Siqing Liu
Amherst College
Mentors: Emery Boose & Barbara Lerner
Group Project: Data Provenance in R

Improving RDataTracker accessibility and functionality

Scientists, in order to turn raw data into usable information, may perform complex manipulations using statistical and other software. Without proper documentation of these manipulations, other scientists can often find it impossible to reproduce and verify results. My mentors Barbara Lerner and Emery Boose are tackling this issue by developing a set of tools that can capture data transformation and history, or as it is more officially termed, data provenance. RDataTracker, our current tool for the R programming language, can automatically execute a R script (a program that can automate tasks) and generate a line-by-line visual representation of what the code does in a Data Derivation Graph, or DDG. Our main goals are extending the functionality of RDataTracker and making it more accessible to scientists. We extended RDataTracker to allow it to work directly with RMarkdown files, a file format that allows for easy formatting and publication of scripts. We also implemented a caching system. This is to save time for scientists who are changing or re-running a program, by saving data that is already calculated and only recalculating parts that have been edited. Overall, we hope to make RDataTracker an easy and useful tool for scientists, so that capturing data provenance is not a burden but an asset.