IPMU APEC Seminar, February 21st 2019.

Towards New Science with the LSST

Envisioning Platforms for Science with Large Datasets

With Eric Bellm, Robert Lupton, Zeljko Ivezic, Andy Connolly, Colin Slater, and many, many, many other colleagues in the LSST Project!

Prof. Mario Juric
DIRAC Institute | eScience Institute | UW Astronomy

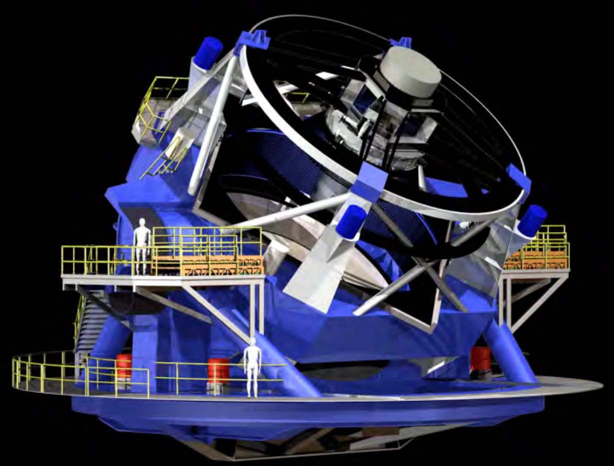
Data Intensive Research in Astrophysics and Cosmology
COLLEGE OF ARTS & SCIENCES | UNIVERSITY of WASHINGTON

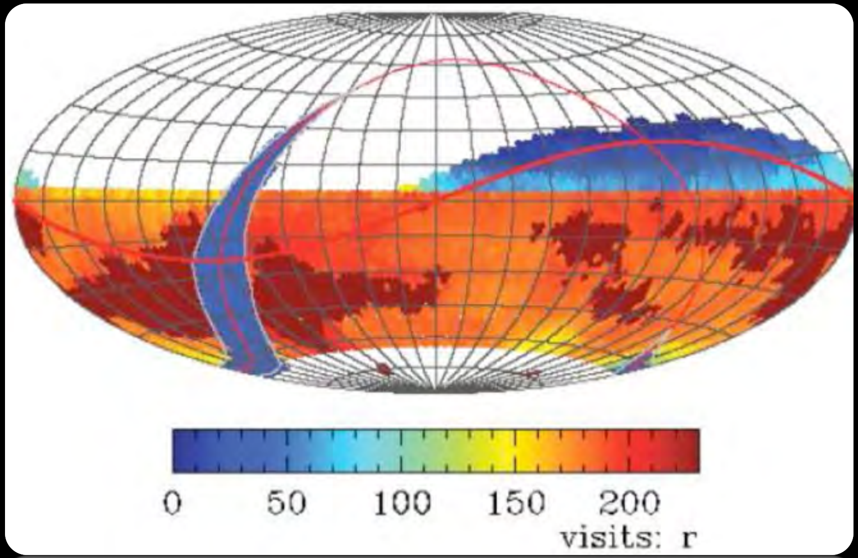
Introducing the LSST

LSST: A Deep, Wide, Fast, Optical Sky Survey

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

ugrizy 0.5-1% photometry



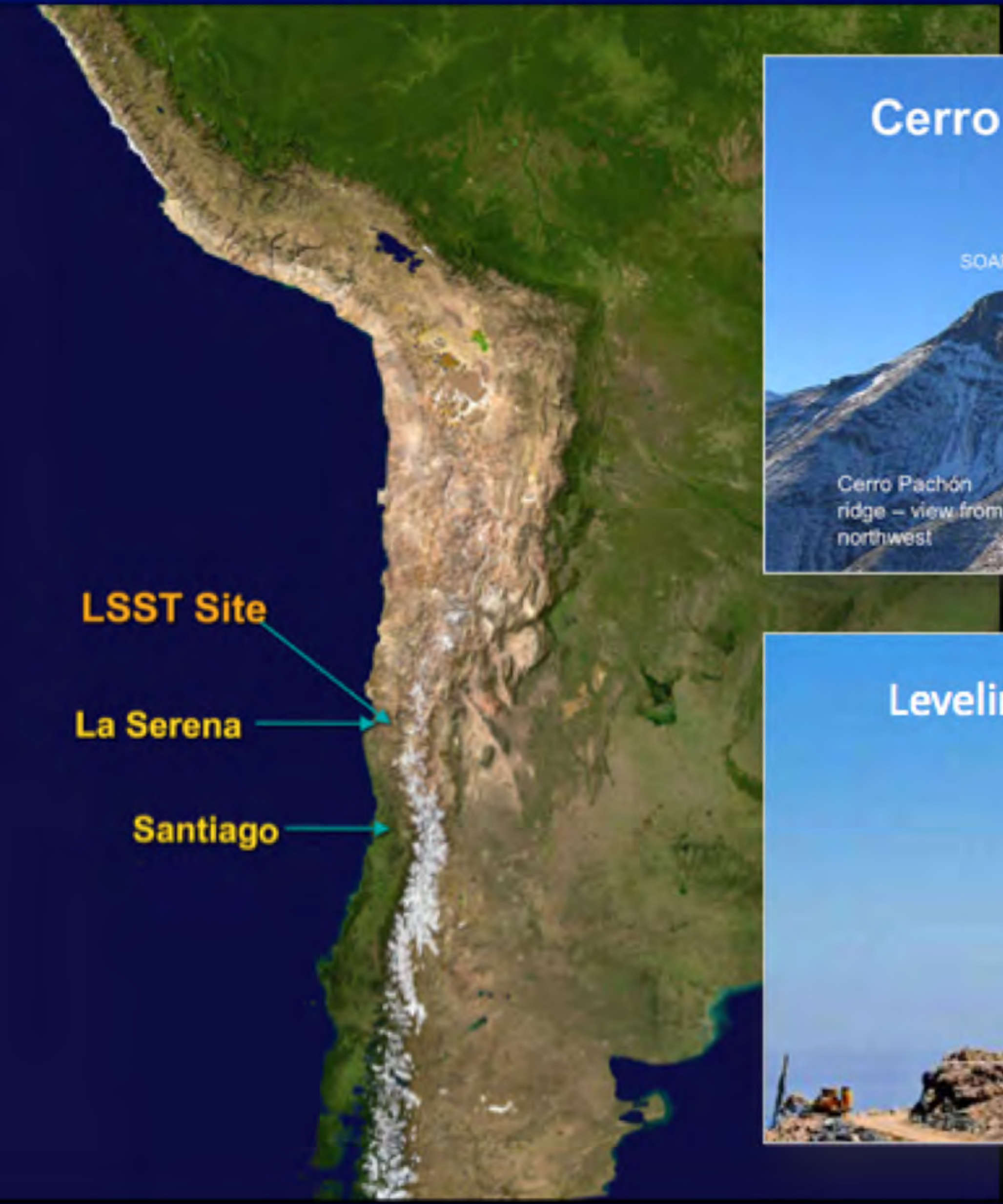


3.2Gpix camera 30sec exp/4sec rd 15TB/night 37 B objects

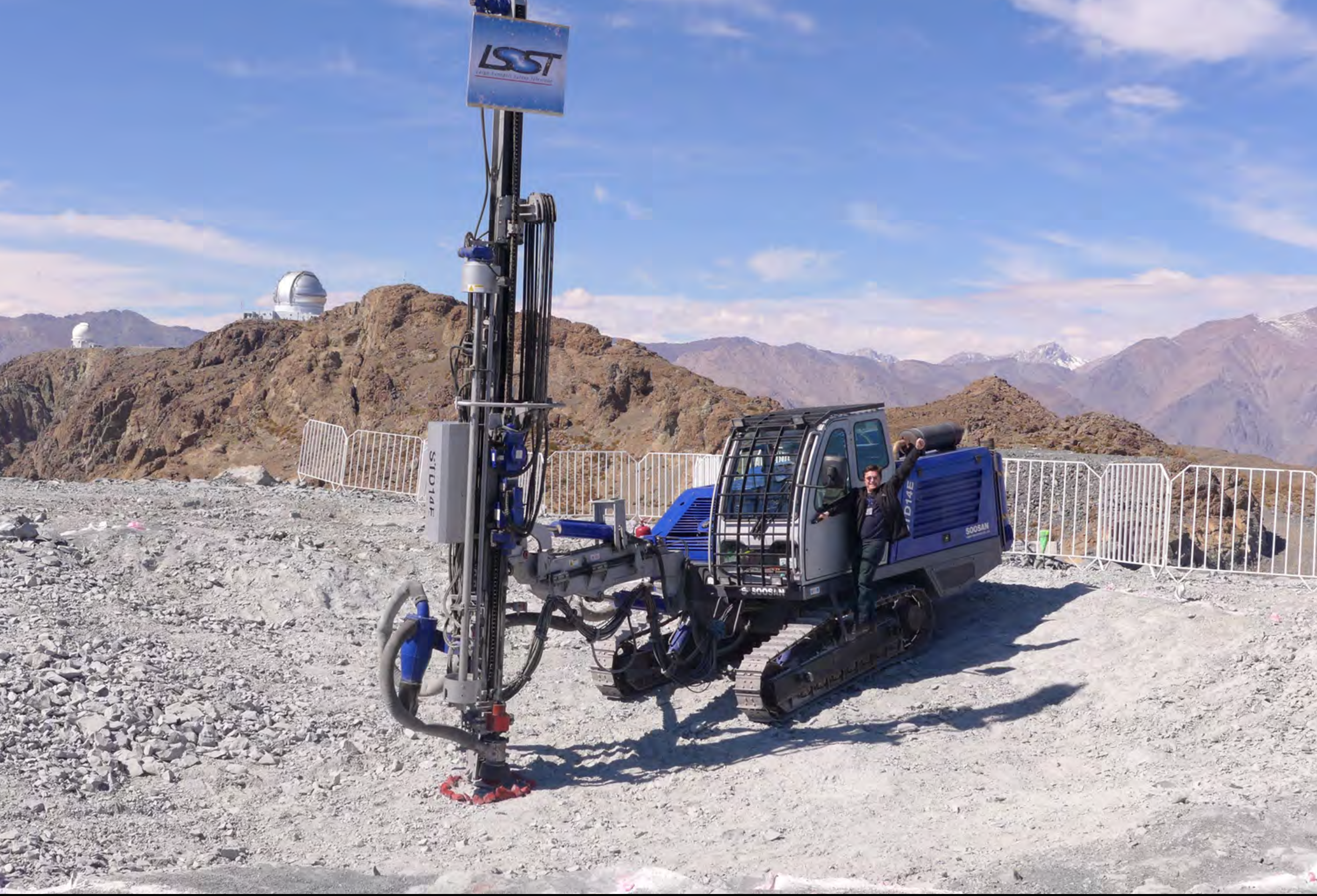
Imaging the visible sky, once every 3 days, for 10 years (825 revisits)

Alerts public, other data proprietary for two years, delivered in bulk afterwards.

Location: Cerro Pachon, Chile

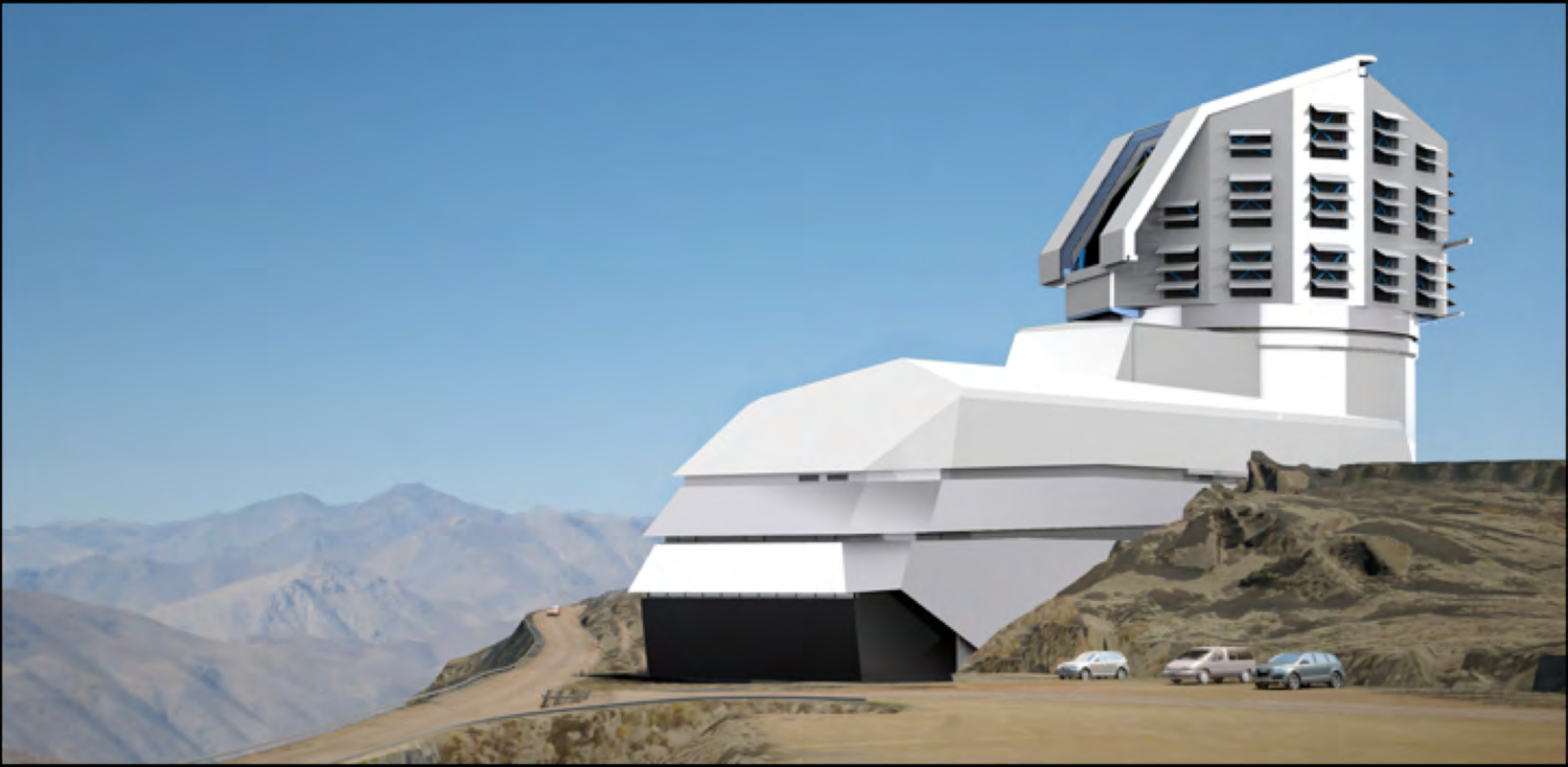


First Stone Ceremony, April 14th 2015



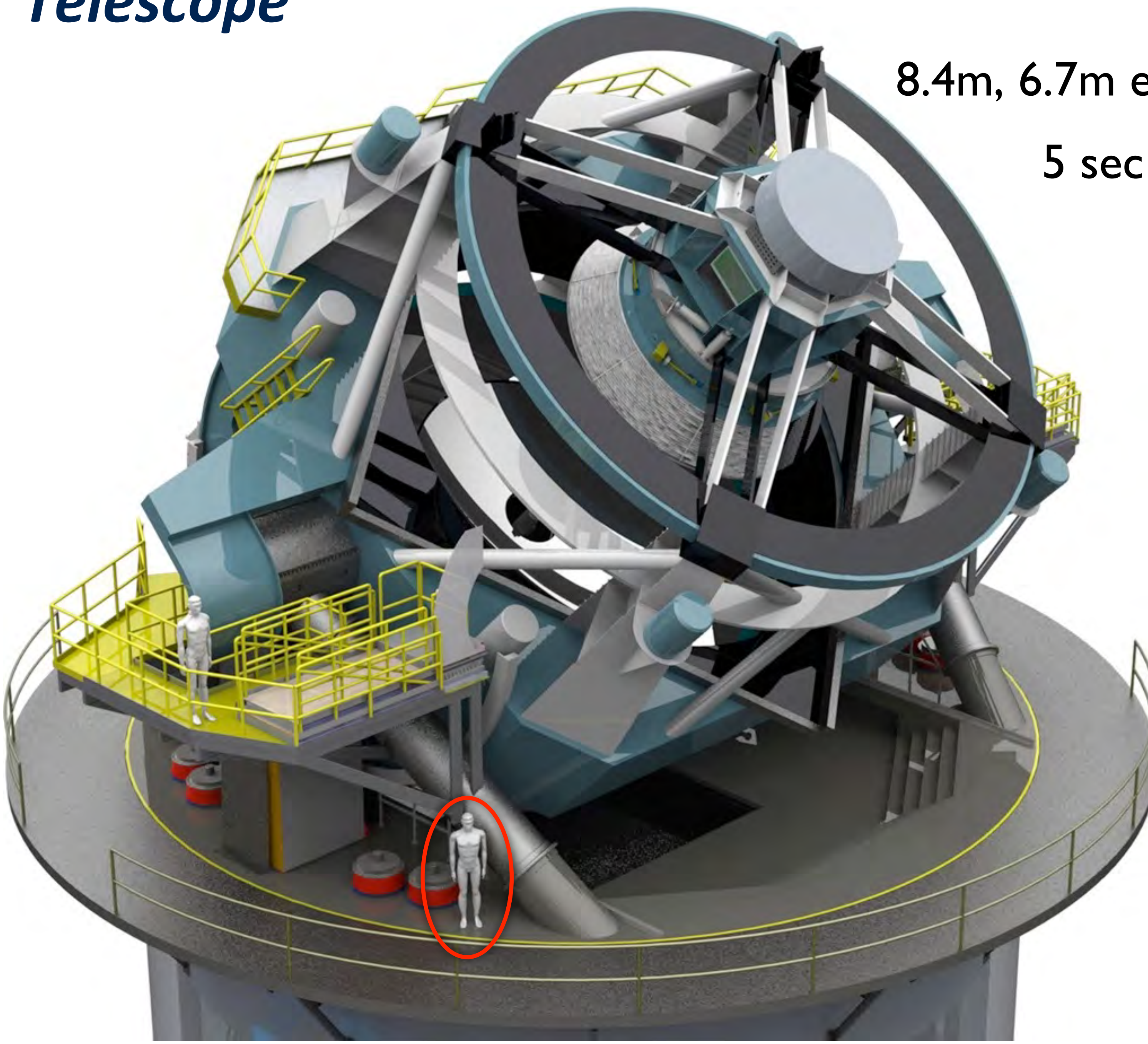
LSST Site, April 14th 2015

LSST Site, May 2018



First Light: 2020

LSST Telescope

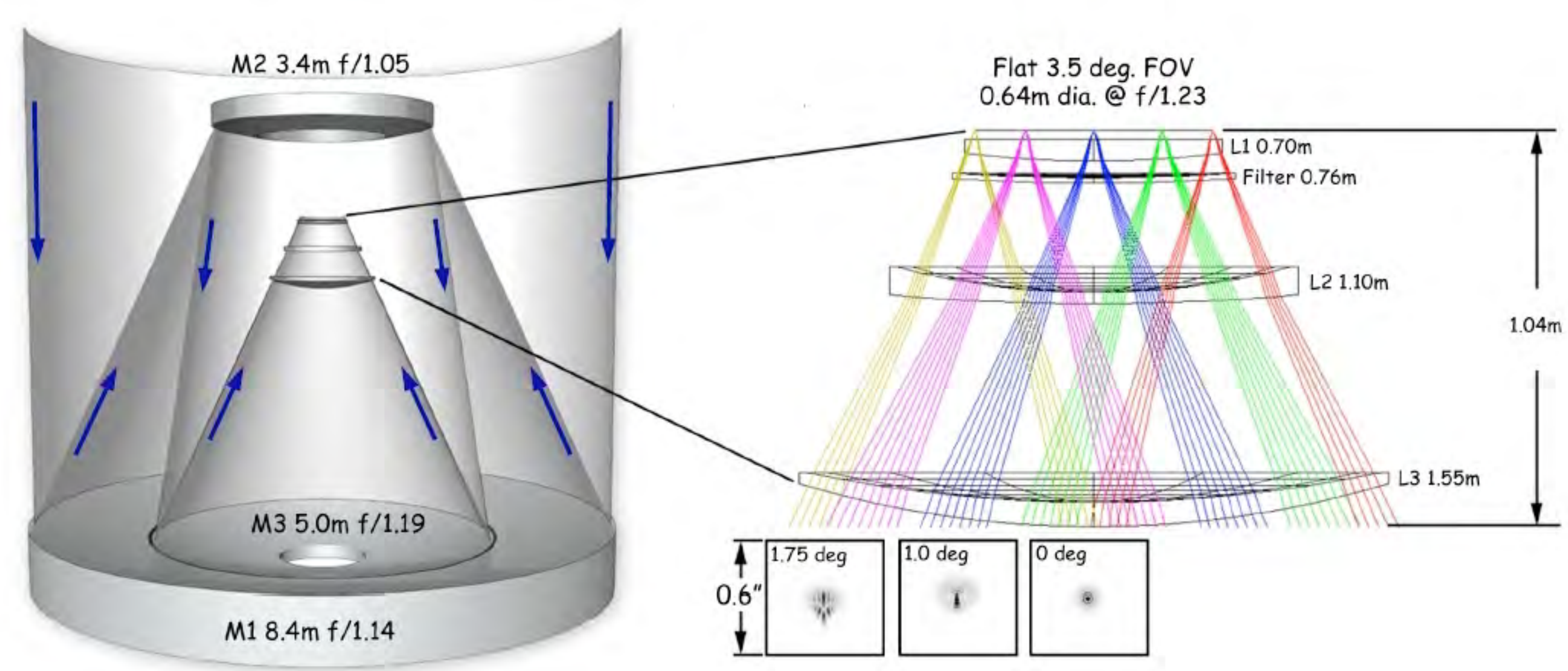


8.4m, 6.7m effective

5 sec slew & settle

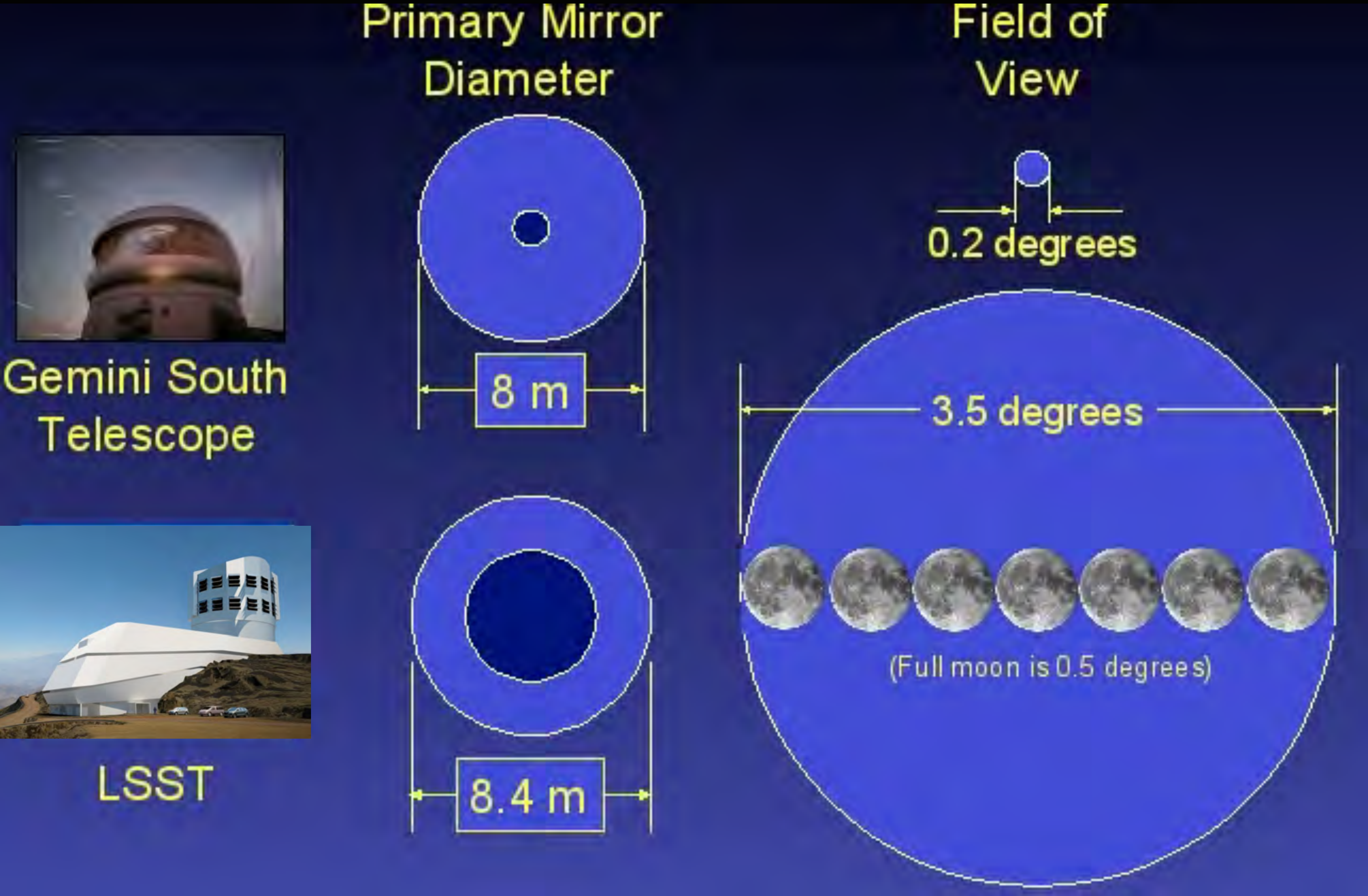
Telescope and Mount Assembly @ Asturfeito in Spain

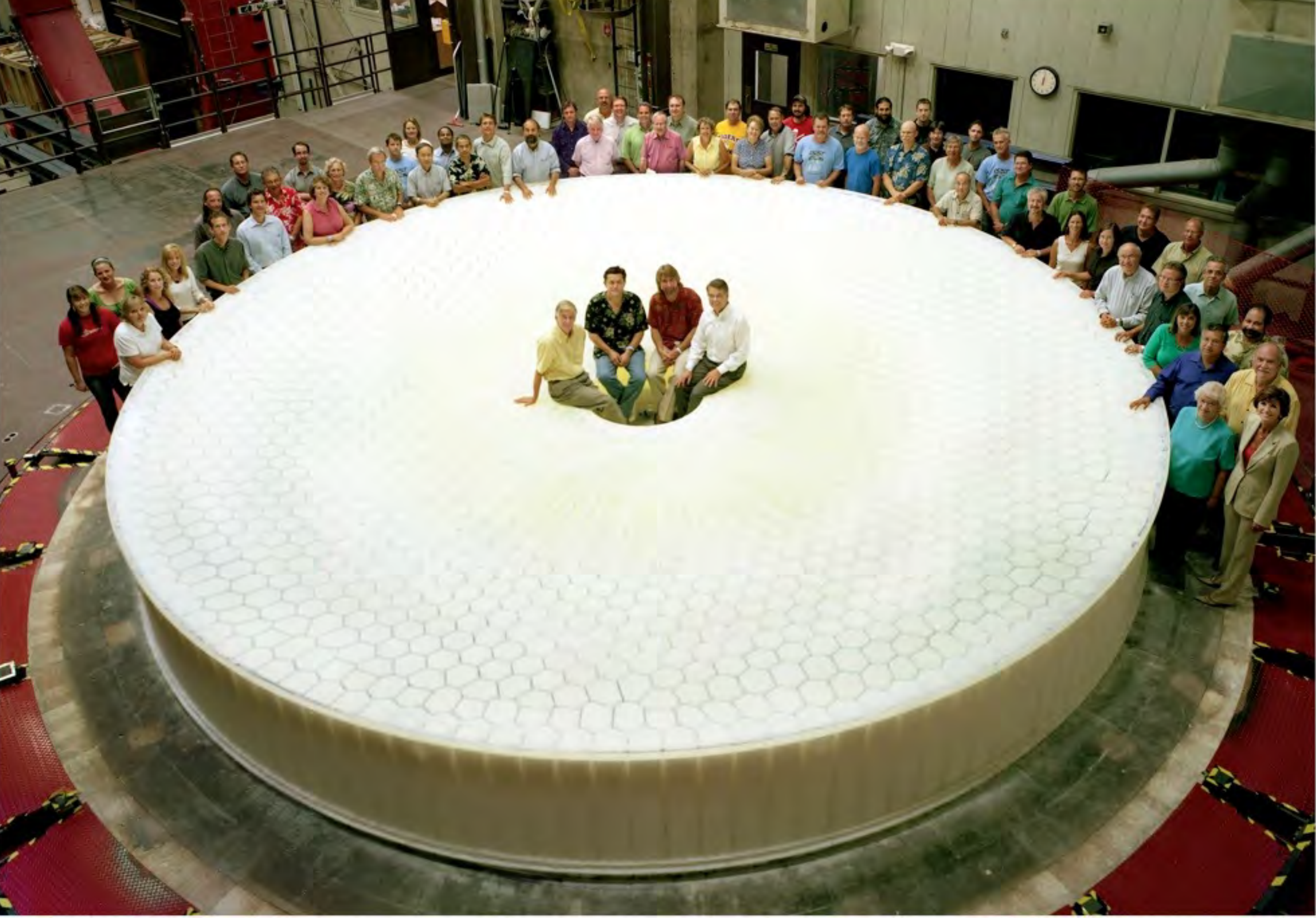
Optical Design



Three-mirror design (Paul-Baker system)
enables large field of view with excellent image quality:
delivered image quality is dominated by atmospheric seeing

The field-of-view comparison: Gemini vs. LSST



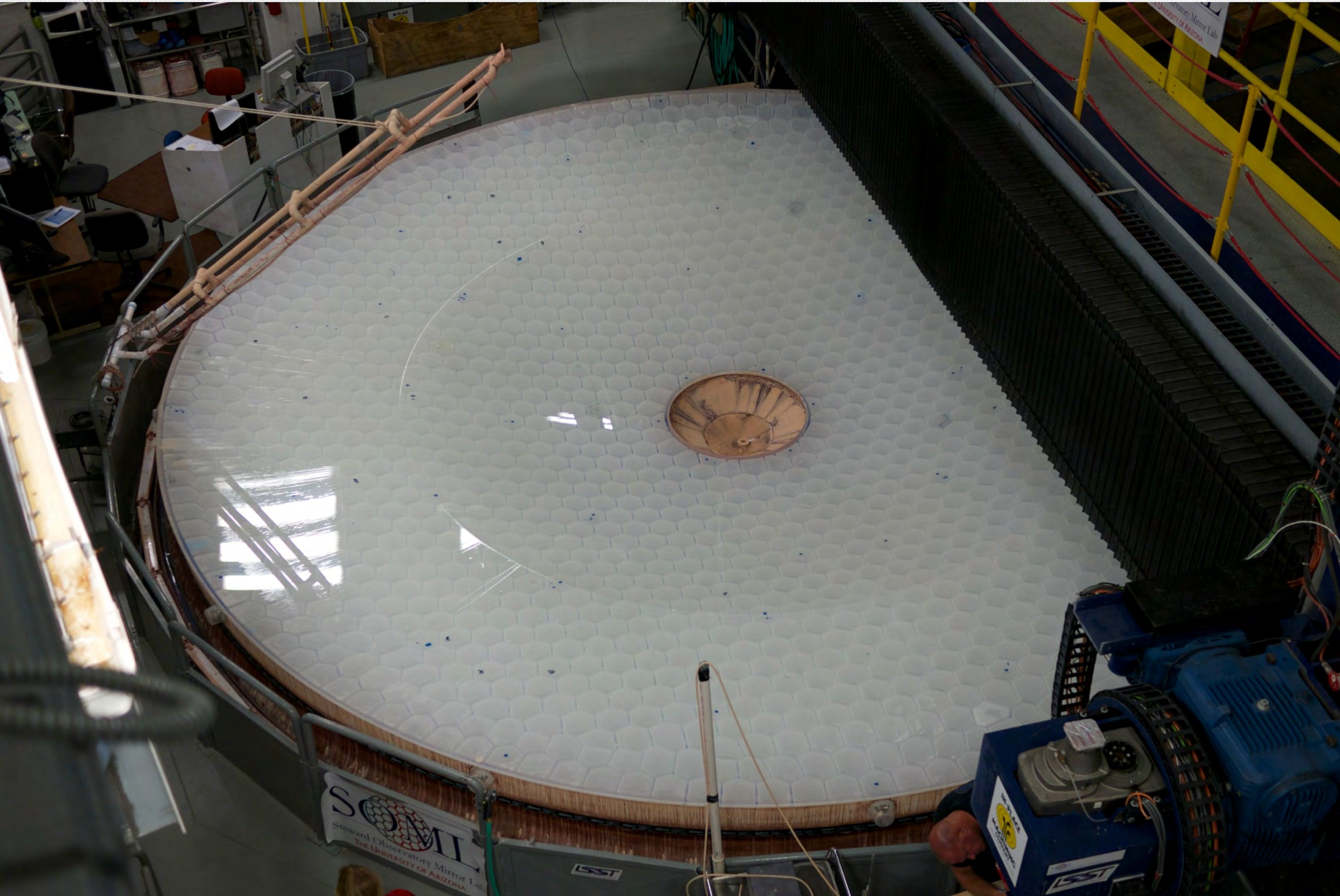


THE UNIVERSITY OF ARIZONA. TUCSON ARIZONA SOML

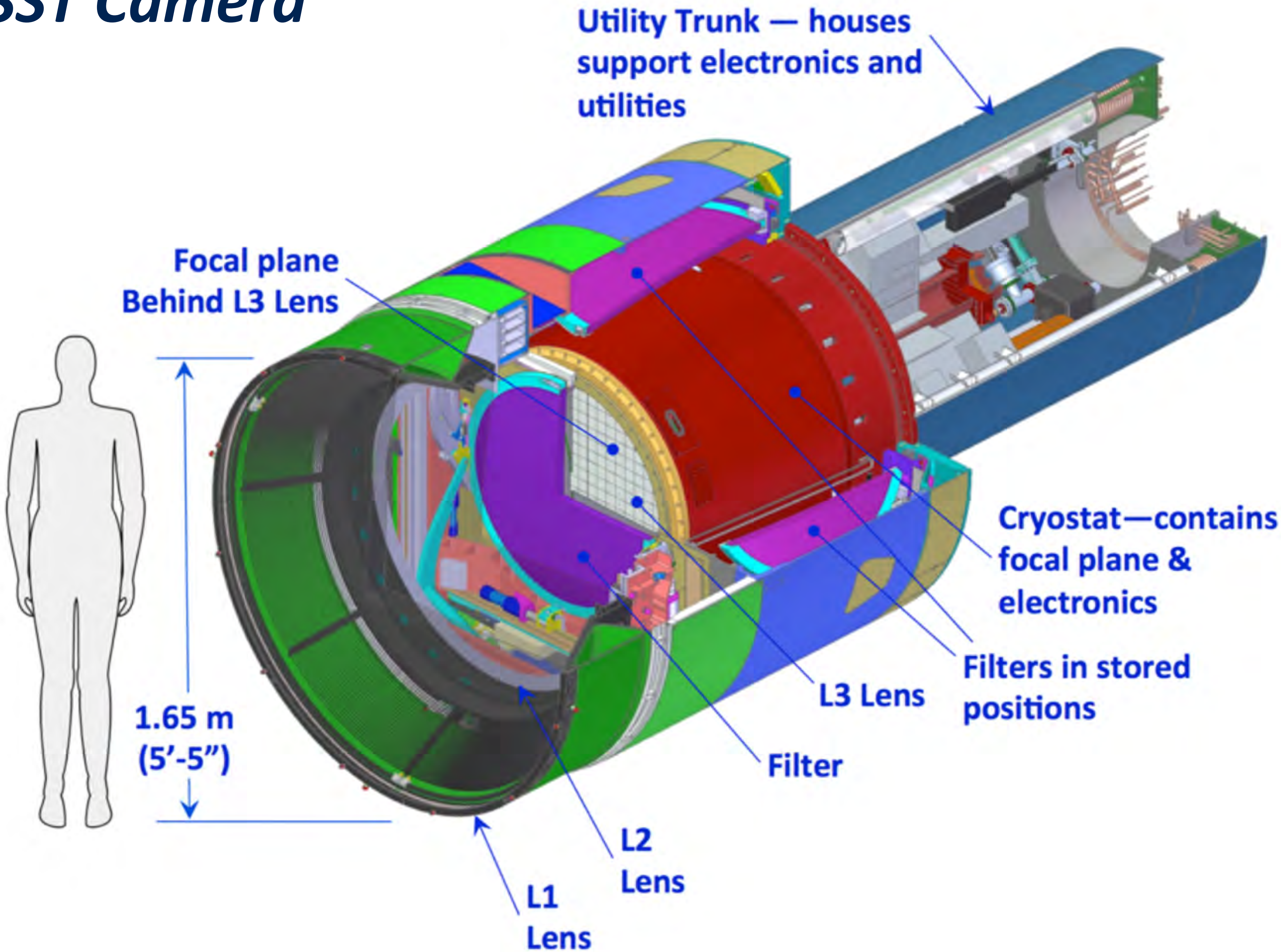
LSST Primary/Tertiary Mirror Blank
August 11, 2008, Steward Observatory Mirror Lab, Tucson, Arizona

LSST Large Synoptic Survey Telescope

Done!


SOML
Steward Observatory Mirror Lab
THE UNIVERSITY OF ARIZONA
LSST

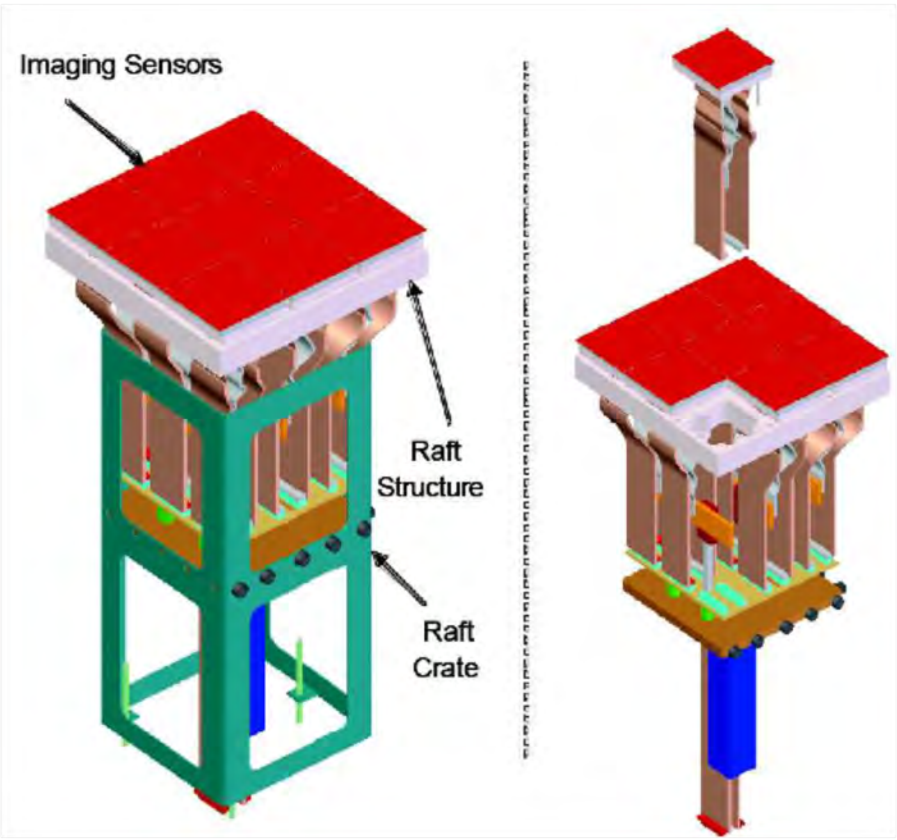
LSST Camera

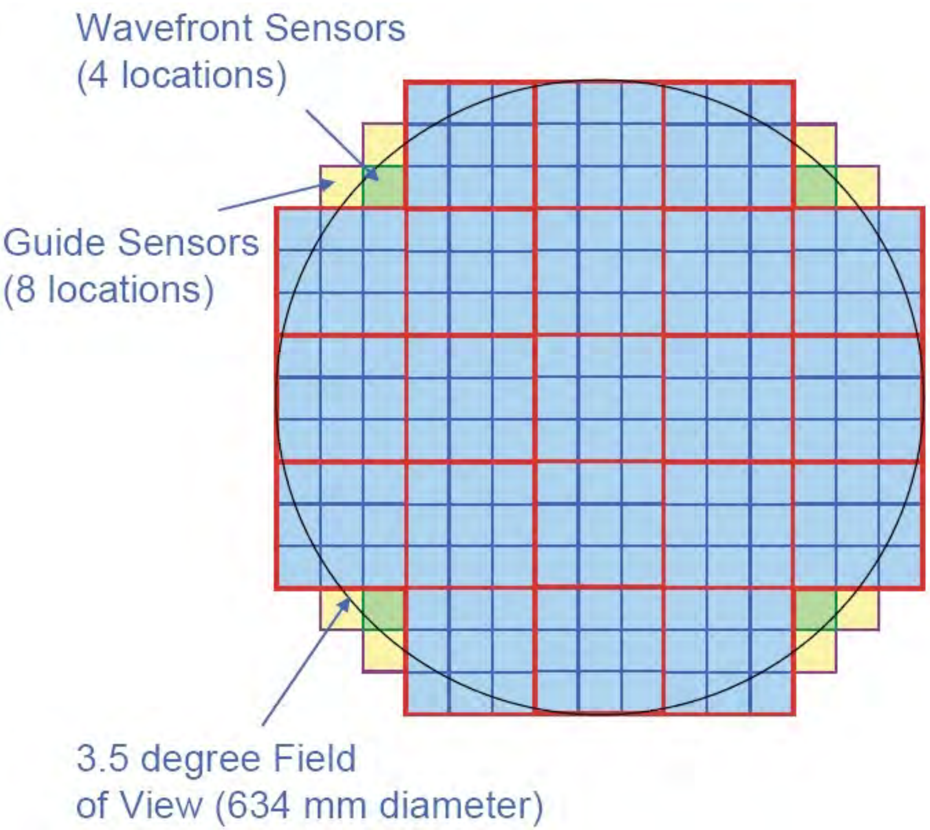


A very large (and massive) camera: 2800 kg, 3.2 Gpix

LSST Camera

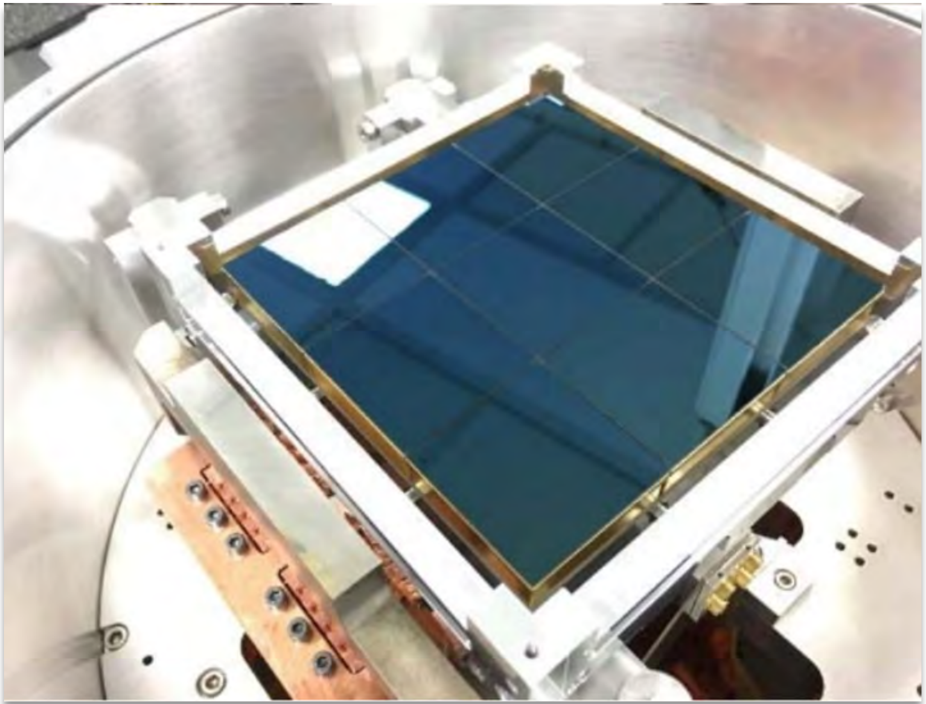
189 CCDs with 16Mpix each, assembled in 3x3 "raft" modules

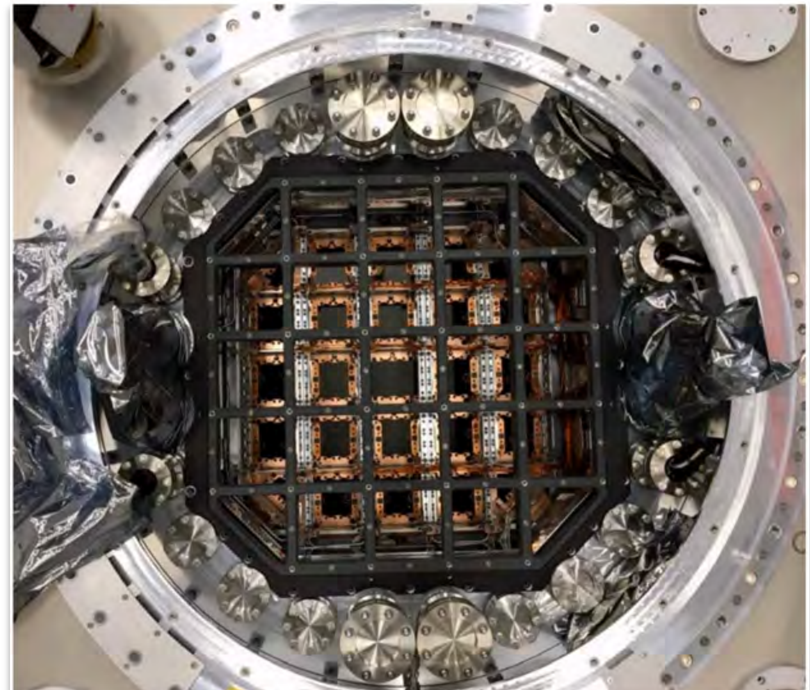




4k x 4k CCDs   3 x 3 CCD "rafts"   21 Raft Camera

Almost there!





- 264 Science and Science Reserve Sensors delivered
- Need 208

- 12 Rafts SLAC Accepted
- 1.6 Gpixels Ready!
- Many in progress
- Need: 21 Science Rafts and 4 corner Rafts

SDSS vs. LSST comparison: LSST=d(SDSS)/dt, LSST=SuperSDSS

3 arcmin is 1/10 of the full Moon's diameter

(almost) like LSST depth (but tiny area)

20x20 arcsec; lensed SDSS quasar (SDSS J1332+0347, Morokuma et al. 2007)

LSST Operations: Sites and Data Flows

Satellite Processing Center

(CC-IN2P3, Lyon, France)

Data Release Production (50%)

CCIN2P3

Archive Site

NCSA

Archive Center

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

Data Access Center

Data Access and User Services

Chilean Sites

Telescope and Camera
Data Acquisition
Crosstalk Correction
Long-term storage (copy 1)
Chilean DAC Entry-point

HQ Site

Science Operations
Observatory Management
Education and Public Outreach

LSST Data Products

- 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.

Prompt

- 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.

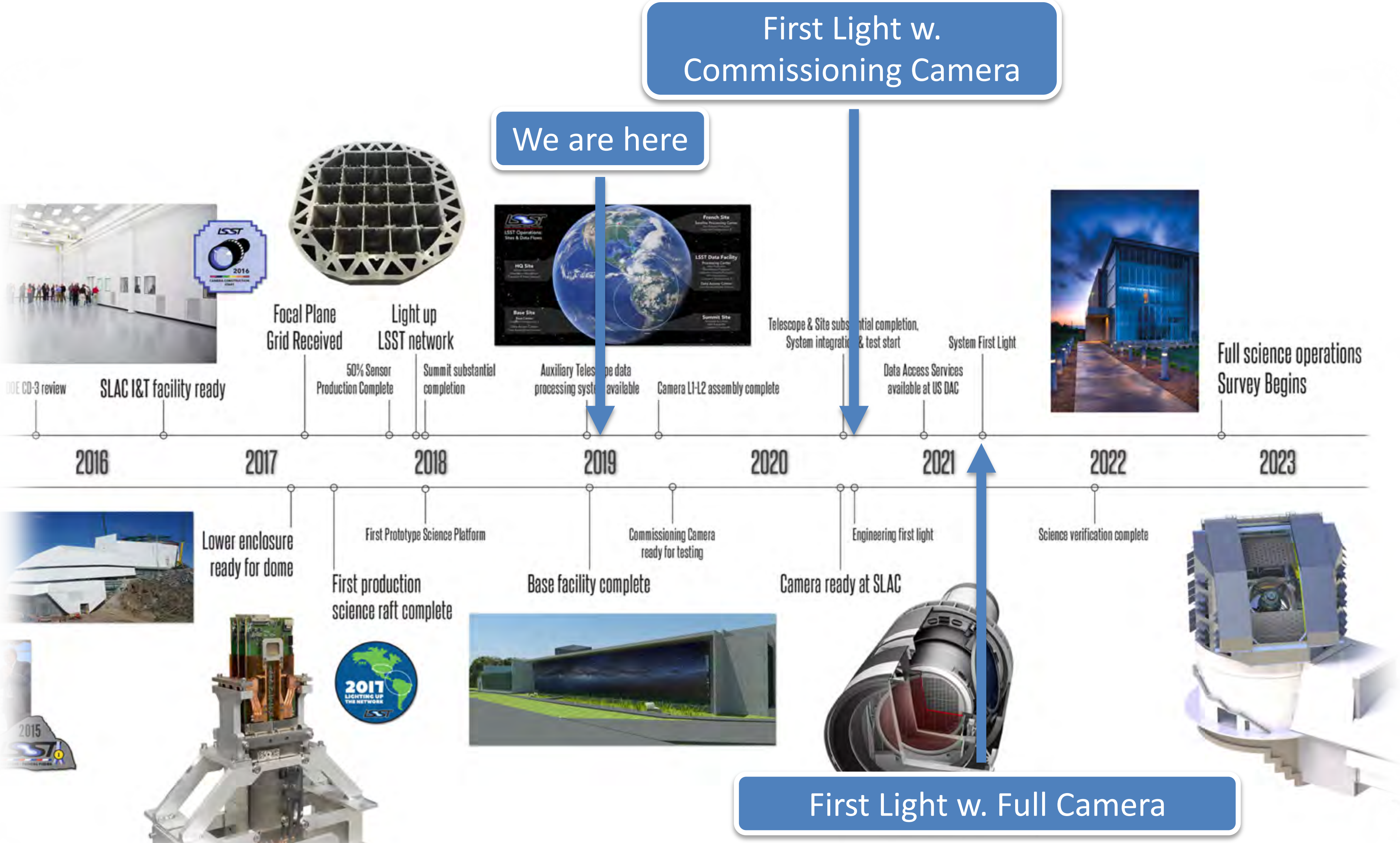
Data Rel.

- User-produced added-value data products (deep KBO/NEO catalogs, variable star classifications, shear maps, …)

User generated

*For more details, see the **“Data Products Definition Document”**, http://ls.st/lse-163*

LSST Is Almost Here!

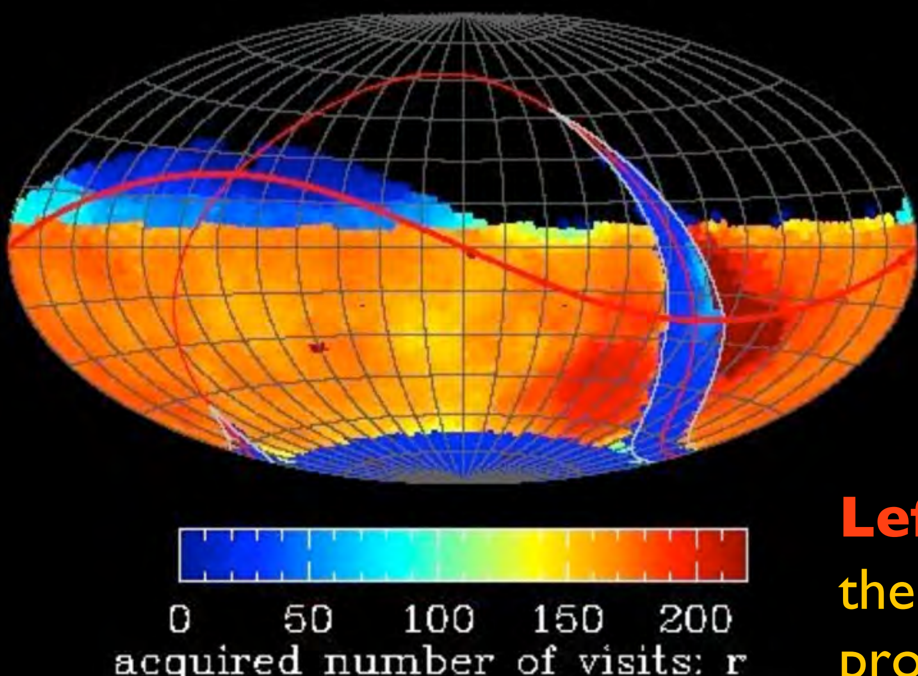


https://www.lsst.org/about/timeline

LSST Science

Basic idea behind LSST: **a uniform sky survey**

- 90% of time will be spent on a uniform survey: every 3-4 nights, the whole observable sky will be scanned twice per night
- after 10 years, half of the sky will be imaged about 1000 times (in 6 bandpasses, ugrizy): a digital color movie of the sky
- ~100 PB of data: about a billion 16 Mpix images, enabling **measurements for 40 billion objects!**



LSST in one sentence:
An optical/near-IR survey of half the sky in ugrizy bands to r~27.5 (36 nJy) based on 825 visits over a 10-year period: deep wide fast.

Left: a 10-year simulation of LSST survey: the number of visits in the r band (Aitoff projection of eq. coordinates)

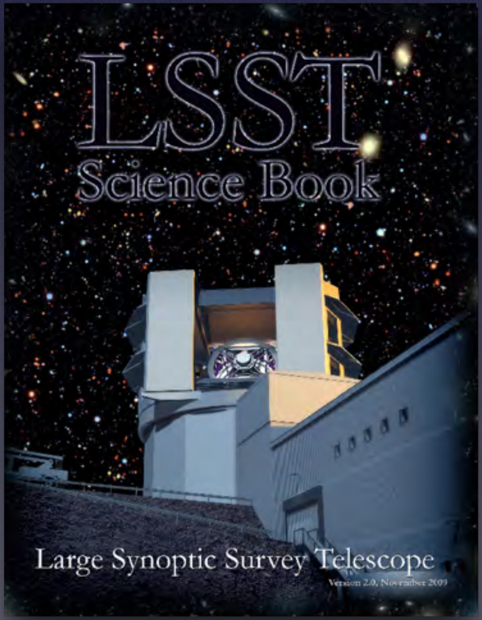
Key Science Themes Enabled by LSST

- Dark matter, dark energy, cosmology (spatial distribution of galaxies, gravitational lensing, supernovae, quasars)
- Time domain (cosmic explosions, variable stars)
- The Solar System structure (asteroids)
- The Milky Way structure (stars)

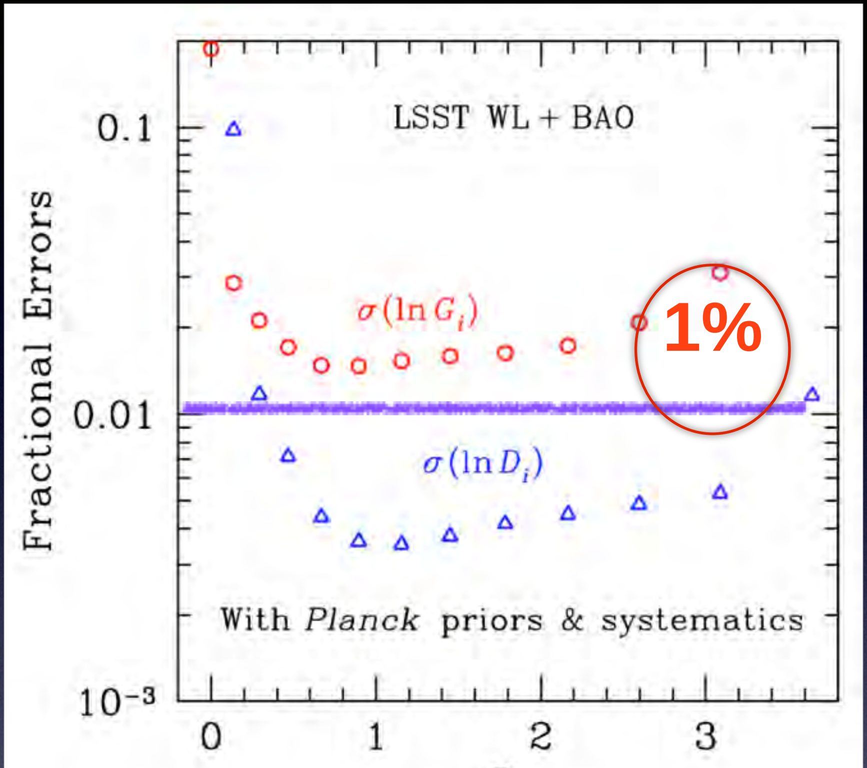
LSST Science Book: arXiv:0912.0201

Summarizes LSST hardware, software, and observing plans, science enabled by LSST, and educational and outreach opportunities

245 authors, 15 chapters, 600 pages

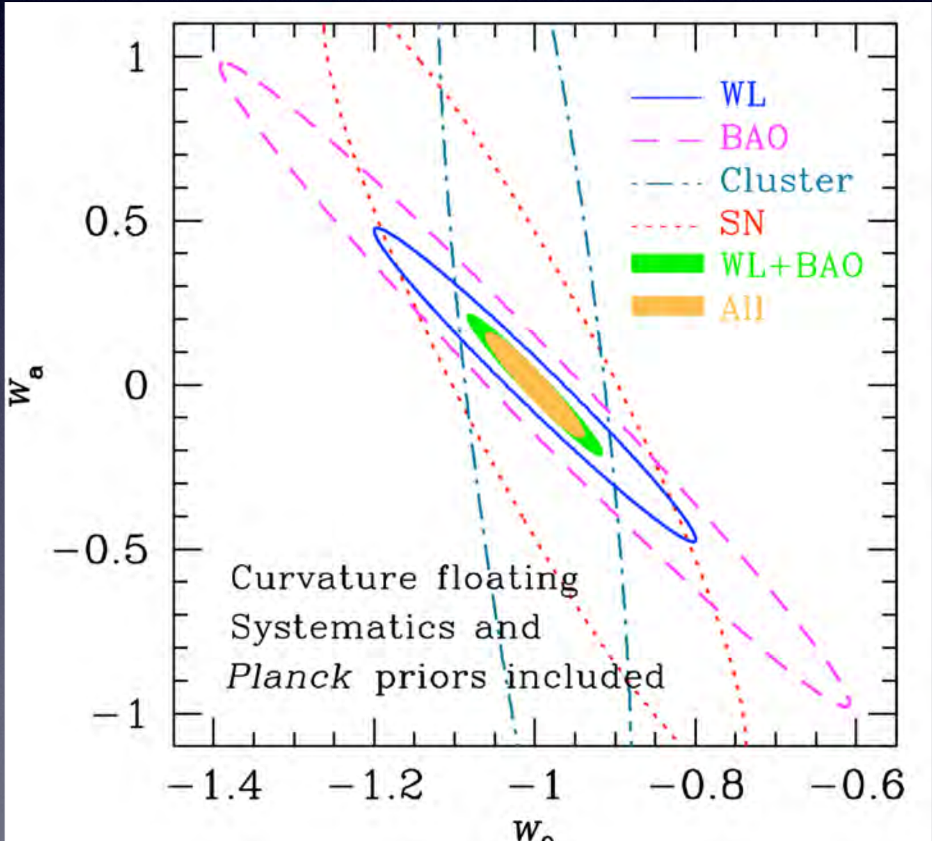


Cosmology with LSST: high precision measurements



LSST Science Book, figure 15.2

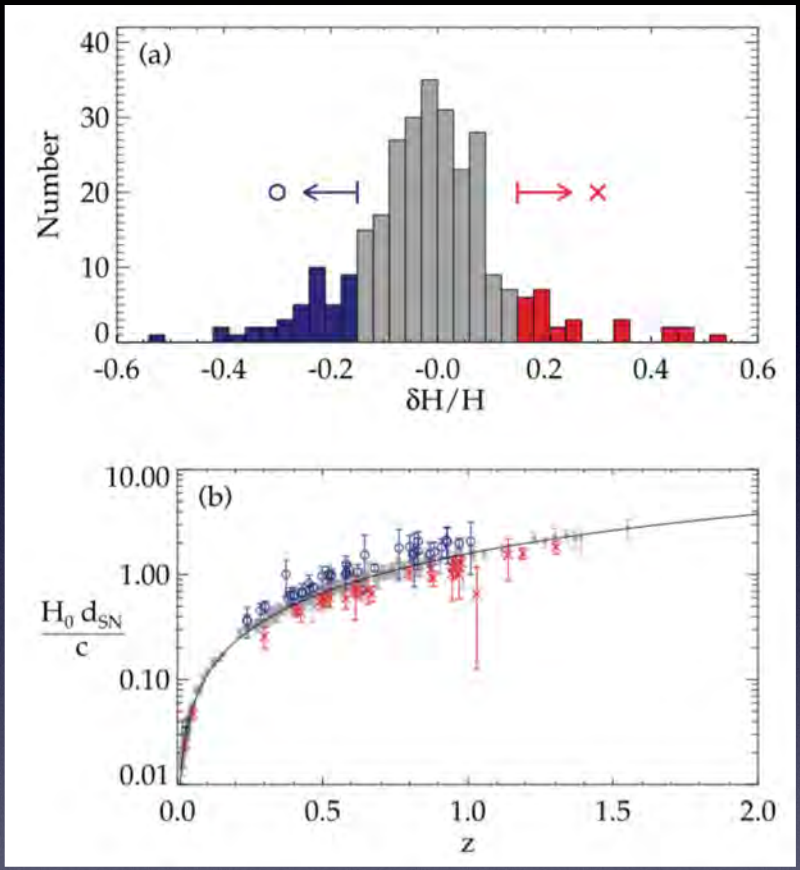
- Measuring distances, H(z), and growth of structure, G(z), with a percent accuracy for 0.5 < z < 3
- Multiple probes are the key!



LSST Science Book, figure 15.3

LSST is designed to be a Stage IV Dark Energy Experiment (DETF)

Cosmology with LSST SNe: is the cosmic acceleration the same in all directions?



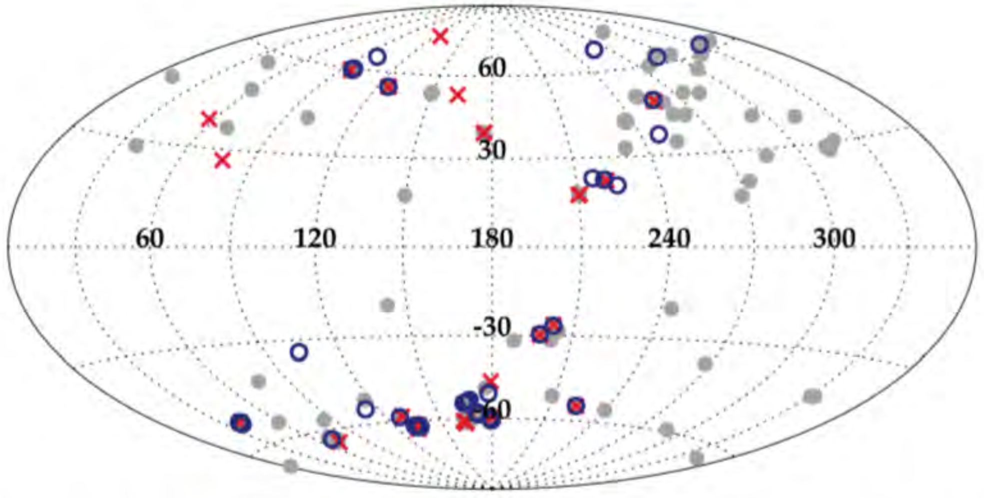


Figure 1. A projection of the spatial distribution of the Union SNe Ia sample in Galactic coordinates. Note the relative uniformity of the points, except around the Galactic plane. The symbols correspond to those in Fig. 2, and are explained in Section 3.1.

Cooke & Lynden-Bell (2009, MNRAS 401, 1409)

Is there spatial structure in the SNe distance modulus residuals for the concordance model?

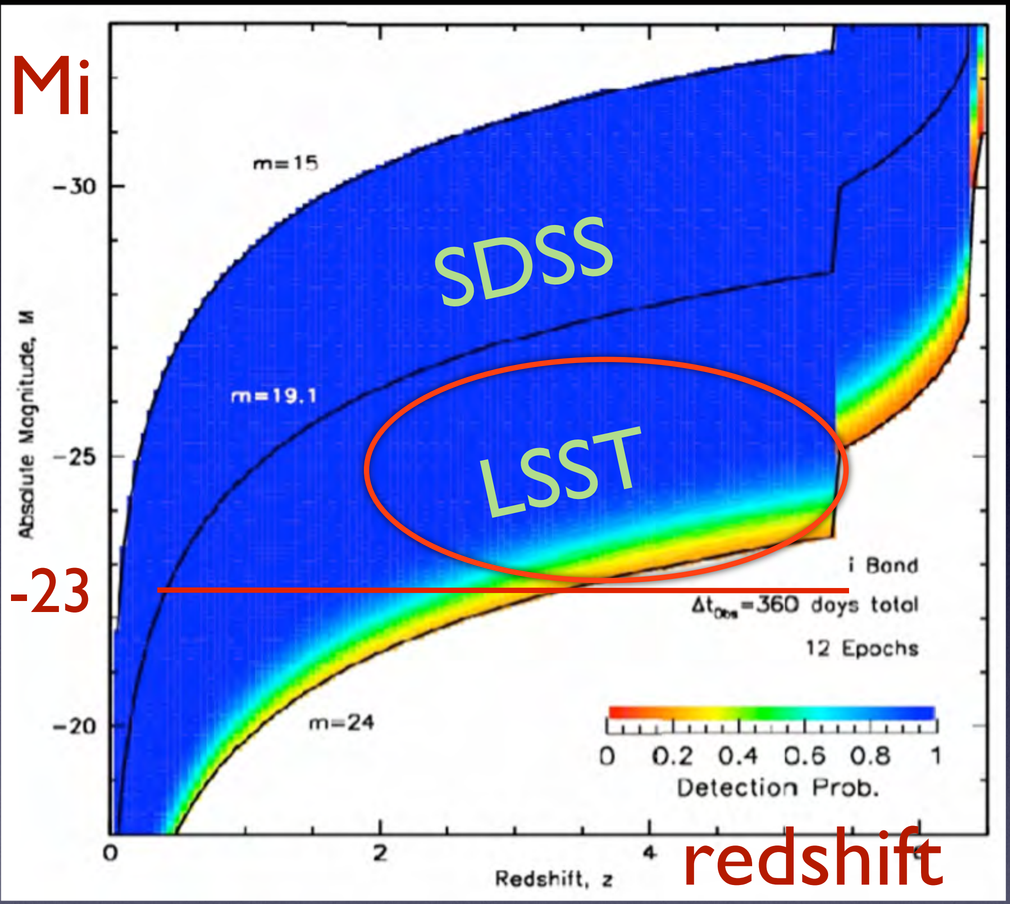
- Even a single supernova represents a cosmological measurement!
- LSST will obtain light curves for several million Type Ia supernovae!

Extragalactic astronomy: low surface brightness objects

Sloan Digital Sky Survey

3x3 arcmin, gri

Extragalactic astronomy: Quasars



Top: absolute magnitude vs. redshift diagram for quasars

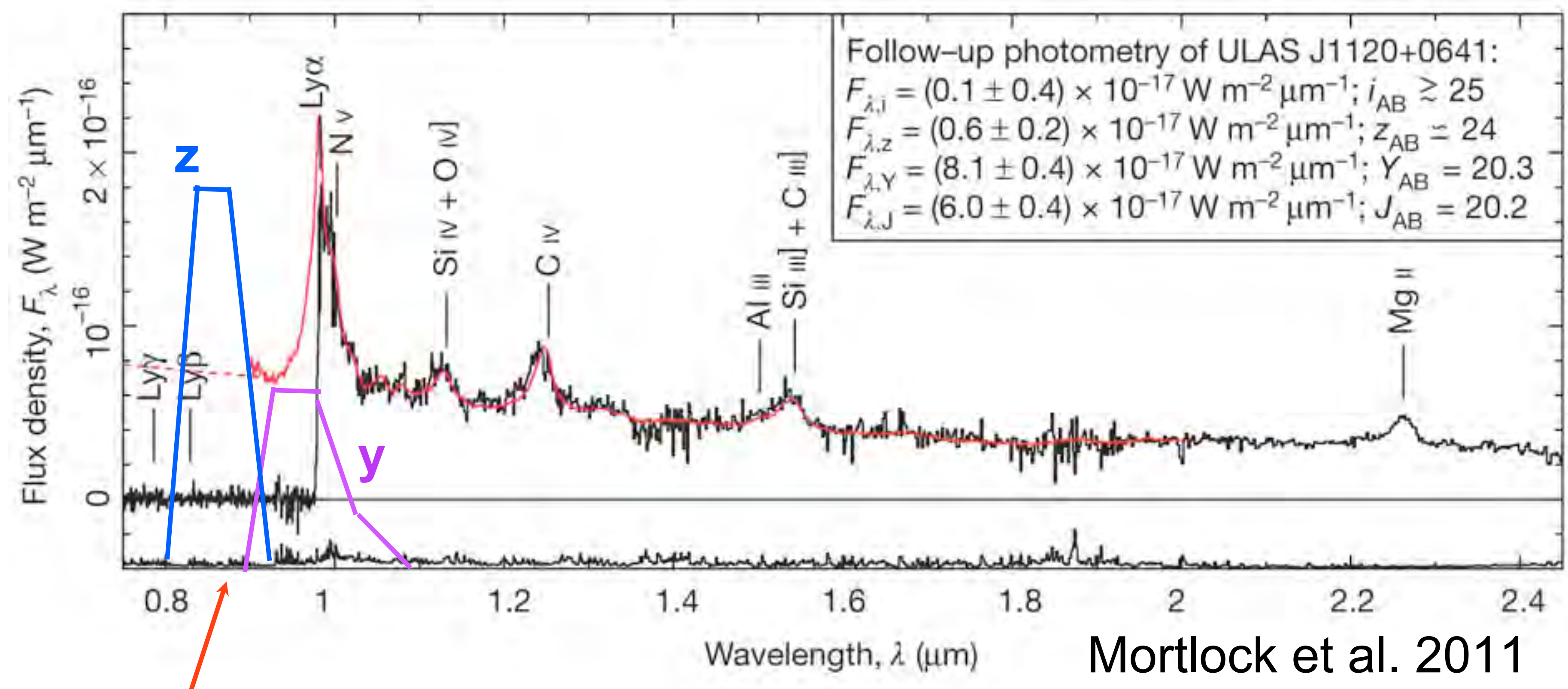
- About 10 million quasars will be discovered using variability, colors, and the lack of proper motions
- The sample will include Mi=-23 objects even at redshifts beyond 3
- Quasar variability studies will be based on millions of light curves with 1000 observations over 10 yrs

Today: ~100s of quasars with 6<z<7.5

Reionization studies!

LSST will detect ~10,000 quasars with 6<z<7.5!

The Highest Redshift Quasar at z=7.085 from UKIDSS



Mortlock et al. 2011

Such a quasar would be detected by LSST as a z-band dropout

(multi-epoch data will greatly help with false positives)

LSST should discover about 1,000 quasars with z>7

Today (2016): one quasar with z>7

Time Domain: objects changing in time

positions: asteroids and stellar proper motions

brightness: cosmic explosions and variable stars

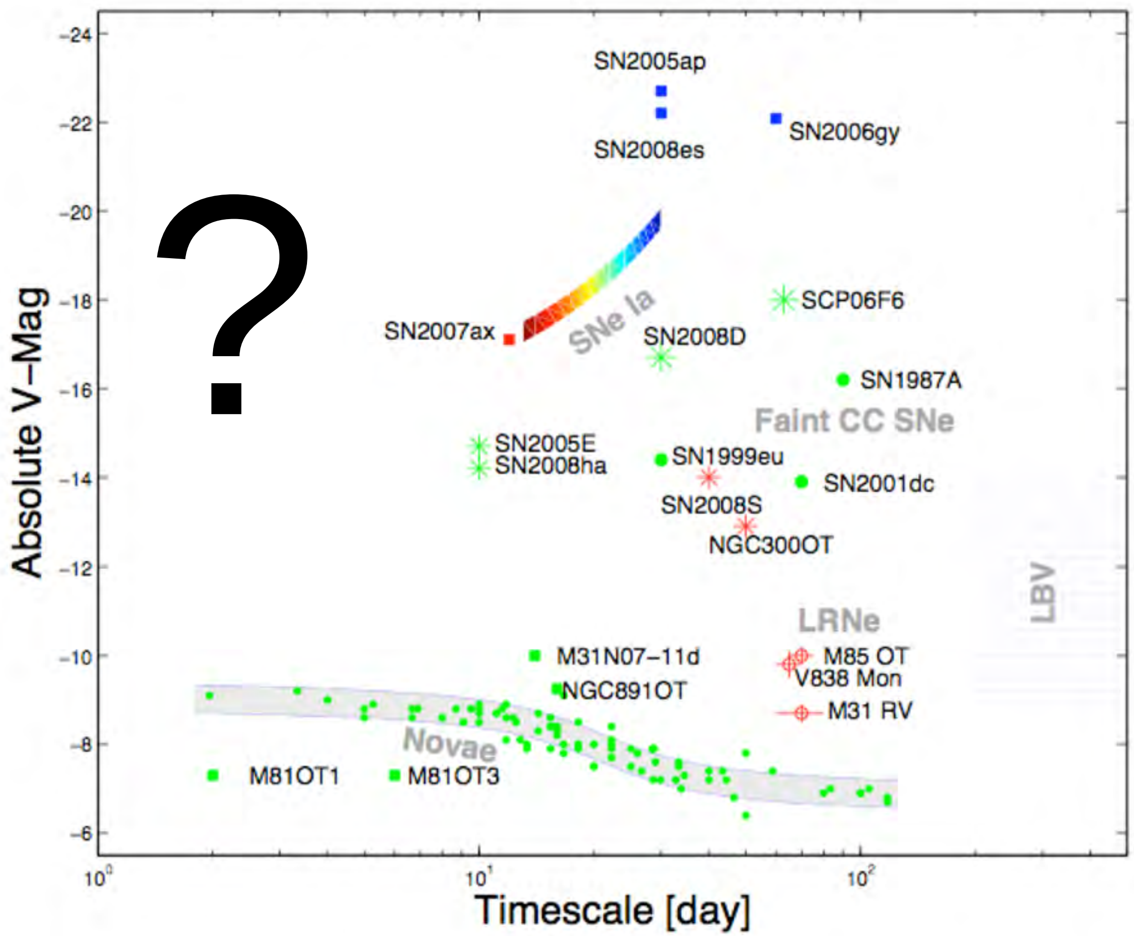


FIG. 29.— The phase space of cosmic explosive and eruptive transients as represented by their absolute V band peak brightness and the event timescale (adapted from Kulkarni et al. 2007).

LSST will extend time-volume space a thousand times over current surveys (new classes of object?)

known unknowns
unknown unknowns

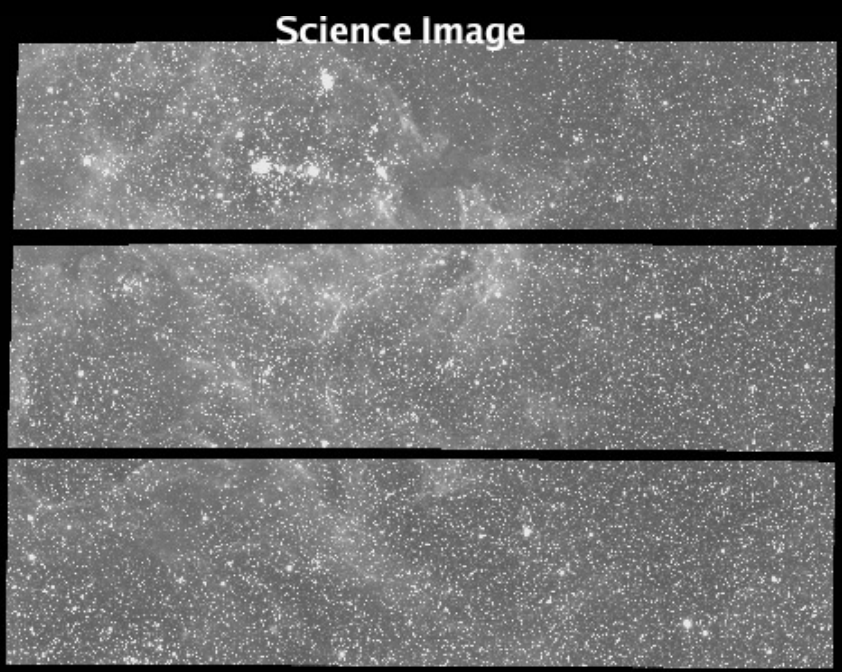
Note: There will be as many variable stars from LSST, as all stars from SDSS !

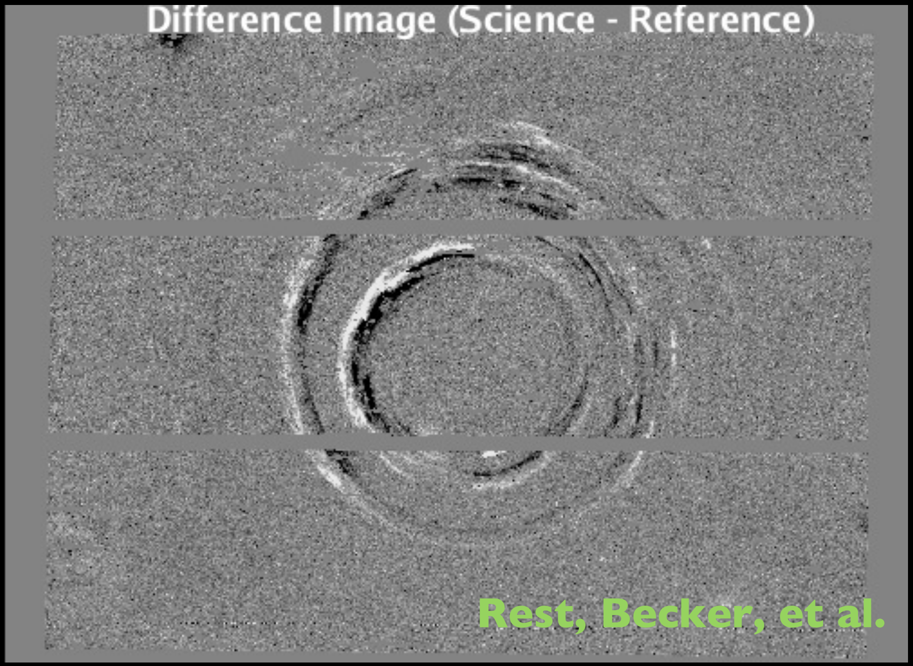
Time Domain: objects changing in time

positions: asteroids and stellar proper motions

brightness: cosmic explosions and variable stars

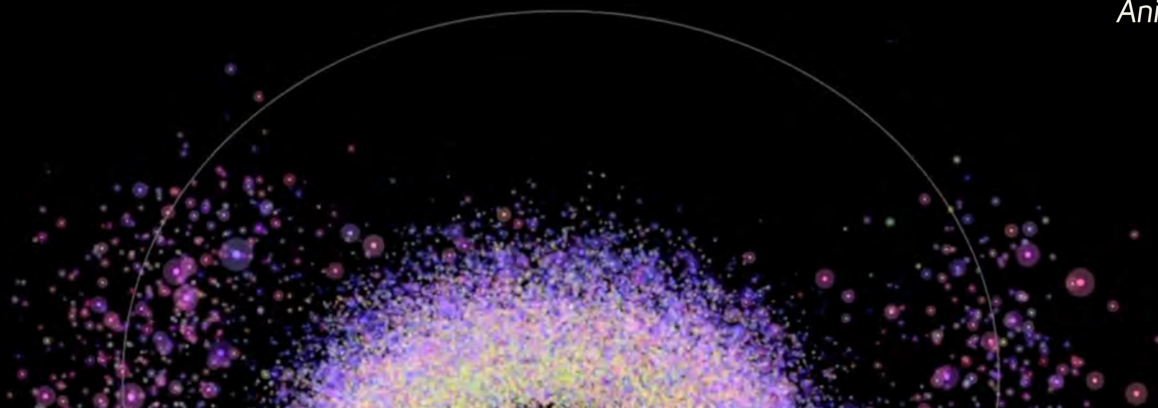
Not only point sources - echo of a supernova explosion:





Census of the Solar System

Animation: SDSS Asteroids
(Alex Parker, SwRI)



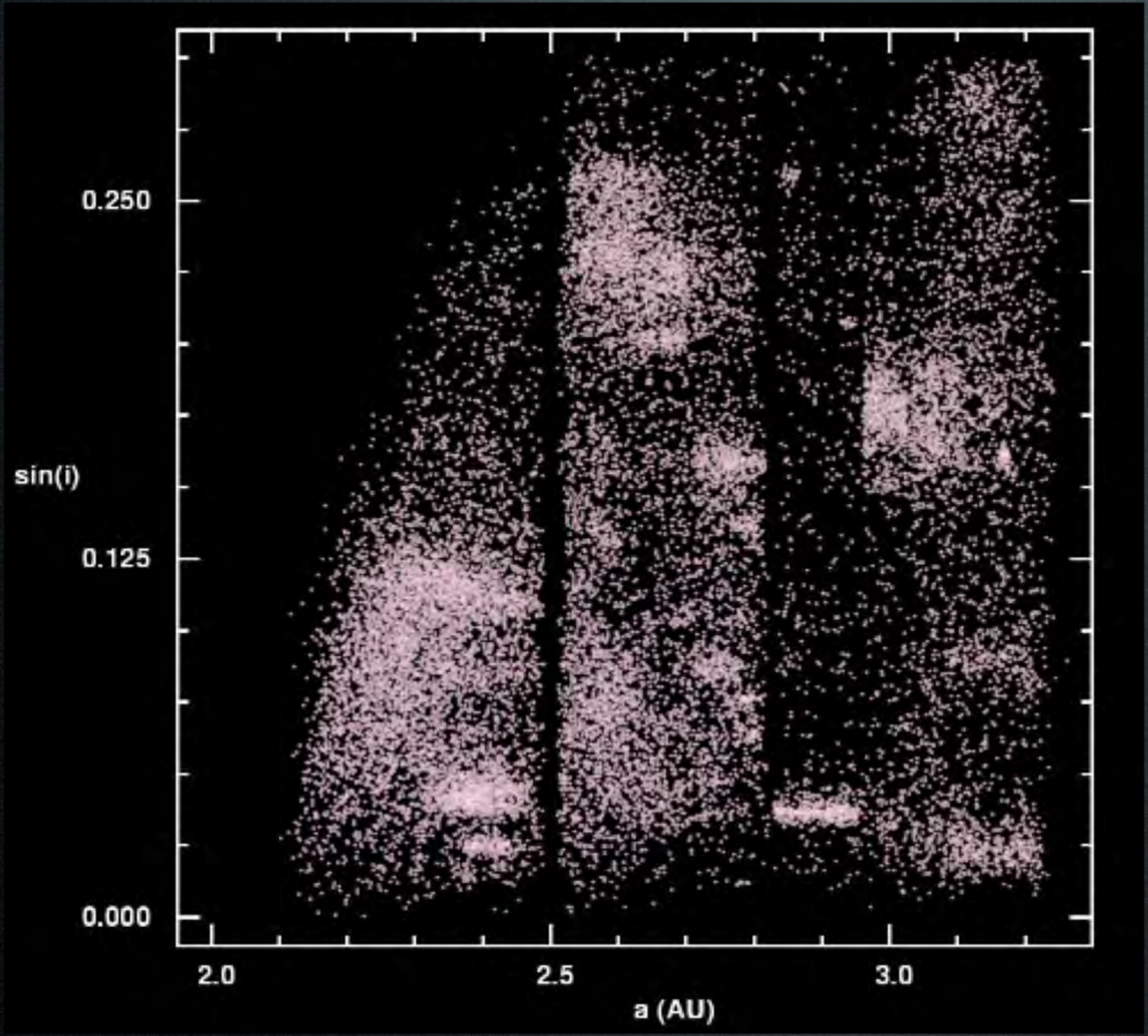
	Currently Known*	LSST Discoveries**	Median number of observations+	Observational arc length+
Near Earth Objects (NEOs)	14,500	100,000	(D>250m) 60	6.0 years
Main Belt Asteroids (MBAs)	650,000	5,500,000	(D>500m) 200	8.5 years
Jupiter Trojans	6000	280,000	(D>2km) 300	8.7 years
TransNeptunian Objects (TNOs) + Scattered Disk Objects (SDOS)	2000	40,000	(D>200km) 450	8.5 years

Estimates: Lynne Jones et al.

Solar System (Small Body) Science with LSST

- **Large statistical samples of asteroids and comets**
- Serendipitous discoveries of rare events or objects
  - Asteroid collisions (P/2010 A2)
  - Retrograde TNO (2008 KV42)
- Discover new, incoming comets even before they become active
- Model shapes of asteroids from measurements of their brightness
- Discover links between different populations
  - How are NEOs and Main Belt asteroids related?
  - Are irregular satellites actually captured TNOs?
- Expand our knowledge of all small bodies to provide a better understanding of the formation and evolution of our Solar System

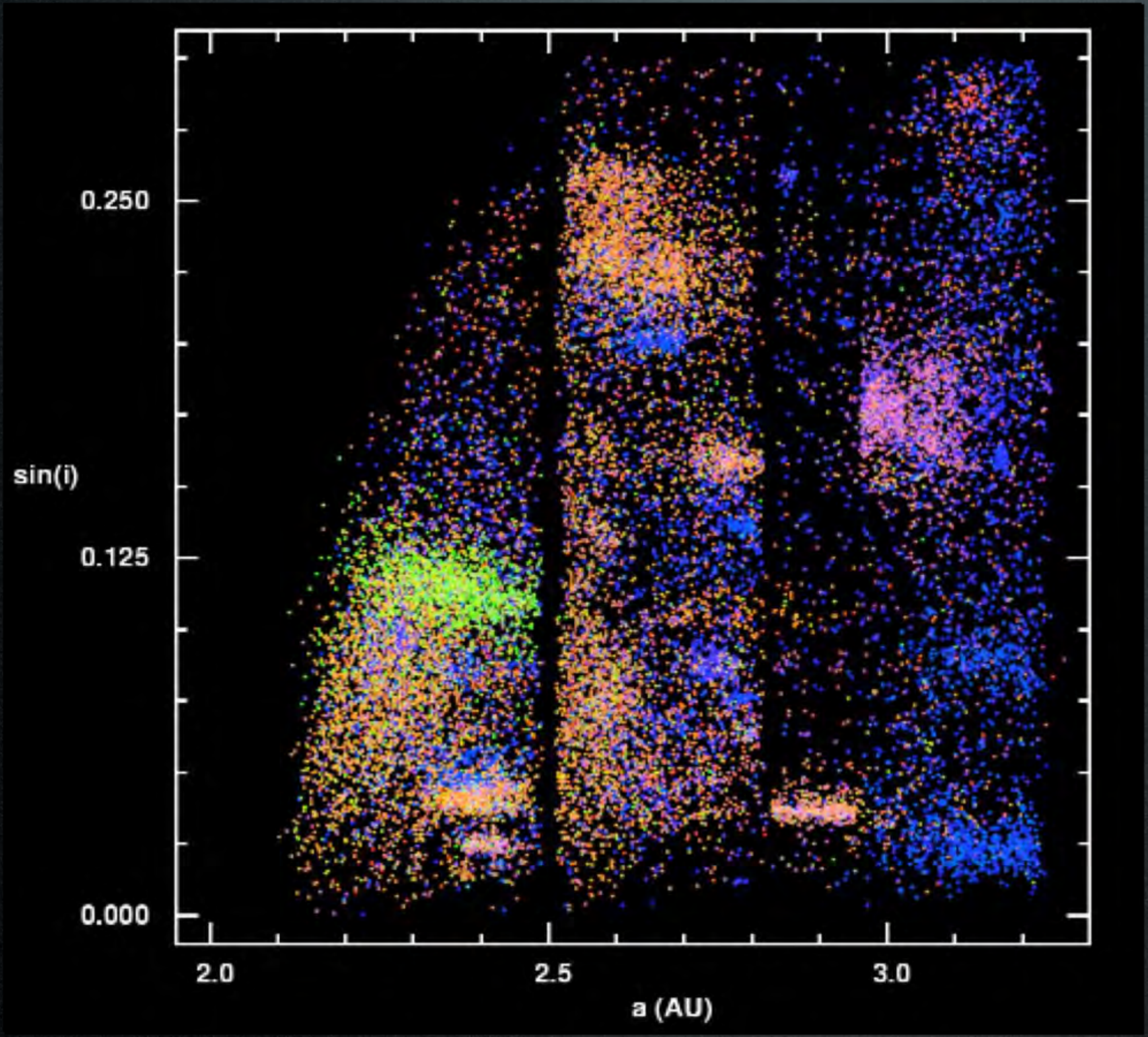
Main-belt Inventory



30,000
Asteroids with
SDSS colors and
proper
orbital elements

(Ivezic, Juric, Lupton 2002)

Main-belt Inventory



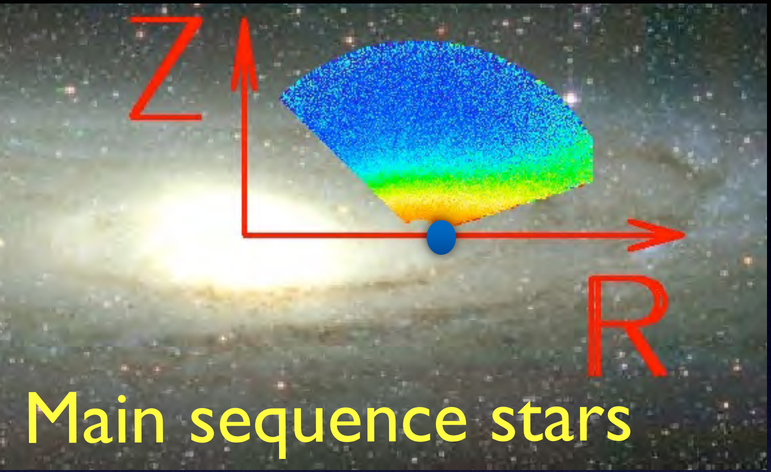
30,000
Asteroids with
SDSS colors and
proper
orbital elements

(Ivezic, Juric, Lupton 2002)

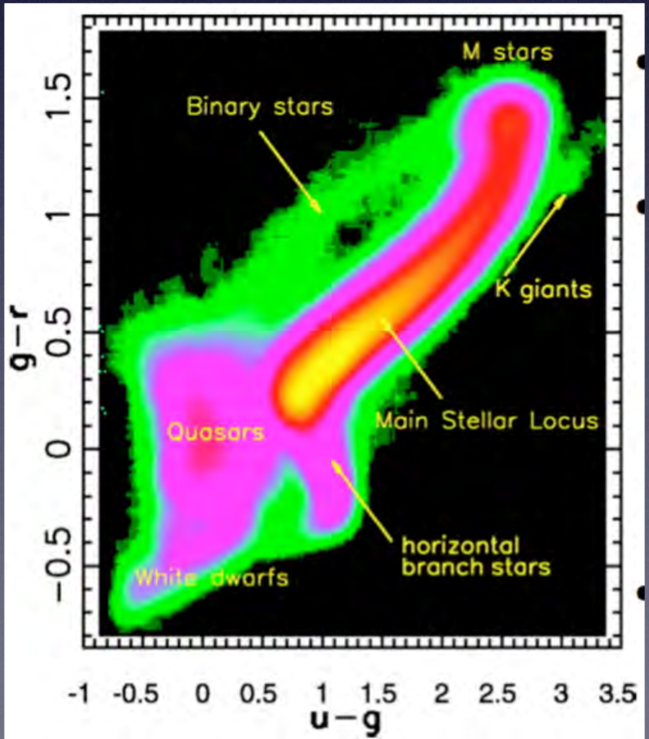
Color-coded with
SDSS colors

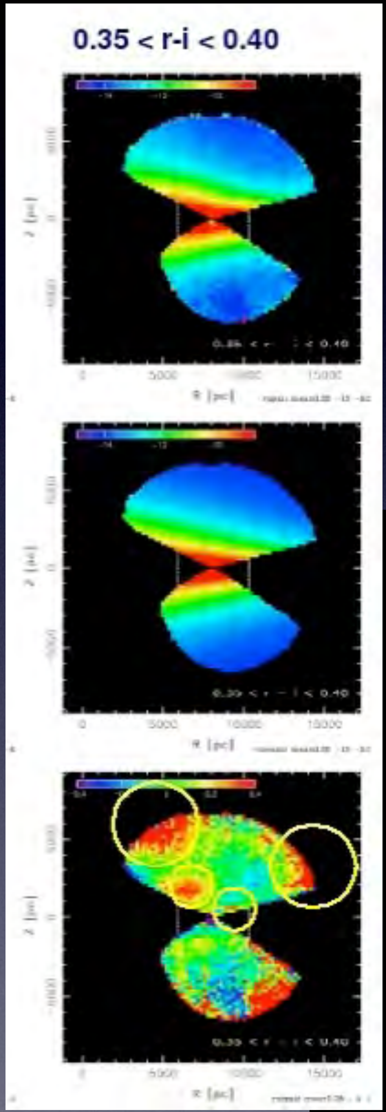
Colors help with the definition of asteroid families.
LSST will also provide color light curves!

The Milky Way structure: 20 billion stars, time domain massive statistical studies!



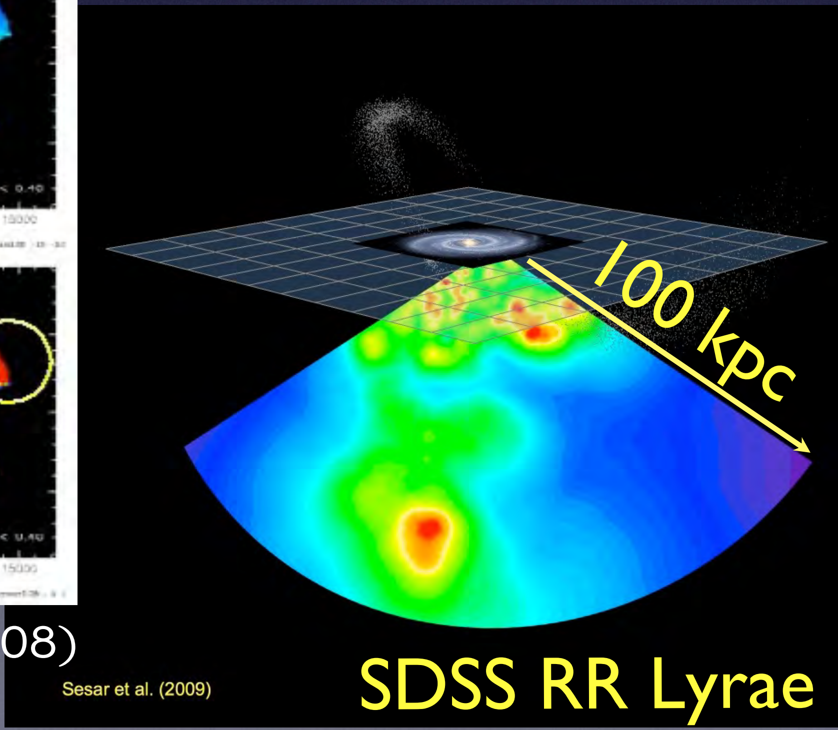
Distance and [Fe/H]:



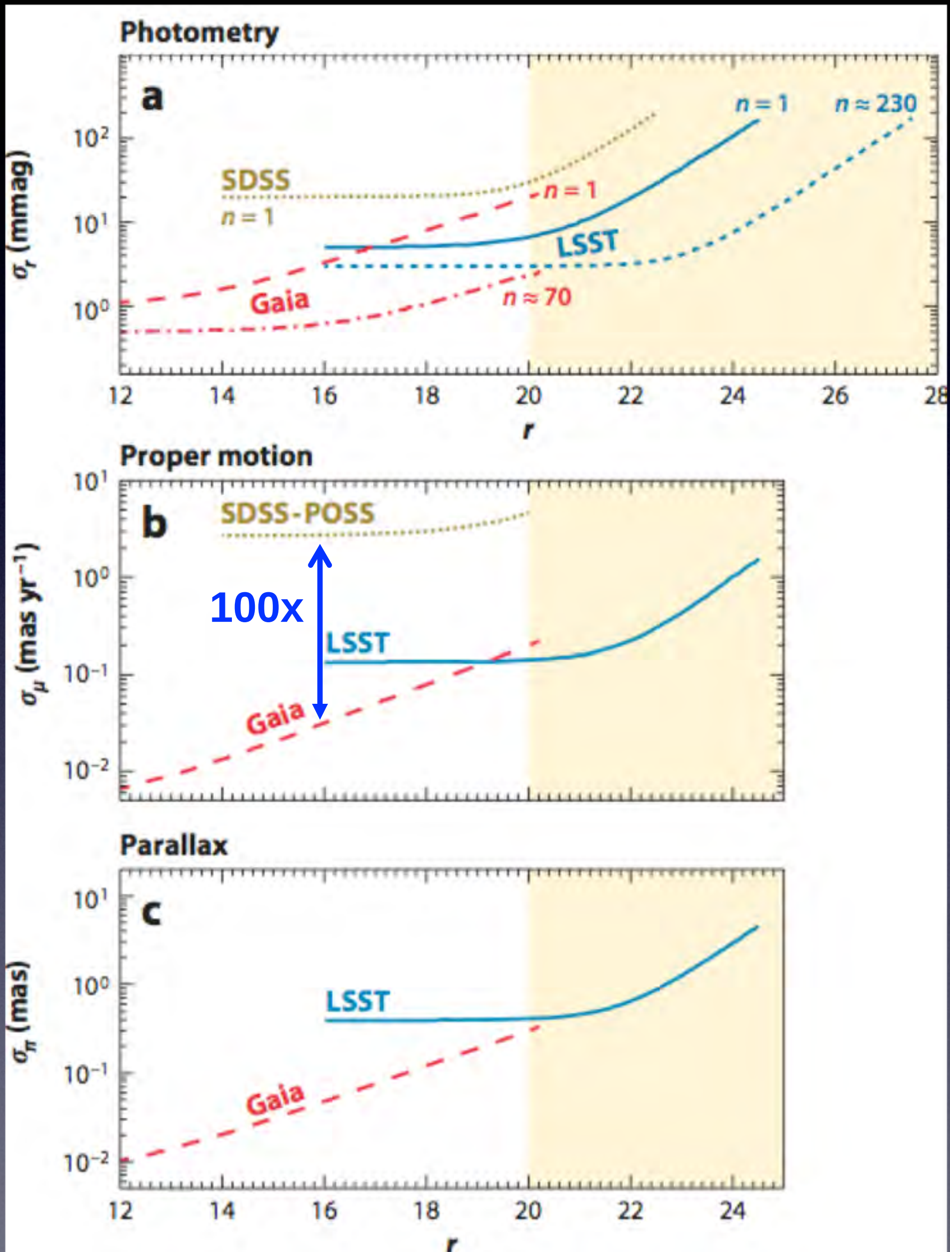


Juric et al. (2008)

Compared to SDSS: LSST can "see" about 40 times more stars, 10 times further away and over twice as large sky area



Gaia vs. LSST comparison



Ivezić, Beers, Jurić 2012, ARA&A, 50, 251

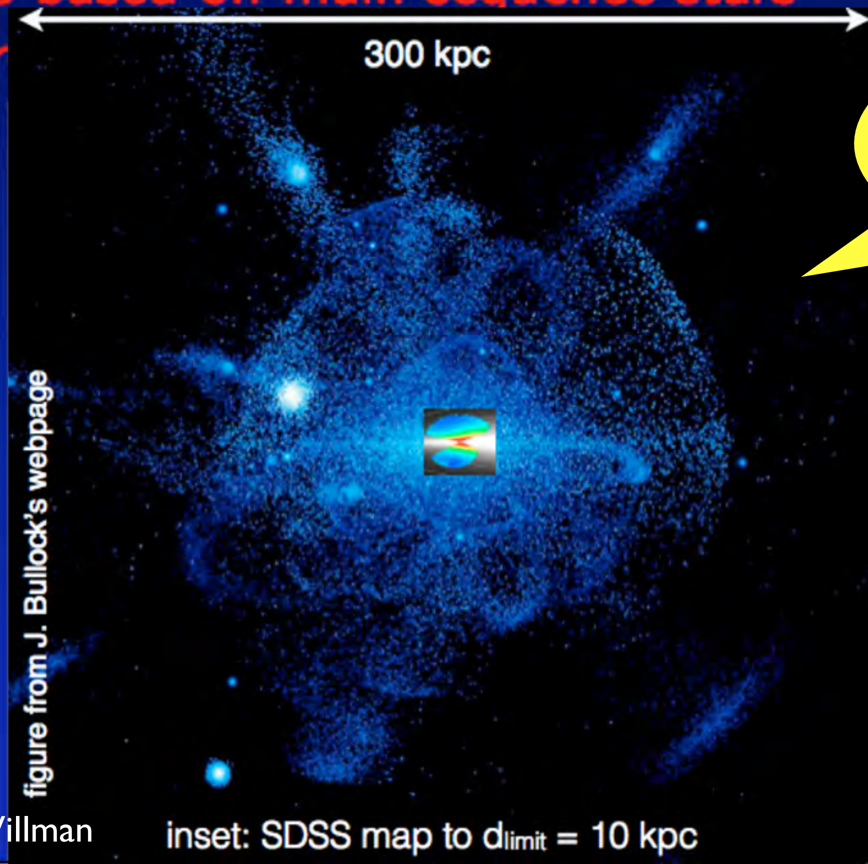
- **Gaia:** excellent astrometry (and photometry), but only to r < 20
- **LSST:** photometry to r < 27.5 and time resolved measurements to r < 24.5
- Complementarity of the two surveys: photometric, proper motion and trigonometric parallax errors are similar around r=20

The Milky Way disk "belongs" to Gaia, LSST will be excellent for the halo (plus very faint and/or very red sources, such as white dwarfs and LT(Y) dwarfs).

The large blue circle: the ~400 kpc limit of future LSST studies based on RR Lyrae

LSST limit for RR Lyrae: 400 kpc

The large red circle: the ~100 kpc limit of future LSST studies based on main-sequence stars (and the curr



200 million stars from LSST!

The small insert: ~10 kpc limit of SDSS and future Gaia studies for kinematic & $[Fe/H]$ mapping with MS stars

montage from B. Willman

Doing Science With LSST

LSST Data Products

- 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.

Prompt

- 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.

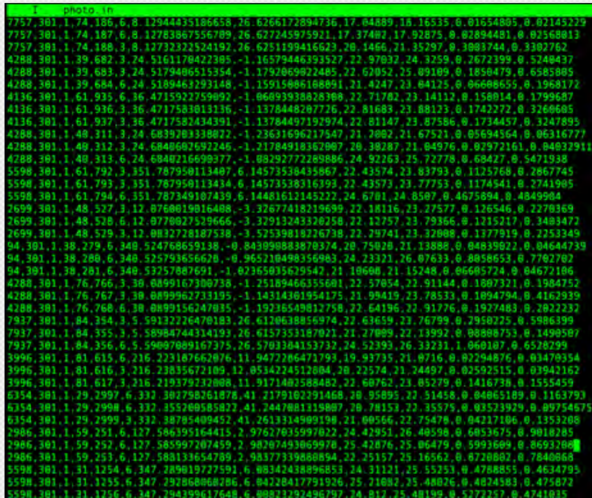
Data Rel.

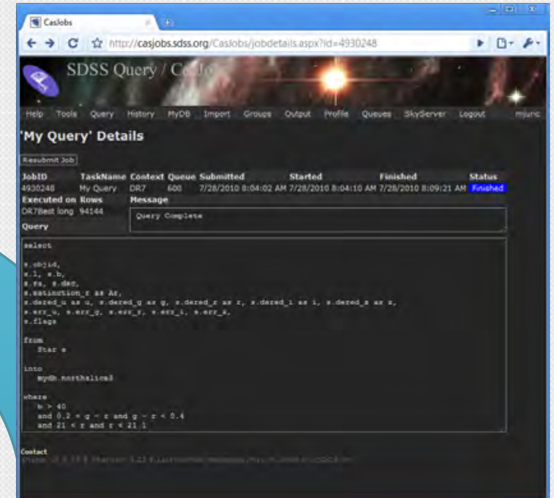
- User-produced added-value data products (deep KBO/NEO catalogs, variable star classifications, shear maps, …)

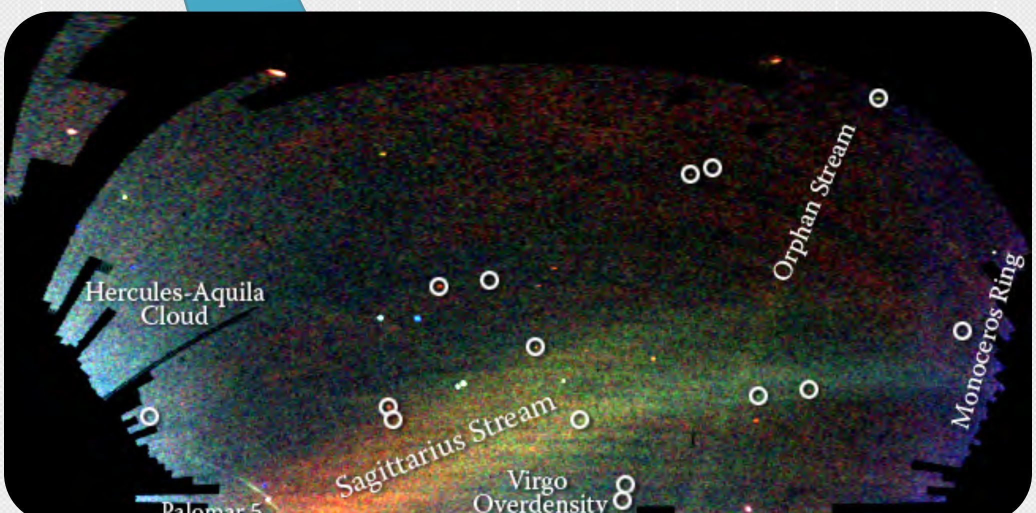
User generated

*For more details, see the **"Data Products Definition Document"**, http://ls.st/lse-163*

Analysis Paradigms: Subset – Download – Analyze







Data Volumes

	ZTF	LSST
Number of detections	1 trillion	7 trillion
Number of objects	1 billion	37 billion
Nightly alert rate	1 million	10 million
Nightly data rate	1.4 TB	15 TB
Alert latency	< 20 minutes	60 seconds

Science analysis code

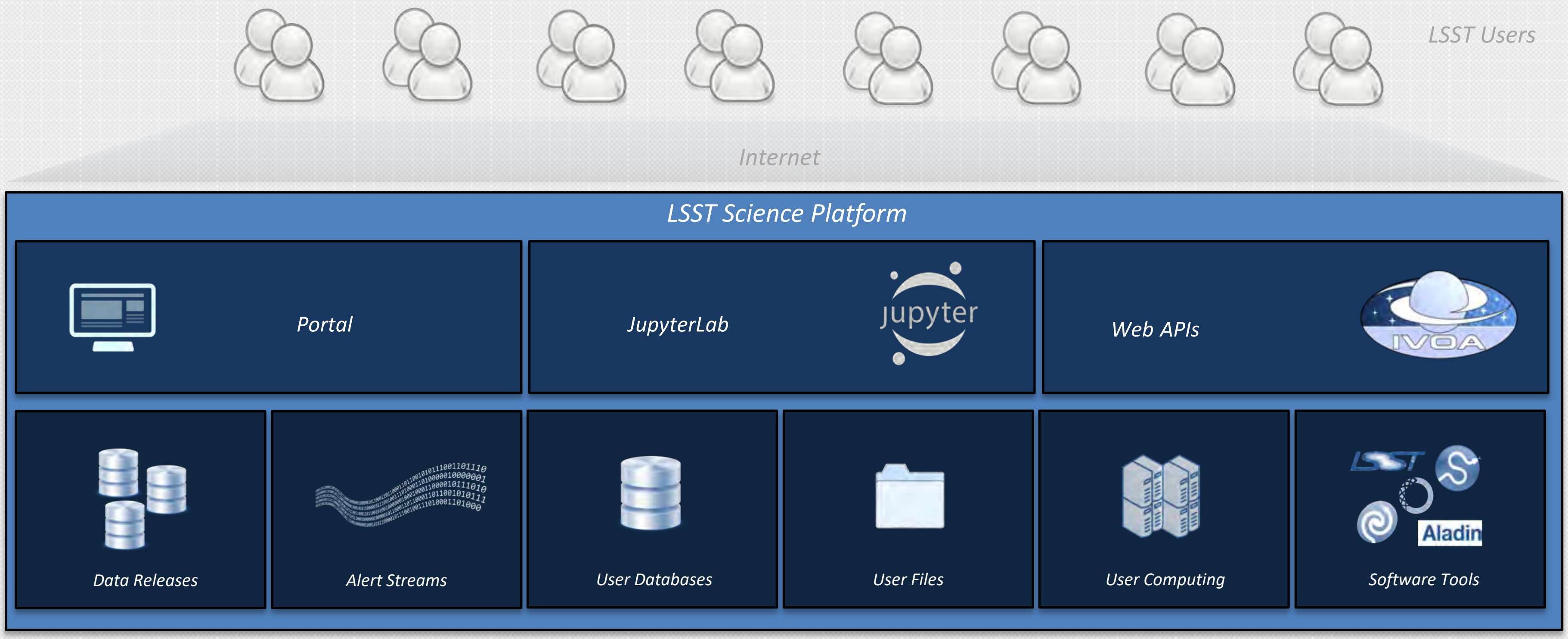
~50kb

If the data is big…

… bring the code to the data.

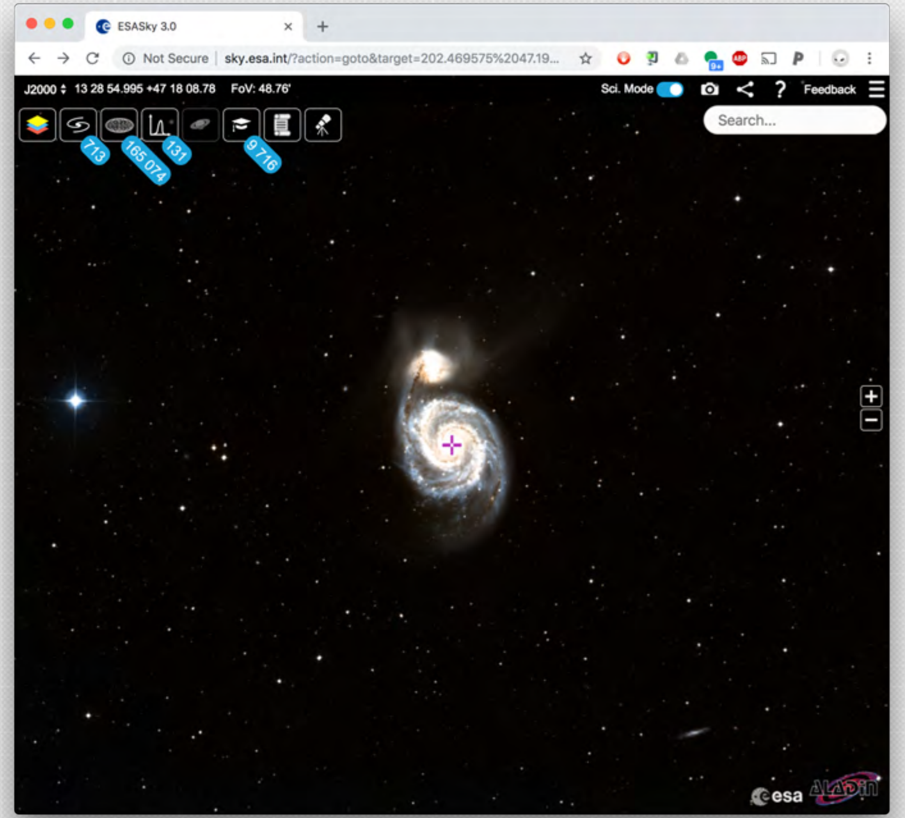
The LSST Science Platform: Accessing LSST Data and Enabling LSST Science

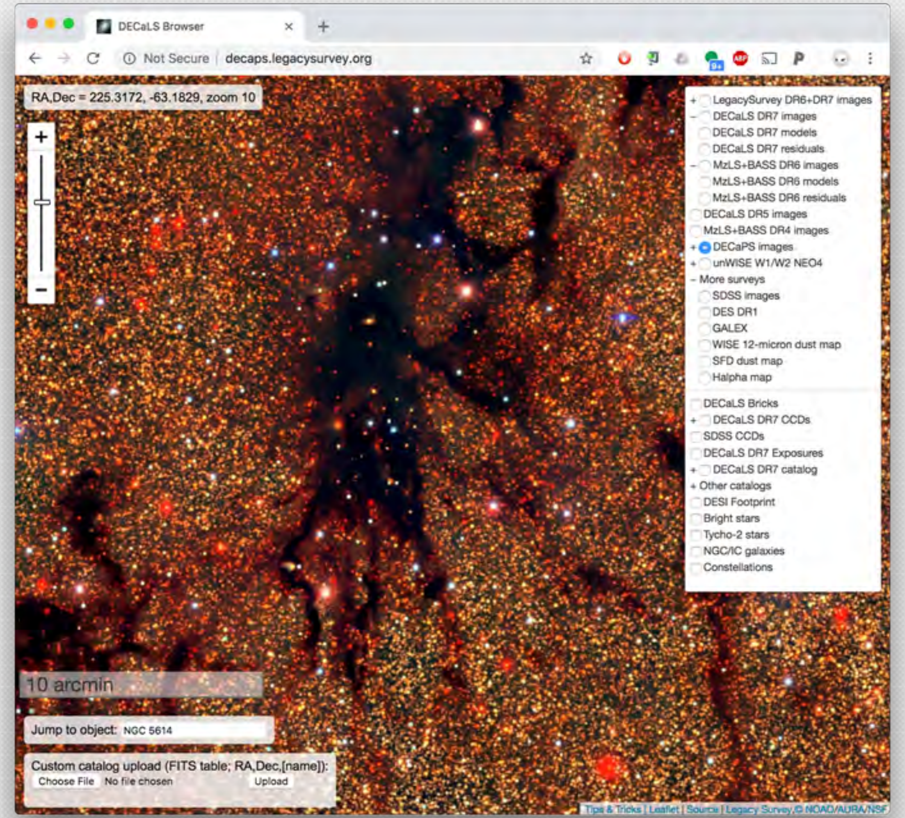
Juric, Dubois-Felsmann & Ciardi (2016)
http://ls.st/lse-319



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

What to expect

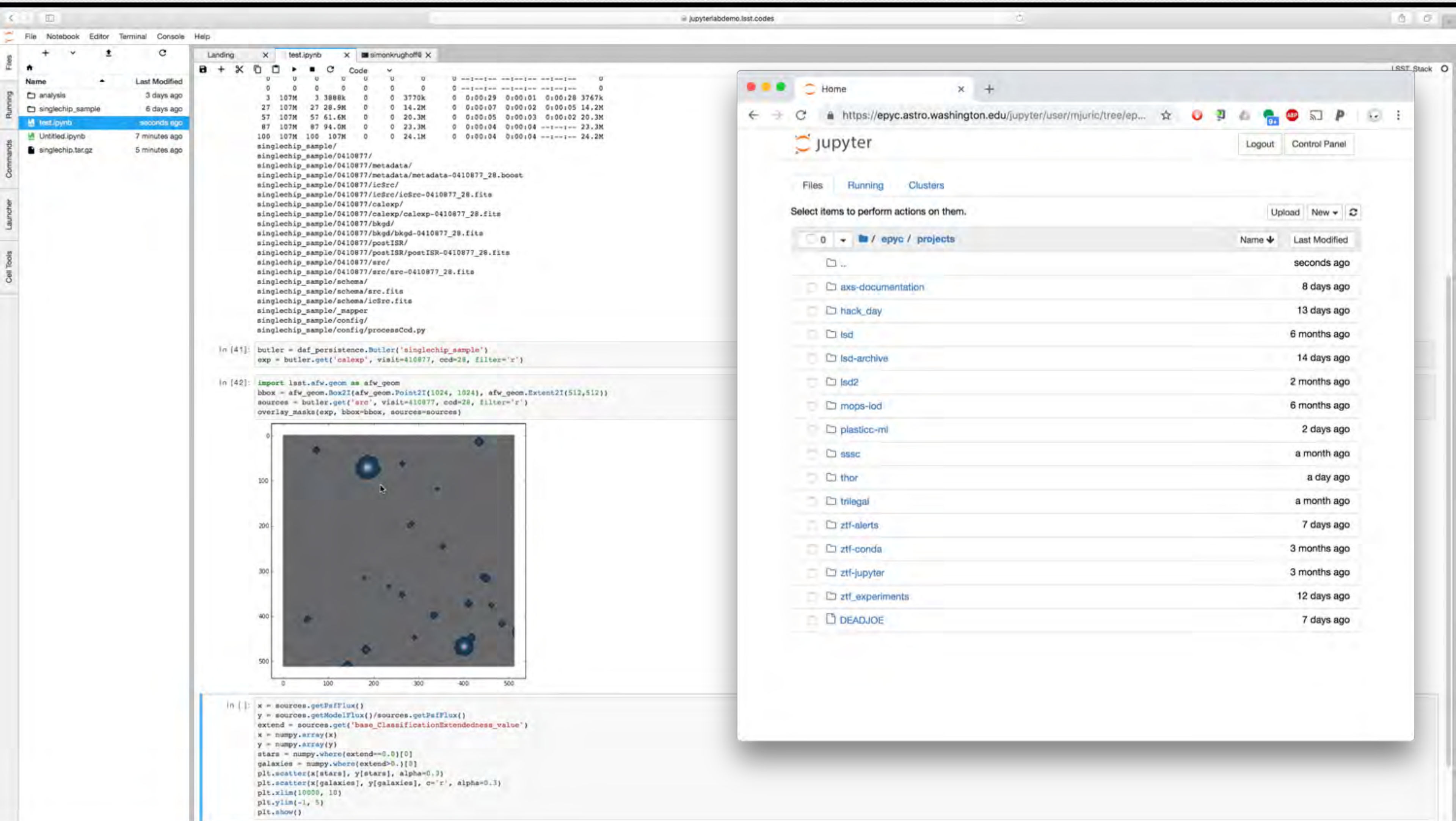




http://sky.esa.int/

http://decaps.legacysurvey.org/

JupyterLab: Next-to-the-data Analysis



YouTube demo of the LSST JupyterLab Aspect Demo: ***http://ls.st/bgt***

Why Jupyter is data scientists

https://www.nature.com/articles/d41586-018-07196-1

nature > toolbox > article

a natureresearch journal

nature
International journal of science

Search E-alert Submit Login

TOOLBOX · 30 OCTOBER 2018

Why Jupyter is data scientists' computational notebook of choice

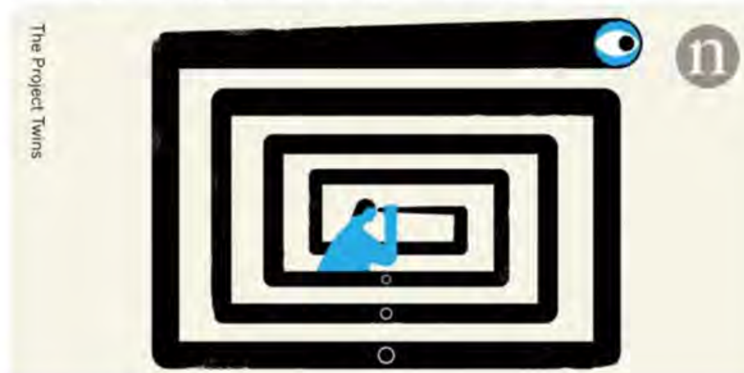
An improved architecture and enthusiastic user base are driving upta
open-source web tool.

Jeffrey M. Perkel

Nature Careers
@NatureCareers

Follow

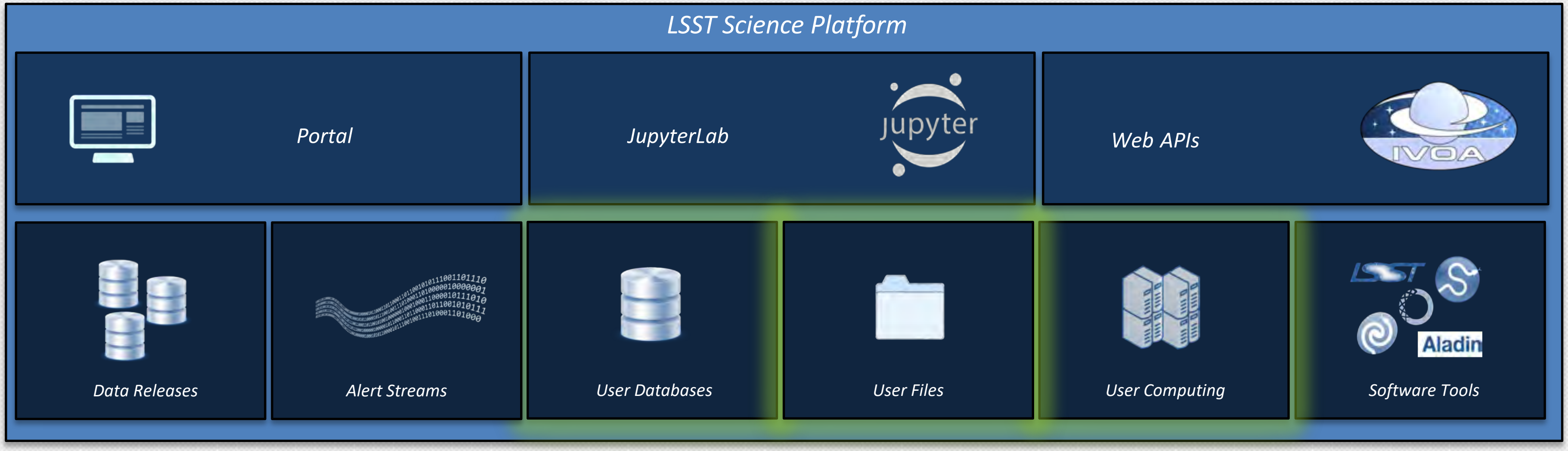
"I've never seen any migration this fast. It's just amazing." -- @mjuric on the rise of @ProjectJupyter in data science



Why Jupyter is data scientists' computational notebook of choice
An improved architecture and enthusiastic user base are driving uptake of the open-source web tool.
nature.com

10:30 PM - 5 Nov 2018

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 is the "LSST Science Cloud" (@ DR2)?

- **Computing:**
  - ~2,400 cores
  - ~18 TFLOPs
- **File storage:**
  - ~4 PB
- **Database storage**
  - ~3 PB

This is shared by all users. We're estimating the number of potential DAC users not to exceed 7500 (relevant for file and database storage).

Not all users will be accessing the computing cluster concurrently. **We are estimating on order of a ~100.**

Though this is a relatively small cluster by 2020-era standards, it will 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.

Think of this as having your own server with a few TB of disk and database storage, right next to the LSST data, with a chance to use tens to hundreds of cores for analysis. <u>It will be excellent for enabling early science!</u>

This kind of approach will become increasingly common for *all* big data archives.

Large-Scale Science in the LSST Era

(my concerns and some potential solutions)

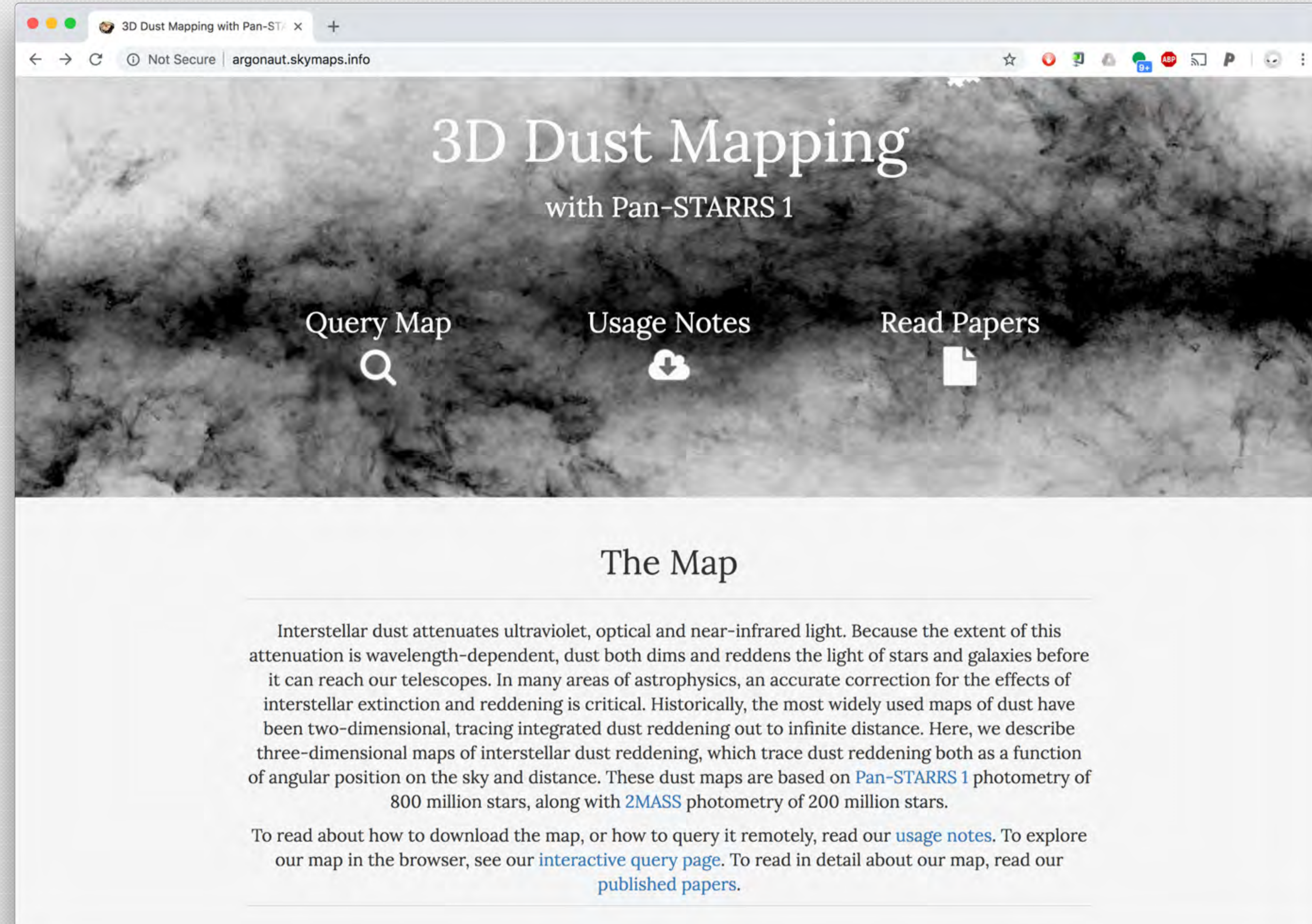
Challenges (part 1)

Better Together

(joining datasets is powerful)

I Want it All

(science demands whole dataset operations)


3D Dust Mapping with Pan-ST
Not Secure
argonaut.skymaps.info
3D Dust Mapping
with Pan-STARRS 1
Query Map
Usage Notes
Read Papers
The Map
Interstellar dust attenuates ultraviolet, optical and near-infrared light. Because the extent of this attenuation is wavelength-dependent, dust both dims and reddens the light of stars and galaxies before it can reach our telescopes. In many areas of astrophysics, an accurate correction for the effects of interstellar extinction and reddening is critical. Historically, the most widely used maps of dust have been two-dimensional, tracing integrated dust reddening out to infinite distance. Here, we describe three-dimensional maps of interstellar dust reddening, which trace dust reddening both as a function of angular position on the sky and distance. These dust maps are based on Pan-STARRS 1 photometry of 800 million stars, along with 2MASS photometry of 200 million stars.
To read about how to download the map, or how to query it remotely, read our usage notes. To explore our map in the browser, see our interactive query page. To read in detail about our map, read our published papers.

Whole Dataset Operations

- Galactic structure: density/proper motion maps of the Galaxy
  - => forall stars, compute distance, bin, create 5D map
- Galactic structure: dust distribution
  - => forall stars, compute g-r color, bin, find blue tip edge, infer dust distribution
- Near-field cosmology: MW satellite searches
  - => forall stars, compute colors, convolve with spatial filters, report any satellite-like peaks
- Variability: Bayesian classification of transients and discovery of variables
  - => forall stars, get light curves, compute likelihoods, alert if interesting
- …

Challenges (part 2)

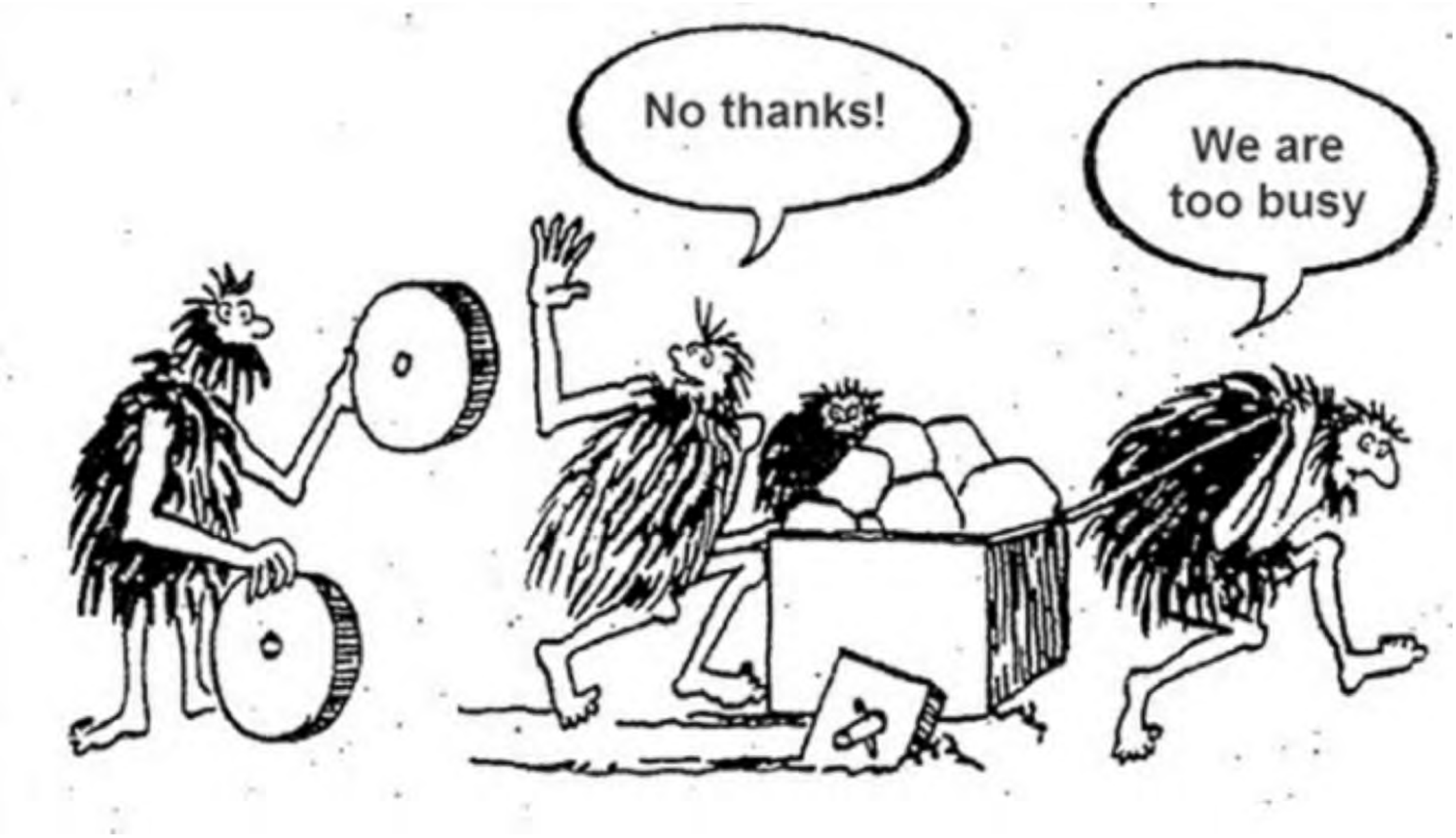
Scalability

(how do I write an analysis code that will scale to thousands of machines and petabytes of data?)

Resources

(where are the resources to run this code?)

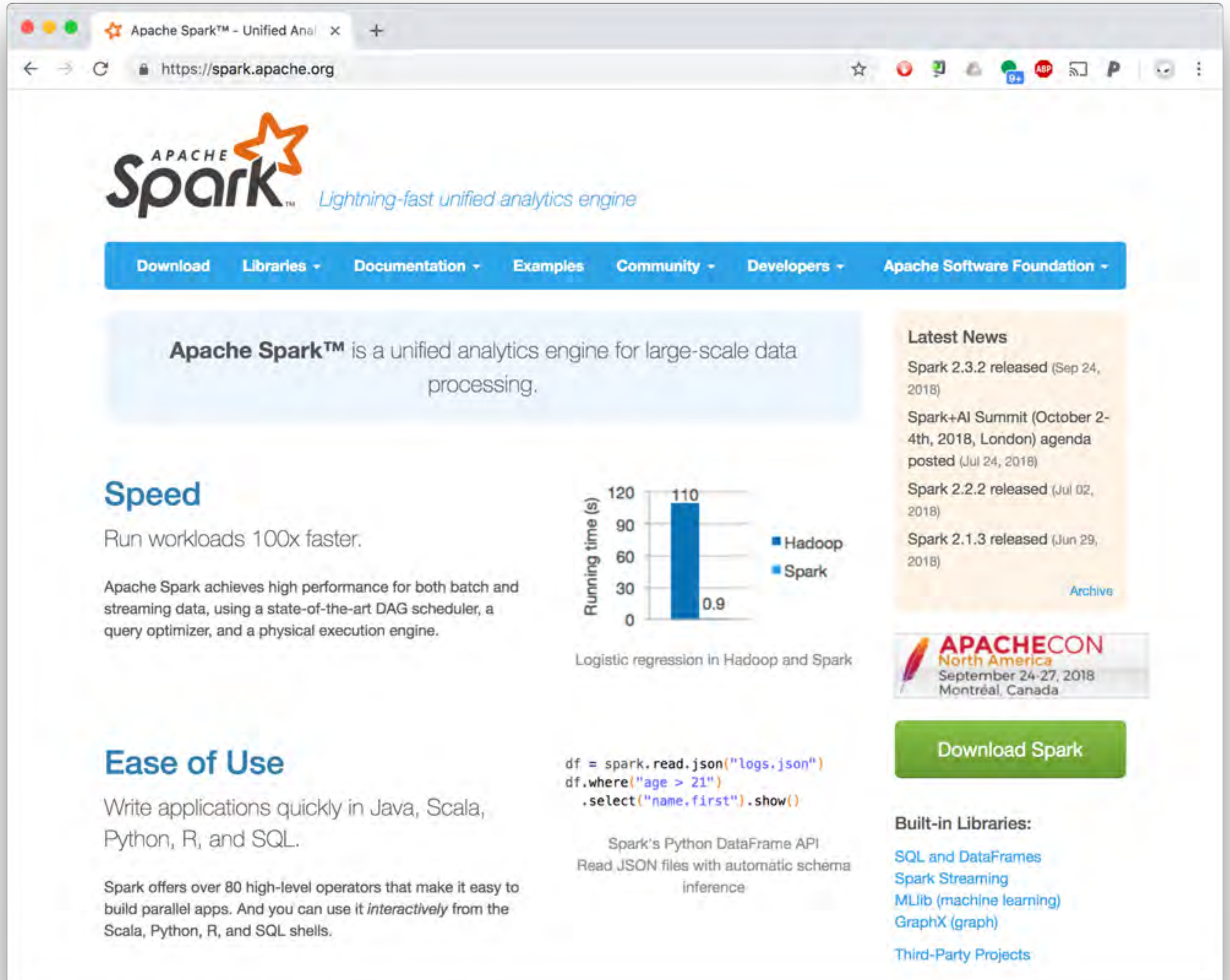
Industry vs. Astronomy (sometimes)



The Big Data Open Source Tools Landscape

https://datafloq.com/big-data-open-source-tools/os-home/

Writing Scalable Applications: MapReduce and Apache Spark



Apache Spark is an open-source distributed general-purpose cluster-computing framework.

Spark provides an interface for programming entire clusters with implicit data parallelism and fault tolerance.

-- Wikipedia

Examples

Pi Estimation

Spark can also be used for compute-intensive tasks. This code estimates π by "throwing darts" at a circle. We pick random points in the unit square ((0, 0) to (1,1)) and see how many fall in the unit circle. The fraction should be π / 4, so we use this to get our estimate.

Python Scala Java

```
def inside(p):
    x, y = random.random(), random.random()
    return x*x + y*y < 1

count = sc.parallelize(xrange(0, NUM_SAMPLES)) \
             .filter(inside).count()
print "Pi is roughly %f" % (4.0 * count / NUM_SAMPLES)
```

https://spark.apache.org/examples.html

Scalability through MapReduce

Dean & Ghemawat (2004)

Map

$\{x_i\}$ ---map--> $\{y_i=f(x_i)\}$

Apply a function *f* to every element of dataset *X*, producing dataset *Y*

Reduce

$\{ (k_i, v_{ij}) \} \rightarrow \{ y_i=(k_i,f(\{v_{ij}\})) \}$

Apply a function *f* to all values with a common key

Example:

{ ("dog", 2), ("dog", 1), ("cat", 3), ("dog", 2), ("cat", 1) }

-> reduce w. *sum()* ->

{ ("dog", 5), ("cat", 4) }

Examples

{ (“dog”, 2), (“dog”, 1), (“cat”, 3), (“dog”, 2), (“cat”, 1) }

-> reduce w. *sum()* ->

{ (“dog”, 5), (“cat”, 4) }

Word Count

In this example, we use a few transformations to build a dataset of (String, Int) pairs called `counts` and then save it to a file.

Python | Scala | Java

```
text_file = sc.textFile("hdfs://...")
counts = text_file.flatMap(lambda line: line.split(" ")) \
             .map(lambda word: (word, 1)) \
             .reduceByKey(lambda a, b: a + b)
counts.saveAsTextFile("hdfs://...")
```

https://spark.apache.org/examples.html

Astronomy Example: Compute Light Curve Features

This works on arbitrarily large datasets!

In [10]:
```
from pyspark.sql.types import ArrayType, FloatType, DoubleType
from pyspark.sql.functions import col, pandas_udf, explode
import pandas as pd

import cesium
from cesium.time_series import TimeSeries
from cesium.featurize import featurize_single_ts, featurize_time_series

##########

features_to_use = ["amplitude", "percent_beyond_1_std", "maximum", "max_slope",
                   "median", "median_absolute_deviation", "percent_close_to_median",
                   "minimum", "skew", "std", "weighted_average"]

ls_features = ["freq1_amplitude1", "freq1_amplitude2", "freq1_amplitude3",
               "freq1_amplitude4", "freq1_freq", "freq1_lambda", "freq1_rel_phase2",
               "freq1_rel_phase3", "freq1_rel_phase4", "freq1_signif", "freq2_amplitude1",
               "freq2_amplitude2", "freq2_amplitude3", "freq2_amplitude4", "freq2_freq",
               "freq2_rel_phase2", "freq2_rel_phase3", "freq2_rel_phase4"]

def featurize_udf(mjd, psfflux):
    feat_outs = []
    for row_mjd, row_psfflux in zip(mjd, psfflux):
        feat_out = featurize_time_series(np.array(row_mjd), np.array(row_psfflux),
                                         features_to_use=features_to_use + ls_features)
        feat_outs.append(feat_out.values.flatten())
    return pd.Series(feat_outs)

##########

feat_udf = pandas_udf(featurize_udf, returnType = ArrayType(DoubleType()))
spark_session.udf.register("FEATURIZE", feat_udf)

pdf = ztf.where("SIZE(mjd)>50").selectExpr("FEATURIZE(mjd, psfflux)").toPandas()
```

Cesium (Naul, 2016), Astronomy eXtensions for Spark (Zecevic+ 2018)

The Result (with apologies for the appallingly poor visualization)

```
In [12]: pdf = ztf.where("SIZE(mjd)>50").limit(10).selectExpr("ADDMJ(mjd, psfflux)").toPandas()

Out[12]: [Row(ADDMJ(mjd, psfflux)=[1925.2211608886719, 0.13978494623655913, 3987.0869140625, None, 344
         6.4375, 152.688720703125, 0.7419354838709677, 136.64459228515625, -2.4431908318547433, 631.64
         34156713688, 3189.848529118364, 313.63354429378256, 14.639380553238073, 1.3644581456964708,
         1.7500900946095723, 1.5124216699725148, 42.94394862399773, 2.5313200999890264, 0.635479246010
         1448, -0.5061265295340045, 5.0589783305487925, 297.2971196481153, 49.95079199367568, 5.377033
         053881004, 2.9892154859975197, 4.01272141380738, -0.11956697923508663, 0.44717905887839726, -
         1.1813749683927528]),
          Row(ADDMJ(mjd, psfflux)=[208.086181640625, 0.17857142857142858, 848.4930419921875, 57922.265
         60191954, 561.7611083984375, 21.10723876953125, 0.7142857142857143, 432.3206787109375, 1.7943
         474250678484, 58.001638792839145, 561.3877334594727, 41.09281808433971, 3.6566340247461695,
         0.484343106761218, 0.3561043472427892, 1.9945387332853752, 36.82950672444141, -2.622502373517
         709, 1.901872187975775, 1.3933600656842617, 3.208573392947105, 26.071430553383156, 1.26390776
         08450804, 0.47282767919988666, 0.11065610514135481, 23.0800641543819, -2.85326837788731, 1.89
         44949539332152, -0.3698982601601857]),
          Row(ADDMJ(mjd, psfflux)=[491.52618408203125, 0.1320754716981132, 1122.197509765625, 32600.79
         9800087658, 402.0052490234375, 31.34088134765625, 0.7924528301886793, 139.1451416015625, 2.88
         3311815955483, 131.10471926999602, 416.95930855229216, 114.34083688267272, 10.76770241852664
         3, 4.481966442155684, 1.2482773020089568, 1.0816044062094867, 24.405764071747054, -1.35716539
         37044816, -2.4773879544783286, 2.77126228292522, 3.23567097891298, 60.04996221378995, 5.37096
         0993975372, 0.8751468632528988, 0.3325519757277276, 0.6780459096769216, -0.8609512237261738,
         -1.5228145978972758, 1.6412947076259528]),
          Row(ADDMJ(mjd, psfflux)=[107.95289611816406, 0.17333333333333334, 303.4608154296875, 15769.0
         08109654327, 146.07757568359375, 16.181289672851562, 0.64, 87.55502319335938, 1.9784213046179
         848, 33.608624825862975, 152.7379244995117, 14.807230580579672, 0.2625194273012284, 0.1712371
         6298087752, 0.1045290010068471, 29.250142510335483, 75.41574345145936, -1.005145004902362, 2.
         1760284474899296, 1.1339028285326511, 2.980258224087021, 18.704151766762745, 1.06178361034372
         75, 0.283697089687251, 0.18280423388998288, 0.965224836036898, 0.24573780352703498, -1.282736
         1427620787, -2.016597738480162]),
          Row(ADDMJ(mjd, psfflux)=[1531.0211791992188, 0.14285714285714285, 4344.10595703125, None, 39
         23.2584228515625, 160.582275390625, 0.6587301587301587, 1282.0635986328125, -2.22651166992487
         46, 582.6092968330478, 3688.5484958224824, 387.9816297071594, 48.53561353536267, 12.997457710
         221171, 6.372125562326946, 3.0098298953254314, 13.648208330625398, 0.4783521707732701, -0.520
         8550287735074, 0.9412748273621686, 6.5045045280108695, 236.76237021479648, 3.799074459364612,
```

Want to try it out?

```
conda install -c conda-forge pyspark
```

ANACONDA

CONDA-FORGE

Scaling with Spark

Today, Spark is being adopted by major players like Amazon, eBay, and Yahoo! Many organizations run Spark on clusters with thousands of nodes. According to the Spark FAQ, the largest known cluster has over 8000 nodes. Indeed, Spark is a technology well worth taking note of and learning about.

But where do I find 8000 nodes?

https://www.toptal.com/spark/introduction-to-apache-spark

+ government-sponsored private clouds (e.g., JetStream)

Cloud services

- Essentially, companies that rent computers (or a few million of them)
  - The same for storage.
- Pay only for what you use (by the second/minute/hour)
- Scalable: ask for 1000 machines, get a 1000 machines
- Becoming cost effective (TCO)
  - Especially “spot” pricing

Meeting the Challenges

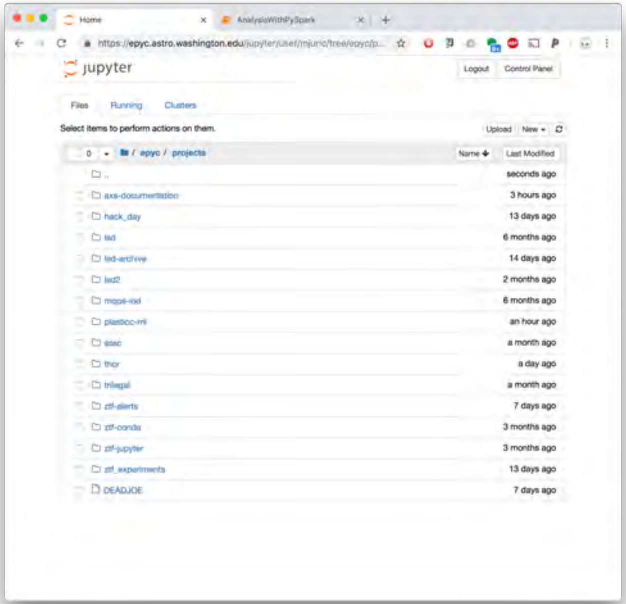
Dataset Storage

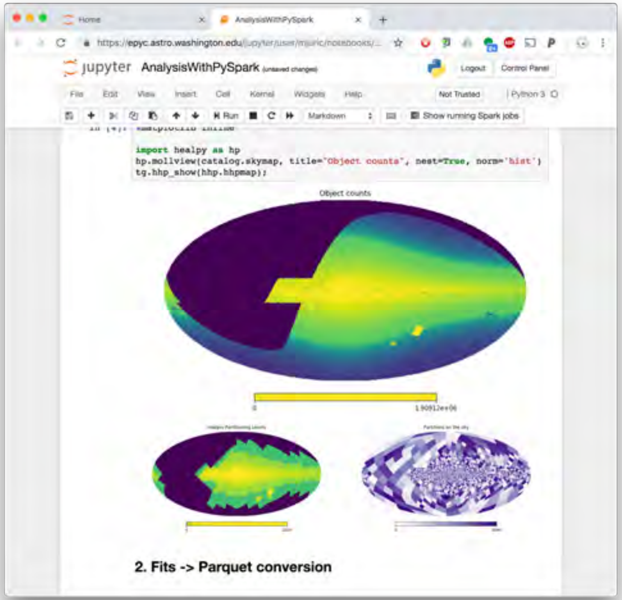
Resources

Scalable Analysis Code

Interface

"Analysis 2025"





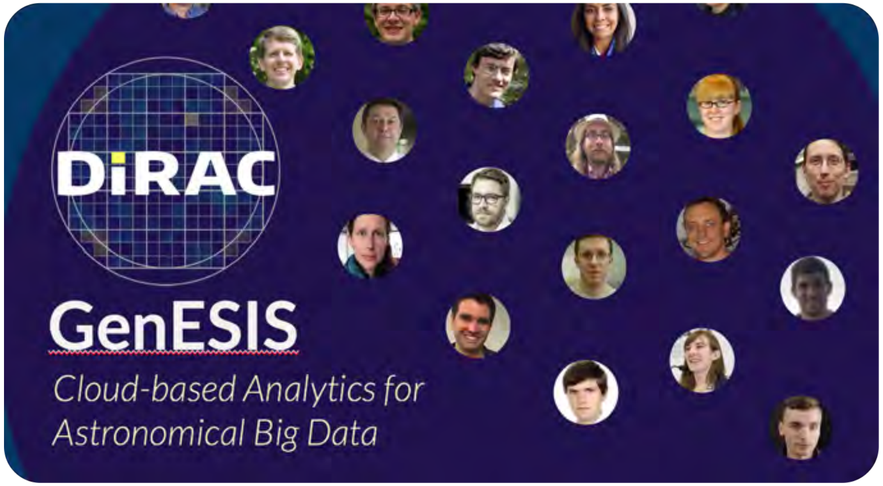
A Number of Projects are Working to Make this Happen

PANGEO

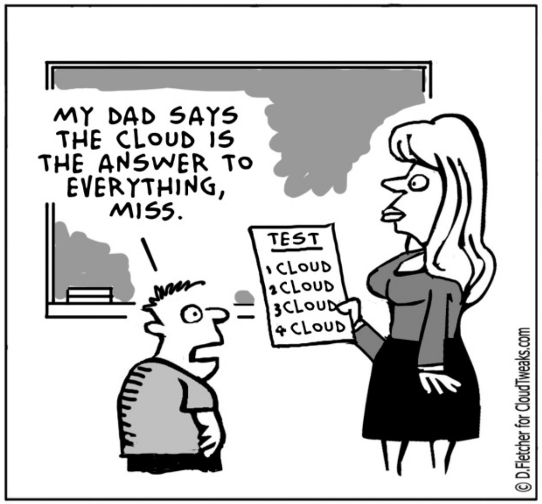
A community platform for Big Data geoscience

http://pangeo.io/

Coming soon w. ZTF !



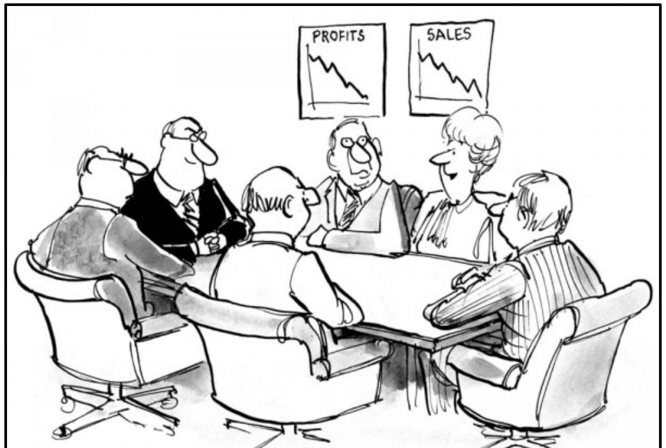
Some Words of Caution



Just like with machine learning / A.I., there's no need to throw cloud at everything.

Small datasets?
Large-ish datasets?

But the *programming model* works across all scales.



"What if we don't change at all ... and something magical just happens?"

The implementation of these technologies is still in its infancy. They change incredibly quickly.

Expect you may need to shift from framework to framework (e.g., Spark → Dask).

That said, the *programming models* change on a much longer timescale (e.g., MR 2004 -> ?).

Summary

> After decades of planning and construction, LSST is coming soon! First light in 2020, science commissioning in 2021, start of operations in October 2022! LSST is around the corner.

> Remote analysis platforms are being set up to "bring the code to the data" and lower the barrier to entry to working with the dataset. Remote access through Jupyter is becoming a standard.

> Large scale (~PB), end-user analysis remain an unsolved problem (both software and resources) in academia. Adopting cloud-ready solutions from the industry is one way forward. MapReduce and related frameworks (Spark, Dask, etc.) will play a large role.