PS2

Colton Steele

February 7, 2023

1 Main Tools of a Data Scientist

The first main section of tools for Data Scientists is measurement. Just like any quantitative discipline, measurement often determines the importance or significance of our takeaways.

Another section of tools for data scientists are the statistical programming languages we use which include R, Python, and Julia. This is not an exhaustive list, but these are the ones that are most commonly used in the data science profession. These languages are scripted languages which are known and created to be easier for humans to interpret and use as compared to compiled languages such as C, C++, or Fortran. Another useful language for data scientists is SQL, or structured query language which is a language primarily used to pull data from databases.

Another great tool for data scientists is the ability to scrape web data. Data is not always available in easy to access datasets, so having the ability to scrape web pages to gather data is instrumental towards accomplishing some tasks. The most common ways this is done is through the use of an API, or application program interface, or through downloading HTML files and parsing their text.

Likely some of the most important tools of a data scientist are data visualization tools such as Tableau or built in visualization functions in programming languages such as ggplot in R. As the common phrase goes, "a picture is worth a thousand words". It is far easier to convey trends or the distribution of data in a visualization rather than purely through numbers or words.

One swath of useful tools for data scientists that pertains to big data is Resilient Distributed Datasets (RDDs). RDD's can be accessed using a cluster of computers and software such as Hadoop or Spark, which separates the data set into manageable chunks.

Models are another important statistical tool for data scientists. Models are useful in accomplishing three objectives: using data to test theories, using the data to predict behavior, and using the data to explain behavior.