STScI | SPACE TELESCOPE SCIENCE INSTITUTE

EXPANDING THE FRONTIERS OF SPACE ASTRONOMY

The Roman SOC Data Management System

Henry Ferguson STScI

Science Operations and the Data Management System

STScI Roman Science Operation Center responsibilities include:

- Planning & scheduling all observations
- Calibration and support of the Wide Field imaging
- The archive (MAST) for all mission data
  - Most Roman science will be archival due to the survey nature of the mission

NASA Astrophysics *Big Data*:

- Data accumulated per week likely to be >>100x *Hubble*
- Both catalogs and pixel-level data sets provide unique science opportunities
- Downloading and processing exceeds resources typically available

Barbara A. MIKULSKI ARCHIVE FOR SPACE TELESCOPES

Science data products from multiple mission partners

- Calibrated and mosaiced images, extracted spectra, catalogs, etc.
- Staged in the cloud and co-located with significant computational resources
- Open source, modular imaging pipeline facilitates custom reprocessing

Data storage & processing

- Cloud-based high-level data processing brings software to the data
- Jupyter Lab environments ease access, sharing and repeatability
- Software environment for the community in sync with mission data processing

AWS

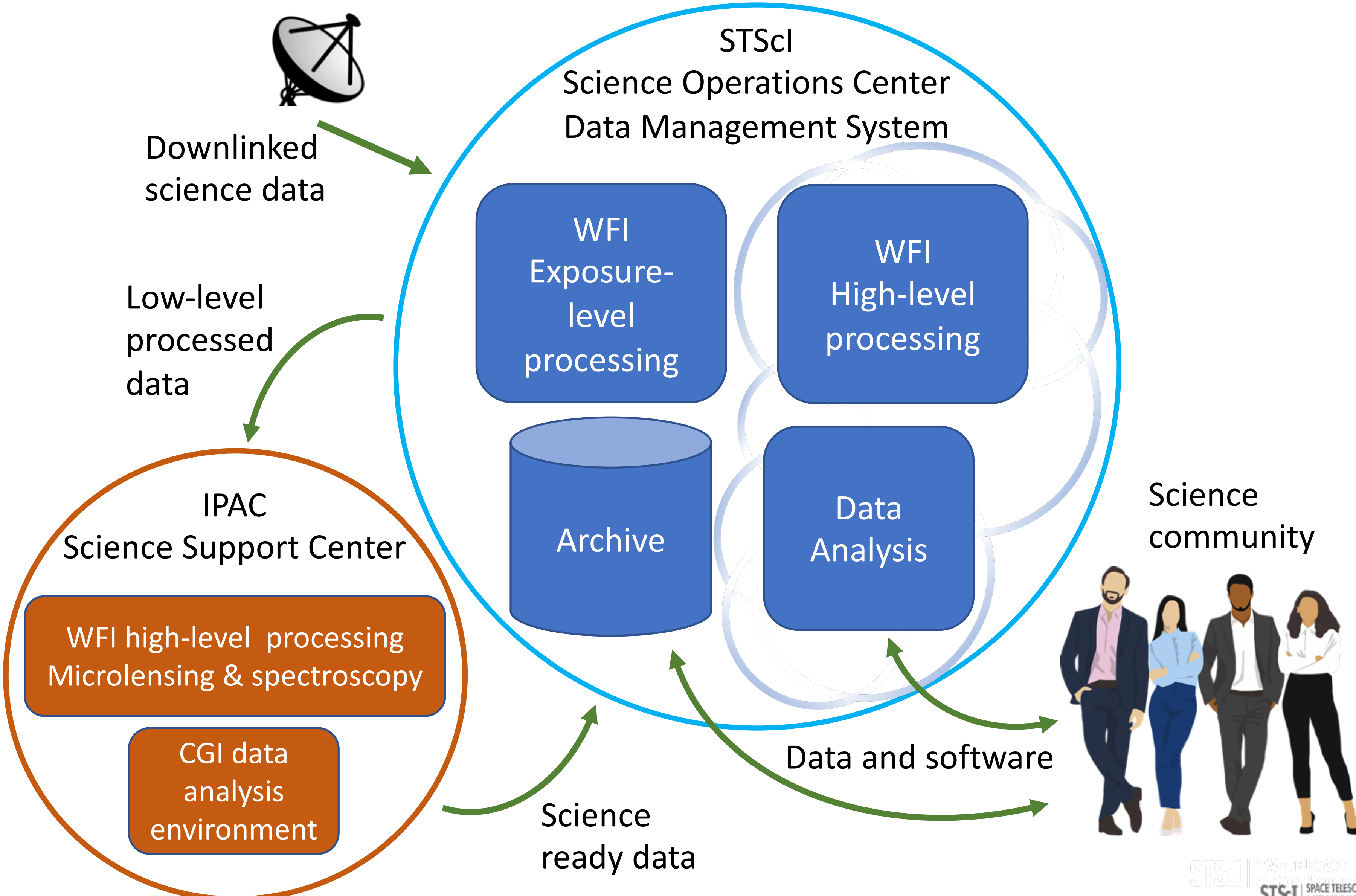
jupyter

STScI | SPACE TELESCOPE SCIENCE INSTITUTE

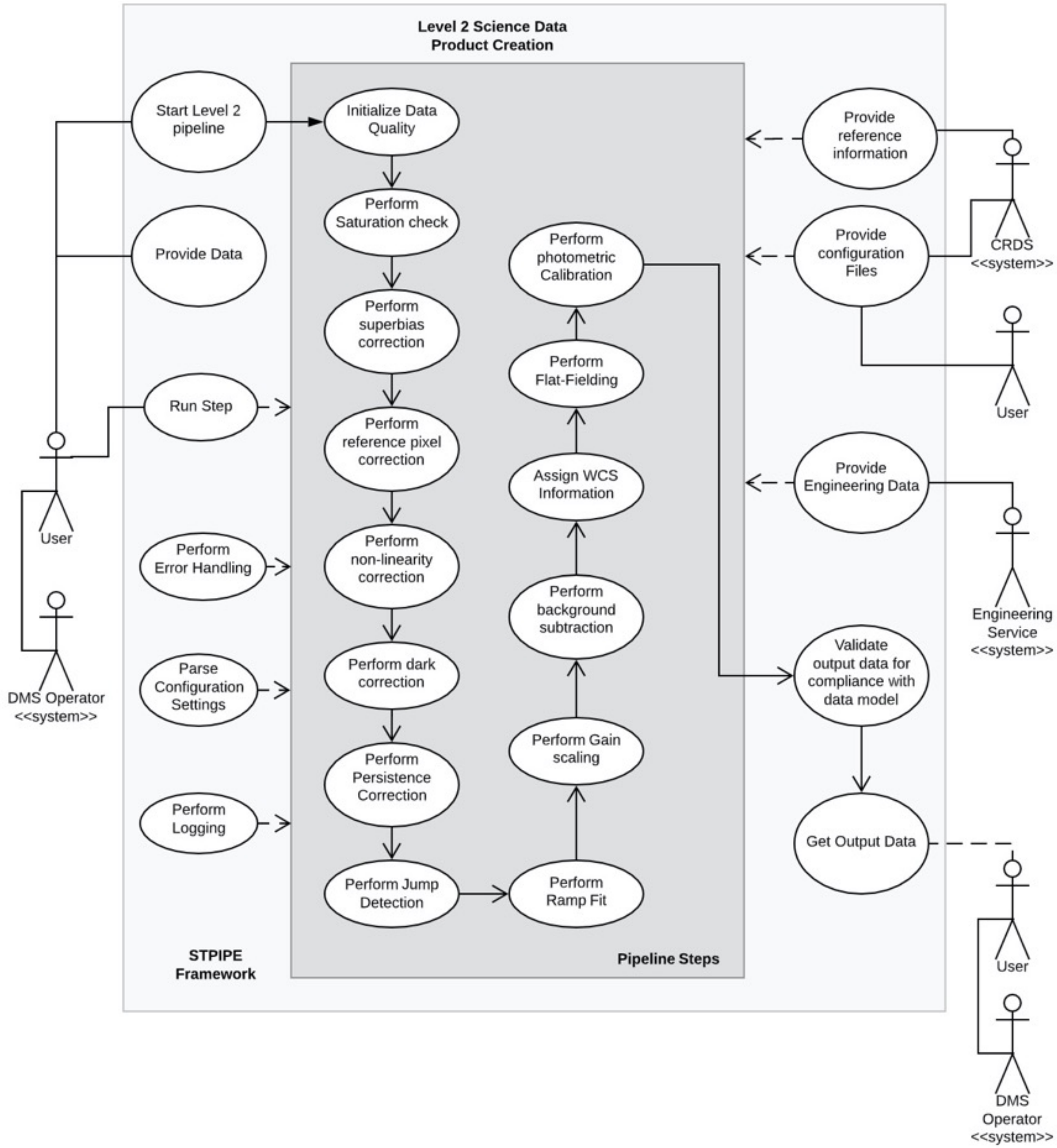
SOC Data Management System

- System approved at PDR had the following attributes:
  1. The JWST science calibration pipeline with minor adaptations
  2. High-level WFI processing using software provided by external science teams
  3. A science platform (HLPP) that allows users to interact with the data and high-level processing software in the cloud
  4. An Archive with HST/JWST/MAST like functionalities
     - Including science data from the SOC, the SSC and high-level community products.
     - Archiving selected WFI and CGI ground test data
     - Storage of all Roman mission data products

Roman Data Management



Exposure Level Processing Flow



Archive

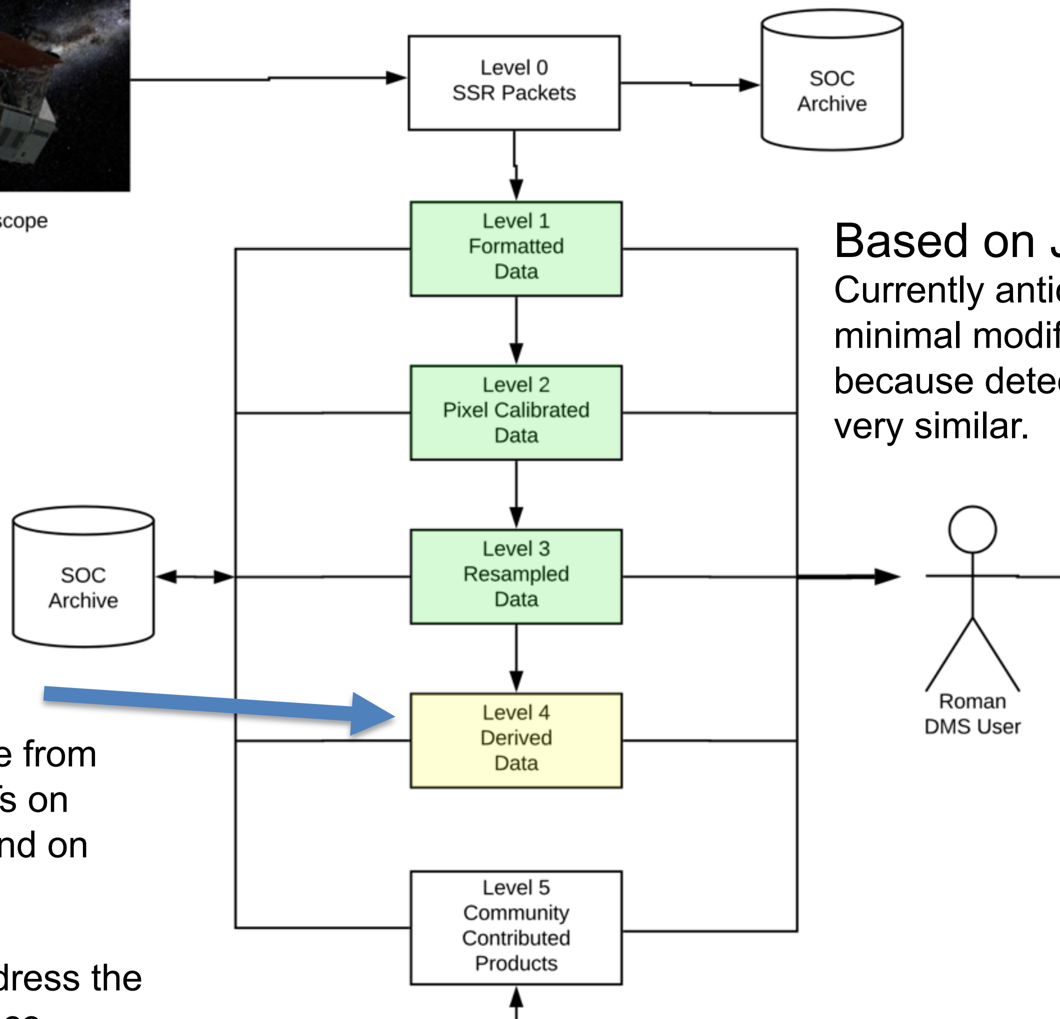
- Roman science data are public
- Users will be able to retrieve science data from MAST.
  - including data from the SOC, the SSC as well as I&T data and high-level community products.
- Expect archive services to evolve
- Currently incorporating Jupyter analysis+visualization tools into the archive for JWST.
  - Improving access to high-level products with services like z.mast and exo.mast.
- Higher-level products (level 2 and beyond) will be available in the cloud as well
  - SOC is currently scoped for cloud hosting of SOC data, not SSC, CGI or community products (although they will be in MAST)

Architecture Trades & Discussions

- Decided at mission level
  - Division of responsibilities between MOC, SOC and SSC
- Past SOC studies
  - Cloud service providers ⟹ AWS
  - Mission Data Formats ⟹ ASDF
  - Science Platform ⟹ JupyterHub
- Underway
  - Database technologies
  - Extent of cloud integration for MOC, SOC and SSC
  - Evolution of cloud vs. on-premises for all missions
  - Management & policies for community use of the cloud resources (later in this talk)

High-Level Data Pipeline Software

Roman Space Telescope



Based on JWST

Currently anticipate minimal modifications because detectors are very similar.

Now in scope for SOC:

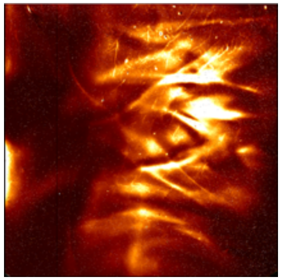
Original plan was to integrate software from SITs. New plan is to work with the SITs on features and algorithms, but not depend on SITs for software deliveries.

Worked closely with the Project to address the highest priority risks for meeting science requirements.

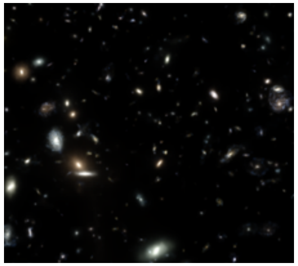
Newly in scope for development

- Simulations
  - Package (based on JWST Mirage) to simulate WFI level-1 data
    - Including the most important instrument signatures
    - Hugely beneficial for testing pipeline instrument-signature removal
  - Idealized simulations for Level 2
    - Enables artificial source injection in the pipeline
    - This is critical for quantifying systematic uncertainties for many different science topics
- Point spread functions
  - Empirical calibrations vs. position and time
  - Queryable library
  - Programmatic access via API

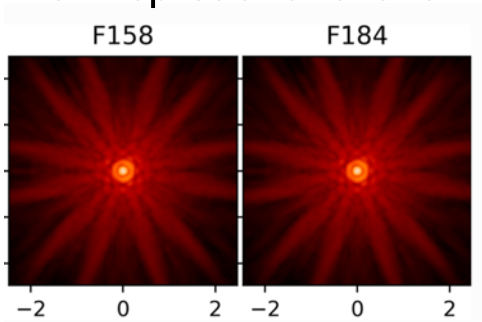
Instrument signatures



Idealized

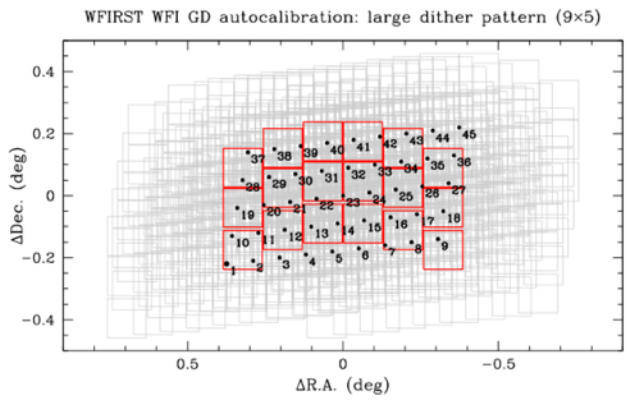


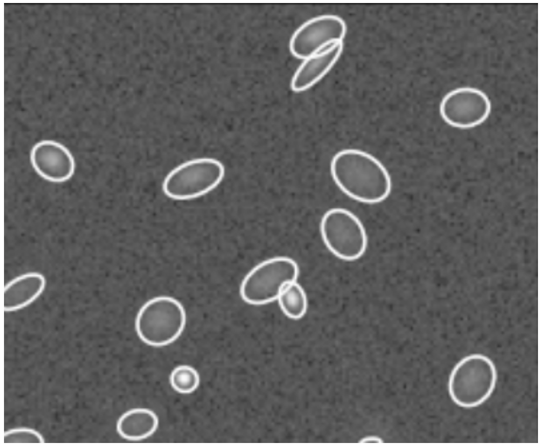
Point spread functions



Newly in scope for development

- Astrometric calibration
  – Referenced to GAIA
  – Consistency between image metadata and catalogs
- Catalogs
  - Catalogs of static sources
    - Matched-PSF multi-band photometry
    - Including photometric redshifts (based only on Roman data)
  - Variable sources
    - Multiple epochs of catalogs
    - Catalogs of difference images
  - Spectroscopic extractions & redshifts (SSC)
    - Matched to sources in imaging catalog





Integrating high-level science data

- SSC responsible for microlensing data & spectroscopy
- SOC and SSC pipelines diverge at the highest levels, but:
  - Data will be distributed through common archive and (ideally) common formats with meta-data
  - HLSS and HLIS data will be integrated and matched per science requirements
  - SOC Science calibration pipeline will be public so the community can use SOC and SSC modules for other applications
  - The goal is for the archive to appears seamless to outside users

Community Contributed Products

- Public data products contributed by the science community are likely to be widely used. Examples include:
  - Joint photometry with complementary data sets
  - Photometric redshifts that use complementary data sets
  - Value-added catalogs of derived properties (e.g. from SED fitting)
  - Hybrid spectroscopic and photometric catalogs
  - Survey-level calibrations
    - Improved astrometry & photometry after constraining for consistency across the full survey
    - Window functions, masks, PSF kernels, etc.
  - Transient-free template images
- Details & cadence to be defined through future community engagement and opportunities

Computing & Data Resource Management for the Science Platform

Work in Progress

Evolution

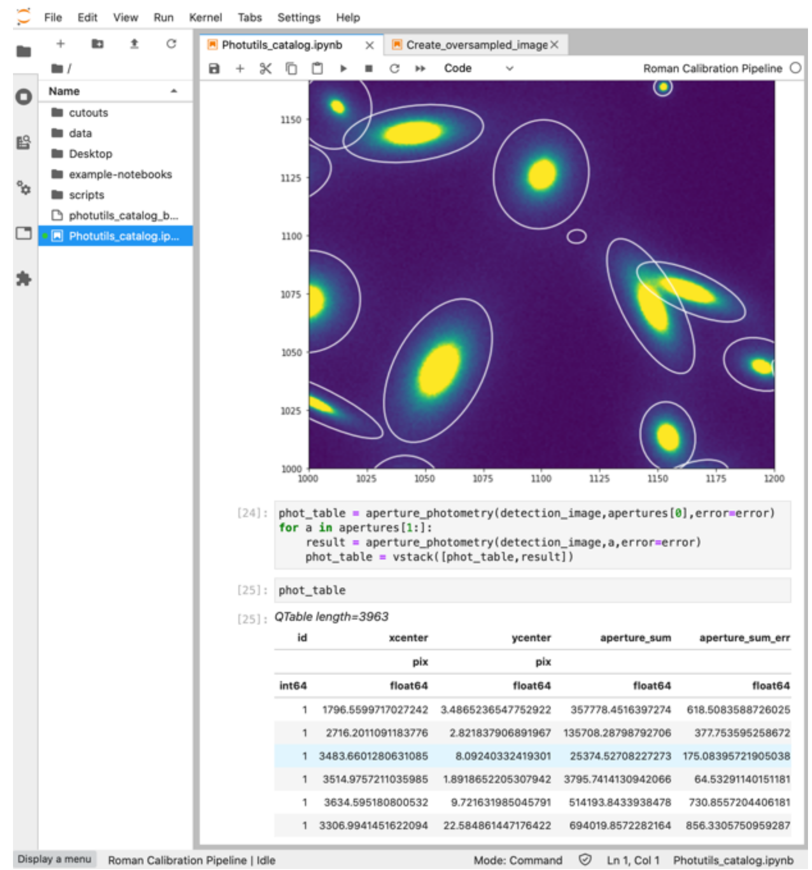
- Major projects are increasingly moving to the cloud
- Funding strategies could/should evolve accordingly
  - Traditional Hubble model is to enable hardware purchases & computing service fees in grants
  - Supercomputing proposals are typically separate from grants
    - NSF XSEDE program; NASA HEC program
  - Pros & cons:
    - May not be efficient to fund cloud allocations in individual small grants
    - Want to avoid double-jeopardy of separate proposals for computing vs. getting observing time or archival funding
    - Want to enable people with funds to be able to use them and work in the same environment
- Cross-institutional collaboration brings opportunities & challenges
  - How to manage access & allocations for non-US astronomers

Why the Cloud?

- Putting both the computing and the science-ready data in the commercial cloud offers the following benefits:
  - Convenient scalability for both data volume and computational demands
  - Flexible solutions for specific computing needs (e.g. GPUs or I/O optimized computing)
  - Lower total costs to NASA relative to multiple on-site installations
- Benefits to the science users include:
  - Efficient access to the data
  - Computing resources for exploratory work are available with no need to write a grant proposal
  - Local IT and software support costs are greatly diminished
  - Easier collaboration with astronomers across institutions
  - A powerful and stable science software environment

Community use of the HLPP

- Log in with your MyST account
- JupyterHub instance
  - Roman science calibration pipeline software installed and configured
  - Full Python + Astropy ecosystem installed and configured
  - Ability to install other packages and your own code
- Flexible, scalable architecture
  - Simple to add CPU & storage
  - High-throughput access to the data
  - Can scale up resources (e.g. GPUs or neural engines) as science needs & technology evolves



Concept for managing Roman platform

- The platform should provide a sufficient base level of resources
  - Most Roman grants wouldn't need a computing line item.
  - Projects needing exceptional resources could still apply for funds
  - Allows much more global optimization for science than case where funds are locked in small grants.
- Tier concept to support most users
  - Very lightweight process for getting access and increasing allocations
  - Will require periodic renewal
  - Will require rather frequent purging of stored data to control costs
- Looking to to enable ~1000 Roman papers/year
  - The Roman-data-intensive work; not necessarily all the computing
  - Not long-term archival storage of intermediate projects

Resources

- Priorities
  - Provide resources for work that needs to access large quantities of Roman data
  - Make it easy for the user community to get access, use & collaborate
  - Cost effectiveness & cost predictability
    - Biggest concern is persistent file-system storage (EFS)
  - Do not *require* use of the platform to do Roman science

- Not part of the concept
  - Providing resources for everything else
    - e.g., simulations & modeling, reducing complementary data
    - Don't plan to micromanage but will set quotas based on Roman data-analysis needs.
  - Providing a dedicated machine-learning environment
    - Could add this later if there is demand
  - Resources for high-school & undergraduate education
    - Using Roman data for coursework
  - Cost sharing with non-US stakeholders
  - CGI data analysis on the same platform
  - TBD
    - Co-location & support for analysis of SSC high-level products on the same platform
    - Co-location & support for community contributed products on the same platform

Relation to the Core Surveys

- Using WFI: weak lensing, supernovae and microlensing
- Have looked at “typical use cases” rather than core-survey workflows
    - Unclear if the teams will work in the cloud or other facilities
        - Depends on which teams are selected and what resources they may already have available
    - Expect the survey teams to do extensive survey-level processing beyond the scale of typical users
    - Infrastructure for this processing may or may not be best associated with this general platform
        - May need co-location with extensive simulations or other data sets
        - May make use of other facilities associated with survey teams
    - Even if in AWS, likely beneficial to optimize based on specific needs
- The concept presented here does not preclude very large “consortium” allocations, if needed
- Or separate AWS accounts, access to NASA HPC facilities, …

Tier Concept

- Entry tier
  - Anyone with a MyST account
  - Examples: Filter the entire HLIS catalog a dozen times, make custom catalogs of few FPAs, registered to an external dataset, Extract & download 1000 cutouts
- Research tier
  - Pro-forma science justification & annual renewal
  - Examples: Custom catalogs on ~ 100 sq deg. Custom processing at the level of 1 minute per FPA for the entire HLIS. Thousands of catalog queries & cutouts.
- Consortium tier
  - Proposals & panel review (lightweight process)
  - Multiple users sharing a single allocation
    - Resource sub-allocations left to the consortium
  - Examples: Signal injection and re-run of pipeline for full HLIS

Summary of Policy Recommendations

- NASA funds the platform through STScI
  - Not via individual grants to PI, because this is simply not worth the administrative overhead of multiple transfers of funds
  - Projects needing exceptional resources could still apply for funds through ROSES process
  - Allow international access in all tiers
- Tier concept to support most users
  - Lightweight process for getting access and increasing allocations
  - Will require user agreements and periodic renewal
  - Regular migration and purging of data to control costs
- Scoped to support ~1000 Roman papers per year
  - Not all computing but stages where proximity of CPU to Roman data is beneficial
  - Not explicitly driven by core survey computational requirements

Big Data Discovery Tools

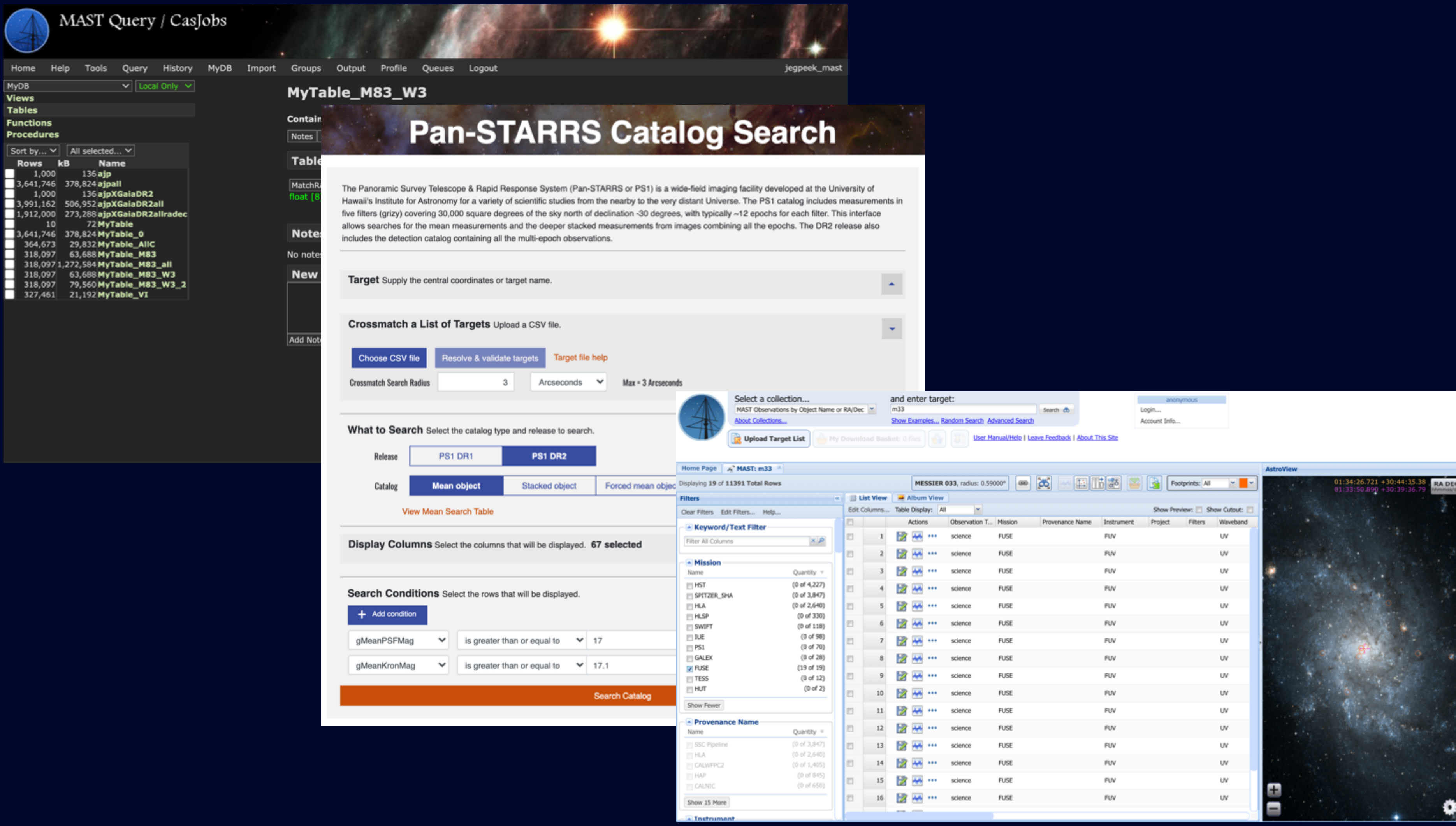
Courtesy of Josh Peek,
STScI Data Science Mission Office

What is in the archive plan already: MAST adapted for Roman

The SOC will reuse existing MAST functionalities. It will not provide new data mining or exploration functions specifically to assist the community in dealing with the challenges of the large Roman data volume, beyond what is provided by the HLPP …

What is in MAST for Roman today?

- CASJobs — The first science platform, built by SDSS, based on SQL
- catalogs.mast — A new, lightweight catalog query webpage
- Discovery Portal — A observation-based search tool
- HLSP system — A robust system for ingesting community data


MAST Query / CasJobs
Home Help Tools Query History MyDB Import Groups Output Profile Queues Logout
jegpeek_mast
MyDB
Local Only
Views
Tables
Functions
Procedures
Sort by...
All selected...
Rows kB Name
1,000 136 ajp
3,641,746 378,824 ajpall
1,000 136 ajpXGaiaDR2
3,991,162 506,952 ajpXGaiaDR2all
1,912,000 273,288 ajpXGaiaDR2allradec
10 72 MyTable
3,641,746 378,824 MyTable_0
364,673 29,832 MyTable_AllC
318,097 63,688 MyTable_M83
318,097 1,272,584 MyTable_M83_all
318,097 63,688 MyTable_M83_W3
318,097 79,560 MyTable_M83_W3_2
327,461 21,192 MyTable_VI
MyTable_M83_W3
Pan-STARRS Catalog Search
The Panoramic Survey Telescope & Rapid Response System (Pan-STARRS or PS1) is a wide-field imaging facility developed at the University of Hawaii's Institute for Astronomy for a variety of scientific studies from the nearby to the very distant Universe. The PS1 catalog includes measurements in five filters (grizy) covering 30,000 square degrees of the sky north of declination -30 degrees, with typically ~12 epochs for each filter. This interface allows searches for the mean measurements and the deeper stacked measurements from images combining all the epochs. The DR2 release also includes the detection catalog containing all the multi-epoch observations.
Target Supply the central coordinates or target name.
Crossmatch a List of Targets Upload a CSV file.
Choose CSV file
Resolve & validate targets
Target file help
Crossmatch Search Radius
3
Arcseconds
Max = 3 Arcseconds
What to Search Select the catalog type and release to search.
Release
PS1 DR1
PS1 DR2
Catalog
Mean object
Stacked object
Forced mean objec
View Mean Search Table
Display Columns Select the columns that will be displayed. 67 selected
Search Conditions Select the rows that will be displayed.
Add condition
gMeanPSFMag
is greater than or equal to
17
gMeanKronMag
is greater than or equal to
17.1
Search Catalog
Select a collection...
and enter target:
Upload Target List
MESSIER 033, radius: 0.5900°
AstroView

What we need: tools for big data discovery

Roman's greatest impact will be through its archival science.

To amplify this science we need tools that allow collaborations to

- quickly find what they are looking for
- explore the image and catalog space
- integrate *Roman* with other large data sets

Here are 5 main areas of expansion to achieve these goals:

1. Big Data Catalog Discovery

2. Big Data Image Discovery

3. Search By Example

4. Big Data Fast Survey APIs

5. Curated Complementary Data

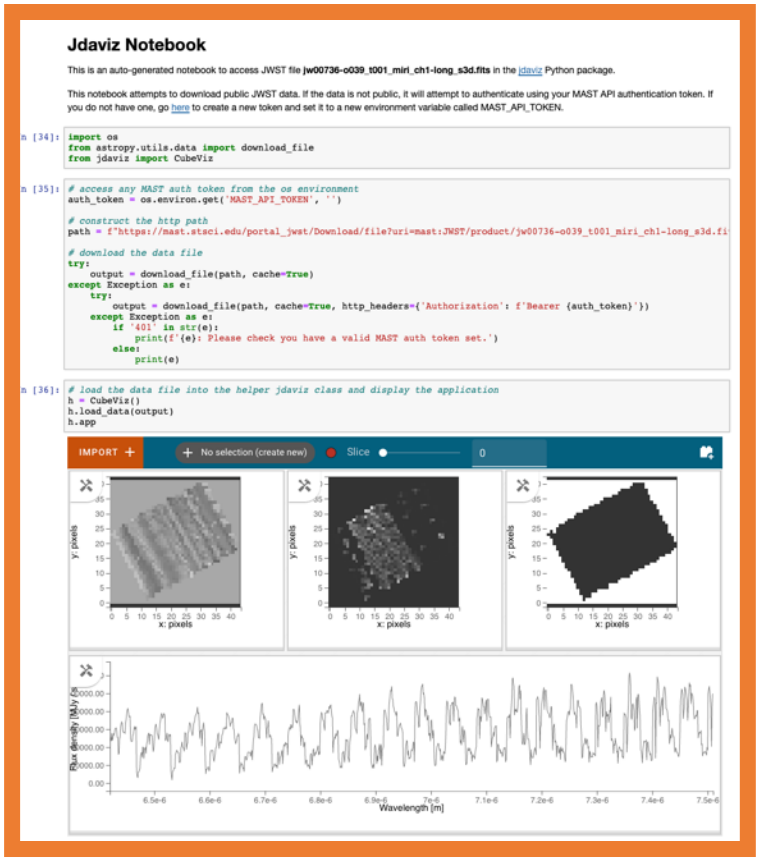
1. Big Data Catalog Discovery

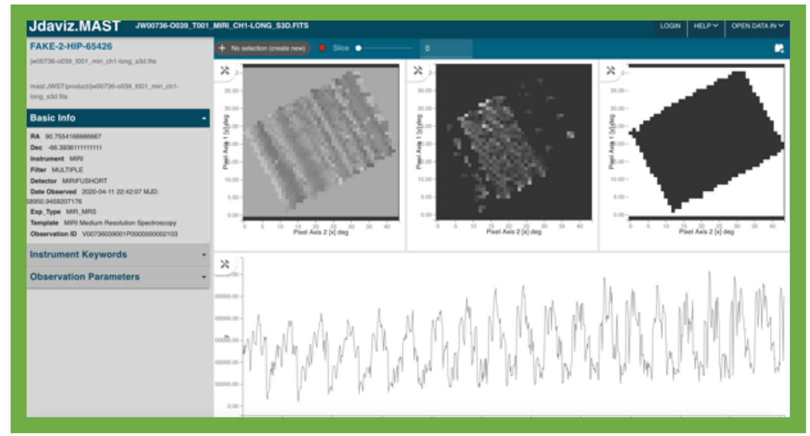
Users will need a fast source catalog query system that integrates with Jupyter

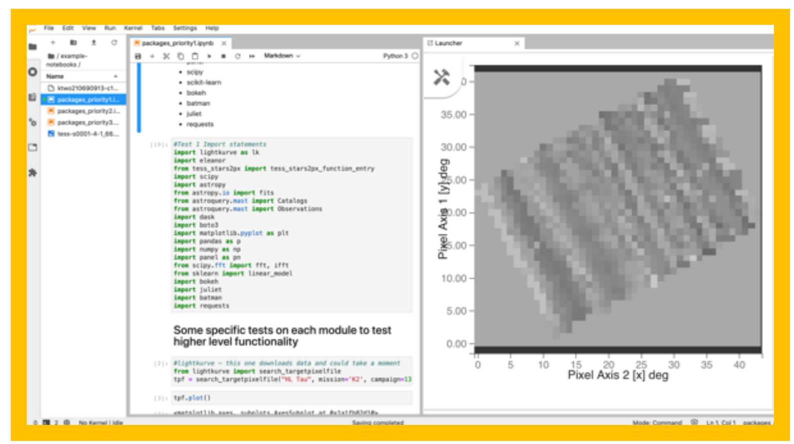
Cloud-based Catalog Data Infrastructure

- Integration with commercial cloud services, Roman Science Platform
- Asynchronous search
- TAP/ADQL endpoints (consistent with *Gaia*, *Rubin*)
- Gigascale Crossmatching (e.g. GIS-Based Greenplum, AXS)
- Jupyter Viz Stack interface: Notebook, Platform, Webpage

The Jupyter Viz Stack: Notebook, Platform, Webpage



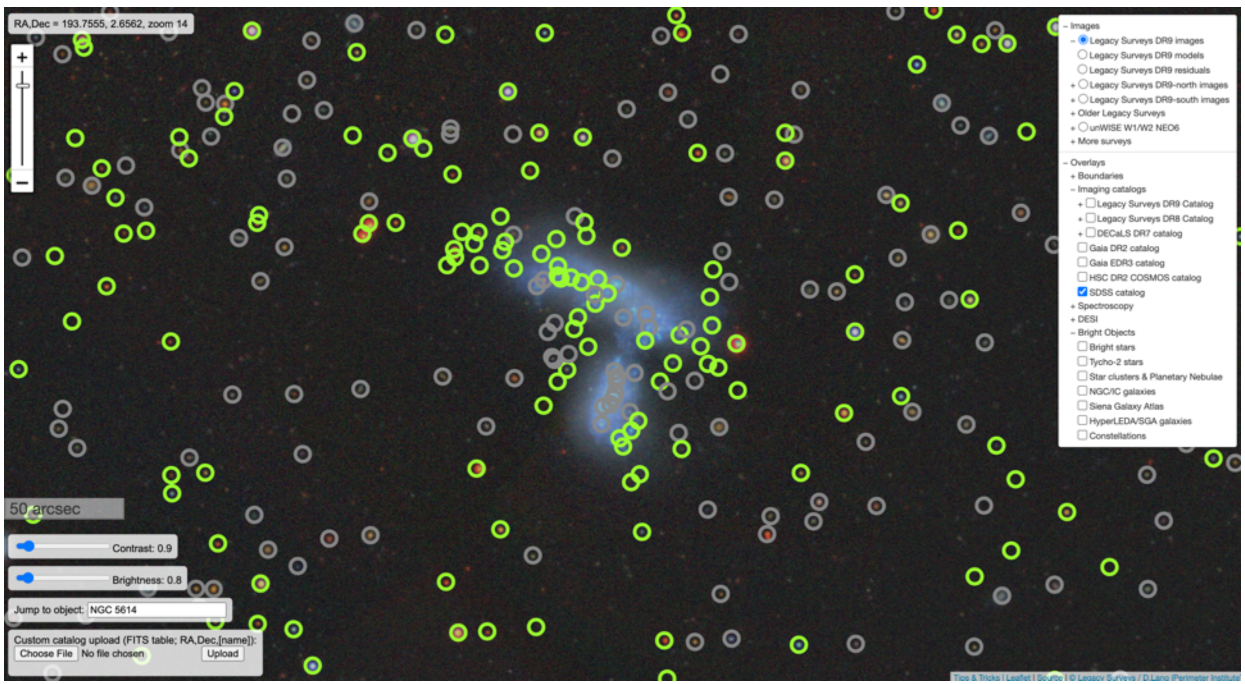




2. Big Data Image Discovery

Users will need to be able to quickly explore the surveys and image cutouts

- **Survey Viewer**

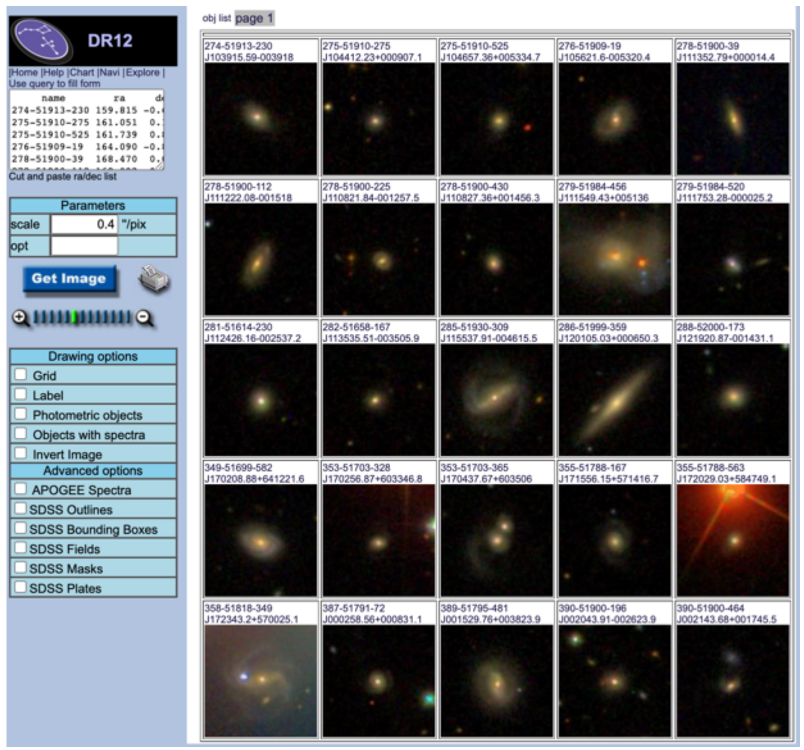


example: legacy survey

2. Big Data Image Discovery

Users will need to be able to quickly explore the surveys and image cutouts

- Survey Viewer
- **Cutout Viewer**

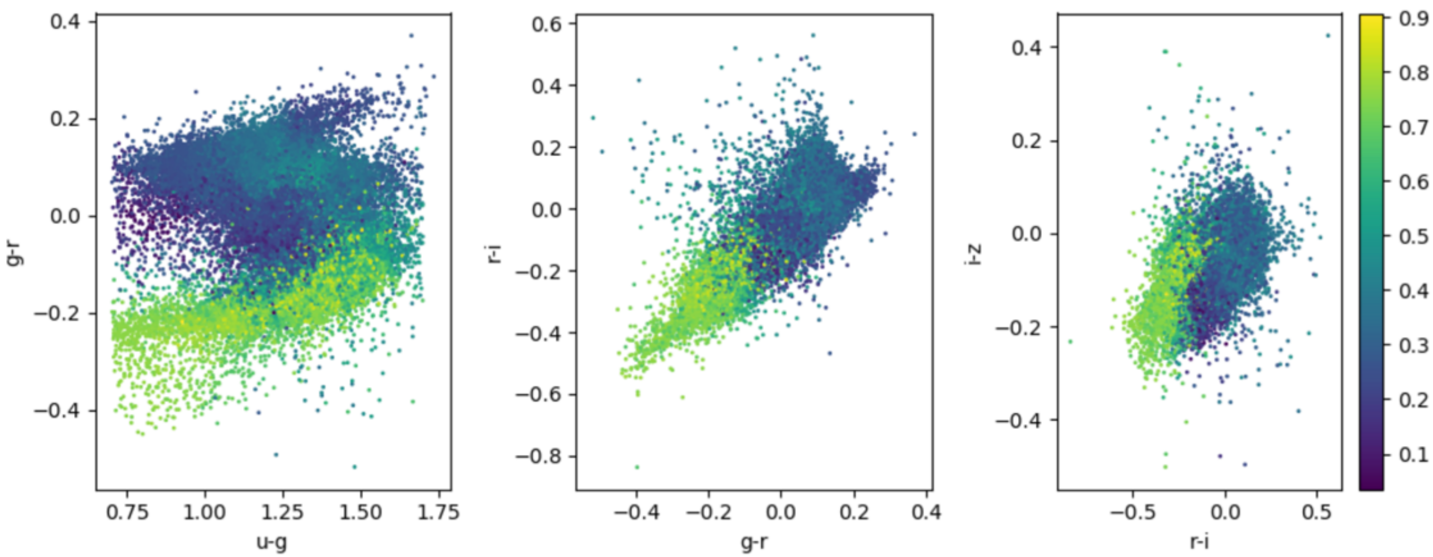


example:
SDSS Imaging

3. Machine Learning Search By Example

By being able to put in the coordinates, metadata, or images of scientifically relevant users can retrieve the full wealth of similar structures within the Roman archive

- Search by sources with catalog metadata (Classical Machine Learning)



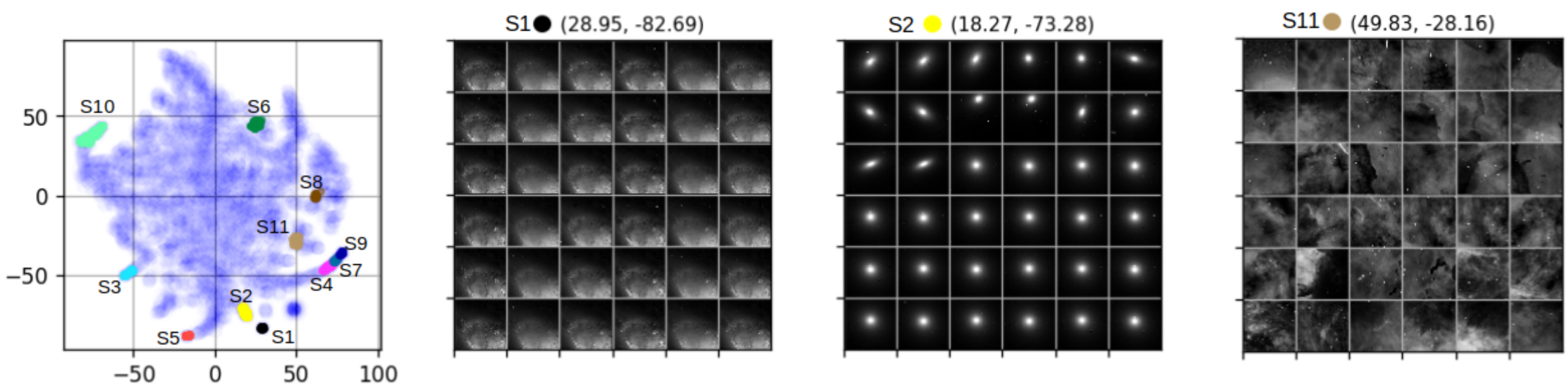
example: finding BHBs with KNN

3. Machine Learning Search By Example

By being able to put in the coordinates, metadata, or images of scientifically relevant users can retrieve the full wealth of similar structures within the Roman archive

- Search by sources with catalog metadata (Classical Machine Learning)

- Search by Image (Deep Learning)

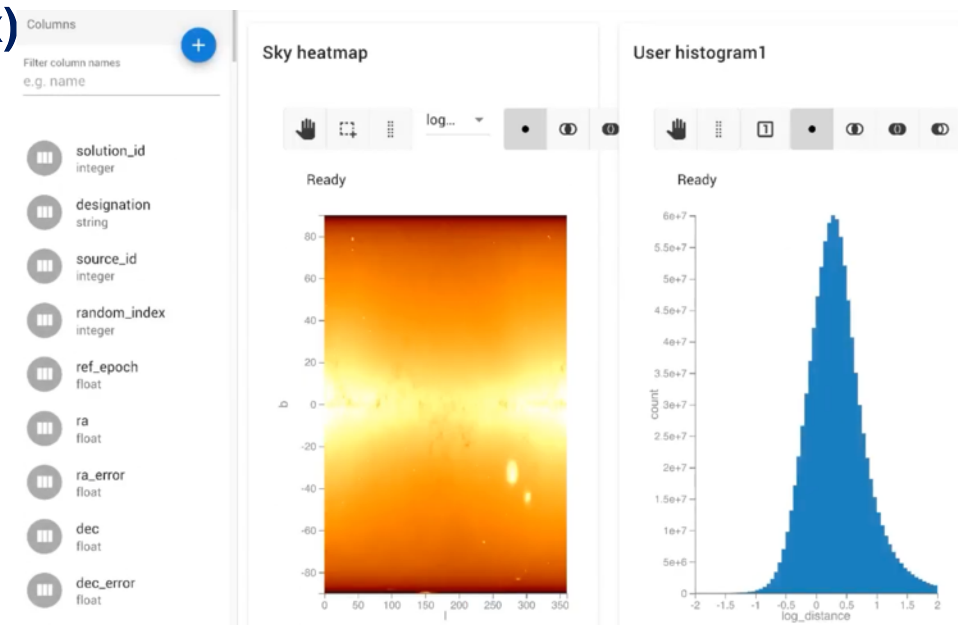


example: Finding similar ACS images with transfer learning

4. Big Data Fast Survey APIs

By providing fast, optimized services for common computationally intensive tasks, Roman can allow for advanced processing workflows both within and beyond the Roman Science Platform

- Histogram API (e.g. AXS, vaex)



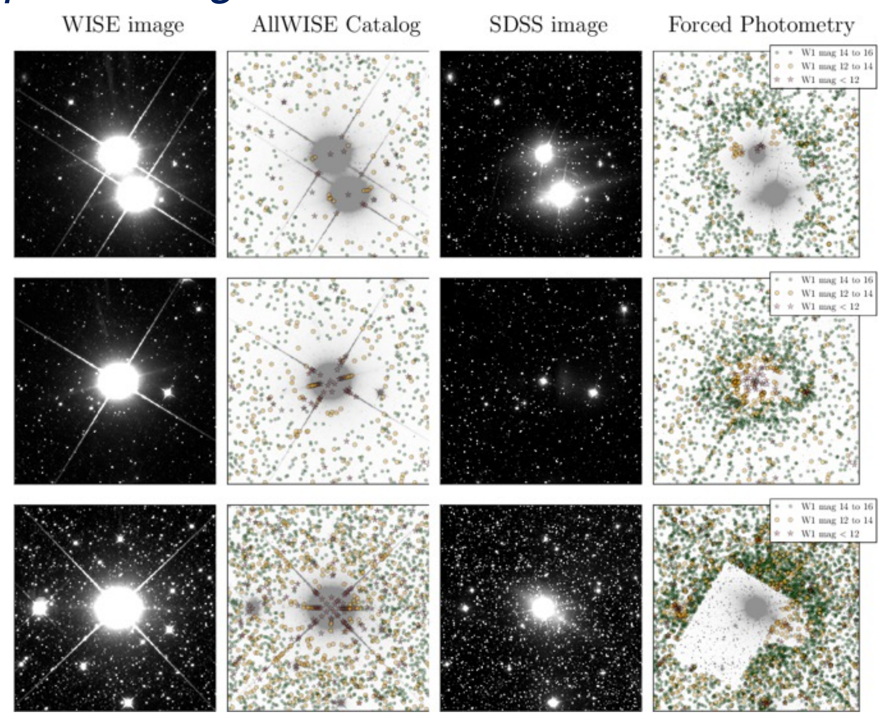
example:
vaex+vuetify on Gaia

4. Big Data Fast Survey APIs

By providing fast, optimized services for common computationally intensive tasks, Roman can allow for advanced processing workflows both within and beyond the Roman Science Platform

- Histogram API (e.g. AXS, vaex)

- Forced Photometry API



example:
Lang, Hogg, and Schlegel 2014

5. Curated Complementary Data

While Roman will be combined with almost every other astronomical data set, hosting copies of specific, well-chosen data sets for fast comparisons will dramatically amplify the scientific impact of Roman

- LSST/Rubin co-adds on *Roman* survey area
- Euclid Vis & IR co-adds on *Roman* survey area
- eROSITA on *Roman* survey area
- Subaru HSC on *Roman* survey area
- HST on *Roman* survey area