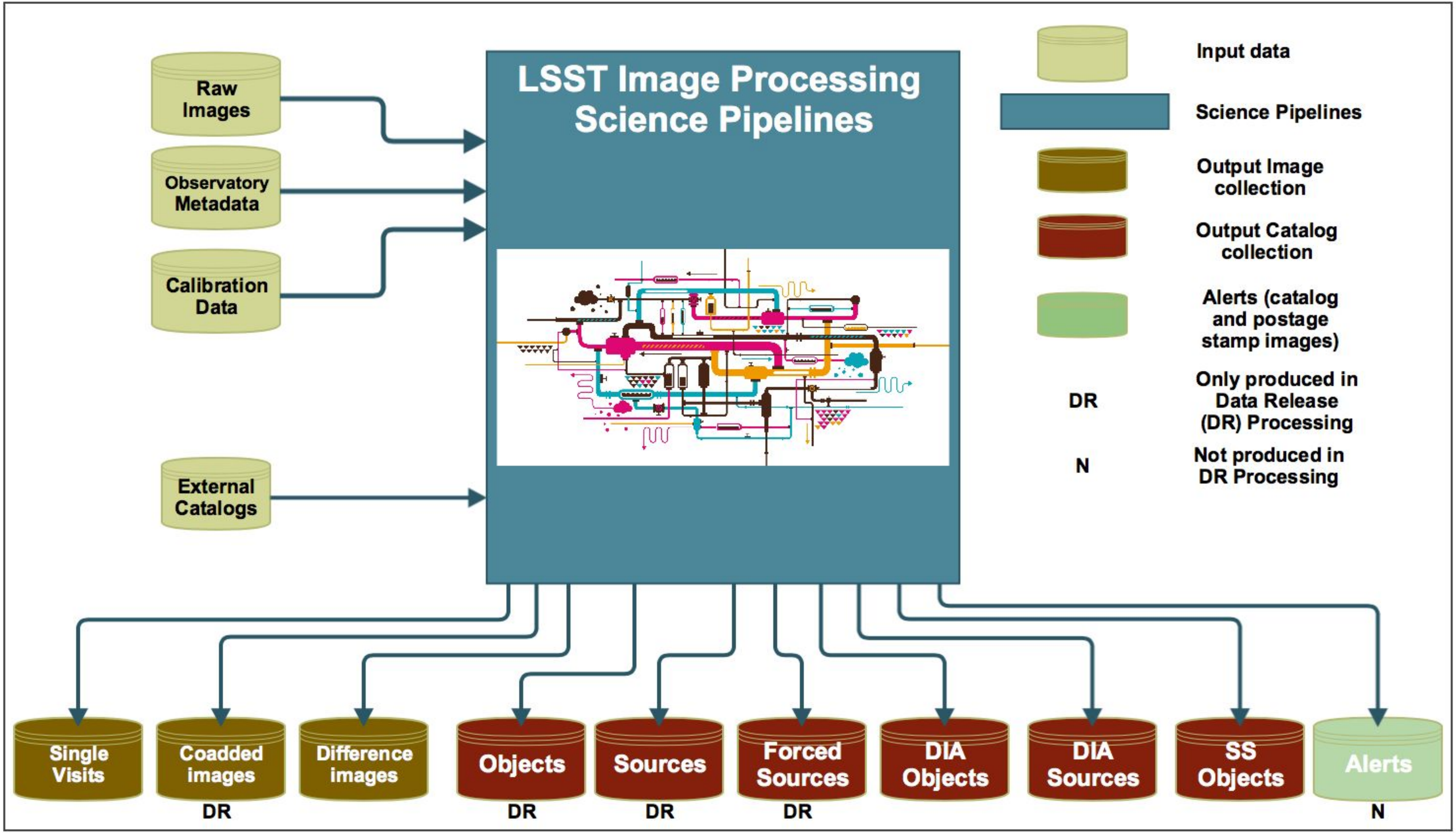
VERA C. RUBIN OBSERVATORY

How DM uses precursor datasets and stories from HS

Yusra AlSayyad

Project and Community Workshop
August 10 2023

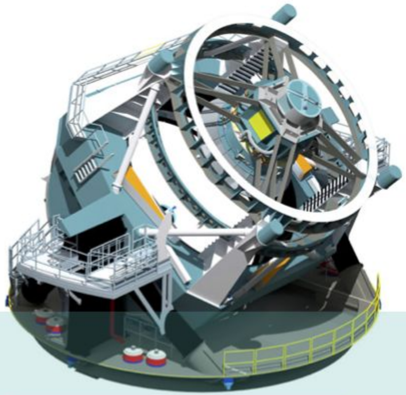


LSST Image Processing Science Pipelines
Raw Images
Observatory Metadata
Calibration Data
External Catalogs
Input data
Science Pipelines
Output Image collection
Output Catalog collection
Alerts (catalog and postage stamp images)
DR
Only produced in Data Release (DR) Processing
N
Not produced in DR Processing
Single Visits
Coadded images
DR
Difference images
Objects
DR
Sources
DR
Forced Sources
DR
DIA Objects
DIA Sources
SS Objects
Alerts
N

You will get data products fast and slow

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

Raw & Processed Visit Images, Difference Images, Templates

Transient and variable sources from Difference Image Analysis

Solar System Objects: ~ 6 million

— 24h via Prompt Products DB

Data Release Data Products

Final 10yr Data Release:

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

— via Data Releases

Rubin Data Access Centres (DACs)

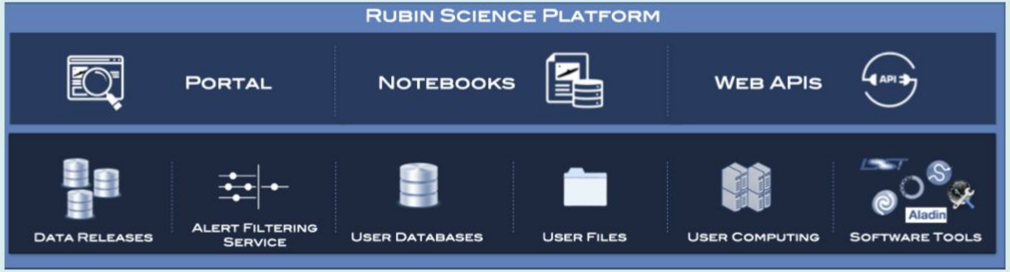
USA (USDF)
Chile (CLDF)
France (FRDF)
United Kingdom (UKDF)

Independent Data Access Centers (IDACs)

Access to proprietary data and the Science Platform require Rubin data rights

Rubin Science Platform

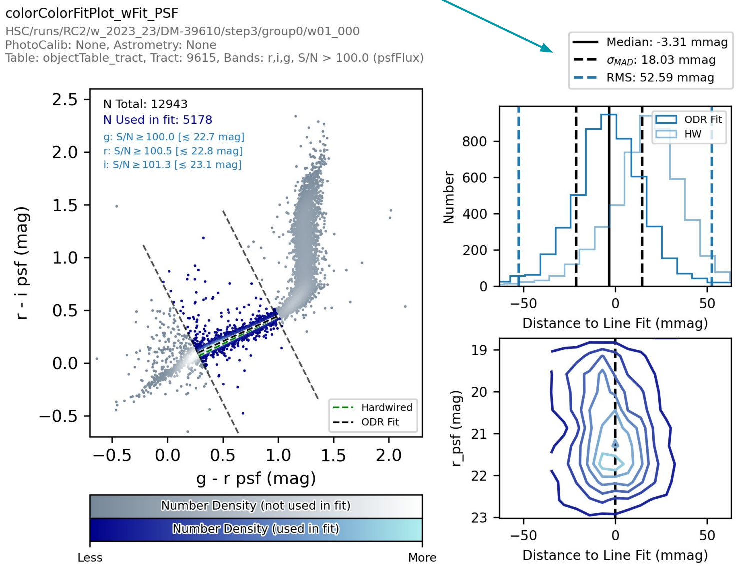
Provides access to LSST Data Products and services for all science users and project staff.

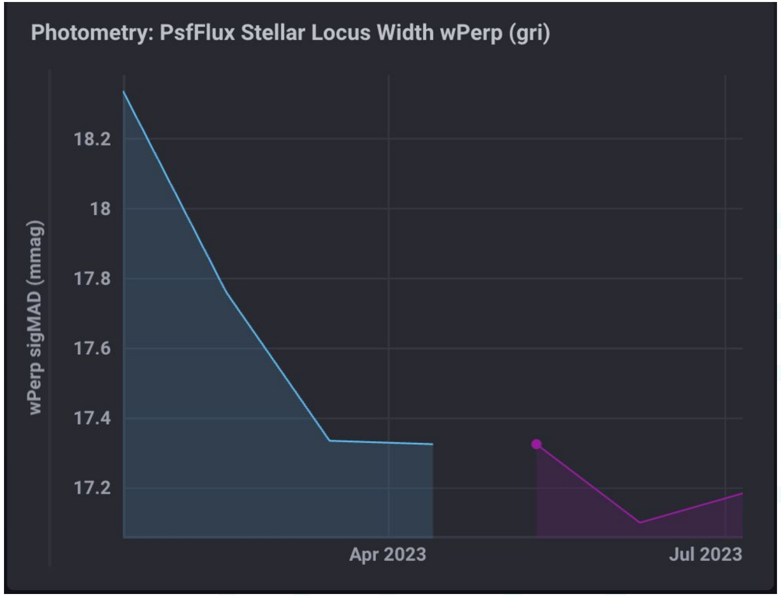


Credit: Leanne Guy

Plots and Metrics are computed during pipeline execution alongside the algorithms

The **metrics** printed on the **plot**, **match** those written to the butler and dispatched to Sasquatch





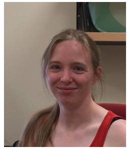
Sasquatch is the Rubin Observatory service for recording, displaying, and alerting on telemetry data and metrics
See sasquatch.lsst.io

We welcome your pull requests to analysis tools

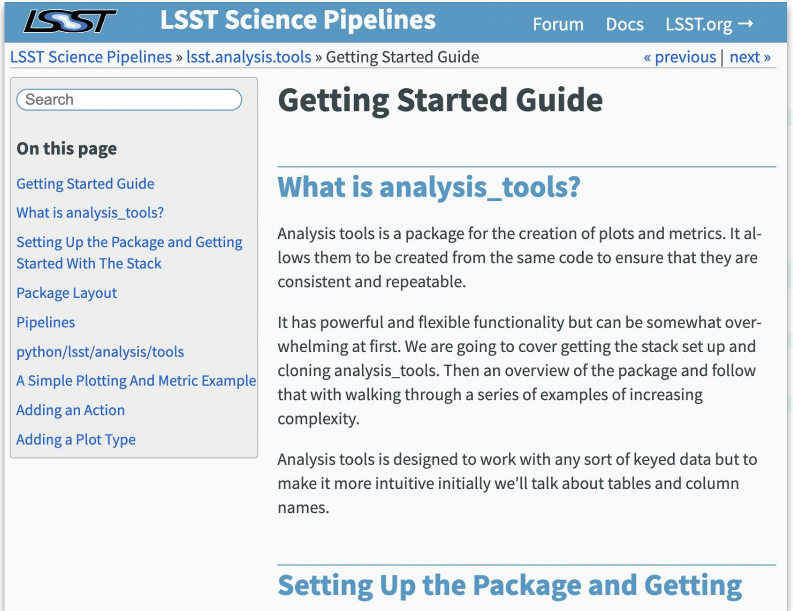
Last year, the metric and plotting code was refactored and redesigned into a framework/package called analysis_tools.

analysis_tools was intentionally designed in a modular way such that members of the science community can contribute analysis code to be automatically run as part of science pipelines processing and generate science performance diagnostic plots and metrics

See tutorials from the [May 2023 Commissioning Science Validation Bootcamp](#) and the new [getting started guide](#)

Nate Lust Sophie Reed

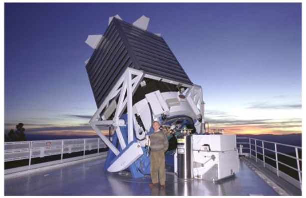


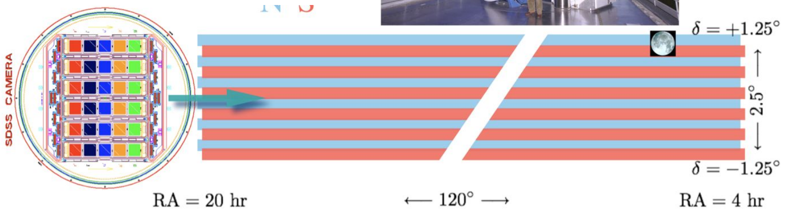
We’ve been running the pipelines on 100s of sq deg precursor dataset since the beginning

In the pre-construction era we called them “data challenges”

In 2012 we were validating the coaddition algorithms and forced photometry **on SDSS Stripe 82** (dmtn-034.lsst.io)
In 2013 we did a joint reprocessing with the FrDF, with some improvements (e.g. background matching)

300 sq deg
ugriz
80 epochs over
10 yr baseline



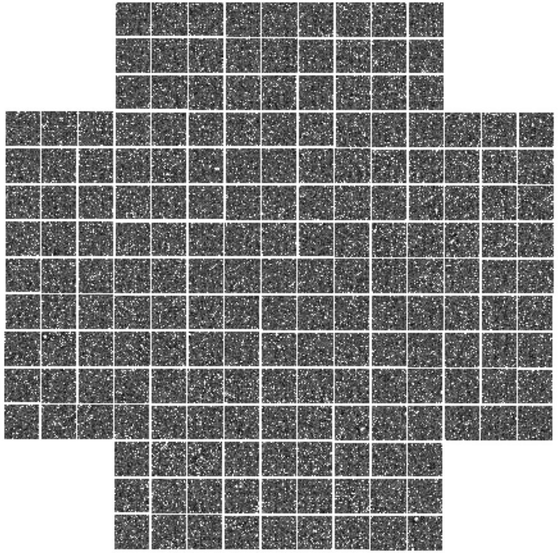


Since then we regularly process precursor datasets from 4 cameras (1 simulated) with more in common with LSSTCam

Rubin AuxTel LATISS

1.2; 6.7 arcmin diam
1 real LSST CCD

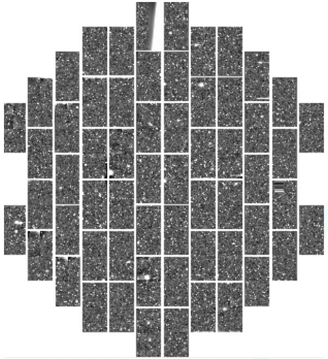
SIMULATIONS **LSST ImSim**
DESC's DC2 Run2.2i



6.5m; 3.5 deg diam
189 4k X 4k CCDs, ugrizy

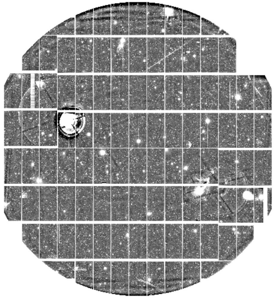
Meredith and Lee speaking to this next

Dark Energy Camera (DECam)



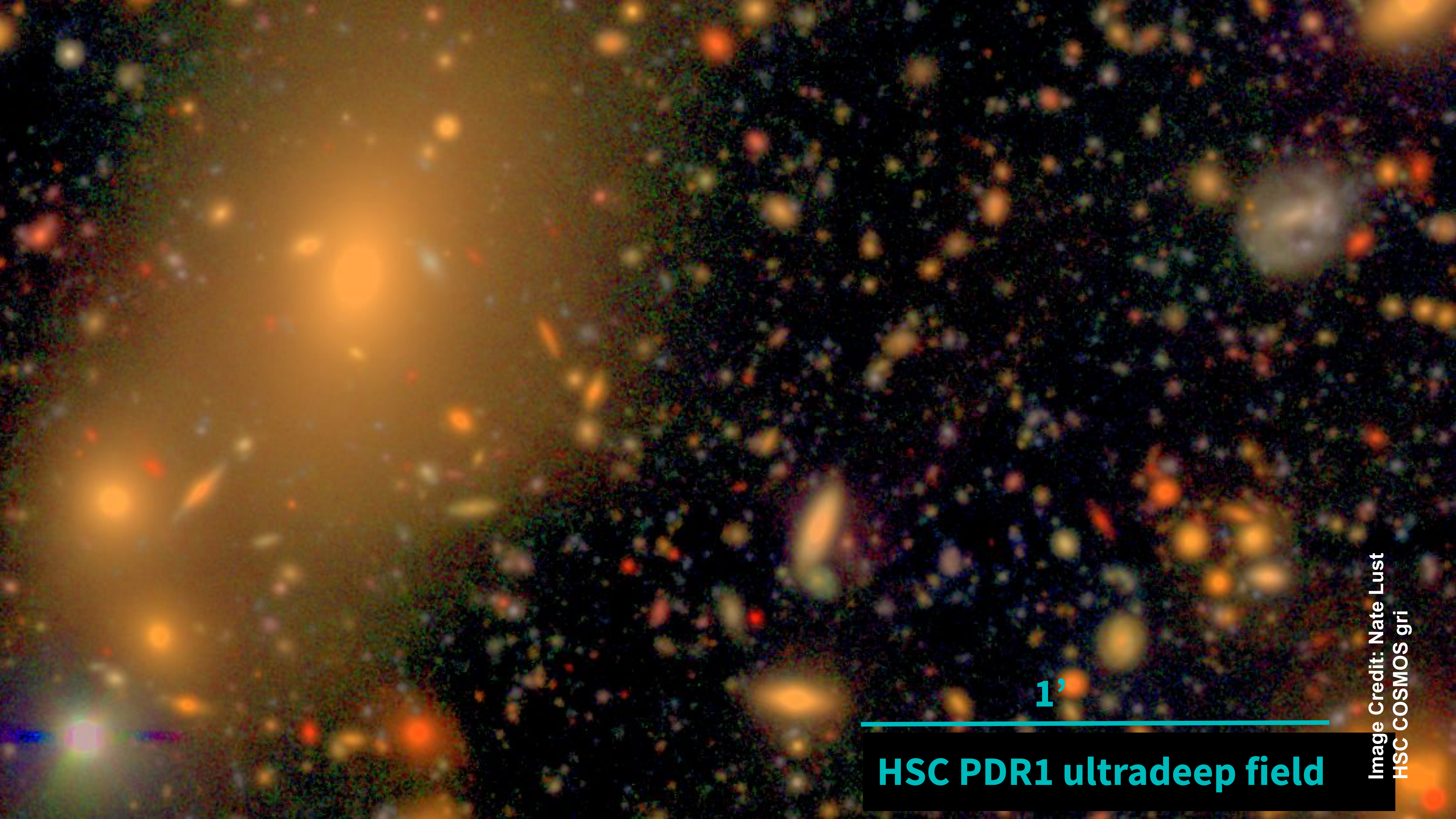
3.9m; 2.2 deg diam
62 2k X 4k CCDs

Hyper Suprime-Cam (HSC) Subaru Strategic Program



8.2m; 1.5 deg diam
103 2k X 4k CCDs
grizy

Figures: Pipeline-processed visit images (PVIs) aka calexps


1'
HSC PDR1 ultradeep field
Image Credit: Nate Lust
HSC COSMOS gri

And the LSST Pipelines are the Hyper Suprime-Cam (HSC SSP) Pipelines

Survey Comparison	**LSST**	**HSC** (Subaru Strategic Program)
Effective Aperture	6.5m	8.2m
Filters	ugrizy	grizy + narrow
Exp time per visit	~30s	~240s
Field of View	10 deg^2 3.5 deg diam	1.8 deg^2 1.5 deg diam
Num CCDs	189 (4k x 4k)	103 (4k x 2k)

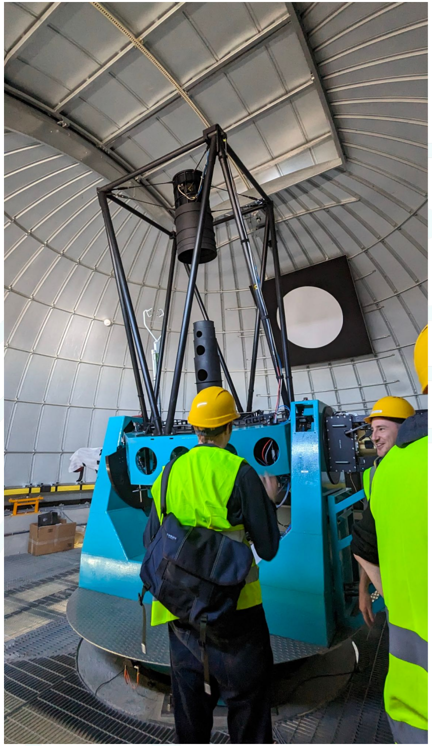
Image Credit: Nate Lust
HSC COSMOS gri

We run pipelines on precursor data in **two modes now**
Fix images vary pipelines vs. Fix pipelines vary images

1) As always, we analyze pipeline performance on **fixed datasets** of 3 sizes and cadences:
   - Small areas (< 1 deg^2) on a **nightly** cadence
   - Medium areas (~10 of deg^2) on a **monthly** cadence
   - Large areas (100s of deg^2) on **annual** cadence

2) And now also run pipelines routinely on **real-time AuxTel observations** via 3 campaigns:
   - **Rapid Analysis**
   - **10am DRP Processing**
   - **Nightly Prompt Processing**

DMTN-091: Test Datasets for Scientific Performance Monitoring



Colin, Merlin, and AuxTel

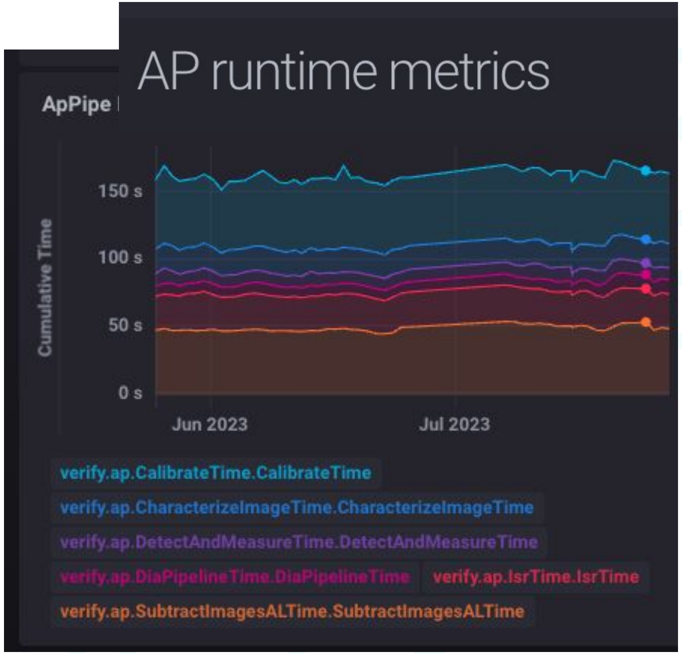
Daily small (<1 deg^2) reruns in continuous integration keep pipelines healthy

Two builds are launched during **each nightly software release** that **produce metrics** dispatched to Sasquatch on [Chronograf](#).

- **rc2_subset** tests the **DRP** and is a one patch subset of **HSC RC2** (next slide) that is the same one as the getting started guide on pipelines.lsst.io
- **ap_verify** runs nightly on a few
  - HSC ccds ap_verify_ci_cosmos_pdr2
  - DECam ccds ap_verify_ci_hits2015
  - ImSim ccds ap_verify_ci_dc2

Tracks science quality metrics and comp perf metrics

Runtime (s) on an ImSim ccd



Date of nightly release

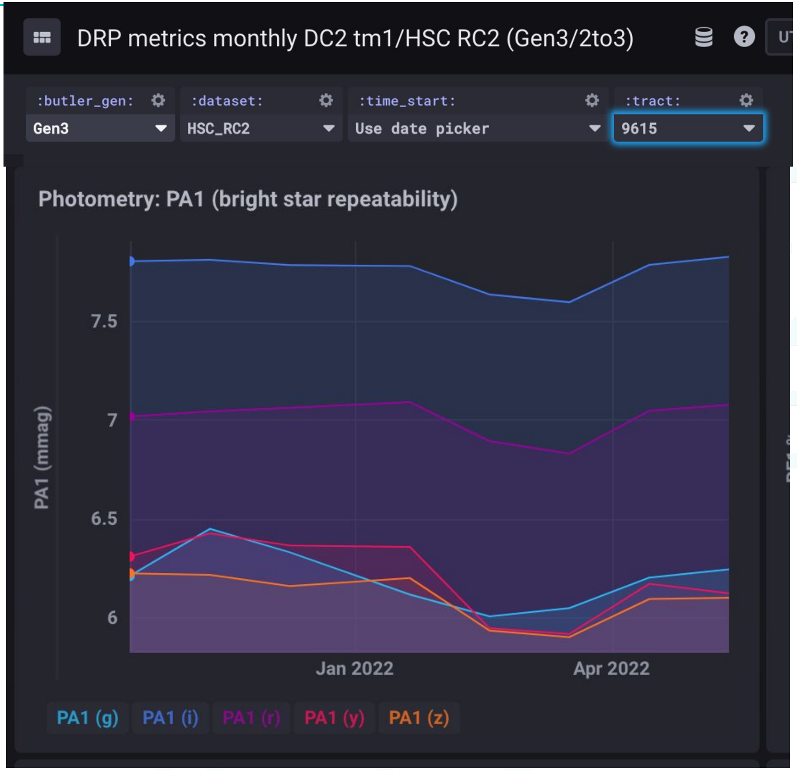
Monthly medium (<10 deg^2) reruns provide a testbed for new algorithms

Monthly continuous integration on datasets of ~5 deg^2

- Tract-sized datasets are the **minimum** to test the speed, robustness, and performance of any algorithmic change.
  …and **plots** as a function of
  …n (i.e., hold data fixed and
  …ange pipeline)
- AP on HSC COSMOS and DECam HiTS (Förster+16)
- DRP on **HSC RC2** (3 tracts) and ImSim **DC2 2.2i test-med-1** (2 tracts at 1.5 yr depth). All bands.

Next talk!

Photometric Repeatability (mmag)



Date of weekly release

Large (100s deg^2) processing campaigns test whole of data management and rare edge cases

Pipelines are used in **external data releases;** the best QA is astronomers publishing papers with your data products:

- **Feb 2017** — **HSC Public Data Release 1 (PDR1)**: 100 deg^2.
- **May 2019** — **HSC PDR2**: 300 deg^2; 1100 deg2 at least one filter.
- **Fall 2021** — **HSC PDR3**: 670 deg^2, 1470 deg^2 at least one filter
- **Jun 2022** — **Rubin Data Preview 0.2 on ImSim data**: 300 deg^2 ugrizy
- **In progress v24.1.0** — **HSC Final Data Release (PDR4)**

Internal reprocessing campaigns integrate all of DM, provide data for characterization reports, and test algorithms for robustness:

- **May 2018** — **HSC PDR1 reprocessing**
- **Feb 2020** — **HSC PDR2 reprocessing**
- **In progress 2023 v24.1.0** — **HSC PDR2 reprocessing**

If you're curious about how pipelines behave on real data of comparable depth, check out the HSC SSP public data releases

- https://hsc-release.mtk.nao.ac.jp/
- Benefits:
    - Real data of comparable depth
    - Same algorithms and flags
- Caveats:
    - The column names are different
    - No Science Platform available
    - Does not include DIA Data Products

Hyper Suprime-Cam Subaru Strategic Program

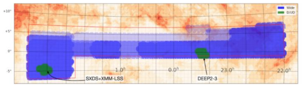
Public Data Release

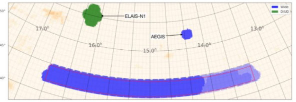
Home | Survey | Processing | Database | Available Data | Data Access | FAQ

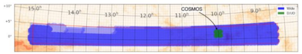
hsc-release.mtk.nao.ac.jp/doc/index.php/available-data__pdr3/

Available Data (PDR3)

HSC-SSP PDR3 includes over 600 square degrees of multi-band data at the nominal survey depth. See the figures below for the survey footprints. The blue and green areas show the Wide and Deep+UltraDeep layers, respectively. The darker blue regions are covered in more filters (max. 5).







Wide	g	r	i	z	y				
exposure (min)	10^{+2}_{-2}	10^{+2}_{-2}	20^{+3}_{-6}	20^{+3}_{-10}	20^{+3}_{-10}				
seeing (arcsec)	$0.79^{+0.09}_{-0.08}$	$0.75^{+0.13}_{-0.09}$	$0.61^{+0.05}_{-0.05}$	$0.68^{+0.08}_{-0.06}$	$0.68^{+0.10}_{-0.08}$				
depth (mag)	$26.5^{+0.2}_{-0.2}$	$26.5^{+0.2}_{-0.2}$	$26.2^{+0.2}_{-0.3}$	$25.2^{+0.2}_{-0.3}$	$24.4^{+0.2}_{-0.3}$				
saturation (mag)	$17.4^{+0.6}_{-0.4}$	$18.1^{+0.5}_{-0.5}$	$18.3^{+0.5}_{-0.3}$	$17.5^{+0.5}_{-0.4}$	$17.0^{+0.5}_{-0.7}$				
area (deg^2)	1332	1298	1264	1299	1209				
Deep+UltraDeep	g	r	i	z	y	$NB387$	$NB816$	$NB921$	$NB1010$
exposure (min)	70^{+21}_{-21}	66^{+17}_{-17}	98^{+46}_{-32}	177^{+130}_{-46}	93^{+23}_{-23}	68^{+13}_{-13}	120^{+30}_{-15}	168^{+14}_{-28}	705^{+45}_{-345}
seeing (arcsec)	$0.83^{+0.05}_{-0.12}$	$0.77^{+0.04}_{-0.04}$	$0.66^{+0.07}_{-0.06}$	$0.78^{+0.02}_{-0.03}$	$0.70^{+0.04}_{-0.05}$	$0.82^{+0.07}_{-0.08}$	$0.70^{+0.07}_{-0.08}$	$0.67^{+0.04}_{-0.04}$	$0.77^{+0.02}_{-0.02}$
depth (mag)	$27.4^{+0.2}_{-0.2}$	$27.1^{+0.1}_{-0.2}$	$26.9^{+0.2}_{-0.3}$	$26.3^{+0.1}_{-0.3}$	$25.3^{+0.2}_{-0.2}$	$25.0^{+0.2}_{-0.2}$	$26.0^{+0.2}_{-0.2}$	$25.9^{+0.2}_{-0.2}$	$24.2^{+0.2}_{-0.5}$
saturation (mag)	$18.0^{+0.4}_{-0.5}$	$18.2^{+0.4}_{-0.4}$	$18.6^{+0.3}_{-0.4}$	$17.7^{+0.3}_{-0.3}$	$17.4^{+0.3}_{-0.3}$	$14.8^{+0.4}_{-0.3}$	$16.8^{+0.4}_{-0.4}$	$16.9^{+0.4}_{-0.3}$	$14.8^{+0.2}_{-0.2}$
area (deg^2)	36	36	36	37	36	30	33	33	5

The table gives a quick overview of the quality of our data. The depths are given as 5 sigma limiting magnitudes for point sources. Area is the area covered in at least 1 exposure in each filter.

Data Retrieval

The data can be retrieved in multiple ways. The simplest way to retrieve catalog data is to use the database. We have online/offline SQL tools. For image data, most users will find hscMap, an online image browser, very useful. For binary files, we have a data search tool as well as image cutout tool. All these tools are summarized in the Data Access page. In order to access the data, you first have to sign up for an account. Before you use our data products, we strongly recommend you to go over the data release paper and the Known Problems page. If you use the HSC data in your publication, please acknowledge us. This site serves only the processed data. Raw data can be retrieved from SMOKA.

Data Quality

We have performed a number of validation tests for our data products. A complete set of the plots can be found here.

Quality Assurance Plots | Stellar Sequence | Star/Galaxy Separation

Large (100s deg^2) processing campaigns test whole of data management and rare edge cases

Pipelines are used in **external data releases;** the best QA is astronomers publishing papers with your data products:

Feb 2017 — **HSC Public Data Release 1 (PDR1)**
100 deg^2.

May 2019 — **HSC PDR2**
300 deg^2; 1100 deg2 at least one filter.

Fall 2021 — **HSC PDR3**
670 deg^2, 1470 deg^2 at least one filter

Jun 2022 — **Rubin Data Preview 0.2 on ImSim data**
300 deg^2 ugrizy

In progress v24.1.0 — **HSC Final Data Release (PDR4)**

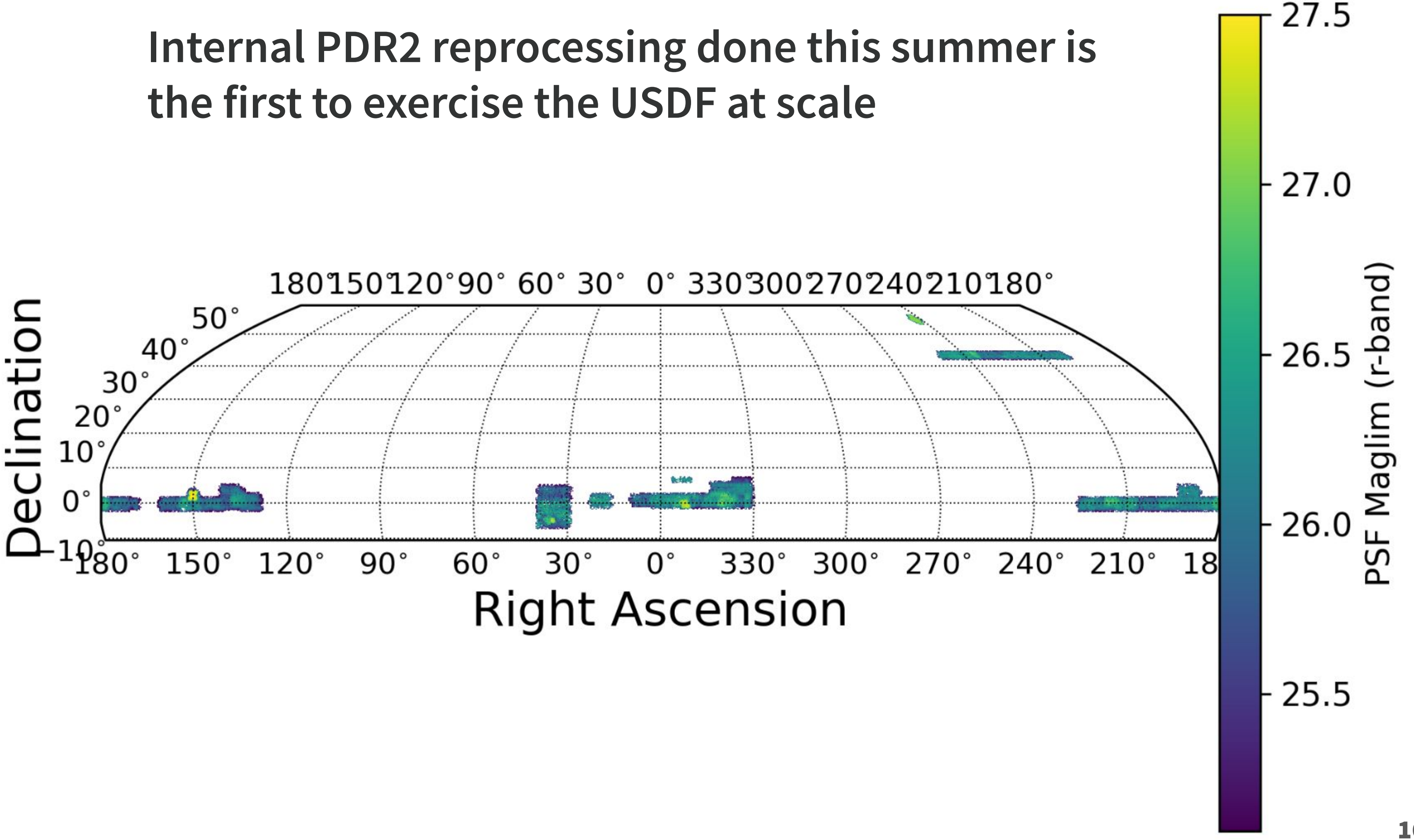
Internal reprocessing campaigns integrate all of DM, provide data for characterization reports, and test algorithms for robustness:

May 2018 — **HSC PDR1 reprocessing**

Feb 2020 — **HSC PDR2 reprocessing**

In progress 2023 v24.1.0 — **HSC PDR2 reprocessing**

Internal PDR2 reprocessing done this summer is the first to exercise the USDF at scale

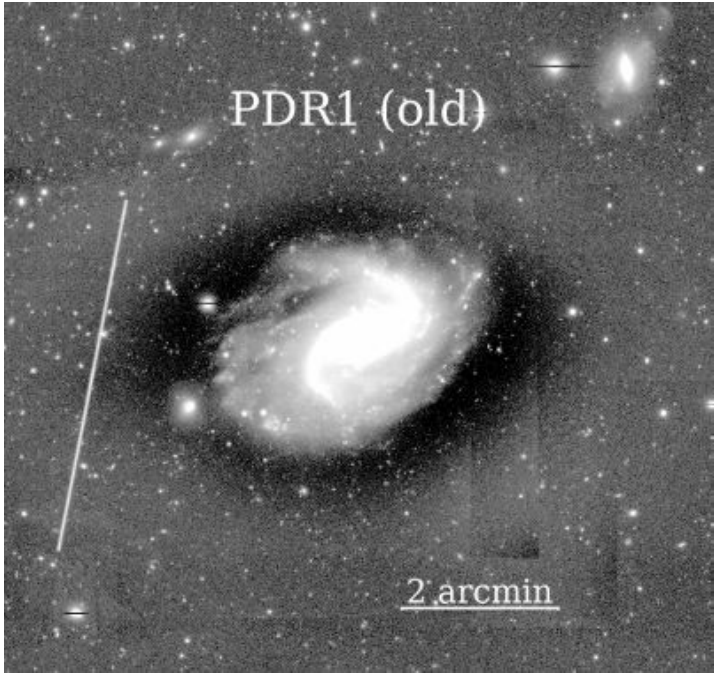


VERA C. RUBIN OBSERVATORY

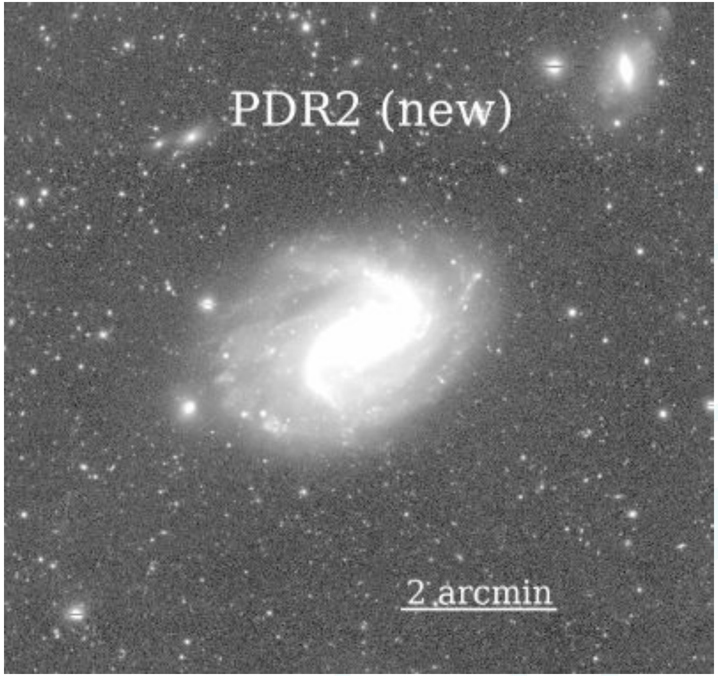
Stories from HSC

(if there's time)

Low Surface Brightness community was happy with PDR2 full focal plane “SkyCorrection”



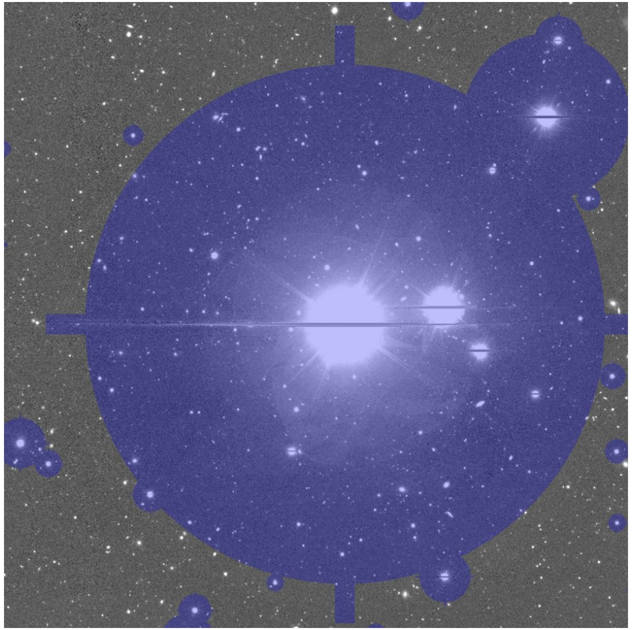
Coadd with
PDR1 Local Background subtraction

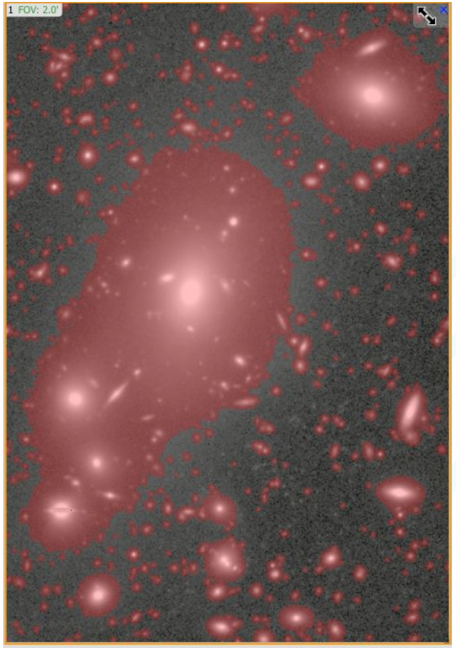


Coadd with
PDR2 Focal Plane Background subtraction

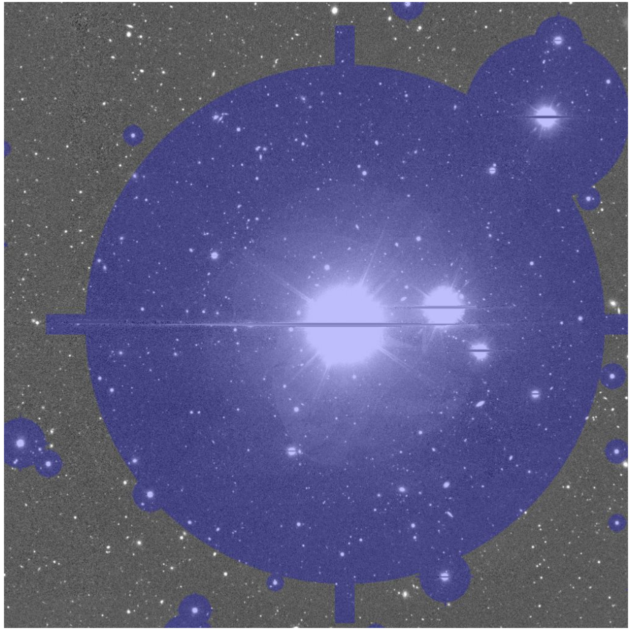
Aihara+19 (PDR2 release paper)

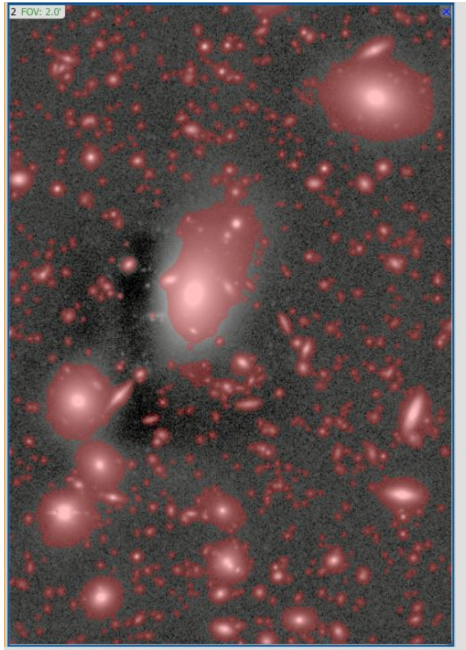
But everyone else was unhappy





But everyone else was unhappy

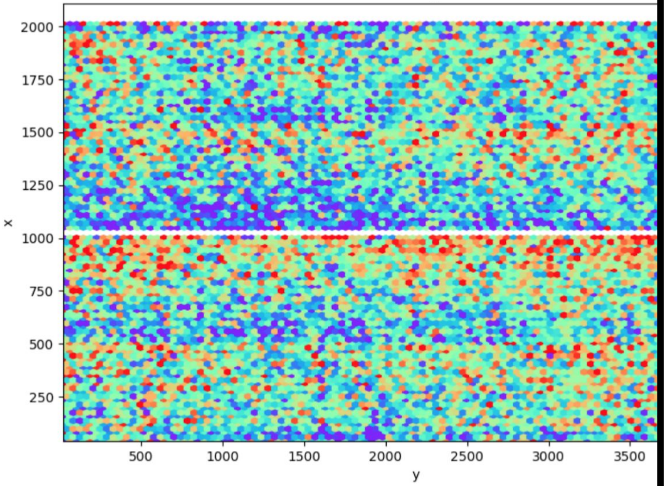


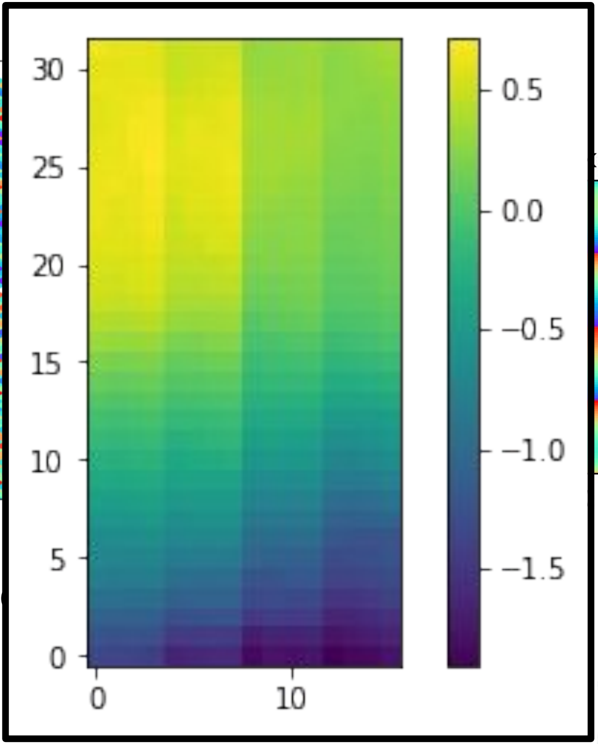


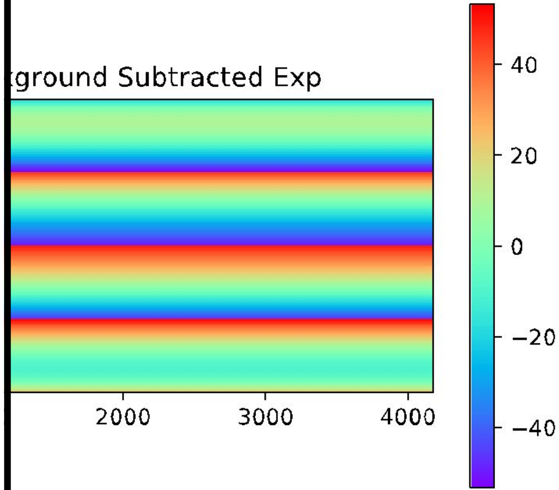
PDR3 adds a very aggressive 128x128 binned spline background subtraction on the deepCoadd_calexps

deepCoadd still available for LSB detection

Large area also alerted us to very small amplitude systematics

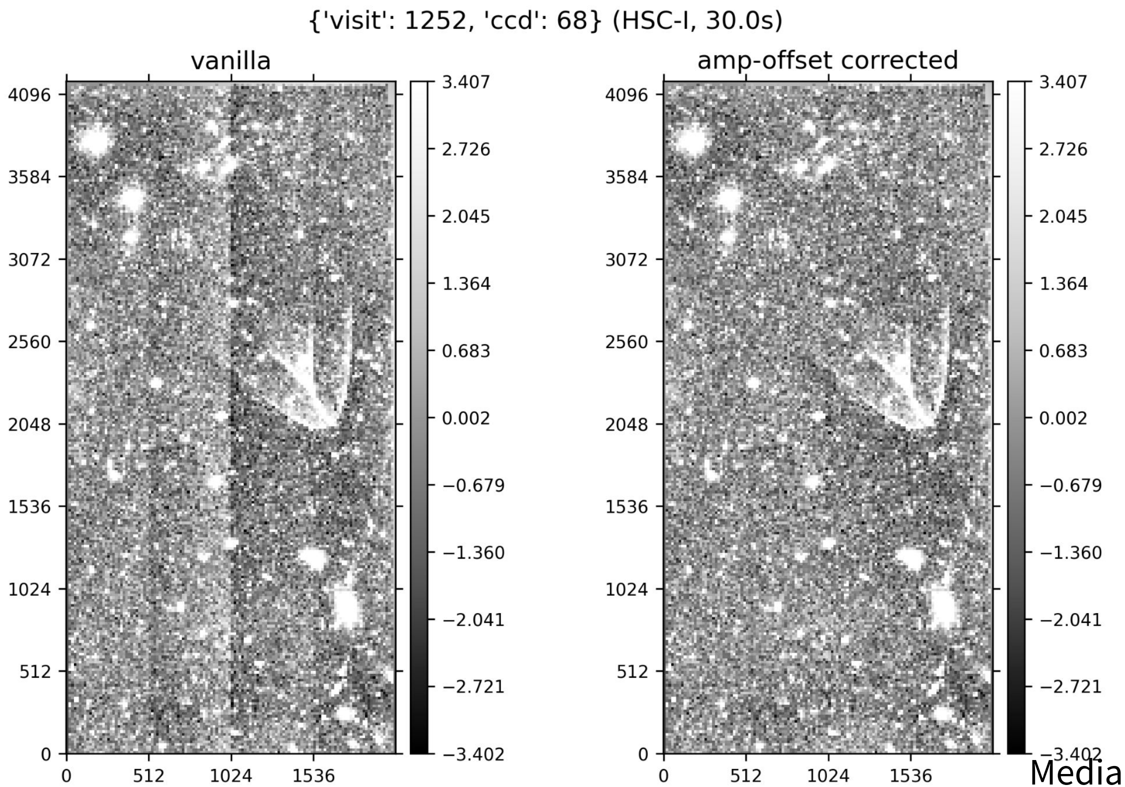




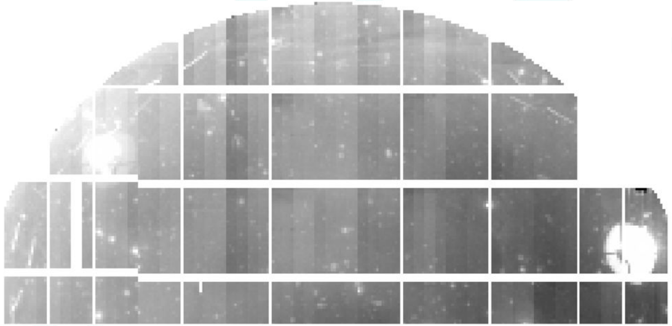


HSC PDR2 g-band photometric resi
stacked by CCD
Figure: Eli Rykoff

Which led to the addition of the the Pan-STARRS “pattern continuity” (a.k.a amp-to-amp matching)



Lee Kelvin



Median binned (128x128) HSC-I visit=1252 (short exp)

Some answers to the discussion questions

Do HSC data releases use a recent version of the LSST pipelines, or rather very similar underlying algorithmic code that was "forked off" some time ago?

- *For HSC PDR1 we **forked** the pipelines hard, meaning that there was development on the fork that was not mirrored upstream. We had a bad time upstreaming changed back to main. Not recommended!*
- *For HSC PDR2/3 we did a **soft fork**: see hscPipe. Bugfixes were applied to the LSST Science Pipelines and backported to the fork.*
- *For HSC PDR4, we are maintaining v24.0.x as the **release branch** on the LSST Science Pipelines. Any bugfixes we know will be needed for PDR4 are backported to v24.*
- *For DP0.2. We maintained the v23.0.x release branch. Any bugfixes needed for DP0.2 were backported. See process outlined here: https://developer.lsst.io/work/backports.html*

When processing precursor data sets, how do you balance keeping up with recent LSST pipelines releases versus maintaining some “stable” choice of LSST pipelines version?

Depends on how long your processing will take:
For >100s of sq deg data release:

- *We maintain a **release branch** on the LSST Science Pipelines. Bugfixes are backported to that release branch.*
- *For DP0.2. We maintained the v23.0.x release branch. Any bugfixes needed for DP0.2 were backported. See process outlined here: https://developer.lsst.io/work/backports.html*

For the monthly reprocessings we’re a little more loose, use w_2023_XX and use ticket branches if we need to patch anything.

For the nightlies, we just use the daily release. If it fails, too bad, try again the next night.

What types of computational resources are your large LSST pipelines processings of precursor data sets deployed on? Cloud? Academic HPC?

Data releases:

- *DP0.2 was produced in google cloud*
  - *ctrl_bps_panda*
  - *storage: s3*
- *HSC public data releases at the NOAJ Mitaka cluster*
  - *ctrl_bps_htcondor*
  - *Condor over PBS*
  - *Storage: gpfs*

Monthlies:

- *USDF, ctrl_bps_panda and ctrl_bps_htcondor over slurm.*

Do you have any tips for deploying the LSST pipelines at scale in an efficient way, for instance database optimizations (or similar) that make a big difference?

For the tiny rc2_subset example in the getting started guide uses sqlite as the registry and pipetask -j. For larger datasets you’ll need a real DBMS (USDF uses postgres for it now) and BPS: https://pipelines.lsst.io/modules/lsst.ctrl.bps

When processing precursor data sets, how do you manage large output data volumes from the LSST pipelines? Do you delete any intermediate products?

- *Yes. Process all the data through singleframe processing, then all the data through calibration, all the data through coaddition etc… rather then region by region all the way to the end. When you get to the end of a stage and validate it, you can delete the intermediates. The* ***per-visit intermediate exposures*** *in particular take space. These include:*

DatasetTypeName	producedInStage	Can be deleted...
icExp	1	After single-frame processing is validated
postISRCCD	1	After single-frame processing is validated
deepCoadd_directWarp	3	After coadds are validated
deepCoadd_psfMatchedWarp	3	After coadds are validated
goodSeeingDiff_templateExp	4	After DIA outputs are validated
goodSeeingDiff_matchedExp	4	After DIA outputs are validated
goodSeeingDiff_differenceTempExp	4	After DIA outputs are validated

Are there any upcoming LSST pipelines developments that those of us working on precursor data sets like DECam and HSC should particularly be on the lookout for?

- *There’s a memory leak in the PiffPsfs that cause trouble when you coadds 100s of visits deep.*
- *QuantumBackedButler (QBB) has landed (w_2023_31) and is being tested now.*

Have you (or any other groups you know of) done multi-instrument forced photometry on precursor data sets using the LSST pipelines?

- *Talk to Raphael Shirley about his joint processing of HSC and VISTA data with the pipelines. He treated the two surveys as different bands of the same survey, and achieved good results.*
- *If you just have a list of ra/decls, check out ForcedPhotCcdFromDataFrameTask*

What resources do you recommend for someone learning to run the LSST pipelines on precursor data?

- *The starting point is the getting started guide on https://pipelines.lsst.io/v/weekly It's a tiny amount of data, just so you can get from start to finish in a day and know where you're going with lots of data.*
- *If you're ready to ingest your data into a Gen3 butler take a look at **Lee's guide.***
- *Then it's off to community.lsst.org for help*

I highly recommend watching Jim's Tour of Middleware talk a the 2022 Pipelines Bootcamp: https://confluence.lsstcorp.org/display/DM/DM+Pipelines+Bootcamp+2022

How actively are DECam calibration procedures/algorithms within the LSST pipelines being developed/upgraded/changed?

What is the status of the LSST pipelines as far as run-time optimization? Is that currently a point of emphasis, or are the LSST pipelines already viewed as meeting their formal requirements in this regard?
Same question but for memory usage rather than run-time.

- *Early on we optimized the most expensive kernels on then-hardware (warping and convolution especially). We've kept an eye on scalability (e.g. swapped jointcal out for GBDES. Working on cell-based coadds). Now hardware/DF We've started profiling the higher level Task code and pipelines.*

Possibly likely question from the audience: have the LSST pipelines been adapted for [fill in the blank] facility that I use/manage/operate yet?