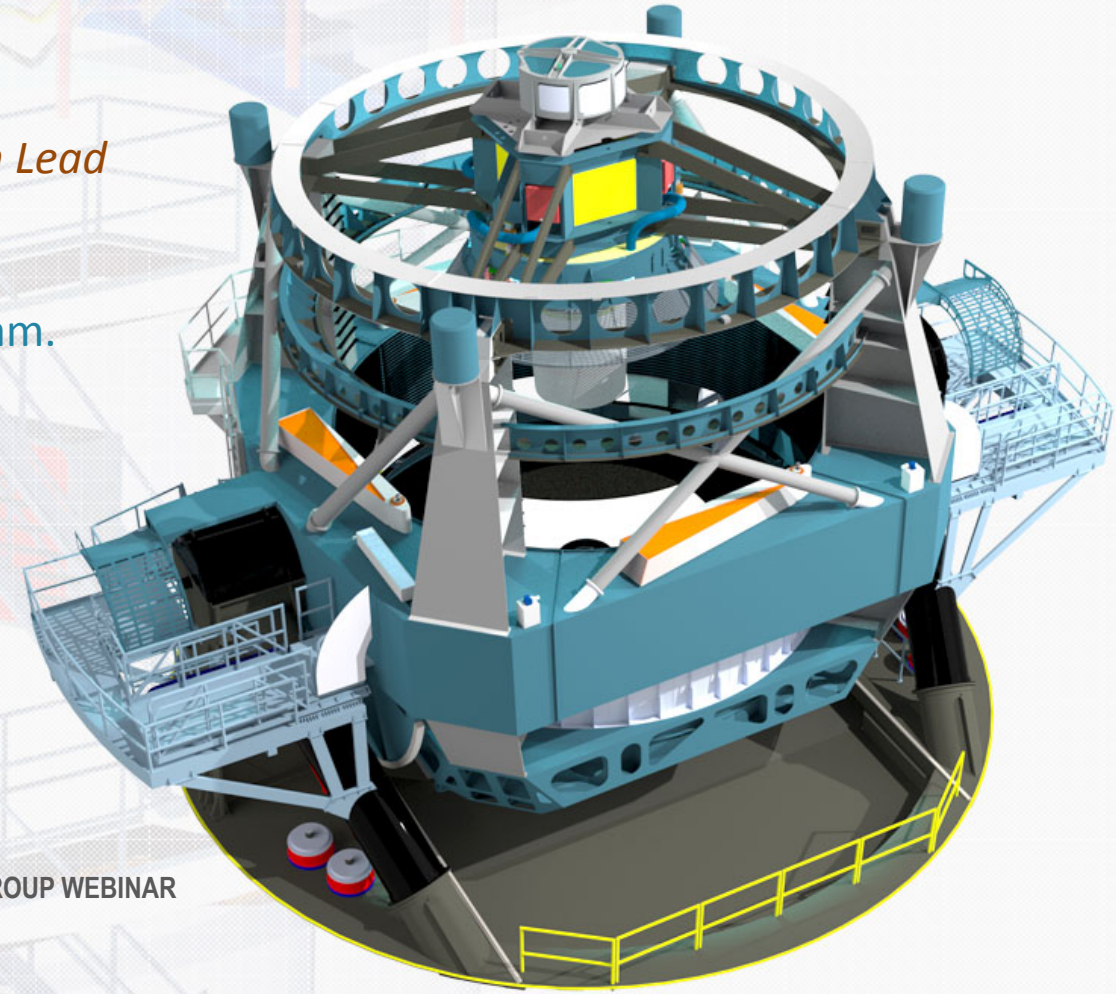
Getting ready for LSST: Services for Access, Exploration, and Analysis of LSST Data

Mario Juric,
University of Washington
LSST Data Management Subsystem Lead

for the LSST Data Management Team.



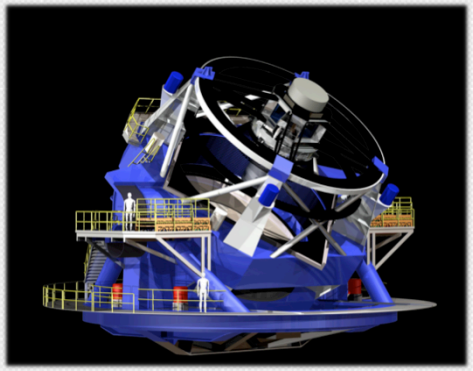
BRAZILIAN LSST PARTICIPATION GROUP WEBINAR
December 7th, 2016

The LSST: A Database of the Sky

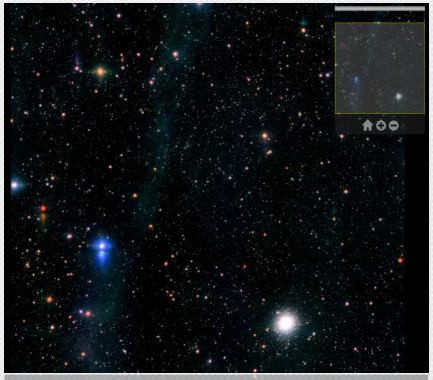
- **A wide (half the sky), deep (24.5/27.5 mag), fast (image the sky once every 3 days) survey telescope. Beginning in 2022, it will repeatedly image the sky for 10 years.**

- The LSST is an integrated survey system. The Observatory, Telescope, Camera and Data Management system are all built to support the LSST survey. There's no PI mode, proposals, or time.

- **The ultimate deliverable of LSST is not the telescope, nor the instruments; it is the fully reduced data. LSST is a *facility* that delivers *data products* and *data access and analysis services*.**



Telescope ➔



Images ➔

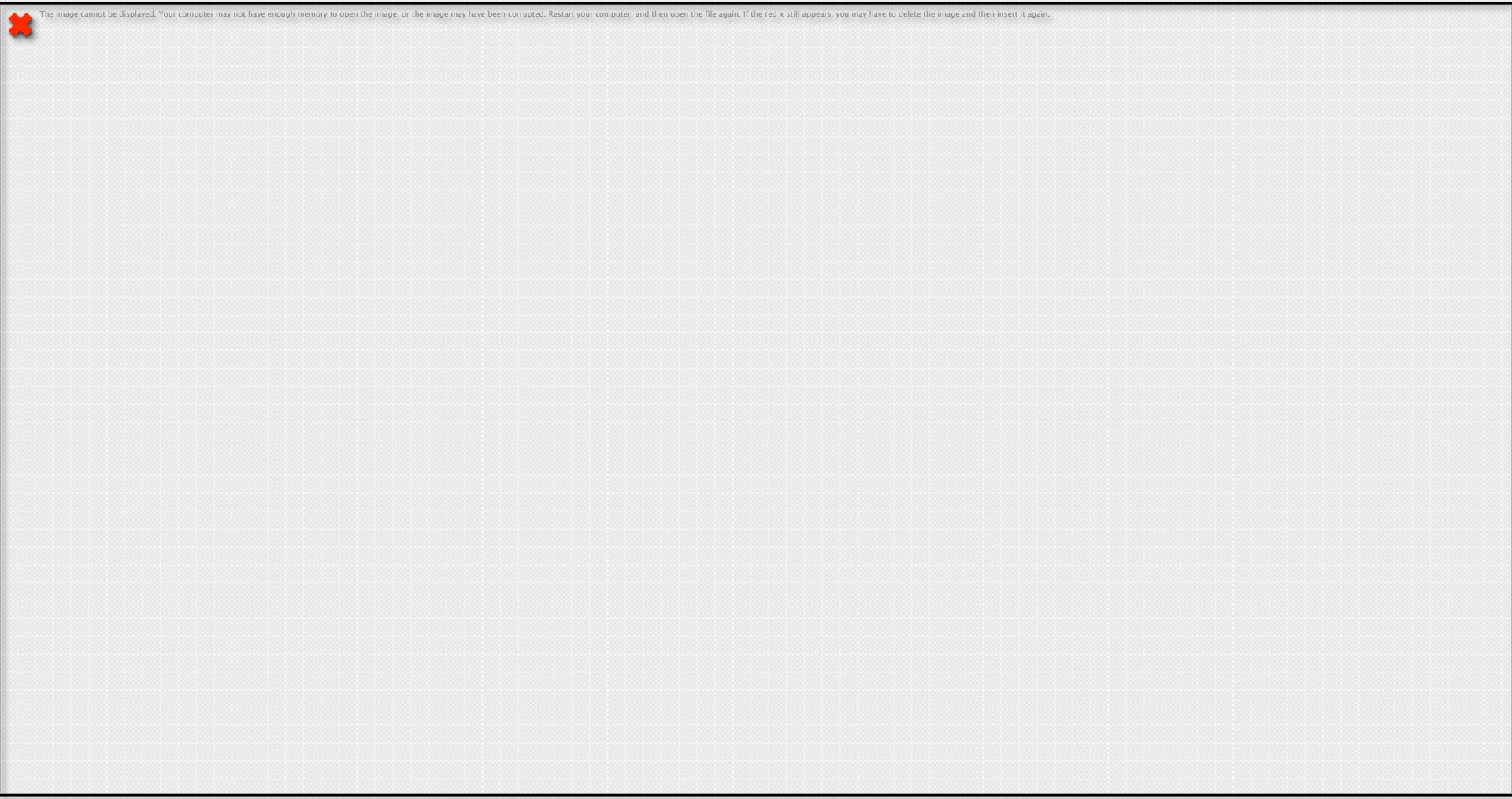
Table 4: Level 2 Catalog `Object` Table

Name	Type	Unit	Description
psRadecTai	double	time	Point source model: Time at which the object was at position `radec`.
psPm	float[2]	mas/yr	Point source model: Proper motion vector.
psParallax	float	mas	Point source model: Parallax.
psFlux	float[ugrizy]	nmgy	Point source model fluxes[58].
psCov	float[66]	various	Point-source model covariance matrix[59].
psLnL	float		Natural *log* likelihood of the observed data given the point source model.
bdRadec	double[2]	degrees	B+D model[60]: (α, δ) position of the object at time `radecTai` in each band.

Catalogs

Location: Cerro Pachon, Chile

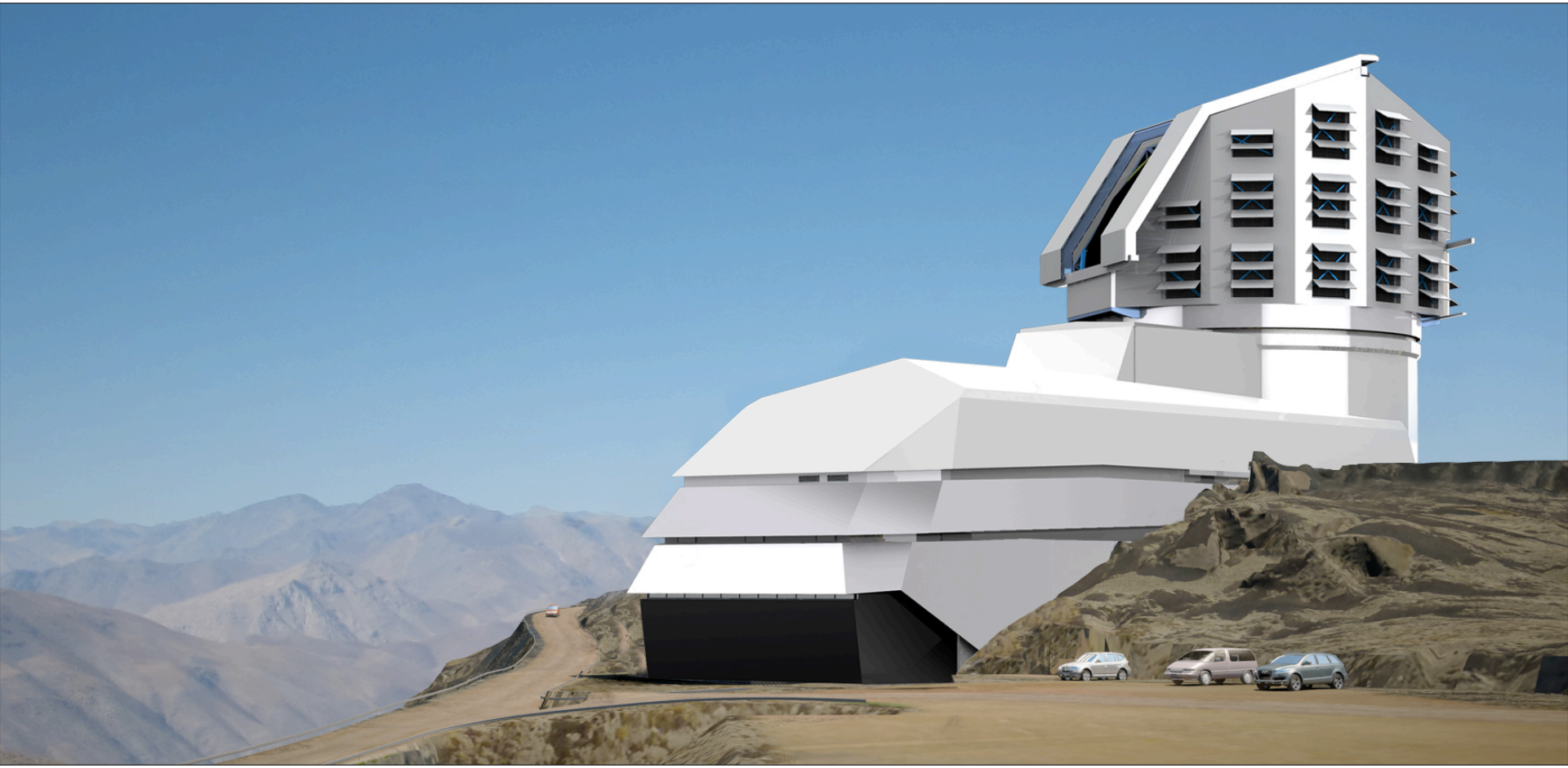
LSST Site (April 14th, 2015)



The Summit, yesterday.

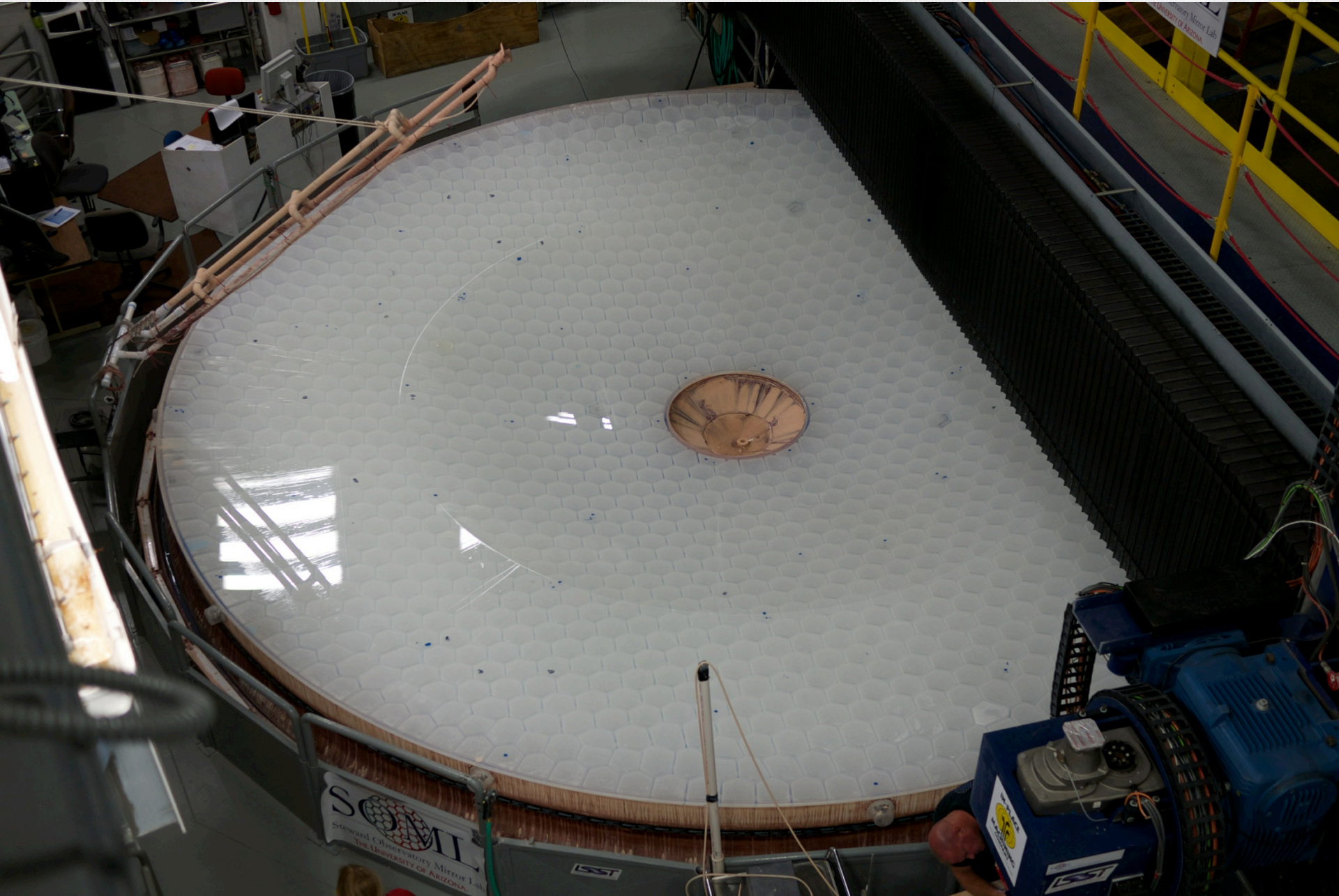
The Support Building and the Pier Taking Shape

LSST Observatory (cca. late ~2018)



We are ~2 years away from the observatory building being close to complete!

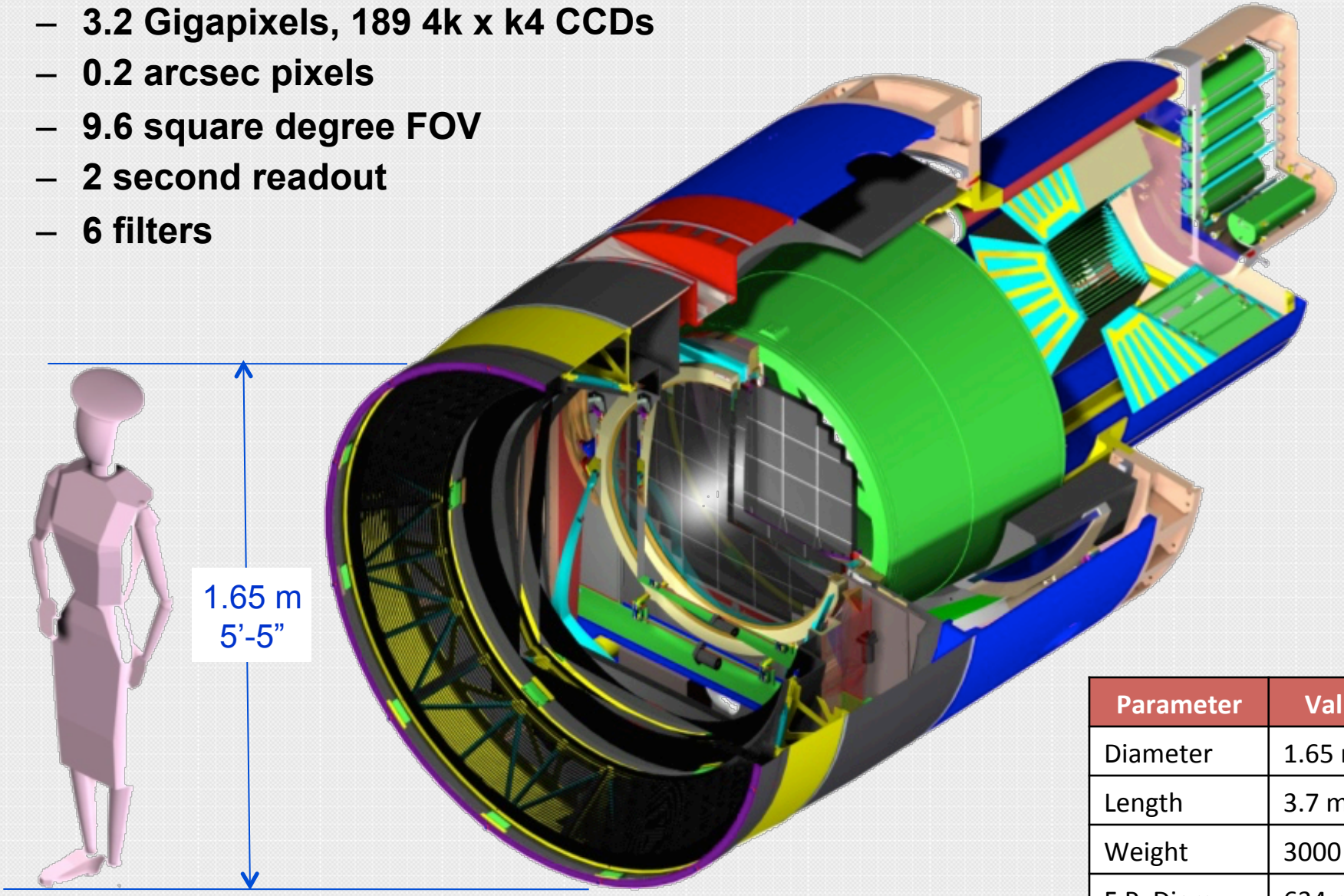
M1/M3 mirror: Done.



Mirror cell is being constructed

LSST Camera

- 3.2 Gigapixels, 189 4k x k4 CCDs
- 0.2 arcsec pixels
- 9.6 square degree FOV
- 2 second readout
- 6 filters



Parameter	Value
Diameter	1.65 m
Length	3.7 m
Weight	3000 kg
F.P. Diam	634 mm

LSST Operations: Sites and Data Flows

Satellite Processing Center
(CC-IN2P3, Lyon, France)

Data Release Production (50%)
French DAC

Archive Site
Archive Center
Alert Production
Data Release Production (50%)
EPO Infrastructure
Long-term Storage (copy 2)

Data Access Center
Data Access and User Services

Summit and Base Sites
Telescope and Camera
Data Acquisition
Crosstalk Correction
Long-term storage (copy 1)
Chilean Data Access Center

HQ Site
Science Operations
Observatory Management
Education and Public Outreach

LSST From a Scientist's Perspective

- A stream of ~10 million time-domain events per night, detected and transmitted to event distribution networks within 60 seconds of observation.
- A catalog of orbits for ~6 million bodies in the Solar System.

Level 1

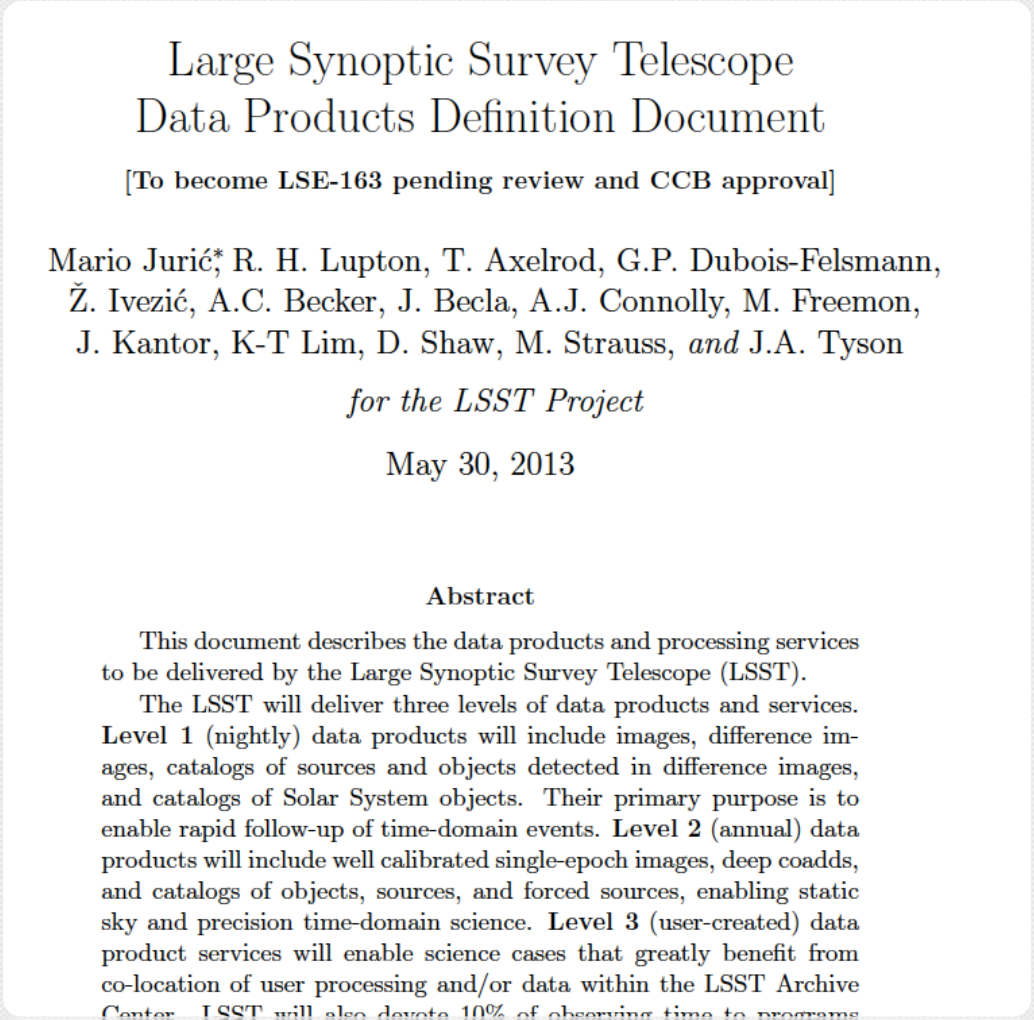
- A catalog of ~37 billion objects (20B galaxies, 17B stars), ~7 trillion observations ("sources"), and ~30 trillion measurements ("forced sources"), produced annually, accessible through online databases.
- Reduced single-epoch, deep co-added images.

Level 2

- Services and computing resources at the Data Access Centers enabling limited analysis, production, and federation of added value products.
- Web APIs enabling the use of remote analysis tools.
- Public LSST pipeline code for deeper insight into LSST data products.

Level 3

The Data Products Definition Document



Large Synoptic Survey Telescope
Data Products Definition Document

[To become LSE-163 pending review and CCB approval]

Mario Jurić*, R. H. Lupton, T. Axelrod, G.P. Dubois-Felsmann, Ž. Ivezić, A.C. Becker, J. Becla, A.J. Connolly, M. Freemon, J. Kantor, K-T Lim, D. Shaw, M. Strauss, *and* J.A. Tyson

for the LSST Project

May 30, 2013

Abstract

This document describes the data products and processing services to be delivered by the Large Synoptic Survey Telescope (LSST).

The LSST will deliver three levels of data products and services. **Level 1** (nightly) data products will include images, difference images, catalogs of sources and objects detected in difference images, and catalogs of Solar System objects. Their primary purpose is to enable rapid follow-up of time-domain events. **Level 2** (annual) data products will include well calibrated single-epoch images, deep coadds, and catalogs of objects, sources, and forced sources, enabling static sky and precision time-domain science. **Level 3** (user-created) data product services will enable science cases that greatly benefit from co-location of user processing and/or data within the LSST Archive Center. LSST will also devote 10% of observing time to programs

LSST Data Products Definition Document

A document giving a high-level description of LSST data products.

http://ls.st/dpdd

Level 1 Data Products: Section 4.

Level 2 Data Products: Section 5.

Level 3 Data Products: Section 6.

Special Programs DPs: Section 7.

Level 1:

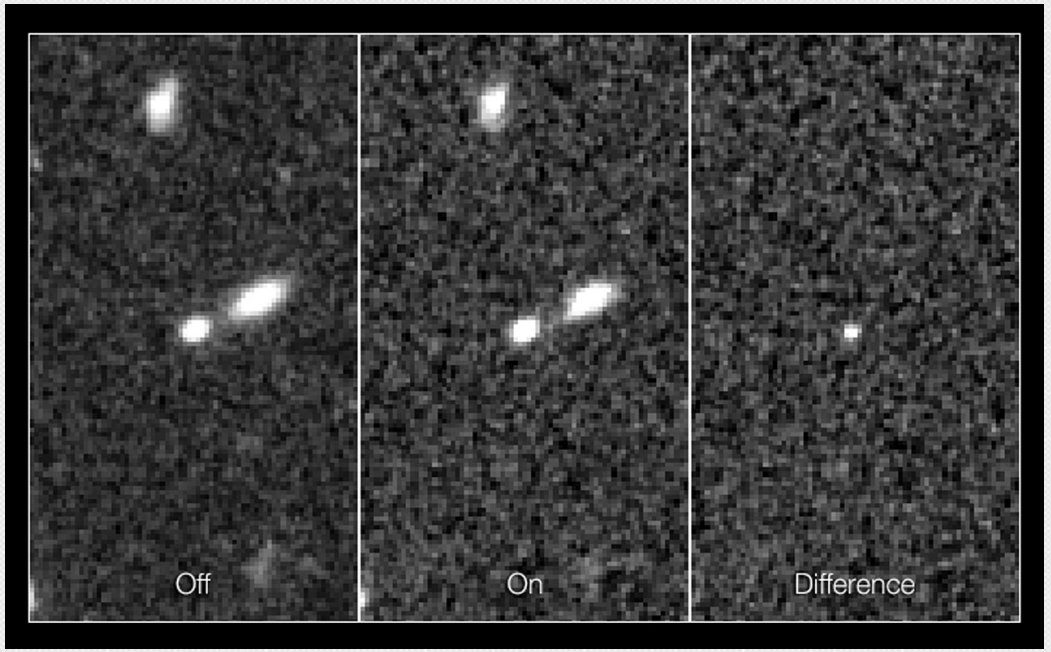
Enabling Discovery and Rapid Follow-up of Time Domain Events

Level 1 Data Products: Time Domain

- **Real-time image differencing as observing unfolds each night**
- Detection performed on image differenced against a deep template
- Measurement performed on the difference image and direct image
- Associated with pre-existing observations and stored in a database

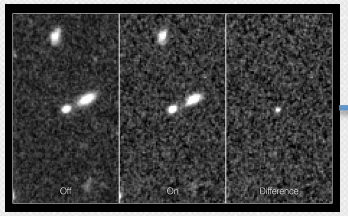
- For every source detected in a difference image, we will emit an "Event Alert" within 60 seconds of observation.

The primary use case is to enable real-time recognition and follow-up of transients of special interest.



CANDELS (http://www.spacetelescope.org/images/heic1306d/)

Level 1 Data Products and Flows



(5 sec)

Difference, detect and measure

(50 sec)

Level 1 Data Store (internal)

Alerts

(60 sec)

Level 1 SQL Database

(24 hrs)

Images

(24 hrs)

Solar System Object Catalog

(24 hrs)

Level 1: Time-Domain Event Alerts

- **Each alert will include the following:**
  - Alert and database ID: IDs uniquely identifying this alert.
  - The photometric, astrometric, and shape characterization of the detected source
  - 30x30 pixel (on average) cut-out of the difference image (FITS)
  - 30x30 pixel (on average) cut-out of the template image (FITS)
  - The time series (up to a year) of all previous detections of this source
  - Various summary statistics ("features") computed of the time series

- **The goal is to transmit nearly everything LSST knows about any given event, enabling downstream classification and decision making *without* the need to call back into LSST databases (thus introducing extra latency)**

- We expect a high rate of alerts, **approaching 10 million per night**.

LSST Alert Filtering Service

- Most end-users will **not** be interested in reception of the full stream, but **only a subset that matches their scientific interest** (e.g., potential SNe candidates, variable stars, or moving objects).

- To support selecting such subsets of alert candidates, **LSST will provide an alert filtering service**. This service will let astronomers create simple *filters* that limit which alerts are ultimately forwarded to them.

- These user defined filters will be possible to specify using an SQL-like declarative language, or short snippets of (likely Python) code (n.b. this is our current thinking, subject to change!).

Example of a User-Defined Filter (a sketch!)

```
# Keep only never-before-seen events within two
# effective radii of a galaxy. This is for illustration
# only; the exact methods/members/APIs may change.

def filter(alert):
    if len(alert.sources) > 1:
        return False
    nn = alert.diaobject.nearest_neighbors[0]
    if not nn.flags.GALAXY:
        return False
    return nn.dist < 2. * nn.Re
```

The user will subscribe to the alert stream by specifying a filtering function such as the one shown above. Once specified, only the alerts for which the function returns True will be forwarded to the user's VOEvent client.

Public VOEvent Brokers and Networks

- **We also anticipate that advanced, public, filtering services – VOEvent brokers – will be established by the community.**

- These may provide advanced functionality such as:
  - cross-correlation of LSST alerts with external catalogs and other alert streams,
  - classification engines,
  - more extensive annotation of alerts,
  - coordination of follow-up groups,
  - incorporation of other contextual information needed to decide on whether a transient is worth following up.

- **We are encouraging the community to self-organize and develop such alert brokers and networks.**
  - **US: ANTARES project led by NOAO**

Level 1: Solar System Objects

- **Solar System objects will be identified and linked together based on compatibility of their observed positions with motion around the Sun.**
  - Enhanced variant of MOPS algorithm; advanced prototype in hand.

- **Planning to:**
  - Identify and link observations of Solar System objects
  - Measure their orbital elements
  - Measure their photometric properties
  - *Expect to provide orbits for > 60% of all NEOs brighter than H=22*

- **Availability: within 24 hours of orbit determination**

Level 2:

Enabling Deep Sky and High-Precision Astrophysics

Level 2: Annual Data Releases

- **Well calibrated, consistently processed, catalogs and images**
  - Catalogs of objects, detections, detections in difference images, etc.

- **Made available in *Data Releases***
  - Annually, except for Year 1
    - Two DRs for the first year of data

- **Complete reprocessing of all data, for each release**
  - Every DR will reprocess ***all*** data taken up to the beginning of that DR

- **Projected catalog sizes:**
  - **18 billion objects** (DR1) → **37 billion** (DR11)
  - **750 billion observations** (DR1) → **30 trillion** (DR11)

Level 2: Archive Contents

- Processed visits ("calibrated exposures")
  - Visit images with instrumental signature removed, background, PSF, zero-point and WCS determined
- Coadds
  - Deep coadds across the entire survey footprint (multiple flavors)

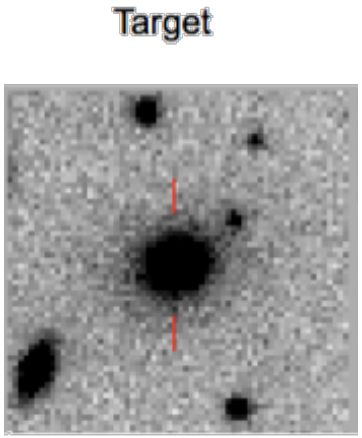
More in DPDD, Section 5.4

- Catalogs of Sources
  - Measurements of sources detected on calibrated exposures
- Catalogs of Objects
  - Characterization of objects detected on multi-epoch data
- Catalogs of ForcedSources
  - Forced photometry performed on all exposures, at locations of all Objects

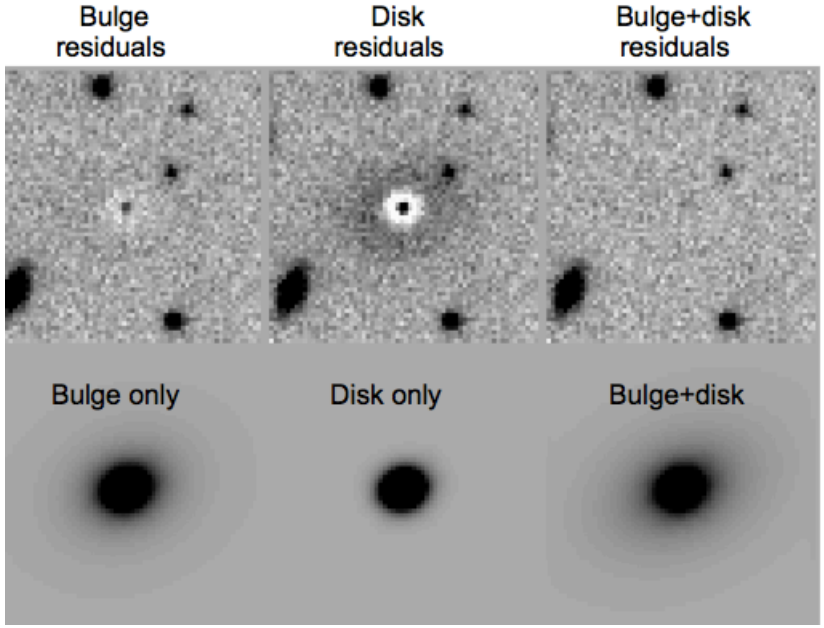
More in DPDD, Section 5.3

LSST Catalog Contents (Level 2)

- **Object characterization (models):**
  - Moving Point Source model
  - Double Sérsic model (bulge+disk)
    - Maximum likelihood peak
    - **Samples of the posterior (hundreds)**
- **Object characterization (non-parametric):**
  - Centroid: (α, δ), per band
  - Adaptive moments and ellipticity measures (per band)
  - Aperture fluxes and Petrosian and Kron fluxes and radii (per band)
- **Colors:**
  - Seeing-independent measure of object color
- **Variability statistics:**
  - Period, low-order light-curve moments, etc.



LSST Science Book, Fig. 9.3



Level 3:

Enabling the Creation of Added-Value Data Products

Level 3: Added Value Data Products

- **Level 3 Data Products: Added-value products created by the community**

- **These may enable science use-cases not fully covered by what we'll generate in Level 1 and 2:**
  - Catalogs of SNe light echos
  - Characterization of diffuse structures (e.g., ISM)
  - Extremely crowded field photometry (e.g., globular clusters)
  - Custom measurement algorithms

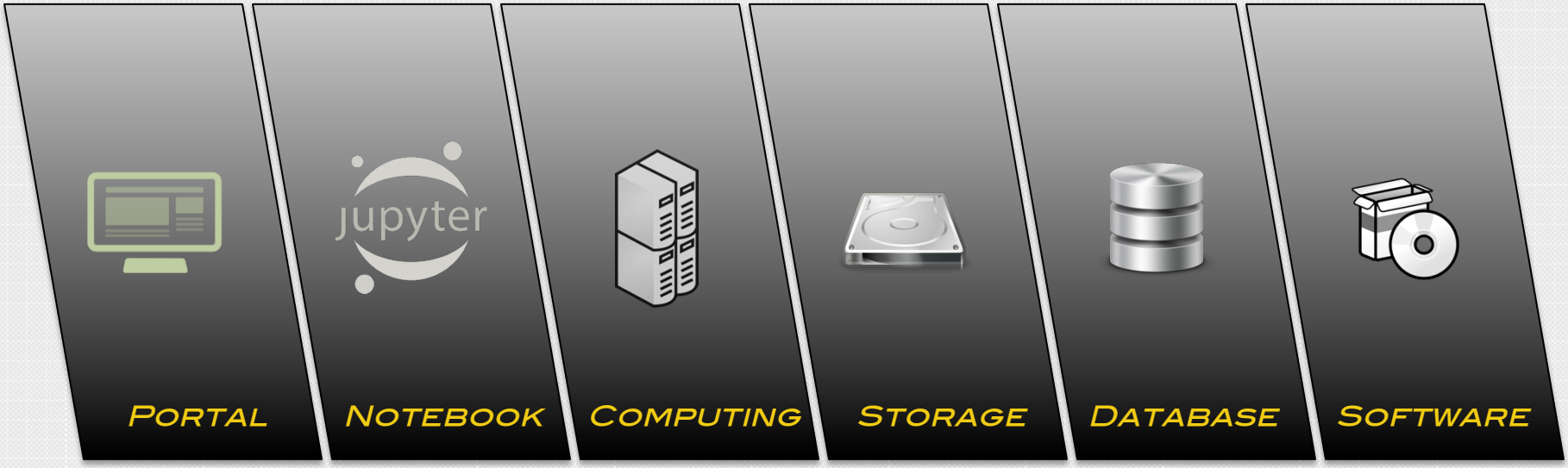
- **We want to make it easier for the community to create and distribute Level 3 products**
  - **Enabling limited end-user analysis and processing at the LSST data center**
  - **User databases and workspaces ("mydb")**
  - (making the LSST software stack available to end-users)

- Level 3 products may be migrated to Level 2 (with owners' permission); this is one of the ways how Level 2 products will evolve.

LSST Data Access Center Services

Accessing, Exploring, and Analyzing LSST Data

Services Offered at the LSST Data Access Center



Key LSST Deliverables: The Data and a Way to Reach It

The LSST will be a facility whose primary mission is to acquire, process, and make available to the data-rights holders the data collected by its telescope and camera. Our primary products are the stream of events alerts (Level 1) and Data Release data products (Level 2).

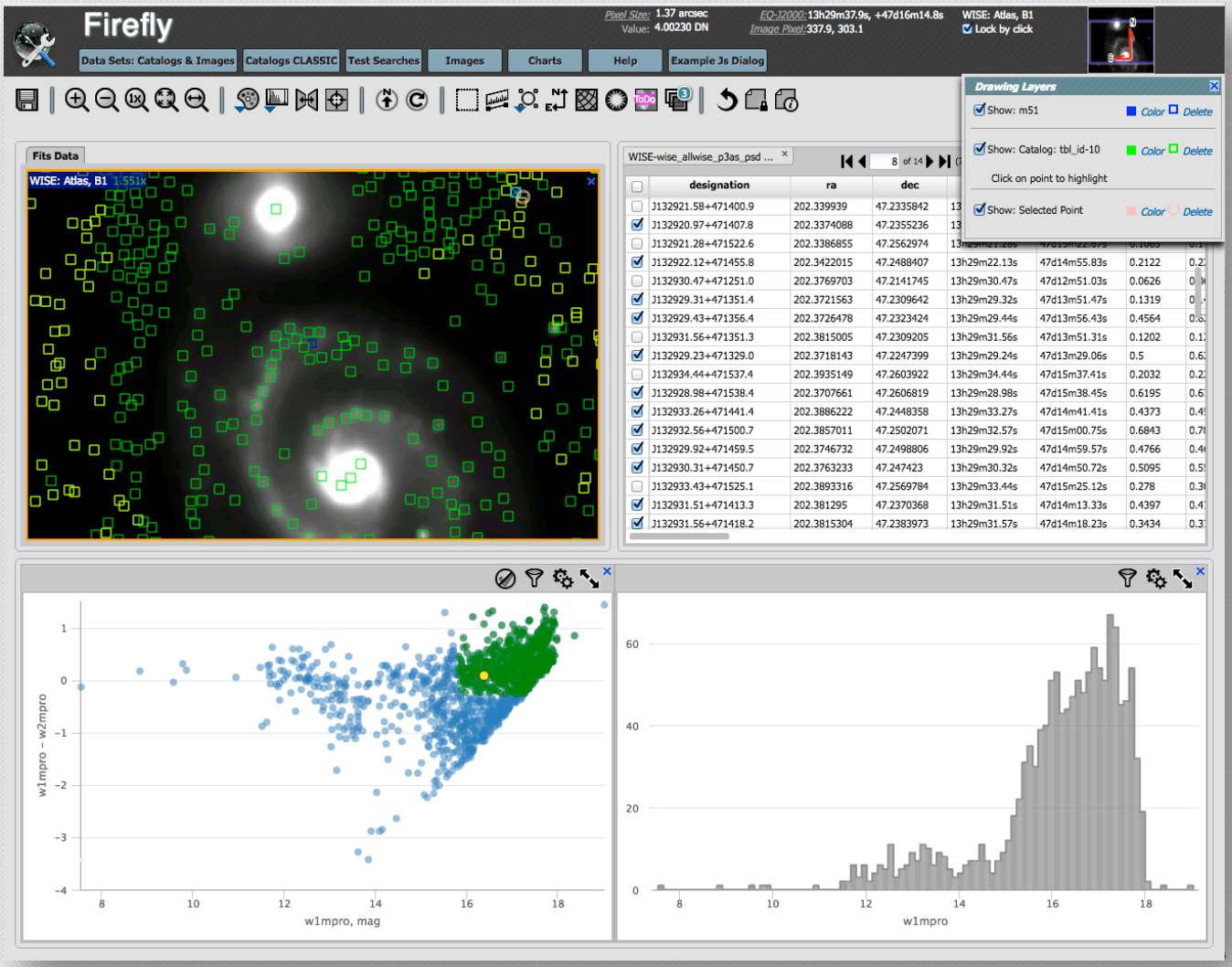
To make those products available and useful to the community, we're building two **Data Access Centers** – one in the U.S. and another one in Chile. These will expose the LSST data to the data rights holders through a number of data access center services.

LSST Portal: The Web Window into the LSST Dataset

The Web Portal to the archive will enable browsing and visualization of the available datasets in ways the users are accustomed to at archives such as IRSA, MAST, or the SDSS archive, with an added level of interactivity.

Through the Portal, the users will be able to view the LSST images, request subsets of data (via simple forms or SQL queries), construct simple plots, and generally explore the LSST dataset in a way that allows them to identify and access (subsets of) data required by their science case.

LSST Portal: The Web Window into the LSST Dataset



The Firefly Web Science User Interface (Wu et al, 2016; ADASS)

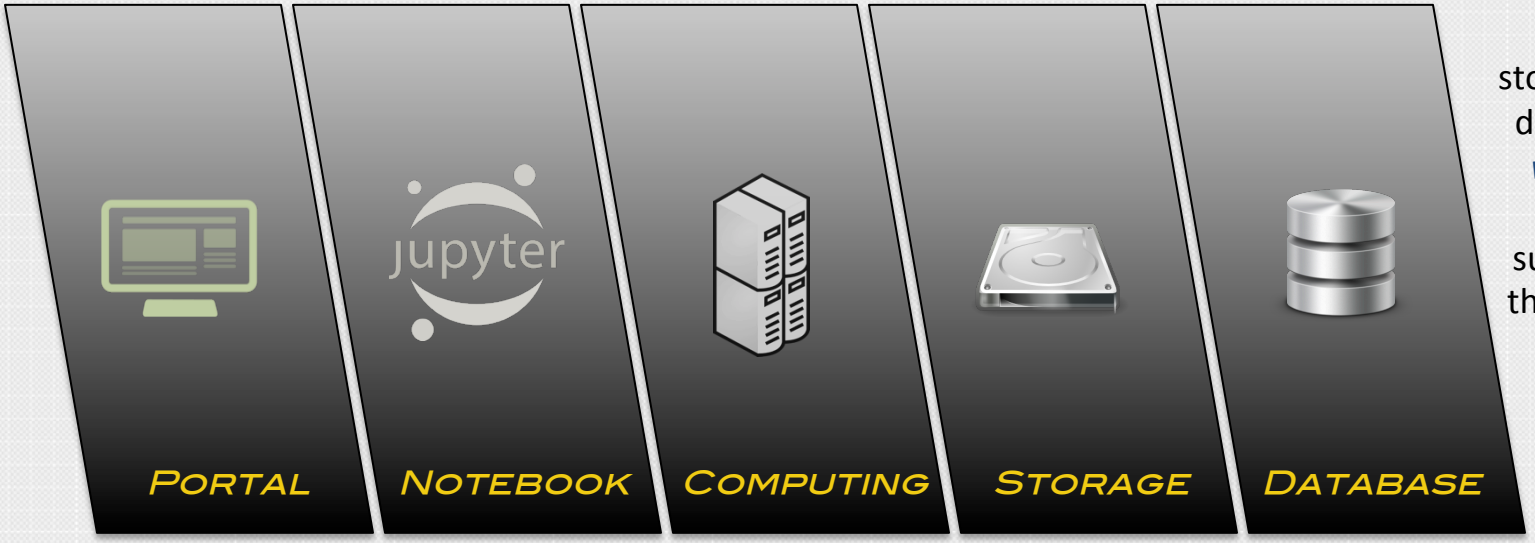
Next-to-the-data Analysis: Jupyter Notebooks

The tools exposed through the Web Portal will permit simple exploration, subsetting, and visualization LSST data. They may not, however, be suitable for more complex data selection or analysis tasks.

To enable that next level of next-to-the-data work, we plan to enable the users to launch their own Jupyter notebooks at our computing resources at the DAC. These will have fast access to the LSST database and files. They will come with commonly used and useful tools preinstalled (e.g., AstroPy, LSST data processing software stack).

This service is similar in nature to efforts such as SciServer at JHU, or the JupyterHub deployment for DES at NCSA.

Computing, Storage, and Database Resources



Computing, file storage, and personal databases (the *"user workspace"*) will be made available to support the work via the Portal and within the Notebooks.

An important feature is that no matter how the user accesses the DAC (Portal, Notebook, or VO APIs) they always "see" the same workspace.

How big are US/Chile DAC user resources (@ DR2)?

- **Computing:**
  - 2,400 cores
  - 18 TFLOPs
- **File storage:**
  - U.S.: 3 PB
  - Chile: 0.8 PB
- **Database storage**
  - U.S.: 2 PB
  - Chile: 0.8 PB

This is shared by all users. At various times we have estimated the number of potential users to be in the low 1000s (relevant for file and database storage).

Not all users will be accessing the computing cluster concurrently. A reasonable guess may be in 10-100 range.

Though this is a relatively small cluster by 2020-era standards, it should be **sufficient to enable preliminary end-user science analyses** (working on catalogs, smaller number of images) and creation of some added-value (Level 3) data products.

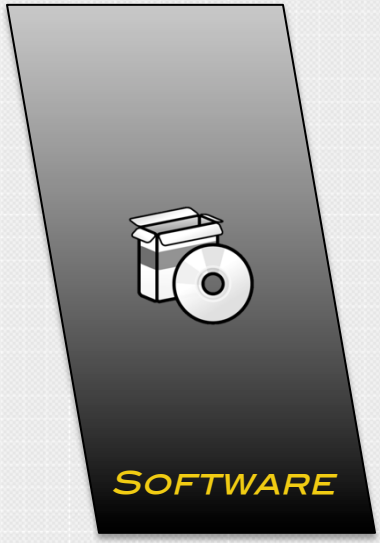
For larger endeavors (e.g., pixel-level reprocessing of the entire LSST dataset), the users will want to use resources beyond the LSST DACs (more later).

LSST Data Access Center Services

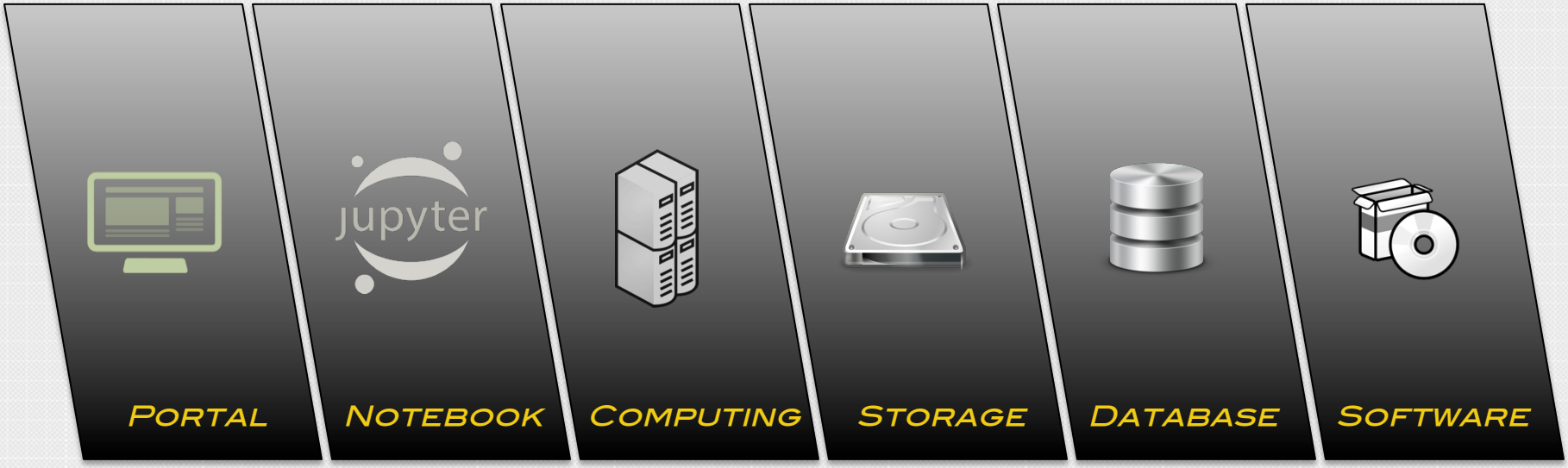
As the final “piece of the puzzle”, we’re also be making available the source code of the LSST data processing software (and configurations used in processing).

This will enhance reproducibility of the LSST data products, as well as provide source-code level of insight into algorithms utilized by LSST data processing.

Having the source code may also enable community efforts extend and apply the LSST codes to projects beyond the LSST. Some efforts, such as processing of HSC Survey data (Miyazaki et al.) or of CFHT-LS (Boutigny et al.), are already under way.



Putting it all together:



How we (think) we will work with LSST data?

- Most users are likely to begin with the Web Portal, to become familiar with the LSST data set and query smaller subsets of data for “at home” analysis. Some may use the tools they’re accustomed to (e.g., TOPCAT, Aladin, AstroPy, etc.) to grab the data using LSST’s VO-compatible APIs.

- Some users may choose to continue their analysis by utilizing resources available to them at the DAC. We’re setting aside some ~10% of our storage and computing resources (~20-100 TFLOPS) to enable that. We will use this to power a JupyterHub-type remote analysis environment and a small HPC-type processing cluster. It’s quite possible that a large fraction of end-user (“single PI”) science may be achievable this way.

- For users who need more, they may be able to apply for more resources at adjacent computing facilities. For example, the U.S. DAC will be located in the National Petascale Computing Facility at National Center for Supercomputing Applications (NCSA). Significant additional supercomputing is expected to be available at the same site (e.g., NPCF currently hosts the Blue Waters supercomputer).

- Finally, rights-holders may build their own DACs or computing facilities to support larger-scale processing, reusing our software (pipelines, middleware, databases) to the extent possible.

LSST: In operations by 2022.