Luis Perez
Harvard College
Mentors: Emery Boose and Barbara Lerner
Group Project: Retracing our steps in the analysis of data

Accessible and efficient data provenance in the R scripting environment

Scientific progress relies on the reproducibility of experiments and data analyses. However, as scientists explore increasingly complex problems, they often require the use of large data sets, complicated data-filling models, and complex analysis methods. Subsequently, the data's provenance — the transformation performed to achieve the reported results — becomes increasingly difficult to record and access, even by expert peers. A Data Derivation Graph (DDG), a structured record of the data manipulation process, is one solution for the problem of accessibility. However, the creation of complete DDGs often involves unfamiliar tools, additional data recording steps, and longer computation time, hindering wide scientific adoption. In our project, our team worked to improve the accessibility, correctness, and efficiency of RDataTracker, a provenance collection package for R. In order to maintain correctness throughout the development process, we created and utilized an automated system for software testing and performance data collection (using Apache Ant). We worked on features to facilitate and reduce the number of annotations needed for the creation of DDGs. Furthermore, we reduced annotation time by improving the effectiveness and correctness of RDataTracker's automatic DDG creation capabilities. We expect this additional ease of data collection to decrease computational performance; however, improvements in saving intermediate data to disk should counteract this performance hit. Through the above improvements in data collection and automation, we expect RDataTracker to become a more useful and accessible tool for scientists, making provenance collection second nature, an important first step in dealing with the increasing complexity inherent to scientific research.