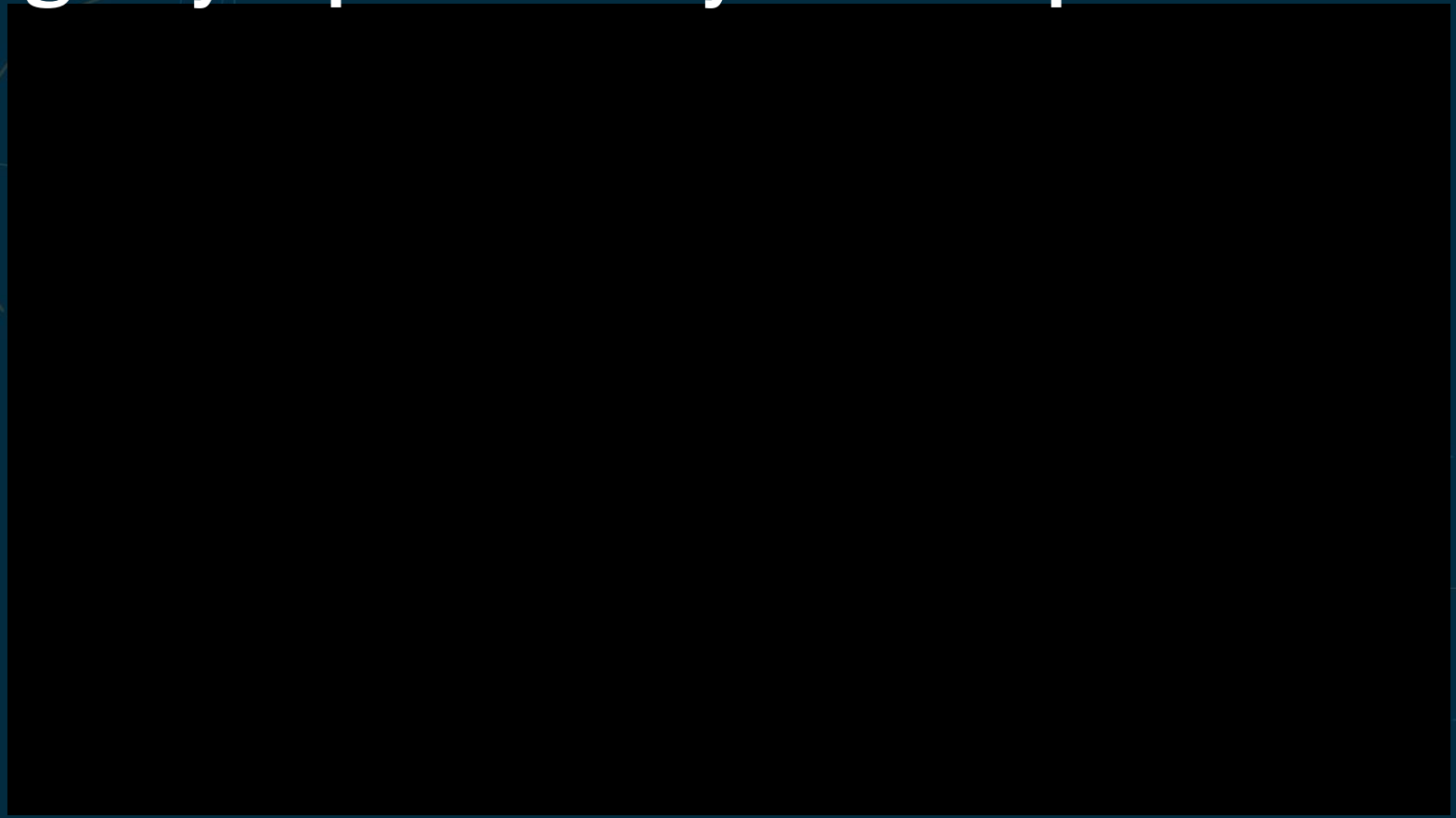
LSST Data Management and Software Challenges

Yusra AlSayyad – LSST/Princeton University

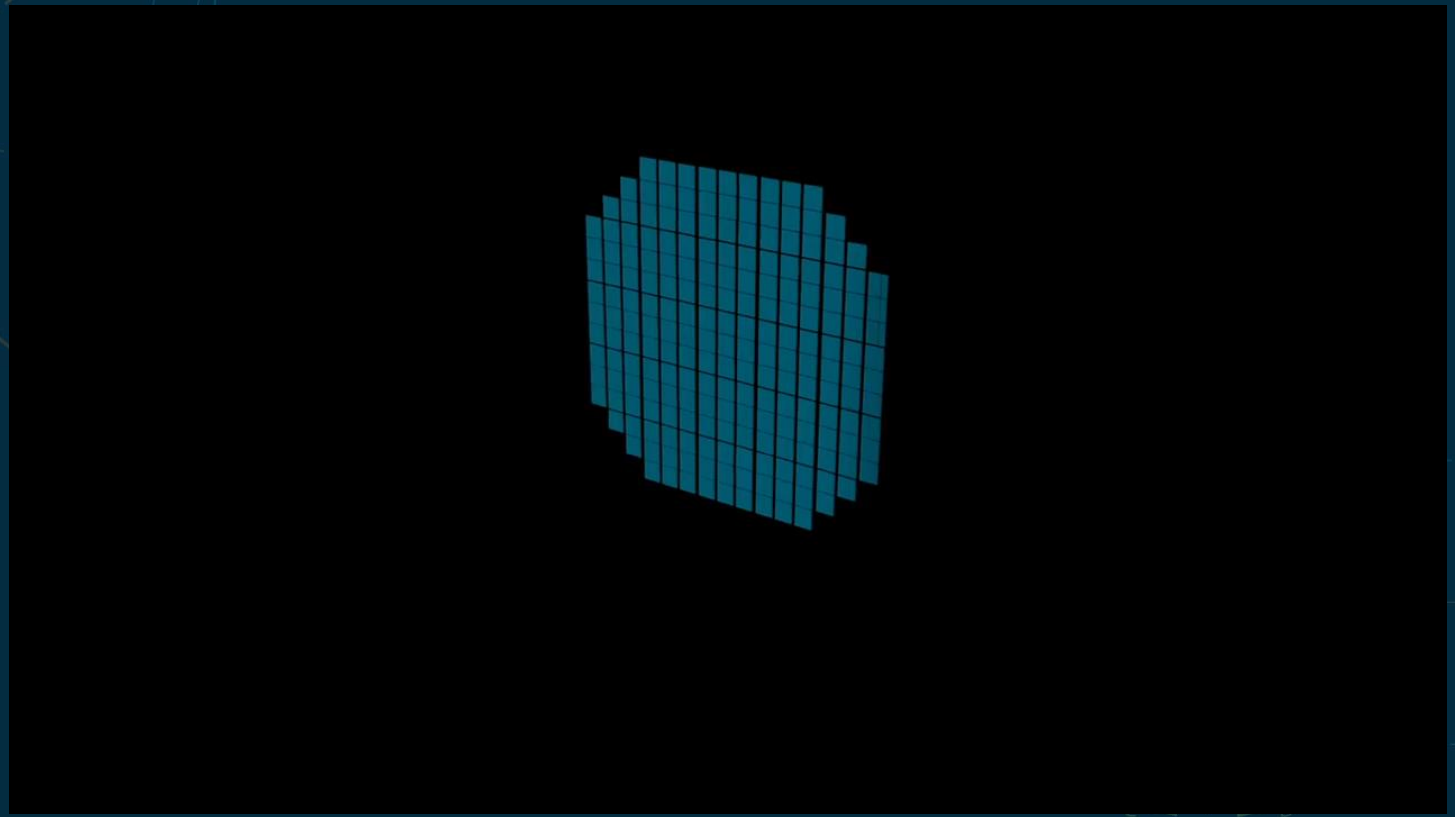
HOW2019 - March 18 2019

Large Synoptic Survey Telescope



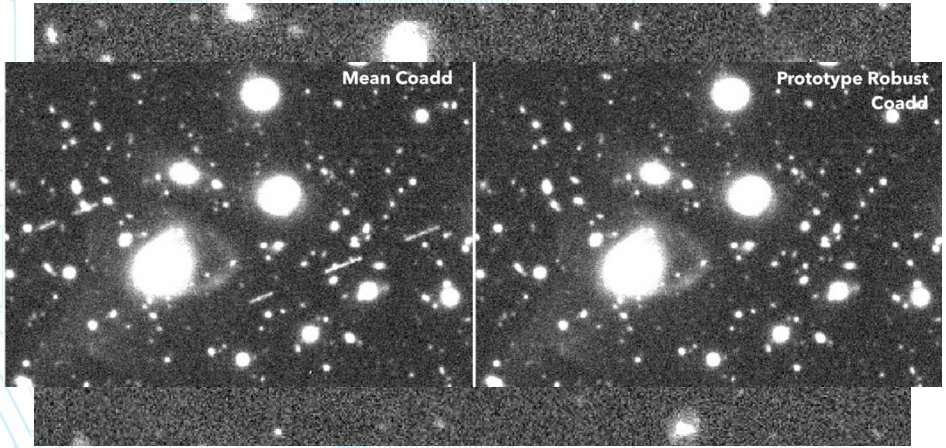
Animated Video credit: SLAC National Accelerator Lab

Large Synoptic Survey Telescope



Animated Video credit: SLAC National Accelerator Lab

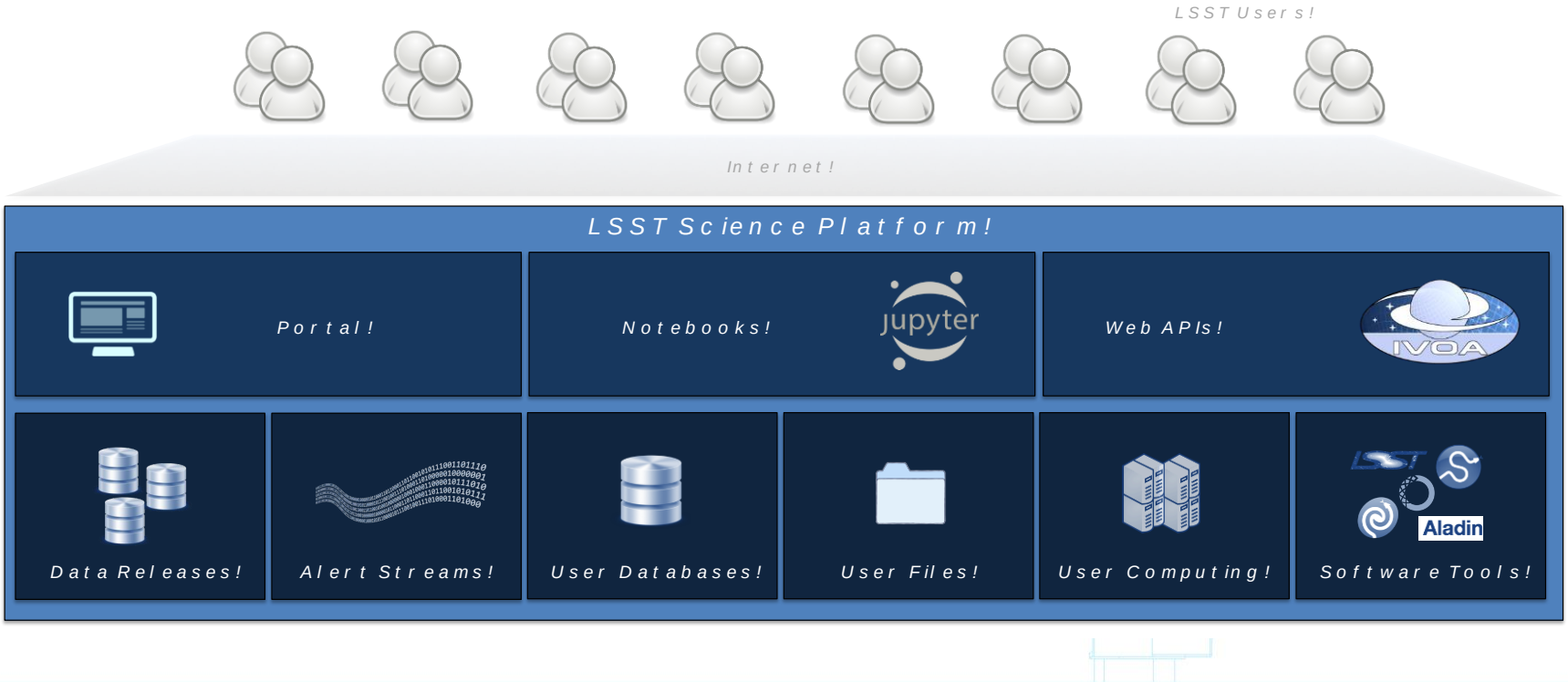
LSST Data Management System



Annual Data Release Products
11 Data releases in 10 years.
Final catalog: 15PB
Final pixels: 500PB

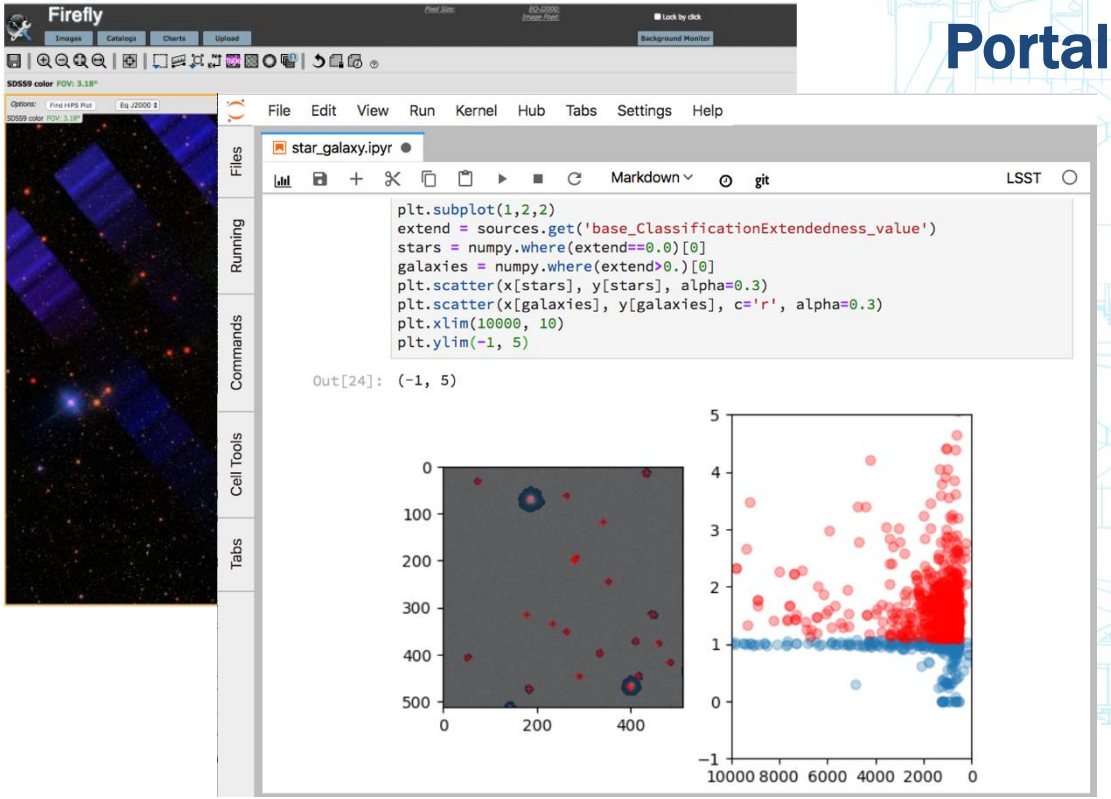
Prompt Data Products via nightly alert streams
~10 million alerts per night
issued within 60 s of shutter close

LSST Data Management System



LSST Science Platform technology emerged

LSST Science Platform



Portal

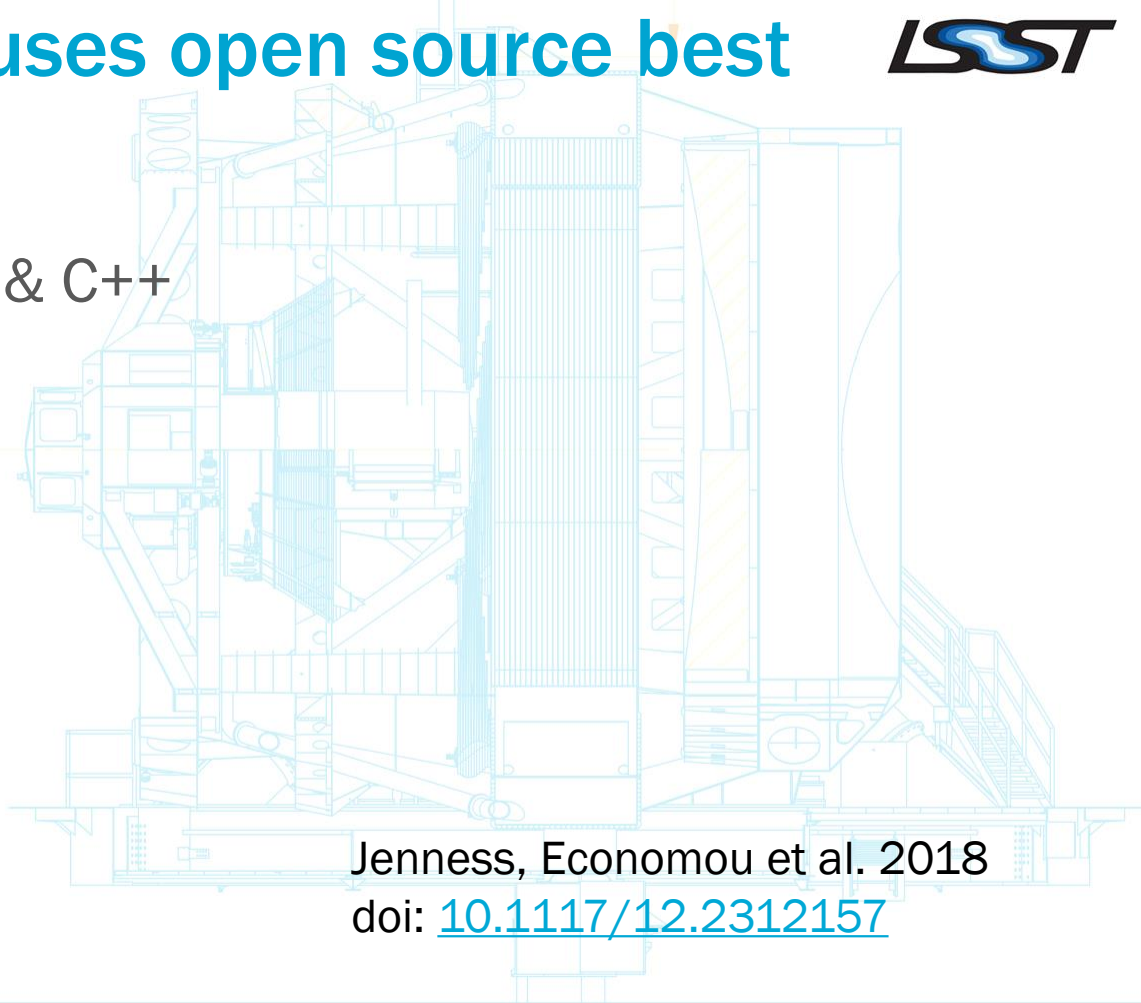
Jupyter Notebooks

Web APIs

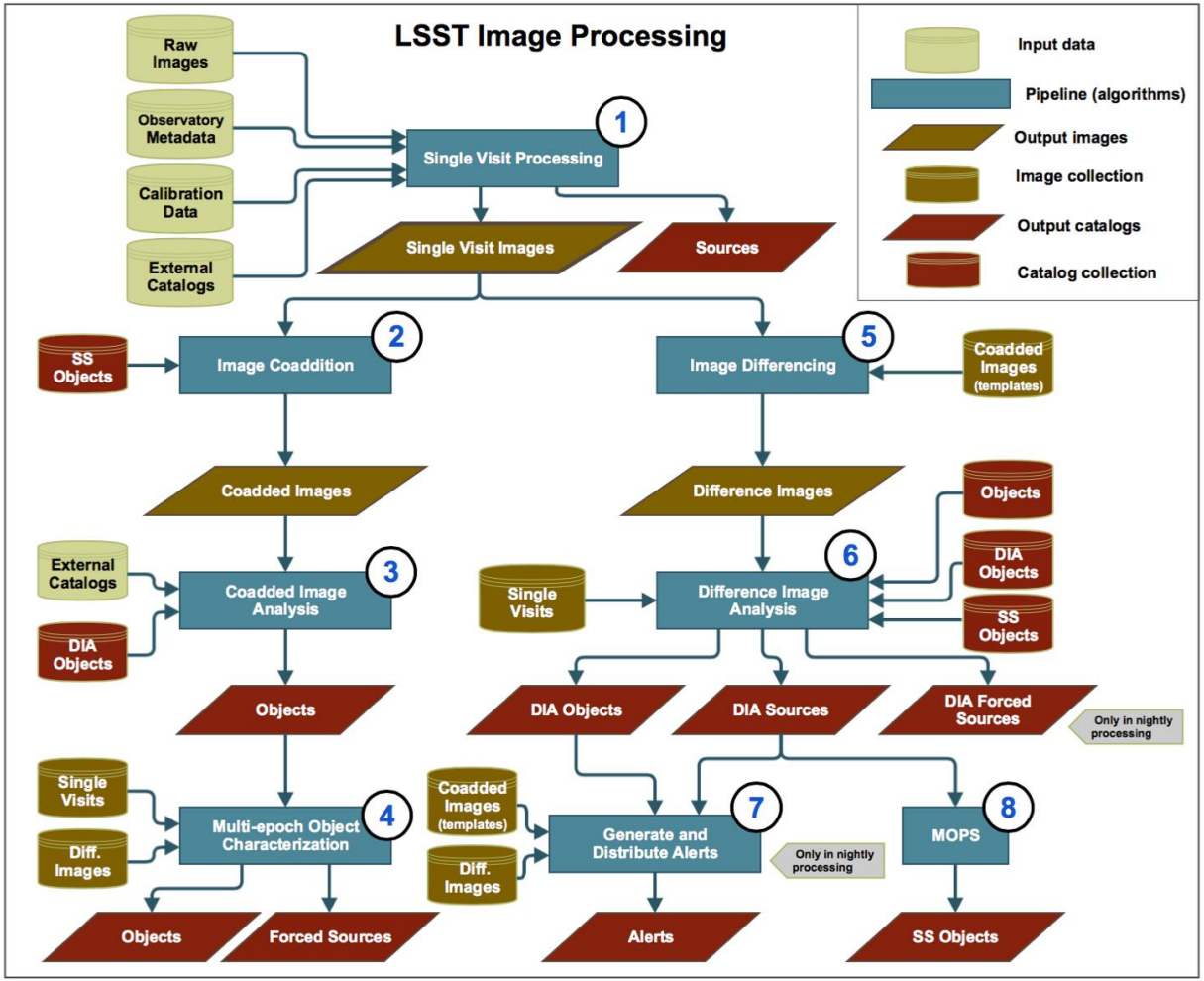
- Data access via IVOA-standard protocols
- Same interfaces that support other aspects

LSST DM software uses open source best practices

~500000 lines of Python & C++
Agile Principles
DevOps Engineering



Jenness, Economou et al. 2018
doi: 10.1117/12.2312157


LSST Image Processing
Raw Images
Observatory Metadata
Calibration Data
External Catalogs
1
Single Visit Processing
Single Visit Images
Sources
Input data
Pipeline (algorithms)
Output images
Image collection
Output catalogs
Catalog collection
SS Objects
2
Image Coaddition
Coadded Images
3
External Catalogs
DIA Objects
Coadded Image Analysis
Objects
Single Visits
Diff. Images
4
Multi-epoch Object Characterization
Objects
Forced Sources
5
Image Differencing
Coadded Images (templates)
Difference Images
6
Single Visits
Difference Image Analysis
Objects
DIA Objects
SS Objects
DIA Objects
DIA Sources
DIA Forced Sources
Only in nightly processing
Coadded Images (templates)
Diff. Images
7
Generate and Distribute Alerts
Only in nightly processing
Alerts
8
MOPS
SS Objects


Data Releases!
Alert Streams!

LSST DM ~100 devs across 5 institutions

PRINCETON UNIVERSITY

NCSA

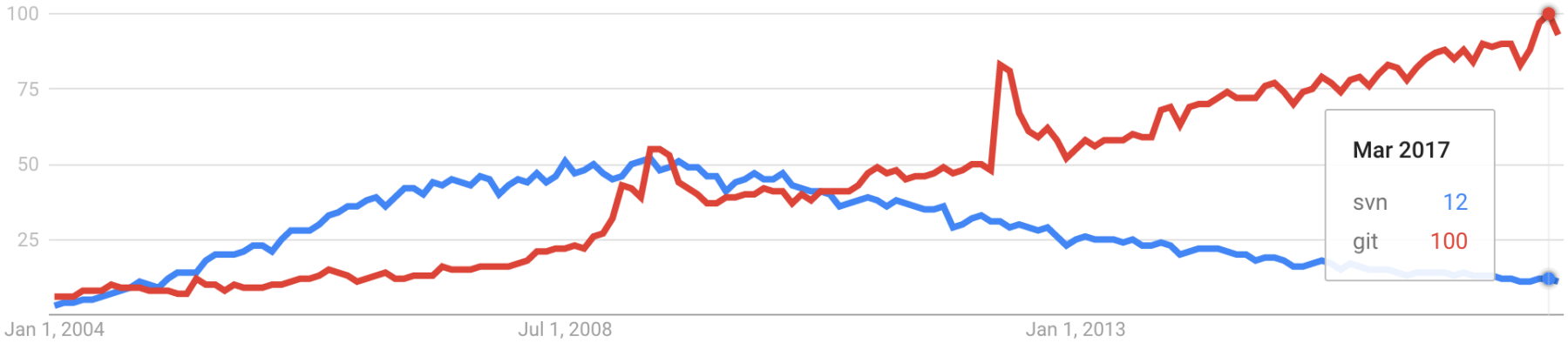
UNIVERSITY of WASHINGTON

SLAC

NOAO

Darling technologies come and go

Google Trends for svn vs. git



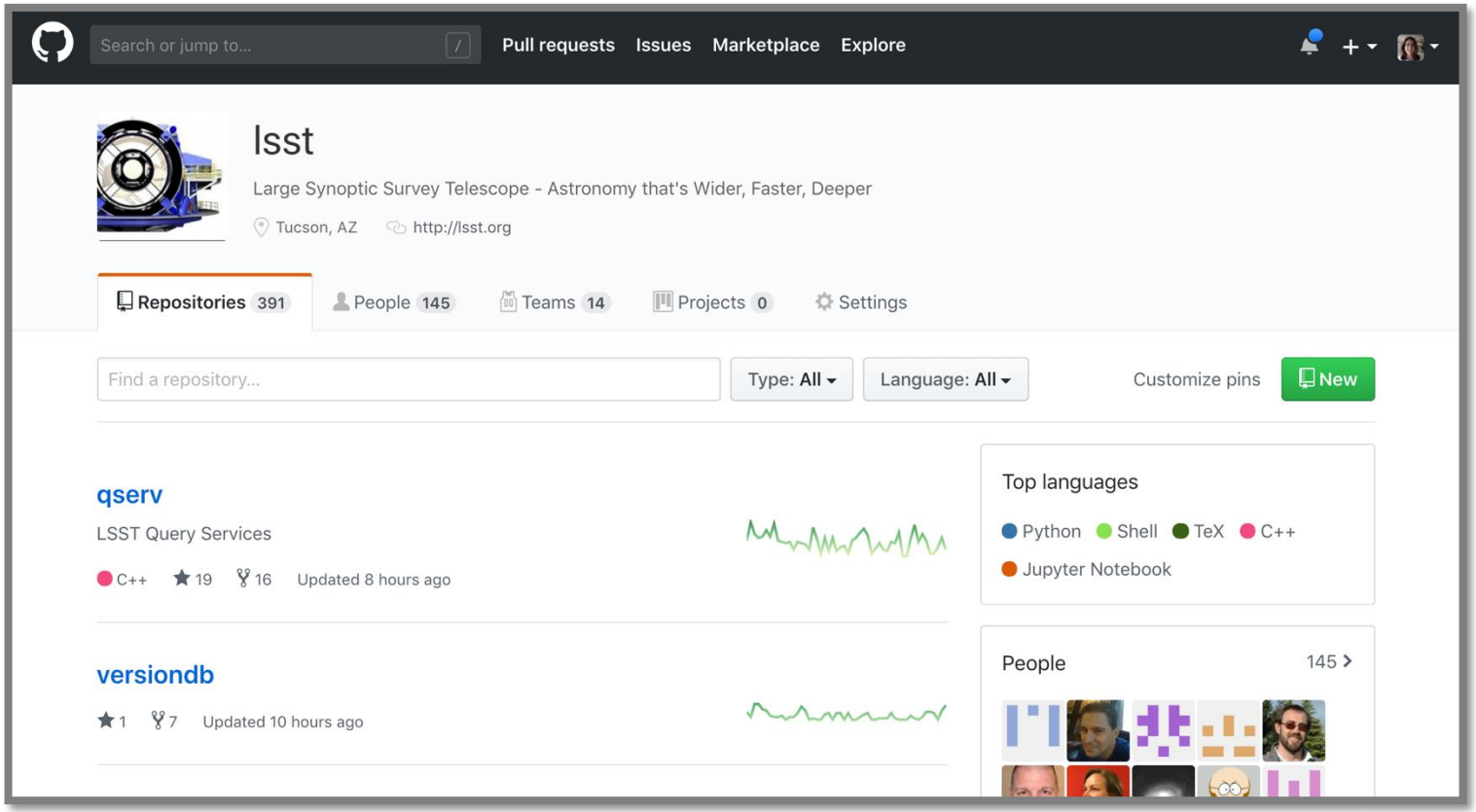
We get value from following open source trends

2011

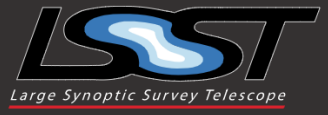
2014

2016

github.com/lsst



DevOps: developer.lsst.io



LSST DM Developer Guide

This is an internal guide for LSST DM staff. It's also openly available so that others can understand how we're building the LSST's data management subsystem.

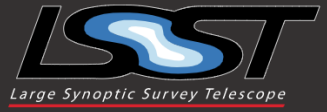
This guide includes a mix of normative requirements and helpful, descriptive, pages. When it's particularly important that you closely follow a standard, we include an annotation box at the top of the page.

Any member of DM can contribute to this guide. It's published from the https://github.com/lsst-dm/dm_dev_guide GitHub repo. Check out the README to get started.

Jump to: Team · Communications · Project documentation · Work management

Development guides: Overview · C++ · Python · Pybind11 · JavaScript · ReStructuredText · DM Stack · Git · Editors · Legal · User documentation style

DevOps: developer.lsst.io



LSST DM Developer Guide

Edition: Current

Change edition

Search docs

TEAM

Onboarding Checklist

Team Culture and Conduct Standards

Empowerment of DM team members

Data Release Production

COMMUNICATIONS

Configuring your GitHub username in your Slack profile

Docs » DM Development Workflow with Git, GitHub, JIRA and Jenkins

Edit on GitHub

DM Development Workflow with Git, GitHub, JIRA and Jenkins

This page describes our procedures for collaborating on LSST DM software and documentation with Git, GitHub and JIRA:

1. Configuring Git for DM development.
2. Using JIRA for agile development.
3. DM GitHub organizations.
4. Policies for naming and using Git branches.
5. Preparing code for review.
6. Reviewing and merging code.

In appendices, we suggest some *best practices* for maximizing the usefulness of our Git development history:

- Commit organization best practices.

Our outstanding challenges

- **Workflow management:** taking shape on 3rd try based on Pegasus
- **Software:** Release management. Quick deployment during commissioning.
- **Data management:** How expectations around data access will change?
  - How to be flexible. Column stores?
- **Workforce** (e.g. careers of the postdocs writing the software)

If you see me later, ask:

- for a demo of the JupterHub/Kubernetes-deployed Science Platform
- about adoption of Machine Learning in the astronomy community and plans in LSST
- why its important to choose technologies supported by the private sector and open source communities

If you see Margaret Johnson (NCSA) around, ask:

- About data facility challenges, data storage and compute plans.
- See talk Wednesday 2-3:30 OSG parallel

In Summary

The HEP and Astro communities share challenges and can learn from each other.

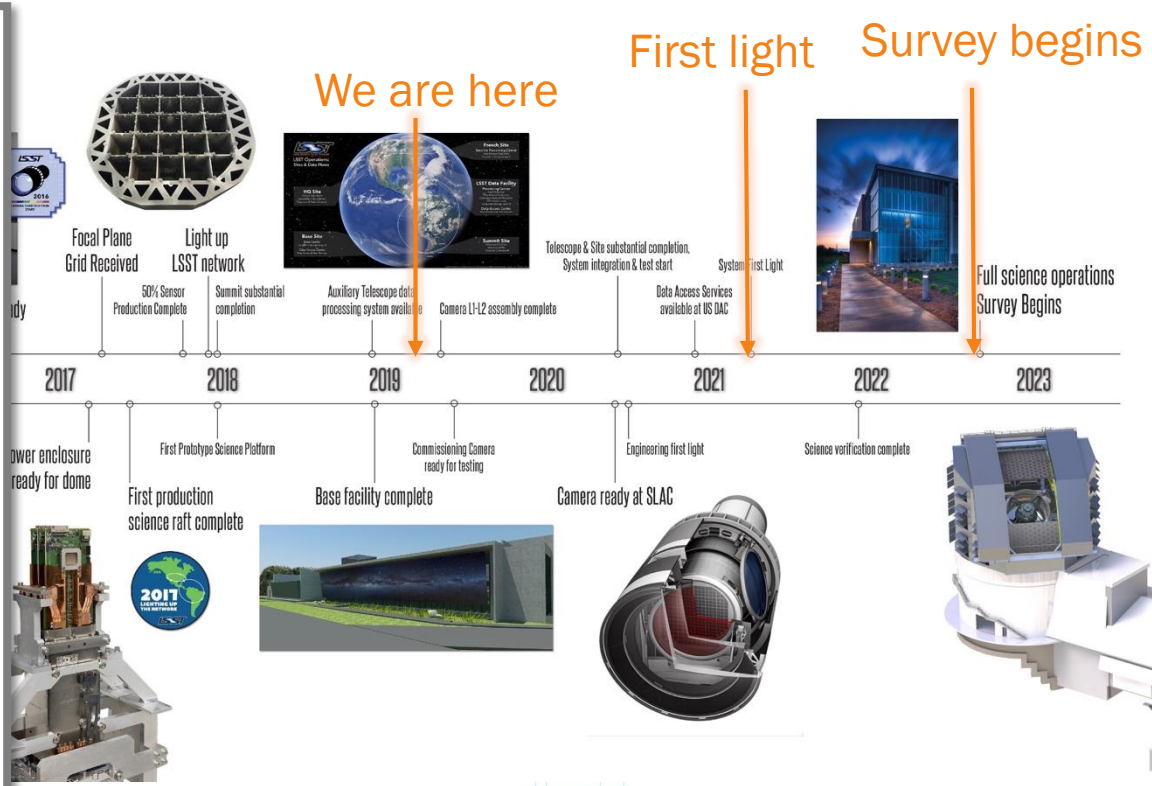
Find us:
@lsst
github.com/lsst
www.lsst.io for docs

www.lsst.org/scientists/keynumbers

Data and compute sizes:

Final volume of raw image data = 60 PB
Final image collection (DR11) = 0.5 Exabytes
Final catalog size (DR11) = 15 PB
Final disk storage = 0.4 Exabytes
Peak number of nodes = 1750 nodes
Peak compute power in LSST data centers = about 2 PFLOPS

18 months until first light
Then we start 2 years of commissioning



Large Synoptic Survey Telescope

LSST Operations: Sites & Data Flows

HQ Site
Tucson, AZ

Science Operations
Observatory Management
Education & Public Outreach

Base Site
La Serena, Chile

Base Center
Long-term storage (copy 1)

Data Access Center
Data Access & User Services

French Site
CC-IN2P3, Lyon, France

Satellite Processing Center
Data Release Production
Long-term Storage (copy 3)

LSST Data Facility
National Center for Supercomputing Applications (NCSA), Urbana-Champagne, IL

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

Data Access Center
Data Access and User Services

Summit Site
Cerro Pachón, Chile

Telescope & Camera
Data Acquisition
Crosstalk Correction