PS2

Steven Plaisance

February 2023

Main Tools of a Data Scientist

- **Measurement -** This is the backbone of quantitative analysis. One must understand measurement error and have a good instinct for engineering and adjusting measurements while minimizing measurement error within their field of study.
- **Programming Languages -** The most common are R, Python and Julia. Code-based analysis results in work that is more reproducible and easier to automate.
- **Web Scraping -** The internet is the most vast and varied data source in the history of mankind. Web scraping collects and converts web data from an unmalleable format (html) to a more usable one (ie. dataframes in R)
- **APIs -** If web scraping is like ordering take-out, then APIs are like getting food delivered right to your door. Requests are made directly from your machine to the provider's database rather than navigating to their webpage and parsing through their html.
- **Big Data Tools -** Some datasets are too large to be processed by a sinlge device. Tools such as Resilient Distributed Datasets (RDDs) solve this issue by dividing the data into smaller pieces and performing operations in parallel across multiple machines.
- **Modeling -** Often times, data scientists are asked to make predictions about future events using existing data. This process is known as modeling and can take many forms, such as regression. Each type of prediction requires a unique modeling approach.
- **Visualization -** Strong visualization is far-and-away the best method to bridge the gap between data scientists and decision-makers. Images are retained at a far greater rate than text, and visualizations can summarize millions of rows and columns into a single image.