Joseph Wonsil

Carthage College
Mentors: Emery Boose, Elizabeth Fong, Barbara Lerner
Group Project: Data Provenance in R

Using Provenance to Make a Better Debugger

Provenance collection tools were created to increase reproducibility in science. RDataTracker is one such program created for provenance collection in R, a popular language for data analysis. Provenance holds a history of data, operations, and computing environment from the execution of a script. The existence of this record leads to transparency and reproducibility in science since an analysis is tracked through its execution. This study focuses on using provenance data in other novel ways. Provenance is, by nature, full of meaningful information about a script. We built a tool that can use it to highlight errors and irregularities. The result is a debugging package for R. We designed the package, provDebugR, to read in provenance and extract helpful data that is then presented to the scientist. Some unique features include tracking variable lineage, monitoring type changes, and searching for errors on Stack Overflow. We developed an interactive console interface similar to that of R's browser function to allow scientists to explore scripts in a more familiar way. They can step line-by-line through their code like a normal debugger; however, the scientist is not actually executing the code. Rather, they are stepping through the history of a previous execution stored in provenance. As a result, they can step forwards and backwards, and are not limited to moving one line at a time. It also provides the functionality to move immediately to any line number of their script. These features are only possible due to the provenance at the package's core.