**Languages of Data Science**

For users who are just starting on their data science journey, the range of available technical options can be overwhelming. There are literally thousands of different programming languages with their own strengths and weaknesses. So, which language should you learn first? The answer to that question depends largely on your needs, the problems you are trying to solve, and who you are solving them for. First and foremost, we recommend considering the Python, R, and SQL languages. However, other languages such as Scala, Java, C++, and Julia with specific features are also popular. JavaScript, PHP, Go, Ruby, and Visual Basic all have their own unique use cases as well. The language you choose to learn will depend on the things you need to accomplish and the problems you need to solve. The problems can be related to the company you work for, what role you have, and the age of your existing application. So, what problems do you need to solve? We’ll have a better insight into the answers to this question as we learn more about the popular languages in the data science industry. There are many roles available for people who are interested in getting involved in data science, and they include: Business Analyst, Database Engineer, Data Analyst, Data Engineer, Data Scientist, Research Scientist, Software Engineer, Statistician, Product Manager, Project Manager, and so on. In this video, you learned that you should select a language to learn depending on your needs, the problems you are trying to solve, and who you are solving them for. The popular languages are Python, R, SQL, Scala, Java, C++, and Julia. JavaScript, PHP, Go, Ruby, and Visual Basic all have their own unique use cases as well. The problems you need to solve can be related to your company, role, and age of the existing application.

**Introduction to Python**

Python is a powerhouse of a language. It is the most widely used and most popular programming language used in the data science industry. According to the 2019 Kaggle Data Science and Machine Learning Survey, ¾ of the over 10,000 respondents worldwide reported that they use Python regularly. Glassdoor reported that in 2019 more than 75% of data science positions listed included Python in their job descriptions. When asked which language an aspiring data scientist should learn first, most data scientists say Python. Let’s start with the people who use Python. If you already know how to program, then Python is great for you because it uses clear and readable syntax. You can develop the same programs from other languages with lesser code using Python. For beginners, Python is a good language to start with because of the huge global community and wealth of documentation. Several different surveys done in 2019 established that over 80% of data professionals use Python worldwide. Python is useful in many areas including data science, AI and machine learning, web development, and Internet of Things (IoT) devices, like the Raspberry Pi. Large organizations that heavily use python include IBM, Wikipedia, Google, Yahoo!, CERN, NASA, Facebook, Amazon, Instagram, Spotify, and Reddit. Python is widely supported by a global community and shepherded by the Python Software Foundation. Python is a high-level, general-purpose programming language that can be applied to many different classes of problems. It has a large, standard library that provides tools suited to many different tasks including but not limited to Databases, Automation, Web scraping, Text processing, Image processing, Machine learning, and Data analytics. For data science, you can use Python's scientific computing libraries like Pandas, NumPy, SciPy, and Matplotlib. For artificial intelligence, it has TensorFlow, PyTorch, Keras, and Scikit-learn. Python can also be used for Natural Language Processing (NLP) using the Natural Language Toolkit (NLTK). Another great selling point for Python is that the Python community has a well-documented history of paving the way for diversity and inclusion efforts in the tech industry as a whole. The Python language has a code of conduct executed by the Python Software Foundation that seeks to ensure safety and inclusion for all, in both online and in-person Python communities. Communities like PyLadies seek to create spaces for people interested in learning Python in safe and inclusive environments. PyLadies is an international mentorship group with a focus on helping more women become active participants and leaders in the Python open-source community.

In this video, you learned that Python uses clear and readable syntax. Python has a huge global community and a wealth of documentation. 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 community has a well-documented history of paving the way for diversity and inclusion efforts in the tech industry as a whole.

**Introduction to R Language**

According to the results of the 2019 Kaggle Data Science survey, which had over ten thousand respondents worldwide, learning three languages can earn you an increment! And R offers many possibilities. Now, Python is open source, while R is free software. Let us compare the open source and free software. The similarities are that both are free to use. Both commonly refer to the same set of licenses. For example, many open-source projects use the General Public License (GNU). Both support collaboration. And in many cases, these terms can be used interchangeably (but not all). Now, let’s discuss the differences between open source and free software. The Open-Source Initiative (OSI) champions open source, while the Free Software Foundation (FSF) defines free software. Open source is more business focused, while free software is more focused on a set of values. So, why R? You should learn R because it is free software. You can use the language in the same way that you contribute to open source. In addition, it allows for private use, commercial use, and public collaboration. R is another language supported by a wide global community of people who want to use the language to solve big problems. Statisticians, mathematicians, and data miners use R to develop statistical software, graphing, and data analysis. R Language's array-oriented syntax makes it easier to translate from math to code for learners with no or minimal programming background. According to Kaggle’s Data Science and Machine Learning survey, most programmers learn R after a few years into their data science career. And R is mostly popular in academia. In addition, companies that use R include IBM, Google, Facebook, Microsoft, Bank of America, Ford, TechCrunch, Uber, and Trulia. R has become the world’s largest repository of statistical knowledge. As of 2018, R has more than 15,000 publicly released packages making it possible to conduct complex exploratory data analysis. R integrates well with other computer languages like C++, Java, C, .Net, and Python. Using R, common mathematical operations like matrix multiplication give immediate results. And R has stronger object-oriented programming facilities than most statistical computing languages.

Now, there are many ways to connect with other R users around the globe. For connecting to other users, you can use communities such as useR, WhyR, SatRdays, and R-ladies. And in addition, you can check out the R project website for R conferences and events.

In this video, you learned that: The Open-Source Initiative (OSI) champions open source, while the Free Software Foundation (FSF) defines free software. 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 none or minimal programming background. And R has become the world’s largest repository of statistical knowledge.

**Introduction to SQL**

SQL is a bit different than the other languages. Officially it is pronounced as “ess cue el” though some call it “sequel”. And while the acronym stands for “Structured Query Language”, many people consider SQL different from other software development languages because it is a non-procedural language. Its scope is limited to querying and managing data. While it is not a “Data Science” language, data scientists regularly use it because it is simple and powerful! Some other facts about SQL are that it is older than python and R by about 20 years. It first appeared in 1974 and was developed at IBM! This language is useful in handling structured data, which is the data incorporating relations among entities and variables. SQL was designed for managing data in relational databases. Here you can see a diagram showing the general structure of a relational database. A relational database is formed by collections of two-dimensional tables, for example, Datasets and Excel Spreadsheets. Each of these tables is then formed by a fixed number of columns and any possible number of rows. However, although SQL was originally developed for use with relational databases, because of its pervasiveness and ease of use, SQL interfaces have also been developed for many NoSQL and big data repositories. The SQL language is subdivided into several language elements, including: Clauses, Expressions, Predicates, Queries, and Statements. So, what makes SQL great? Knowing SQL will help you get many different jobs in data science, such as a business and data analyst. This knowledge is also a must in data engineering. When performing operations with SQL, the data is accessed directly, without needing to copy the data separately, which can considerably speed up workflow executions. SQL behaves like an interpreter between you and the database. SQL is an American National Standards Institute (or ANSI) standard, which means if you learn SQL and use it with one database, you can apply your SQL knowledge to many other databases easily. Now, many different SQL databases are available, including the following: MySQL, IBM DB2, PostgreSQL, Apache Open Office Base, SQLite, Oracle, MariaDB, Microsoft SQL Server, and more. The syntax of the SQL you write may change based on the relational database management system you are using. If you want to learn SQL, you should focus on a specific relational database and then plug into the community for that specific platform. In addition, there are many available great introductory courses on SQL! In this video, you learned that: SQL is different from other software development languages because it is a non-procedural language. SQL’s scope is limited to querying and managing data. SQL was designed for managing data in relational databases. SQL behaves like an interpreter between you and the database. And if you learn SQL and use it with one database, you can apply your SQL knowledge to many other databases easily.

**Other Languages for Data Science**

Previously, we reviewed Python, R, and SQL. In this lesson, we will review some other languages that have compelling use cases for data science. Scala, Java, C++, and Julia are probably the most traditional data science languages. However, JavaScript, PHP, Go, Ruby, Visual Basic and many others have found their place in the data science community. Let us go through some notable highlights about a few of them. Java is a general-purpose tried and tested object-oriented programming language. It has huge adoption in the enterprise space and was designed to be fast and scalable. Java applications are compiled to bytecode and run on the Java Virtual Machine or JVM. Some notable data science tools built with Java include: Weka for data mining, Java-ML for machine learning, Apache MLlib makes machine learning scalable, and Deeplearning4 for deep learning. Hadoop is another application of Java which manages data processing and storage for big data applications running in clustered systems. Scala is a general-purpose programming language that provides support for functional programming and is a strong static type system. The Scala language was constructed to address the shortcomings of Java. It is also inter-operable with Java as it runs on the JVM. The name Scala is a combination of scalable and language. This language is designed to evolve with the requirements of its users. For data science, the most popular program built with Scala is Apache Spark. Spark is a fast and general-purpose cluster computing system that provides APIs, which make parallel jobs easy to write. It has an optimized engine that supports general computation graphs. Spark includes Shark, which is a query engine, MLlib for machine learning, GraphX for graph processing, and Spark Streaming. It was designed to be faster than Hadoop. C++ is a general-purpose programming language. It is an extension of the C programming language or "C with Classes.” C++ improves processing speed, enables system programming, and provides broader control over the software application. Many organizations that use Python or other high-level languages for data analysis and exploratory tasks rely on C++ to develop programs that feed data to customers in real-time. For data science, TensorFlow is a popular Deep Learning library for dataflow that was built with C++. Although C++ is the foundation of TensorFlow, it runs on a python interface, so users don’t require the knowledge of C++ to run it. MongoDB is a NoSQL database for big data management that was built with C++. Caffe is a deep learning algorithm repository built with C++ with Python and Matlab bindings. A core technology for the world wide web, JavaScript is a general-purpose language that extended beyond the browser with the creation of Node.js and other server-side approaches. Javascript is NOT related to the Java language. For Data Science, undoubtedly TensorFlow.js is the most popular implementation. TensorFlow.js makes machine learning and deep learning possible in Node.js as well as in the browser. TensorFlow.js was also adopted by other open-source libraries including brain.js and machinelearn.js. Another implementation of JavaScript for Data Science is R-js. The project R-js has re-written linear algebra specifications from the R Language into typescript. This sets the foundation for future projects to implement more powerful math base frameworks like Numpy and SciPy of Python. Typescript is a superset of JavaScript. Finally, Julia was designed at MIT for high-performance numerical analysis and computational science. Julia provides speedy development like Python or R, while producing programs that run as fast as C or Fortran programs. It’s compiled which means that Julia code is executed directly on the processor as executable code. It calls C, Go, Java, MATLAB, R, Fortran, and Python libraries, and has refined parallelism. Julia as a language is only 8 years old, written in 2012, but there is a lot of promise for its future impact on the data science industry. One great application of Julia for Data Science is JuliaDB, which is a package for working with large persistent data sets. In this video, you learned that Data science tools built with Java include Weka, Java-ML, Apache MLlib, and Deeplearning4. For data science, a popular program built with Scala is Apache Spark, that includes Shark, MLlib, GraphX, and Spark Streaming. For data science, TensorFlow, MongoDB and Caffe were built with C++. Programs built for Data Science with JavaScript include TensorFlow.js and R-js. One great application of Julia for Data Science is JuliaDB.