@merrdiff

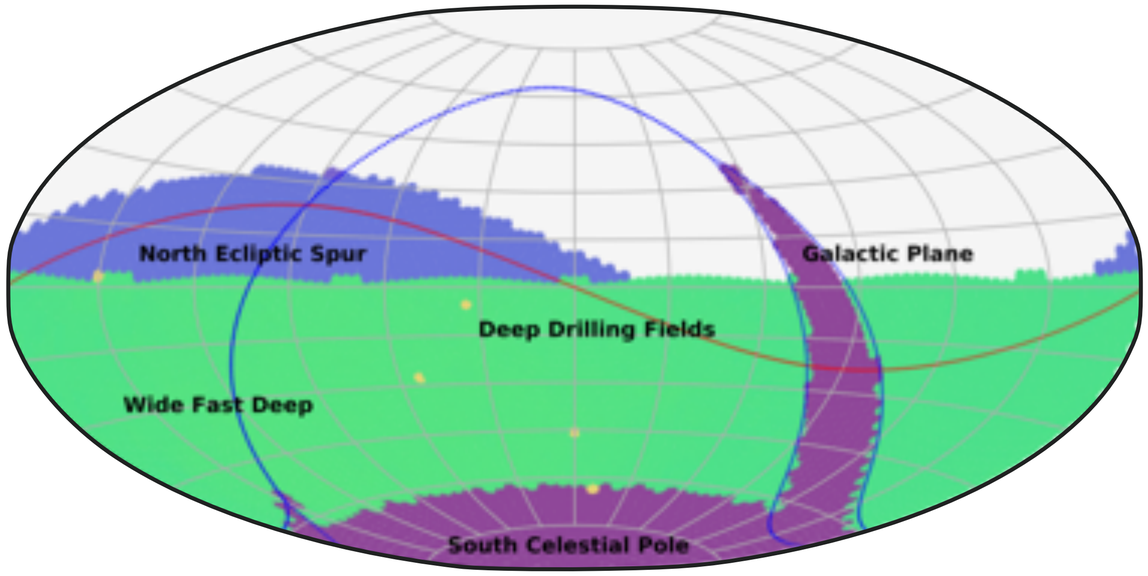
@mrawls

Preparing for Big Data from Vera C. Rubin Observatory

Meredith Rawls | SOMACHINE 2021 | April 21, 2021

The Legacy Survey of Space and Time

- The 10-year survey from Rubin Observatory
- High-resolution movie of the night sky
- Uniform map of southern sky every three days
- Data and software for the community
- Science from solar system to dark energy



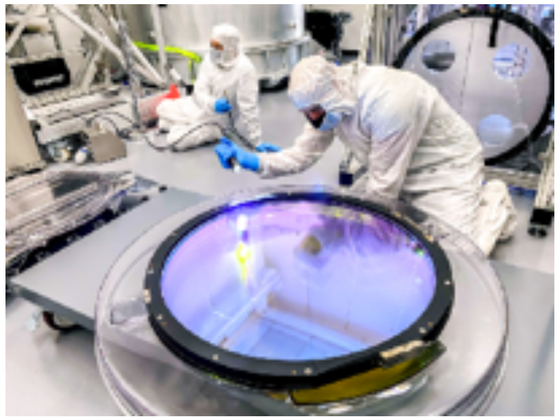
Number of visits (*g*, *r*, and *i* bands)

People make Rubin Observatory possible

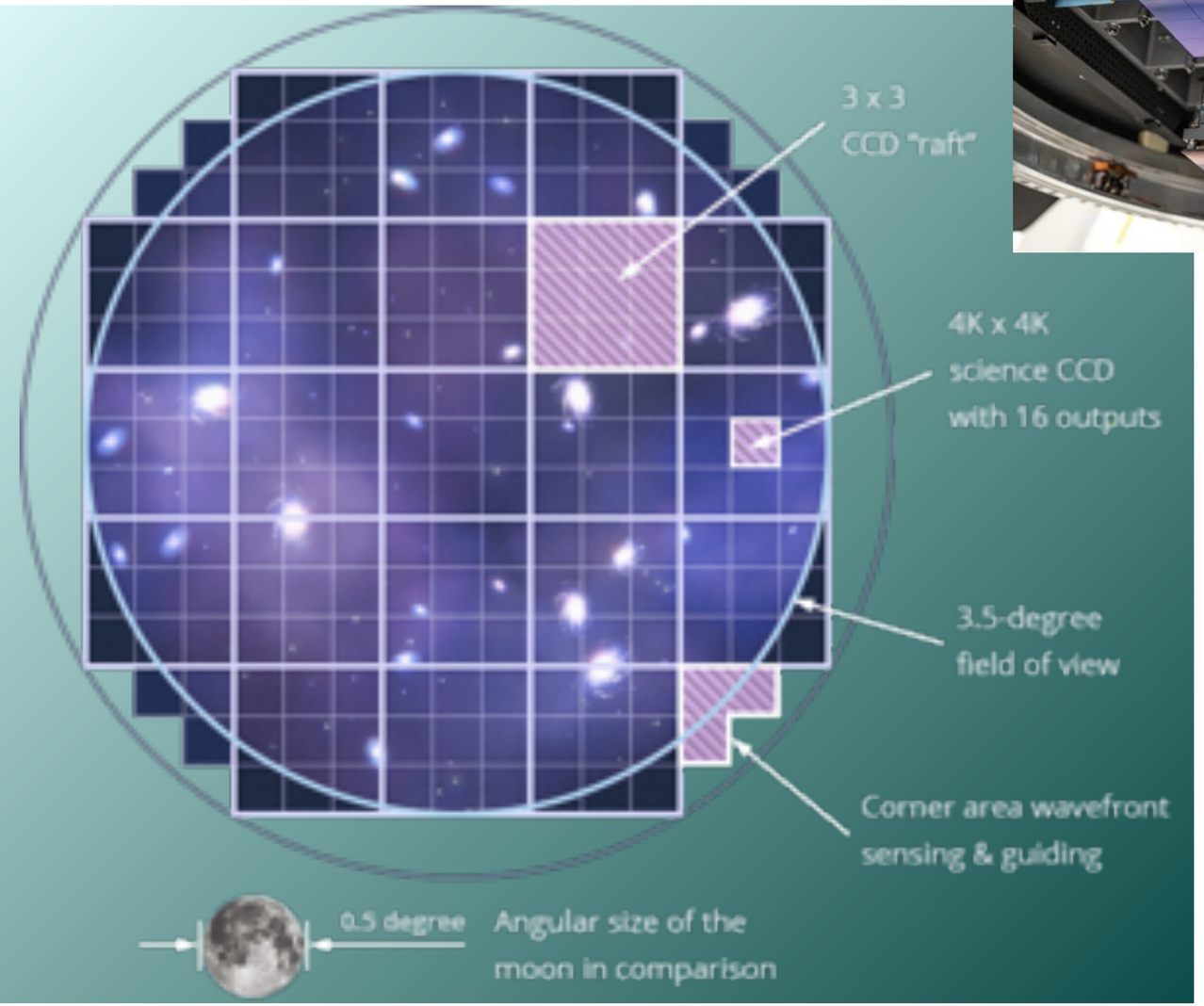
This presentation, much like the project,
is a collaborative (and international) effort!

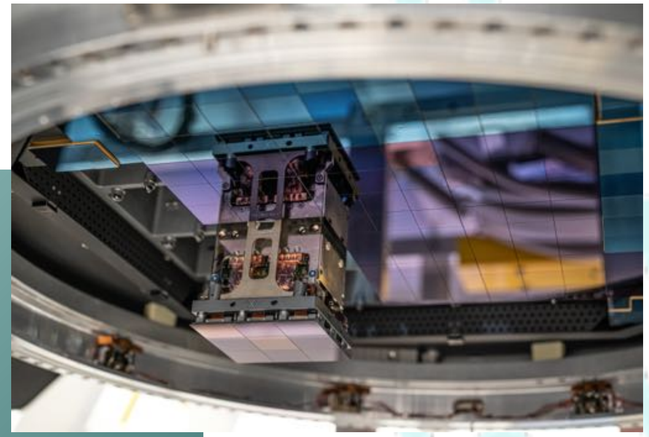
Legacy Survey of Space and Time

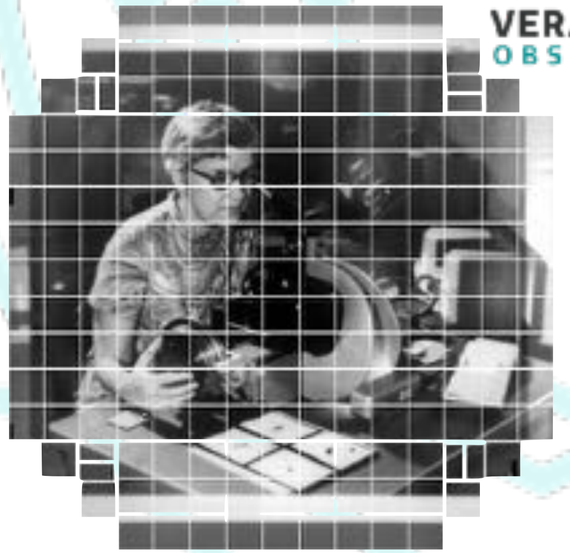
Rubin Observatory under construction



3.2-gigapixel camera!

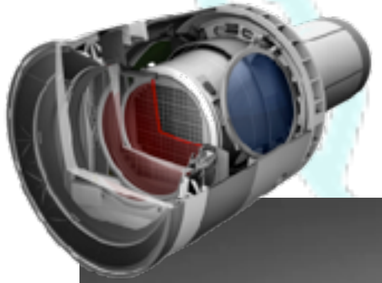


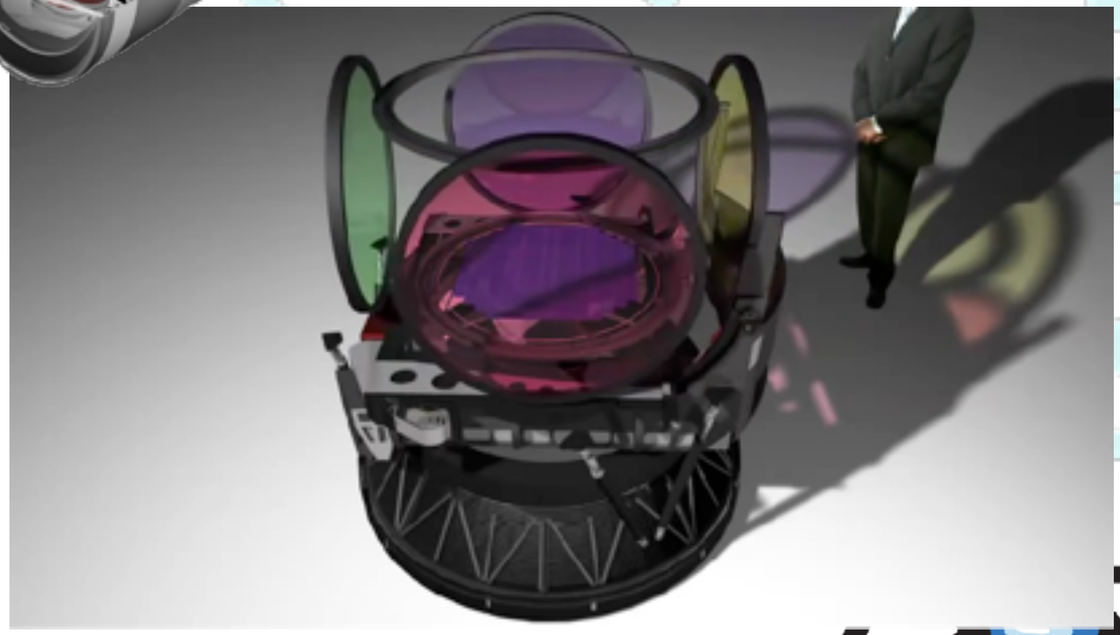




VERA C. RUBIN OBSERVATORY

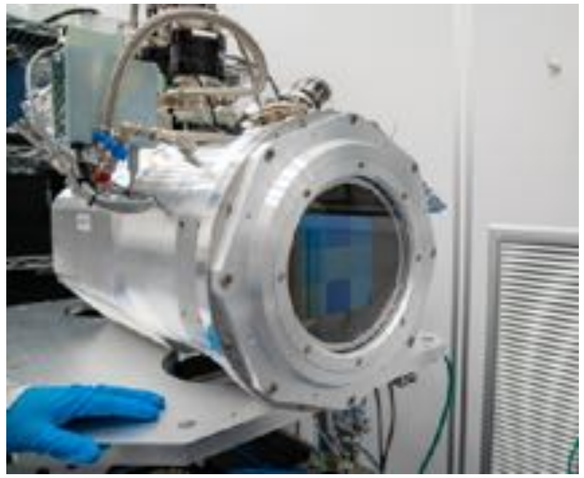
SLAC

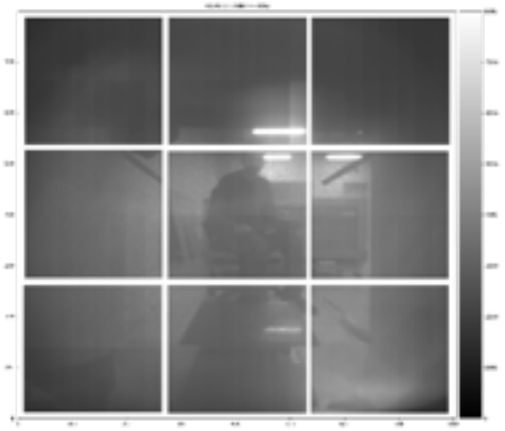


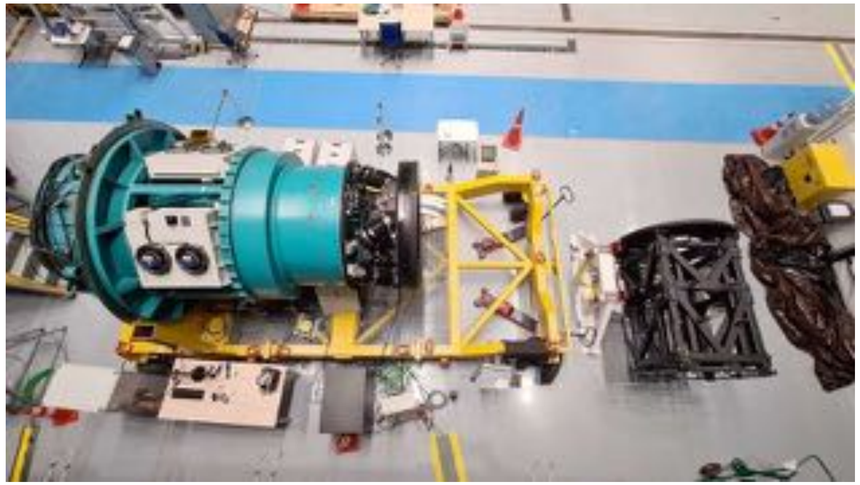


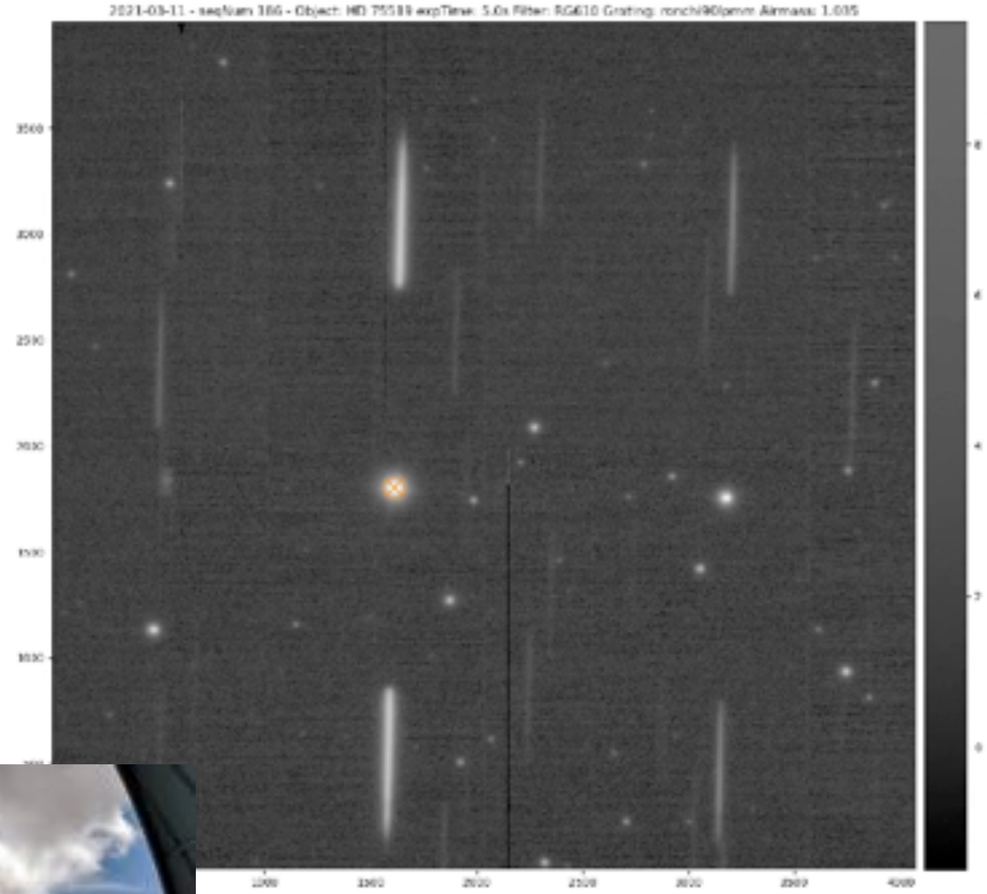
LSST Legacy Survey of Space and Time

Commissioning Camera & Auxiliary Telescope





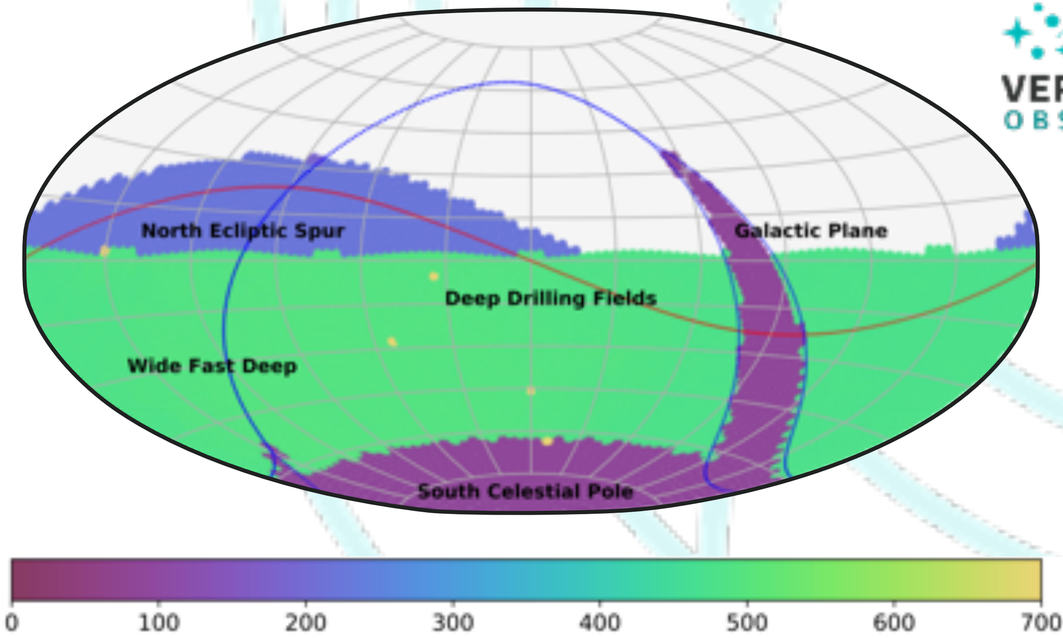




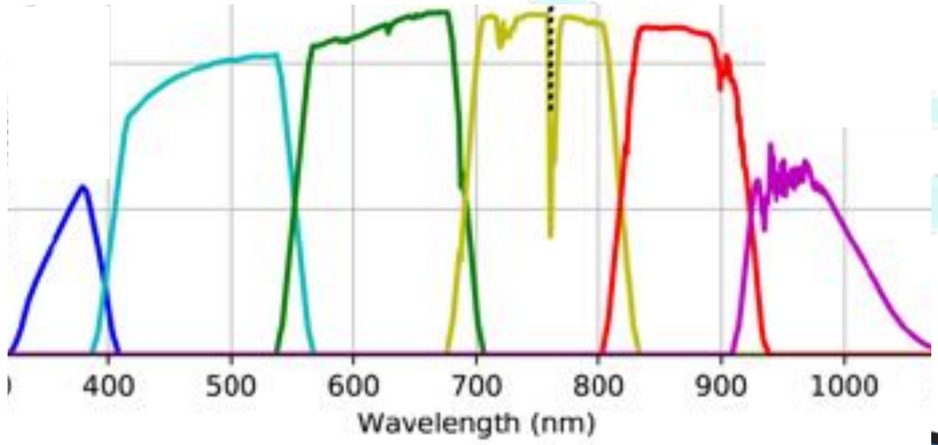
Meredith Rawls | SOMACHINE 2021 | April 21, 2021

LSST Survey Strategy

- Baseline Survey Strategy in development must meet science goals (details: ls.st/srd)
  - Dark energy and cosmology
  - Transient universe
  - Solar System
  - Milky Way and local volume
- Wide-Fast-Deep area of 18,000 deg^2, 825 visits per field over 10 years, and same-night same-field re-visit “pairs”
- Account for hardware realities (slews, filters)
- **Maximizing scientific return is a hard problem!**



Number of visits (*g*, *r*, and *i* bands)

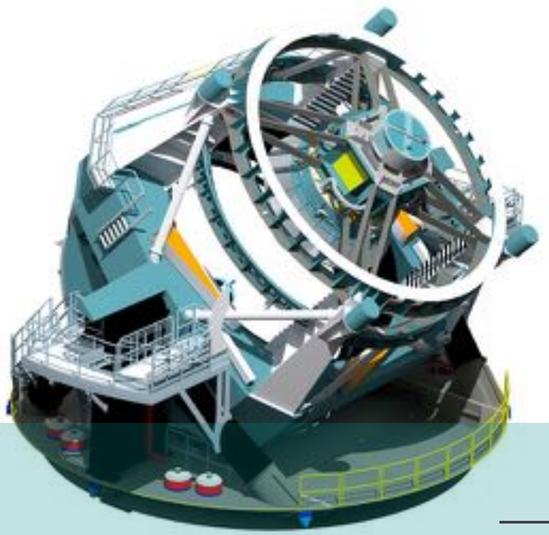


How will we manage all that data?!

VERA C. RUBIN OBSERVATORY

Raw Data: 20TB/night

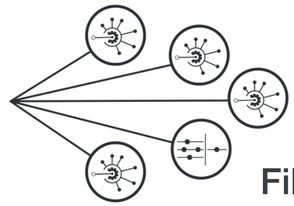
Sequential 30s images covering the entire visible sky every few days



Prompt Data Products

Alerts: up to 10 million per night

60s via nightly alert streams



Community Brokers

LSST Alert Filtering Service

Results of Difference Image Analysis (DIA): transient and variable sources

Solar System Objects: ~ 6 million

24h via Prompt Products Database

Data Release Data Products

Final 10yr Data Release: Michitaro Koike

- Images: 5.5 million x 3.2 Gpx
- Catalog: 15PB, 37 billion objects

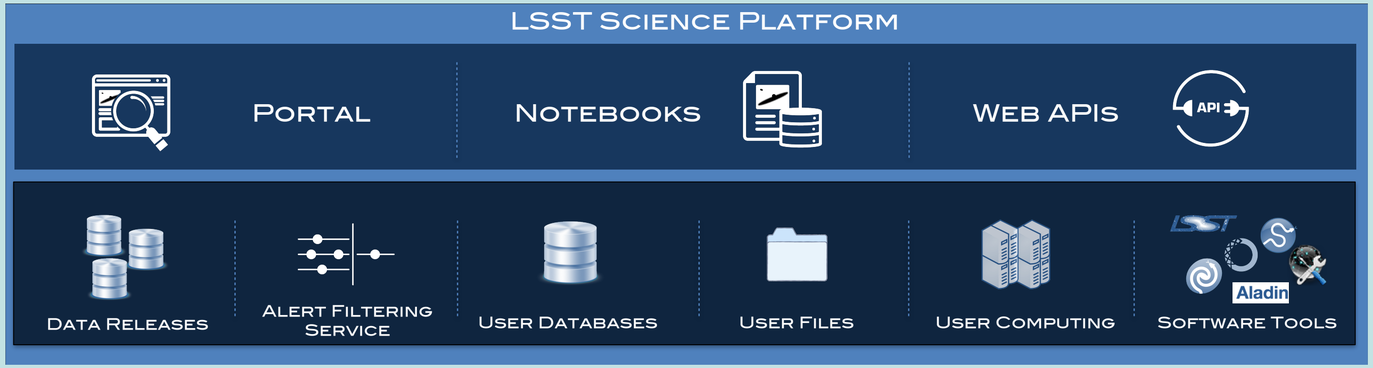
via Data Releases

LSST DACs (Chile & NCSA)

Independent DACs (iDACs)

LSST Science Platform

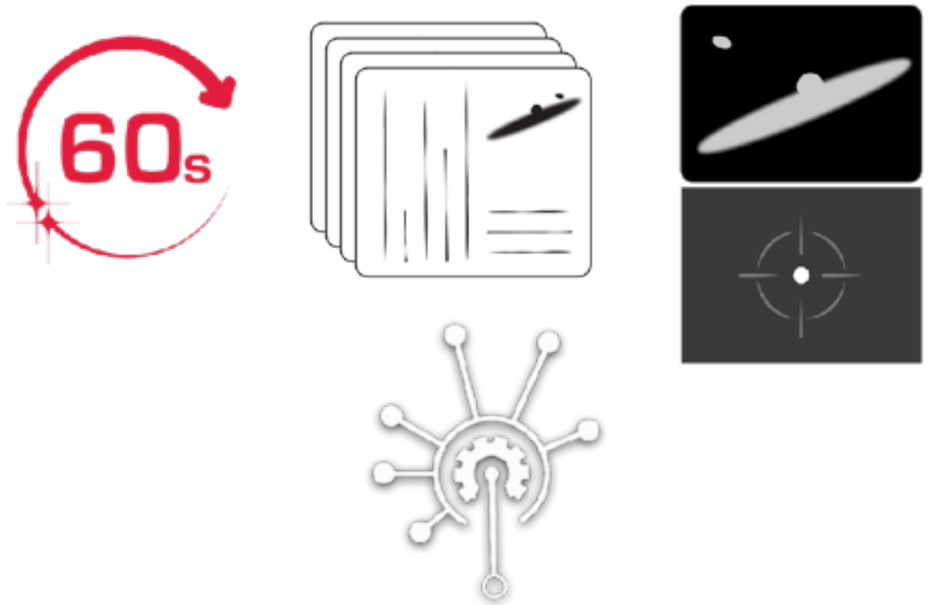
Provides access to LSST Data Products and services for all science users and project staff



Alerts (Real Time) & Prompt Data Products (Nightly)

VERA C. RUBIN OBSERVATORY

DIA = **D**ifference **I**mage **A**nalysis



- **DIA Sources**
  - astrometry, photometry, shape
  - signal-to-noise ratio, spuriousness
- **DIA Objects** (~12 month history)
  - proper motion, parallax, mean flux
  - variability parameters
- **Solar System Object info**
- **Image stamps**
  - both difference and template images
  - flux, variance, mask frames with metadata

Community Alert Brokers

ANTARES

Lasair

Queen's University Belfast

ZTF

ALeRCE
Automatic Learning for the Rapid Classification of Events

10 million alerts and
20 TB data per night

LSST
Legacy Survey of Space and Time

Data Releases Data Products (Annual)

- **Reprocessed DIA Source and DIA Object catalogs**
  - characterization (e.g., variability or moving)
  - forced photometry in all difference images
- **Source and Object catalogs**
  - built from direct images and deep stacks
  - forced photometry in all direct images
  - derived parameters for objects
- **Reprocessed Images**
  - single-visit, template, and difference
  - deep stacks

37 billion cataloged objects
5.5 million 3.2-gigapixel images
15 PB database after 10 years
~100 PB including images

The software is at least as important as the science

- Python modules with some C++ to process, generate, and serve images, catalogs, and alerts
- Active, open source development: github.com/lsst
- Documentation and a tutorial: pipelines.lsst.io
- Process data from Subaru Hyper Suprime-Cam (HSC), Blanco DECam, and more
- Join the “Stack Club” for tutorials, notebooks, and workflows: stackclub.readthedocs.io
- Visit the Community Forum: community.lsst.org

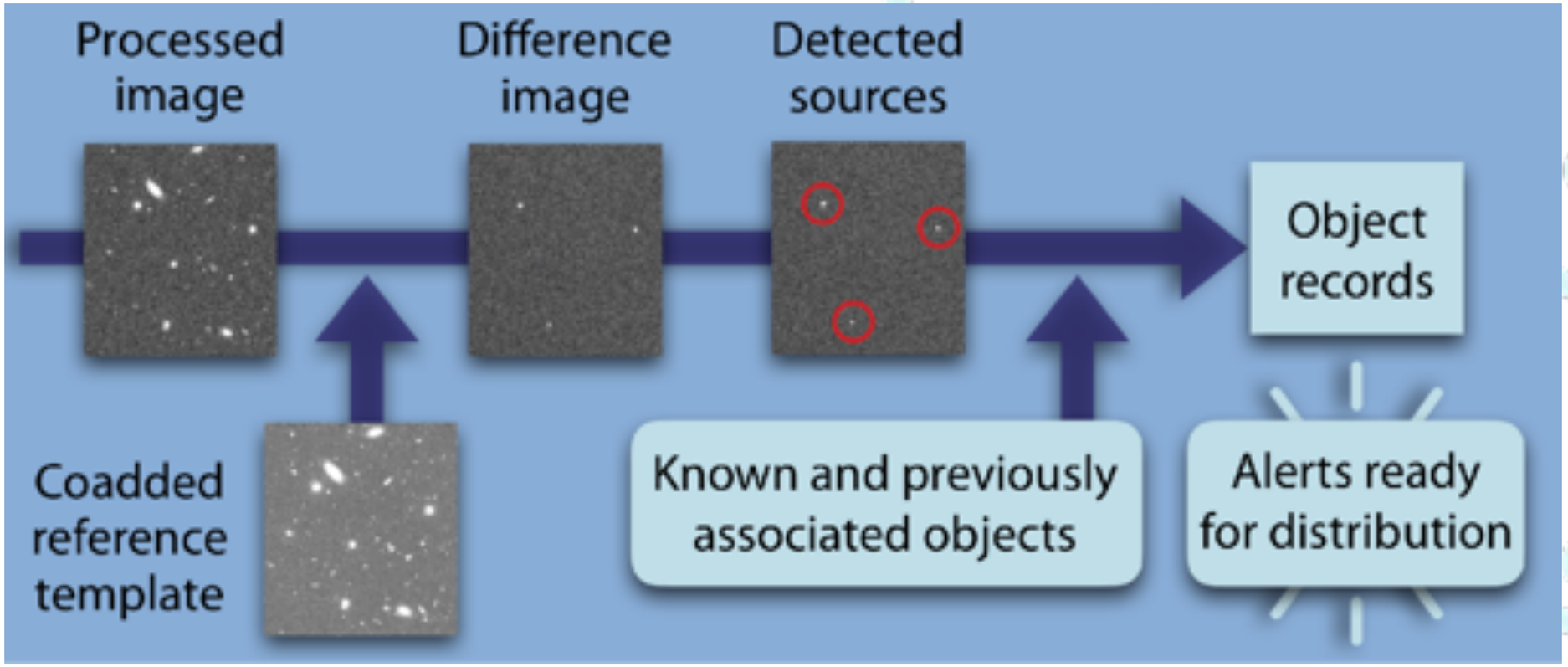
LSST Community

Legacy Survey of Space and Time

Rubin looks awesome, but what do *you* do?

Building a pipeline for prompt data products

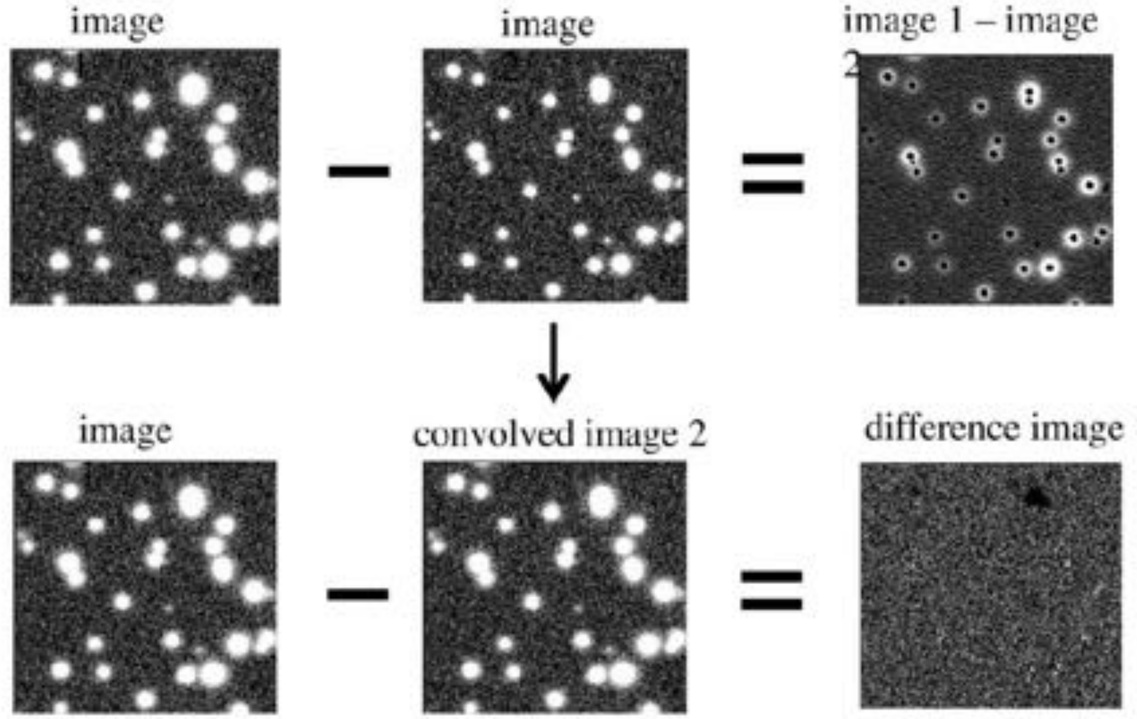
- The data management team at UW focuses on **difference imaging** and **alert production**



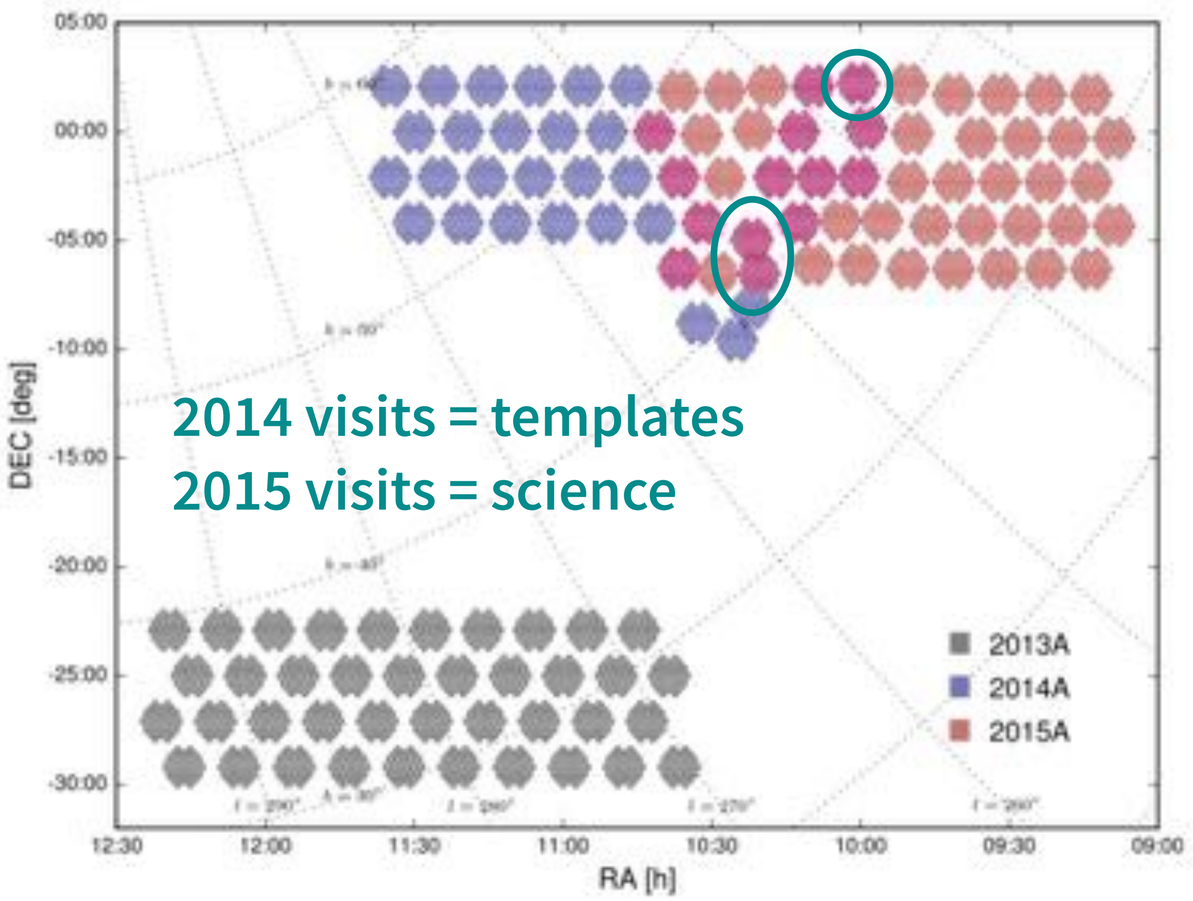
60s

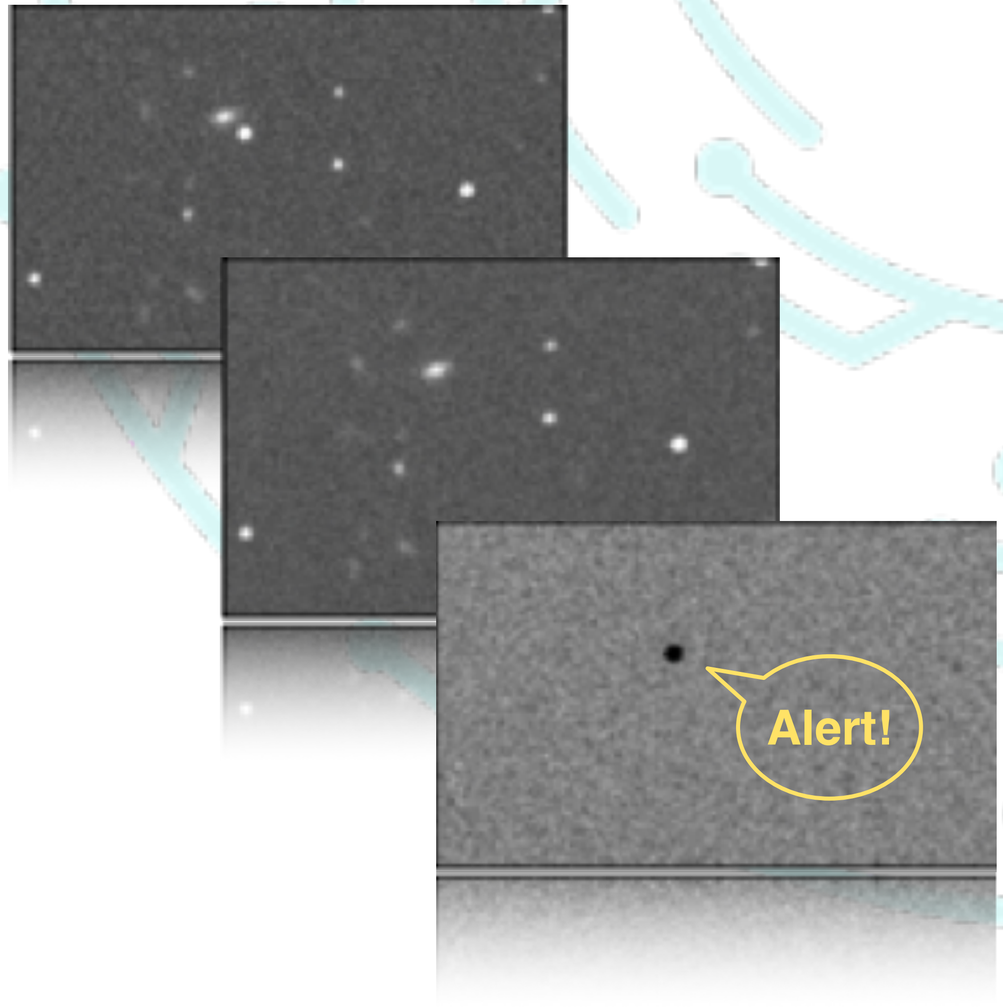
Difference imaging is harder than you think

- Given two images, “template” and “science,” can naively measure fluxes and subtract
- For faint sources, sky noise dominates
- In crowded fields, measuring flux can be tricky
- Images with identical seeing (PSF) would fix this
- One approach: remove non-variable sources by convolving one image to match the other
- **There’s no such thing as a perfect template!**
  - Correlated noise, poor seeing, few coadded exposures, range of air masses, sky is not static, etc.
- More: Alard & Lupton (1998) and Zackay, Ofek, & Gal-Yam (2016)



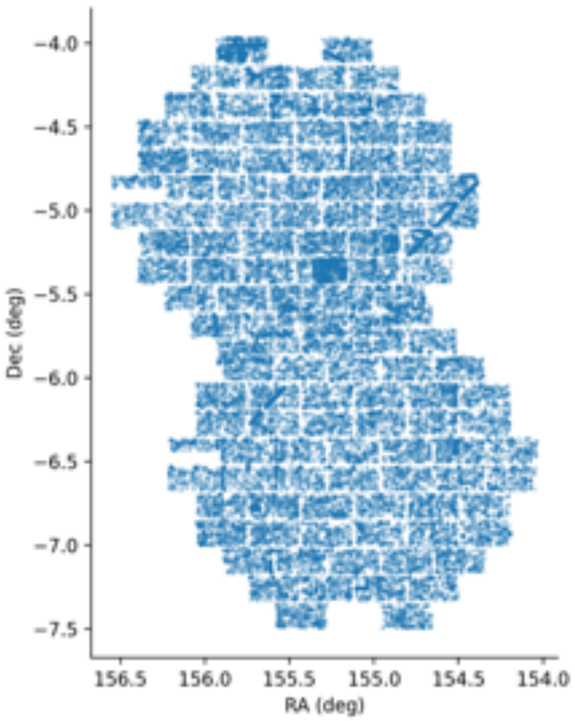
Running Rubin software on precursor data

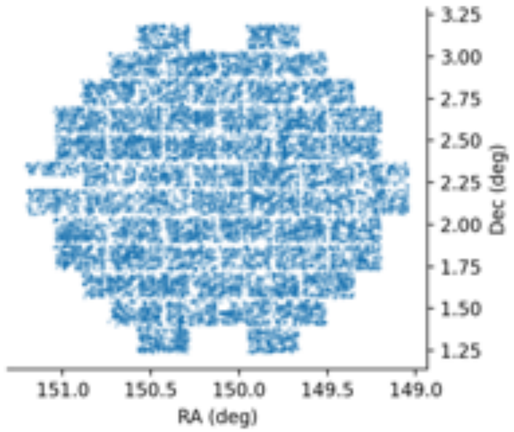




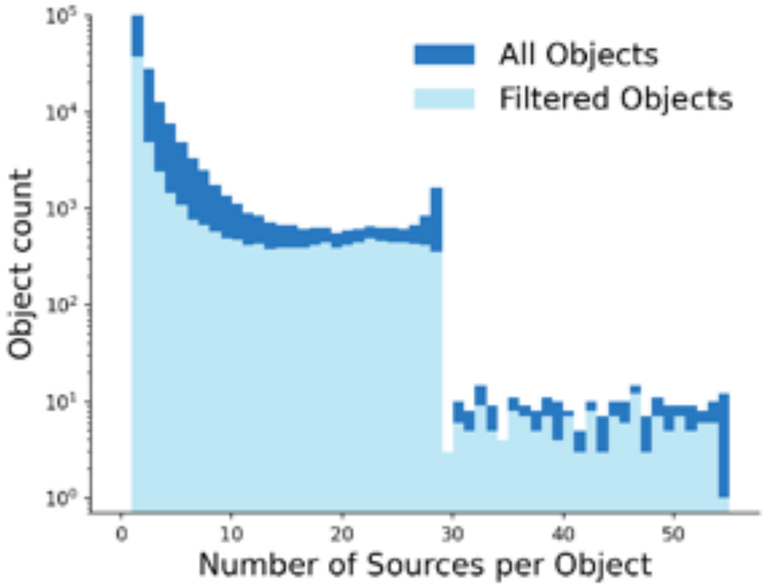
Blanco DECam HiTS Survey (Förster+ 2016)

Associating transient sources into objects





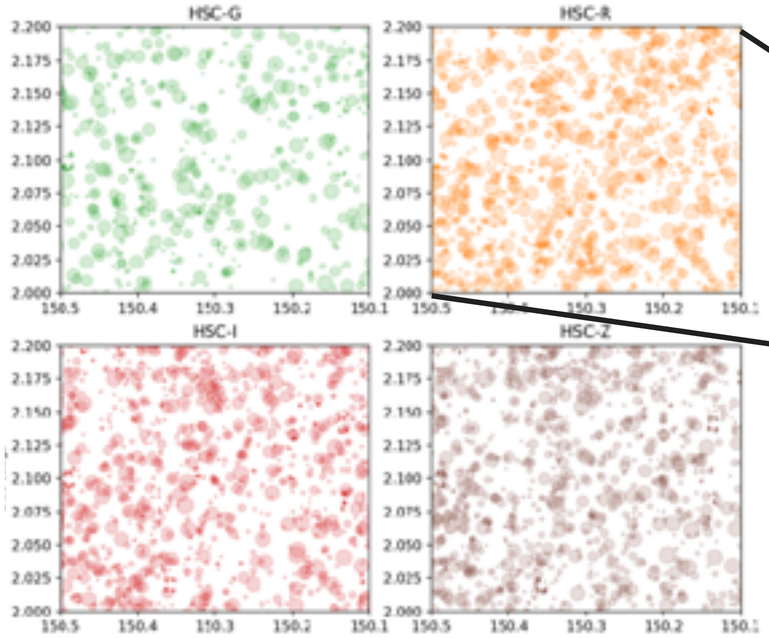
Objects on the sky
in g-band only



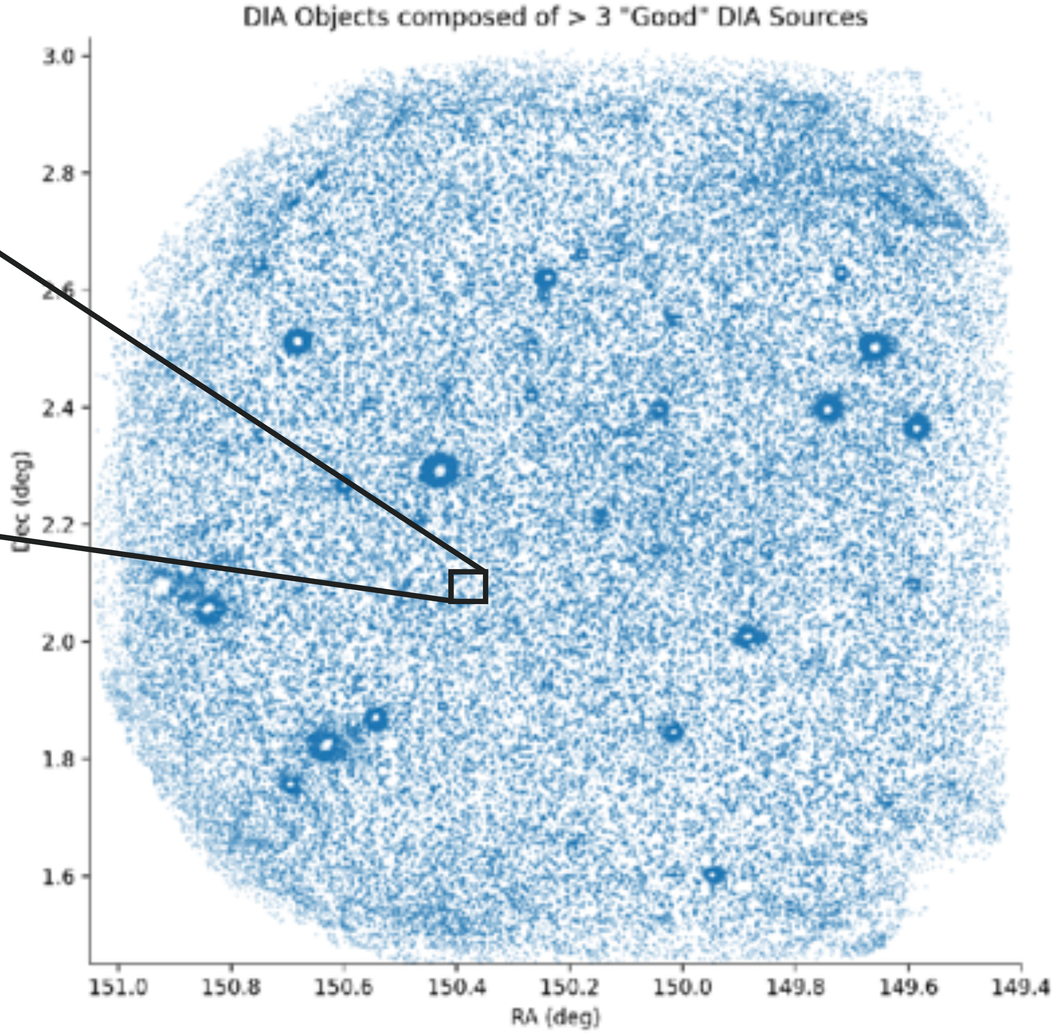
False positives (red)
flagged near CCD edges

Hyper Suprime-Cam sources & objects

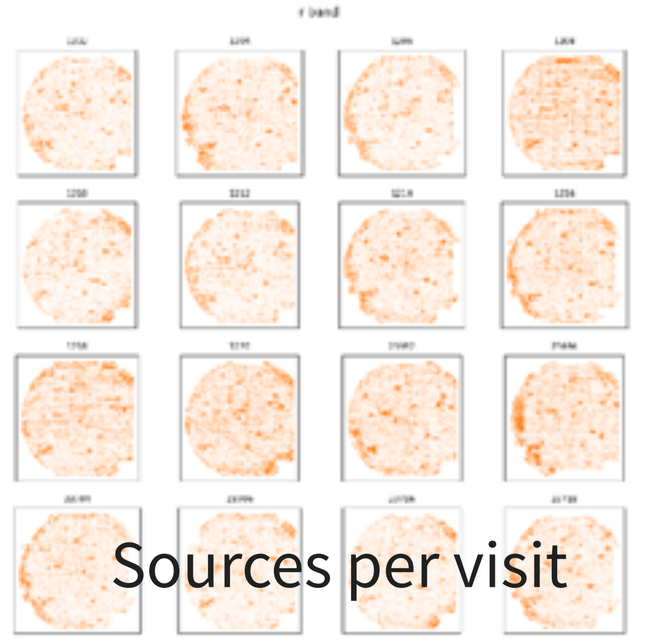
There’s no substitute for trying it out on real data



Objects on the sky
in four bands, zoomed

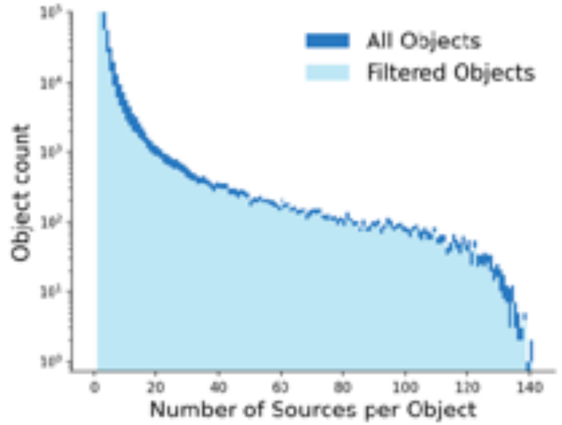


Objects on the sky
in all bands



Sources per visit

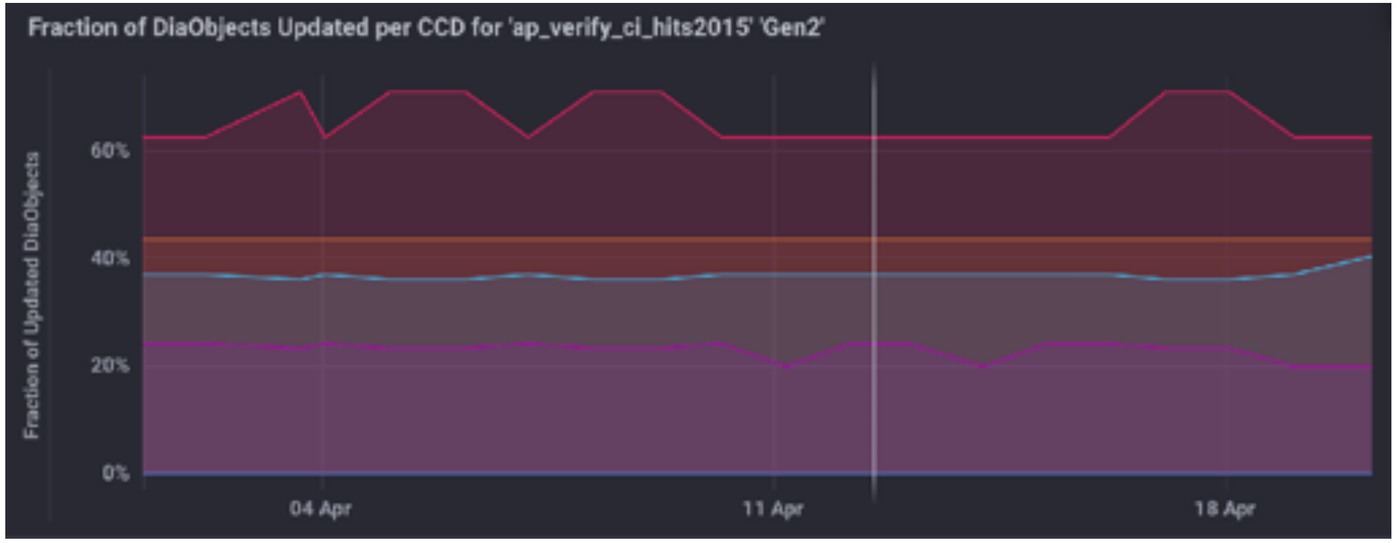
Sources per object



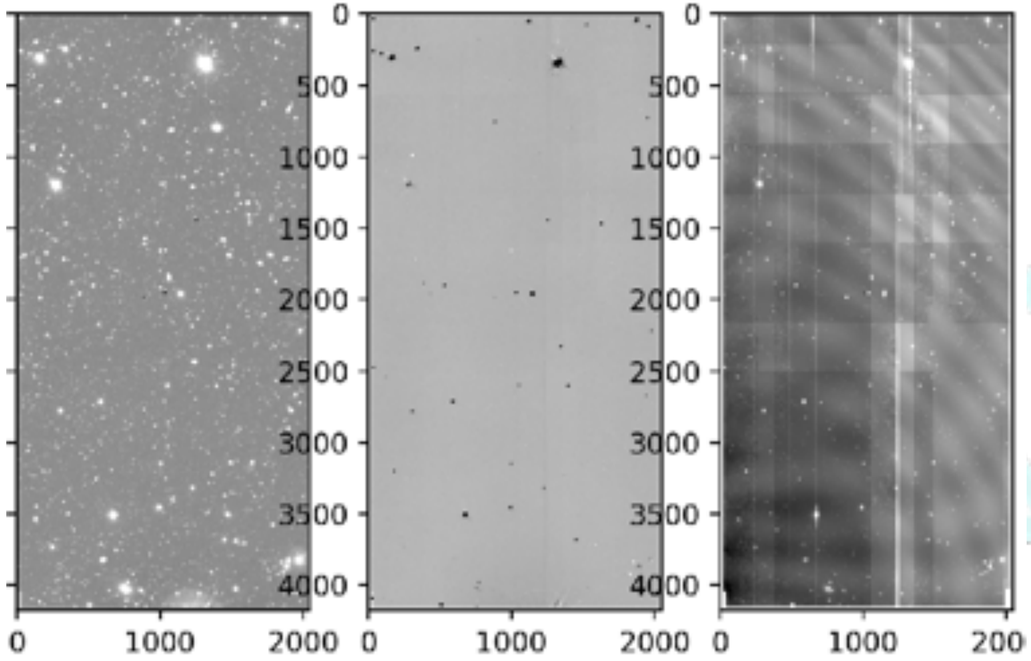
Legacy Survey of Space and Time

Testing difference imaging performance

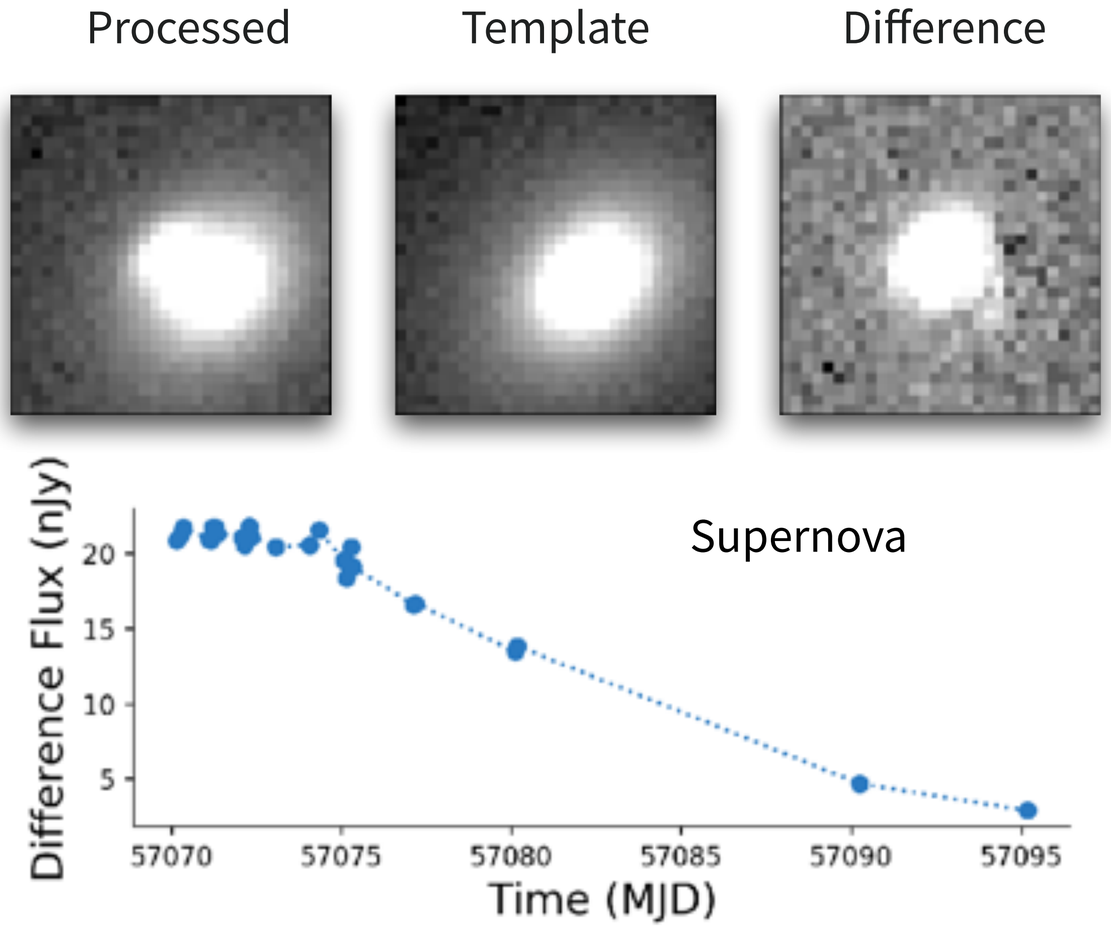
- Which algorithm performs better in which situations (A&L 1998 or ZOGY 2016)?
- Pre-convolve the science image with its PSF?
- Match science to template, or vice versa?
- Run a decorrelation algorithm at the end?
- Scale variance for template, or for difference image?

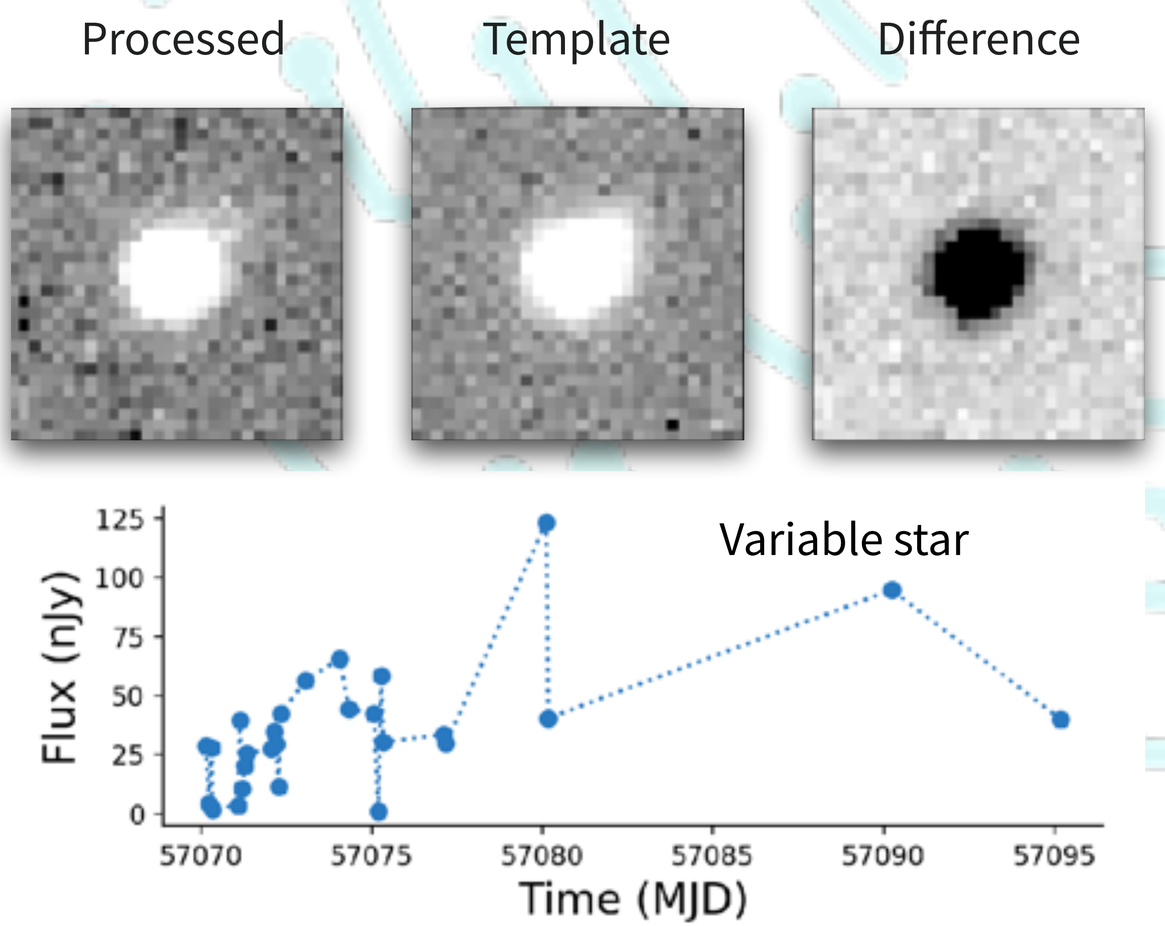


Science Image | Difference Image | Difference Variance

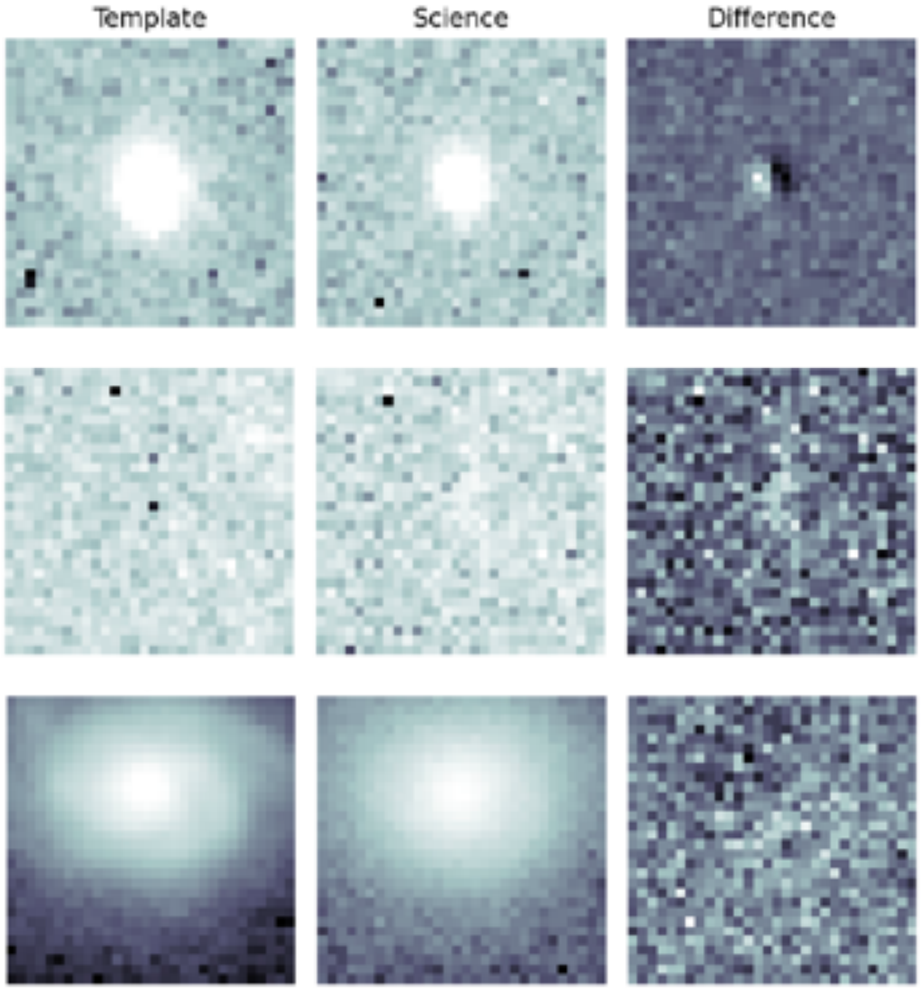


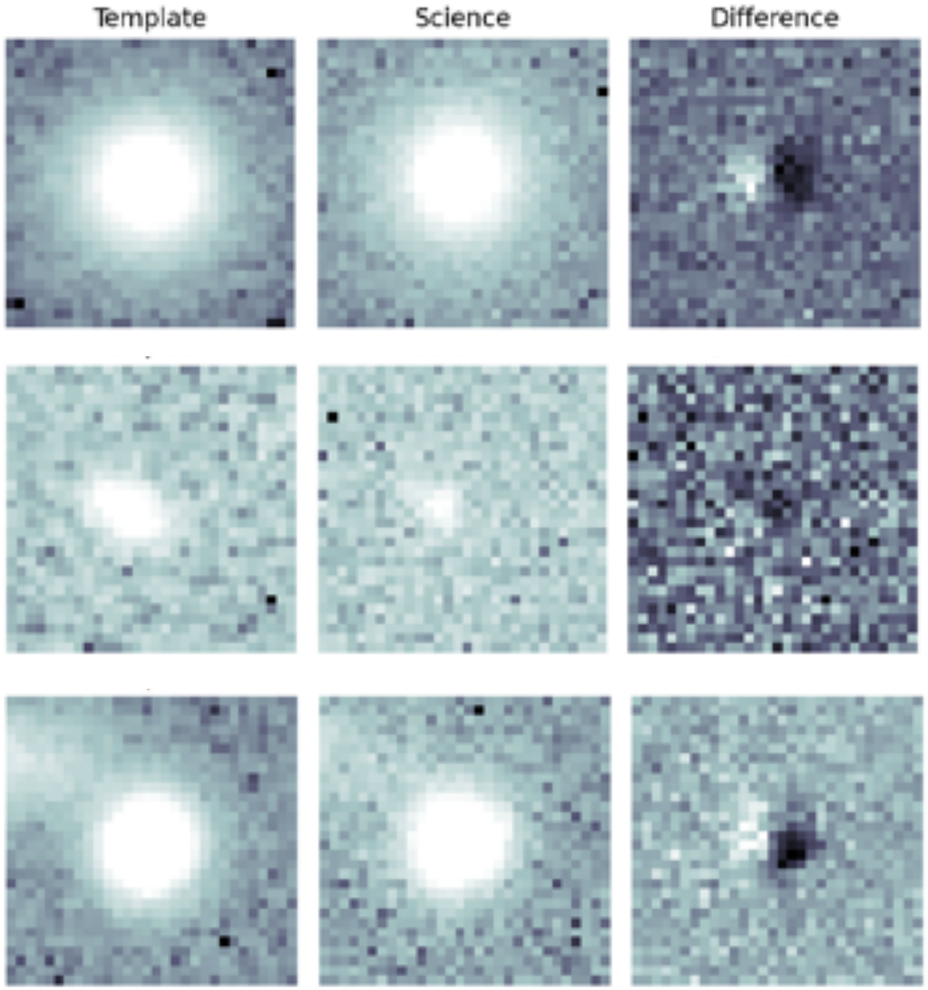
Constructing light curves from precursor data





More cutouts coming to a Zooniverse near you

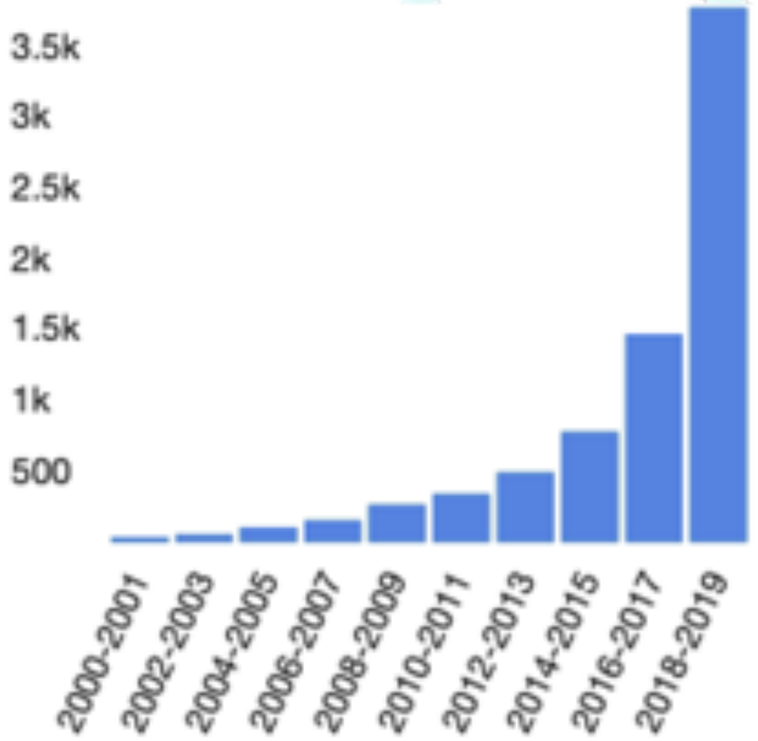




Rubin will enable machine learning studies

- Our Science Pipelines use analytic algorithms, not “black boxes”
- The idea is to **enable science** rather than do machine learning for scientists
- Rubin will provide lots of well-curated and consistently-processed data

- In ten years:
  - 20B galaxies
  - 17B resolved stars
  - 6M orbits of solar system bodies
- Each night:
  - 10M alerts on average



ADS publications per year mentioning machine learning

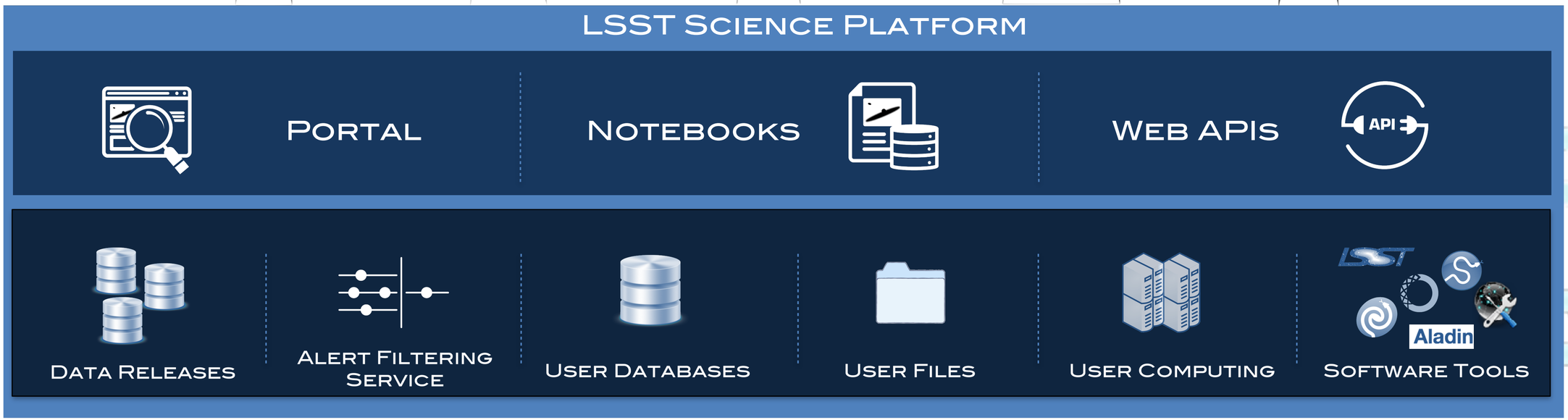
That’s interesting, so how can I get some data?

How to: *Not* Download All The Images

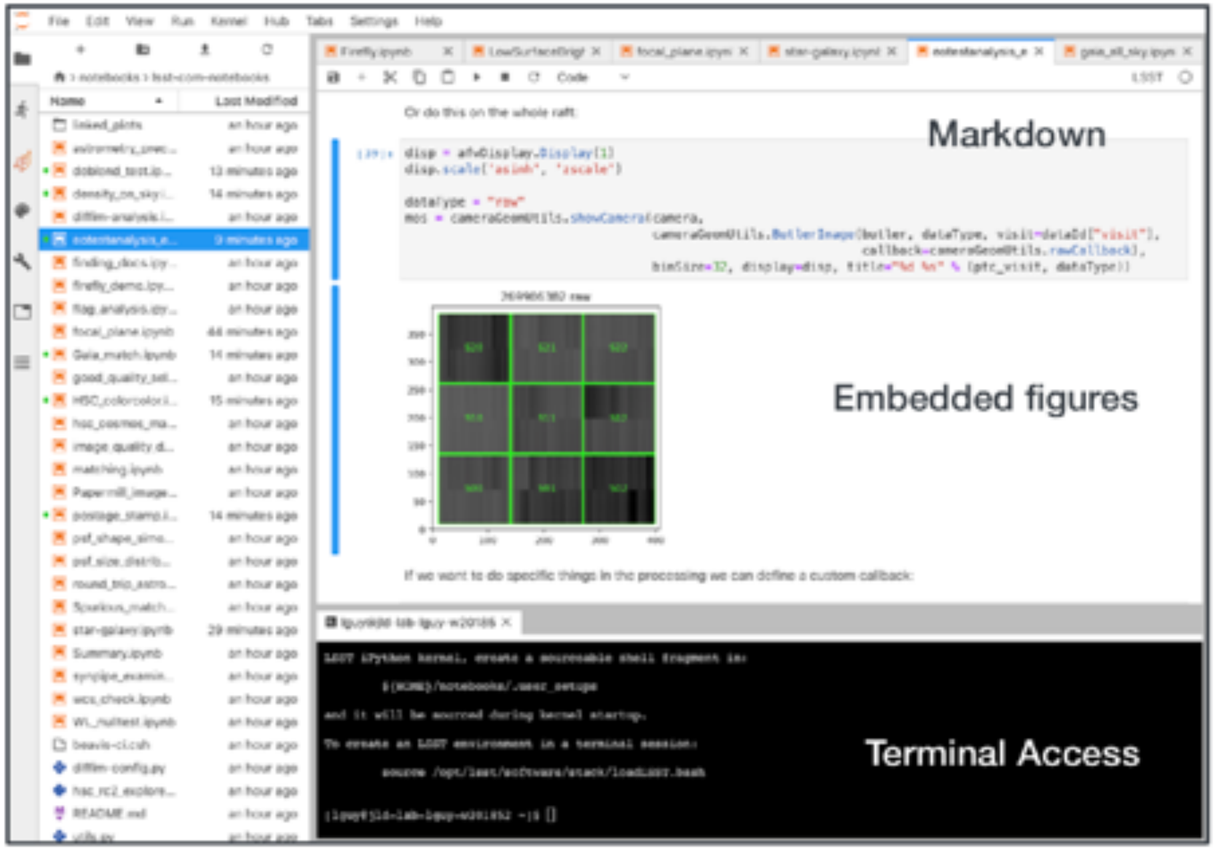
exploratory analysis and visualization of the LSST archive

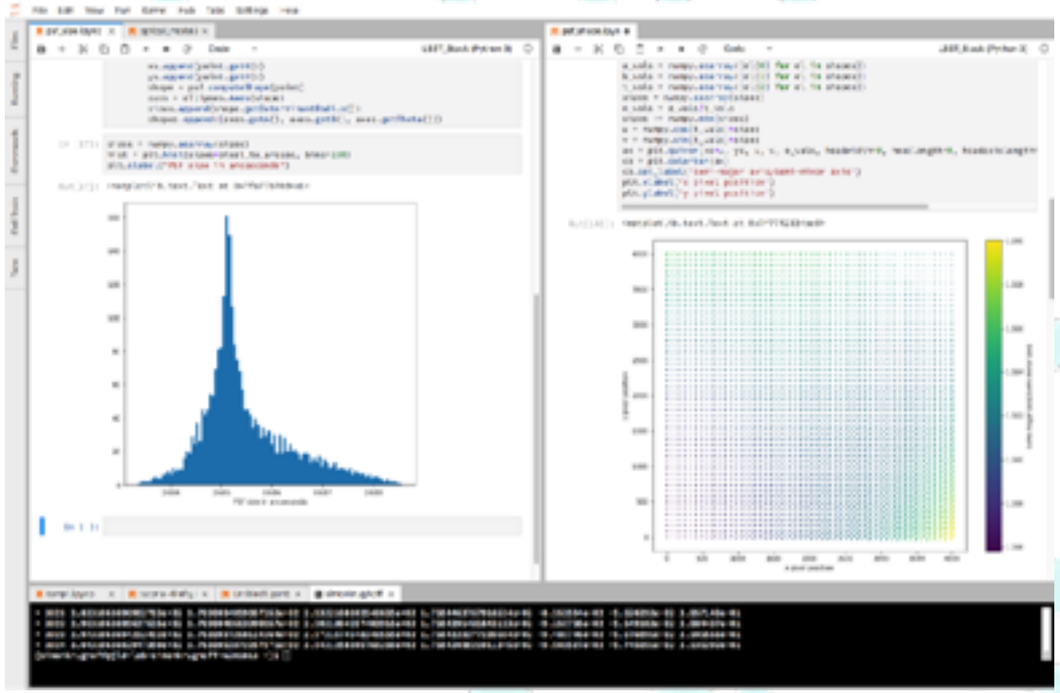
'next-to-data' analysis and create added-value data products

remote access to archive via industry-standard APIs



Rubin Science Platform: bring the code to the data





nb.lsst.io (JupyterLab running in JupyterHub)

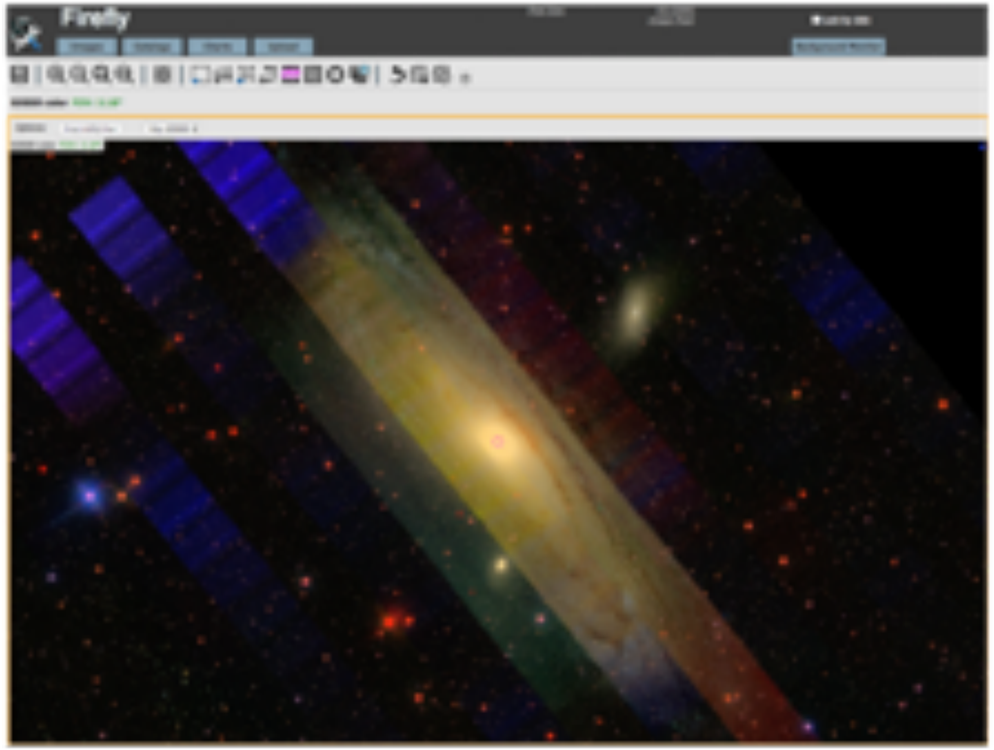
LSST Science Platform Notebook Aspect Documentation

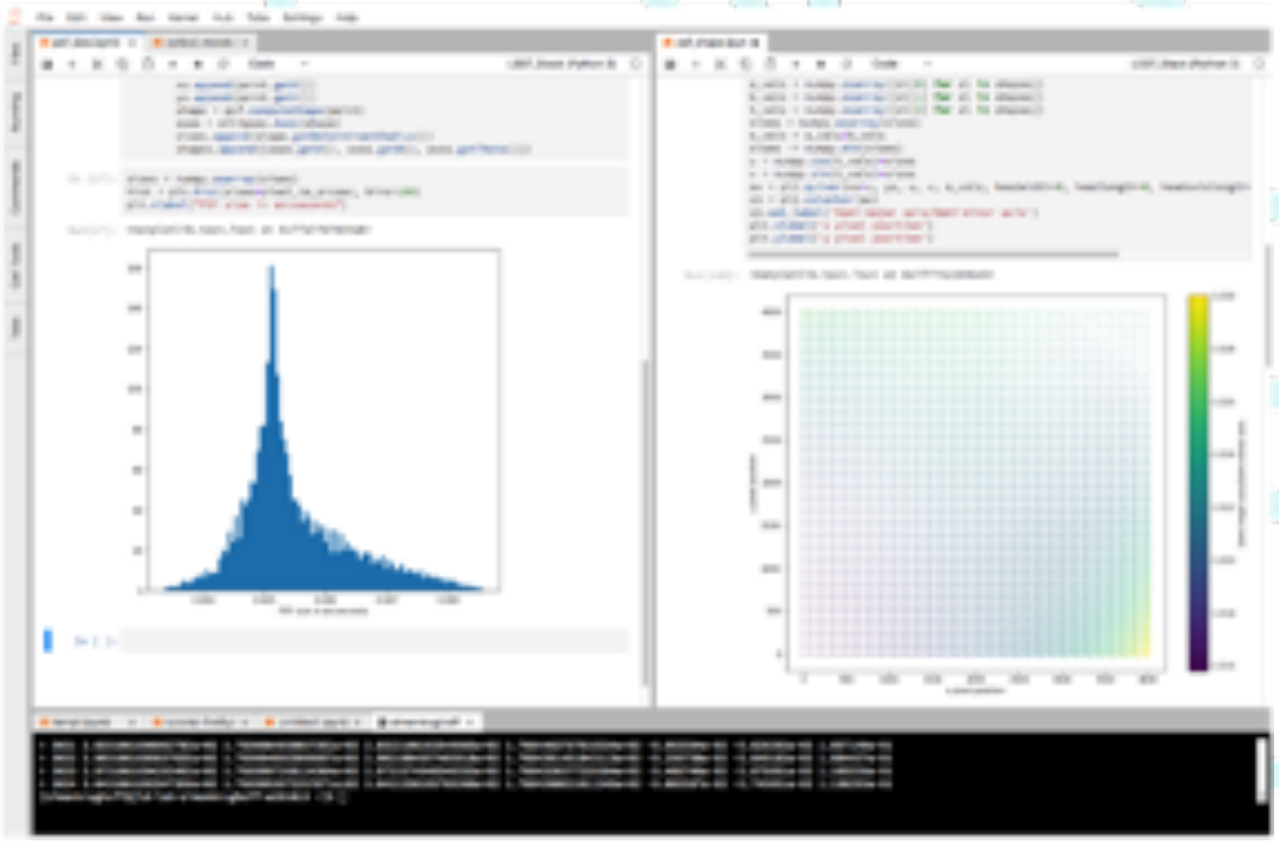
The Notebook Aspect enables you to do your science at the LSST Data Facility, with the LSST Science Pipelines, a full suite of development tools, and your own Python code. The Notebook Aspect is powered by the JupyterLab project.

Legacy Survey of Space and Time

Rubin Science Platform: visualization built in

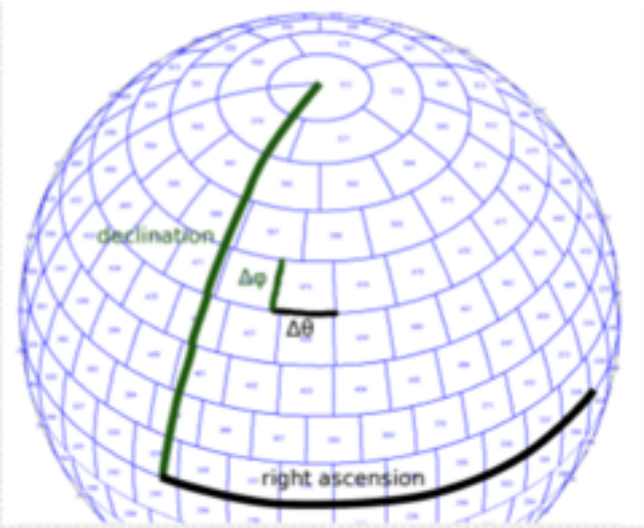
- Firefly web portal enables interactive data exploration beside a notebook





Rubin Science Platform: queries galore

- Developing a database query system, Qserv
- A shared-nothing SQL database to handle huge spatial joins and cross-matching
- Open source and leverages existing tools
- Learn more: [ls.st/LDM-135](ls.st/LDM-135)




```
[16]: df = pd.read_sql_query("""
          SELECT coord_ra, coord_dec
          FROM dc2_object_run2_2i_dr6_wfd.dpdd_forced
          WHERE i_base_PsfFlux_instFlux BETWEEN 0.00000 and 0.00001
      """, conn)
      df
```

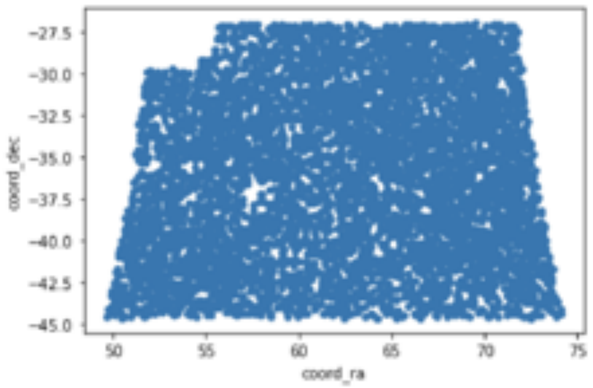
[16]:

	coord_ra	coord_dec
0	49.661000	-44.480226
1	51.676920	-44.494506
2	50.963874	-44.553124
3	50.477251	-44.571932
4	50.597745	-44.548080
...	...	...
7046	67.423669	-27.197836
7047	67.443431	-27.499548
7048	67.239164	-27.490121
7049	69.985264	-27.391659
7050	70.085414	-27.183343

7051 rows × 2 columns

```
[19]: df.plot.scatter('coord_ra','coord_dec')
```

[19]: <AxesSubplot:xlabel='coord_ra', ylabel='coord_dec'>



Community Engagement & Data Rights

- Data rights holders (all US & Chilean astronomers plus international partners via in-kind contributions) will have access to the Rubin Science Platform
- Annual data releases become public **after 2 years**
- A subset of data will be public via the Education & Public Outreach team, including for citizen science projects
- **Alert stream and software are fully public** 60s
- Students can apply for LSSTC DSFP in spring/summer

The Science Pillars of Rubin Observatory and the LSST

LSST Survey Cadence Optimization

Data Products for Time Domain Astronomy

How to Get (More) Involved with Rubin Observatory

Annual Data Releases and User-Generated Data Products

Community Participation in Data Preview 0

The Rubin Science Platform: an Environment for Data Access and Analysis

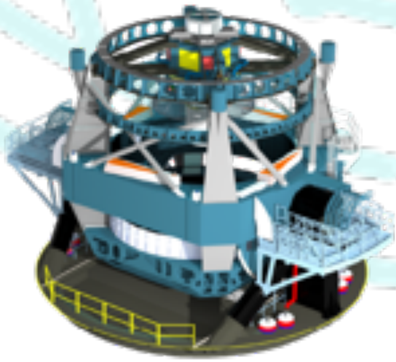
Community Engagement and Support for Science

Bilingual in English and Spanish

Data Previews begin this summer

- DP0 is simulated “LSST-like” data products in the Rubin Science Platform, June 2021
  - Based on the LSST DESC DC2 Simulated Sky Survey, lsstdesc-portal.nersc.gov
- Access to Rubin staff and **up to 300 scientists with data rights**, applications due Apr 30
- Enable operations team to scale up and enable community to prepare for early science
- DP0 is simulated data with no difference imaging products or alerts
- DP1 and DP2 will use real data from commissioning and include high-latency alerts
- More on DP0: community.lsst.org/tag/dp0
- **Latest on int’l data rights** via in-kind contributions: community.lsst.org/tag/in-kind

Community Participation in Data Preview 0



Collaborations are where the science happens

- Science Collaborations are networks to support data exploration
- Design & test user-generated data products, pipelines, brokers
- Source of expert opinion and analysis for Rubin project
- Welcome and train new scientists with Rubin tools
- Some collaborations welcome new members independent of data rights status (this may vary)
- **No** membership fees or time commitment requirements
- New joint onboarding program and other equity initiatives

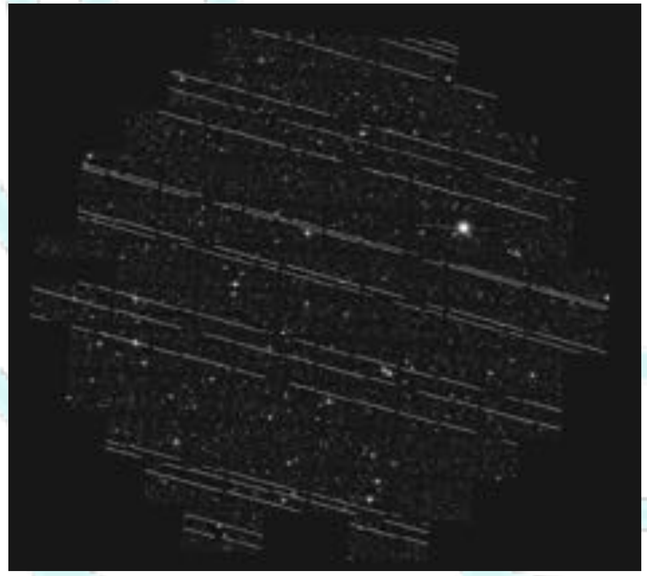
The SCs aspire to be an inclusive & nurturing environment for scientists interested in pursuing LSST-based science

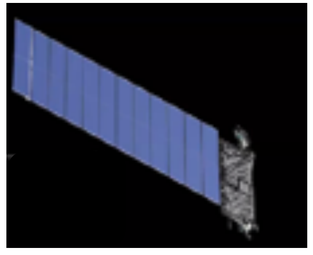
TVS DESC GalaxiesSC SLSC SMWLV SSSC AGN ISSC

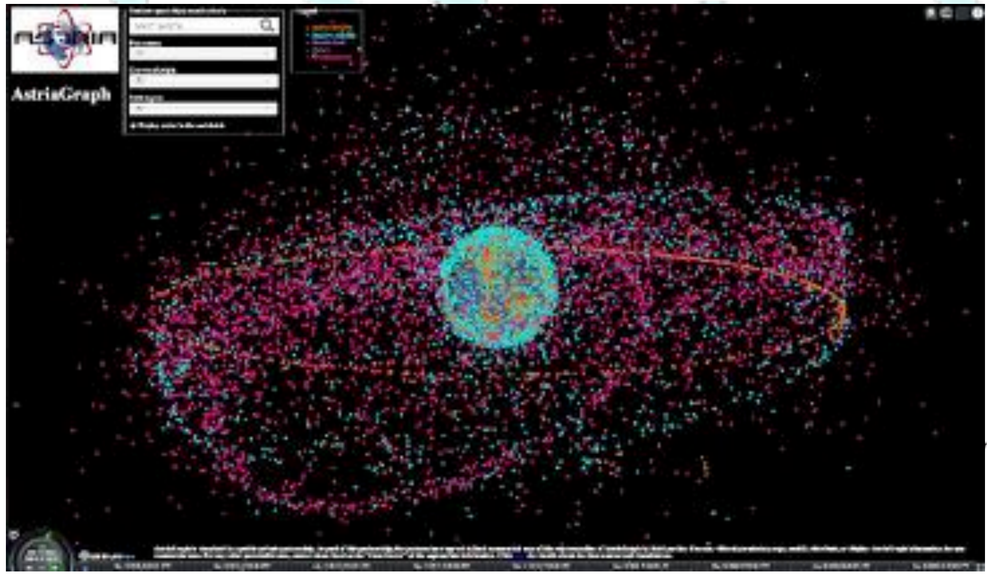
The data and the science sound exciting — are there any potential problems?

Hugely increasing rate of new LEO satellites

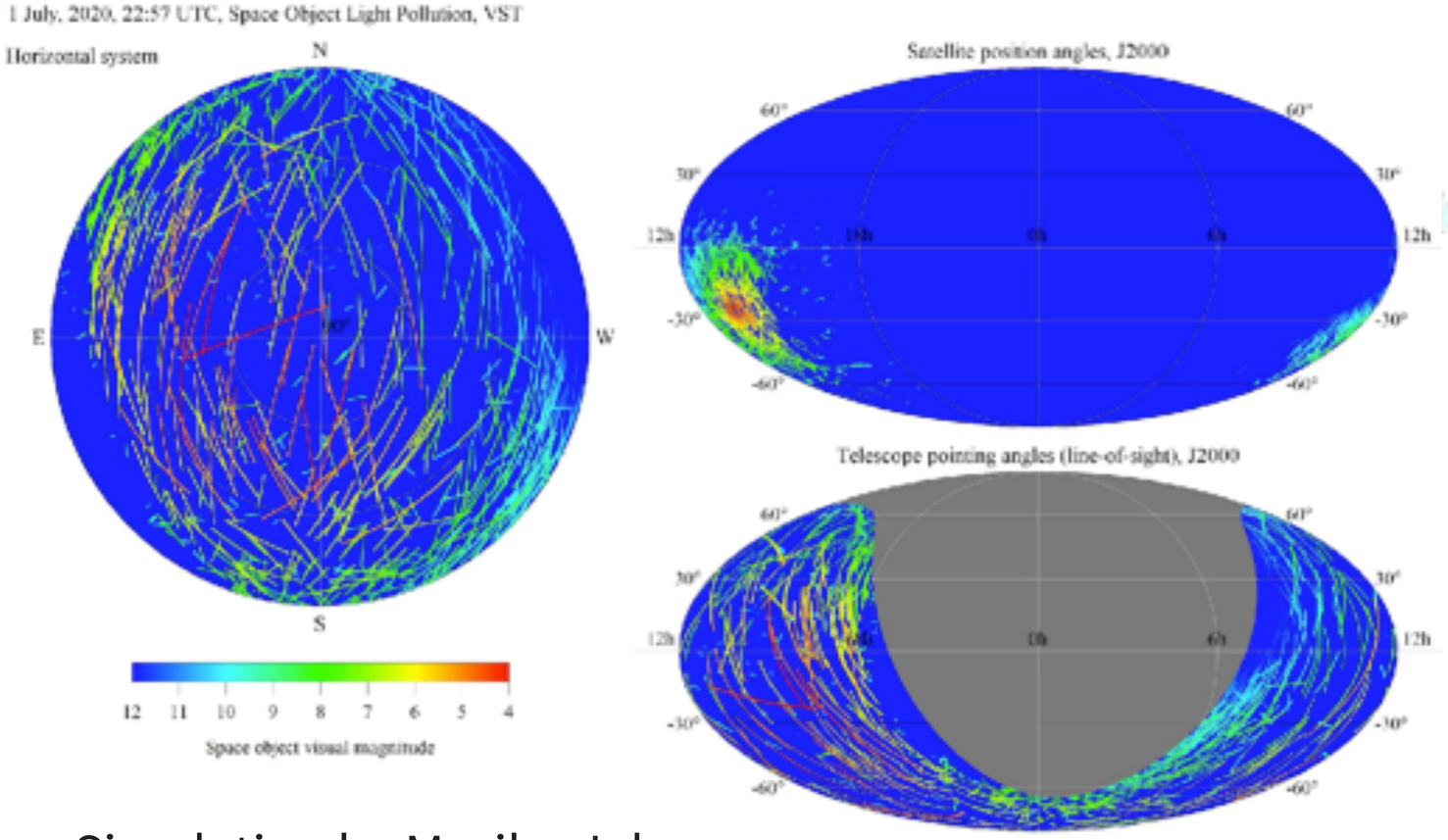
- Most prevalent near twilight, but some **illuminated all night long**
- Impacts are worst for large wide-field ground-based facilities
- New satellites launching and old ones de-orbiting
- **Key recommendations** include: designing to 7th mag or dimmer; sharing timely, accurate, and precise trajectories; low altitudes
- Must consider significant impacts to IR, sub-mm, **radio**, stargazers worldwide, photographers, etc.
- **Funding for crucial mitigation work still unclear**



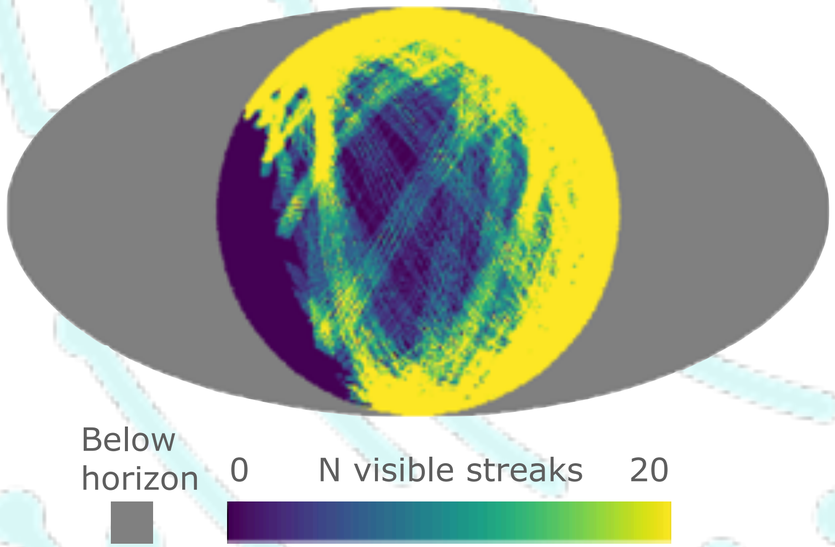


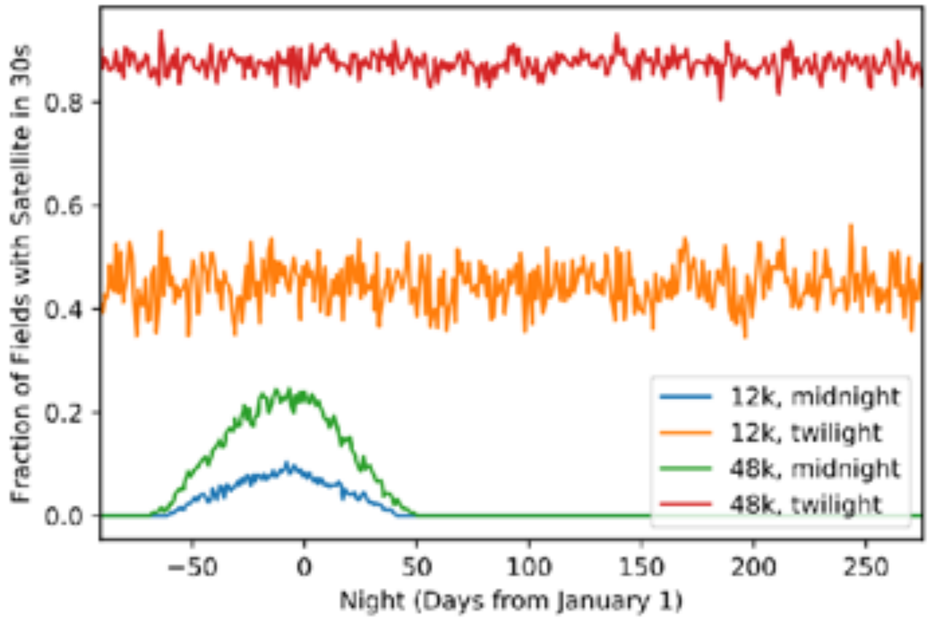


Rubin's potential for discovery is also its vulnerability to LEO satellites



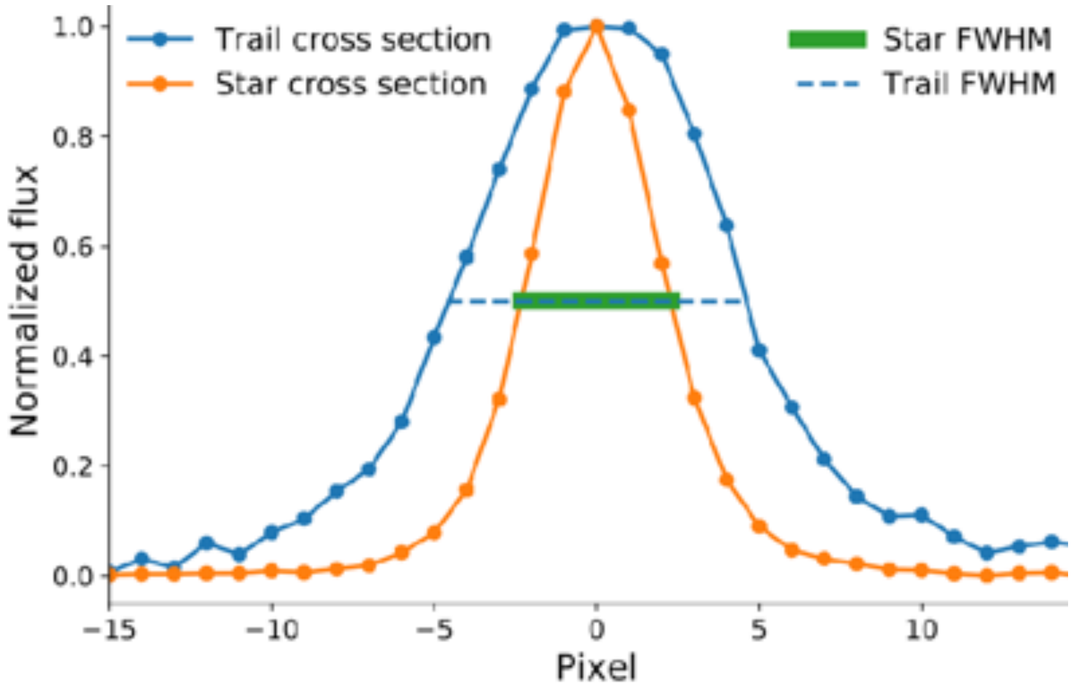
Simulation by Moriba Jah

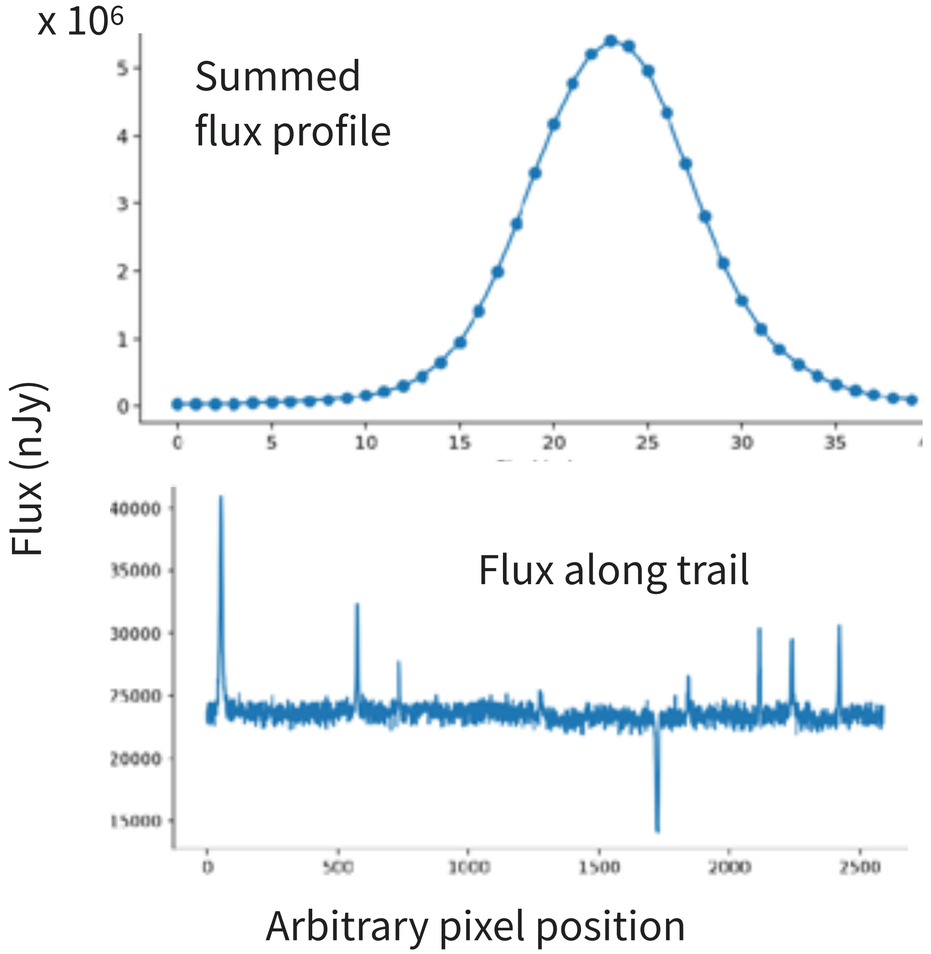




Lower satellites have lower peak brightness

- Higher satellites, > 600 km altitude, are worse!
- **Slower speed + sharper focus = more time on each pixel**
- Most bright streaks don't saturate detectors, but subtle effects like non-linear crosstalk can still affect science





One science impact: finding “killer” asteroids

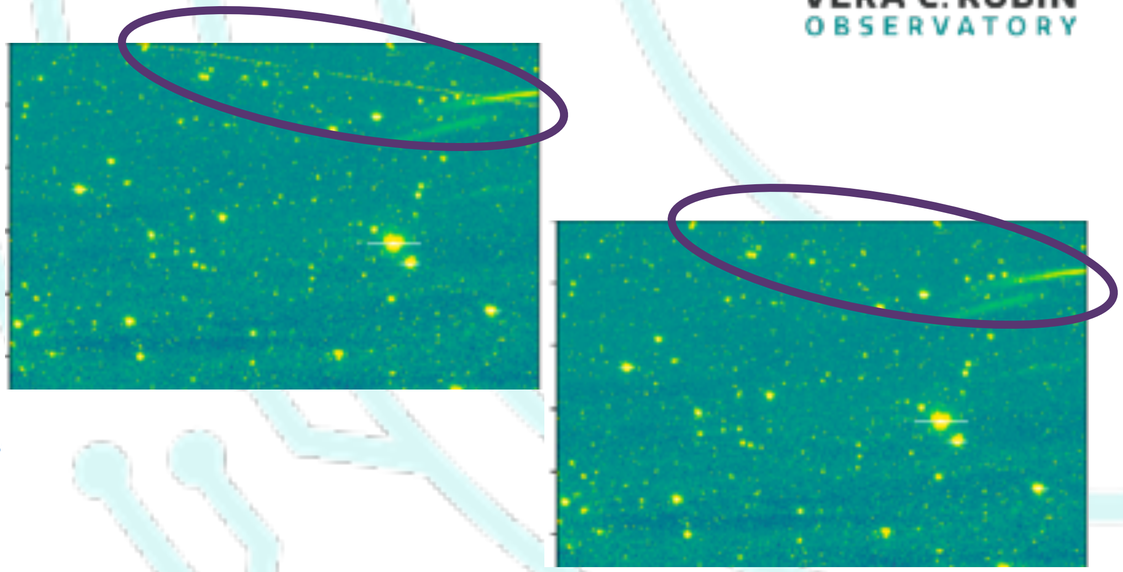
- To discover an asteroid and map its orbit, we need at least 3 images close in time
- Most near-Earth asteroids are best observed in twilight, when most satellites are visible
- If a satellite streaks through one of the images, we lose the opportunity to measure an orbit

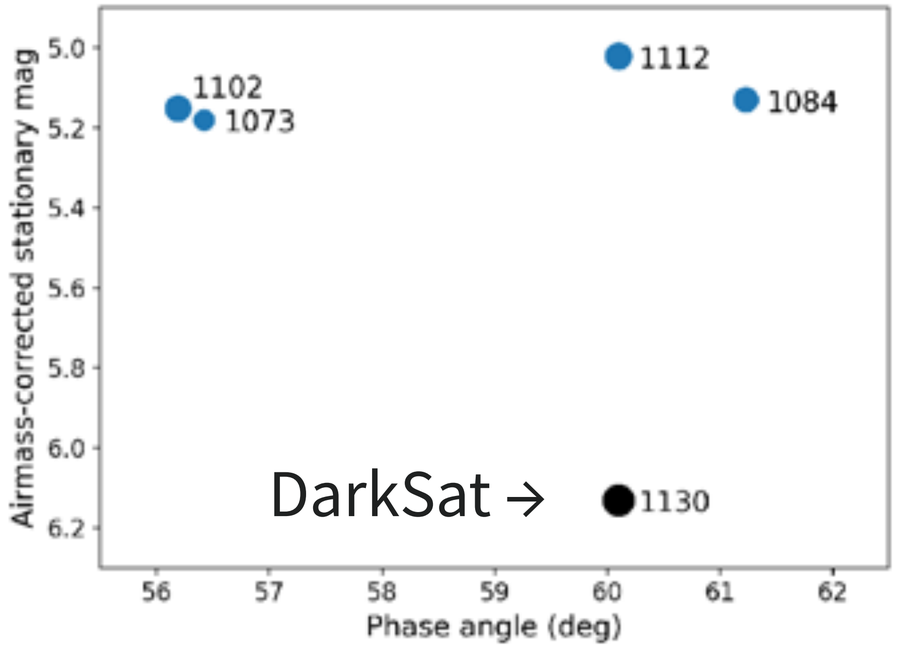
Visualization by Jonny Hyman

More: youtu.be/TifUa8ENQeses

An uneasy future with 50–100k LEO satellites

- Astronomers are in conversation with operators
- DarkSat and VisorSat Starlinks are ~6th V mag
- **All mitigations are presently voluntary**
- SATCON1: aas.org/satellite-constellations-1-workshop-report
- Look for SATCON2 in July 2021
- Dark & Quiet Skies: iau.org/news/announcements/detail/ann21002
- Software can mask trails in stacked images,
  but lines of unusable pixels can still affect science
- Observational efforts are ongoing, but this takes
  time and $ away from the science we want to do
- **We won't know what we don't discover**





Rubin LSST → Big Data → Science is a huge collaborative effort

- Despite the growing LEO satellite problem, Rubin Observatory's LSST will revolutionize astronomy
- 10-year sky survey from Chile begins 2023
- Multi-color time-resolved faint sky map and real-time finder of changing things
- 37B cataloged objects and 10M nightly alerts
- Alert stream and software are fully public, and annual data releases are public after 2 years
- Rubin Science Platform is a window to the data
- Get ready for petabytes of astronomy

Thank you!