# LSST Data Management: Processing, Pipelines and Workflows

Simon Krughoff,
LSST Data Management, UW

Robyn Allsman,
Yusra AlSayyad,
Tim Axelrod,
Jacek Becla,
Andrew Becker,
Steve Bickerton,
Jim Bosch,
Bill Chickering,
Andy Connolly,
Greg Daues,
Gregory Dubois-Fellsman,
Mike Freemon,
Andy Hanushevsky,
Mario Juric,
Fabrice Jammes,
Lynne Jones,
Jeff Kantor,

Kian-Tat Lim,
Dustin Lang,
Ron Lambert,
Robert Lupton,
Serge Monkewitz,
Jon Myers,
Russell Owen,
Steve Pietrowicz,
Ray Plante,
Paul Price,
Andrei Salnikov,
Dick Shaw,
Schuyler Van Dyk,
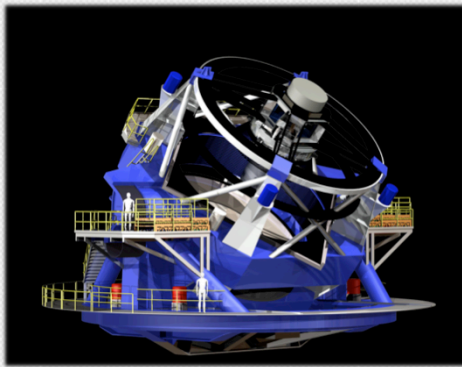*Daniel Wang*

and the LSST Project Team
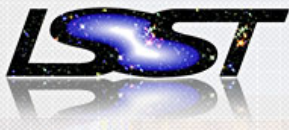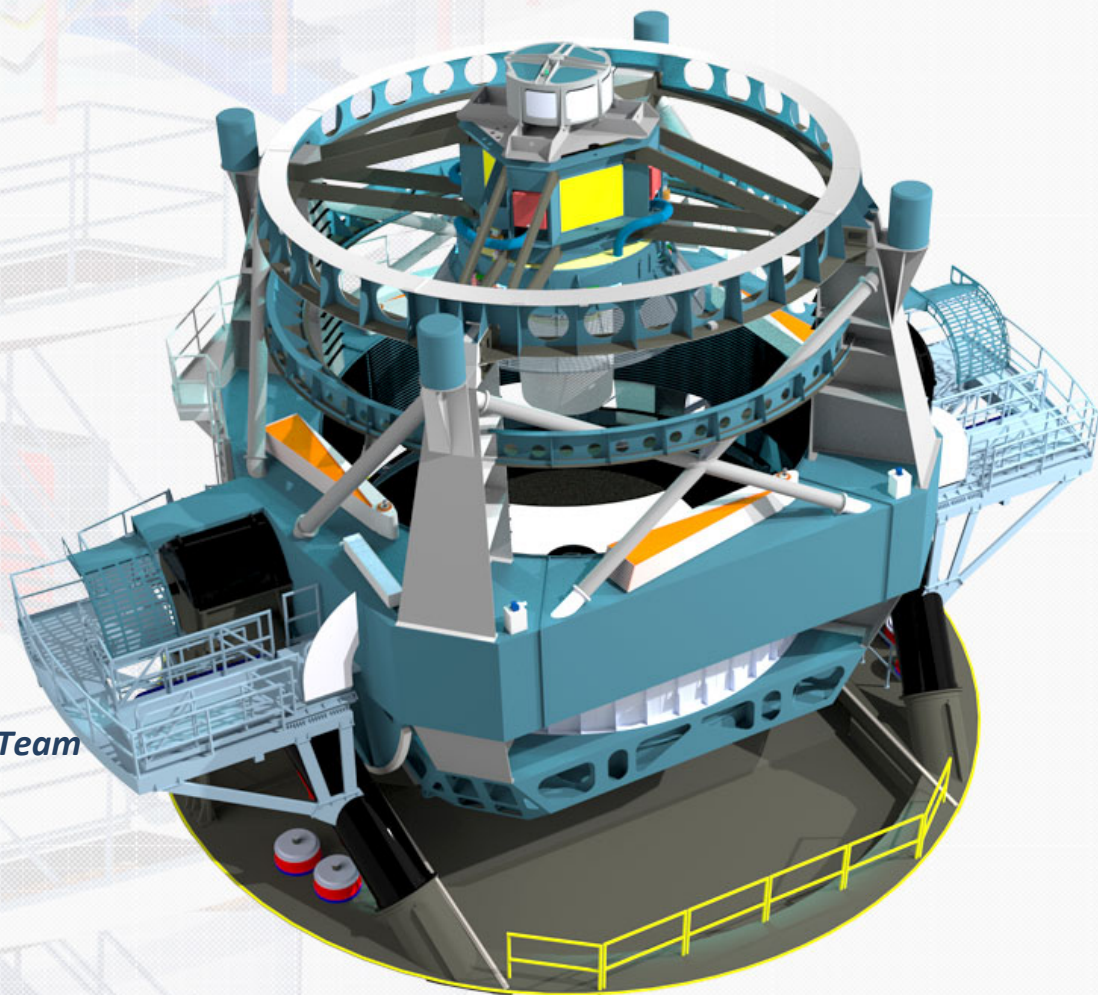
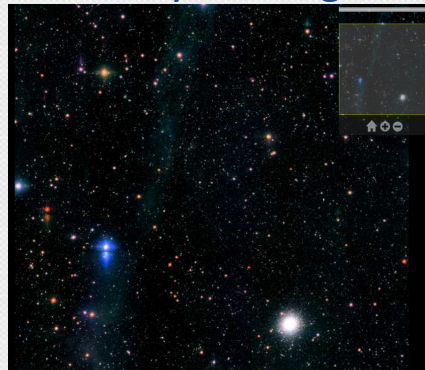OSG ALL HANDS
March 25, 2015

A Dedicated Survey Telescope

- **A wide (half the sky), deep (24.5/27.5 r-band 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 <u>fully reduced data</u>.**
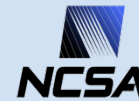 - All science will come from survey catalogs and images

Telescope ➔ Images ➔ Catalogs

LSST Operations: Sites and Data Flows

Archive Site
NCSA

Archive Center
Alert Production
Data Release Production
Calibration Products Production
EPO Infrastructure
Long-term Storage (copy 2)

Data Access Center
Data Access and User Services

Dedicated Long Haul Networks

Two redundant 40 Gbit links from La Serena to Champaign, IL (existing fiber)

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

Infrastructure: Petascale Computing, Gbit Networks
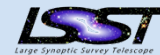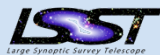
> The computing cluster at the **LSST Archive** (at NCSA) will run the processing pipelines.

- *Single-user, single-application, dedicated data center*
- *Process images in real-time to detect changes in the sky*
- *Produce annual data releases*

Archive Site and U.S. Data Access Center
NCSA, Champaign, IL

Base Site and Chilean Data Access Center
La Serena, Chile
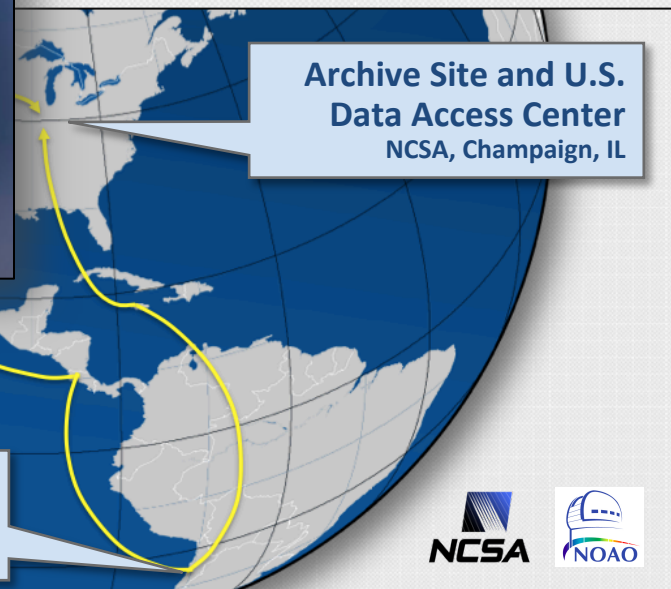
Long Haul Networks to transport data from Chile to the U.S.

- *200 Gbps from Summit to La Serena (new fiber)*
- *2x40 Gbit (minimum) for La Serena to Champaign, IL (protected, existing fiber)*

NCSA

NOAO

LSST Pipelines

- 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.

Alert Stream

- 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.
- Deep co-added images.

Data Release

- Services and computing resources at the Data Access Centers to enable user-specified custom processing and analysis.
- Software and APIs enabling development of analysis codes.

Value Added

Level 1: Transients Alerts

- **LSST computing is sized for 10M alerts/night (average), 10k/visit (average), 40k/visit (peak)**
  - Dedicated networking for moving data from Chile to the US
  - New image differencing pipelines with improved algorithms

- **Will measure and *transmit with each alert*:**
  - position
  - flux, size, and shape
  - light curves in all bands (up to a ~year; stretch: all)
  - variability characterization (eg., low-order light-curve moments, probability the object is variable)
  - cut-outs centered on the object (template, difference image)

Level 1 Processing: System Architecture
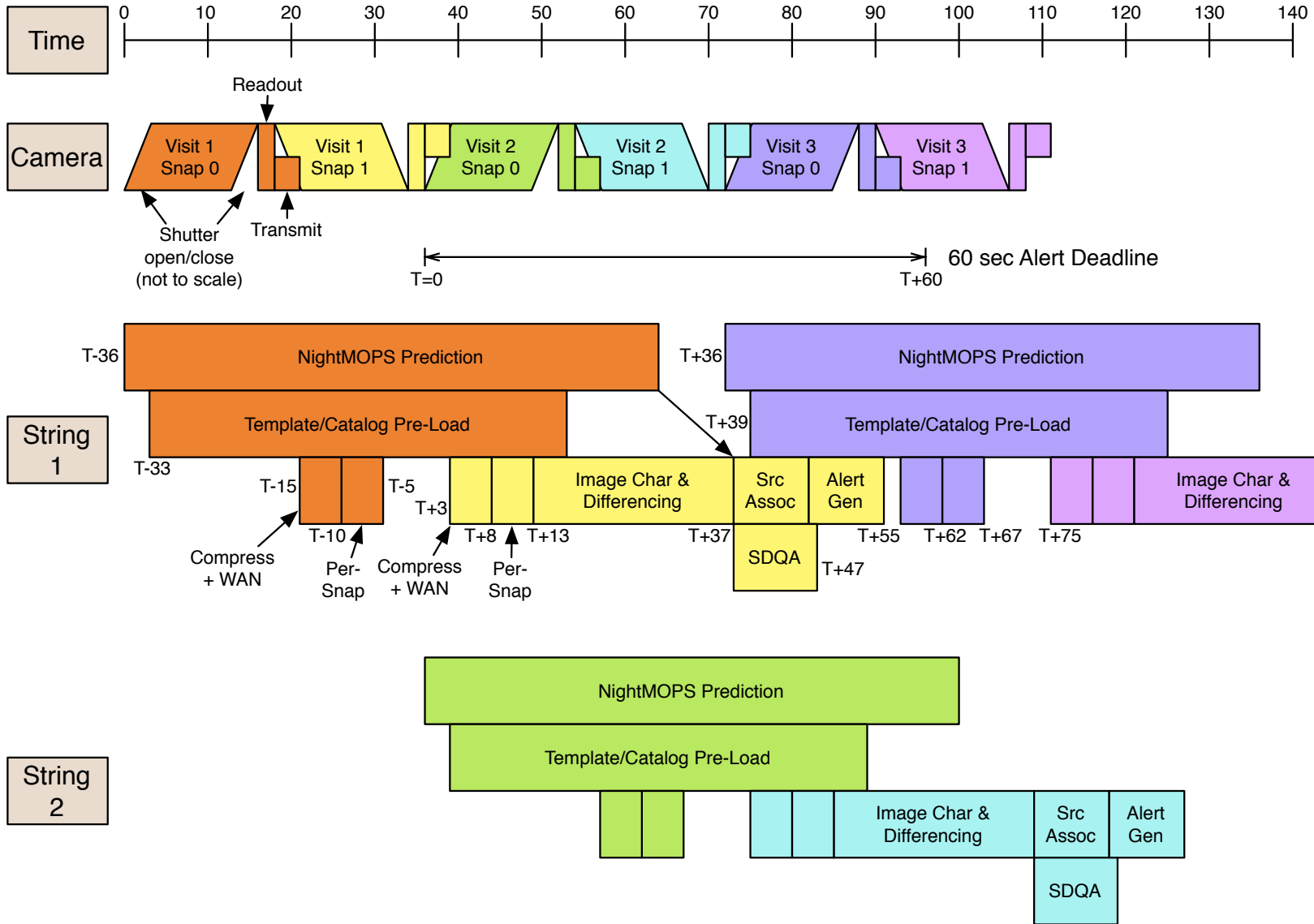
Level 2: Annual Data Releases

- **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)

Data Release Processing: Logical Flow

DRP begins with processing, detection, and measurement on single frames, generating *Sources*.

These are used to photometrically and astrometrically calibrate the survey.

A series of *coadds* is built next, where *Objects* are detected.

Detections on co-adds are *deblended* and *associated* to form a master object list.

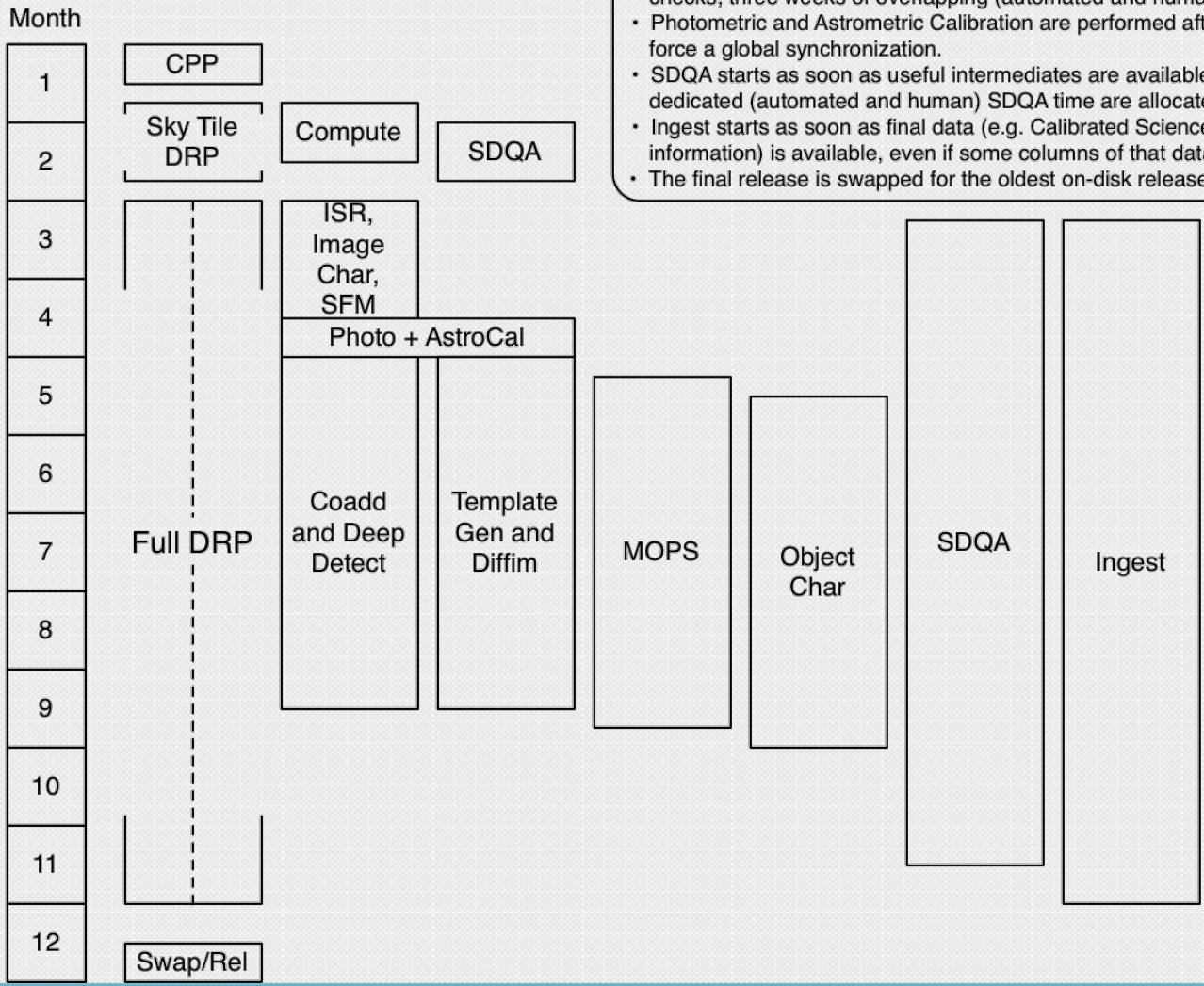The objects are simultaneously characterized in all observed epochs (*MultiFit*).

Time variability is characterized by independent measurement of *Forced Sources* in individual epochs.
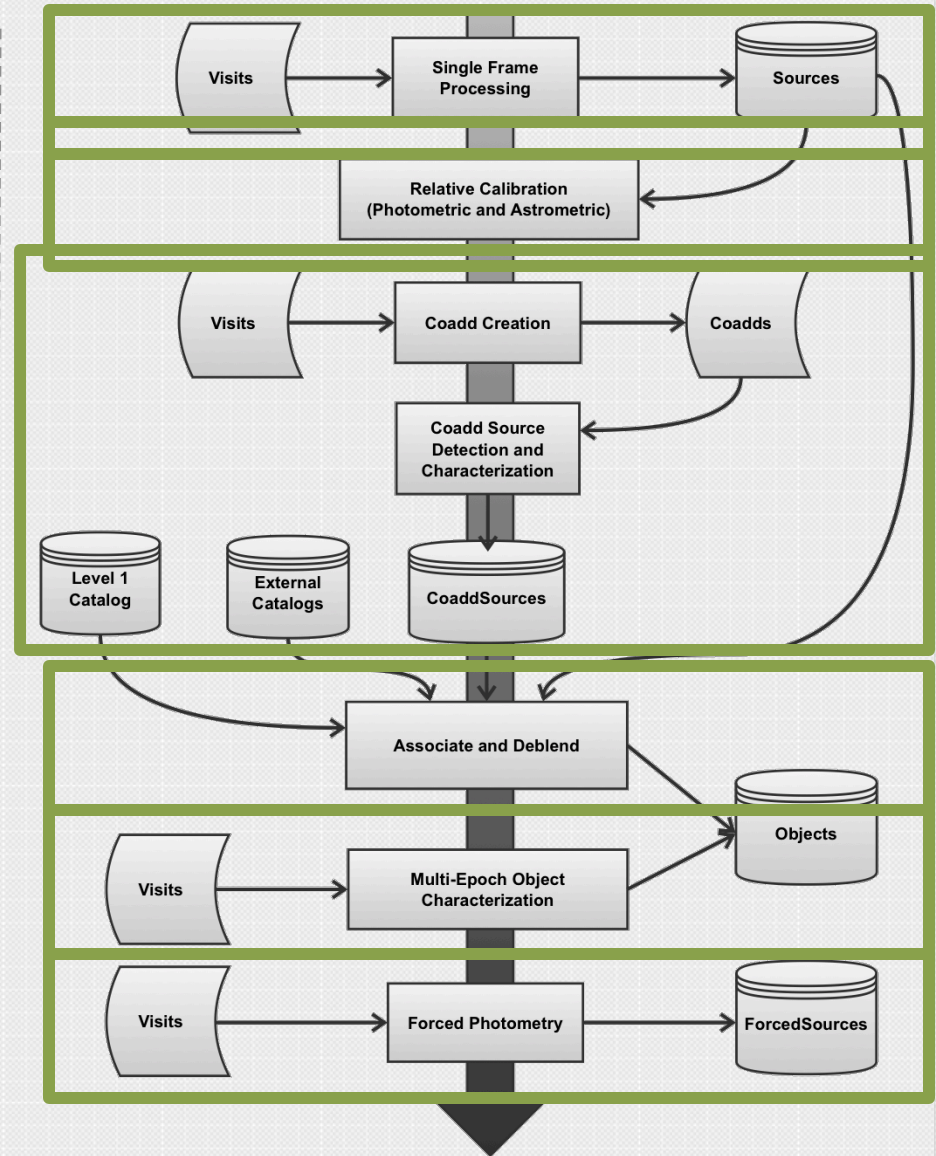
Data Release Processing: Physical Flow

Notes:
- CPP is the Calibration Products Production, generating Master Calibration Images and the Calibration Database.
- A single sky tile is processed first through the entire DRP for initial testing and quality checks; three weeks of overlapping (automated and human) SDQA time are allocated.
- Photometric and Astrometric Calibration are performed after single-frame processing and force a global synchronization.
- SDQA starts as soon as useful intermediates are available. Six additional weeks of dedicated (automated and human) SDQA time are allocated.
- Ingest starts as soon as final data (e.g. Calibrated Science Exposure metadata and Source information) is available, even if some columns of that data will be modified later.
- The final release is swapped for the oldest on-disk release at the end of the year.

Month

1	CPP
2	Sky Tile DRP	Compute	SDQA
3	Full DRP	ISR, Image Char, SFM
4		Photo + AstroCal
5		Coadd and Deep Detect	Template Gen and Diffim	MOPS	Object Char	SDQA	Ingest
6						
7						
8						
9						
10						
11						
12	Swap/Rel

Level 3: Enabling User-created Data Products

- **Products created by the community using LSST's software, services, or computing resources.**
  - **Our task: enabling the creation of L3 products with infrastructure and software**

- **For use-cases not fully enabled by Level 1 and 2 processing:**
  - Reprocessing images to search for SNe light echos
  - Characterization of diffuse structures (e.g., ISM)
  - Extremely crowded field photometry (e.g., globular clusters)
  - Custom measurement algorithms

- **Enabling Level 3:**
  - User databases and workspaces ("mydb")
  - Enabling user computing at the LSST data center
    - processing that will greatly benefit from co-location with the LSST data
  - Making the LSST software stack available to end-users

Beyond LSST: Opportunities for Survey Projects

- **LSST DM's primary mission is to build the data processing system for LSST. We are scoped as required to achieve that.**

- That said, the optimal design for LSST (that balances technical and programmatic risks) makes this system _**reusable**_ and largely _**general purpose**_.
  - Necessary to deal with real-world hardware
  - Necessary to be able to process precursor data
  - Necessary to enable science ("Level 3") software to be written on top of it

- **Attractive opportunities exist for using the LSST stack components on existing and future data sets.**
  - Possibilities: SDSS, CFHT-LS, PanSTARRS, HSC, DES, WFIRST, Euclid, …
  - Good basis for analysis frameworks (LSST DESC)
  - **Leveraging a 100M+ NSF investment in large survey data management**

- **The benefits feed back to LSST:** more users, less bugs, better understanding, shorter path to science.

New Algorithms: Background-matched co-add of SDSS Stripe 82 in the vicinity of M2.

Background matching preserves diffuse structures.

Generated with LSST pipeline prototypes.

Figure:
5 sq. deg.
background-matched
coadd composite

(g,r,i)
~55 epochs

Region: Aqr
Galactic lat = -35.0

Slide: Yusra AlSayyad

http://moe.astro.washington.edu/sdss/

Finding Out More

- LSST Data Management Home:
  - https://confluence.lsstcorp.org/x/QYEf
  - http://dm.lsst.org

- Installing LSST stack:
  - https://confluence.lsstcorp.org/display/LSWUG/LSST+Stack+Installation

- Mailing list – dm-users (see below for a ful list):
  - https://confluence.lsstcorp.org/display/DM/Communication+and+Links

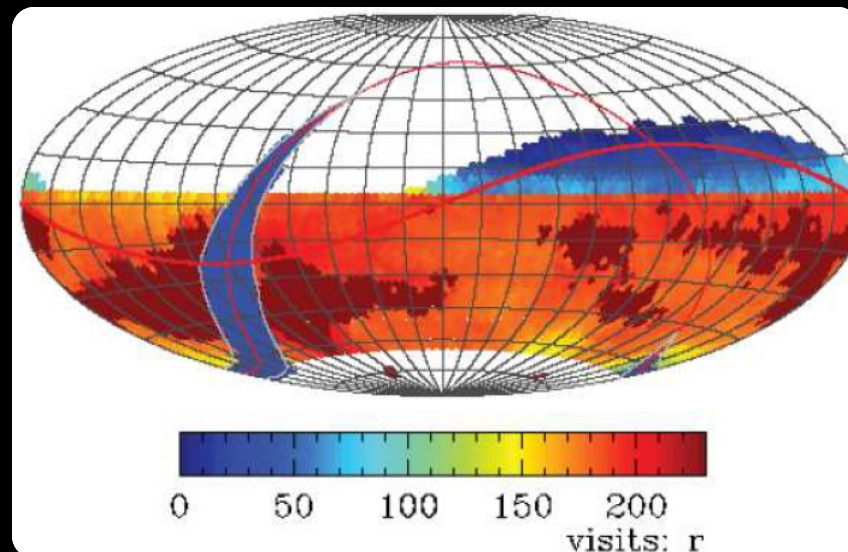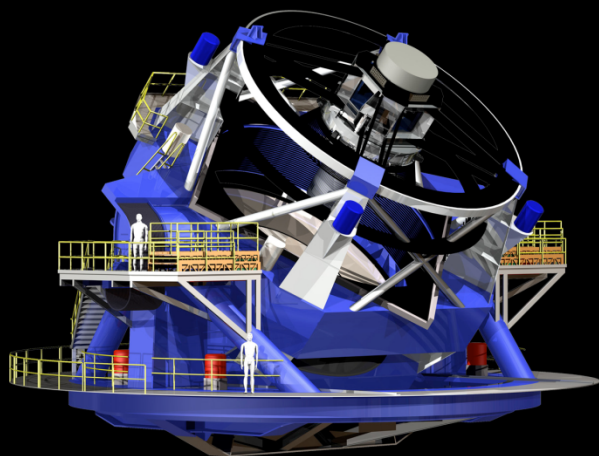8.4m telescope   18000+ deg^2 10mas astrom.   r<24.5 (<27.5@10yr)

ugrizy   0.5-1% photometry

3.2Gpix camera   30sec exp/4sec rd   15TB/night   37 B objects

Imaging the visible sky, once every 3 days, for 10 years (825 revisits)