PS2 Felkner

Beth Felkner

February 2024

1 Tools of a Data Scientist

- There are five main tools of a data scientist that we talked about: measurement, statistical programming languages, web scraping, techniques for handling big data sets, visualization tools.

- Measurement is how insights/policy can be constructed. In the case of measuring a baseball hitter's effectiveness, you have to decide what metric you want to use: AVG, OBP, RBI, strikeout rate, walk rate, etc.

- The three main languages we will use in this class are R, Python, and Julia, but there are also others like Stata etc. R, Python, and Julia are all scripted languages which means they are easier for humans to read but sacrifice performance efficiency.

- Webscraping refers to collecting publicly accessible data from webpages. You can do this either through or an API, or by directly parsing the HTML code.

- To handle large data sets, data scientists use Resilient Distributed Datasets. This means that a large data set is broken up into many smaller data sets and each set is operated on in parallel, using a program like Spark. SQL is the language most commonly used for management of large databases.

- Data visualization allows you to look at the data in multiple dimensions and easily identify outliers. The main packages for this are ggplot2, matplotlib, Plots.jl in their respective languages. You can also use a software like Tableau, or Power BI which my company loves to use.