PS2-Wardwell

lwardwell

February 2025

Data Science Tools

- **Measurement:** Any time we study something, we measure it. How we measure it effects the outcome.
- **Statistical Programming Languages:** Generally grouped into Compiled and Scripted Languages. Also separated into Open Source and Proprietary.
  - **Compiled -** Code is run all at once, more painful to write in, once compiled performance is better.
  - **Scripted -** Code can be run in pieces, more human readable, lower performance.
  - **Open Source -** Free to use and improved by the community (Examples: R, Python, Julia)
  - **Proprietary -** Cost to use and maintained by a company (Stata, Matlab, SAS)
- **Visualization Tools:** Usually included with the softwares noted above. Some are better than others at visualization.
- **Big Data Management Software:** Used to manage large datasets (Resilient Distributed Datasets)
- **Data Collection Tools:** For example, Webscraping - Using an application program interface (API) to download data or downloading HTML files and parsing the text to extract data.