Jennifer Johnson

Middlebury College

Mentors: Matthew Lau, Emery Boose, & Barbara Lerner

Group Project: Data Provenance in R

Collecting Provenance in Python

The reproducibility of results is an essential quality of scientific research. Reproducibility allows researchers to more easily verify results and expand on each other's work. However, data analyses often lack complete and high quality documentation, making them incomprehensible to other researchers, and sometimes even to the author. It is easy to forget the logic of a complex analysis. Data provenance is a detailed record of the processes performed in an analysis and greatly improves its transparency. Tools that collect provenance could help researchers keep projects organized for collaborators, reviewers, and themselves. The package RDataTracker collects provenance for R scripts, and the Java application DDG Explorer displays it as a Data Derivation Graph. However, R is not the only language used for data analysis. Python and R have different uses and strengths in data analyses, and can complement each other when used together. The goal of this project was to collect provenance for Python scripts in a similar manner to RDataTracker. This feature relies on an existing tool for collecting Python provenance, noWorkflow, developed by researchers at the Universidade Federal Fluminense in Brazil and at New York University. Their goal was to provide provenance for projects that do not use a workflow system for organization. In this project, noWorkflow provenance was converted into a format that is compatible with DDGExplorer. Because of the popularity of Python, a tool that collects Python provenance could be a first step to increase awareness of provenance and the value of reproducibility in scientific research.