Katerina Poulos
Mount Holyoke College
Mentors: Emery Boose and Barbara Lerner
Group Project: Data Provenance in R

Incorporating data analysis workflow into data provenance to increase accessibility

Scientific progress requires the validation, authentication, and reproduction of results to uphold the quality of scientific research. One technique to make large datasets and complex data analyses more intelligible is to capture the data provenance, or the history of a digital object. The current tools developed by the Harvard Forest team to capture and analyze provenance are RDataTracker and DDGExplorer, respectively. These tools create and use a Data Derivation Graph (DDG), a metadata structure that captures data provenance. However, these DDGs may be difficult to comprehend for many R scripts as comprehensibility of a DDG may be inversely proportional to the size and complexity of the script. The objective of this project is to make provenance shown by the DDG more intelligible by breaking it down into multiple sections that reflect typical data analysis workflow. This automated system takes advantage of a common workflow of an R script developed by analyzing various R scripts from the Harvard Forest community. The current interpretation of the typical data analysis workflow was identified and consists of: initialization, data preparation, data analysis, plotting, and output. Building upon existing tools from RDataTracker, my program will provide a higher abstraction to R scripts with little to no user intervention by labelling the portions of the DDG that correspond to these different data analysis activities. By compartmentalizing provenance into comprehensible bits using lingo known by all scientists, a potentially large and complex DDG will be collapsed into more readily coherent sections which will increase accessibility.