A Serverless Tool for Platform Agnostic Computational Experiment Management

Gregory Kiar[a,b,†], **Shawn T. Brown**[a], **Tristan Glatard**[*,c], **and Alan C. Evans**[*,a,b,d]

[a]Montreal Neurological Institute, McGill University, Montreal, Canada; [b]Department of Biomedical Engineering, McGill University, Montreal, Canada; [c]Department of Computer Science, Concordia University, Montreal, Canada; [d]Department of Neurology and Neurosurgery, McGill University, Montreal, Canada; [*]Co-senior author; [†]Corresponding author: greg.kiarmcgill.ca

This manuscript was compiled on May 29, 2018

Please provide an abstract of no more than 250 words in a single paragraph. Abstracts should explain to the general reader the major contributions of the article. References in the abstract must be cited in full within the abstract itself and cited in the text.

Keyword 1 | Keyword 2 | Keyword 3 | ...

- Computational sciences are becoming more and more common and necessary.
- Standards are emerging to aid in the reproducibility and shareability of tools and data
  - Boutiques
  - BIDS
  - BIDS apps
- Virtualization tools make analysis software increasingly portable
  - Docker
  - Singularity
- Platforms enable running workflows at scale on a variety of computational resources
  - CBRAIN
  - LONI
  - Nipype
- Execution provenance is becoming increasingly focal and we are recognizing its importance
  - NIDM (neuroscience prov)
  - ReCAP (infrastructure stats/prov)
  - Reprozip (file i/o prov)
- Tools have varying use-cases and barriers for adoption
  - CBRAIN/LONI are designed for production-level pipelines
  - Nipype is complex for tool consumption or simple workflow construction
  - NIDM is very rich and requires deep integration with the tool
  - ReCAP monitors machine resources in virtual machine-based clouds
  - Reprozip has limited compatibility when run around containers, depending on infrastructure
- Clowdr accessibly leverages these approaches where possible and builds-up pipelines with increased deployability, provenance, and shareability
  - Accessible deployment environment closer to development
  - Makes tool consumption very easy
  - Records rich cpu and memory provenance everywhere
  - Records reprozip provenance whenever possible
  - Enables apps/containers that leverage NIDM, Nipype, Reprozip, etc., internally to do their thing, and only adds further richness to provenance records
  - Provides accessible web interface to browse, download, and share executions

Methods

- Data awareness with BIDS
- Cluster and cloud interface with SLURM and Amazon APIs (and extensible)
- Containerization with Singularity or Docker
- Parameter sweeping with boutiques/clowdr
- Tool encapsulation with Boutiques
- Provenance capture using reprozip*, memprofile, cpu-timing
- Data sharing and publication with Flask
- Figure 1: workflow diagrams (figures from main body of poster)

Significance Statement

Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of speciality. The primary goal of the Significance Statement is to explain the relevance of the work in broad context to a broad readership. The Significance Statement appears in the paper itself and is required for all research papers.

GK did things, SB provided help and advice, TG provided help and advice and co-supervised, AE co-supervised.

The authors declare no conflicts of interest in this work.

- Supplement repos?
  - Dockerfile
  - Boutiques descriptor
  - Invocations
  - Clowdr command
  - Dataset

Results.

- Figure 2: instructions infographic (i.e. steps to use clowdr) (figure on left panel of poster)
- Listing 1: installation and execution instructions (listing on left panel of poster)
- Figure 3: we ran, find provenance "here" (i.e. clowdr share) (PING or HCP processing examples)
  - running ndmg on hcp data (compute canada)
  - 1-voxel analysis (compute canada cloud?)
  - Bids example (amazon)
- Figure 4: example provenance analysis (i.e. instance size optimization) (evaluating HCP/PING with ndmg)
  - A) Mem usage comparisons
  - B) CPU call-histograms comparisons

Discussion.

- other uses of provenance information
  - Reprozip trace comparisons (cite)
  - Extrapolate for informed decision making on cloud resource selection (cite)

Format. Many authors find it useful to organize their manuscripts with the following order of sections; Title, Author Affiliation, Keywords, Abstract, Significance Statement, Results, Discussion, Materials and methods, Acknowledgments, and References. Other orders and headings are permitted.

References. References should be cited in numerical order as they appear in text; this will be done automatically via bibtex, e.g. (1) and (2, 3). All references should be included in the main manuscript file.

Data Archival. PNAS must be able to archive the data essential to a published article. Where such archiving is not possible, deposition of data in public databases, such as GenBank, ArrayExpress, Protein Data Bank, Unidata, and others outlined in the Information for Authors, is acceptable.

Digital Figures. Figure ?? shows an example of how to insert a column-wide figure. To insert a figure wider than one column, please use the `\begin{figure*}...\end{figure*}` environment. Figures wider than one column should be sized to 11.4 cm or 17.8 cm wide. Use `\begin{SCfigure*}...\end{SCfigure*}` for a wide figure with side captions.

Single column equations. Authors may use 1- or 2-column equations in their article, according to their preference.

To allow an equation to span both columns, options are to use the `\begin{figure*}...\end{figure*}` environment mentioned above for figures, or to use the `\begin{widetext}...\end{widetext}` environment as shown in equation ?? below.

Please note that this option may run into problems with floats and footnotes, as mentioned in the cuted package documentation. In the case of problems with footnotes, it may be possible to correct the situation using commands `\footnotemark` and `\footnotetext`.

Supporting Information (SI). The main text of the paper must stand on its own without the SI. Refer to SI in the manuscript at an appropriate point in the text. Number supporting figures and tables starting with S1, S2, etc. Authors are limited to no more than 10 SI files, not including movie files. Authors who place detailed materials and methods in SI must provide sufficient detail in the main text methods to enable a reader to follow the logic of the procedures and results and also must reference the online methods. If a paper is fundamentally a study of a new method or technique, then the methods must be described completely in the main text. Because PNAS edits SI and composes it into a single PDF, authors must provide the following file formats only.

Appendices. PNAS prefers that authors submit individual source files to ensure readability. If this is not possible, supply a single PDF file that contains all of the SI associated with the paper. This file type will be published in raw format and will not be edited or composed.

ACKNOWLEDGMENTS. Please include your acknowledgments here, set in a single paragraph. Please do not include any acknowledgments in the Supporting Information, or anywhere else in the manuscript.

References

1. Belkin M, Niyogi P (2002) Using manifold stucture for partially labeled classification in *Advances in neural information processing systems.* pp. 929–936.
2. Bérard P, Besson G, Gallot S (1994) Embedding riemannian manifolds by their heat kernel. *Geometric & Functional Analysis GAFA* 4(4):373–398.
3. Coifman RR, et al. (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America* 102(21):7426–7431.