Connor Gregorich-Trevor

Grinnell College
Mentors: Matthew Lau, Emery Boose, & Barbara Lerner
Group Project: Data Provenance in R

Data Provenance in R and Python Across Multiple Scripts

New research often involves extending a previous scientific study. To do this, a scientist may need to recreate and modify the procedure or methods of an existing study in order to obtain results. Extension is important, since it is one of the foundations for research and increasing scientific knowledge. Scientific studies can contain information that is crucial to their replication and extension, such as how the data in the study were collected, the conditions under which they were obtained, and what processes and transformations they went through. These details and the complete history of data is known as data provenance. Many studies lack information about how their data were collected or processed, yielding little in the way of provenance and making them very difficult to replicate or verify the reliability of. In my research, I worked with Emery Boose, Barbara Lerner, and Matthew Lau on RDataTracker and DDG Explorer. RDataTracker is a program which captures provenance of programs and analyses run in the R programming language, and DDG Explorer creates visualizations of the program flow. I modified these programs so that they can be used to track data through workflows composed of multiple R and Python scripts by matching input and output files. With this new feature, scientists will be able to gather provenance when using a combination of different languages to interact with their data, as well as focus on provenance across multiple scripts, leading to greater transparency and easier replication.