Outline of Proposed Research

With an explosion in data collection across nearly all scientific disciplines, we are entering the big-data era of scientific exploration. From the Sloan Digital Sky Survey [1] in cosmology, to the UK Bio-Bank [2] and many others in Neuroscience, initiatives are being launched to federate this data-tsunami, many with particular emphasis and commitment towards open data-sharing. As access to data has increased drastically, it has become apparent that a lack of reproducibility in data analysis methods or collection paradigms is becoming a plague in many current disciplines of science [3]. Unless addressed, our analyses will continue to fail to do justice to the extraordinary datasets we possess. Computational sciences in particular have an opportunity to develop more rigorous practices and provide tools which lower the barrier for producing reproducible tools and analyses. I propose to develop and release an open-source platform which enables researchers to reproduce and distribute computational pipelines across a variety of environments both locally and in the cloud, empowering researchers to create, evaluate, and share more reliable, and therefore more impactful scientific tools, datasets, and discoveries.

The proposed platform will provide three modes of operation: *prototype*, *parallelize*, and *publish*. The *prototype* mode will enable users to develop and run their tool on local resources by leveraging virtual containers as available in Docker [4] and Singularity [5]. While pipeline accuracy may be evaluated in a manner suitable for a given domain, reproducibility and stability will be evaluated through provenance graphs of execution information created at runtime using Reprozip [6]. Provenance graphs will be compared through a variety of graph statistical measures [7] to demonstrate the consistency of a tool across executions, and identify sources of instability therein.

Once tools are deemed stable and ready to be deployed across large cohorts of data, the *parallelize* mode will allow users to distribute these pipelines on high performance computing (HPC) environments. Initially, this will be done using the Amazon Web Services Batch service, which acts as a computing cluster in the cloud; this will be later extended to operate within the Compute Canada HPC environment through CBRAIN [8]. This integration will accelerate the adoption of new tools and datasets into CBRAIN, and increase the accessibility of this platform through the existing web interface and user-base exposed by CBRAIN. Integrations with data management services such as LORIS [9] and Datalad will also increase the accessibility of the datasets they expose.

Finally, the *publish* mode will launch a web server which provides a website and RESTful API for users to interrogate or share the execution information of their tool, as well as the derived data results.

The platform resulting from my proposal will be the only known tool which integrates the development, evaluation, deployment, and distribution/publication of tools, datasets, and analyses all together, and in the cloud. This project will provide a foundation for the next stages of my PhD: the terascale evaluation of neuroimaging pipelines in the cloud, and ultimately the principled development of neuroinformatics analyses based on the joint optimization of accuracy, reproducibility, and stability.

References

[1] York DG. The sloan digital sky survey. The Astronomical Journal. 2000 Sep;120(3):1579.

[2] Sudlow C. UK biobank. PLoS medicine. 2015 Mar 31;12(3):e1001779.

[3] Baker M. 1,500 scientists lift the lid on reproducibility. Nature News. 2016;533(7604):452.

[4] Merkel D. Docker. Linux Journal. 2014;2014(239):2.

[5] Kurtzer GM, et al. Singularity. PloS one. 2017; 12(5):e0177459.

[6] Chirigati F, et al. Reprozip. ICMD 2016 (pp. 2085-2088). ACM.

[7] Dullien T, Rolles R. Graph-based comparison of executable objects. SSTIC. 2005 Jun;5(1):3.

[8] Sherif T, et al. CBRAIN. Frontiers in neuroinformatics. 2014;8.

[9] Das S, et al. LORIS. Frontiers in neuroinformatics. 2011;5.