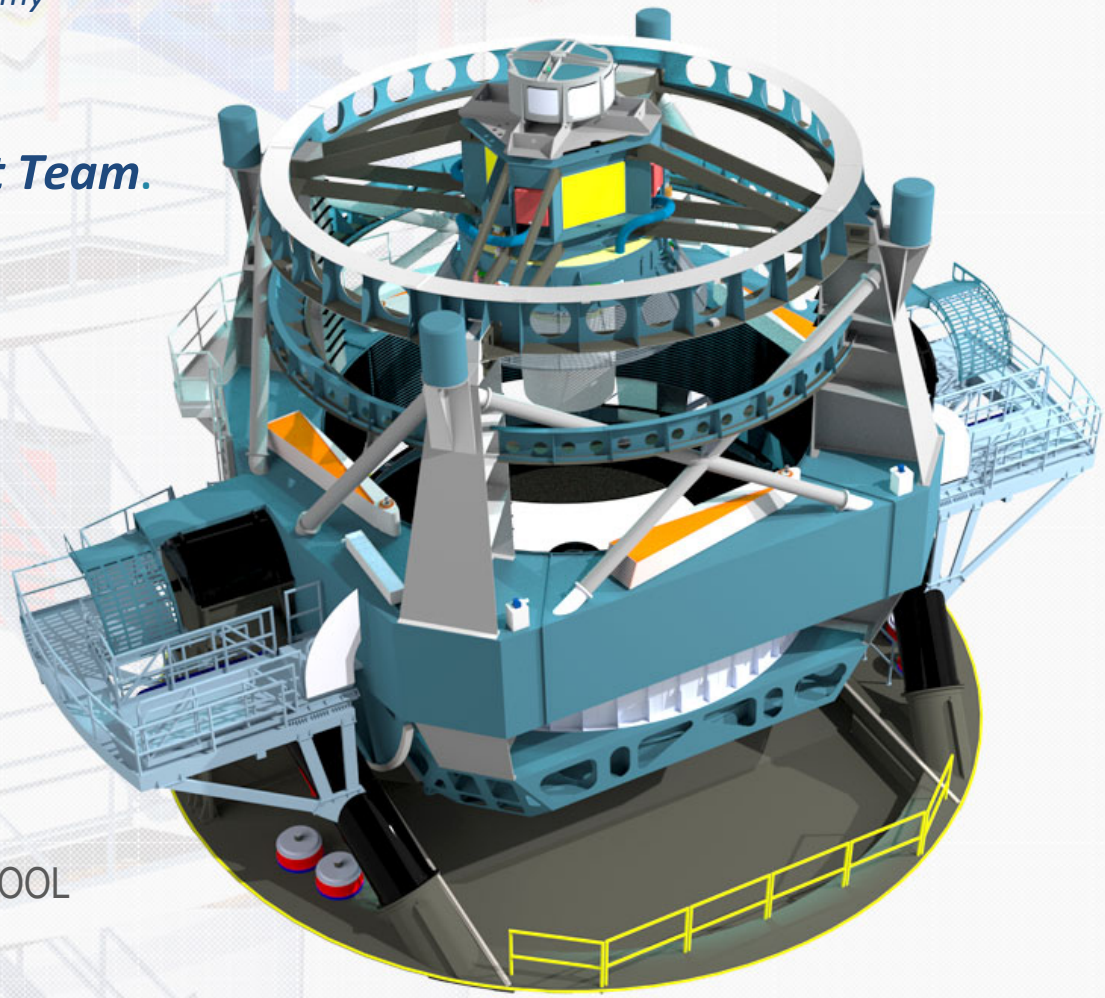
The Era of Large Surveys: What will LSST deliver (and how to think about it)

Mario Juric
University of Washington, Professor of Astronomy
LSST Data Management Subsystem Lead

and the LSST Data Management Team.



DESC 2016 SUMMER SCHOOL
Oxford, UK, July 16th, 2016

Finding out more

- ***"The LSST Data Management System"*** paper
  (Juric et al. 2015; http://arxiv.org/abs/1512.07914)

- **"The Data Products Definition Document"** paper
  (Juric et al. 2015; http://ls.st/dpdd)

- ***"LSST: from Science Drivers to Reference Design and Anticipated Data Products"*** paper
  (Ivezic et al. 2008; https://arxiv.org/abs/0805.2366)

- http://dm.lsst.org and http://community.lsst.org

Polls, questions, referenda …

- Pat: “Make sure the audience is engaged!”
- MJ: “*Maximum effort!*  “

Polls, questions, referenda …

- Pat: “Make sure the audience is engaged!”
- MJ: “*Maximum effort!*  “

There will be many questions & quick quizzes during this lecture. We’ll capture & show the results in real time!

Open:

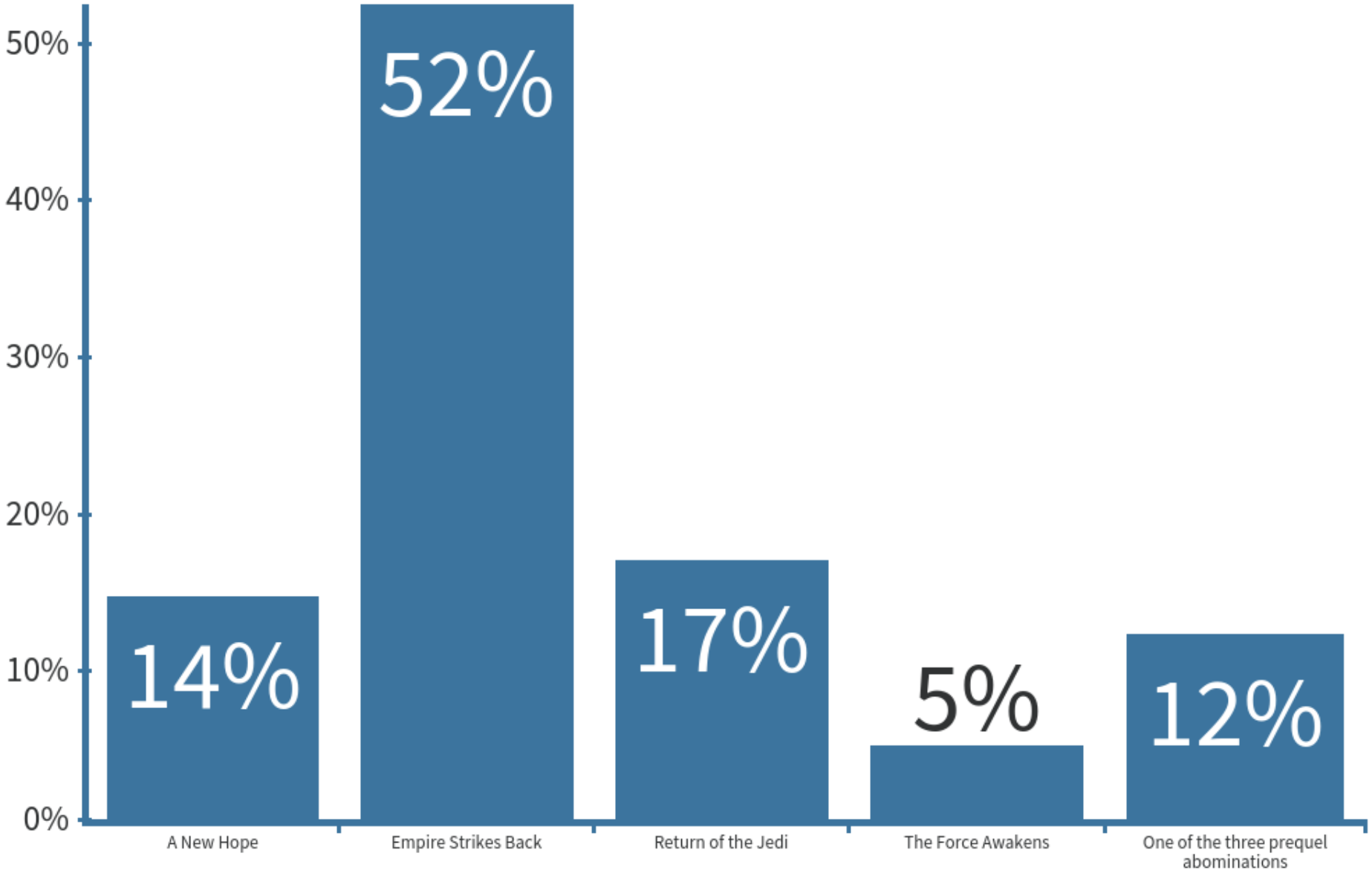
`PollEv.com/lsst`

… and keep it open.

n.b: works great on your cell phone!

Favorite Star Wars?

When poll is active, respond at **PollEv.com/lsst**

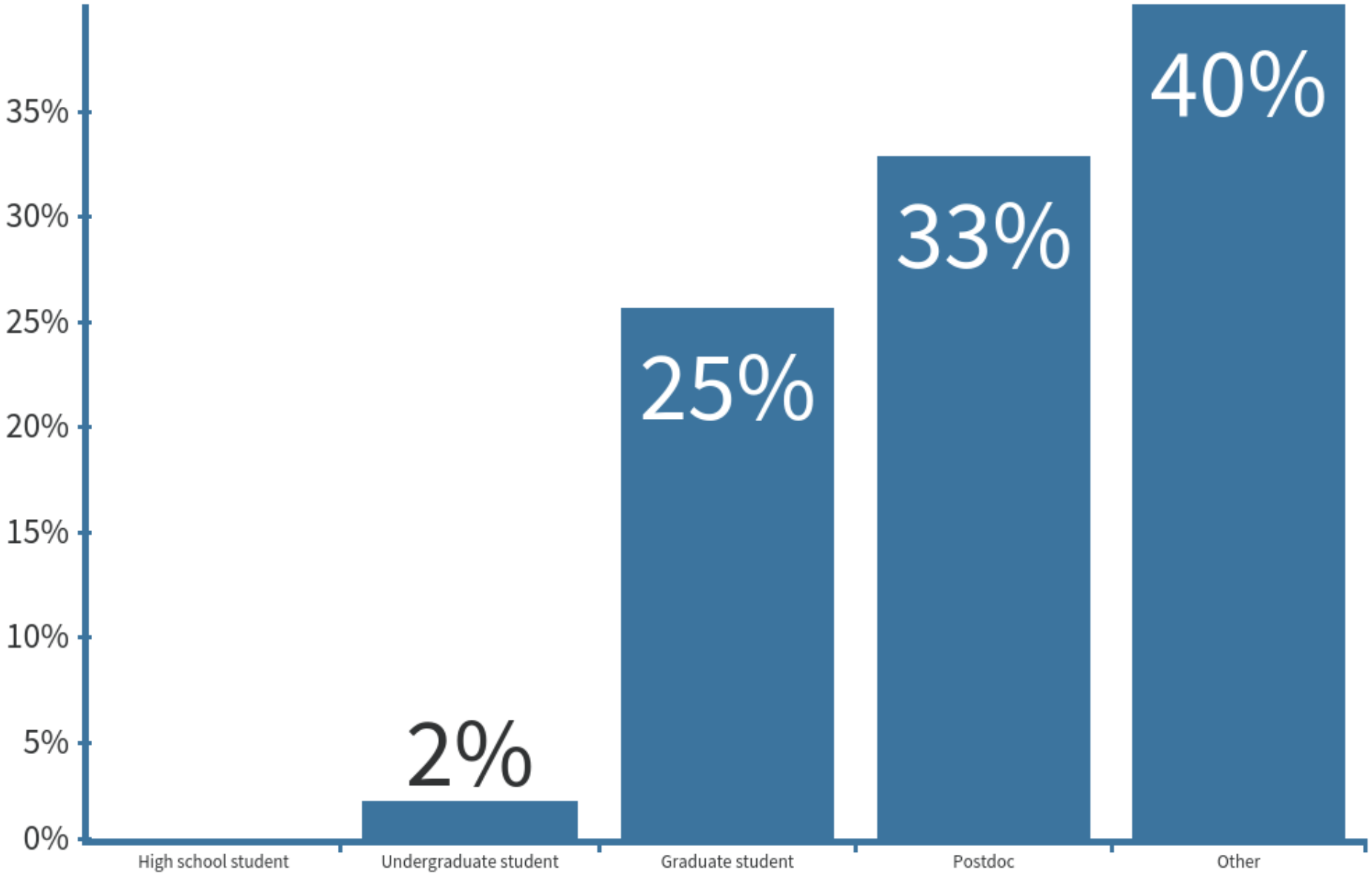


Your first name?

When poll is active, respond at **PollEv.com/lsst**

Stage of your career

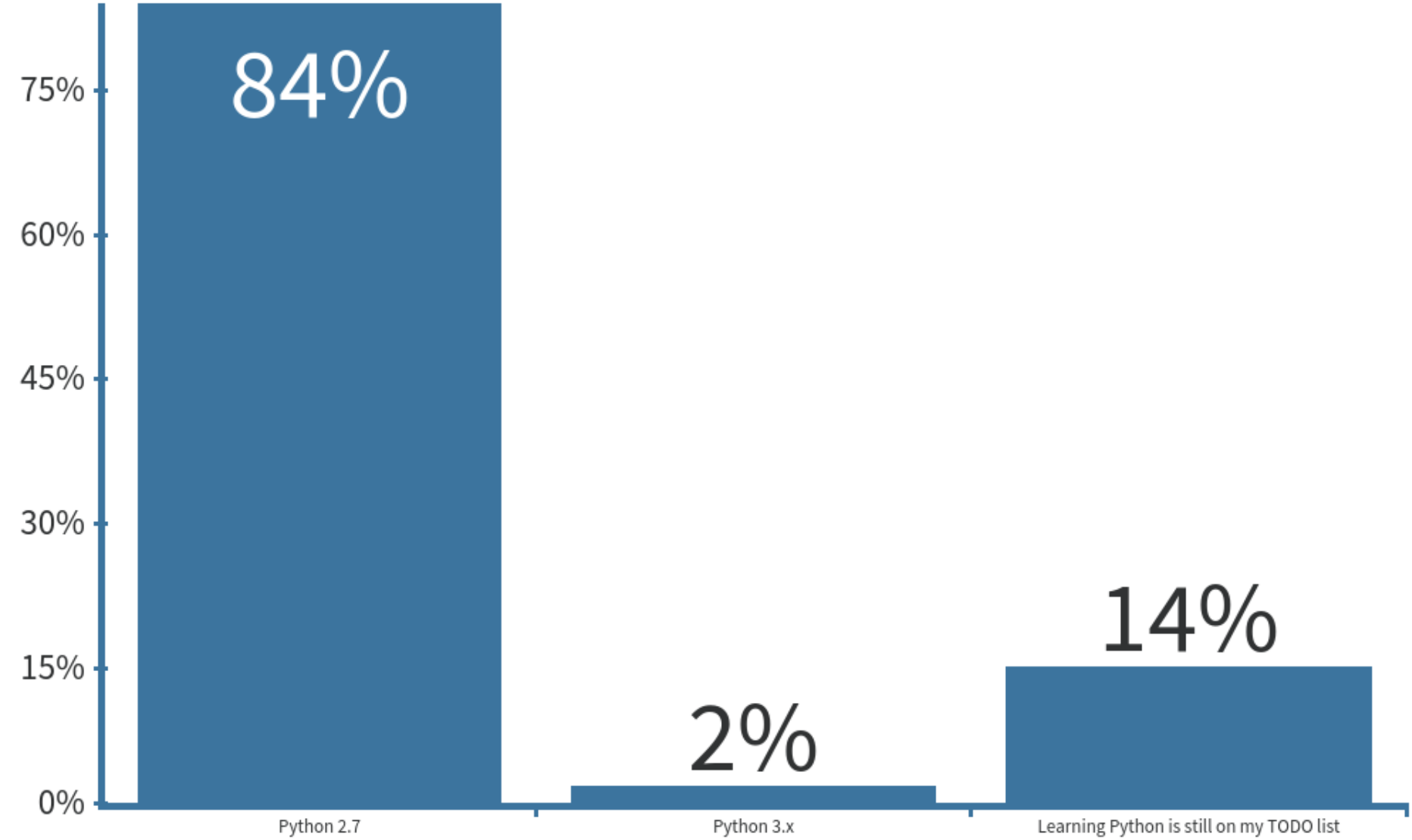
When poll is active, respond at **PollEv.com/lsst**



The version of Python I'm most comfortable with

When poll is active, respond at **PollEv.com/lsst**

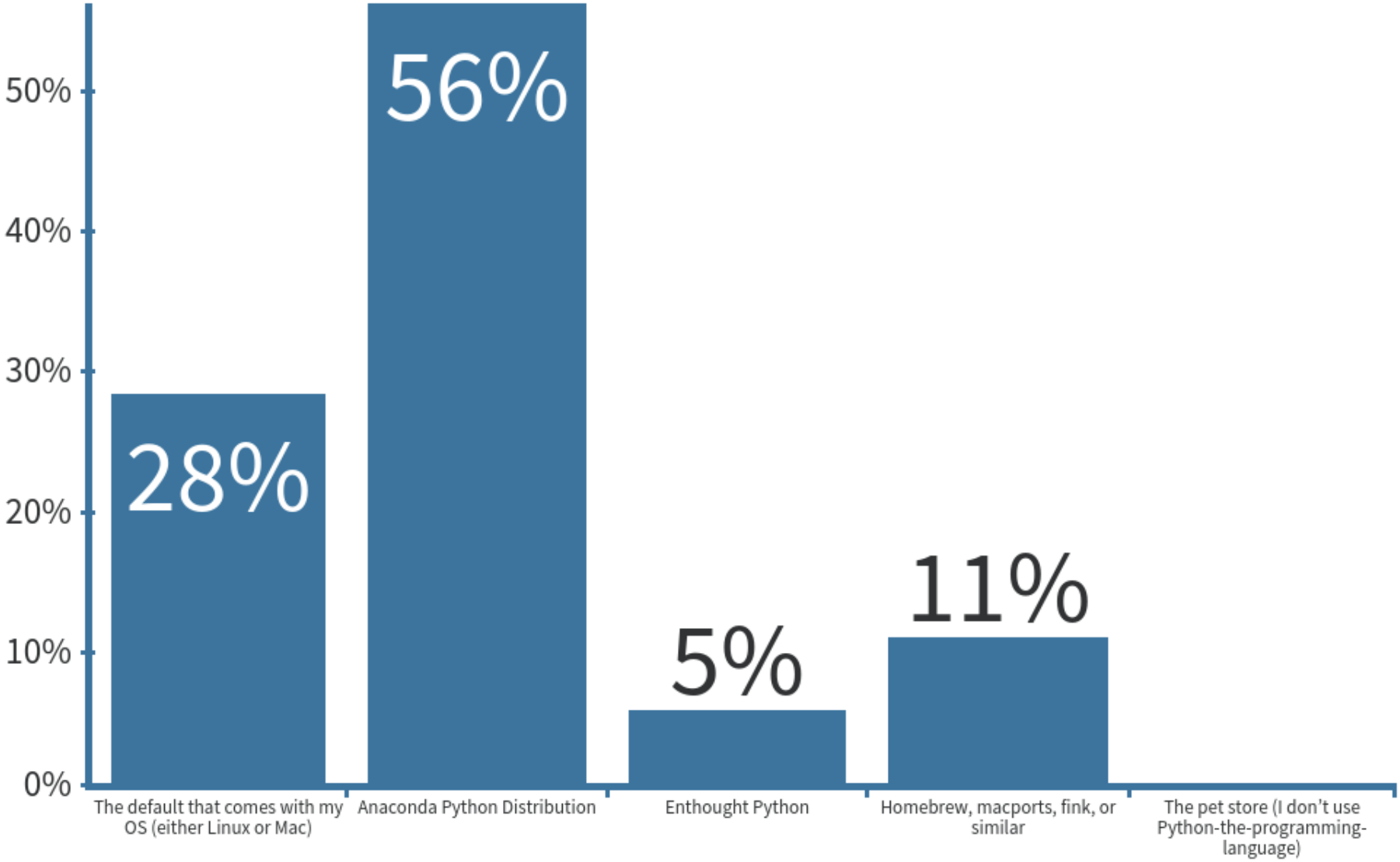
Text **LSST** to **020 3322 5822** once to join



How I got my Python

When poll is active, respond at **PollEv.com/lsst**

Text **LSST** to **020 3322 5822** once to join



- Large Surveys (and Why They're Different)

Hipparchus of Rhodes (180-125 BC)

Discovered the precession of the equinoxes.

Measured the length of the year to ~6 minutes.

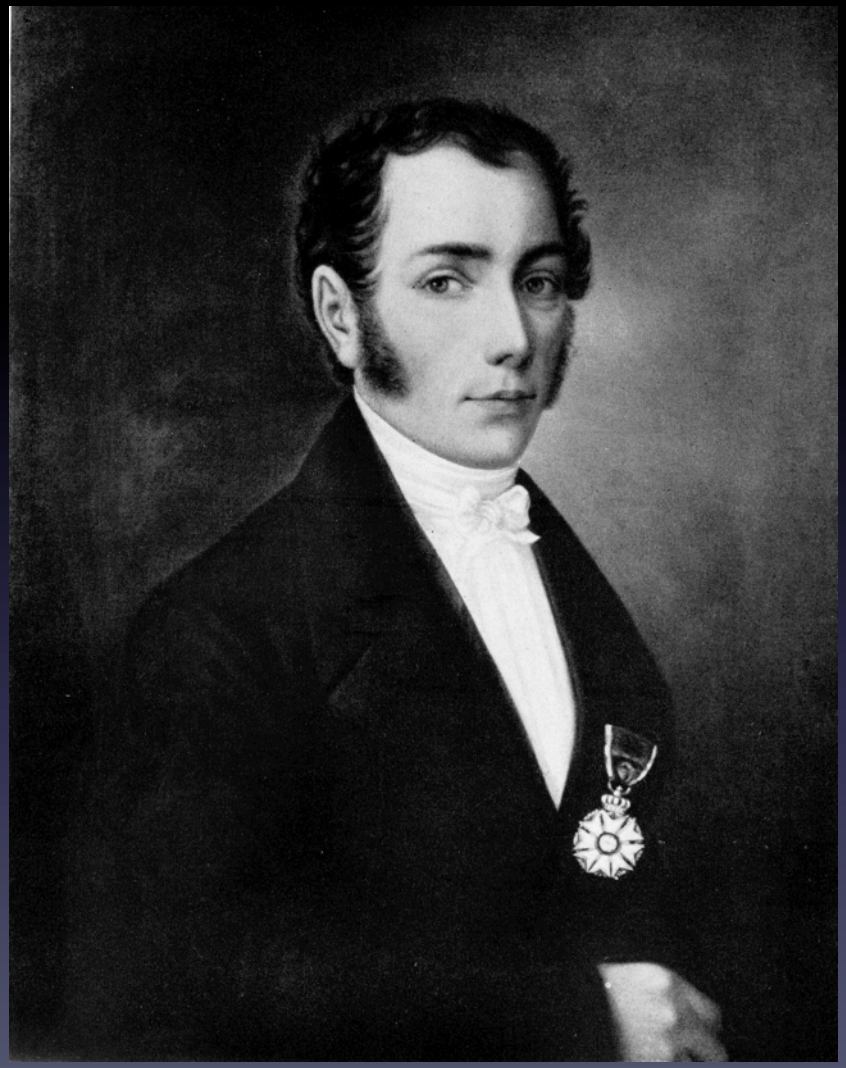
In 129 BC, constructed one of the first star catalogs, containing about 850 stars.

n.b.: also the one to blame for the magnitude system …

Galileo Galilei (1564-1642)

Researched a variety of topics in physics, but called out here for the introduction of the *Galilean telescope*.

Galileo's telescope allowed us for the first time to *zoom in* on the cosmos, and study the individual objects in great detail.



Joseph von Fraunhofer (1787-1826)

Mounted a prism in front of an objective of a small telescope, and pointed it to the Sun.

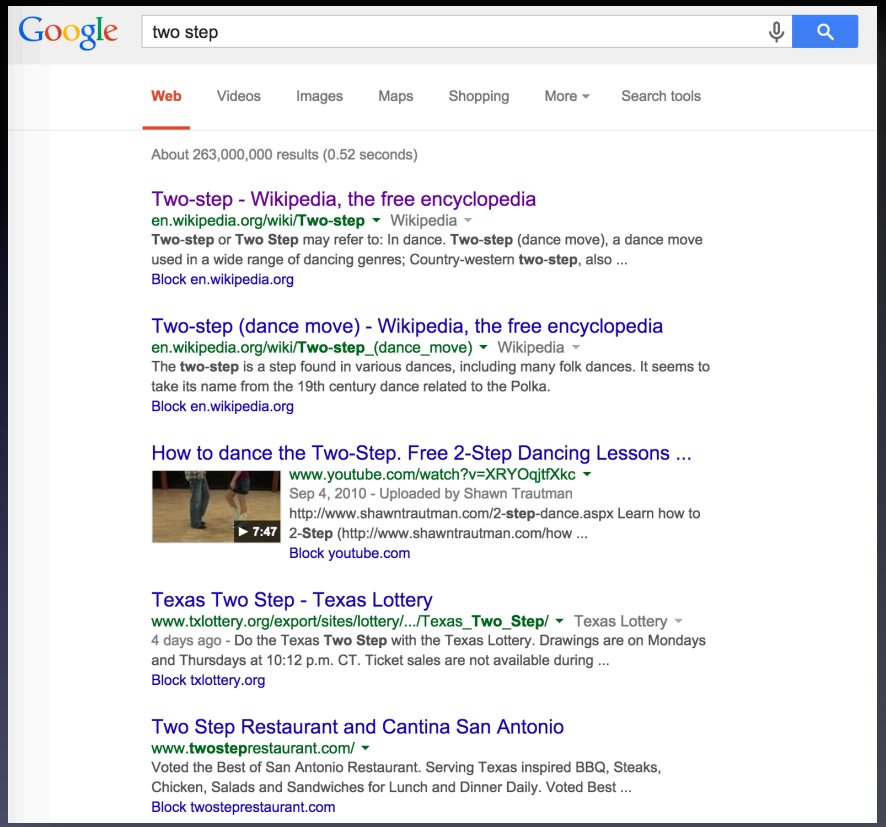
In 1859, Kirchhoff and Bunsen understood what Fraunhofer saw.

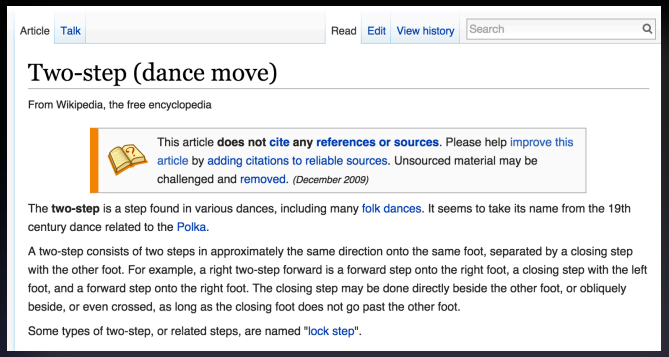
The birth of modern astrophysics!

The Astrophysics Two-Step

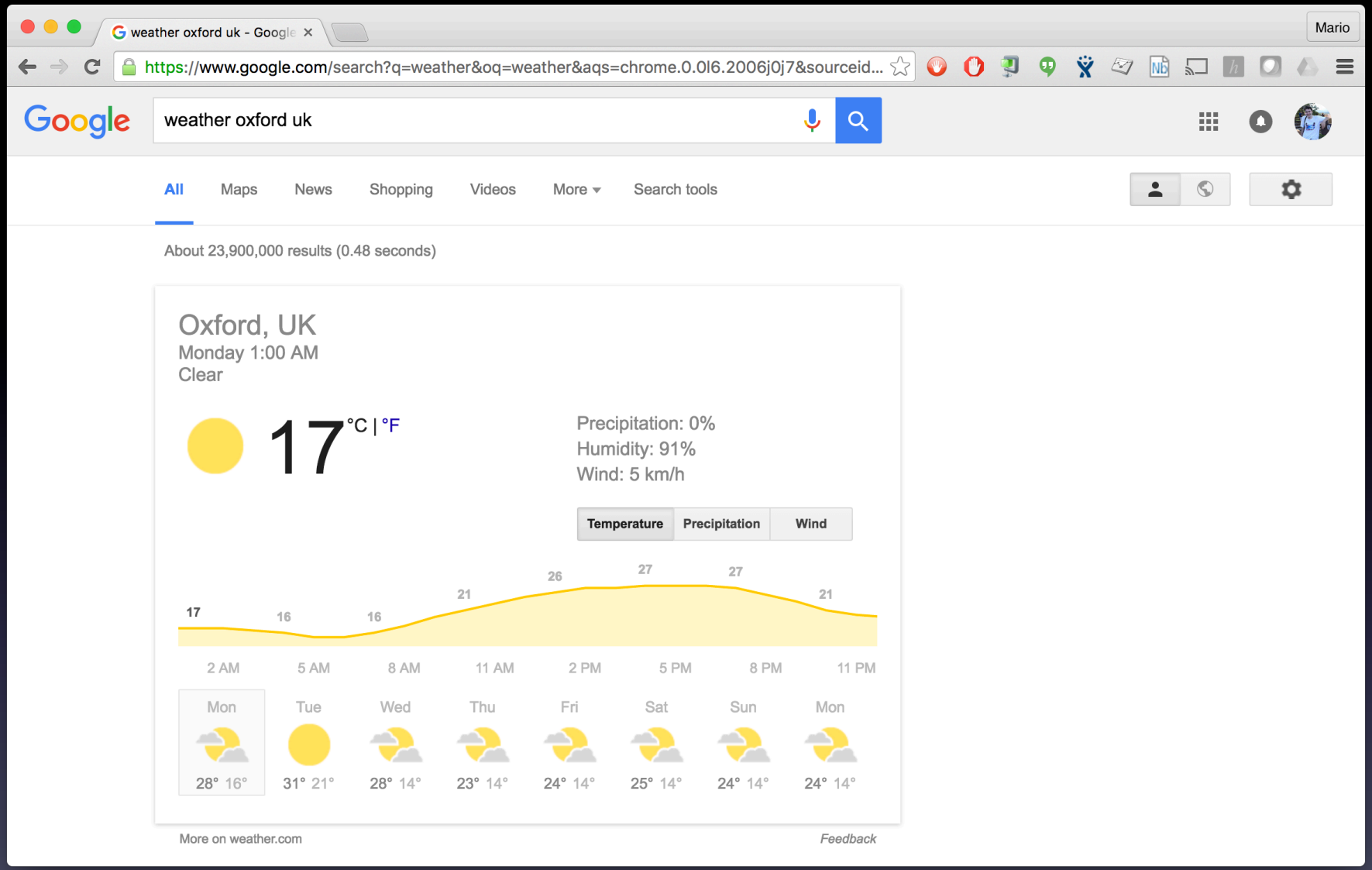
- Surveys
  - Construct catalogs and maps of objects in the sky. Focus on coarse classification and discovering targets for further follow-up.
- Large telescopes
  - Acquire detailed observations of a **few** representative objects. Understand the details of astrophysical processes that govern them, and **extrapolate that understanding to the entire class**.

Analogy: Web Search





Google's index is a catalog of the Web. We use it to "zoom in" on individual entries to find out more.



It's more than just a catalog of pointers – more and more, Google itself collects, processes, indexes, visualizes, and serves the actual information we need.

More and more often, our "research" begins and ends with Google!

Entering the Era of Massive Sky Surveys

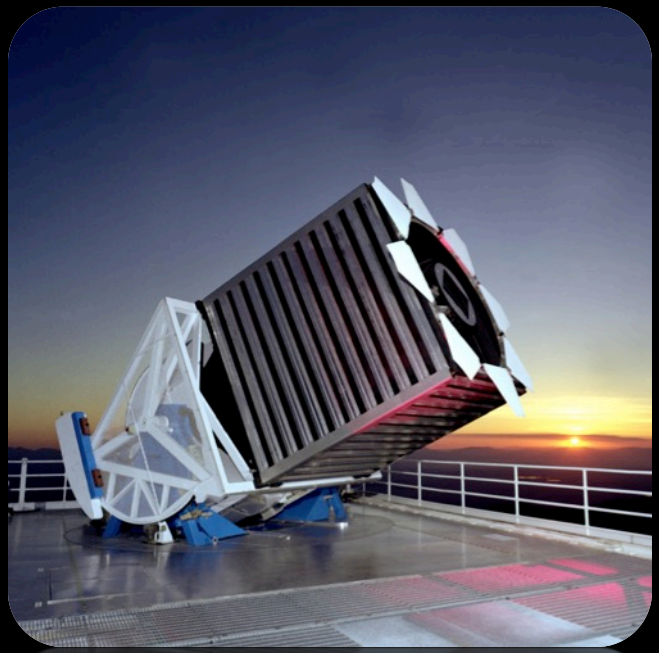
- There's a close parallel with large surveys in astronomy, in scale, quality, and richness of the collected information
  - Scale: We're entering the era when we can image and catalog the entire sky
  - Quality: The measurements can be as precise as those taken with "pointed" observations (used to be ~5-10x worse)
  - Richness: Those catalogs contain not only positions and magnitudes, but also shapes, profiles, and temporal behavior of the objects.

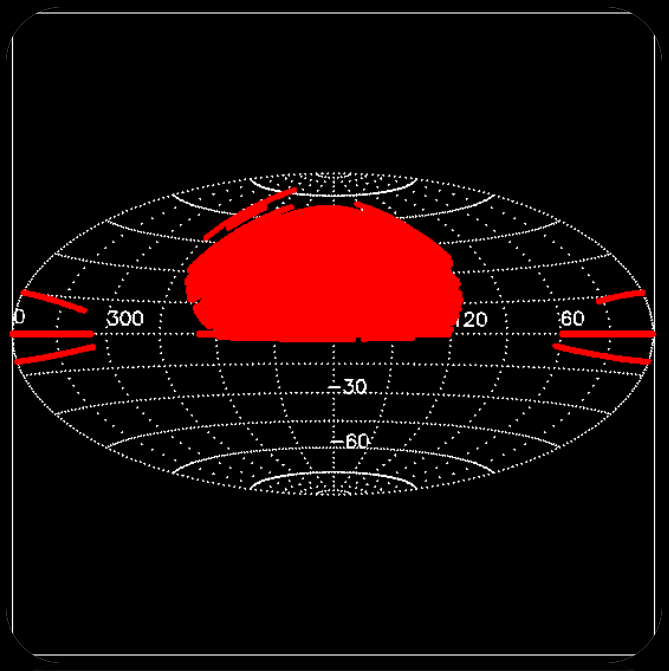
- Quite often, the research can begin and end with the survey.
- This is what makes large surveys of today **not just bigger, but better, more information rich, and therefore different**. **Big data is more about complexity and optimal knowledge extraction, than PBs.**

Sloan Digital Sky Survey

2.5m telescope >14500 deg^2 0.1" astrometry r<22.5 flux limit

5 band, 2%, photometry for >460M objects
Millions of spectra





10 years of ops: ~10 TB of imaging

Panoramic Survey Telescope and Rapid Response System

1.8m telescope 30,000 deg^2 50mas astrometry r<23 flux limit

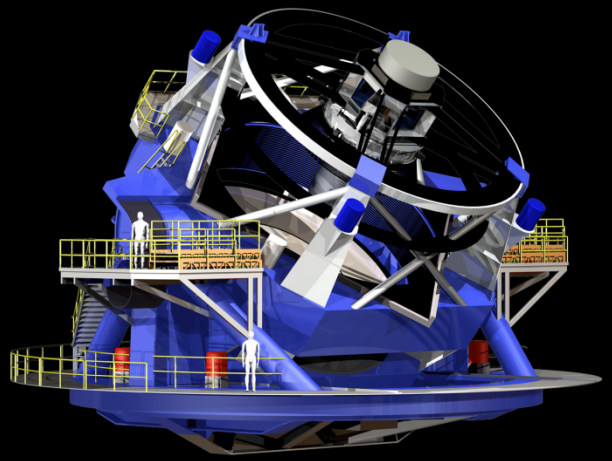
5 band, better than 1% photometry (goal)

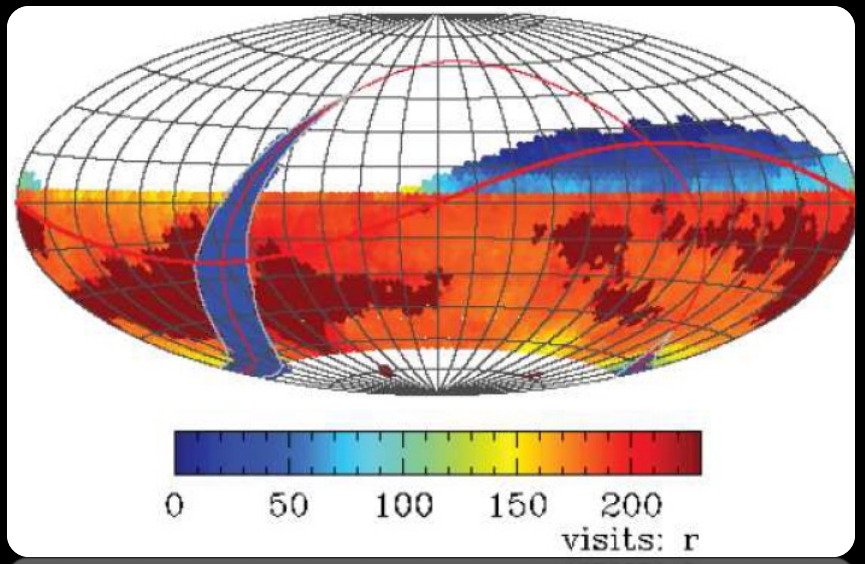
~700 GB/night

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

8.4m telescope | 18,000+ 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)

Location: Cerro Pachon, Chile

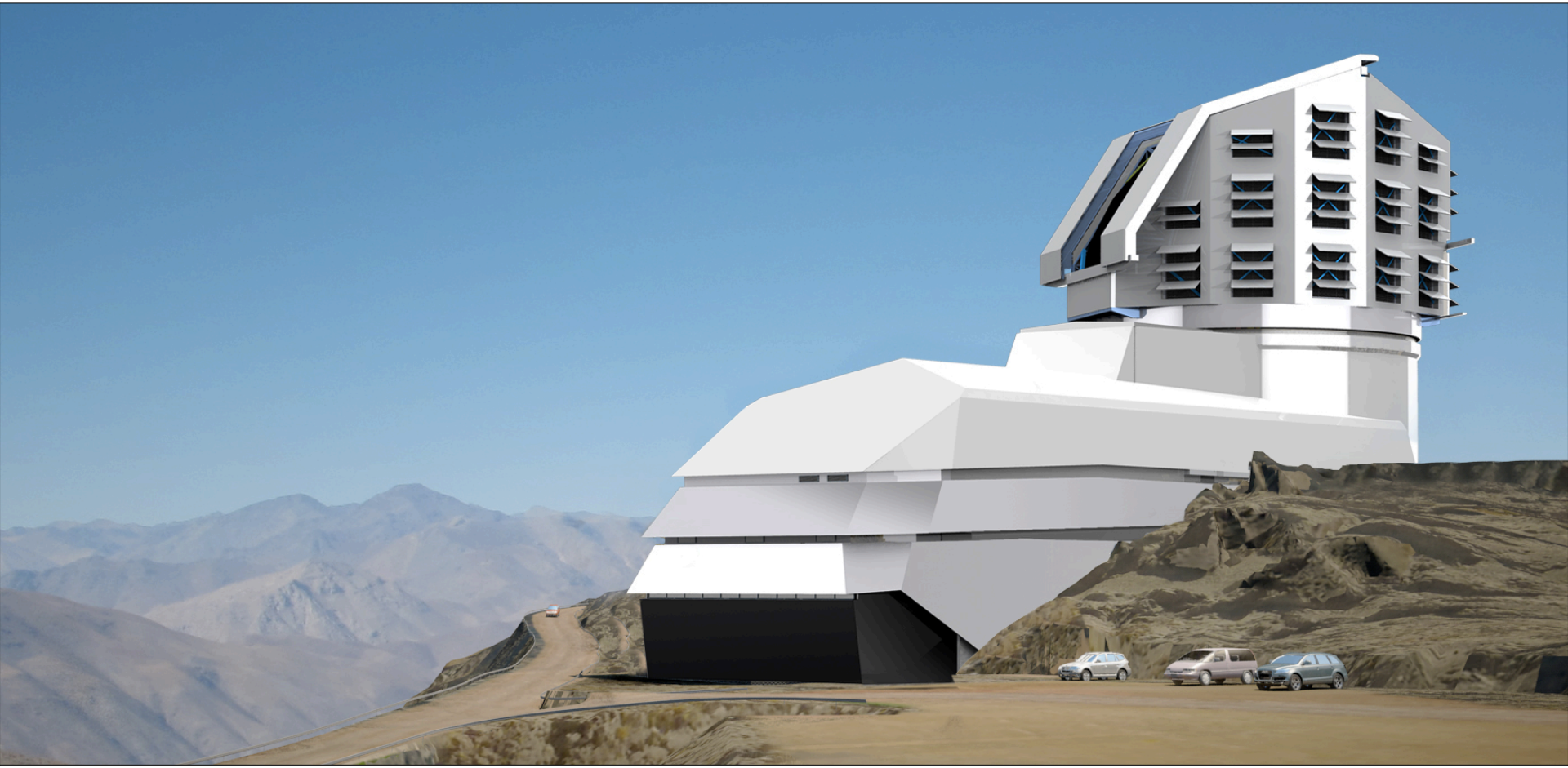
LSST Site (April 14th, 2015)

The Summit, yesterday.

The support building

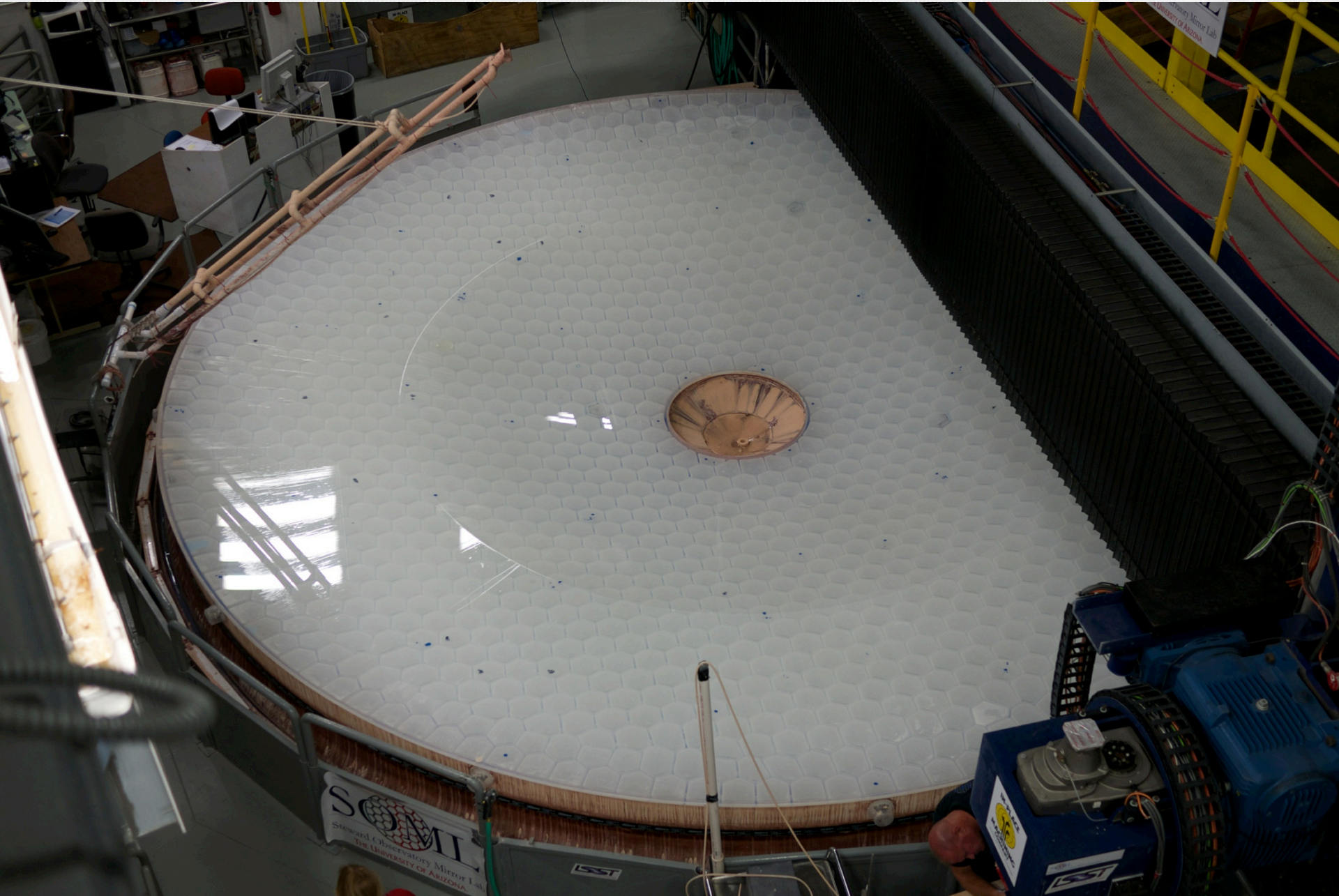
The Pier

LSST Observatory (cca. late ~2018)



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

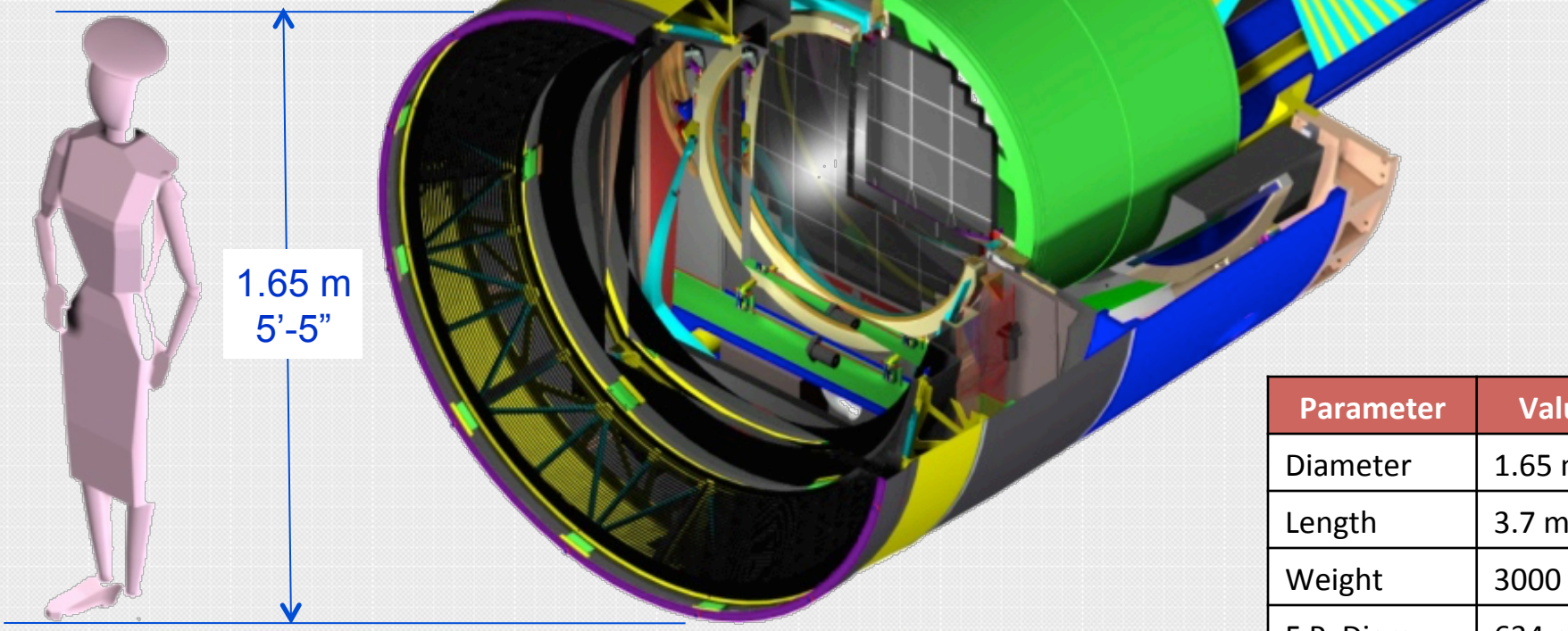
M1/M3 mirror: Done.



Mirror cell is being constructed

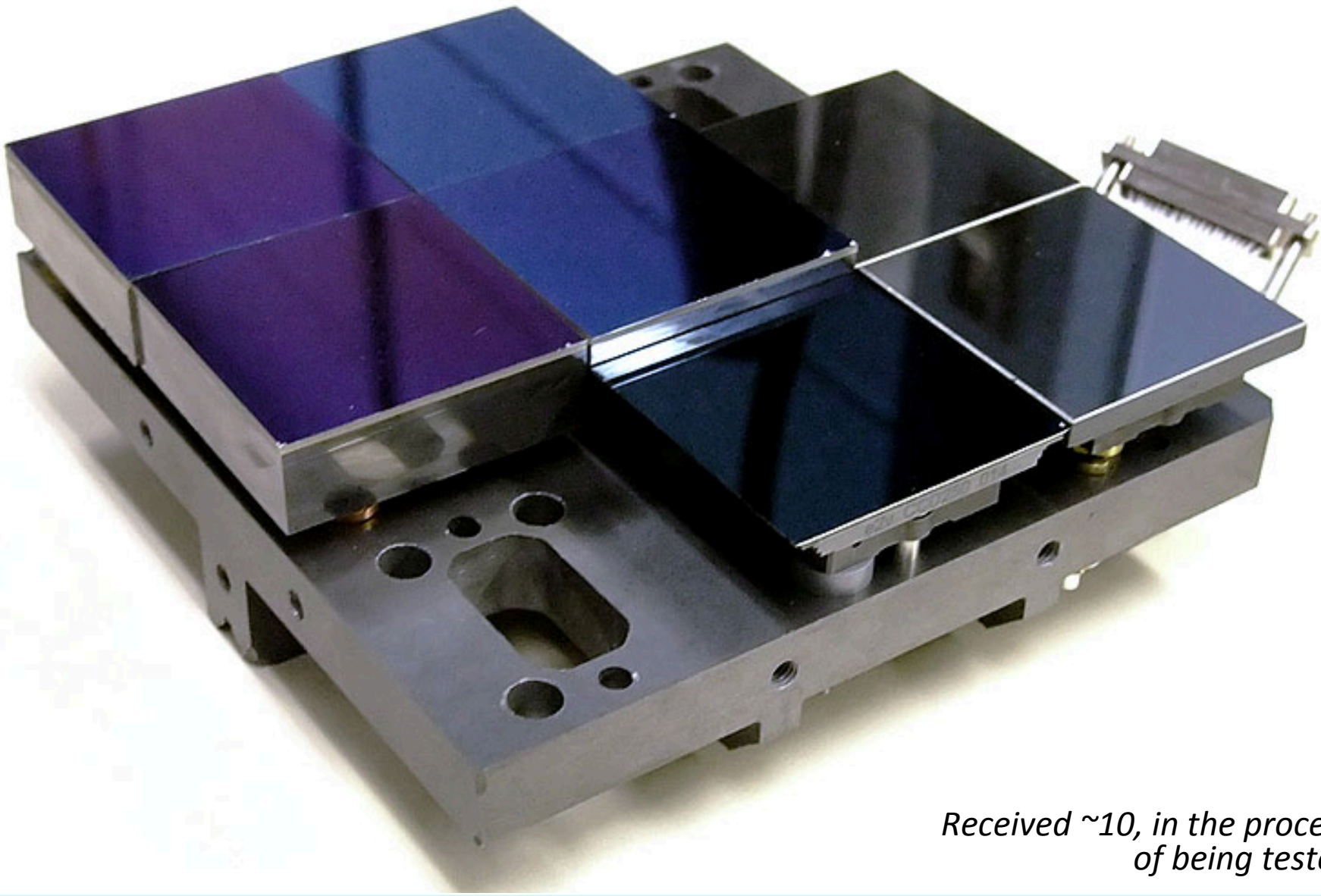
LSST Camera

- 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

Sensors (CCDs)



Received ~10, in the process of being tested

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

What will LSST deliver?

When poll is active, respond at **PollEv.com/lsst**

Database

(Dataset, Catalog, …)

The primary deliverable of LSST is not the telescope, nor the instruments, nor the raw data; *it is the fully reduced data products*.

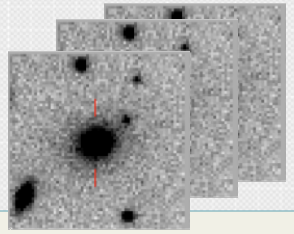
LSST is a *facility* that delivers *data products* and *data access and analysis services*.

LSST's #1 Data Management Challenge:

Creating useful general-purpose data products, while minimizing information loss.

From Data to Knowledge

Computationally (and cognitively) expensive, science-case speciific



And metadata!

Users Model ← *inference* – Data Facility

Users Model ← *inference* – Catalog ← Data Processing – Data

Users Users Facility Facility Facility

Computationally cheaper, Easier to understand, Science-case speciific

- Computationally expensive, general
- Reprojection; may or may not involve compression
- Almost always introduces some information loss
- Data Processing == Instrumental Calibration + Measurement

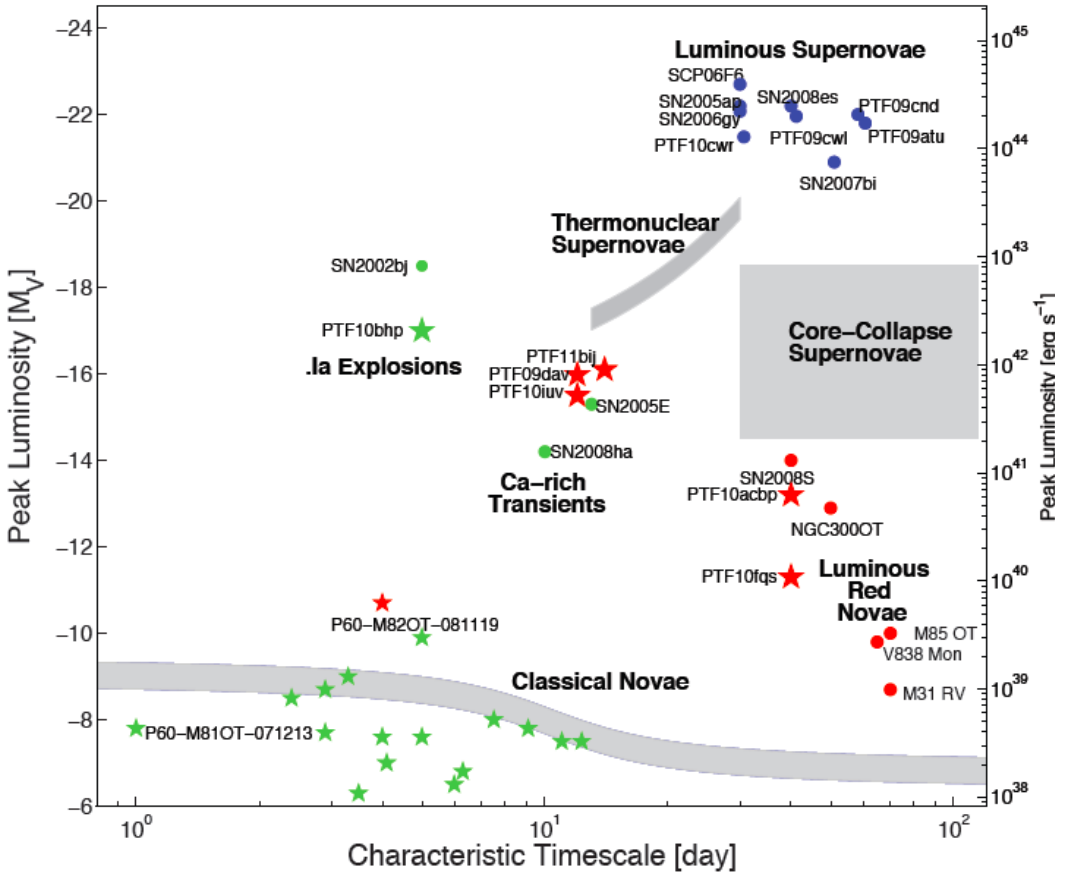
Guiding Principles for LSST Data Products

- There are virtually infinite options on what quantities (features) one can measure on images. But if catalog generation is understood as a *(generalized) cost reduction tool*, the guiding principles become easier to define:

1. **Maximize science enabled by the catalogs**
   - Working with images takes time and resources; a large fraction of LSST science cases should be enabled by our catalog data products.
   - Be considerate to the user: provide even sub-optimal measurements if they will enable leveraging of existing experience and tools
2. **Minimize information loss**
   - Provide (as much as possible) estimates of likelihood surfaces, not just single point estimators
3. **Provide and document the transformation (the software)**
   - Measurements are becoming increasingly complex and systematics limited; need to be maximally transparent about *how* they're done

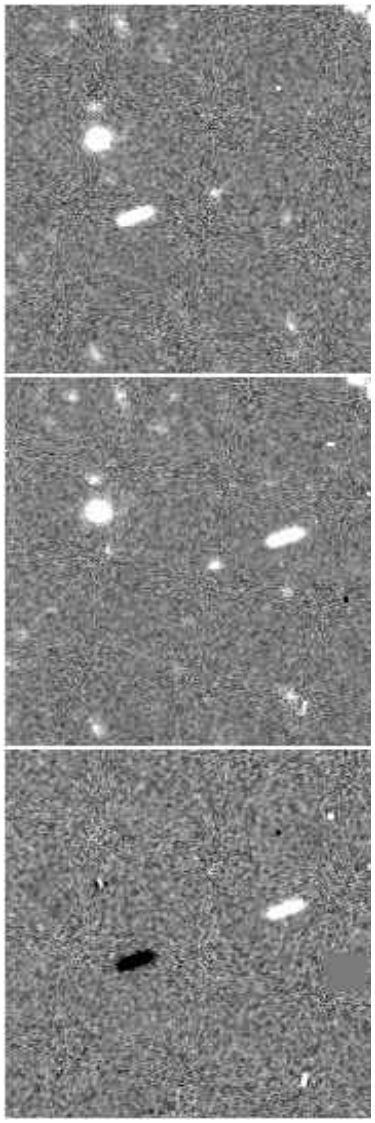
Discoveries we need to enable

- Time domain science
  - Nova, supernova, GRBs
  - Source characterization
  - Instantaneous discovery
- Census of the Solar System
  - NEOs, MBAs, Comets
  - KBOs, Oort Cloud
- Mapping the Milky Way
  - Tidal streams
  - Galactic structure
- Dark energy and dark matter
  - Strong Lensing
  - Weak Lensing
  - Constraining the nature of dark energy



Discoveries we need to enable

- Time domain science
  - Nova, supernova, GRBs
  - Source characterization
  - Instantaneous discovery
- Census of the Solar System
  - NEOs, MBAs, Comets
  - KBOs, Oort Cloud
- Mapping the Milky Way
  - Tidal streams
  - Galactic structure
- Dark energy and dark matter
  - Strong Lensing
  - Weak Lensing
  - Constraining the nature of dark energy



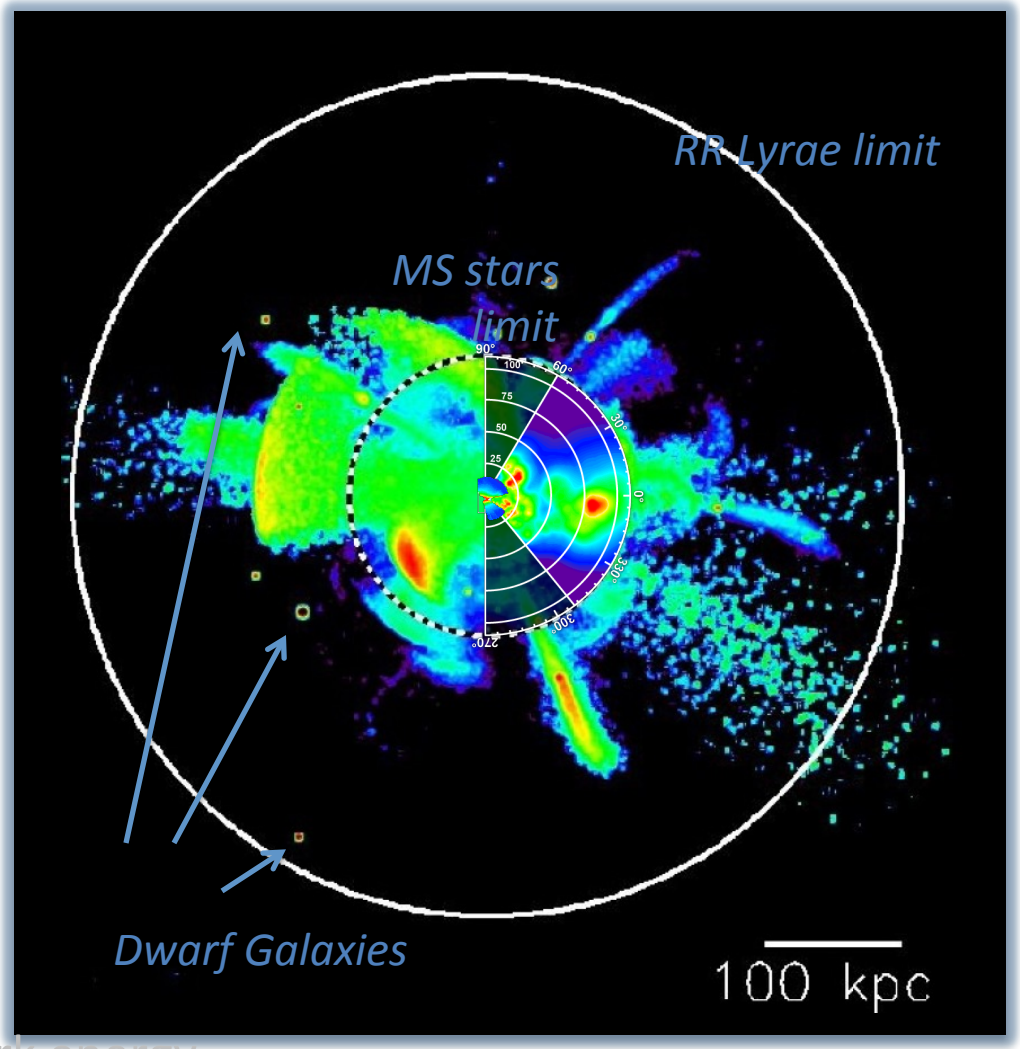
Exposure 1

Exposure 2

Exposure 1
-
Exposure 2

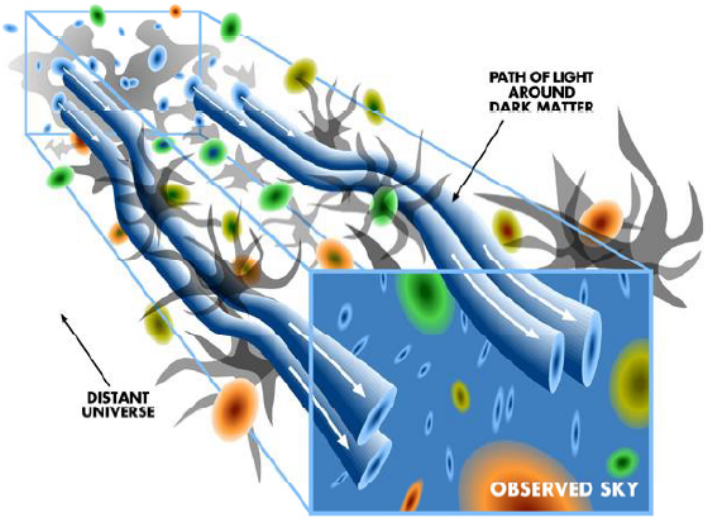
Discoveries we need to enable

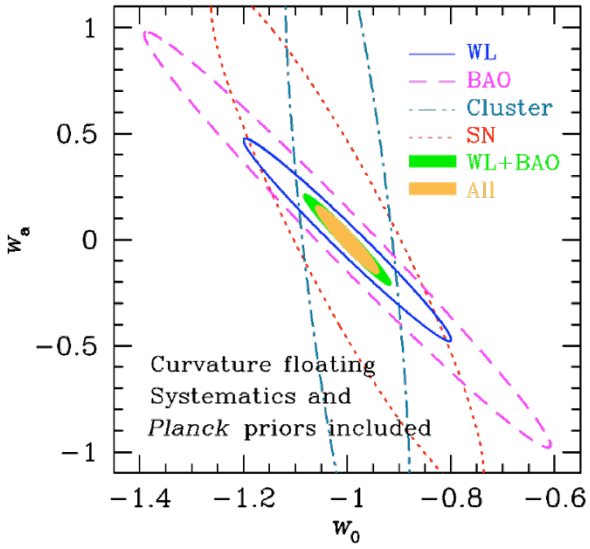
- Time domain science
  - Nova, supernova, GRBs
  - Source characterization
  - Instantaneous discovery
- Census of the Solar System
  - NEOs, MBAs, Comets
  - KBOs, Oort Cloud
- Mapping the Milky Way
  - Tidal streams
  - Galactic structure
- Dark energy and dark matter
  - Strong Lensing
  - Weak Lensing
  - Constraining the nature of dark energy



Discoveries we need to enable

- Time domain science
  - Nova, supernova, GRBs
  - Source characterization
  - Instantaneous discovery
- Census of the Solar System
  - NEOs, MBAs, Comets
  - KBOs, Oort Cloud
- Mapping the Milky Way
  - Tidal streams
  - Galactic structure
- Dark energy and dark matter
  - Strong lensing
  - Weak lensing
  - **Constraining the nature of dark energy**





What LSST will Deliver:
A Data Stream, a Database, and a (small) Cloud

- 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 single-epoch detections ("sources"), and ~30 trillion forced sources, produced annually, accessible through online databases.
- Deep co-added images.

Level 2

- 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.

Level 3

Level 1:

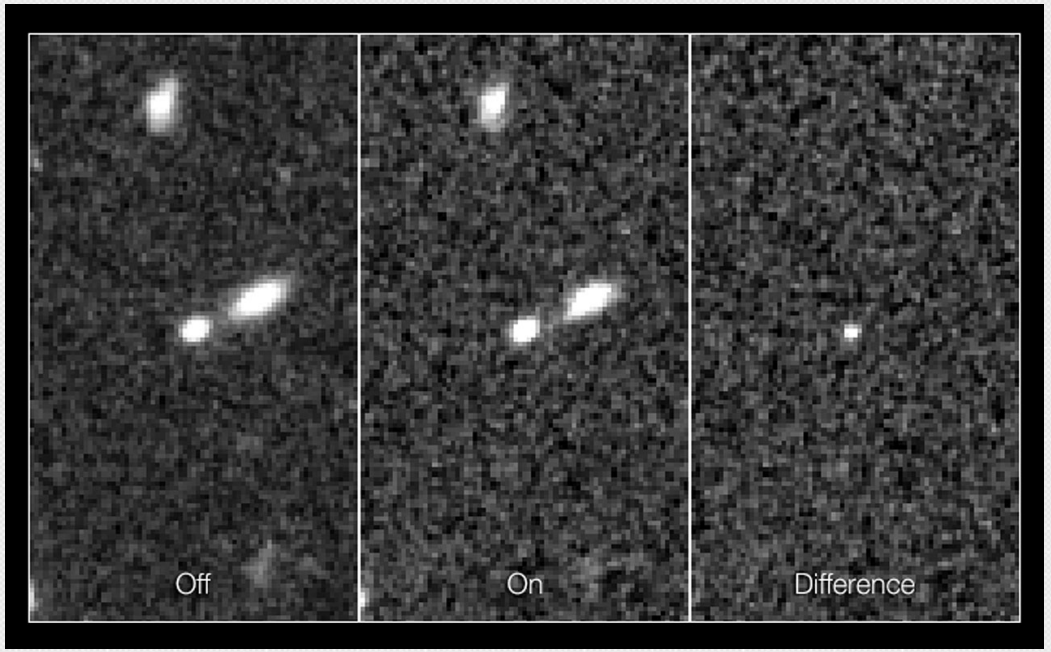
Enabling Discovery and Rapid Follow-up of Time Domain Events

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

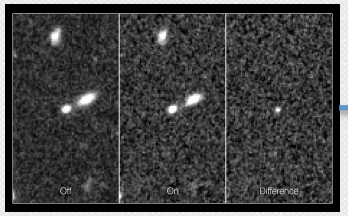
- For every source detected in a difference image, we will emit an "Event Alert" within 60 seconds of observation.

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.

Homework

Try to think of, and sketch out (in Python-like pseudo-code), your own filter to select a class of objects of interest to you.

Use the LSST Data Products Document (http://ls.st/dpdd) and the LSST Baseline Schema Browser (http://ls.st/tg2) to find out what features are available to select on.

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)

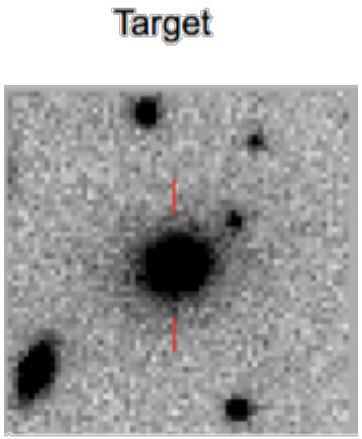
More in DPDD, Section 5.4

- 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

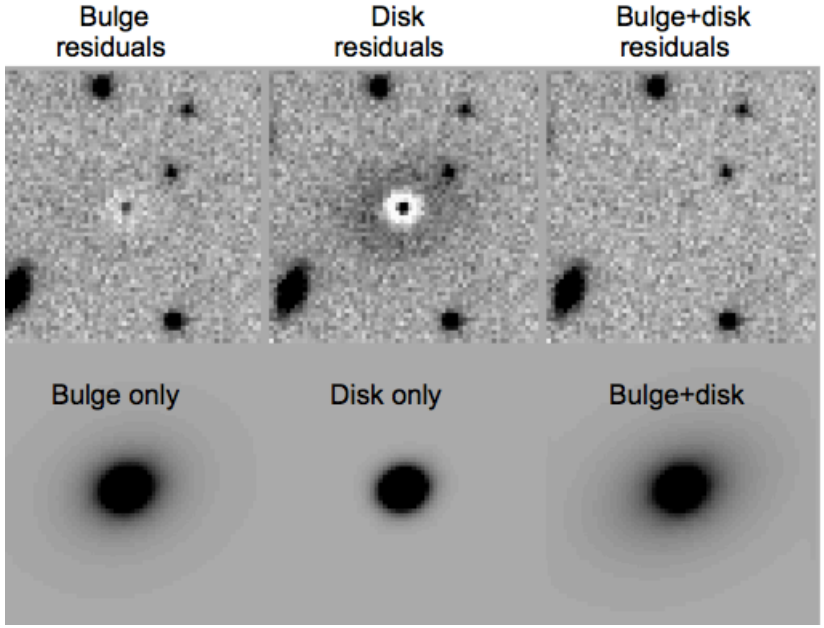
More in DPDD, Section 5.3

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

Think of something that will be impossible, or very difficult, to do with Level 1/2 data products

When poll is active, respond at **PollEv.com/lsst**

"Combine/compare data with previous non-LSST survey data"
1 day ago

"Joint modeling of extended and point like sources: "scene modeling" for high accuracy supernova and strong lens light curves"
1 day ago

"Sky OH variability"
1 day ago

"nebulae"
1 day ago

"Detailed properties of large galaxies"
1 day ago

"Systematic studies of pixel based algorithms."

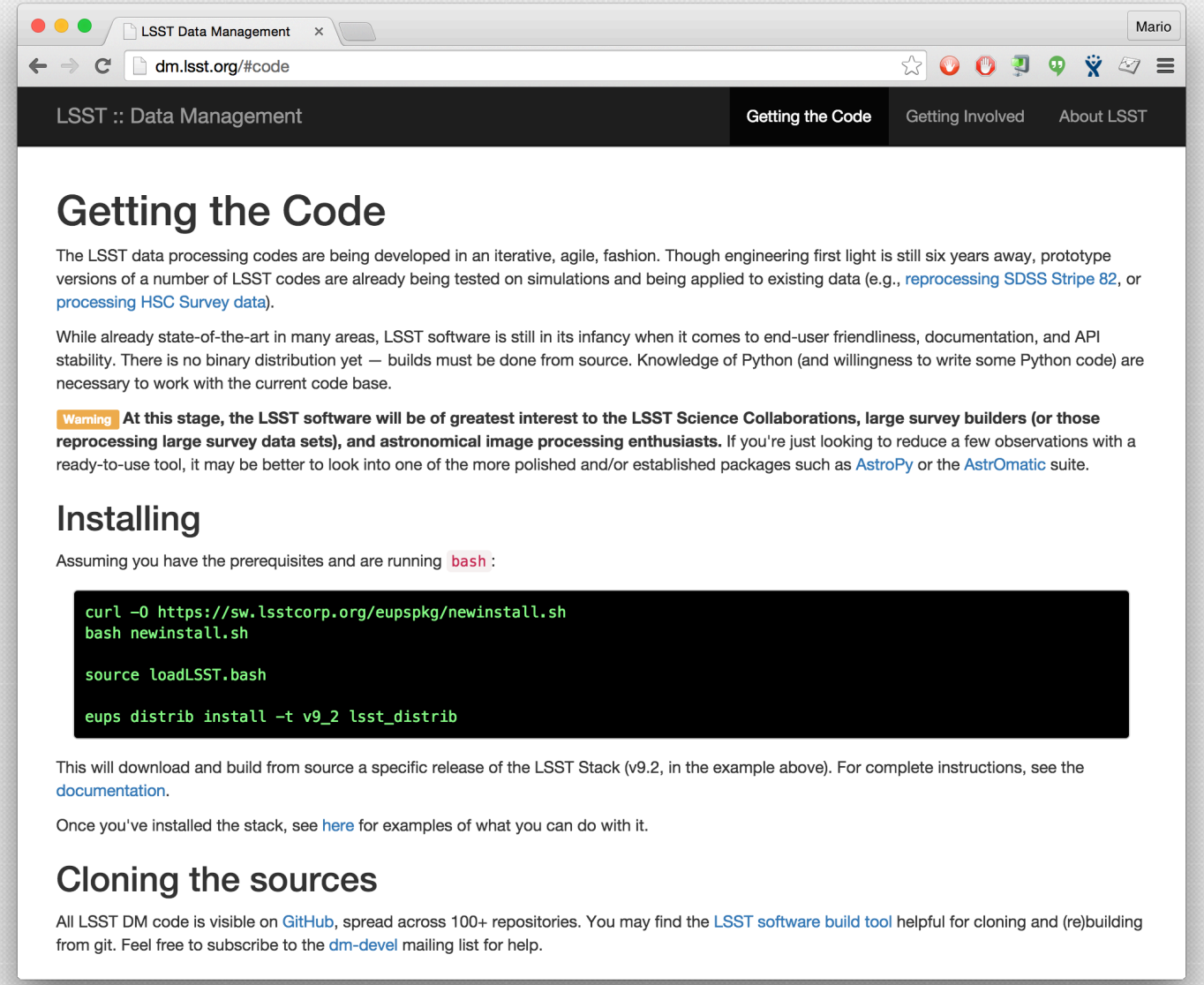
Level 3: Added Value Data Products

- **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

- **The LSST wants to make it easier for the community to create and distribute Level 3 products**
  - Enabling end-user analysis and processing at the LSST data center (JupyterHub cloud + a small HPC resource)
  - User databases and workspaces ("mydb")
  - Making the LSST software stack available to end-users

Software Prototypes are already available



http://dm.lsst.org

and

http://pipelines.lsst.io

The latest release includes conda-installable binaries!

WARNING:

Still under HEAVY development !!!!!

Summary

- The scale and precision of today's surveys is changing astronomy. Survey data is becoming abundant, well characterized, information rich, and complex.

- The primary deliverable of LSST is not the telescope, nor the instruments, nor the raw data; it is the fully reduced data products. **LSST is a facility that delivers data products and data access and analysis services.** Try to become familiar with the products and service LSST will deliver.

- What LSST will deliver:
  - **Level 1; a real-time data stream:** a stream of alerts to changes in the sky, issued within 60 seconds of each observation. Designed for rapid discovery and follow-up.
  - **Level 2; a database:** an annually processed catalog and a collection of fully reduced images, to enable systematics-limited science.
  - **Level 3; a small "cloud"**: a way to run queries, analysis, and additional computation in the LSST data center, next to the data.

- Along with the data products, the **LSST will also deliver the software used to create them, that you can modify and rerun yourself**. This may spur novel kinds of specialized data analyses.