List of Tools for Data Scientists

Daniel Carpenter

January 2020

1 Measurement

- Measuring information in the form of data

2 Statistical Programming Languages

- R
- Julia
- Python

3 Web Scraping

- Pulling information from webpage
- Use CTRL + 'u' to show page's code
- API's can pull data, but sometimes are limited by host
- Parsing is an alternate when API's are not allowed, but IP tracking is present.
- Languages: R, Julia, Python

4 Handling large data sets

- RDD: Resilient Distributed Datasets - uses Hadoop or Spark to manage dataset.
- SQL = Structured Query Language to handle dataset manipulation and storage.

5 Visualization

- ggplot2 (R)
- matplotlib (Python)
- Plots.jl (Julia)
- Tableau: popular software; does not allow good integration for coding.

6 Modeling

- Use the data to test theories. For example, Amazon may wonder ”Are women more likely than men to subscribe to Prime?” Without data this is simply a ”hunch” or a belief.

- Use the data to predict behavior. For example, many companies need to optimize their inventory management so that the right amount of inventory is in the right stores at the right time.

- Use the data to explain behavior. This is a bit more difficult, because it goes beyond prediction. Once a company or government knows the ”why” behind consumer/citizen behavior, it can optimize its behavior in response. (Note here that the ”why” implies a causal relationship—this requires having a particular type of data or requires making additional assumptions and making use of additional statistical methods than are used for prediction.)