




# 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 analyze time series data with the RSP?** (Eric Bellm, Neven Caplar)
- **How do we make time series data accessible to researchers?** (Mario Juric, Colin Slater)
- **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

VERA C. RUBIN OBSERVATORY

Science Platform Research

Inference

Algorithms

https://www.lsstcorporation.org/lincc/frameworks

# 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 remotely**)
- **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.


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

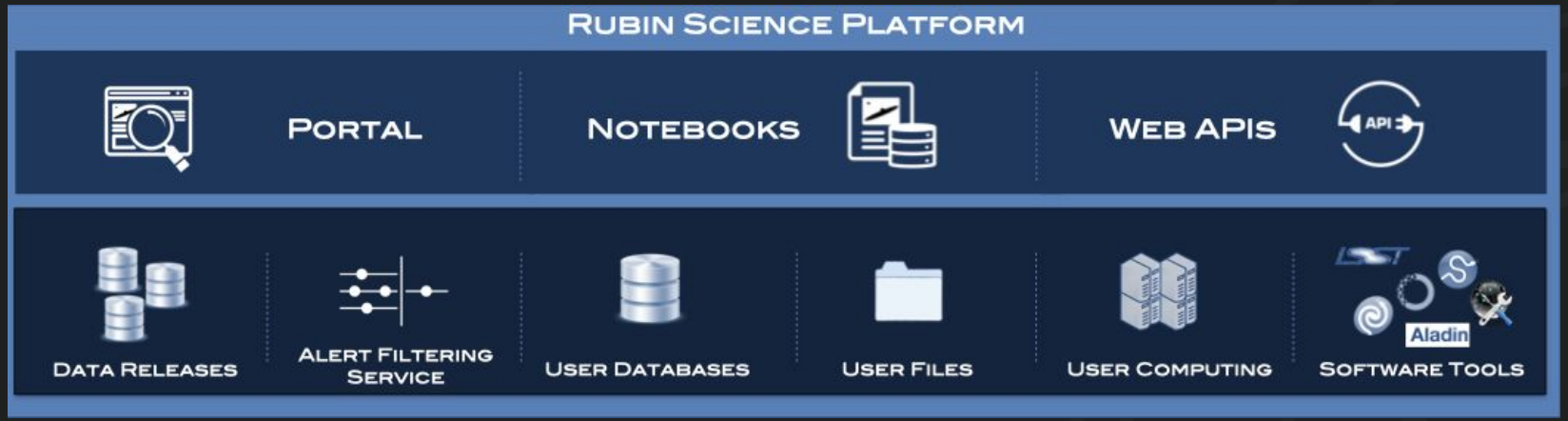
- Enable peta-scale analysis of LSST data
- Exploratory analysis via browsing & visualisation
- Enable discovery –'bring the analysis to the data'
- Supports User-Generated product creation
- Integration with extant archives via IVOA protocols
- Collaborative working environment
- Provision of backend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

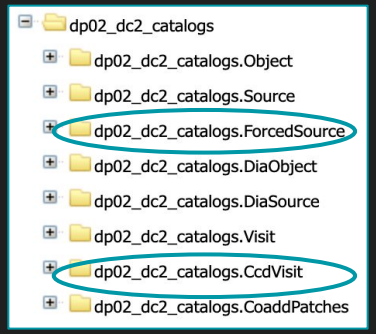
**DP0 Data Set**

- 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**

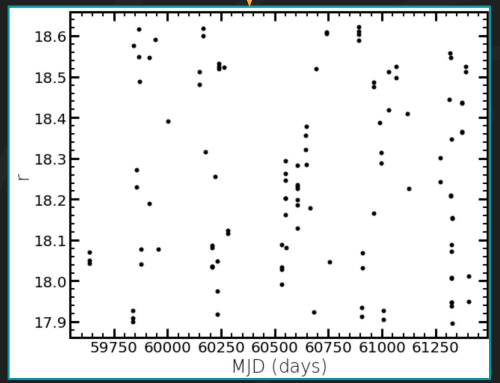
- DP0.1, June 2021: *DC2 as processed by the DESC available in the RSP*
- DP0.2, June 2022: *DC2 as reprocessed by Rubin Data Production available in RSP*

# DP0: Single time series analysis

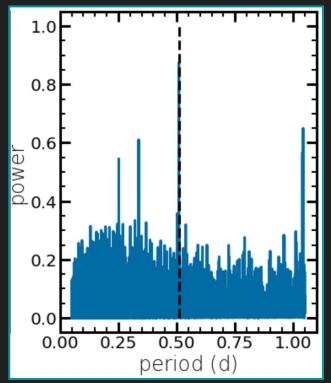


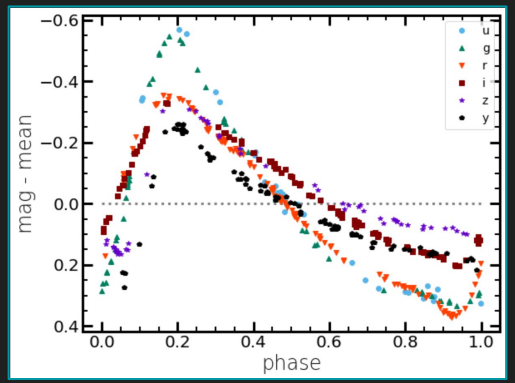

```
# ADQL TAP query joining CcdVisit & ForcedSource tables
and selecting a single Object

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

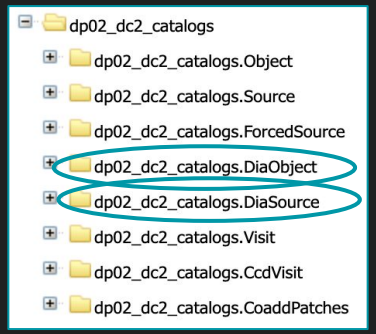





*Notebook credit: Jeff Carlin, : 07b_Variable_Star_Lightcurves*

# DP0: Identify candidate variables

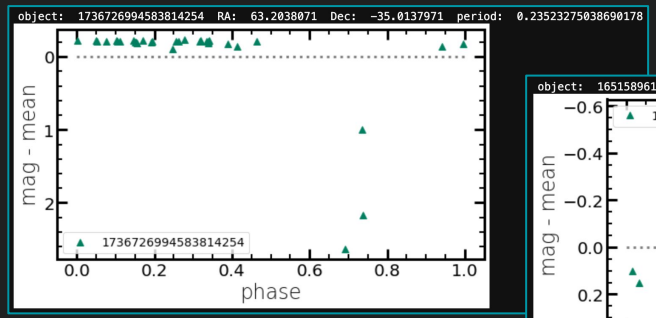


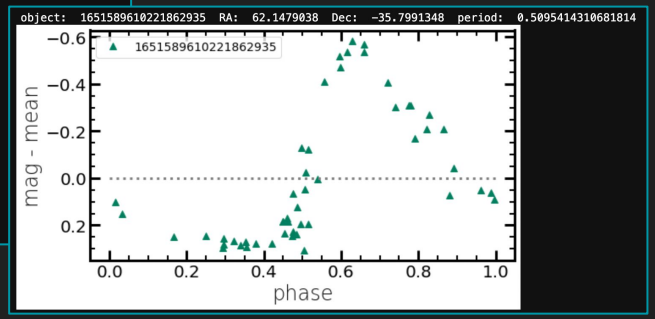

**# ADQL TAP query joining DiaObject and DiaSource tables and applying selection criteria**

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






*Notebook credit: Jeff Carlin, : 07b_Variable_Star_Lightcurves*

VERA C. RUBIN OBSERVATORY

# Time Series in the RSP

Gregory Dubois-Felsmann, Caltech/IPAC
RSP Product Owner

Rubin Observatory PCW, 11 August 2022

NSF AURA U.S. DEPARTMENT OF ENERGY SLAC CHARLES AND LISA SIMONYI FUND FOR ARTS AND SCIENCES LSST CORPORATION

# Time Series in the LSST Data Model

At the most basic level, there isn't a specific "time series" entity in the DPDD data model.

Instead, time series are generally represented by a combination of information from:

- A *single* row in a summary-over-time table:
  Object, DIAObject, SSObject
- A corresponding *set* of rows in a row-per-epoch table:
  ForcedSource, Source, DIASource, SSSource

Matching elements are linkable by the use of object IDs as foreign keys in JOIN operations.

# DIAObject to DIASources link on DP0.2

dp02_dc2_catalogs.DiaObject...

1 of 3 (1 - 10

| diaObjectId long | ra (deg) double | decl (deg) double | nDiaSources long | radecTai double | gPSFluxMean double | gPSFluxMeanErr double |
|---|---|---|---|---|---|---|
| 1736269597746659454 | 60.4311859 | -35.0911744 | 232 | 61394.1793962 | 249.1326373 | 89.5250095 |
| 1736269597746659447 | 60.3953983 | -35.1199186 | 178 | 61394.1793962 | 2149.6076468 | 76.5249222 |
| 1736269597746659442 | 60.3718296 | -35.0977601 | 152 | 61394.1615222 | -2095.7588751 | 110.8328194 |
| 1736269597746659594 | 60.3880964 | -35.1184075 | 74 | 61394.1606082 | 1757.4983003 | 65.7734034 |
| 1736269597746659751 | 60.3722054 | -35.1076471 | 29 | 61394.1606082 | -504.5312189 | 112.1243113 |

dp02_dc2_catalogs.DiaObject - data... | dp02_dc2_catalogs.DiaSource ...

1 of 1 (1 - 74 of 74)

| diaObjectId long | diaSourceId long | filterName char | midPointTai double | psFlux (nJy) double | psFluxErr (nJy) double | apFlux (nJy) double | apFluxErr (nJy) double | snr double | ccdVisitId long | ra (deg) double | decl (deg) double |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1736269597746659594 | 21650450713411730 | r | 59634.0870892 | -162.4165258 | 493.3907271 | -474.4172627 | 566.1506331 | -0.83797 | 40327107 | 60.3880926 | -35.1184083 |
| 1736269597746659594 | 103695629347192947 | r | 59839.3345202 | 782.4670571 | 491.0154481 | -463.4886483 | 682.508526 | -0.6790958 | 193148161 | 60.3880885 | -35.1184321 |
| 1736269597746659594 | 104089161631269026 | r | 59840.2634672 | 129.6234774 | 481.6366517 | -103.3563425 | 570.0203827 | -0.1813204 | 193881172 | 60.3880992 | -35.1184027 |
| 1736269597746659594 | 113880060884156582 | r | 59867.2389572 | -606.504622 | 498.3523518 | -711.1592372 | 816.5674336 | -0.8709131 | 212118143 | 60.3880568 | -35.1184211 |
| 1736269597746659594 | 114629457515380850 | r | 59869.1437982 | 1377.7409913 | 467.870113 | 169.5410355 | 544.6786202 | 0.311268 | 213514003 | 60.3880994 | -35.1183974 |
| 1736269597746659594 | 115139499377295429 | i | 59870.2191692 | 2268.7542599 | 451.2647375 | 1791.4775525 | 771.5089272 | 2.3220444 | 214464030 | 60.3882083 | -35.1183562 |
| 1736269597746659594 | 115190038794338442 | i | 59870.2628022 | 2547.3883046 | 499.9141078 | 114.1604098 | 817.3599102 | 0.1396697 | 214558167 | 60.3882059 | -35.1183591 |
| 1736269597746659594 | 117672458487595165 | g | 59876.1526762 | 2310.4245603 | 358.1040394 | 836.1425164 | 392.5975848 | 2.1297701 | 219182034 | 60.3880227 | -35.1184146 |
| 1736269597746659594 | 118074634930225305 | r | 59877.1308142 | -463.2969718 | 465.400646 | -1325.8760761 | 546.9054646 | -2.4243245 | 219931146 | 60.3881175 | -35.118393 |
| 1736269597746659594 | 130193935812264085 | r | 59914.1503912 | -2787.6273709 | 458.2782084 | -1526.8971649 | 668.821203 | -2.2829681 | 242505103 | 60.3881838 | -35.1184036 |
| 1736269597746659594 | 145986031806578912 | g | 59958.1119582 | 1715.5003027 | 379.7710178 | 460.1891598 | 387.3741091 | 1.1879709 | 271920174 | 60.3880918 | -35.1184118 |
| 1736269597746659594 | 146004282733232378 | g | 59958.1288562 | 585.2172472 | 375.220935 | 19.4731102 | 386.4123448 | 0.0503946 | 271954169 | 60.3881031 | -35.1184134 |
| 1736269597746659594 | 220646458704003180 | i | 60162.3682742 | 1272.4296824 | 678.9425943 | 1674.0205101 | 1142.1047422 | 1.465733 | 410986056 | 60.3881125 | -35.1183818 |

# Searchable by position as well as object ID

- Searches based on *Object IDs rely on the pipeline software having performed associations on the single-epoch detections (Source, DIASource)
  - ForcedSource, of course, has an inherently “perfect” association – the measurements were taken explicitly driven by the parent object
- There are judgement calls inherent in those associations
- You can get an independent take on that by searching the single-epoch tables by position instead of ID
- The tables are/will be indexed to facilitate access either way

# Data are not segregated by filter

- The “Object-like” tables are not only summaries over time but also over filter
- The associated “Source-like” tables will generally contain data from multiple filters (a single filter per row)
- Plotting or otherwise analyzing time-series data requires dividing the single-epoch data by filter

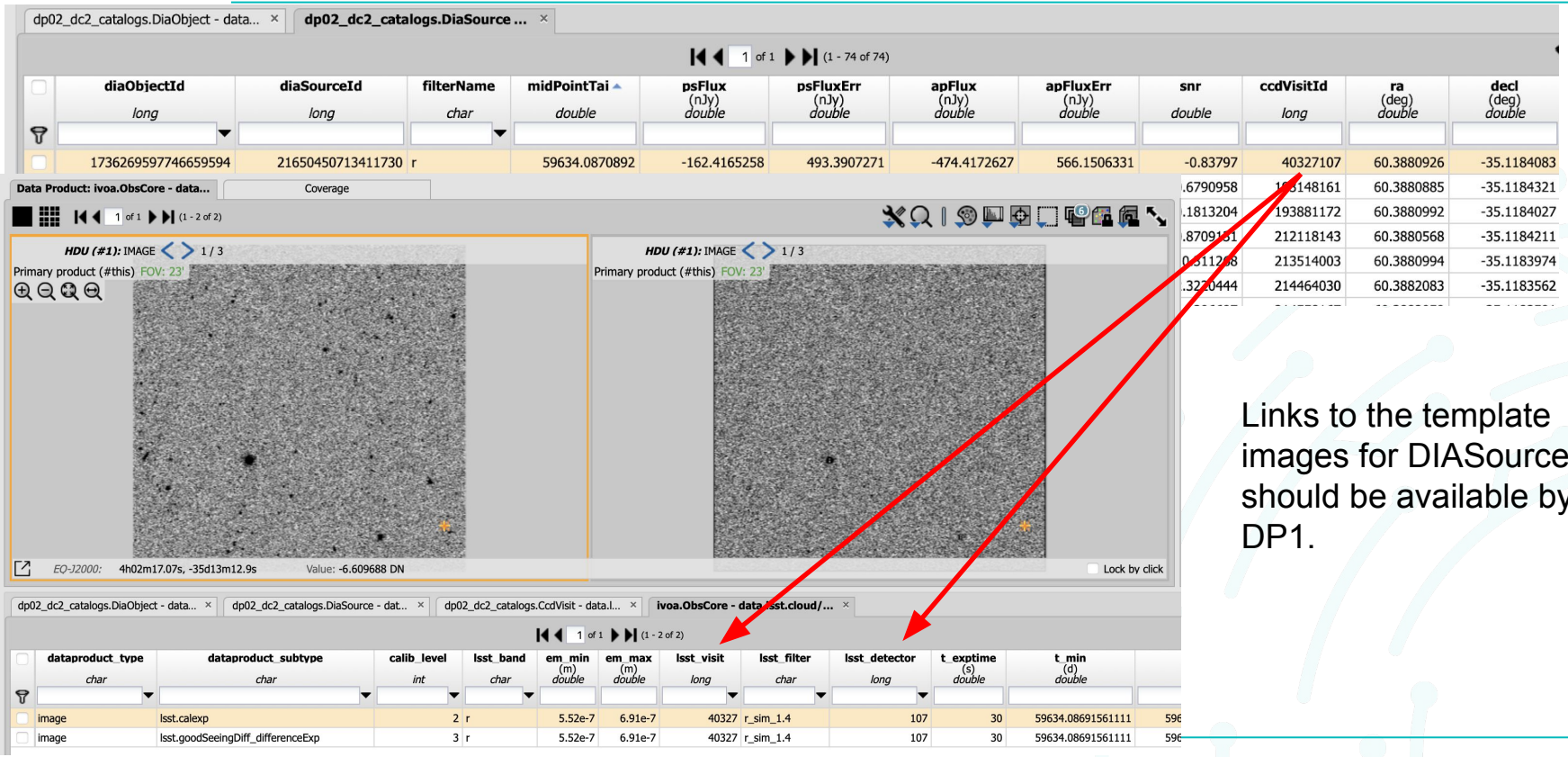
# Variability metrics and features

- The “Object-like” tables are planned to contain summary data on variability
  - E.g., in the DPDD for Object:

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

- Eric Bellm is speaking about the development of these metrics in this session

# Links to images




Links to the template images for DIASources should be available by DP1.

# … and links to image metadata

dp02_dc2_catalogs.DiaObject - data... | dp02_dc2_catalogs.DiaSource ...

1 of 1 (1 - 74 of 74)

| diaObjectId long | diaSourceId long | filterName char | midPointTai double | psFlux (nJy) double | psFluxErr (nJy) double | apFlux (nJy) double | apFluxErr (nJy) double | snr double | ccdVisitId long | ra (deg) double | decl (deg) double |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1736269597746659594 | 21650450713411730 | r | 59634.0870892 | -162.4165258 | 493.3907271 | -474.4172627 | 566.1586331 | -0.83797 | 40327107 | 60.3880926 | -35.1184083 |
| 1736269597746659594 | 103695629347192947 | r | 59839.3345202 | 782.4670571 | 491.0154481 | -463.4886483 | 682.508526 | -0.6790958 | 193148161 | 60.3880885 | -35.1184321 |
| 1736269597746659594 | 104089161631269026 | r | 59840.2634672 | 129.6234774 | 481.6366517 | 103.3563425 | 570.0203827 | -0.1813204 | 193881172 | 60.3880992 | -35.1184027 |
| 1736269597746659594 | 113880060884156582 | r | 59867.2389572 | -606.504622 | 498.3533518 | -711.1592372 | 816.5674336 | -0.8709131 | 212118143 | 60.3880568 | -35.1184211 |
| 1736269597746659594 | 114629457515380850 | r | 59869.1437982 | 1377.7409913 | 467.870113 | 169.5410355 | 544.6786202 | 0.311268 | 213514003 | 60.3880994 | -35.1183974 |
| 1736269597746659594 | 115139499377295429 | i | 59870.2191692 | 2268.7542599 | 451.2647375 | 1791.4775525 | 771.5089272 | 2.3220444 | 214464030 | 60.3882083 | -35.1183562 |

dp02_dc2_catalogs.DiaObject - data... | dp02_dc2_catalogs.DiaSource - dat... | dp02_dc2_catalogs.CcdVisit - d...

1 of 1 (1 - 1 of 1)

| band char | ccdVisitId long | darkTime (s) double | detector long | expMidpt char | expMidptMJD (d) double | expTime (s) double | obsStart char | obsStartMJD double | physical_filter char | psfSigma (pixel) float | seeing (arcsec) double | skyBg (DN) float | skyNoise (DN) float | skyRotation (deg) double |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| r | 40327107 | 30 | 107 | 2022-02-24 02:05:24.5089 | 59634.0870892 | 30 | 2022-02-24 02:05:09.5089 | 59634.0869156 | r_sim_1.4 | 1.81463 | 0.8544857 | 988.611 | 38.2808 | 86.8747148 |

- Seeing, PSF metrics, background levels, etc.

# Data Access

Fundamentally, tabular data in the RSP will be available in two ways:

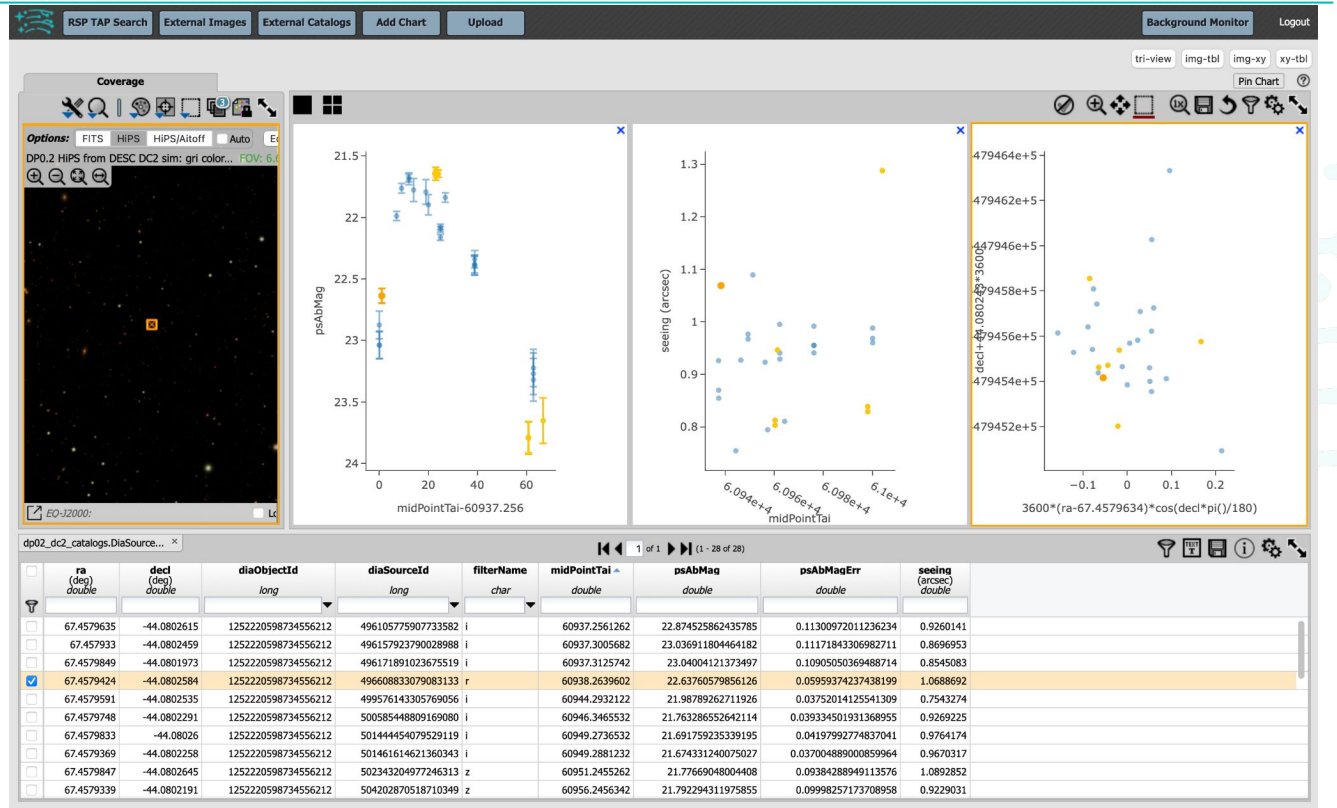
- Via ADQL database queries, mediated by an IVOA TAP service
  - This is part of the “API Aspect” of the Science Platform
  - It is accessible from both the Portal Aspect and the Notebook Aspect (as illustrated by DP0.2 tutorials)

```
SELECT diaObjectId,diaSourceId,filterName,midPointTai,psFlux,psFluxErr,apFlux,apFluxErr,snr,ccdVisitId,ra,decl
  FROM dp02_dc2_catalogs.DiaSource WHERE (diaObjectId = 1736269597746659594)
```

- As spatially-sharded Parquet files, accessible via the “Butler” middleware
  - Internally, this is aimed at the Notebook Aspect and at usability in the user batch system
  - Externally, we are working with the IVOA on means to support this via appropriate metadata and API Aspect services – mainly for access from beyond the RSP

# Example from DP0.2

From the DP0.2 Portal Lightcurve tutorial




| ra (deg) double | decl (deg) double | diaObjectId long | diaSourceId long | filterName char | midPointTai double | psAbMag double | psAbMagErr double | seeing (arcsec) double |
|---|---|---|---|---|---|---|---|---|
| 67.4579635 | -44.0802615 | 1252220598734556212 | 4961057759077335 82 | i | 60937.2561262 | 22.874525862435785 | 0.11300972011236234 | 0.9260141 |
| 67.457933 | -44.0802459 | 1252220598734556212 | 4961579237900289 88 | i | 60937.3005682 | 23.036911804464182 | 0.11171843306982711 | 0.8696953 |
| 67.4579849 | -44.0801973 | 1252220598734556212 | 4961718910236755 19 | i | 60937.3125742 | 23.04004121373497 | 0.10905050369488714 | 0.8545083 |
| 67.4579424 | -44.0802584 | 1252220598734556212 | 4966088330790831 33 | r | 60938.2639602 | 22.63760579856126 | 0.05959374237438199 | 1.0688692 |
| 67.4579591 | -44.0802535 | 1252220598734556212 | 4995761433057690 56 | i | 60944.2932122 | 21.98789262711926 | 0.03752014125541309 | 0.7543274 |
| 67.4579748 | -44.0802291 | 1252220598734556212 | 5005854488091690 80 | i | 60946.3465532 | 21.763286552642114 | 0.039334501931368955 | 0.9269225 |
| 67.4579833 | -44.08026 | 1252220598734556212 | 5014444540795291 19 | i | 60949.2736532 | 21.691759235339195 | 0.04197992774837041 | 0.9764174 |
| 67.4579369 | -44.0802258 | 1252220598734556212 | 5014616146213603 43 | i | 60949.2881232 | 21.674331240075027 | 0.037004889000859964 | 0.9670317 |
| 67.4579847 | -44.0802645 | 1252220598734556212 | 5023432049772463 13 | z | 60951.2455262 | 21.77669048004408 | 0.093842889491135 76 | 1.0892852 |
| 67.4579339 | -44.0802191 | 1252220598734556212 | 5042028705187103 49 | z | 60956.2456342 | 21.792294311975855 | 0.09998257173708958 | 0.9229031 |

# Data Access philosophy

- You can think of these, TAP/ADQL access and Parquet-shard access, as “query” and “bulk” analysis interfaces, respectively.
    - However, remember that the Qserv database behind the TAP system is optimized for performing full-table scans efficiently, not just “targeted”, indexed searches, so some types of bulk analysis will work just fine as TAP queries.
    - The value of this TAP access is very strongly tied to the usefulness of the object-level metrics and feature extraction that we are able to do, because…
    - It is still not thought to be realistic to do full time-series analysis (e.g., template-based searches, periodograms, etc.) in the database, so…
- A robust system for analyzing the Parquet data in bulk is essential; and
- Your input on appropriate feature analyses to perform would be most valuable.

# API Aspect direct support for time series

- In addition to the basic ability to query the tables via TAP, the RSP will include value-added services in the API Aspect to retrieve the Source-like time-series data based on an object-like ID without the user having to construct ADQL
  - In other words, interfaces with URL query parameters like “?OBJECTID=1736269597746659594”
- We will also have services to return image data and metadata based on a source
  - These will generally just be trivial convenience wrappers for TAP queries
- Queries on the Object-like and Source-like tables will contain IVOA DataLink annotations that enable client software (e.g., PyVO) to follow these links easily
- The Portal Aspect will provide “one-click” UI support for this
- Prototypes will be deployed in DP0.2 later this year - stay tuned for tutorials!

# Future directions for the APIs - I

- At present there have been many proposals, but there is no fully worked-out community standard for a “time series” query result that combines summary/feature information with the actual time-dependent photometry
- We are engaged with the IVOA in standards development efforts and will actively promote progress in this area
- We have a service architecture in the RSP that will facilitate rapid prototyping, and following a standard if/when it develops
- In the mean time, as we move to the DR1 era we will progressively annotate queries on the time series tables with more metadata, following existing standards

# Future directions for the APIs - II

- The service architecture will easily allow for additional, value-added services like periodogram calculation to be linked to light curve data in the API Aspect - and then easily discovered from the Portal Aspect.
  - We expect to enhance the environment along these lines during operations.
  - LINCC contributions could be made available to users of all Aspects by wrapping them in this (Python) framework

- Project requirements also call for us to provide a forced-photometry-on-demand service. This will also be readily available through the API Aspect and Portal Aspect, in addition to being straightforward to code using the Science Pipelines stack in the Notebook Aspect.

# Future directions for the APIs - III

- We have been communicating with the IVOA on formalizing support for Parquet in a way that preserves interoperability with existing and future IVOA metadata and data models…
- And that facilitates access to very large, wide-area spatially-sharded catalog datasets through existing services like ObsTAP and the Registry.

- Mario Juric will also be talking about this

# Computation on time series in the RSP

- We are obligated by project requirements to provide a bulk-data-analysis environment for Rubin/LSST data users
- This was deliberately left vaguely defined in early project design documents because we recognized how rapidly the software and big-data-architecture landscape was changing
- We are only explicitly committed to provide a “conventional batch” service
- However, we recognize the importance of providing a structured environment in the RSP for bulk catalog analysis (e.g., Dask, Spark), and we have successfully prototyped Dask access from the RSP. This is still under investigation and we will have more to say about this at future meetings.

# 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 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

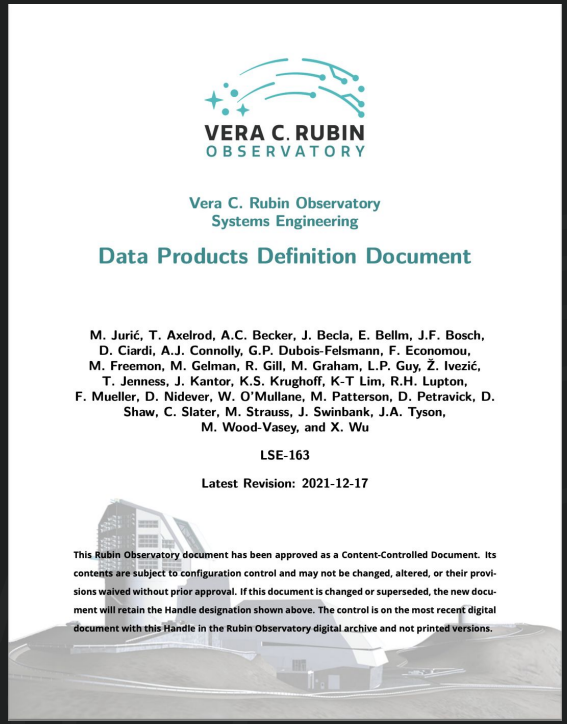
- computed on 12 months of DIASources during Prompt Processing (< 60 second latency)
- included in alerts & the Prompt Products Database

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.

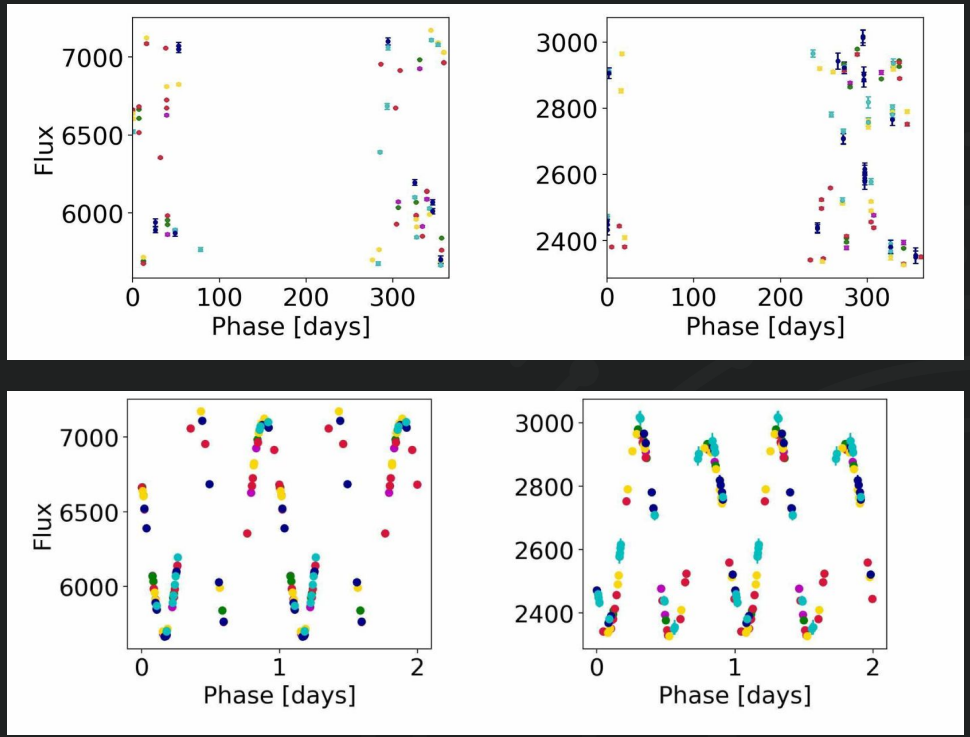
ls.st/dpdd

# 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

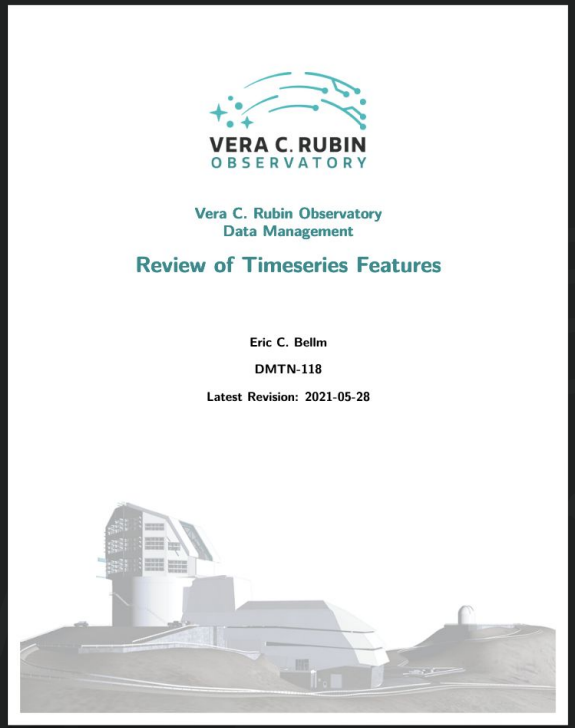
# Rubin is working to develop a feasible and useful feature set for AP.

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




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

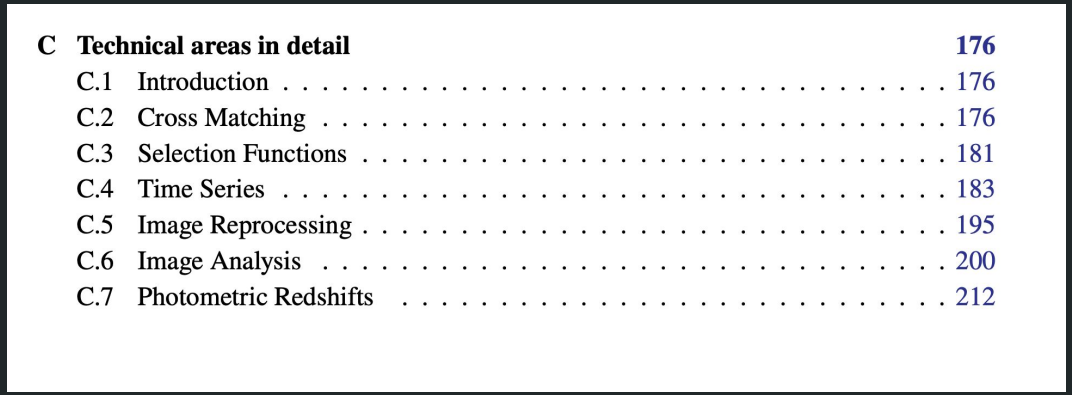
- **Explosive transients**
  - SuperNNova
    - Implement unit and integration tests
    - Major rewrite to make compatible with pytorch updates
    - Optimize running of the code on the alert type data products
- **Lensing**
  - Lensastronomy
    - JAX and adaptive mesh supported micro-lensing code; 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.)?

How do we make time series data accessible to researchers for large-scale analyses?

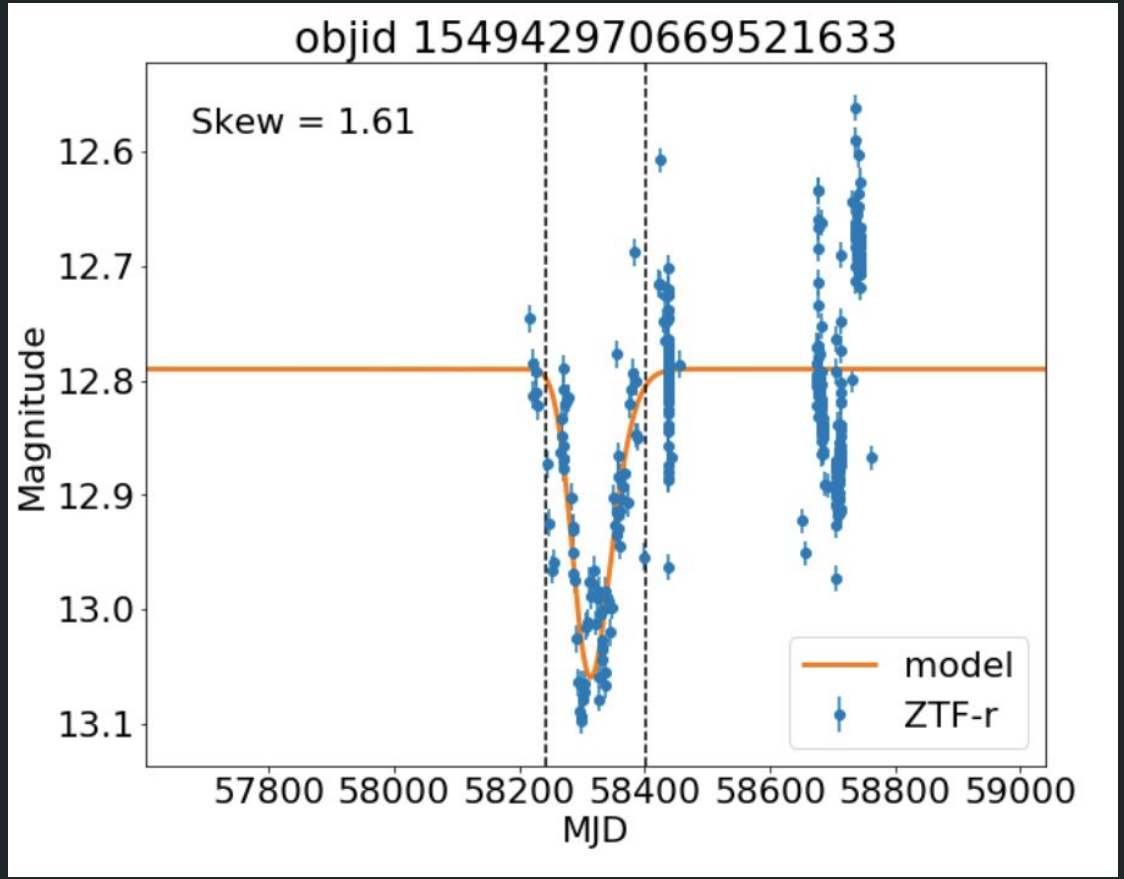
Mario Juric & Colin Slater



# Example

Left: a search for “dipper” objects (stars that exhibit dimming episodes, inconsistent with binarity), in 2Bn light curves of ZTF by running a custom filter:

*Stetzler et al. (2022)*



# Analysis at (whole-dataset) scale

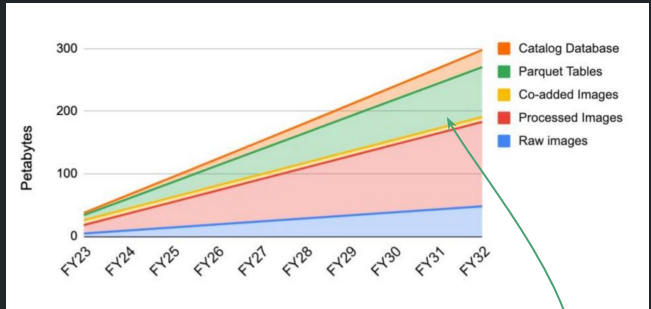


FIGURE 1: Volume of different datatypes over time from DMTN-135 data.

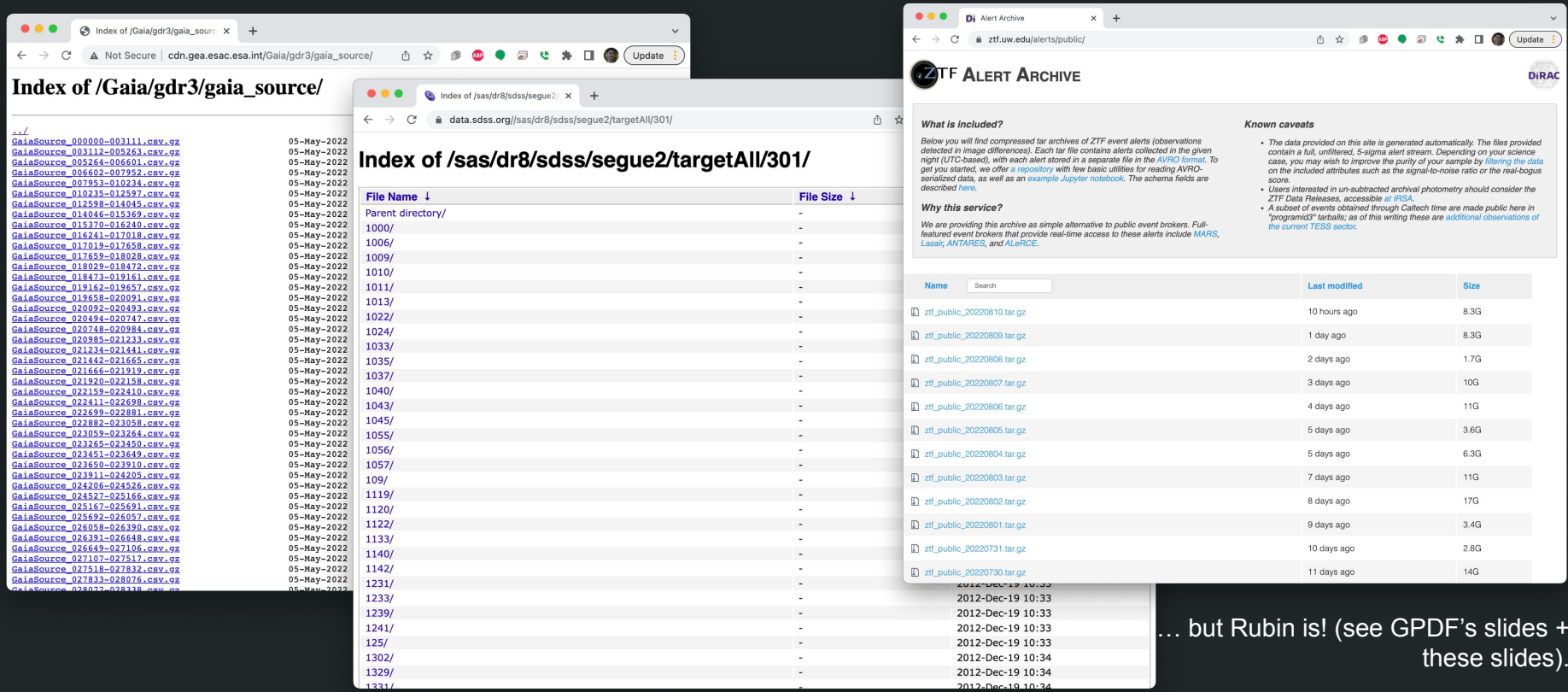
- Relational databases are not ideal for this type of work; it's better to operate on files.

- Rubin plans to provide data in the Parquet (https://parquet.apache.org/) file format both to support bulk download and large-scale analyses.
  - Why Parquet: it the de-facto standard format for tabular big-data analyses. Supported by all major analytics frameworks from Spark, Dask, to Pandas and astropy.
  - It is well suited to keeping data in the cloud (i.e., in "object stores").

```
import pandas as pd
pd.read_parquet('example_pa.parquet', engine='pyarrow')
```

- The dataset will be partitioned – delivered in multiple files, enabling tools to operate in parallel.
  - Think 1,000 cores, each analyzing ~40 files worth of light curves.
  - Or a massively parallel download, for bulk distribution.

# How to partition? Historically, we haven't generally given this much thought...




… but Rubin is! (see GPDF's slides + these slides).

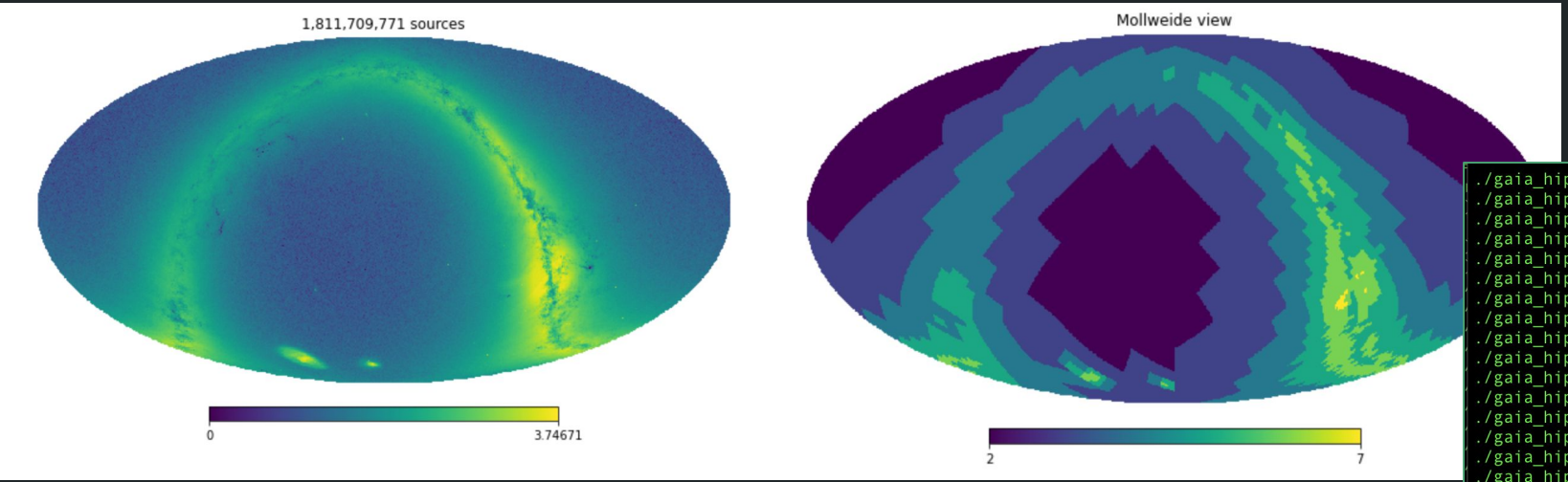
# User desiderata

1. Keep all data related to the same object together (i.e., the whole time series).

2. Partition spatially. Each file should have objects that are close together, enabling spatial queries.

3. Ensure similarly-sized partitions (files). When processing in parallel, each file should take roughly same time to process.

# A Solution: Hierarchical Healpix Partitioning (“HiPSCat”)




*Gaia DR2 Catalog Counts (log scale)*




*Visualization of file storage (color = healpix level)*
*3933 partitions of similar size (128-256 MB)*

*Layout on “disk”:* 👇

```
./gaia_hipscat/Norder5
./gaia_hipscat/Norder5/Npix0
./gaia_hipscat/Norder5/Npix1
./gaia_hipscat/Norder5/Npix10
./gaia_hipscat/Norder5/Npix100
./gaia_hipscat/Norder5/Npix1000
./gaia_hipscat/Norder6/Npix10000
./gaia_hipscat/Norder6/Npix10001
./gaia_hipscat/Norder6/Npix10002
./gaia_hipscat/Norder6/Npix10003
./gaia_hipscat/Norder5/Npix10004
./gaia_hipscat/Norder5/Npix10005
./gaia_hipscat/Norder5/Npix10006
./gaia_hipscat/Norder5/Npix10007
./gaia_hipscat/Norder5/Npix10008
./gaia_hipscat/Norder5/Npix10009
./gaia_hipscat/Norder5/Npix1001
./gaia_hipscat/Norder5/Npix10010
./gaia_hipscat/Norder5/Npix10011
./gaia_hipscat/Norder7/Npix10012
./gaia_hipscat/Norder7/Npix10013
./gaia_hipscat/Norder7/Npix10014
./gaia_hipscat/Norder7/Npix10015
./gaia_hipscat/Norder5/Npix10016
./gaia_hipscat/Norder5/Npix10017
./gaia_hipscat/Norder5/Npix10018
./gaia_hipscat/Norder5/Npix10019
```

Extension of the widely used IVOA HiPS standard.

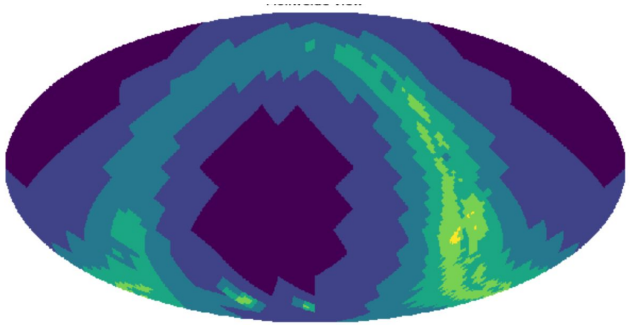
# Some benefits

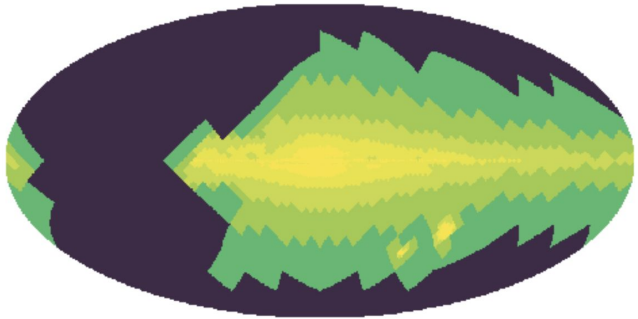
1. Can keep all data related to the same object together. ✔
   - Open implementation question: how? Series of rows? Or arrays?
2. Partitioned spatially. Very easy to locate a file an object w. (ra, dec) is stored in. ✔
3. Ensures similarly-sized partitions (files). Partitions are hierarchically split to the next healpix order if they exceed a pre-set size limit. ✔

4. Minimal extension of an already well-understood standard (HiPS)
5. Can re-uses (significant existing) HiPS infrastructure, including mirroring and caching.
6. Can be utilized by (HiPS) image viewers for overplotting of catalog data (including full light curves)

7. Enables high-level functionality via HTTP-only interface (no special servers)
8. and...

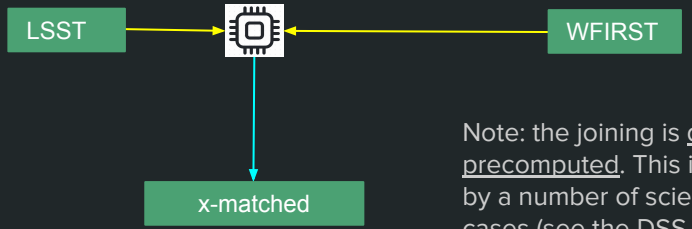
# : Enables Catalog-to-Catalog Joining and Crossmatching





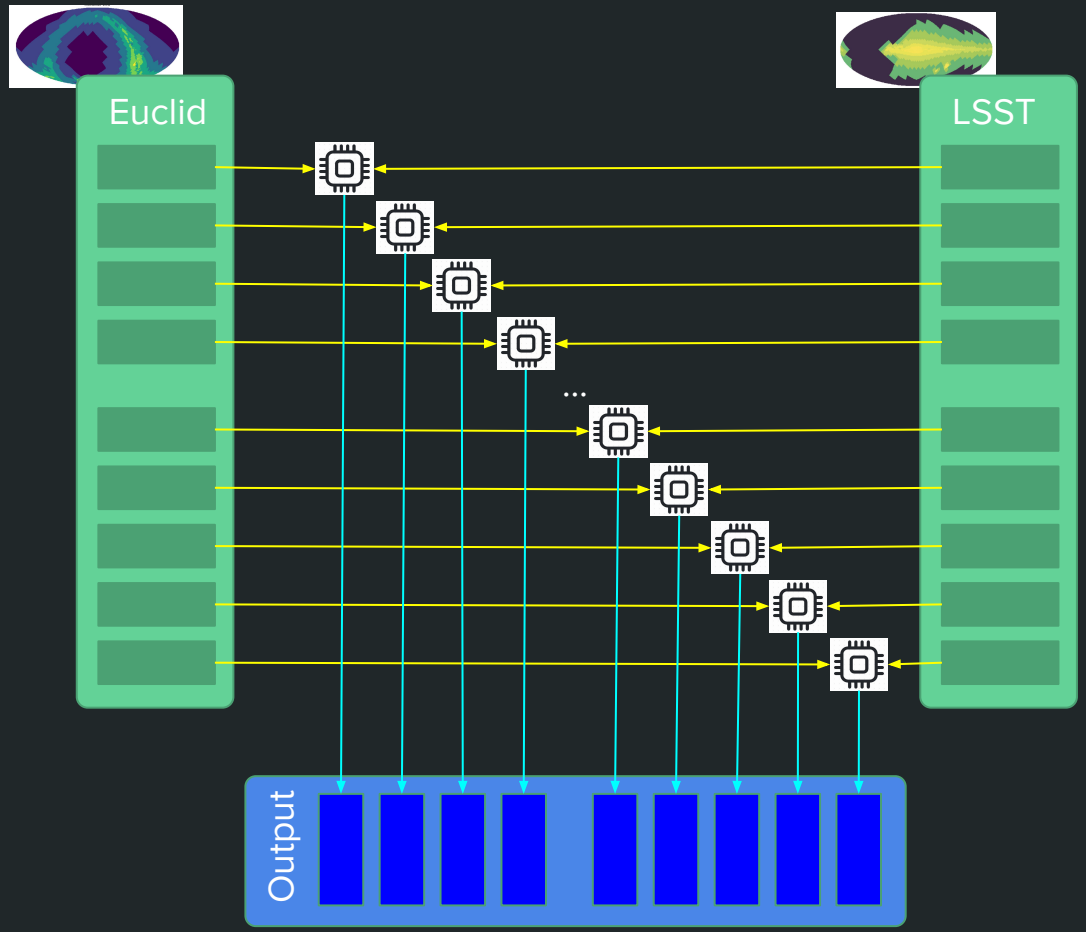
If two or more surveys have catalogs published following this format (ideally on the cloud), **highly parallel, on-the-fly, joining and cross-matching becomes possible**.

Right: the Python that you'd write to cross-match and jointly analyze two large-scale catalogs.




Note: the joining is dynamic, not precomputed. This is required by a number of science use cases (see the DSS whitepaper for details)

# Towards scalable multi-catalog analysis




```
euclid = spark.table("s3://ipac/euclid/objects.hcat")
lsst = spark.table("s3://noirlab/lsst/objects.hcat")

euclid
        .selectPencilBeam(
                ra=60*u.deg, dec=30*u.deg, radius=20*u.deg
        )
        .crossmatch(lsst, radius=1.0)
        .filter(selectDropouts)
        .saveAsTable("s3://mydb/candidates.hcat")
```

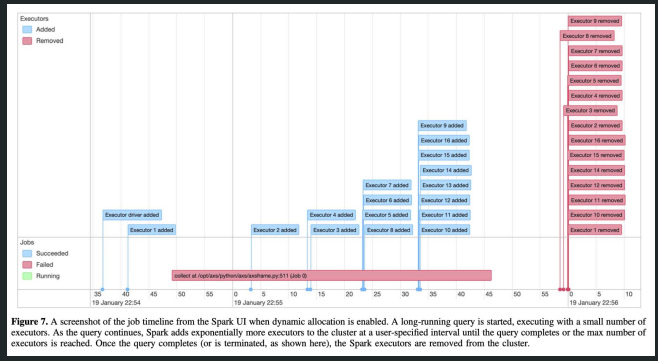
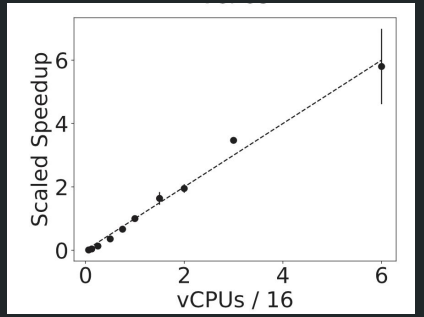


**Figure 7.** A screenshot of the job timeline from the Spark UI when dynamic allocation is enabled. A long-running query is started, executing with a small number of executors. As the query continues, Spark adds exponentially more executors to the cluster at a user-specified interval until the query completes or the max number of executors is reached. Once the query completes (or is terminated, as shown here), the Spark executors are removed from the cluster.

Above: exponential scale-up of nodes until all data are processed.

Right: weak scaling (nearly perfect).

(Stetzler et al. 2022, using Spark)



# Progress

- Learning what others are doing, collecting feedback, possible alternatives, and fellow travelers. This is very much an idea at its initial stages.
  - Been testing these concepts for the past ~3 years (more in Zecevic+ 2019, Stetzler+ 2022)

- To explore it further, looking to:
  - Develop a Python implementation prototype/testbed
    - Creation/query
    - Cross-matching
  - Developing a Dask-based large-scale processing example...
  - ... to be followed by Spark, if successful.

- Goal: A prototyped concept ready for IVOA discussions / Rubin evaluation.

# Discussion questions

1. **How are other groups thinking about the problem of large-scale data/time series storage? What other efforts are out there?**
2. Tool support and (more) use-cases. What to focus on first?

3. Thoughts about this general approach. Is HiPS the right standard to start from? Or an industry-standard format (e.g., Delta Lake, Iceberg, ...)?
4. Cross-matching support – how to efficiently handle objects at partition boundaries.
5. Updates, appends, transactions, etc...

***Interested? Join #lincc-dataformats on the LSSTC Slack.***

# 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

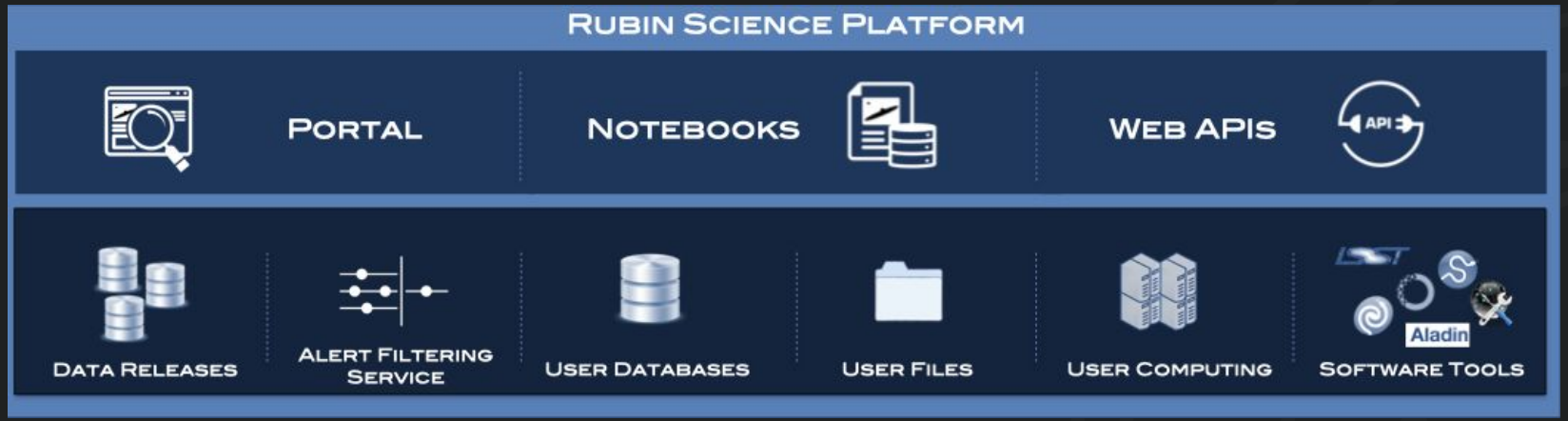
- LSE-319: Science Platform Vision Document
- LSE-61: Data Management System Requirements
- LDM-554: Data Management LSST Science Platform Requirements
- LDM-542: Science Platform Design
- DMTN-202: Use cases and science requirements on a user batch facility
- DMTN-086: Next-to-the-Database Processing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

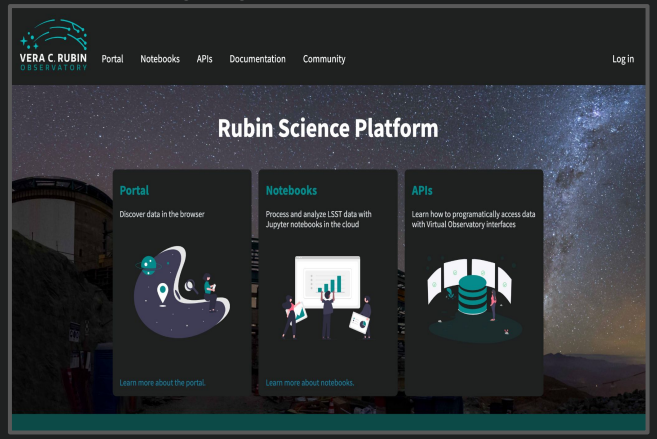
**API Aspect**
Remote access to the LSST archive via Virtual Observatory interfaces



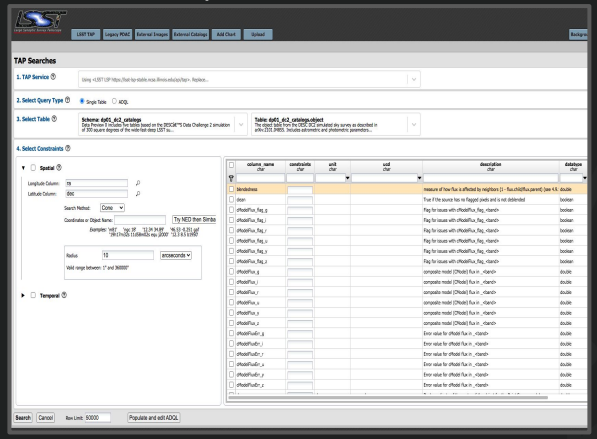
# The DP0-Era Rubin Science Platform

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.

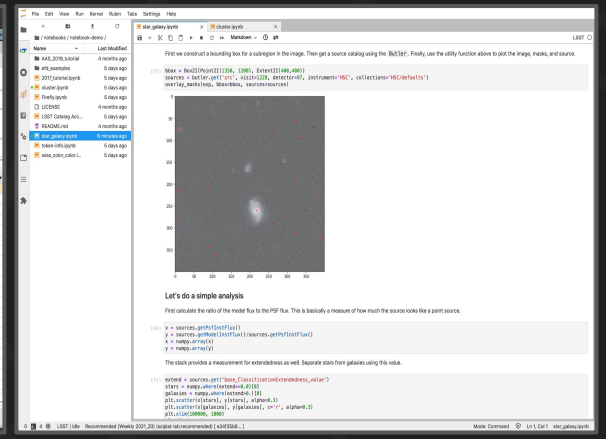
RSP Landing Page




RSP Portal Aspect




RSP Notebook Aspect






*Slide credit: Melissa Graham*

# 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

## Community forum https://community.lsst.org/