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] *Rc²ai*

The canonical entity for reproducible computing and reporting is a Docker or Singularity computational container, which supports R Markdown. The image, which instantiates the computational container, should have all the components required to reproduce the report ranging from a stable release of Linux to the application software such as R and RStudio, versioned appropriately. The code underlying the report can be embedded in the image generated by a 'Dockerfile' or it can be cloned from GitHub or another repository. Likewise the data may be available within the container, but it is more likely imported. One possibility is to have the data within a database available through another container. In this case, the compute and the database containers can be built using docker-compose. Shell scripts can be used to push the images to Docker Hub.
The report itself is often generated from an 'Rmarkdown' file, which is not only self documenting but also supportive of several output formats. The report can be rendered directly if it is simple enough, but more generally it can be automated through a 'make' file. If a more powerful computational environment is required, e.g., a Spark cluster, the appropriate drivers should be installed in the base computational image. This has the advantage of allowing a single compute node embedding Spark, which might be adequate for reproducibility. These concepts are illustrated by 'rspark', an experimental edge computing environment consisting of RStudio and a Postgres containers capable of connecting to a containerized Spark cluster or more generally a cloud-based cluster. The Rstudio container is based on R running on Ubuntu 18.04 from the Rocker Project. Although RStudio is the main interface for programming, it is also possible to use an alternate editor, such as Vim as.
Keywords: Reproducible report, Docker, R markdown, Docker-compose, Spark