables, much like how a library organizes its books on shelves. These tables have rows (like individual books) and columns (like the information about each book, such as title, author, and genre). SQL helps you find, add, or change information in these tables easily.

**An Engaging Analogy** Imagine you are at a restaurant, and the menu is like a database. When you want to order food, you use a specific language to communicate with the waiter (SQL). You might say, "I would like to see the pasta options," which is similar to querying the database to get specific information. The waiter then brings you the pasta section of the menu (the relevant data). Just like that, SQL allows you to interact with the database to get the information you need without having to sift through everything manually.

summary of the video on SQL and relational databases:

* **SQL Overview**: SQL stands for Structured Query Language, and it is a non-procedural language used for querying and managing data in relational databases.
* **Relational Databases**: These databases consist of two-dimensional tables, similar to datasets or Excel spreadsheets, organized into rows and columns.
* **Key Features**:
  + SQL allows direct access to data without needing to copy it, speeding up workflows.
  + It acts as an interpreter between users and the database.
  + SQL is standardized by the American National Standards Institute (ANSI), making it applicable across various database systems.
* **Common SQL Databases**: Examples include MySQL, PostgreSQL, Oracle, and Microsoft SQL Server.
* **Learning SQL**: Focusing on a specific relational database is recommended for effective learning.

This video emphasizes the importance of SQL in data science and its widespread use in various job roles.

# **Other Languages for Data Science**

"Other Languages for Data Science." Here’s a simplified explanation of the key points:

1. **Programming Languages**: The video discusses several programming languages used in data science, including:
   * **Java**: A widely used language in businesses, known for its speed and ability to handle large data.
   * **Scala**: A language that combines features of Java and functional programming, often used with Apache Spark for big data processing.
   * **C++**: Known for its speed and control, it's used in applications like TensorFlow for deep learning.
   * **JavaScript**: Commonly used for web development, it also has tools like TensorFlow.js for machine learning.
   * **Julia**: A newer language designed for high-performance numerical analysis, making it fast for data science tasks.
2. **Applications**: Each language has specific tools and libraries that help in data science tasks, such as data analysis, machine learning, and big data management.

# **Summary**

**Module 2 Summary**

Congratulations! You have completed this module. At this point in the course, you know:

* You should select a language to learn depending on your needs, the problems you are trying to solve, and whom you are solving them for.
* The popular languages are Python, R, SQL, Scala, Java, C++, and Julia.
* For data science, you can use Python's scientific computing libraries like Pandas, NumPy, SciPy, and Matplotlib.
* Python can also be used for Natural Language Processing (NLP) using the Natural Language Toolkit (NLTK).
* Python is open source, and R is free software.
* R language’s array-oriented syntax makes it easier to translate from math to code for learners with no or minimal programming background.
* SQL is different from other software development languages because it is a non-procedural language.
* SQL was designed for managing data in relational databases.
* If you learn SQL and use it with one database, you can apply your SQL knowledge with many other databases easily.
* Data science tools built with Java include Weka, Java-ML, Apache MLlib, and Deeplearning4.
* For data science, popular program built with Scala is Apache Spark which includes Shark, MLlib, GraphX, and Spark Streaming.
* Programs built for Data Science with JavaScript include TensorFlow.js and R-js.
* One great application of Julia for Data Science is JuliaDB.