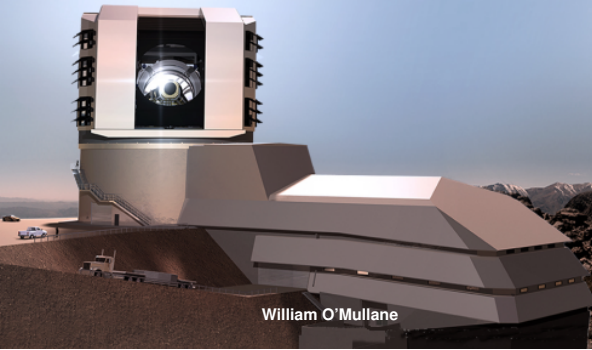
LSST Data Management

William O'Mullane input from Željko Ivezić
DM Project Manager

Towards Science in Chile with LSST
Santiago de Chile 14th December 2017



Outline

Introduction project status

Data Management

Data Access/ Science Platform

DM planning

Conclusion

Site shaping up

Artist impression

Photo December 2017

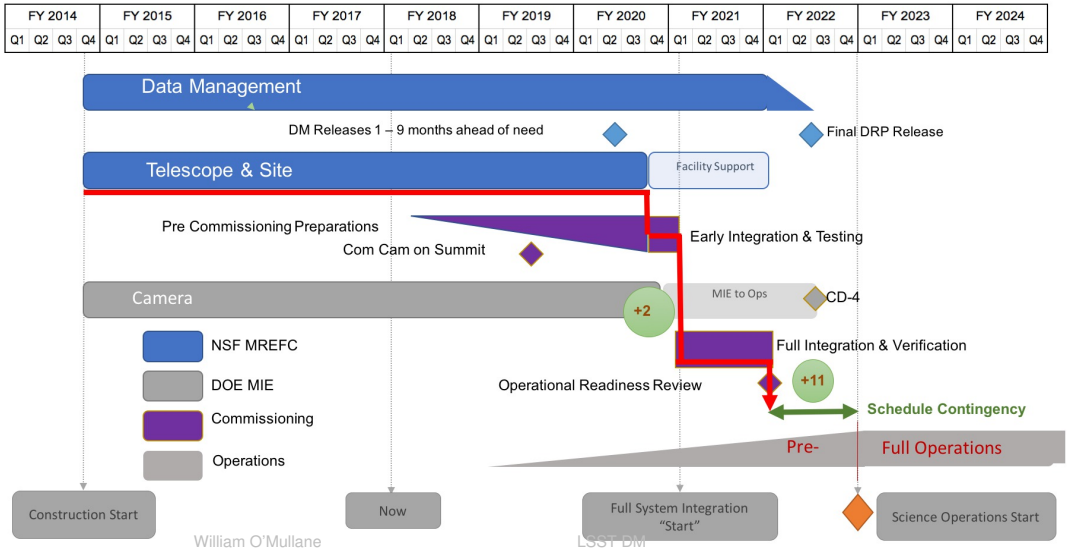
- Prime contract to finish Jan 2018.
- Azimuth rail sections aligned and grouted - Dome completion planned mid 2018
- Network partially in place, internal cabling being done

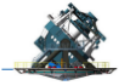
Break Bulk and ≈ 70 containers

$96M of equipment to site in 2018

Identification	Items	Dimensions
		Metres (each)
M1M3 Mirror	1	9.15 x 9.15 x 3.05
M1M3 Primary Cell	1	9.20 x 9.20 x 3.05
VacuLift (containerized)	1	9.20 x 9.20 x 1.00
M1M3 Surrogate Mirror (containerized)	1	9.20 x 9.20 x 1.53
M1M3 Cell Cart (containerized)	1	9.20 x 9.20 x 1.53
Plus estimated 10 FEU		
Coating Plant - To be Disassembled at Coquimbo	1	9.00 x 9.50 x 5.00
Plus estimated 30 FEU		
TMA - Azimuth Ring Sector	4	11.20 x 2.70 x 0.54
TMA - Main Support A & B	2	13.93 x 6.61 x 4.58
TMA - Circular Beam CRA & CRB	2	8.66 x 2.64 x 1.03
TMA - Elevation Brace A & B	2	10.48 x 3.90 x 2.53
TMA - Trunion A & B	2	2.90 x 2.88 x 1.71
TMA - Cradle A & B	2	7.95 x 3.09 x 0.98
TMA - Top End Pier 1 thru 4	4	5.58 x 1.70 x 4.02
TMA - Spider Spindle + TBR + Integrator 1	1	10.20 x 8.70 x 2.80
TMA - Integrator 2 & 3	2	2.61 x 2.61 x 2.51
TMA - M2 Surrogate Mass	1	3.50 x 3.50 x 1.50
TMA - Camera Lifting Fixture	1	7.80 x 3.40 x 2.60
TMA - Cover Trunions	1	3.20 x 3.20 x 1.00
Plus estimated 30 FEU		

High level plan





Potentially lots of data for DM

Data Production Milestone	Completion Date
First calibration data from Auxiliary Telescope	02 Aug 2018
First on-sky and calibration images with ComCam	29 Jan 2020
Sustained scheduler driven observing with ComCam	11 May 2020
Images from Camera re-verification at Summit Facility	16 Jun 2020
First on-sky and calibration data from Camera+Telescope	18 Nov 2020
Sustained scheduler driven observing with Camera+Telescope	08 Feb 2021
Start Science Verification mini-Surveys	30 Mar 2021

DM Challenges

LSST will provde A large (100 PB) database and sophisticated analysis tools: for each of 40 billion objects there will be about 1000 measurements (each with a few dozen measured parameters)

1. Large data volume
2. Large number of objects
3. Highly multi-dimensional space
4. Unknown statistical distributions
5. Time-series data
6. Truncated, censored and missing data
7. Unreliable quantities (e.g. unknown systematics and random errors)

Outline

Introduction project status

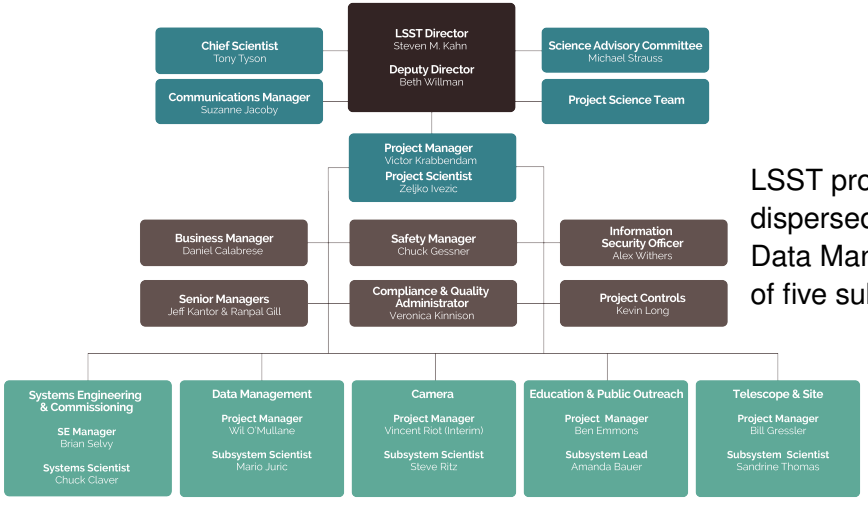
Data Management

Data Access/ Science Platform

DM planning

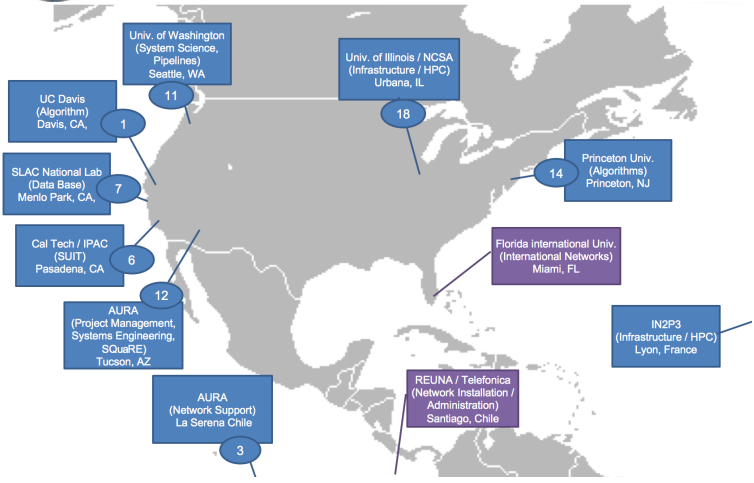
Conclusion

LSST org chart - where DM fits



LSST project is large and dispersed
Data Management is just one of five subsystems.

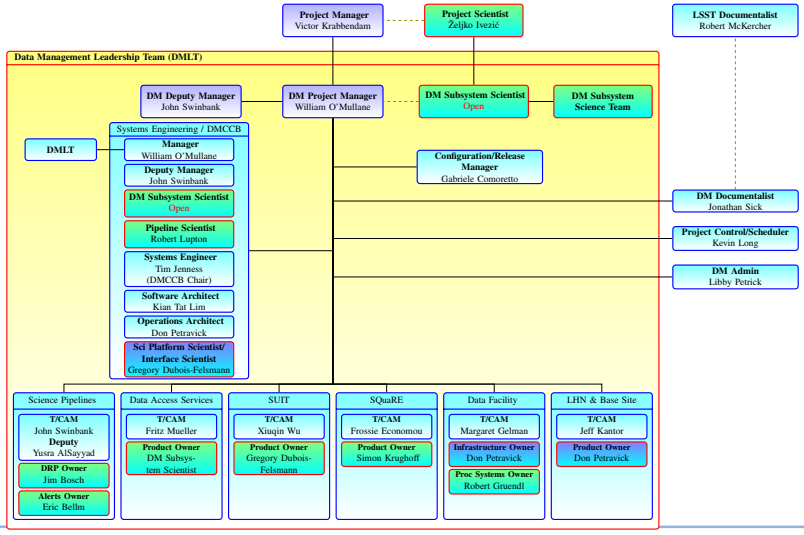
Data management



DM Mission :
Stand up operable, maintainable, quality services to deliver high-quality LSST data products for science, all on time and within reasonable cost.

LSST DM development is distributed across the Americas.
Plus we have partners like IN2P3

DM Organization

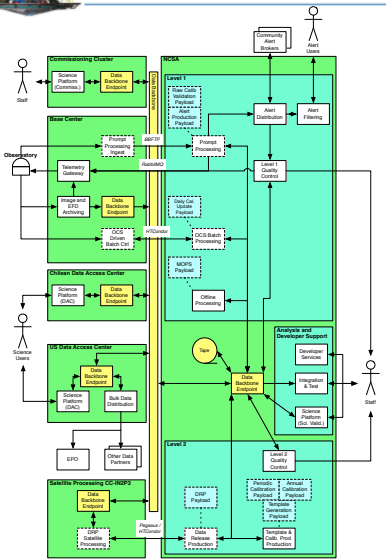


DM leadership meet two times a year and have a weekly telecon.

Technical mangers have a *standup* every Tuesday and Friday.

Toughest thing in any project is communication.

DM build and deploy



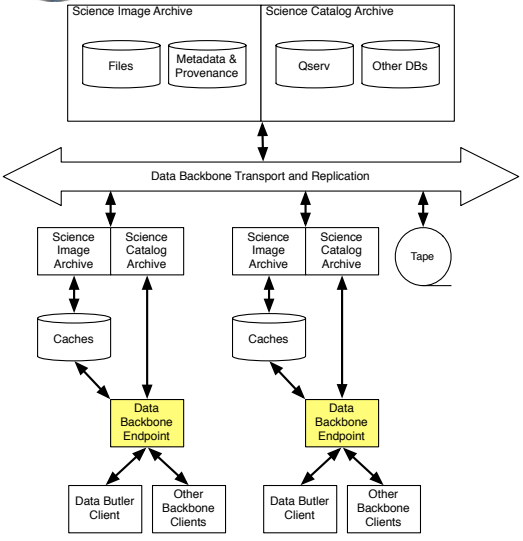
DM must build everything to get LSST products (see `http://ls.st/dpdd`) to the users.

- large data sets (20TB/night)
- complex analysis
- aiming for small systematics
- Science Alerts in under 2 minutes .. (aiming for 1 minute)

About $\frac{1}{2}$ million lines of code (C++/python)

diagram K.T. Lim

Data Backbone

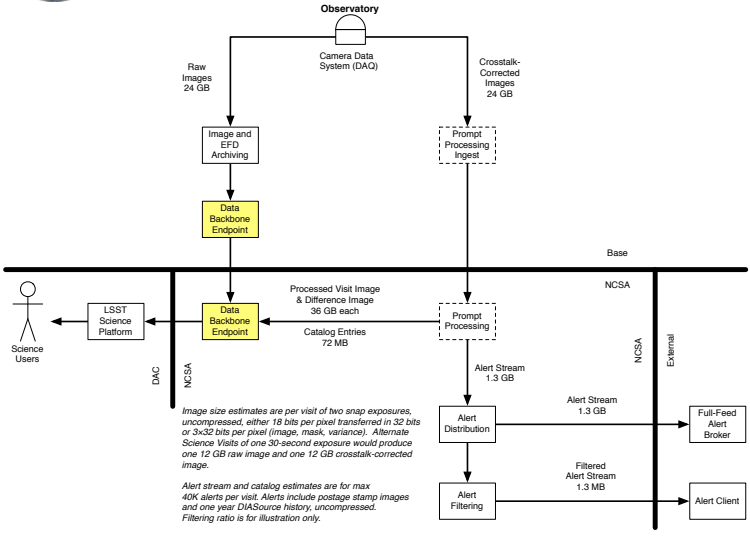


One small box on the previous slide was Data Backbone.
That hides several things

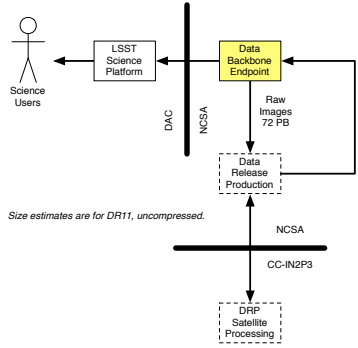
- Qserv - the LSST end user database
- All the networks : we now have fiber to the mountain and from La Serena to NCSA (two routes)

diagram K.T. Lim

Data flow

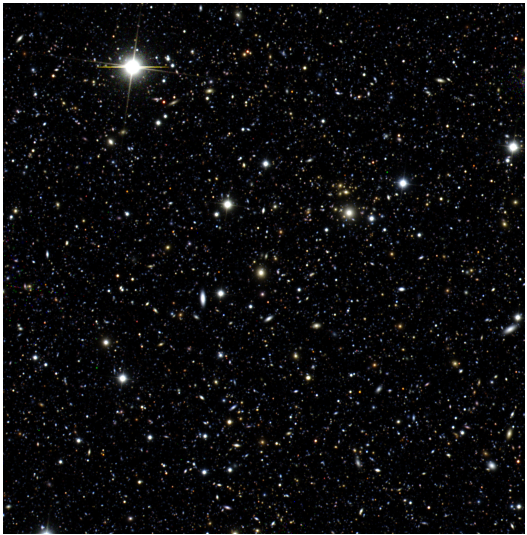


Lots to do every night ..
Plus annually there is a data release



Images from K.T. Lim

All comes down to images . . .



False color from 3 simulated filter images
From just one of 109 CCDs

To work with that there is the *LSST Stack*
`https://pipelines.lsst.io/`
Current version 14.0 release in November 2017.
Python 3 and C++ 14.

It all sits on machines

Prototype Data Access Center Machines at NCSA

- NCSA
  - GPFS 2 PB
  - Common batch Computing - 2304 cores (48×48)
  - use of common NCSA VSphere infrastructure
  - NCSA tape commons (currently in Blue Waters)
  - Fast (100Gbs) links to ESNET,I2,MREN.
- Supporting :
  - Developer spaces and experimentation (Kubernetes), PDAC, etc.
  - LSST Level one test stand (OCS simulator, WAN Emulator, EFD prototype).
- Currently Amazon for builds .
- IN2P3 - full QSERV at least that I know of and more ..

LSST Operations - Distributed

100 - 200 Gbps international links

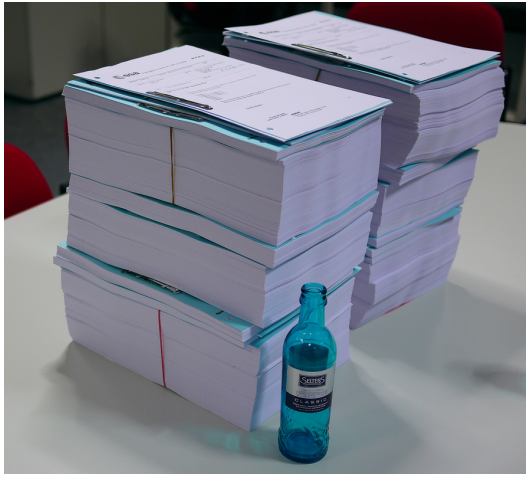
40 - 200 Gbps summit base

See (LSE-78)

Jeff Kantor

Emily Acosta

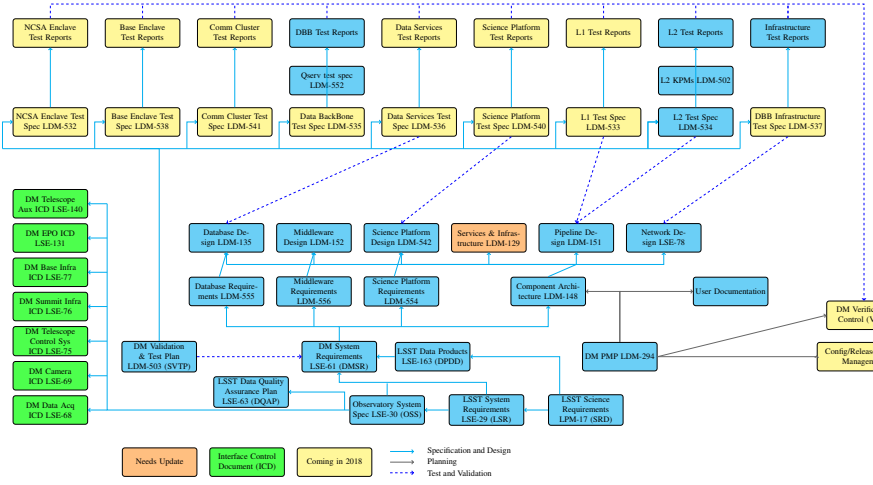
Inevitably we must document ..



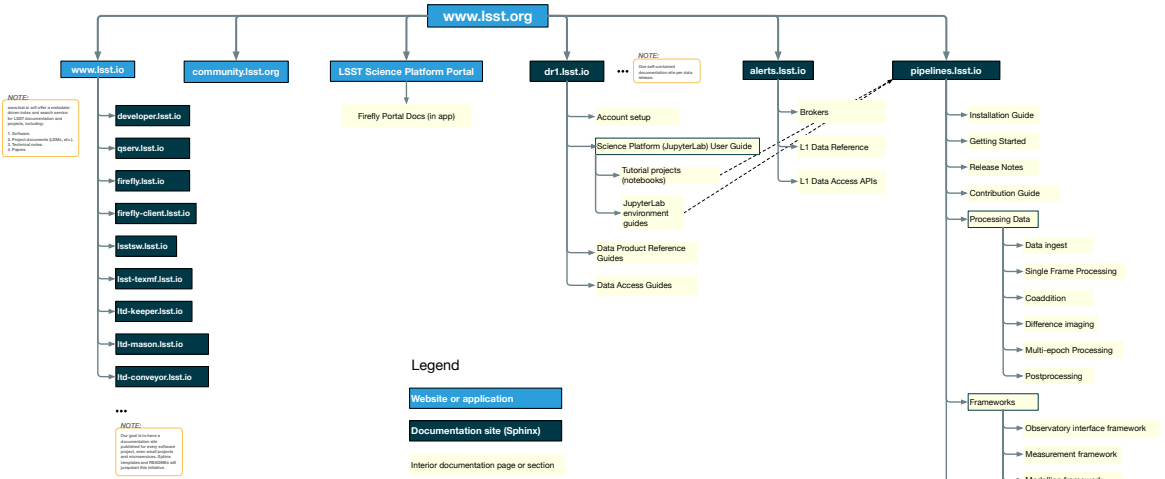
Gaia Flight Operations Procedures (FOP) paper copy in case the computers fail - could be useful!

But we should avoid *write only* documents.

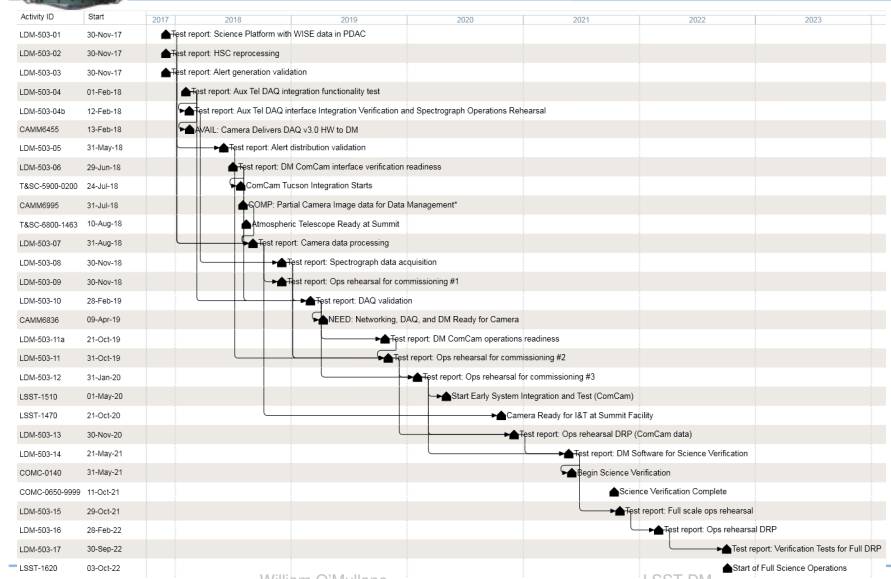
Current DM Doc Tree



Website for users



Big push on verification



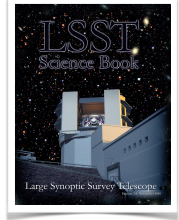
Across all of LSST verification is a big topic right now.

Preparing for LSST: Science Collaborations

- 9 LSST science collaborations, covering the science topics envisioned for LSST:
  - AGN; Dark Energy; Galaxies; Stars, Milky Way and Local Volume; Solar System; Statistics and Informatics; Strong Lensing; Supernovae; Transients/Variable Stars
- Preparing the LSST for the community, and the community for LSST
- More than 400 unique members!

All Contributors with data rights can join !

Initial collaboration work: LSST Science Book
(LSST Science Collaboration, 2009)



Outline

Introduction project status

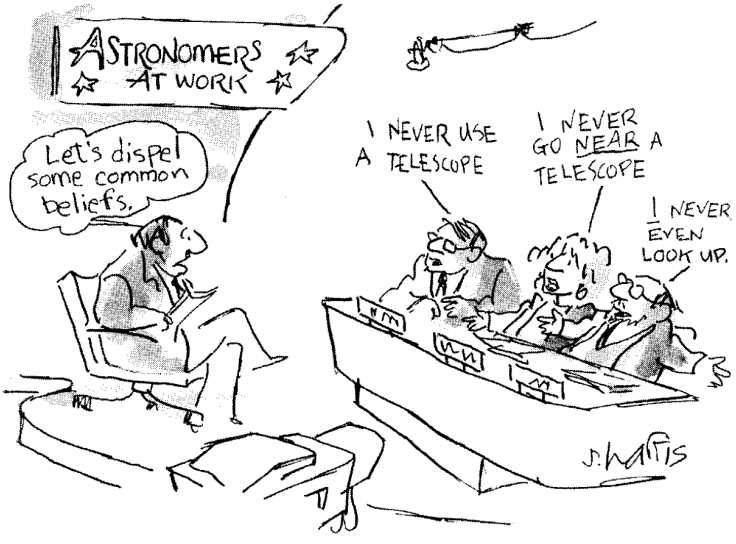
Data Management

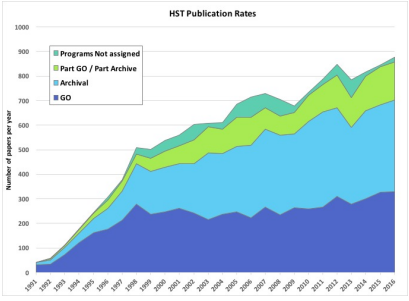
Data Access/ Science Platform

DM planning

Conclusion

The Era Of Surveys



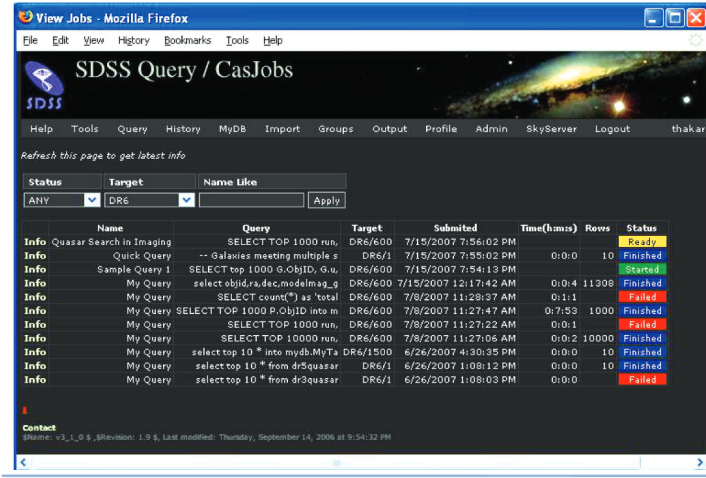


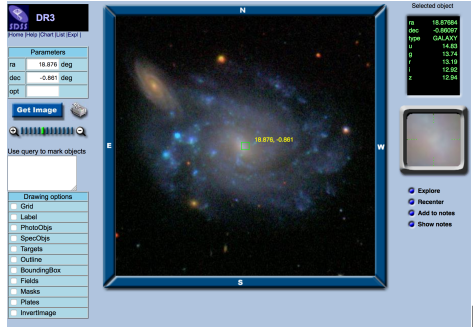
`https://archive.stsci.edu/hst/bibliography/pubstat.html`

. . . indicates archival research probably play an important role in the scientific success of XMM-Newton Ness et al. (2014)

First serious go at code to data

Early in the millennium - CasJobs (C#)(OMullane et al., 2005)

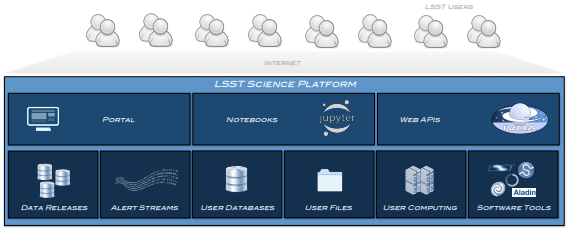




N evolved

http://www.sciserver.org/

LSST Science Platform



For DR2:

- Computing:2,400 cores ($\approx$ 18 TFLOPs)
- File storage: $\approx$ 4PB (VOSpace)
- Database storage: $\approx$ 3PB (MYDB)

The Science Platform has three user facing aspects: the Portal (novice), the JupyterLab (intermediate), and the Web APIs (expert and remote tools).

We enable access to the Data Releases and Alert Streams, and support next-to-the data analysis and Level 3 product creation using the computing resources available at the Data Access Center (DAC).

Mario Juric

Concept has been around a year or two - name suggested in Nov 16

PDAC - Portal, Notebook

Add a URL for downloading the image by ID

```
In [16]: df_coadds['img_url'] = df_coadds.deepCoaddId.map(lambda x:
                'http://lsst-qserv-dax01.ncsa.illinois.edu:5000/image/v0/deepCoadd/id?id=' + str(x))
```

Verify that `df_coadds` has at least five entries for the five filters

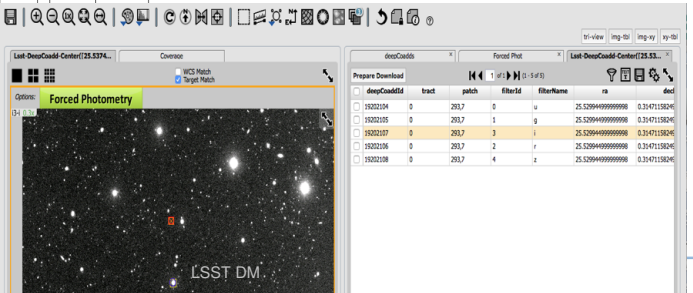
```
In [15]: df_coadds
```

Out[15]:

	deepCoaddId	tract	patch	filterId	filterName	ra	decl	htmId20	equinox	raDeSys	...	corner3Ra	corner3Decl	corner4Ra
0	19202104	0	293,7	0	u	25.529945	0.314712	17147177442654	2000.0	ICRS	...	25.416645	0.422899	25.416645
1	19202105	0	293,7	1	g	25.529945	0.314712	17147177442654	2000.0	ICRS	...	25.416645	0.422899	25.416645
3	19202106	0	293,7	2	r	25.529945	0.314712	17147177442654	2000.0	ICRS	...	25.416645	0.422899	25.416645
2	19202107	0	293,7	3	i	25.529945	0.314712	17147177442654	2000.0					
4	19202108	0	293,7	4	z	25.529945	0.314712	17147177442654	2000.0					

5 rows × 34 columns

- Results available in portal and notebook
- Notebook can drive portal
- Forced Photometry
- NeoWise and Stripe82 currently loaded

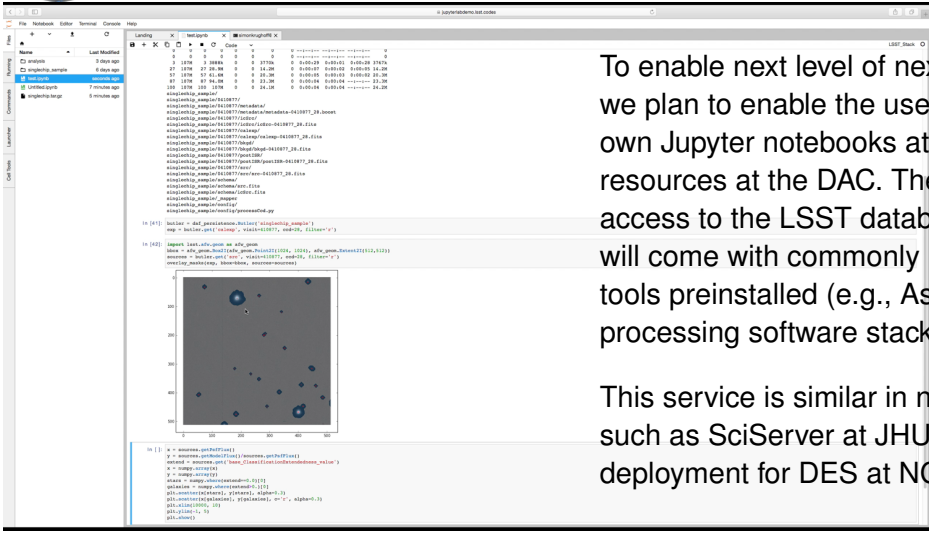


Chilean Data Access Centre

- Similar to US DAC with Science Platform installed but hosted in La Serena
- Chilean Grid authentication to be integrated with the LSST system (LSE-279)
- about 10% of DAC resources
  - Computing: 240 cores ($\approx$ 1.8 TFLOPs)
  - File storage: $\approx$ 400 TB (VOSpace or equivalent)
  - Database storage: $\approx$ 300 TB (user database workspace)
- Plus raw data and data products

See (LDM-572)

JupyterLab Aspect



To enable 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.

APIs and Web Services

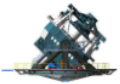
- Web APIs: VO standard protocols where possible; custom extensions where necessary; proposed back to IVOA when applicable
  - Catalog & other tabular data (dbserv) (TAP 1.1, ADQL 2.1, CAOM2)
  - Image data (imgserv) (SIA V2)
  - Metadata (metaserv)
- Python objects (Data Butler) - abstract access in pipelines
- SQL database (Qserv) - MPP Shared noting data base built at Stanford.

JupyterLab demo

- Full stack (choice of versions) no installation needed
- All source available `https://github.com/lsst-sqre/jupyterlabdemo`
- Deploy on your own K8s cluster `https://github.com/lsst-sqre/jupyterlabdemo/tree/master/tools/deployment`
- Versions with Stack 14 available for trial during the workshop `https://jupyter.nlhpc.cl/lsst/hub/home`
- Friday morning we can discuss more - but the time then is short for the *tutorial.ipynb* so start before if you can.

Evolution of the platform

- Documentation
  - The direction is now documented in LSE-319
  - More formal requirements in LDM-554
  - Design is in LDM-542
  - First user reports in DMTR-22
- Coming soon
  - Portal focus on exploration/discovery of data products
  - full sky browsing e.g. something like HiPS
    `http://www.ivoa.net/documents/HiPS/`
  - Authentication integrated with existing credentials
  - Quotas on CPU and Disk .. has to be done..
  - Data Upload to user space
  - Scheme to include collaborator code in Stack (thus in platform)
    - proper way to pull any module or code into notebook kernel from say github
- To be used in commissioning by the commissioning team
  - Boot camps soon



Outline

High Level Goals

- 2017: Prototyping data access and first access to hardware
  - Dec: Prototype notebooks, private databases for Science Platform
  - Dec: Robust coadd available in Stack
- 2018: Prototypes for various processes and databases - “Minimum Viable System”
  - Jan: Jointcal replace meas_mosaic
  - Jun: At-scale test of Alert Distribution
  - Aug: Mountain base network up
  - Oct: Spectrograph data acquisition
  - Dec: Prototype QA/Commissioning Environment
- Dec 2019: ComCam L1, L2 Production
- Dec 2019: Base Center Integration Complete
- Jun 2020: Camera L1, L2 Production
- Jul 2021: US Data Access Center Integrated
- Jun 2022: Chilean Data Access Center Integrated (LDM-572)

Test plans to confirm milestone completion are under development.

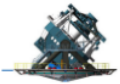
Commissioning Start Requirements

November 2019: DM for Commissioning (minimum required for start of commissioning with ComCam): **(See LSE-79 §3.3 and table 8)**

- Pipeline: single-frame measurement including ISR, ghost masking, cosmic ray detection, PSF estimation, astrometric and photometric calibration, background estimation, single-frame deblending, master calibration image generation, atmospheric characterization
- Services: archiving, EFD transformation, Data Backbone for files (Base/NCSA), telemetry gateway, OCS-controlled batch, offline processing
- LSST Science Platform on Commissioning Cluster: Notebook Aspect, image access, user file storage, batch computing

Milestones:

- LDM-503-9 – 2018-11-30: on the right track with beta software.
- LDM-503-11 – 2019-10-31: verification of ComCam commissioning requirements.



Outline

Introduction project status

Data Management

Data Access/ Science Platform

DM planning

Conclusion

Conclusion

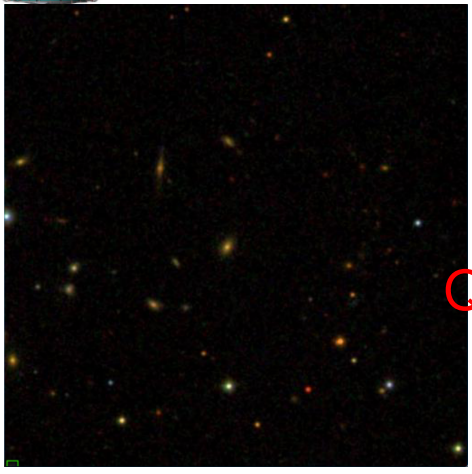
- Current astronomical surveys are changing the way we do astronomy
- LSST will soon follow perhaps ushering a complete change in how we approach end user data interaction
- LSST is on track and data is coming sooner than you might think
- There are plenty of challenges
- Verification and Validation on radar for now
- Looking forward to the first LSST images !

Motto for the future from LSST Project Scientist:

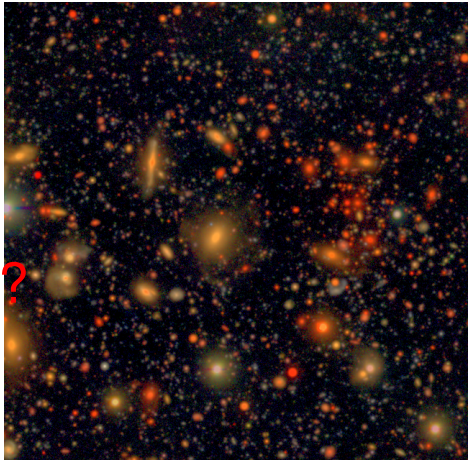
Ask Not What Data You Need To Do Your Science,
Ask What Science You Can Do With Your Data.

Željko Ivezić

The END



Questions?



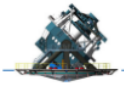
$\sim 3.5'$ SDSS image

HSC image (COSMOS)

g,r(1.5 hrs) ,i(3 hrs) PSF matched co-add (≈ 27.5)

Images:Lupton and HSC colaboration see also Lupton et al. (2004)

http://www.lsst.org http://community.lsst.org



Outline

Reference material

Acronyms I

Acronym	Description
ADQL	Astronomical Data Query Language
API	Application Programming Interface
CCD	Charge-Coupled Device
CPU	Central Processing Unit
DAC	Data Access Center
DM	Data Management
DOE	Department of Energy
FOP	Flight Operation Procedure (Plan)
ITC	Information Technology Center
IVOA	International Virtual-Observatory Alliance
JHU	Johns Hopkins University
LSST	Large Synoptic Survey Telescope
MPP	Massive Parallel Processing
NCOA	National Center for Optical and infrared Astronomy
NCSA	National Center for Supercomputing Applications
NSF	National Science Foundation
PB	PetaByte
PDAC	Prototype Data Access Center
PSF	Point Spread Function
SDSS	Sloan Digital Sky Survey
SIA	Sagnac Interferometer Assembly

Acronyms II

SLAC	Stanford Linear Accelerator Center
SQL	Structured Query Language
TAP	Table Access Protocol
TFLOP	Tera FLOP
VO	Virtual Observatory

References

[LDM-554], Ciardi, D., Dubois-Felsmann, G., 2017, *Science Platform Requirements*, LDM-554, URL https://ls.st/LDM-554

[LSE-79], Claver, C., The LSST Commissioning Planning Team, 2017, *System AI&T and Commissioning Plan*, LSE-79, URL https://ls.st/LSE-79

[LDM-542], Dubois-Felsmann, G., Lim, K.T., Wu, X., et al., 2017, *LSST Science Platform Design*, LDM-542, URL https://ls.st/LDM-542

[LSE-319], Jurić, M., Ciardi, D., Dubois-Felsmann, G., 2017, *LSST Science Platform Vision Document*, LSE-319, URL https://ls.st/LSE-319

[LSE-78], Lambert, R., Kantor, J., Huffer, M., et al., 2017, *LSST Observatory Network Design*, LSE-78, URL https://ls.st/LSE-78

LSST Science Collaboration, 2009, ArXiv e-prints (arXiv:0912.0201), ADS Link

Lupton, R., Blanton, M.R., Fekete, G., et al., 2004, Publications of the ASP, 116, 133 (arXiv:astro-ph/0312483), doi:10.1086/382245, ADS Link

Ness, J.U., Parmar, A.N., Valencic, L.A., et al., 2014, Astronomische Nachrichten, 335, 210 (arXiv:1311.5751), doi:10.1002/asna.201312001, ADS Link

[LDM-572], O'Mullane, W., Petravick, D., 2017, *Chilean Data Access Center*, LDM-572, URL https://ls.st/LDM-572

OMullane, W., Li, N., Nieto-Santisteban, M., et al., 2005, *Batch is back: CasJobs, serving multi-TB data on the Web*, Tech. rep., Microsoft, Microsoft Technical Report MSR TR 2005 19 (arXiv:cs/0502072), ADS Link

[DMTR-22], Suberlak, K., Ivezić, Ž., The PDAC Team, 2017, *Prototype Data Access Center: User Report*, DMTR-22, URL https://ls.st/DMTR-22

[LSE-279], Withers, A., 2017, *Concept of Operations for Unified LSST Authentication and Authorization Services*, LSE-279, URL https://ls.st/LSE-279