PCW2022: LINCC Data to Software to Science

Andy Connolly (UW)
Rachel Mandelbaum (CMU)
Jeremy Kubica (CMU)

White paper:
https://arxiv.org/abs/2208.02781

LINCC Data to Software to Science Agenda

- **Outcomes from the DSS meeting** (Rachel Mandelbaum and Andy Connolly)
- **Current and planned RSP functionality for time series data** (Leanne Guy and Gregory Dubois Feldsman)
- **How do we make time series data accessible to researchers?** (Mario Juric)
- **How do we analyze time series data with the RSP?** (Eric Bellm, Neven Caplar)
- **LINCC summary** (Jeno Sokoloski)

The LINCC Frameworks Project

LSST Interdisciplinary Network For Collaboration And Computing

A collaboration between UW, CMU, LSSTC, U Pitt, and NOIRLab to build software systems for key LSST science

PIs: Andy Connolly (UW), Rachel Mandelbaum (CMU)
Director of Engineering: Jeremy Kubica (CMU)

Science software infrastructure: combining user algorithms & code, astro packages, and industry tools to build scalable science analysis packages

Additional LINCC faculty here at the PCW: Mario Juric (UW), Michael Wood Vasey (Pitt)

LSST Science Pipelines

Science Platform Research

Inference

Algorithms

New LINCC Frameworks Team Members

Software Engineering Team

- Jeremy Kubica (**at PCW)**
- Carl Christofferson (TL: UW)
- Max West
- Doug Branton
- Drew Oldag
- Emmanuel Sarpong
- **4 more to come**

Project Scientists

- Colin Chandler (**at PCW**)
- Neven Caplar (**at PCW**)
- Sam Wyatt
- Alex Malz (**at PCW**)
- **1 more at CMU, to be hired**
- **2 more to come from the University of Pittsburgh**

Workshop: From Data to Software to Science with the Rubin Observatory LSST

Workshop goals:

1. Enabling *interactive development* of exciting scientific use cases for early LSST data, and identifying the common computational/technical challenges and enabling technologies associated with them.
2. Promoting the development of a broad and inclusive community of researchers engaged with LINCC Frameworks.

Program design, plenary talk content, and communication channels for the meeting were developed with both goals in mind.

https://indico.flatironinstitute.org/event/2777/

Science use cases

Divided the science into 7 research areas (not a 1:1 mapping to the LSST Science Collaborations)

- Solar System Science: 6 cases (active asteroids, TNOs)
- Local Universe Static Science: 5 cases (IMF, accreted stellar pops, dwarf gals)
- Local Universe Variable and Transient Science: 9 cases (YSO, microlensing)
- Extragalactic Static Science: 7 cases (morphologies, extinction, LSB dwarfs)
- Extragalactic Variable Science: 8 cases (AGN, lensing)
- Extragalactic Transient Science: 7 cases (SNe, TDEs, classification)
- Cosmology: 6 cases (weak lensing, SNe classification, spectroscopic followup)

~50 use cases for science in the first 2 years of Rubin

	Cross-matching	Photo-z	Selection functions	Time series	Image reprocessing	Image analysis
Cosmology	✓✓	✓✓	✓✓	✓✓	✓	✓
Extragalactic static	✓✓	✓✓	✓✓		✓✓	✓
Extragalactic transient	✓✓	✓✓	✓	✓✓	✓	✓
Extragalactic variable	✓✓	✓	✓	✓✓	✓	✓
Local Universe transient & variable	✓✓		✓	✓✓		
Local Universe static	✓✓		✓✓		✓	✓
Solar system	✓		✓✓	✓✓	✓	✓✓

Table 1. Table highlighting the connection between scientific and technical areas discussed at the workshop. Rows are science areas while columns are for infrastructure capabilities. A double checkmark (✓✓) signifies that some infrastructure capability is essential to enable a particular scientific area, while a single checkmark (✓) signifies that the infrastructure capability would enhance or expand scientific discovery within that area but is not necessary to enable all of it.

Common technical areas identified at the meeting

1. Scalable Cross-matching: real-time (low-latency) positional matching of ~10k sources to ~10 catalogs of ~1Bn sources; offline/batch match and join of ~1Bn sources to catalogs of ~1Bn sources.

2. Photometric redshifts: run and update photo-z’s tailored to specific science cases; outputting PDFs for error estimates (~10TB for LSST data); run in parallel

3. Selection function determination: build on DM selection function capabilities; extend to broad science cases (scalar and vector selection functions)

4. Scalable job execution system: run time series, image analysis, classification, model fitting at an LSST scale ~1Bn sources in parallel

Common technical areas identified at the meeting

5. Sky image access and reprocessing at scale: reprocessing of subsets of images (cutouts and full-focal plane data); requires scalable data access services, processing infrastructure, and processing software (built from DM software)

6. Object image access and analysis at scale: processing individual (object-level) images (e.g. deblending, classification); requires scalable image cutout service of arbitrary size; ability to link results to archival data; run in parallel

7. Time series analysis support infrastructure: extract features and classify the captured time-series; enable parametric and model fitting; enable anomaly detection; run in parallel; store, link, and update outputs

We want your feedback on the white paper! (https://arxiv.org/abs/2208.02781)

Does the whitepaper miss any high priority technical cases?

What gaps do you see or functionality that we should focus on?

Are you already working on any of these technical cases and infrastructure?

Are you looking to collaborate on any of the use cases?

Are there starter projects (1-3 months) that would enable science for you today?

The rest of the session will be devoted to more in-depth discussion of time series analysis

Current and planned RSP functionality for time series

Leanne Guy and Gregory Dubois Felsmann

Rubin Science Platform (RSP)

A set of integrated web applications & services deployed at Data Access Centers through which the scientific community will access, visualize, subset and perform next-to-the-data analysis of Rubin Data products.

LSST

LARGE SYNOPTIC SURVEY TELESCOPE

Large Synoptic Survey Telescope (LSST)
Systems Engineering

LSST Science Platform Vision Document

M. Jurić, D. Ciardi, G.P. Dubois-Felsmann, L.P. Guy

LSE-319

Latest Revision: 2019-07-24

https://ls.st/lsp

This LSST document has been approved as a Content-Controlled Document. Its contents are subject to configuration control and may not be changed, altered, or their provisions waived without prior approval. If this document is changed or superseded, the new document will retain the Handle designation shown above. The control is on the most recent digital document with this Handle in the LSST digital archive and not printed versions.

lse-319.lst.io

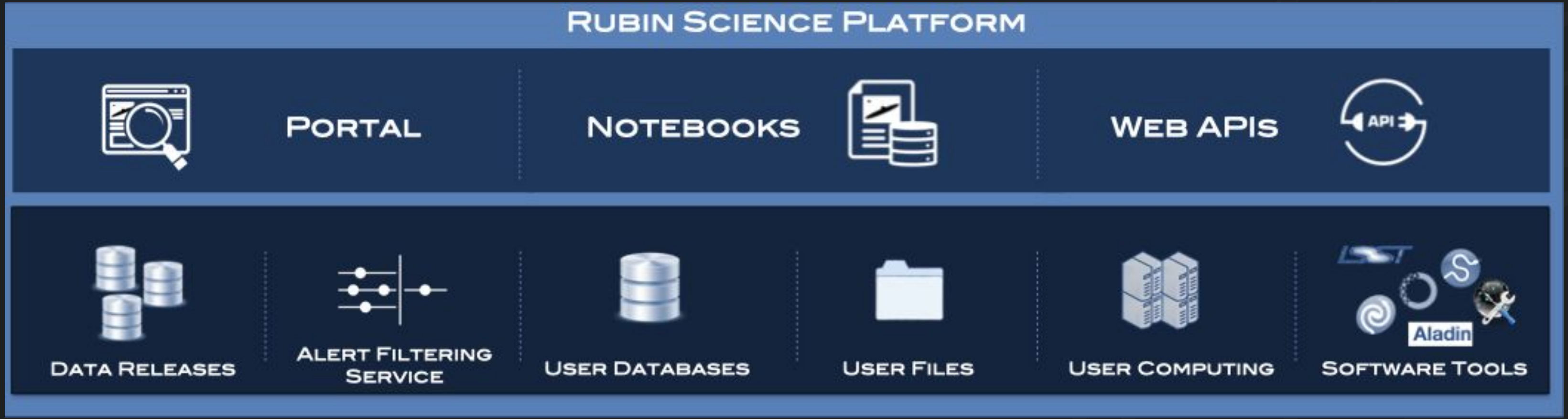
- Enable peta-scale analysis of LSST data
- Exploratory analysis via browsing & visualisation
- Enable discovery –'bring the analysis to the data'
- Supports User-Generated product creation
- Integration with extant archives via IVOA protocols
- Collaborative working environment
- Provision of backend computation & analysis resources

Rubin Science Platform – Three Aspect Design

Portal Aspect
Exploratory analysis and visualization of the LSST archive

Notebook Aspect
In-depth ‘next-to-data’ analysis and creation of added-value data products

API Aspect
Remote access to the LSST archive via Virtual Observatory interfaces



Data Preview 0 (DP0)

DP0 is the first of three planned data previews between now and Operations.

Rubin's DP0 Goals

- enable the community to prepare for early LSST science with the RSP
- test integration of the LSST science pipelines and the RSP
- use feedback on data products and RSP functionality to inform future development

DP0 Data Set

- simulated LSST-like images and catalogs from the DESC's Data Challenge 2 (DC2)
- future DP data sets will be based on LSST commissioning data from Rubin Observatory

DP0 Timeline

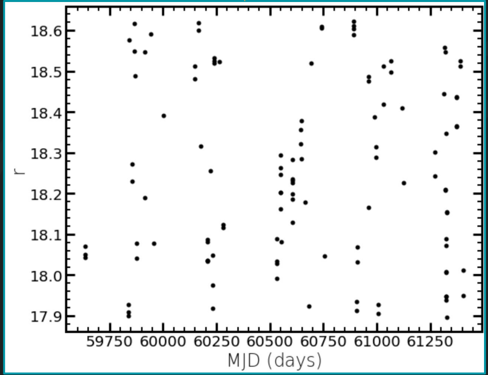
- DP0.1, June 2021: *DC2 as processed by the DESC available in the RSP*
- DP0.2, June 2022: *DC2 as reprocessed by Rubin Data Production available in RSP*

DP0: Single time series analysis

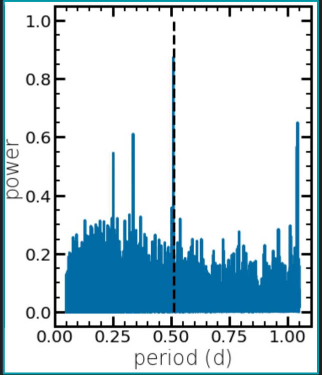
- dp02_dc2_catalogs
  - dp02_dc2_catalogs.Object
  - dp02_dc2_catalogs.Source
  - dp02_dc2_catalogs.ForcedSource
  - dp02_dc2_catalogs.DiaObject
  - dp02_dc2_catalogs.DiaSource
  - dp02_dc2_catalogs.Visit
  - dp02_dc2_catalogs.CcdVisit
  - dp02_dc2_catalogs.CoaddPatches

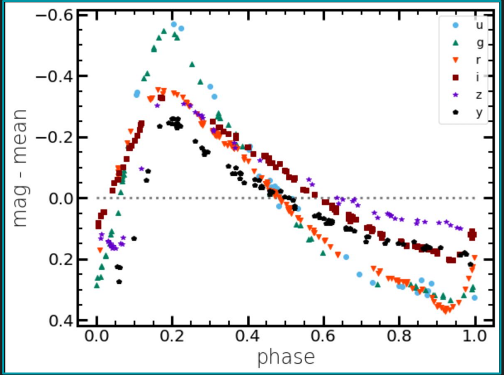
```
# ADQL TAP query joining CcdVisit & ForcedSource tables
and selecting a single Object

SELECT  src.ccdVisitId, src.band, visinfo.expMidptMJD
scisql_nanojanskyToAbMag(psfFlux) as psfMag,
FROM dp02_dc2_catalogs.ForcedSource as src
JOIN dp02_dc2_catalogs.CcdVisit as visinfo
ON visinfo.ccdVisitId = src.ccdVisitId
WHERE src.objectId = 1651589610221899038
```

	ccdVisitId	band	psfMag	expMidptMJD
29	2334102	u	20.119284	59583.120963
23	5882102	y	18.420364	59588.091815
313	7999130	z	18.357822	59591.081810
81	8030161	z	18.376561	59591.097111
323	12467085	y	18.321208	59597.089947
...	...	...	...	...



Group by band





VERA C. RUBIN OBSERVATORY

DP0: Identify candidate variables

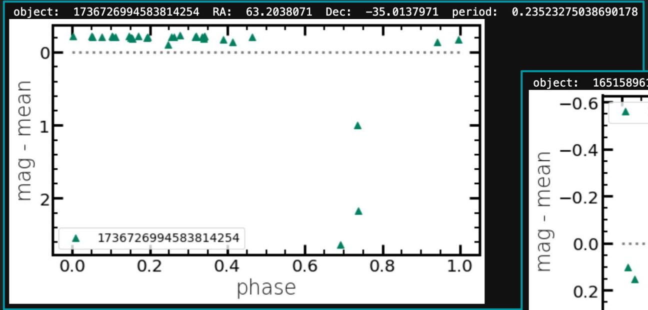
- dp02_dc2_catalogs
  - dp02_dc2_catalogs.Object
  - dp02_dc2_catalogs.Source
  - dp02_dc2_catalogs.ForcedSource
  - dp02_dc2_catalogs.DiaObject
  - dp02_dc2_catalogs.DiaSource
  - dp02_dc2_catalogs.Visit
  - dp02_dc2_catalogs.CcdVisit
  - dp02_dc2_catalogs.CoaddPatches

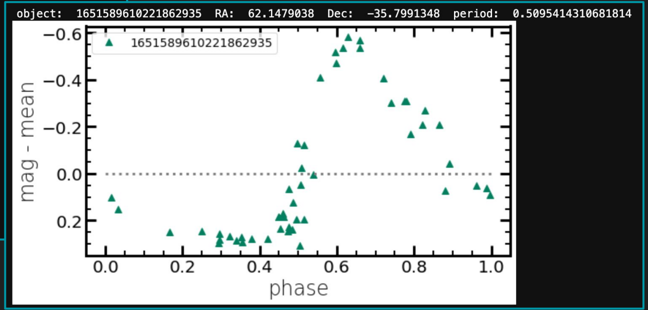
 # ADQL TAP query joining DiaObject and DiaSource tables and applying selection criteria

1. g-band measurements only
2. sigma_flux/flux > 0.25 -- the scatter in measured fluxes is larger than 25% relative to the mean
3. sigma_flux/flux < 1.25 -- the scatter in measured fluxes is no larger than 125% relative to the mean
4. 18 < gmag < 23 -- mean g magnitude between 18-23
5. gPSFluxNdata > 30 -- at least 30 observations in g band
6. gPSFluxStetsonJ > 20 -- StetsonJ index greater than 20
7. within 5 degrees of our chosen RA, Dec position

14 uniques DiaObjects that match the criteria

Compute periodograms and phase fold





Discussion questions

- What gaps do you see in tools that will be available? Which additional functionality would you like to see? What things should go directly into the RSP?
- Are there other systems besides the RSP that you are planning to use for time series analysis?

* This slide is about systems used for analysis; after subsequent presentations we'll talk about more specific analysis algorithms

How do we make time series data accessible to researchers?

Mario Juric

Discussion questions

1. What data formats do you need for lightcurves?
2. What additional metadata do you need to add (annotations from classifiers)?
3. How do you plan to share the output?

How do we analyze time series data with the RSP?

Eric Bellm, Neven Caplar

Rubin will pre-compute time series features in both Alert and Data Release Production.

Alert Production: Difference Image (DIAObject) lightcurve features

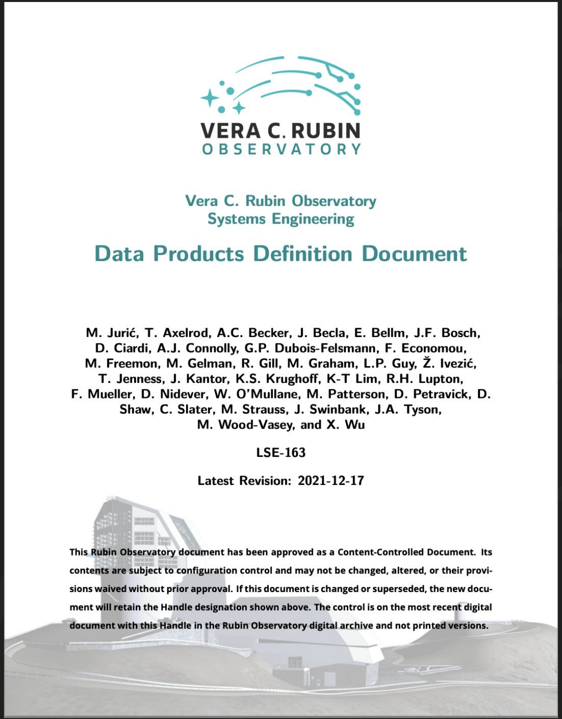
- computed on 12 months of DIASources during Prompt Processing (< 60 second latency)
- included in alerts & the Prompt Products Database

Data Releases: both difference & direct imaging (Object) features

- computed on all DIAForcedSources and ForcedSources during DRP
- included in Data Release catalogs

The Data Products Definition Document allocates space:

lcPeriodic	float[6×32]	Periodic features extracted from DIASource light-curves using generalized Lomb-Scargle periodogram [Table 4, 17][48].
lcNonPeriodic	float[6×20]	Non-periodic features extracted from DIASource light-curves [Table 5, 17].


VERA C. RUBIN OBSERVATORY

Vera C. Rubin Observatory
Systems Engineering

Data Products Definition Document

M. Jurić, T. Axelrod, A.C. Becker, J. Becla, E. Bellm, J.F. Bosch, D. Ciardi, A.J. Connolly, G.P. Dubois-Felsmann, F. Economou, M. Freemon, M. Gelman, R. Gill, M. Graham, L.P. Guy, Ž. Ivezić, T. Jenness, J. Kantor, K.S. Krughoff, K-T Lim, R.H. Lupton, F. Mueller, D. Nidever, W. O'Mullane, M. Patterson, D. Petravick, D. Shaw, C. Slater, M. Strauss, J. Swinbank, J.A. Tyson, M. Wood-Vasey, and X. Wu

LSE-163

Latest Revision: 2021-12-17

This Rubin Observatory document has been approved as a Content-Controlled Document. Its contents are subject to configuration control and may not be changed, altered, or their provisions waived without prior approval. If this document is changed or superseded, the new document will retain the Handle designation shown above. The control is on the most recent digital document with this Handle in the Rubin Observatory digital archive and not printed versions.

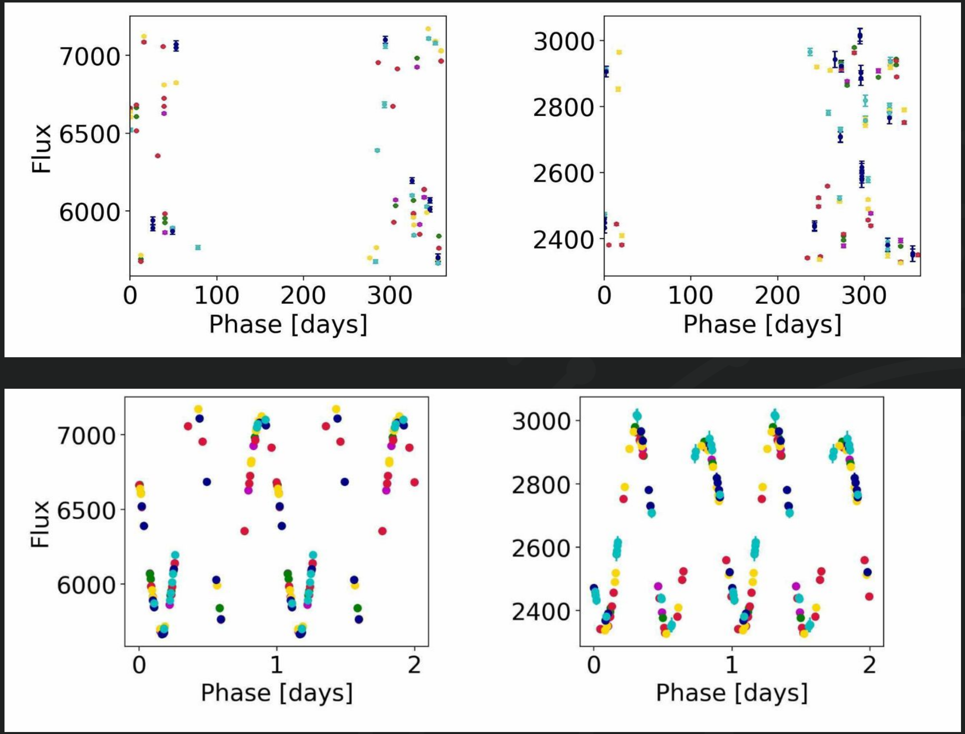
ls.st/dpdd

In AP, features will be constrained by latency, data, and compute.

High-level 60 second latency requirement ⇒

- 12 months of history = ~80 epochs total (6 filters)
- Limited CPU & memory
- Only a few seconds to compute features!

Still enough data for features useful for query, alert filtering, user classification



Andy Tzanidakis

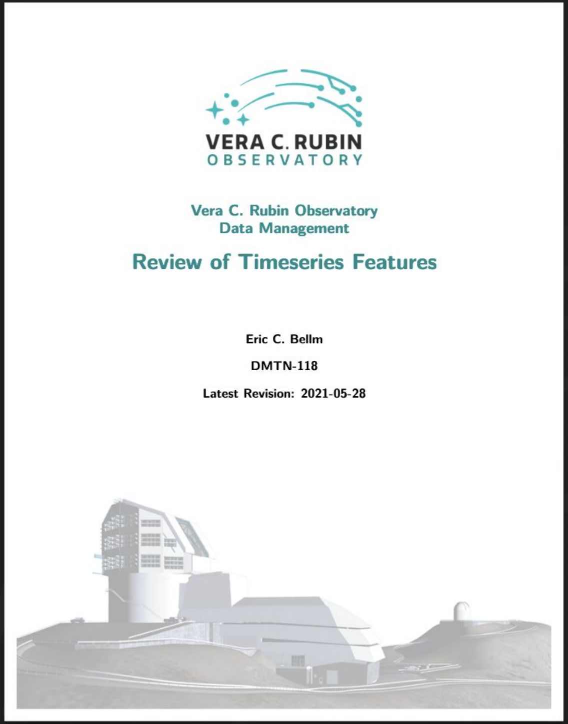
Rubin is working to develop a feasible and useful feature set for AP.

Provide general-purpose feature set:

- Generic summary statistics
- Basic period estimation
- Transient parameterization
- Charaterize aperiodic variability

DMTN-118 discusses technical considerations & open questions

ls.st/fkr is a work-in-progress draft feature set; see discussion on community.lsst.org


VERA C. RUBIN OBSERVATORY

Vera C. Rubin Observatory
Data Management

Review of Timeseries Features

Eric C. Bellm

DMTN-118

Latest Revision: 2021-05-28

DMTN-118

DRP features are likely a superset of AP's.

DRP feature computation is less constrained:

- Computed on forced photometry from entire survey
- More flexibility with computation environment and latency

DRP features will likely evolve from Data Release to Data Release based on community usage and feedback.

Expect we'll start with the AP feature set in DP2-DR1 era.

Good opportunity for discussion/exchange with LINCC.

Places LINCC can help users compute richer features than Rubin can deliver.

Specialized science cases

- E.g., searches for very short period binaries; changing periods

Computationally intensive tasks

- Large-scale fitting of template lightcurves
- MCMC

GPU implementations

Rubin is not trying to develop a general-purpose timeseries library but we are interested in discussions with others who might be

Initial time series projects

- **What are the requirements for a good candidate?**
  - Code applicable for LSST science and scale
  - Active developers/groups using the code (I.e., good and readily deployable ideas)
  - Possibility for a tight interaction between scientists and programmers

We have identified 31 existing timeseries codes in these broad areas:

- **Explosive transients**
  - *Transient Classifiers*
  - *Lightcurve fitting for SN standardization*
- **Lensing**
  - *Microlensing*
  - *Strong lensing*
- **AGN**
- **Periodograms**

Examples of ideas by current maintainers

- **Explosive transients**
  - SuperNNova
    - Implement unit and integration tests
    - Major rewrite to make compatible with pytorch updates
    - Optimize running of the code on the alert type data products
- **Lensing**
  - Lensastronomy
    - JAX and adaptive mesh supported micro-lensing code; JAX for established macro-lensing codes
- **AGN**
  - EzTao
    - Optimize current JAX implementation, stress testing
    - Mutliband analysis
- **Periodograms**
  - Astropy implementations
    - Add possibility for mutliband analysis

Discussion questions

1. What features do you need to measure on time series data?
2. What algorithms do you want to run on time series data (clustering, modeling, etc.)? What software packages are you currently using that apply these algorithms to time series data?
3. How will you sample lightcurves for analysis (uniform random, based on features, no-sampling, etc.)?

LINCC summary

Jeno Sokoloski

LINCC Frameworks is a key pillar of LINCC

An LSSTC initiative

Goal of LINCC: provide the astrophysics community with the tools, training, and collaborative opportunities – beyond and complementary to those provided by the federally funded project – to enable Rubin LSST to fulfill its potential.

Strategy: with input from LSSTC member institutions, the SCs, and the broader astrophysics community, seek private funding to build programs that have broad community impact and could not be carried out by a single university or PI.

Launched: LINCC Frameworks and the LSSTC Catalyst Fellowship Funded by the John Templeton Foundation.

More to come!

RSP Requirements and Design Documents

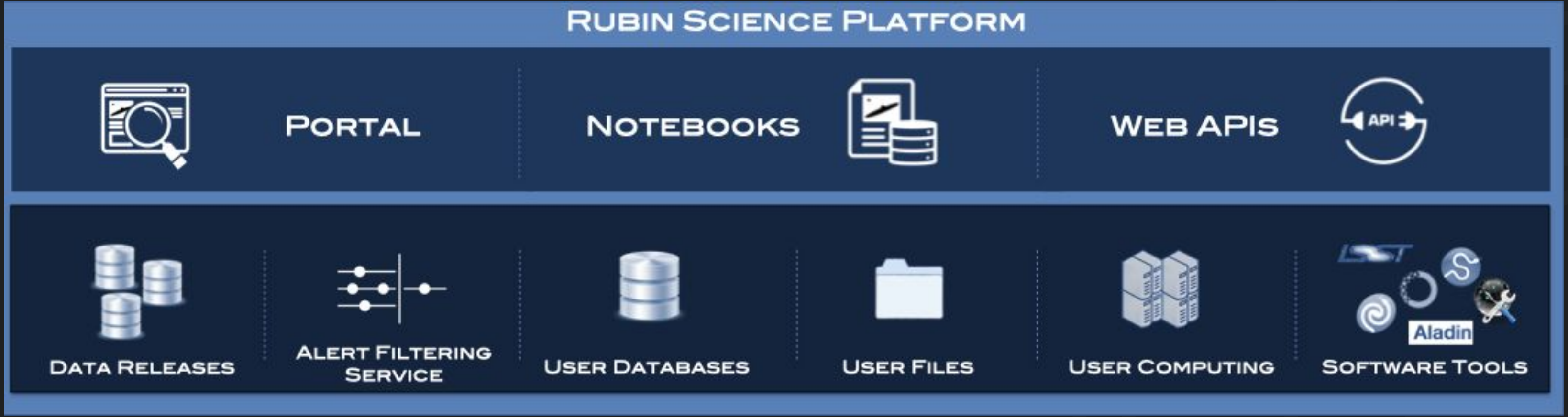
- LSE-319: Science Platform Vision Document
- LSE-61: Data Management System Requirements
- LDM-554: Data Management LSST Science Platform Requirements
- LDM-542: Science Platform Design
- DMTN-202: Use cases and science requirements on a user batch facility
- DMTN-086: Next-to-the-Database Processing Use Cases

Rubin Science Platform – Three Aspect Design

Portal Aspect
Exploratory analysis and visualization of the LSST archive

Notebook Aspect
In-depth ‘next-to-data’ analysis and creation of added-value data products

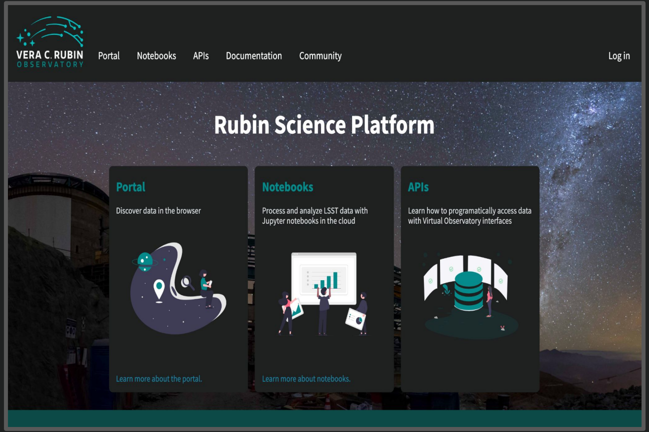
API Aspect
Remote access to the LSST archive via Virtual Observatory interfaces



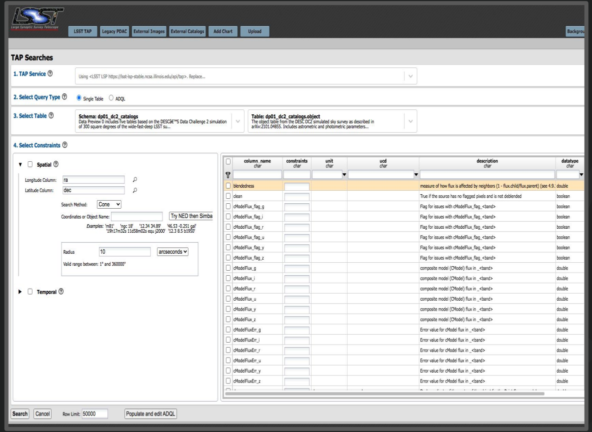
The DP0-Era Rubin Science Platform

The DP0-era RSP provides delegates with access to the data set via the Portal, Notebook, and API Aspects. All three aspects have tools to query, subset, visualize, and analyze the DP0 data set, as well as documentation and tutorials for users. The LSST Science Pipelines (and many other common software packages) are pre-installed in the Notebook environment.

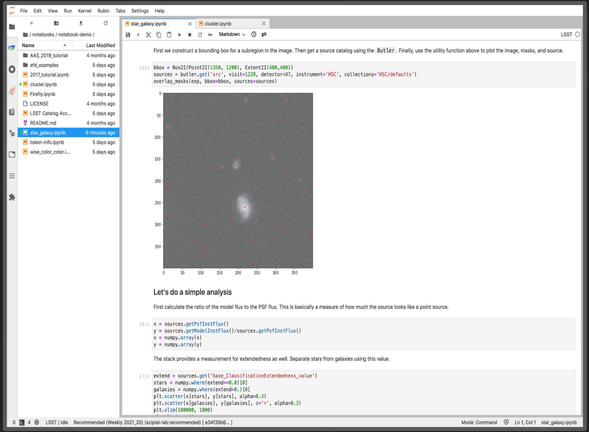
RSP Landing Page



RSP Portal Aspect



RSP Notebook Aspect



Data Preview Schedule and Data Products

Rubin Baseline Data Release Scenario	Jun 2021	Jun 2022	Mar 2024 - Jul 2024	Dec 2024 - Mar 2025	Oct 2025 - Jan 2026	Oct 2026 - Jan 2027	Nov 2027 - Jan 2028	Oct 2028 - Jan 2029
	DP0.1	**DP0.2**	**DP1**	**DP2**	**DR1**	**DR2**	**DR3**	**DR4**
Data Product	**DC2 Simulated Sky Survey**	**Reprocessed DC2 Survey**	**ComCam On-Sky Data**	**LSSTCam On-Sky Data**	**LSST First 6 Months Data**	**LSST Year 1 Data**	**LSST Year 2 Data**	**LSST Year 3 Data**
Raw images	☑	☑	☑	☑	☑	☑	☑	☑
DRP Processed Visit Images and Visit Catalogs	☑	☑	☑	☑	☑	☑	☑	☑
DRP Coadded Images	☑	☑	☑	☑	☑	☑	☑	☑
DRP Object and ForcedSource Catalogs	☑	☑	☑	☑	☑	☑	☑	☑
DRP Difference Images and DIASources	☐	☑	☑	☑	☑	☑	☑	☑
DRP ForcedSource Catalogs including DIA outputs	☐	☑	☑	☑	☑	☑	☑	☑
PP Processed Visit Images	☐	☐	☑	☑	☑	☑	☑	☑
PP Difference Images	☐	☐	☑	☑	☑	☑	☑	☑
PP Catalogs (DIASources, DIAObjects, DIAForcedSources)	☐	☐	☑	☑	☑	☑	☑	☑
PP Alerts (Canned)	☐	☐	☑	☑	☑	☑	☑	☑
PP Alerts (Live, Brokered)	☐	☐	☐	☑	☑	☑	☑	☑
PP SSP Catalogs	☐	☐	☑	☑	☑	☑	☑	☑
DRP SSP Catalogs	☐	☐	☐	☐	☑	☑	☑	☑

Image credit: Phil Marshall

Getting Help - Community Forum