Reproducible Computing and Reporting in a Complex Software Environment

E. J. Harner[a], C. Grant[b], and M. Lilback[b]

[a] *West Virginia University,* [b] *Rc^{2}ai*

Statistical reports consisting of code-based text, data analyses, visualizations, and simulations can be 'containerized,' creating a transparent reproduction of the computational workflow that produced the results. With the advent of container platforms such as Docker and Singularity, and their integration with cloud technologies and standalone operating systems, entire project environments can be archived, providing later access to the exact data, methods, and configurations used to generate reports.

The code underlying the report, e.g., an R Markdown file, should be cloned from GitHub (or a similar repository) directly into the container to provide versioning as the document is iteratively edited. R Markdown is not only self documenting, but also supports several input languages and output formats. The report can be rendered directly in RStudio, if available, but in the end it should be automated through a 'makefile'. R packages and other artifacts needed for the analysis can be added to the running container as needed. A shell script can be used to push the resulting reproducible image to Docker Hub

At this point, we have been describing a single computational container, which supports document creation. More generally, the data tables should be stored in a containerized database (SQL, NoSQL, etc.), which supports versioning and thus reproducibility. Further, if a more powerful computational environment is required, e.g., a containerized Spark cluster potentially scaled with Kubernetes, could be used.

These concepts are illustrated by 'rspark', an experimental edge computing environment consisting of a compute and a database container capable of connecting to a containerized Spark cluster or, more generally, a cloud-based cluster. The compute container is based on R running on Ubuntu 18.04 from the Rocker Project.

Keywords: Reproducible report, Docker, R markdown, Postgres, Spark