2021-04-02 Meeting notes

Agenda

- Roman Data Management system - Harry Ferguson

Attendees

Alice Shapley, David Spergel, Dimitri Mawet, George Helou, Gregory Mosby, Harry Ferguson, James Rhoads, Jan Tauber, Jason Rhoads, Jeffrey Kruk, Jessie Christiansen, Jessica Lu, John Mackenty, Joshua Schlieder, Julie McEnery, Keith Bechtol, Ken Carpenter, Megan Donahue, Neil Zimmerman, Neill Reid, Peter Melchior, Rachel Akeson, Roeland van der Marel, Ryan Hickox, Sangeeta Malhotra, Saurabh Jha, Zeljko Ivezic

Minutes

Harry Ferguson on Roman SOC Data Management System

This is "big data" regime for astrophysics: Data accumulated per week likely >100x Hubble. Downloading and processing exceeds typically available resources.

Cloud-based high-level data processing - brings software to the data.

The concept approved at PDR:
-JWST pipeline with adaptations
-External science teams provide high-level WFI processing software this was identified as a risk.
-A science platform (HLPP) allows users to interact with data on cloud.
-Archive with functionality like MAST

Archive expected to have notebook interaction and visualization

High level (Level 2 and beyond) will be stored in cloud.

Architecture trades
Division of responsibilities between MOC, SOC, SSC unlikely to change
Cloud service provider likely AWS
Data format is ASDF (Advanced Scientific Data Format) - better metadata capability than FITS. Already being used for JWST (ASDF packaged as FITS).
Science platform - likely JupyterHub but this landscape is evolving quickly enough that it remains undecided

Since PDR, an increase in pipeline scope at the SOC: SOC no longer depends on software deliveries from SITs.
Also new: Level 1 data simulations, idealized Level 2 data simulations with source injection - critical for quantifying systematic uncertainties. This may not extend all the way into weak lensing survey needs.

Point spread functions - empirical calibrations vs position and time. Queryable library with API.
Astrometric calibration referenced to Gaia.
Catalogs of static sources and based on difference images

Level 2 simulations and tools will be publicly available.

SSC is responsible for microlensing & spectroscopy high-level data processing.

We hope there will be community-contributed products. For example: joint photometry with complementary data sets, photometric redshifts, SED fits, etc.

Computing and data resources

Computing and data resource management is a work in progress.

The main advantage of the cloud approach is convenient scalability. Most likely, lower total costs to NASA relative to multiple on-site installations. Roman grants won't need a separate computing resource budget.

Tiered user concept, ranging from lightweight needs to increasing allocations, periodic renewal/purges. Looking to enable ~1000 Roman papers/year.

Features explicitly out of scope of the cloud platform concept: CGI data (covered by SSC), high school and undergrad coursework support, dedicated machine learning environment.

Archive plan

Expect Roman's greatest impact to be through archival science

Archive should enable both catalog-based and image-based discovery

Complementary data (Rubin, Euclid, eROSITA, Subaru HSC, HST): may co-locate Rubin co-adds overlapping on Roman survey areas. Limited functionality for transient science.

The SSC will manage archival research proposal opportunities

Comments/questions

Q: What about time domain products?
A: Expect a limited baseline of time-domain analysis. Users who create their own time-domain tools can easily share them with the community through the cloud platform.

Q: How do we expect to run contributed analysis at scale?
A: So far, we expect the tools are modular enough, open source-based, that the interface will not require much management. Could have working groups, or personnel at SOC to assist with integration.

Q: Is the RSIG being asked to comment on this plan?
A: Some decisions already made, but we may need help prioritizing if we run up against resource limits

Based on Hubble experience, the decisions made by Roman here may transform how the community works as a whole.

What will people have to know to take advantage of this system? Will Rubin already educate the community on this paradigm?
We come after Rubin, so we can benefit from learning the community's experience with Rubin data.
Need very well documented APIs - Sloan is a good example to follow.

The concept for the Roman platform is you won't need your own AWS account, and are insulated from having to really know anything about AWS.

Next meeting we will begin discussing the future science teams.