Data Management Plan

Project Data Types

Project outcomes will include publications of algorithmic development, spectral–temporal models of astrophysical variability, and a web–based framework for classification of ongoing survey data. Testing and verification datasets will be collected either from open data programs, or through proprietary access (with no intention to archive 3rd party primary data).

We will conduct all software development on the publicly available open-source collaboration website `http://github.com`, where many libraries that serve as the foundation for this project are already hosted. The website has become a leading collaboration platform; it enables distributed users to download code and contribute back to the project, thanks to a powerful peer–review system. The underlying version control system behind the website is Git, created to support the development of the Linux kernel. By design, Git includes – with every project download – its entire development history, ensuring that even if central repositories on github were to disappear overnight, thousands of cryptographically verifiable, perfect copies of the project's history exist across the internet. Hence, there is no single point of failure that could cause a loss of the project's data; it is also possible to switch to a different hosting system without loss.

We will, as a safety layer, always host a mirror of all Git repositories and media files of the project in a central server. This provides an easy to locate, trusted copy of the active development sites in case of e.g. temporary service outages.

We will write software with an eye towards integration in current (PTF) and future (LSST) real–time classification streams. This means implementation in the Python and (optionally) C++ languages.

Data Standards

The input training data will be stored and consumed in Virtual Astronomical Observatory[1] defined format. We do not currently plan on making this collected archive available to the general community, but adopting these standards will make our API design (and any unanticipated collaboration) straightforward. It is unclear if the ST surfaces are themselves trivially VAO–compatible, but we will make every effort to represent them as such, to the extent possible. This would ensure that no proprietary tools are necessary to view, use, or contribute back to any of the project outcomes.

Privacy and Intellectual Property Issues

Our project does not need collection of personally sensitive data; we can thus make all of our outcomes available under open source licensing terms without any requirements of confidentiality.

Re–Use and Re–Distribution Licensing Terms

All outcomes will be made available to the public following the principles set forth in the Reproducible Research Standard proposed by Stodden [1]:

- All code we develop for this project will be made available under the terms of the open source BSD license, [2] whenever possible (i.e., as long as it is completely original code or derived from

[1] `http://www.usvao.org/`

[2] `http://www.opensource.org/licenses/bsd-license.php`

similarly licensed code). In the event that we contribute to projects with existing licenses that are not compatible with the BSD license, such as LGPL- or GPL-licensed projects, we will contribute to these projects under the terms of their own licenses.

- Educational materials (e.g. documentation) and other media will be released under the terms of the Creative Commons Attribution License CC BY, [3] except in cases where we may reuse materials released under a more restrictive license such as the Attribution-ShareAlike CC BY-SA [4] (used by e.g. Wikipedia). Such materials would be released under the terms of the original license, in compliance with its original terms.

- Since US Copyright law prevents the copyright of raw facts, any data generated as part of our project will be released under the Creative Commons CC0 terms, [5] i.e., fully released to the public domain without further copyright claims.

Long-Term Archival Plans

While Git's distributed nature effectively uses the entire internet as a backup system, as indicated above we will use our server to host a mirror of all of our materials on the UC Berkeley and University of Washington networks. We also expect extant and developing classification streams to uptake our models, which will ensure they persist (and potentially evolve) in an active environment.

References

[1] Stodden, V. (2009). Enabling Reproducible Research: Open Licensing For Scientific Innovation. *International Journal of Communications Law and Policy*, vol. 13.

[3] `http://creativecommons.org/licenses/by/3.0`
[4] `http://creativecommons.org/licenses/by-sa/3.0`
[5] `http://creativecommons.org/about/cc0`