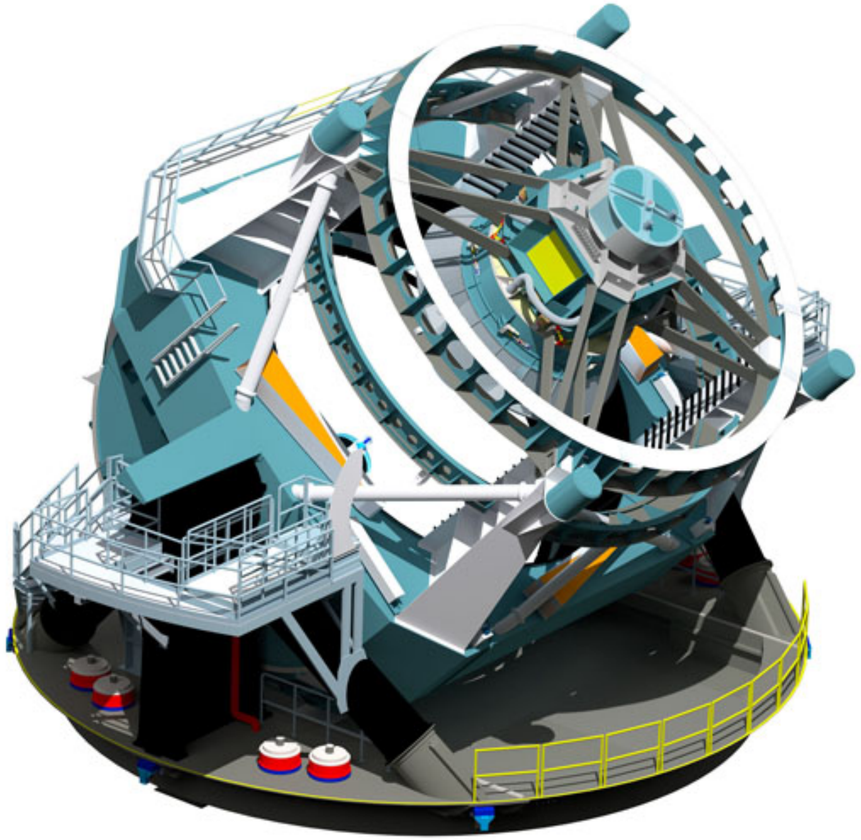
LSST and DESC
Data Access and Analysis

Michael Wood-Vasey
University of Pittsburgh

LSST DESC
Dark Energy Science Collaboration

LSST in a Nutshell

The LSST is an integrated survey system designed to conduct a decade-long, ‘wide-fast-deep’ time-domain survey of the optical sky.



Key Characteristics

- 8.4 m primary mirror (6.5 m effective)
- Wide 9.6 deg^2 field-of-view
- High étendue (AΩ) of 319 m^2 deg^2
- 3.2 Gpix camera, 6 filters: ugrizy
- Fully automated data processing system
- 20 TB data and up to 10 million alerts every night
- Interactive data access and analysis environment

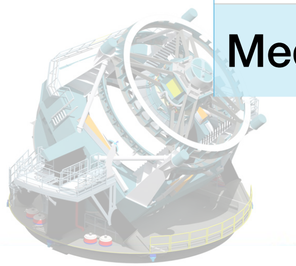
2022 - 2032

LSST will enable a wide variety of complementary scientific investigations

High Level Requirements

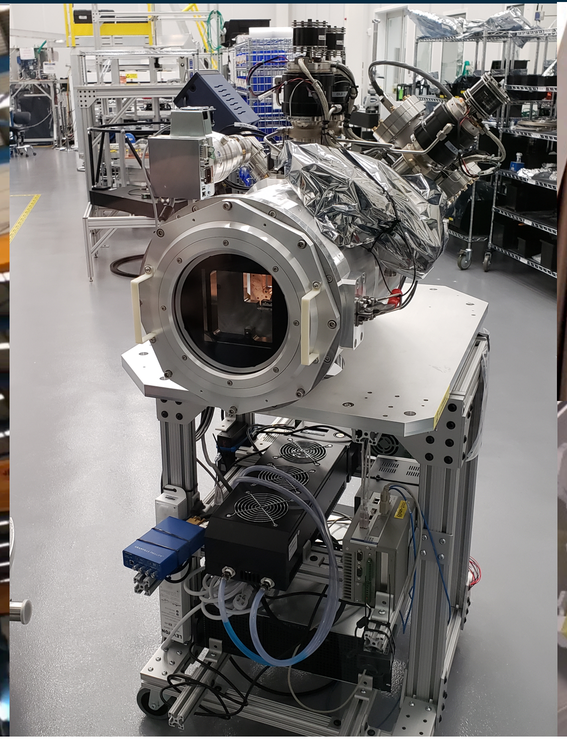
Requirements determined by constraints from the four science drivers

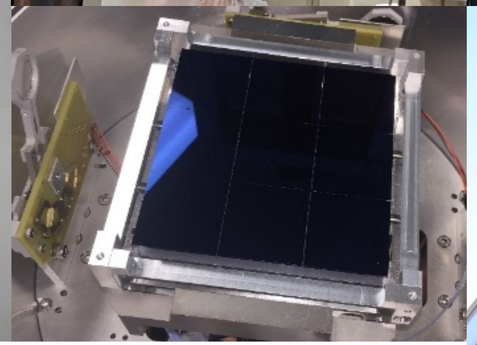
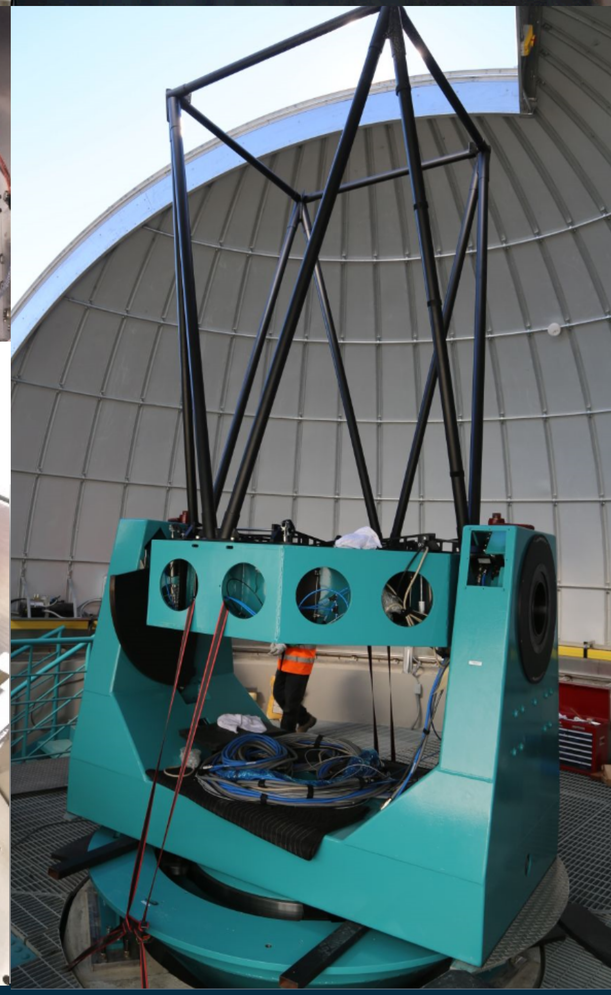
Survey Property	Specification (Design Value)
Survey Area & Median Number of Visits/Pointing	18,000 deg^2 with 825 visits
Filter set	6 filters (ugrizy) from 320 - 1050 nm
Single (Alternate) visit	2x15s (1x30s) exposures
Single Visit Depth (SNR > 5)	u=23.5, g=24.8, r=24.4, i=23.3, y=22.1
Astrometry (single visit)	10 mas relative, 50 mas absolute
Photometry	0.5% relative, 2% absolute
Proper Motion	0.2 mas/yr @ r = 20.5, 1.0 mas/yr @ r = 24.0
Median delivered image quality (seeing)	~0.7 arcsec FWHM



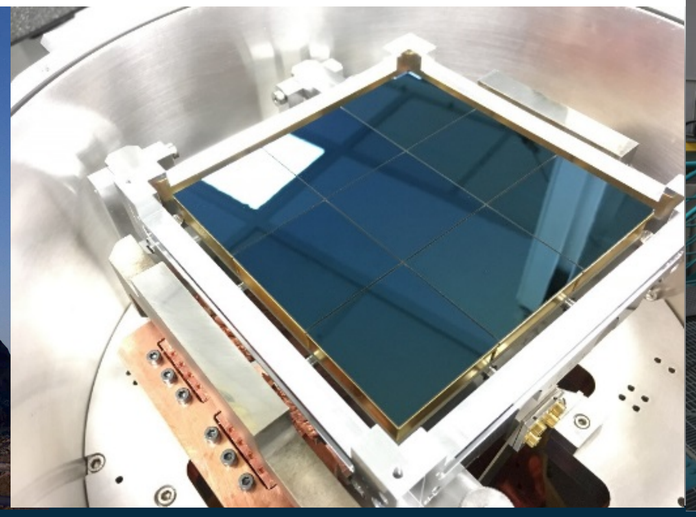
…. also drive the design of the data processing system

Ivezić et. el (2008)

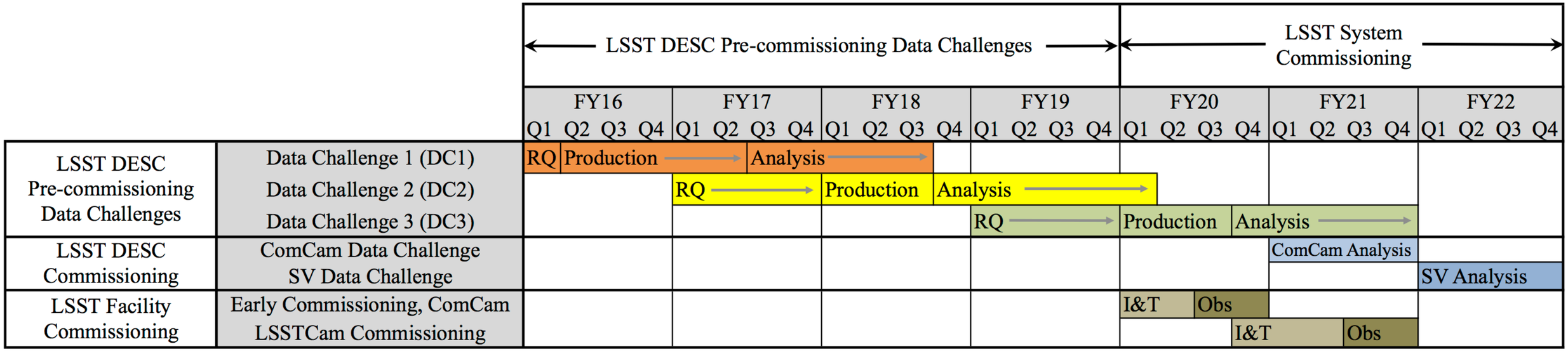
LSST Status Update

Dark Energy Science Collaboration

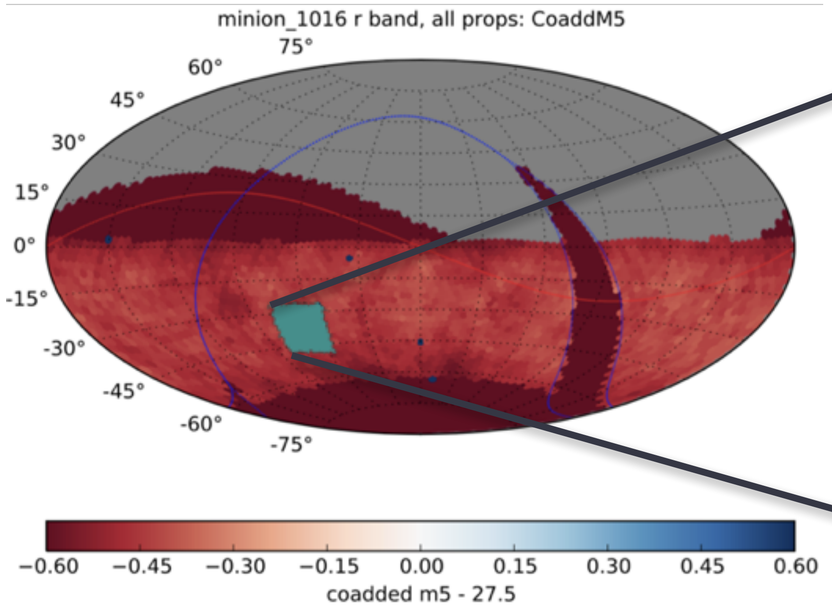
- 200 Full Members+ 600 interested Members
- https://lsstdesc.org
- Focused on Dark Energy Science with LSST
- DESC Science Requirements Document https://arxiv.org/pdf/1809.01669.pdf
- DESC Science Roadmap https://lsstdesc.org/sites/default/files/DESC_SRM_V1_4.pdf

- Data Challenges to prepare for LSST
- DC1 (2016-2018)
- DC2 (2017-2020)
- DC3 (2019-2021)
- LSST ComCam (2021)
- LSST Science Verification (2022)
- LSST Full Operational Data 2023. DR1 2023 - 6 months, DR2 2024 - 1 year.)

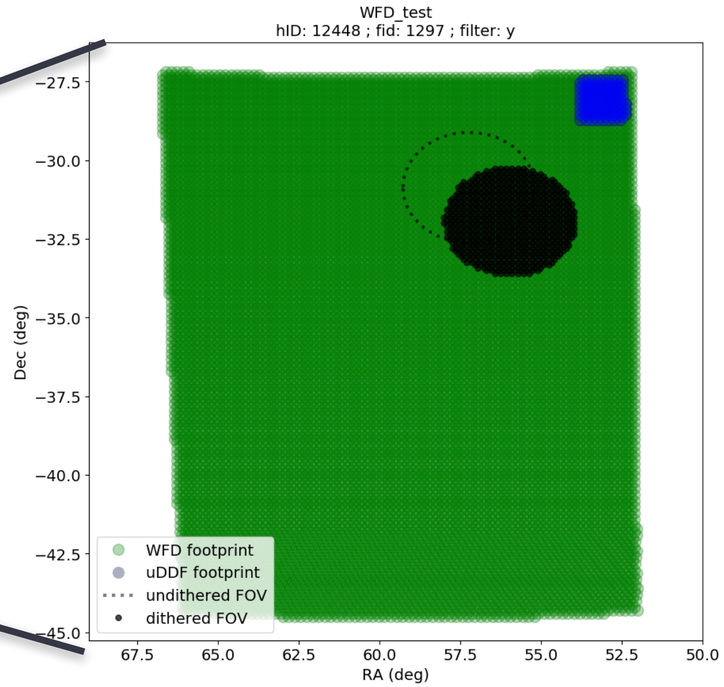


DESC Data Challenge 2

- Overall: 75,000 visits in DC2 = ~7.5M sensor visits
  - Similar in scale to planned LSSTCam Science Validation survey
  - Long-lasting legacy value, e.g., possible DC3 reprocessing activities
- Major supercomputing and human resources required!



DC2 patch (from DESCQA, Mao et al. 2018)

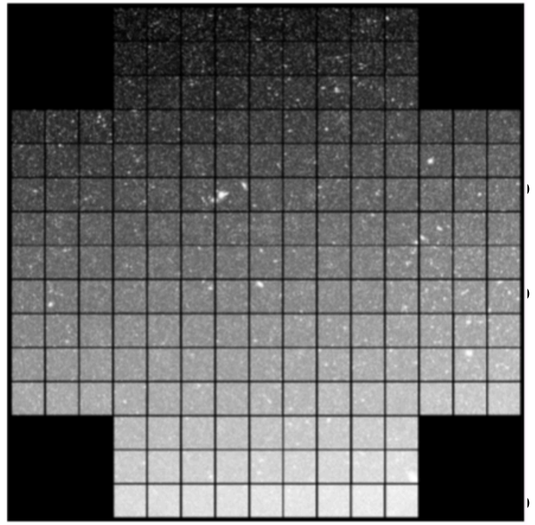


WFD and DDF dithers, Run 2, Image Credit: H. Awan

DESC Data Challenge 2

- Full end-to-end simulation and processing to enable testing and development of
  - Analysis and processing pipelines
  - Mitigation strategies for systematic effects
    - E.g., sensor defects, blending, atmospheric effects,
  - Data access strategies
- Design
  - Extragalactic sky: 5,000 sq deg input catalog
  - Static sky (WL, CL, LSS, PZ) with images:
    - 300 sq deg “main survey” area
    - 10 years ugrizy Wide-Fast-Deep (WFD) cadence
  - Time domain (SN, SL) and deep analyses:
    - 1 sq deg “Deep Drilling Field (DDF)” embedded in corner of main survey
    - 10 years ugrizy WFD + DDF visits
- Start with 3-year simulated survey, which deep enough for useful tests of many science cases!

Image Simulation Properties



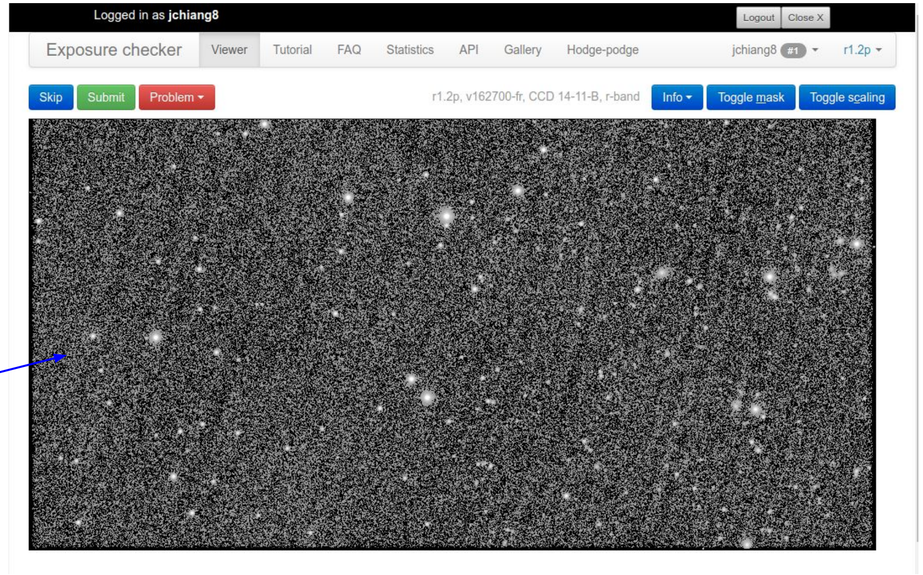
Simulated LSST focal plane

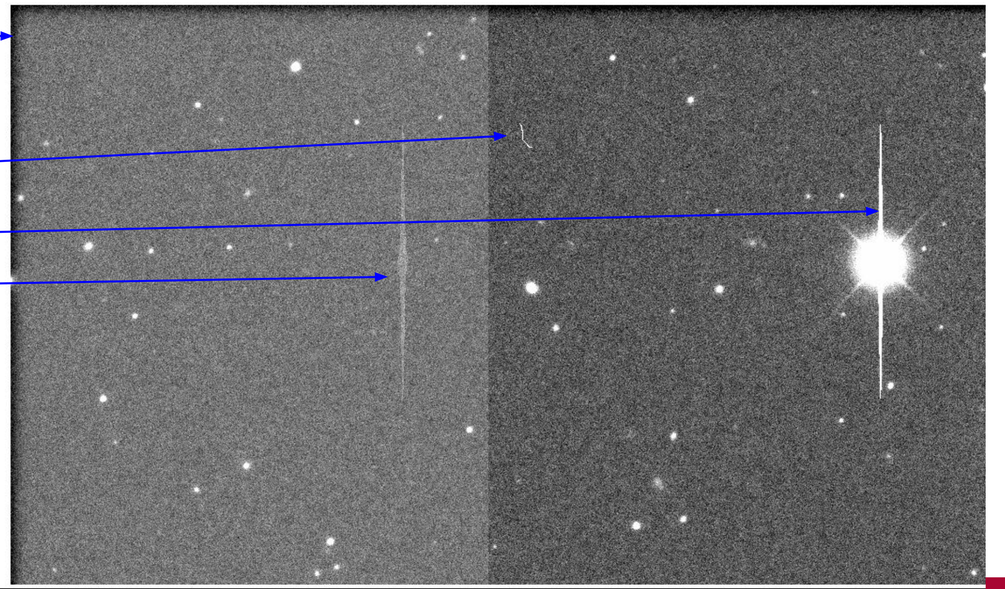
- Objects
  - DC2 galaxies (clustering, WL)
  - AGNs
  - Stars from CatSim
  - Sprinkled time-varying objects (lensed AGNs, SNe)

PSF modeled with atmospheric effects and optical distortions

Sensor effects:
  - Treerings
  - Brighter-fatter
  - Edge rolloff

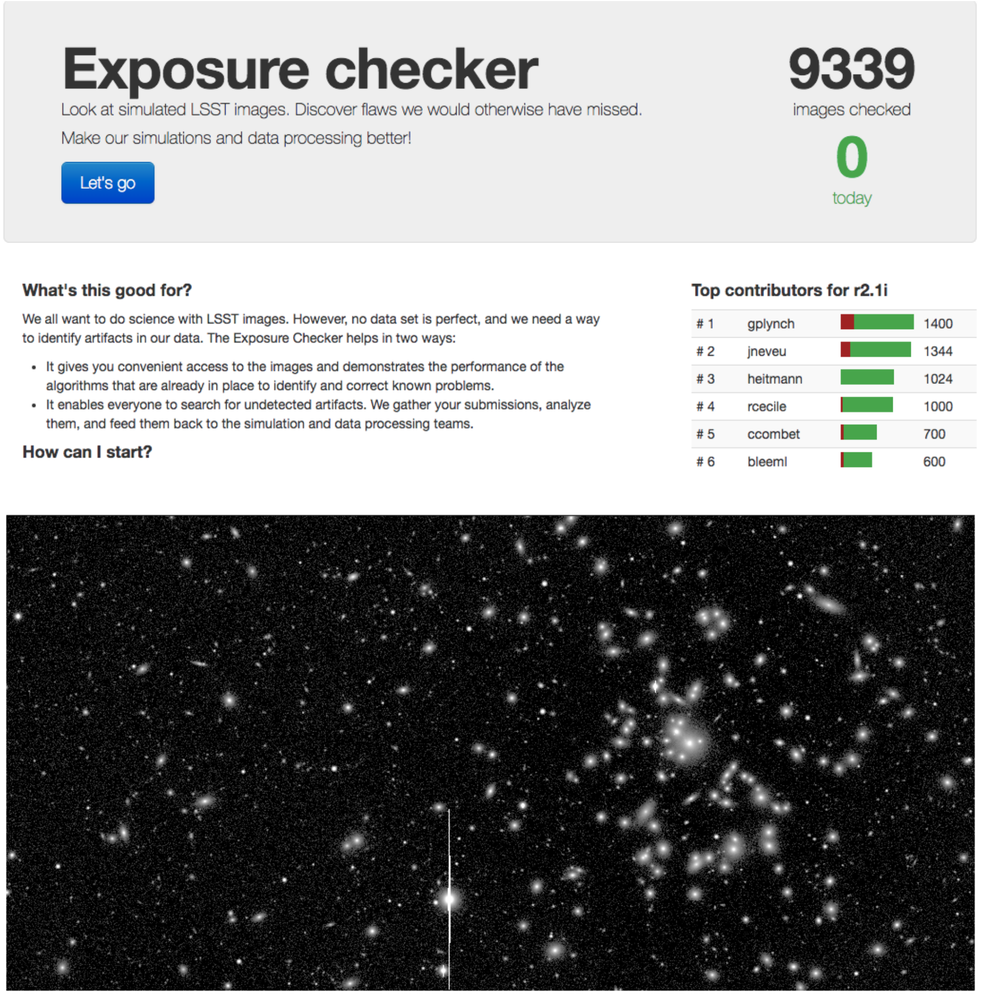
Instrument signature features
  - Vignetting
  - Cosmic Rays
  - Saturation and bleed trails
  - Crosstalk
  - Bright/dark defects
  - Charge transfer inefficiency





Adapted DES Exposure Checker for DC2

- Most common issues with background estimation in crowded regions and spurious cosmic ray masking (not major issues for DC2 science)
- Useful feedback on the tools and the data
- Excellent for quick diagnosis in early image simulation productions.

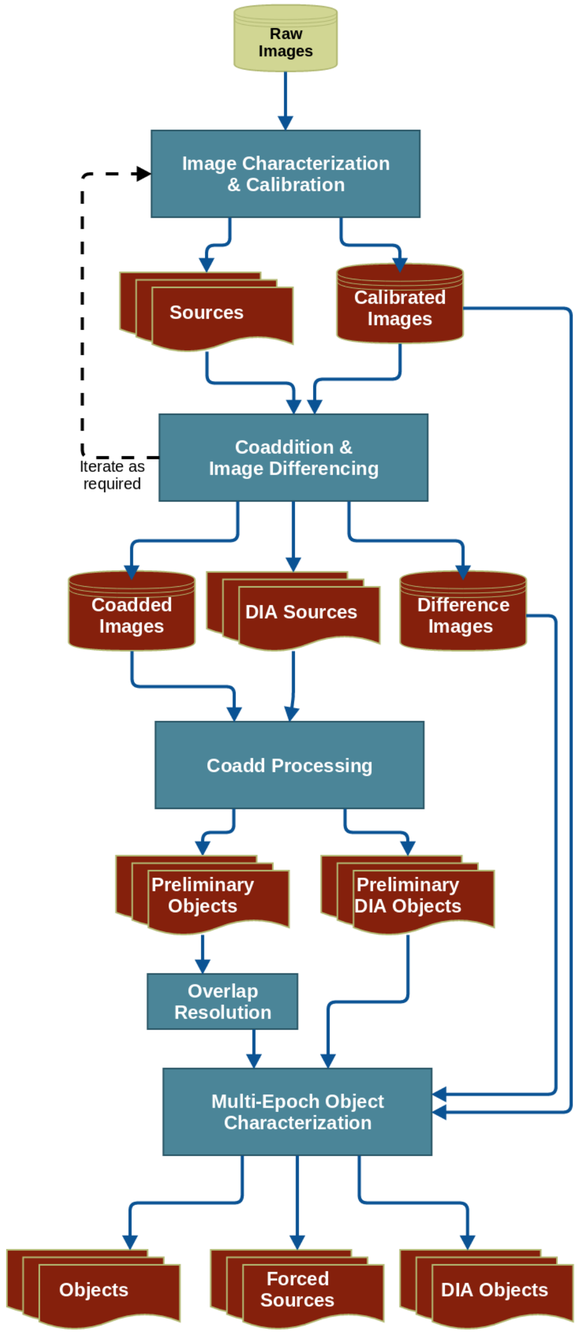


(Slide courtesy of Alex Drlica-Wagner)

Data Release Data Products

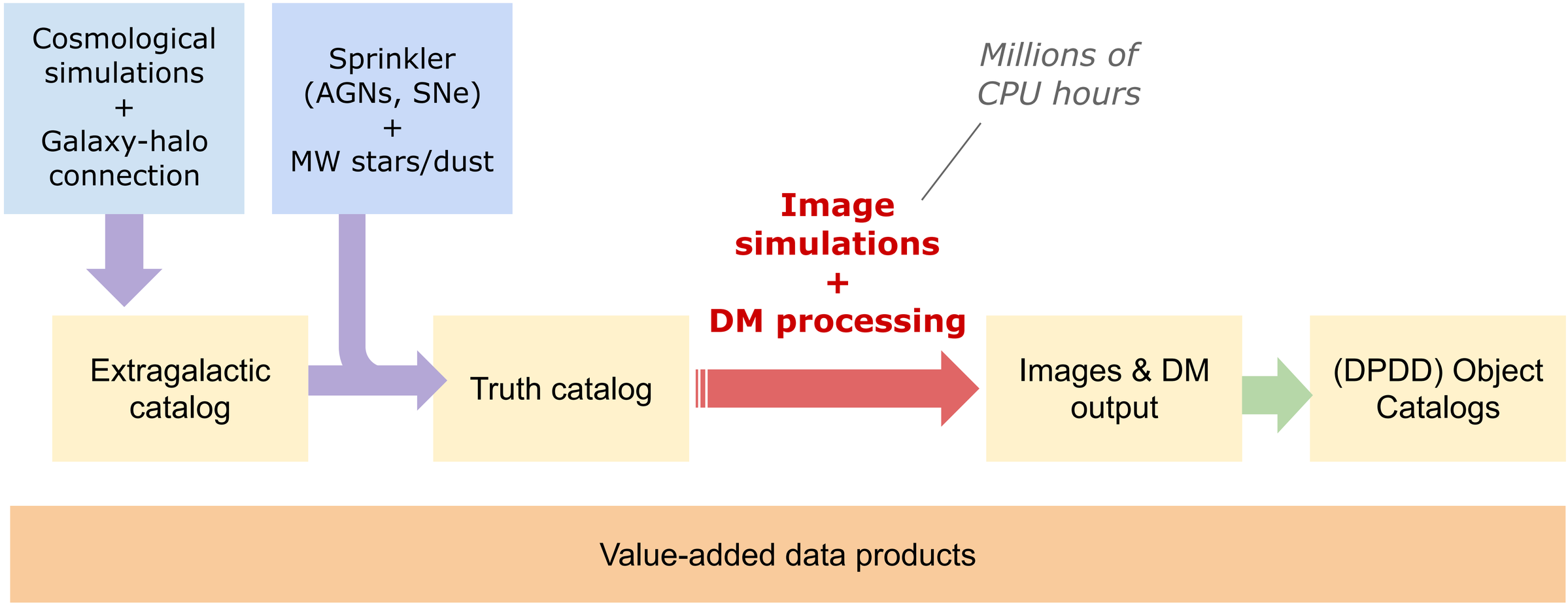
LSST will produce a ‘science-ready’ database of measurements can be analysed without the need for complex image processing

Images	**Single-Visit**	fully processed 2 x 15s exposures processed visit images (PVIs)
	Co-add	short-period (yearly & full survey) best seeing & deepest (unless the same) one per filter (ugrizy), and ‘multi-color’ transient-free template images
Catalogs	**Source** **(single-epoch)**	*Detections in single-visit images:* positions, aperture fluxes, point source fluxes, de-blending results
	Forced Source **(single-epoch)**	*In all single-visit images for all sources:* point-source fluxes, forced photometry
	Object **(associations of sources representing astrophysical phenomena)**	*All sources as well as detections in co-added images:* positions, aperture & point-source fluxes, de-blending results, model fits seeing-independent *galaxy* colours



LSST Data Products Definition Document

DC2 Key Data Products



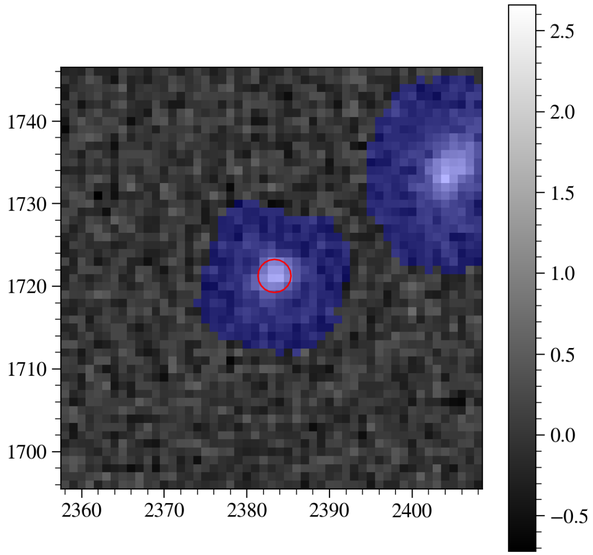
Overview of DC2 Data Products

- CosmoSim team prepares extragalactic catalog
- ImSim team simulations images at NERSC and Argonne
- DM DC2 team processes images at IN2P3 and NERSC
- Data Access Task Force prepares data products for use by DESC members

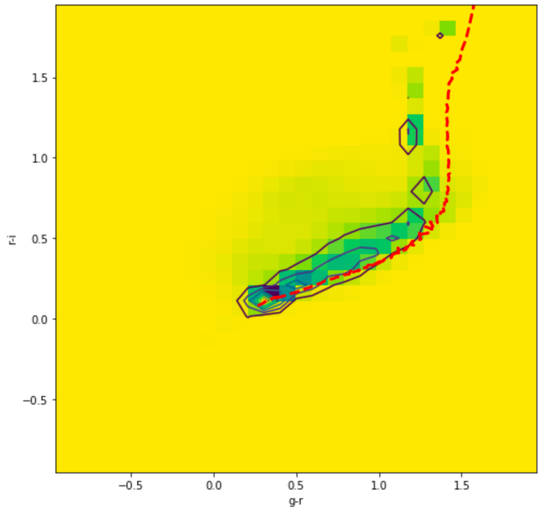
Name	Description	Access
Extragalactic catalogs	Mock catalogs from cosmological simulations, including galaxies, shears,...	GCR
Truth catalogs	Truth information about all objects in the simulated images, including stars, AGNs, SNe,...	GCR
DM science pipeline outputs	Range of data products from the DM pipeline, including raw images, calibrated exposures, ...	Butler
DPDD(-like) object catalogs	Merge static-sky coadd catalogs with forced-position photometry at fixed RA, Dec across bands	GCR, pandas, spark

DESC Data Access

- MWV and Dominique Boutigny Co-Lead Data Access Task Force
- Provide data access for DESC
  - Different data access methods are being explored:
    - GCR: Generic catalog reader
    - Qserv (project): Inject data at IN2P3 (SLAC instance)
    - PostgreSQL mirroring HSC set-up, spark
    - hdf5/parquet files



Postage stamp for set RA, Dec, Image Credit: Michael Wood-Vasey



Color-color plot, Image Credit: Joanne Bogart

Example Jupyter Notebooks on DC2

- https://github.com/LSSTDESC/DC2-analysis/blob/rendered/tutorials/object_pandas_stellar_locus.nbconvert.ipynb
- https://github.com/LSSTDESC/DC2-analysis/blob/rendered/tutorials/object_spark_1_intro.nbconvert.ipynb
- https://github.com/LSSTDESC/DC2-production/blob/master/Notebooks/object_catalog_performance_dask_parquet.ipynb

DESC DC2 Data Volumes

		Run 1.2 25 sq. deg - 2,000 visits		Run 2.1 300 sq. deg - 75,000 visits	
Type	Product	Size	# Files	Size	# Files
Images	Raw	5 TB	300,000	125 TB	7,500,000
Images	Processed	30 TB	1,000,000	750 TB	21,500,000
Tables	Object	4 GB	1	100 GB	1
Tables	Source	30 GB	2,000	1,125 GB	75,000
Tables	Forced Source	10 GB	2,000	500 GB	75,000

I quote images volume in TB and catalog in GB both because those are they natural scales
But also because image volume determines disk storage, which is a TB-scale question
While catalog volume determines memory size, which is a GB-scale question.

DESC DC2 Data Access

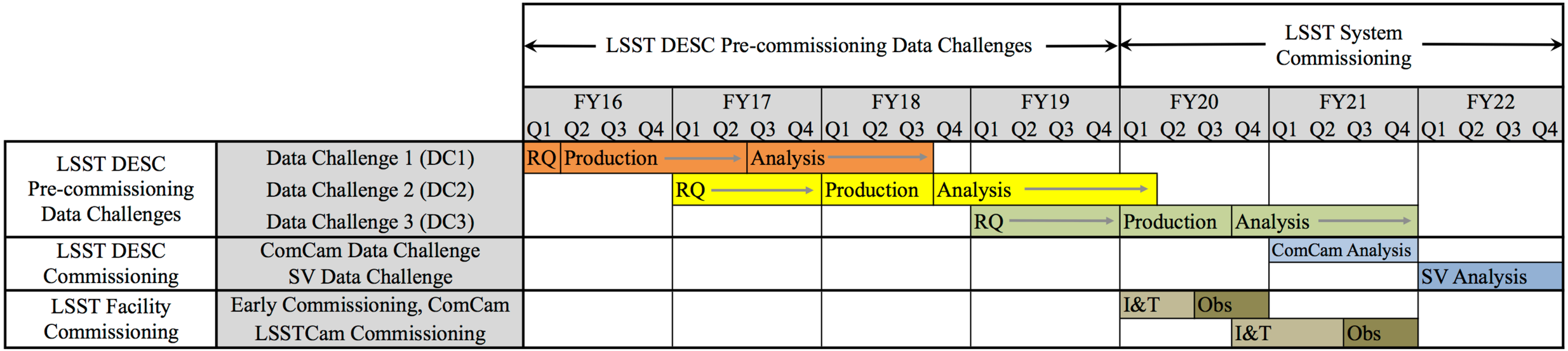
- Centered at NERSC
- Plan to extend to IN2P3.
  - Once that’s done, likely easy extend to additional DESC partners.
- Jupyter Notebooks main user and exploration interaction
- Intensive pipelines, such as galaxy-galaxy, galaxy-shear, shear-shear correlations base on more programatic access.

DESC DC2 Data Volumes

		Run 1.2 25 sq. deg - 2,000 visits		Run 2.1 300 sq. deg - 75,000 visits	
Type	Product	Size	# Files	Size	# Files
Images	Raw	5 TB	300,000	125 TB	7,500,000
Images	Processed	30 TB	1,000,000	750 TB	21,500,000
Tables	Object	4 GB	1	100 GB	1
Tables	Source	30 GB	2,000	1,125 GB	75,000
Tables	Forced Source	10 GB	2,000	500 GB	75,000

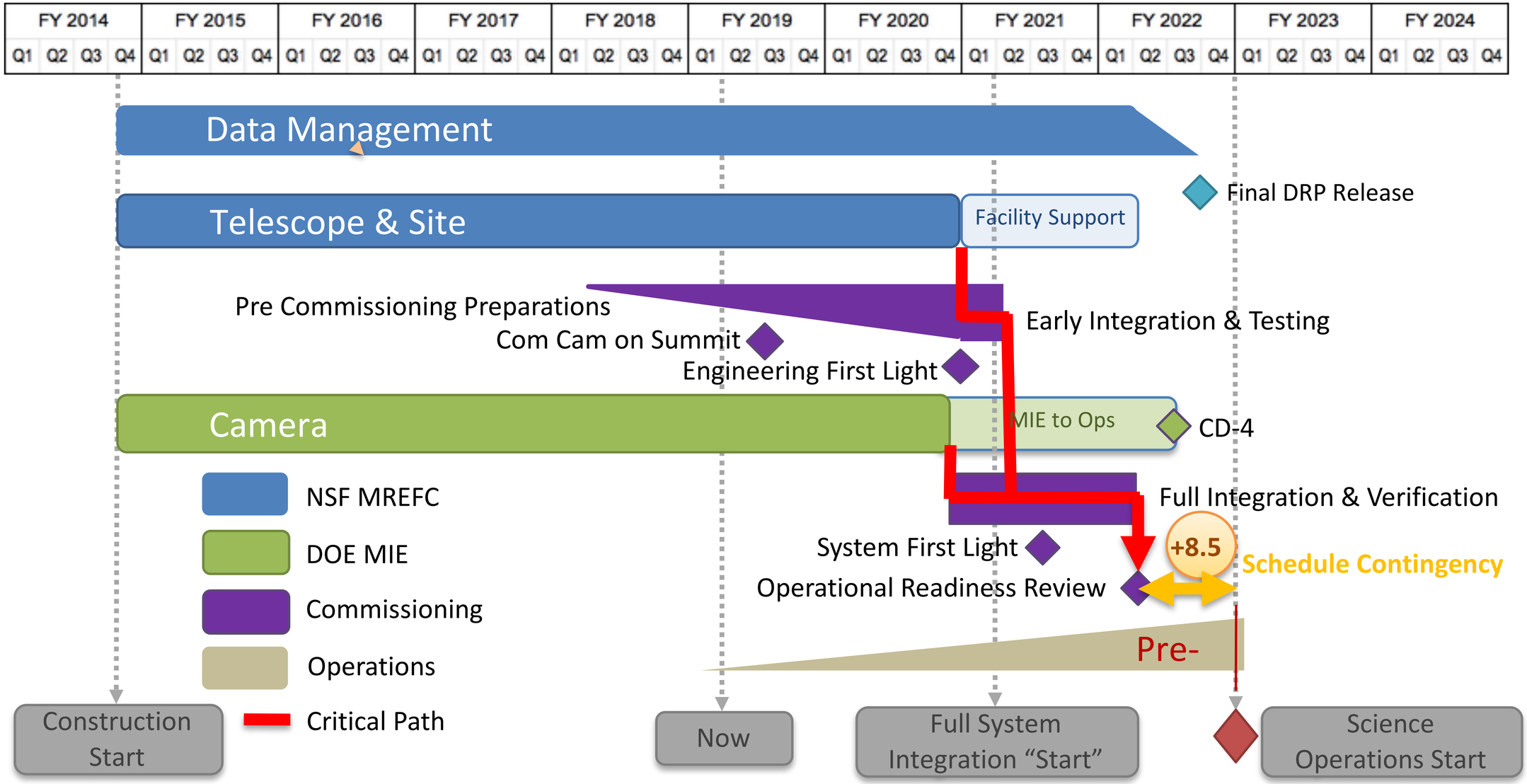
I quote images volume in TB and catalog in GB both because those are they natural scales
But also because image volume determines disk storage, which is a TB-scale question
While catalog volume determines memory size, which is a GB-scale question.

DESC Data Challenge Timeline Toward LSST Start



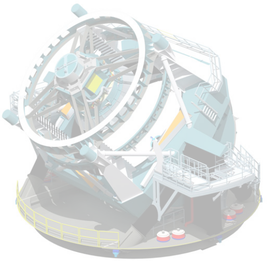
Extra Slides

LSST Schedule



LSST Key Dates

Milestone	Date
Start of On-Sky Data from Auxiliary Telescope	Aug 2019
Start of On-Sky & Calibration Data with ComCam	Oct. 2020
Sustained Observing with ComCam	Feb. 2021
Start of On-Sky & Calibration Data with LSSTCam	July 2021
Sustained Observing with LSSTCam	Oct. 2021
Start of Science Verification Surveys	Dec. 2021
Operations Readiness Review	Mar 2022



Data Product Categories

The LSST data products are organized into three main categories.

Prompt Data Products

Real Time Difference Image Analysis (DIA)

- A stream of ~10 million time-domain events per night (Alerts), transmitted to event distribution networks within 60s of camera readout.
- Images, Object and Source catalogs derived from DIA, and an orbit catalog for ~6 million Solar System bodies within 24h.
- Enables discovery and rapid follow-up of time domain events

Data Release Data Products

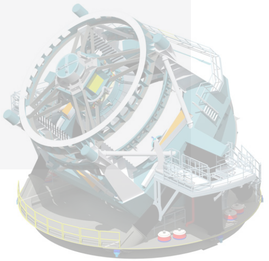
Reduced single-epoch & deep co-added images, catalogs, reprocessed DIA products

- Catalogs of ~37 billion objects (20 billion galaxies, 17 billion stars), ~7 trillion sources and ~30 trillion forced source measurements.
- 11 Data Releases, produced ~annually over 10 years of operation
- Accessible via the LSST Science Platform & LSST Data Access Centers.

User Generated Data Products

User-produced derived, added-value data products

- Deep KBO/NEO, variable star classifications, shear maps, etc …
- Enabled by services & computing resources at the LSST DACs and via the LSST Science Platform (LSP).
- 10% of LSST computing resources will be allocated for User Generated data product storage & processing.



LSST Data Product Categories & DM Data Products & LSST Key Numbers

Nomenclature

Processed Visit Image: LSST image from a single visit with the background subtracted.

Source: single detection of an astrophysical object in an image,

Forced Source: measurement a source, with one or more parameters held fixed

Object: association of Sources by coordinate, an astronomical object (star, galaxy)

Source Association: association of sources at different epochs, or passbands, with an Object

Prompt Processing: processing of the nightly stream of raw images, including Difference Imaging Analysis, Alert Production. This processing generates ***Prompt Data Products***.

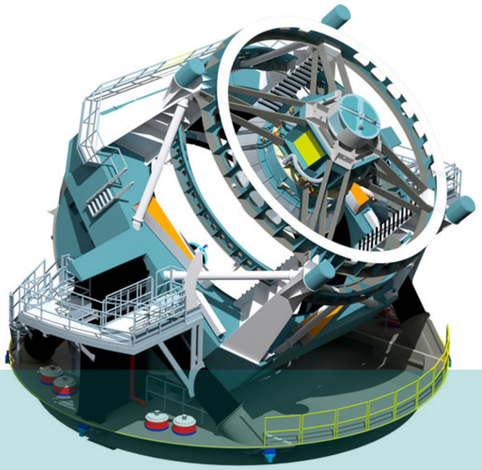
Alert Production (AP): component of Prompt Processing that processes and calibrates images to identify DIASources and DIAObject. This processing generates ***Alerts***.

Data Release Production (DRP): A (re)processing all of the accumulated LSST images.This processing generated the ***Data Release data products***.

LSST Data Management System

Raw Data: 20TB/night

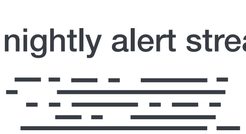
Sequential 30s images covering the entire visible sky every few days

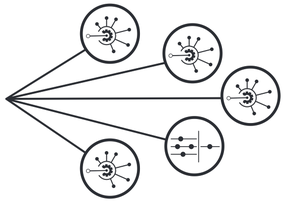


Prompt Data Products

Alerts: up to 10 million per night

60s via nightly alert streams





LSST Alert Filtering Service

Community Brokers

Raw, calibrated, and difference images and their source and object catalogs

Solar System Objects: ~ 6 million

24h via Prompt Products Database

Data Release Data Products

Final 10yr Data Release:

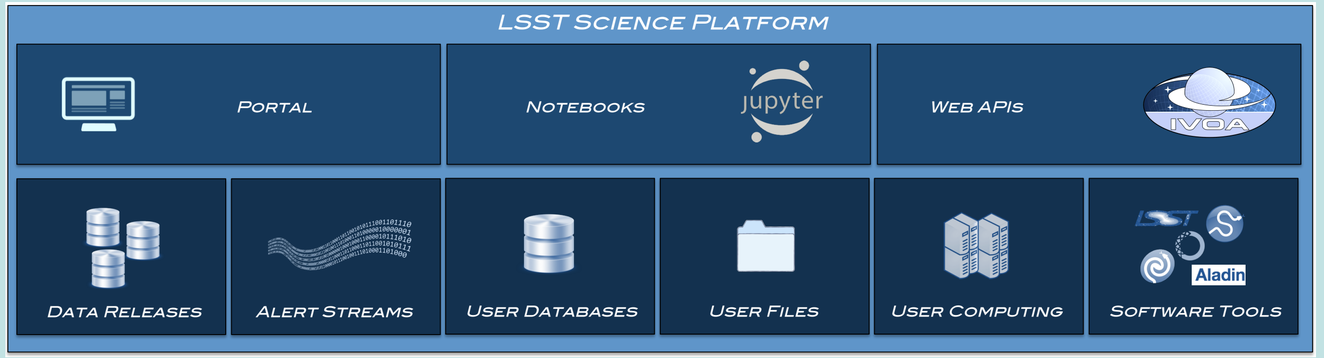
- Images: 5.5 million x 3.2 Gpx
- Catalog: 15PB, 37 billion objects

via Data Releases

LSST & Independent Data Access Centers

LSST Science Platform

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



The LSST Science Platform

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

Portal Aspect

exploratory analysis and visualization of the LSST archive

JupyterLab Notebook Aspect

in-depth ‘next-to-the-data’ analysis & creation of added-value data products

Web API Aspect

remote access to the LSST archive via industry-standard APIs