Domino: A Platform-as-a-Service for Industrialized Data Analysis

Nick Elprin[1*]

1. Domino Data Lab, Inc.
*Contact author: nick@dominodatalab.com

Keywords: Cloud computing, reproducibility, version control

As data volumes have increased and analytical techniques have become more sophisticated, the tools necessary to do industrialized analysis — analysis that is scalable, reproducible, and collaborative — have lagged in their ease of use. We have built Domino, a Platform-as-a-Service for data analysis, to equip a larger group of users with functionality that has typically been inaccessible to people without engineering abilities and/or a massive amount of time to set up infrastructure and plumbing.

Although Domino is language agnostic, it has particularly deep integration with *R*, including integration with RStudio and first-class support for packages from CRAN and other repositories.

Domino address three core areas of functionality:

(1) It lets you run your *R* code (or *Python*, *Julia*, *Matlab*, and more) in the cloud without any setup or configuration. Domino handles AMI and package management, job distribution and secure data transfer. It allows you to change your hardware with one-click, or to distribute your analysis across multiple machines.
(2) It automatically keeps a revisioned history of your project — code, data, and results — so you can browse and reproduce past work. Unlike traditional source control, Domino tracks large data files, and creates a first-class association between your results artifacts (e.g., charts) and the code/data that produced them.
(3) It facilitates collaboration so you can easily share results and co-author analyses.

In this talk we will demonstrate Domino's core functionality and describe its architecture, with an emphasis on the technical challenges involved in enabling reproducible work (e.g., an immutable, revisioned data store for large files). We will also describe some case studies highlighting how Domino is being used in the real world.