Mayfly – Rapid, Accessible, Reproducible Research

Nelson Auner and Cody Buntain
ne.auner@gmail.com and cbuntain@cs.umd.edu

June 19, 2014

1 Introduction

In the midst of the current data explosion and the subsequent complexity of analysis pipelines, it is becoming increasingly difficult to reproduce scientific results in a manual fashion. These issues cumulate into a significant barrier to verification and validation, leading to a growth in publication and wide-spread acceptance of erroneous results with limited recourse for correction. The Open Science Data Cloud (OSDC) has significant potential to alleviate these issues with consistent and shared access to duplicate data stores, but it currently lacks straightforward methods for online publication or facilitating access to research results. Fortunately, as data science becomes more main stream, and barrier to entry comes down, many solutions for facilitating analytics like IPython's Notebook and R-Studio are now available. Additionally, the significant adoption of mobile devices has lead to a plethora of solutions for file sharing across multiple systems: Dropbox, Google Drive, Microsoft's Azure, and Apple's iCloud are all examples of such solutions. Integrating these analytics and storage solutions with the OSDC has real power to capitalize on the both platforms in a way that could cause real impact in the scientific community and lead to a better quality of research across the field. To that end, this project aimed to integrate Dropbox's sharing capabilities, tools for research presentation in Python and R, and the OSDC platform to support rapid, accessible, and reproducible research.

2 Technical Details

We achieved this project's goals by creating a small suite of tools for publishing research code and results online and for quickly installing and configuring existing software analytic packages. As mentioned, several popular programming languages for data analysis now have such analytic platforms specifically for supporting rapid and demonstrable code. Of particular interest here are IPython's Notebook environment and R's Knitr package, which enable simple authoring of HTML-based web reports for code and results. Our suite of tools automatically installs these environments as well as Dropbox's client on OSDC virtual machines in a nearly hands-free manner. After performing any sort of analysis in either IPython or R, the researcher can then leverage our Dropbox-based publication tool to convert their analysis workflows and results to simple HTML pages and automatically publish them directly through Dropbox's public folder API. The researcher is then given a URL, such as `https://dl.dropboxusercontent.com/u/66442241/kdd.html`, she can then share with colleagues for review.

3 Example Workflow

4 Conclusions