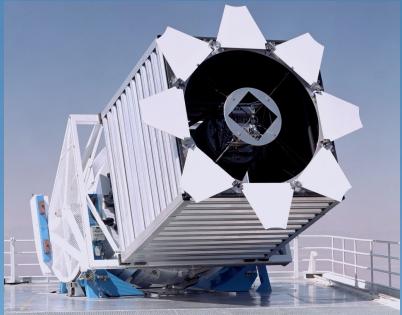


Big Data Sets in Astronomy

Željko Ivezić, University of Washington

LSST



SDSS



Gaia



Large Synoptic Survey Telescope (LSST)

SDSS:

a digital color map
of the night sky

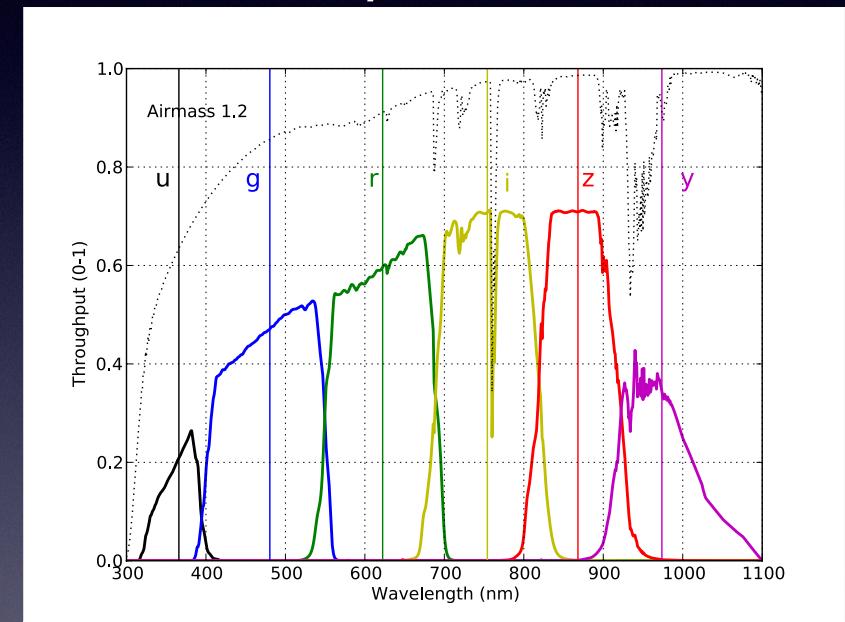
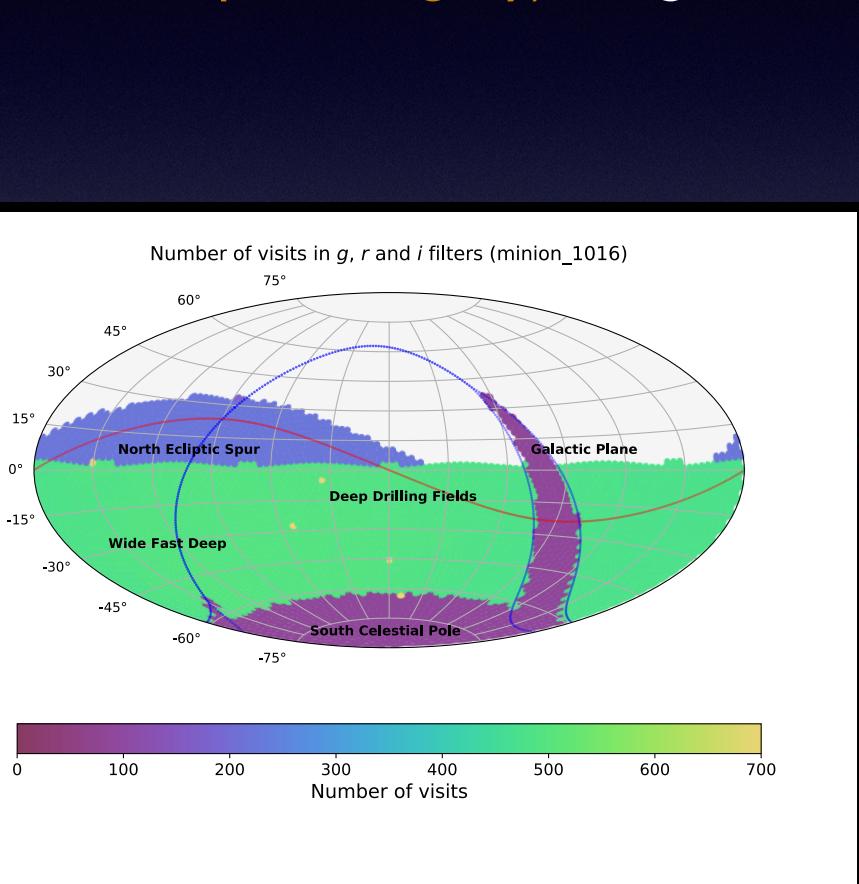
LSST:

a digital color
movie of the sky



Basic idea behind LSST: a uniform sky survey

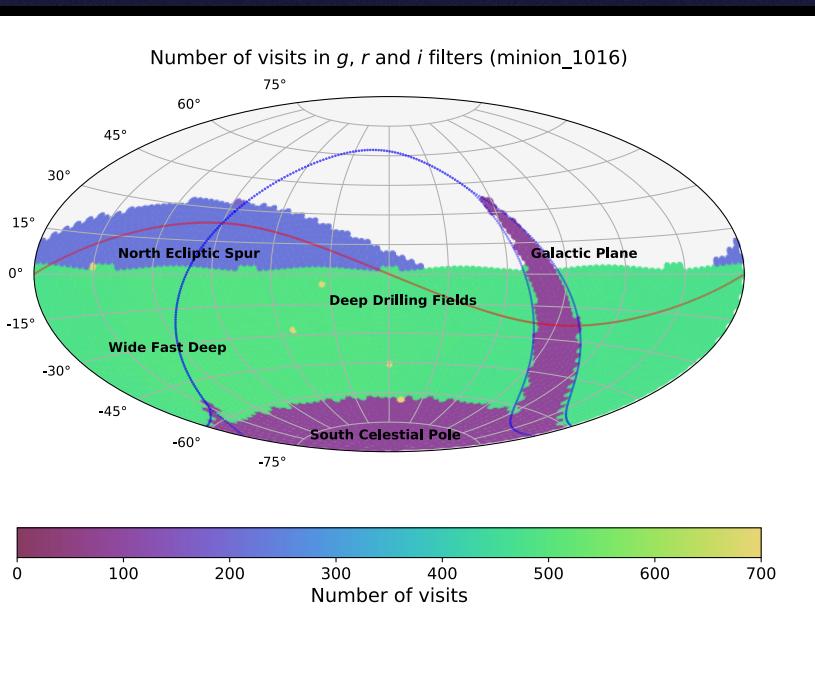
- 90% of time will be spent on a uniform survey: every 3-4 nights, the whole observable sky will be scanned twice per night
- after 10 years, half of the sky will be imaged about 1000 times (in 6 bandpasses, ugrizy): a digital color movie of the sky



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

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 $r \sim 27.5$ (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)

Outline

- LSST science drivers
 - cosmology (dark matter and dark energy)
 - time domain
 - the Milky Way structure
 - the Solar System structure
- Rapid tour of LSST and status report
 - multi-color time-resolved faint sky map
 - 20 billion stars and 20 billion galaxies
- Data analysis challenges ahead of us
 - large data sets
 - complex analysis
 - aiming for small systematics

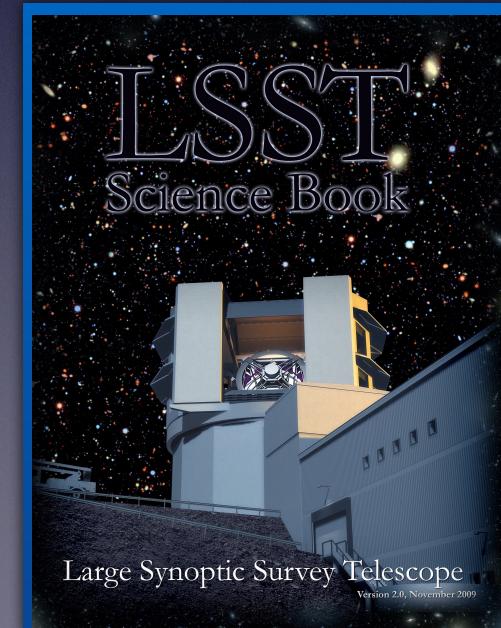
LSST Science Themes

- 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

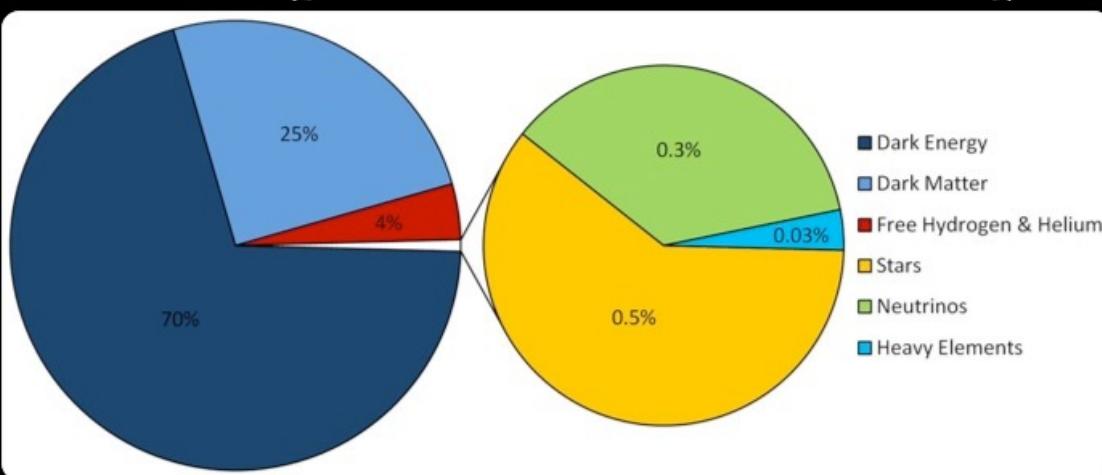
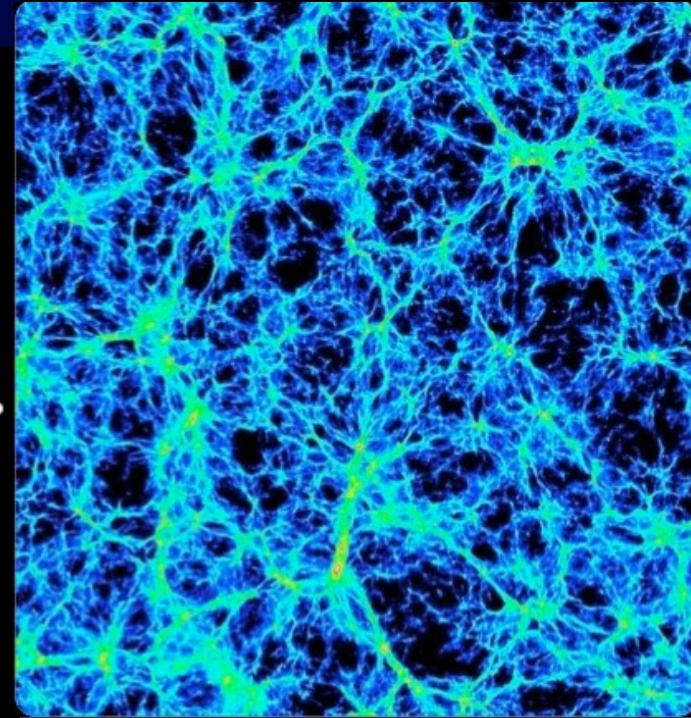
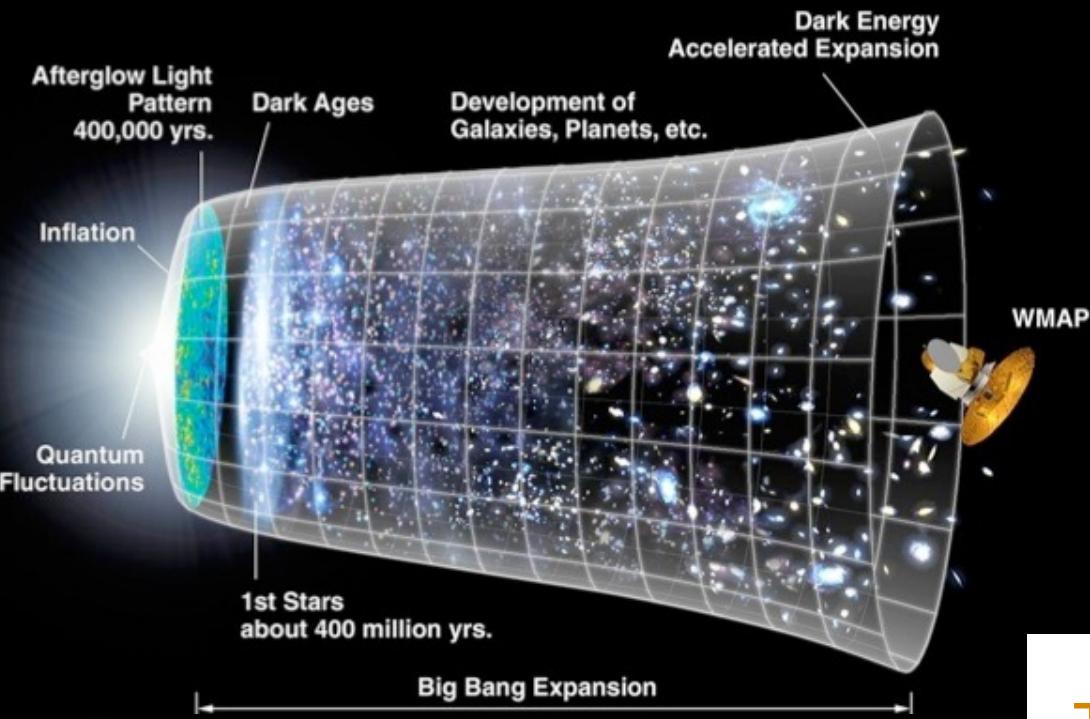
Summarizes LSST hardware, software, and observing plans, science enabled by LSST, and educational and outreach opportunities

245 authors, 15 chapters, 600 pages



New Cosmological Puzzles

Λ CDM: The 6-parameter Theory of the Universe



The modern cosmological models can explain all observations, but need to **postulate** dark matter and dark energy (though gravity model could be wrong, too)

Modern Cosmological Probes

- Cosmic Microwave Background
(the state of the Universe at the recombination epoch, at redshift ~ 1000)
- Weak Lensing: growth of structure
- Galaxy Clustering: growth of structure
- Baryon Acoustic Oscillations: standard ruler
- Supernovae: standard candle

Except for CMB, measuring $H(z)$ and growth of structure $G(z)$
 $H(z) \sim d[\ln(a)]/dt$, $G(z) = a^{-1} \delta \rho_m / \rho_m$, with $a(z) = (1+z)^{-1}$

Killer asteroids: the impact probability is not 0!

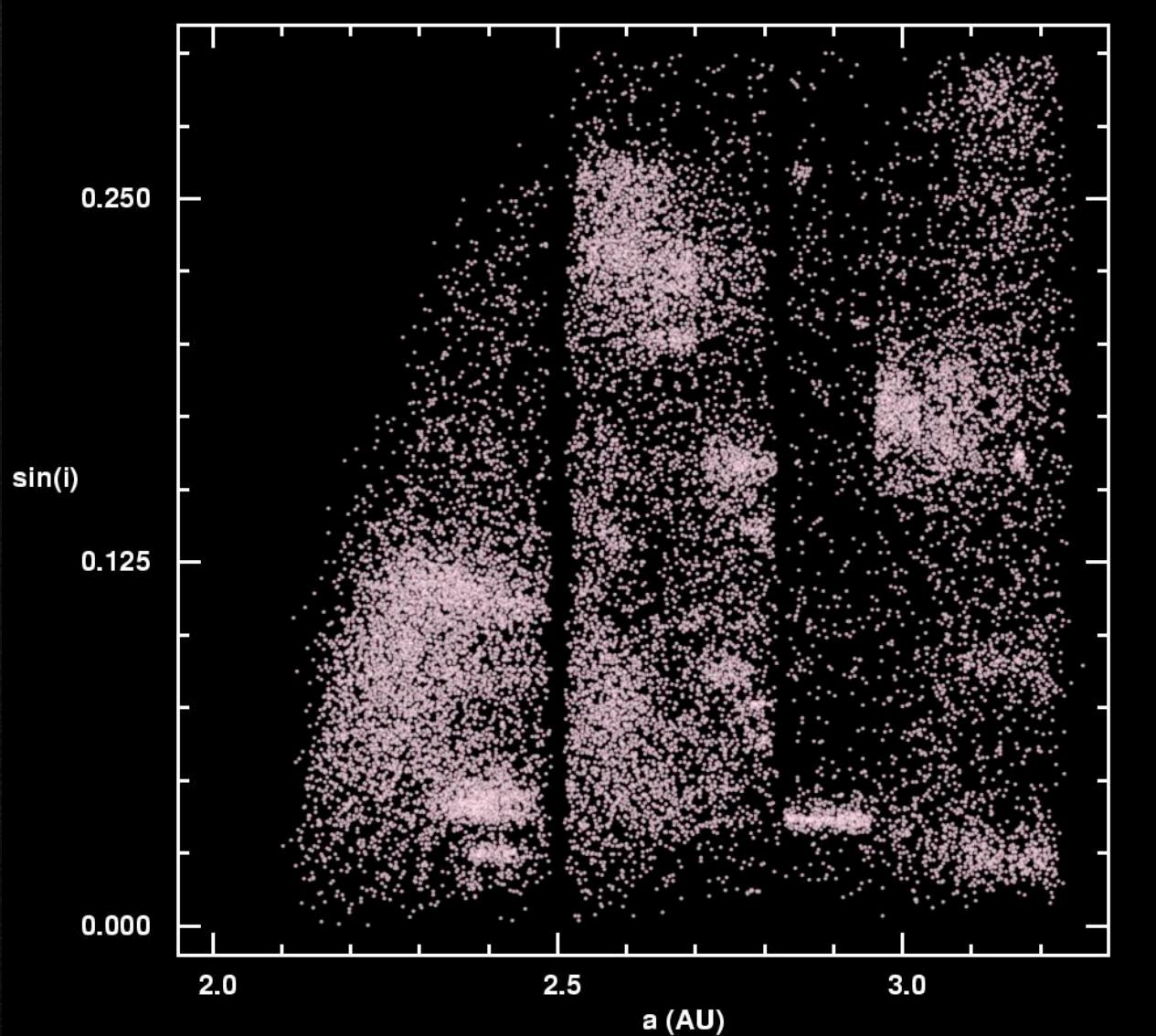


The Barringer Crater,
Arizona: a 40m
object 50,000 yr. ago



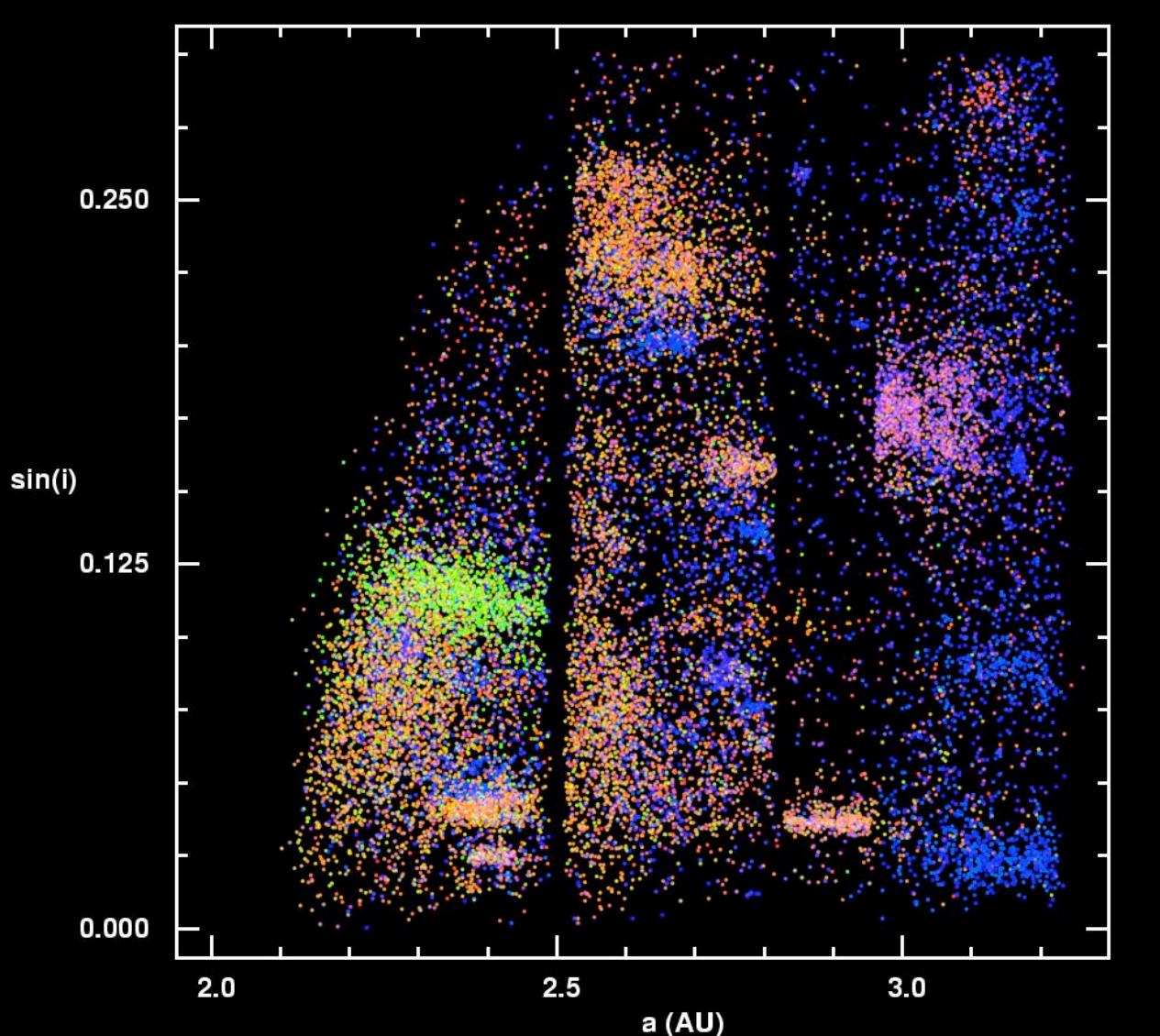
LSST is the only survey capable of delivering completeness specified in the 2005 USA Congressional NEO mandate to NASA (to find 90% NEOs larger than 140m)

Main-belt Inventory



30,000
Asteroids with
SDSS colors and
proper
orbital elements
(Ivezic, Juric, Lupton 2002)

Main-belt Inventory

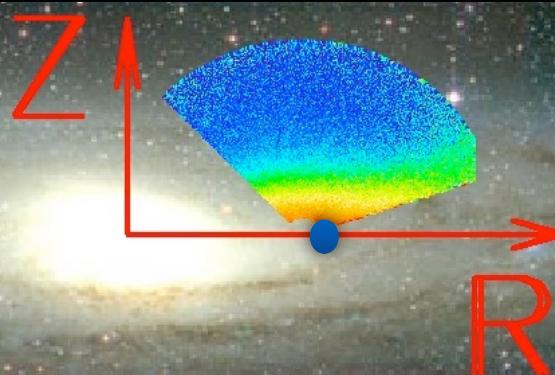


30,000
Asteroids with
SDSS colors and
proper
orbital elements
(Ivezic, Juric, Lupton 2002)

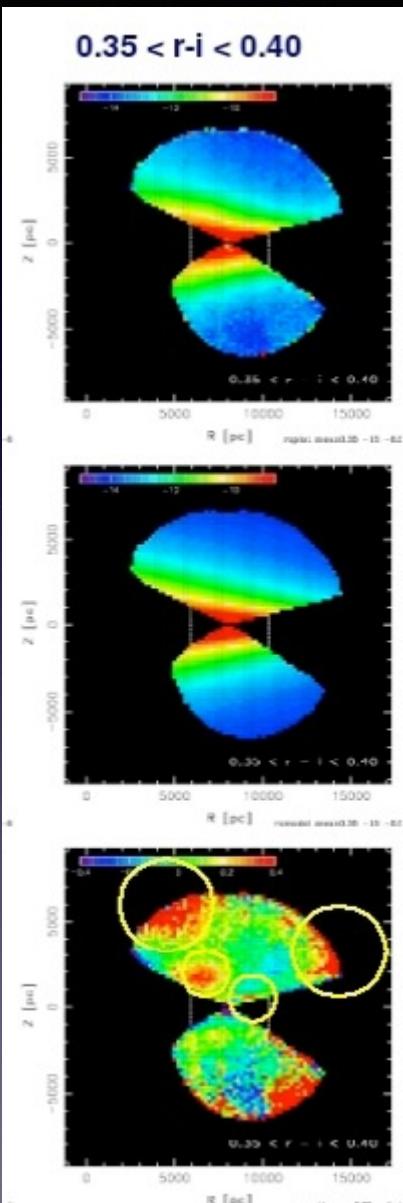
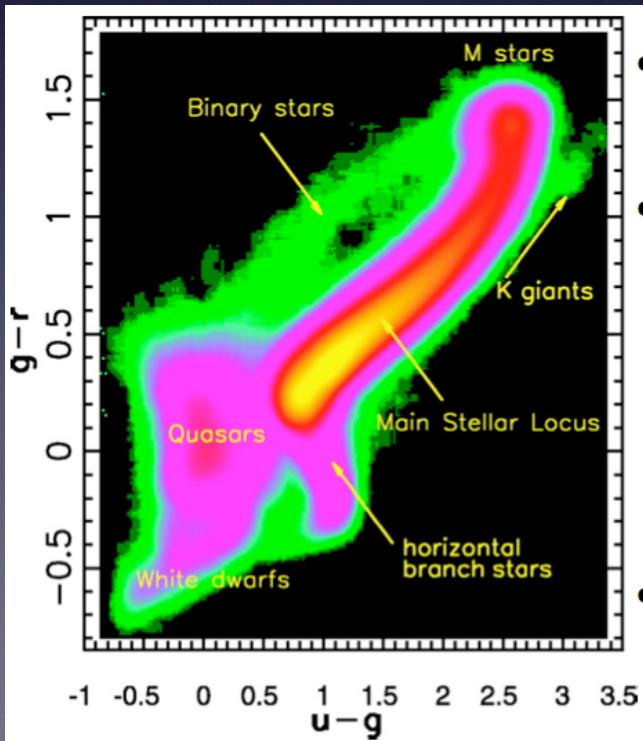
Color-coded with
SDSS colors

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

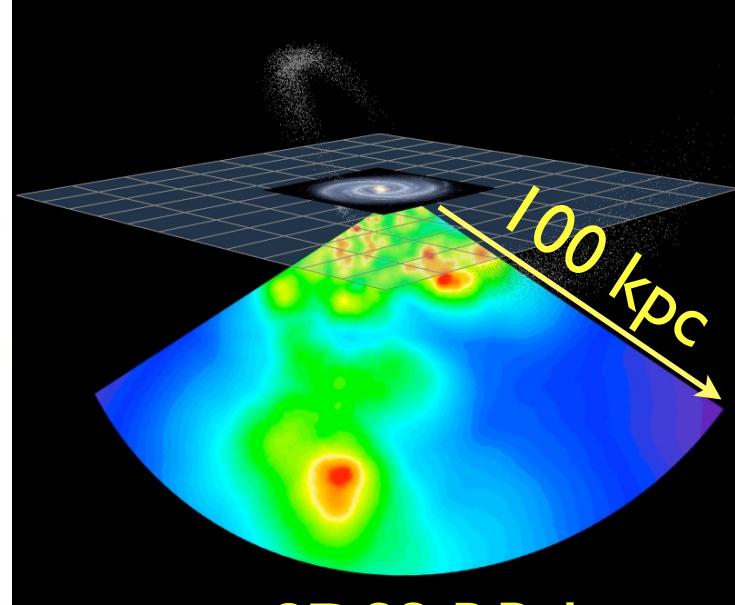
The Milky Way structure: 20 billion stars, time domain massive statistical studies!



Main sequence stars
Distance and [Fe/H]:



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

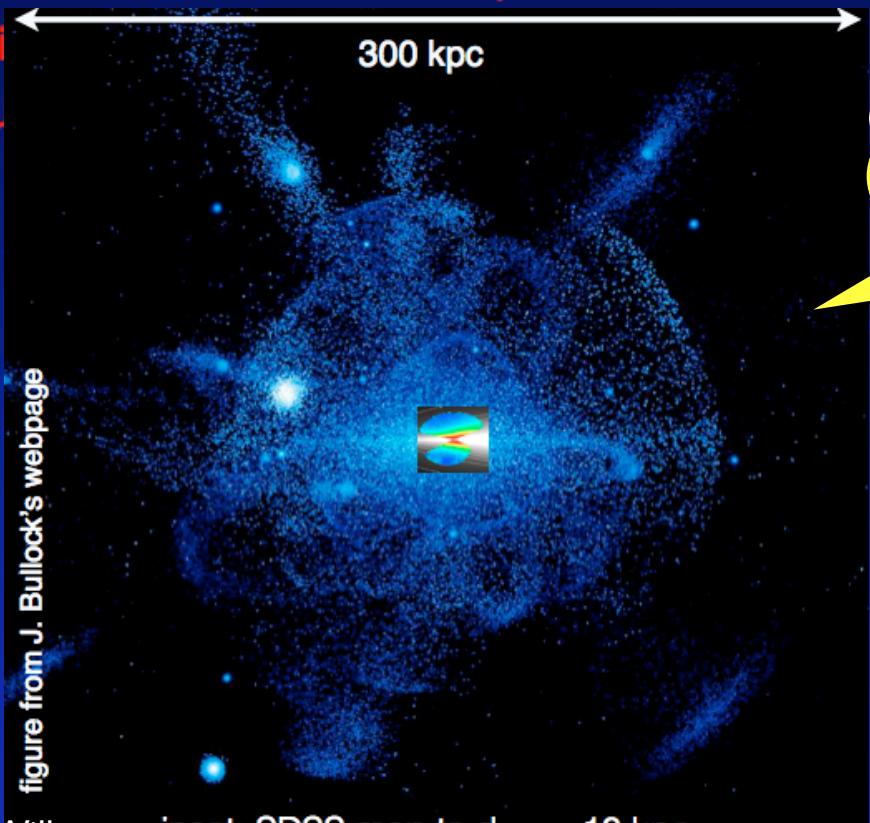


Sesar et al. (2009)

SDSS RR Lyrae

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

The large red circle: the ~100 kpc limit of future LSST studies
(and the current limit of SDSS)



montage from B. Willman

inset: SDSS map to $d_{\text{limit}} = 10 \text{ kpc}$

LSST limit for RR Lyrae: 400 kpc

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

SDSS

gri

3.5'x3.5'

r~22.5



HSC

gri

3.5'x3.5'

r~27

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

like LSST
depth (but
tiny area)

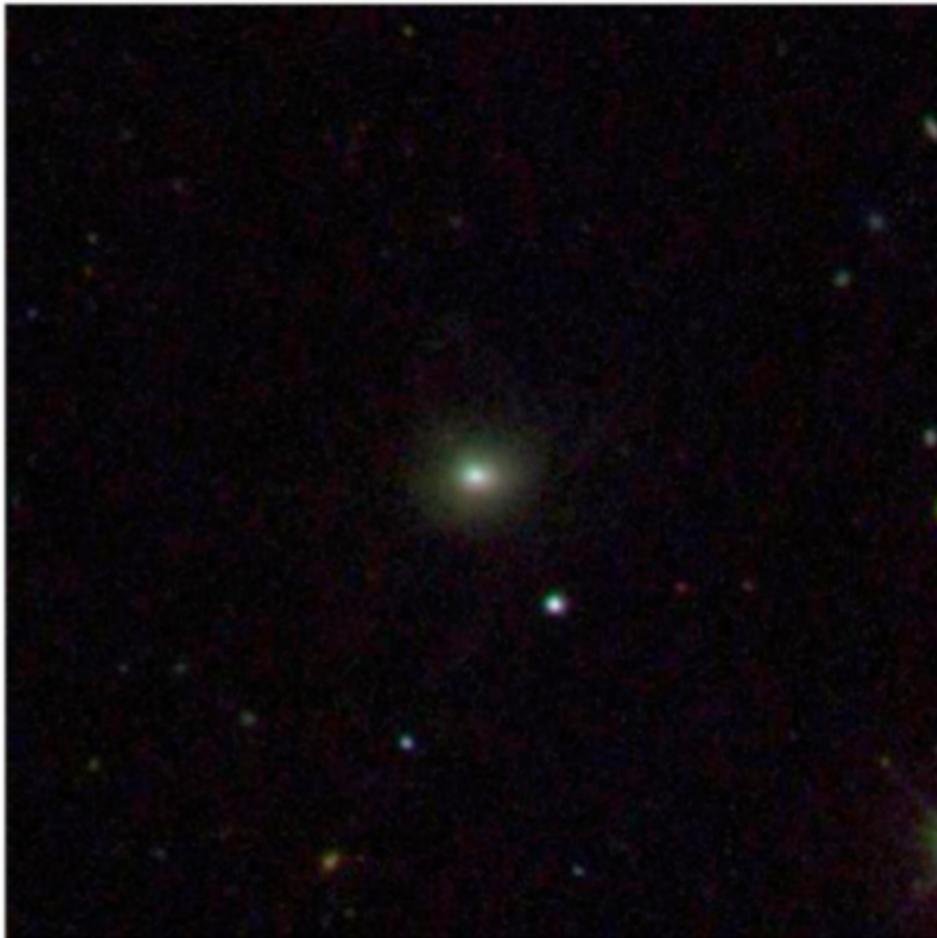
LSST will
deliver 5
million such
images



Extragalactic astronomy: faint surface brightness limit

SDSS

3x3 arcmin, gri



MUSYC r ~26

(almost) like LSST
depth (but tiny area)



Gawiser et al

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



Gemini South
Telescope

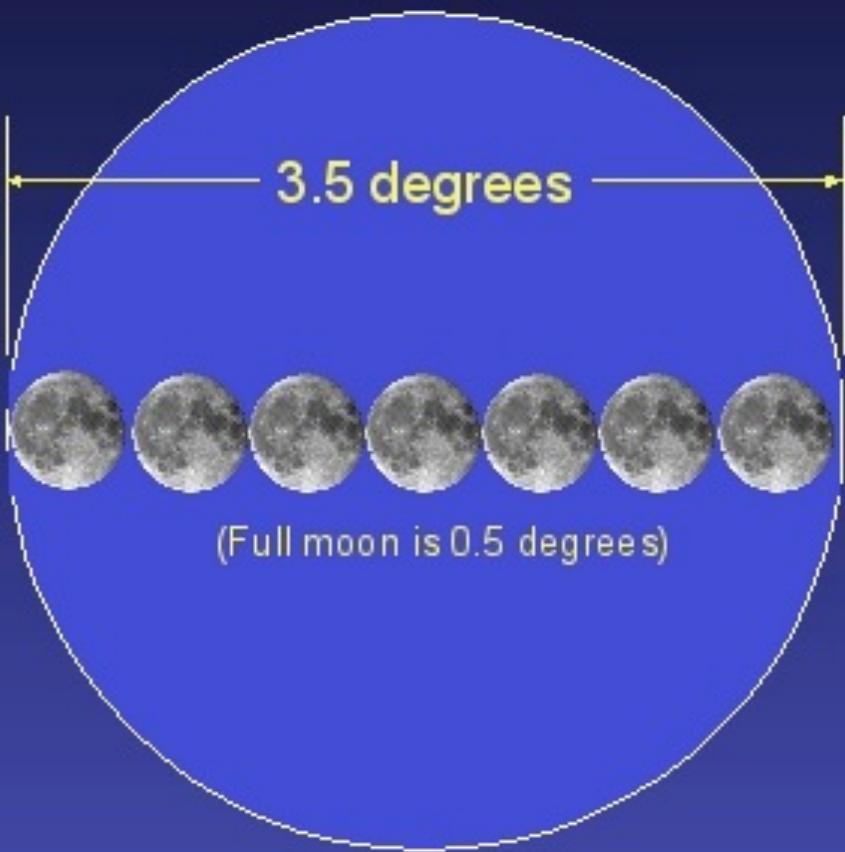


LSST

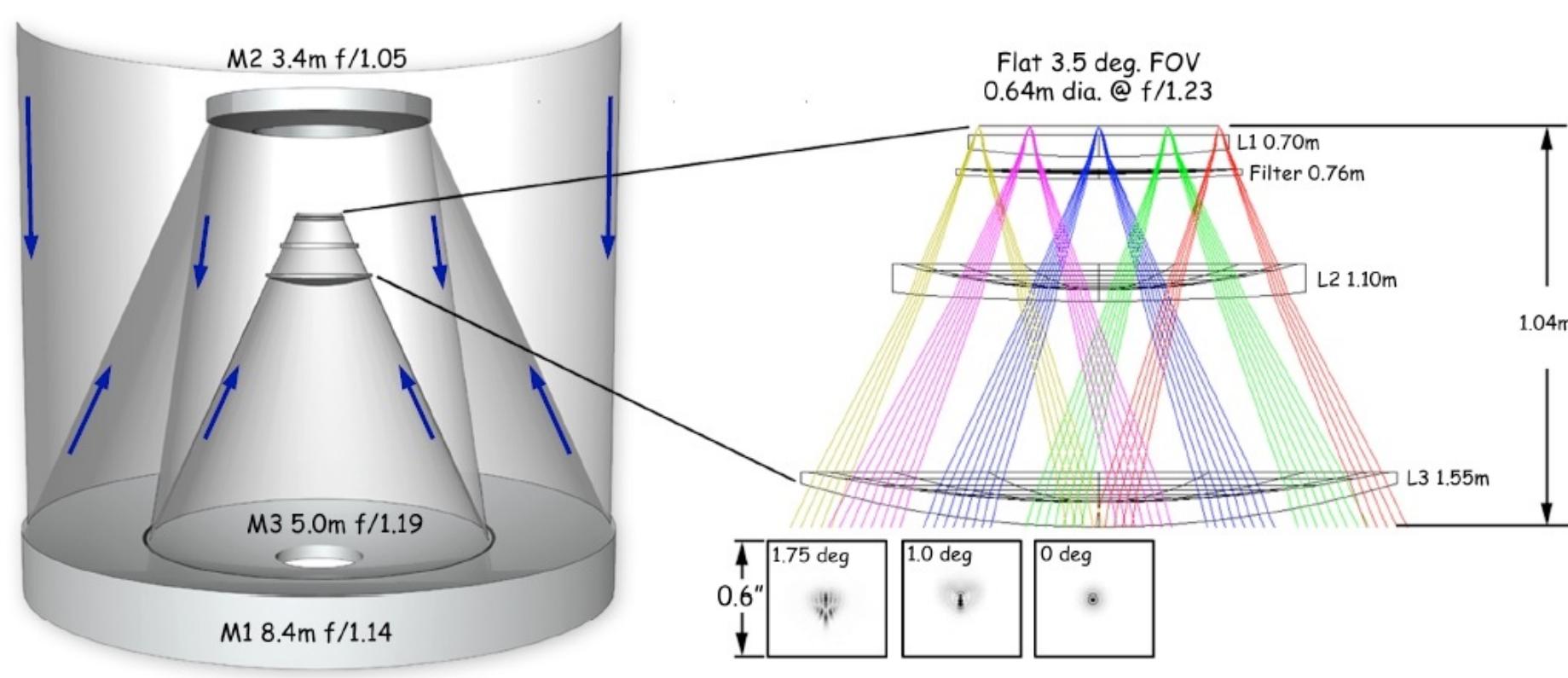
Primary Mirror
Diameter



Field of
View



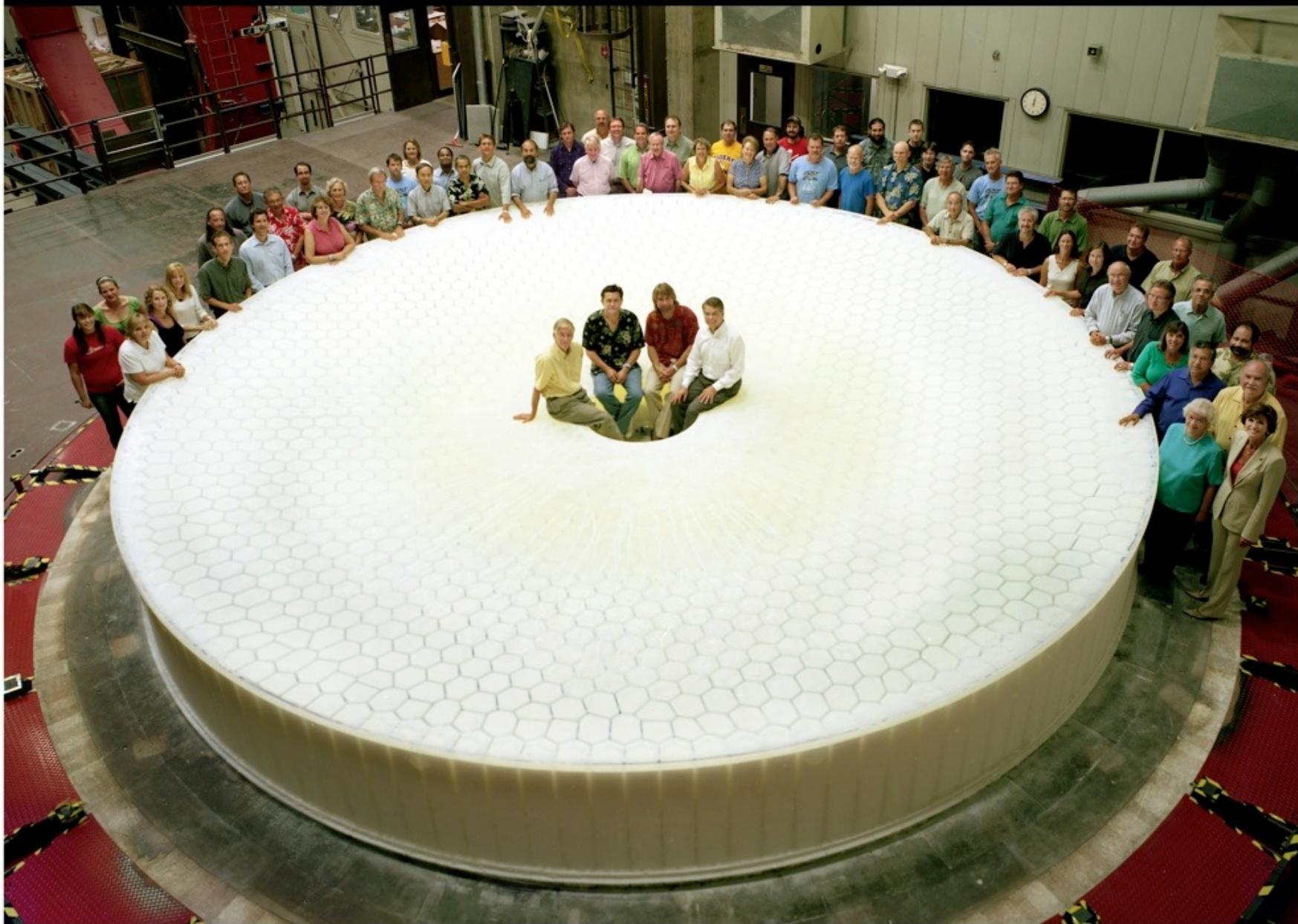
Optical Design for LSST

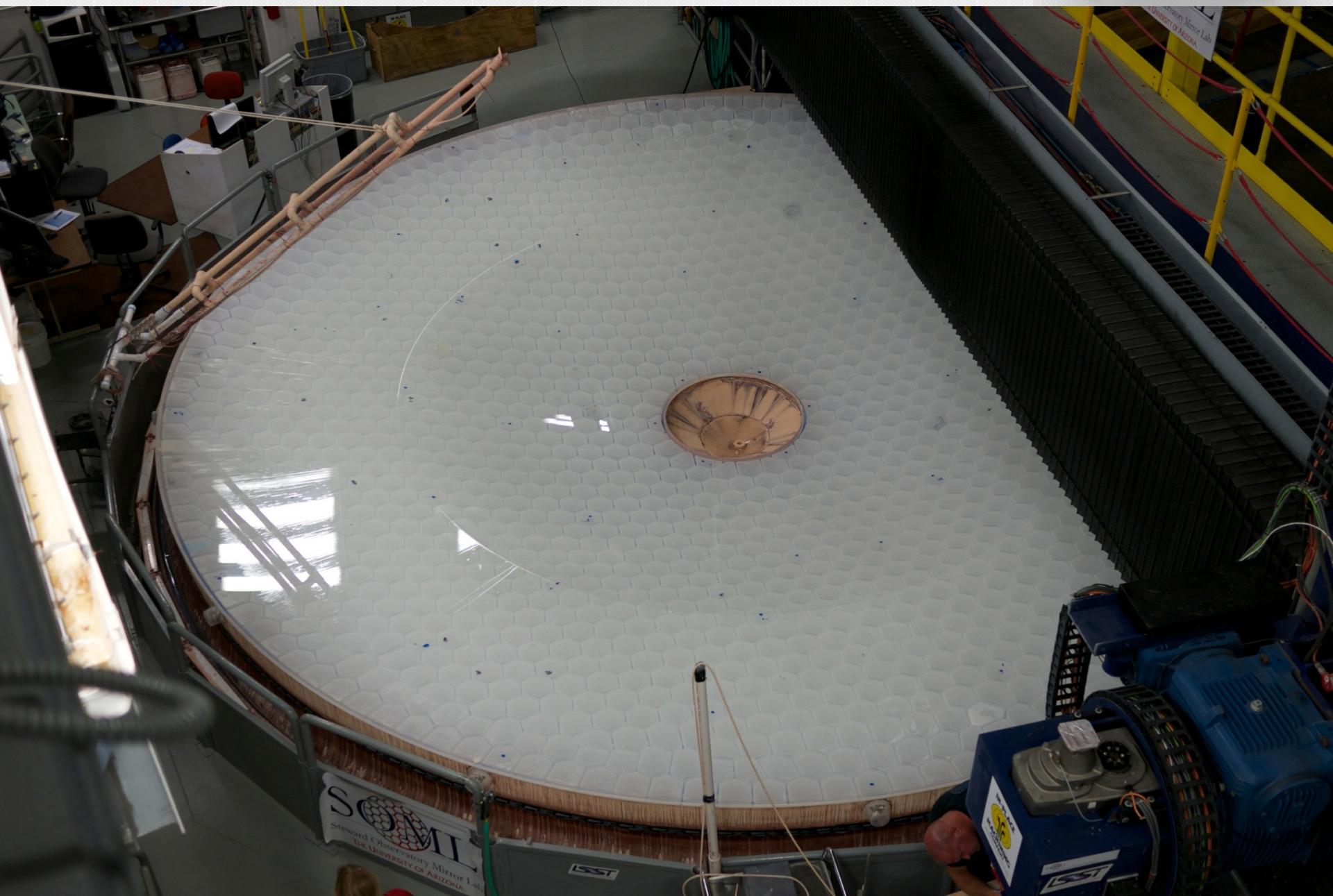


Three-mirror design (Paul-Baker system)
enables large field of view with excellent image quality:
delivered image quality is dominated by atmospheric seeing



Large Synoptic Survey Telescope

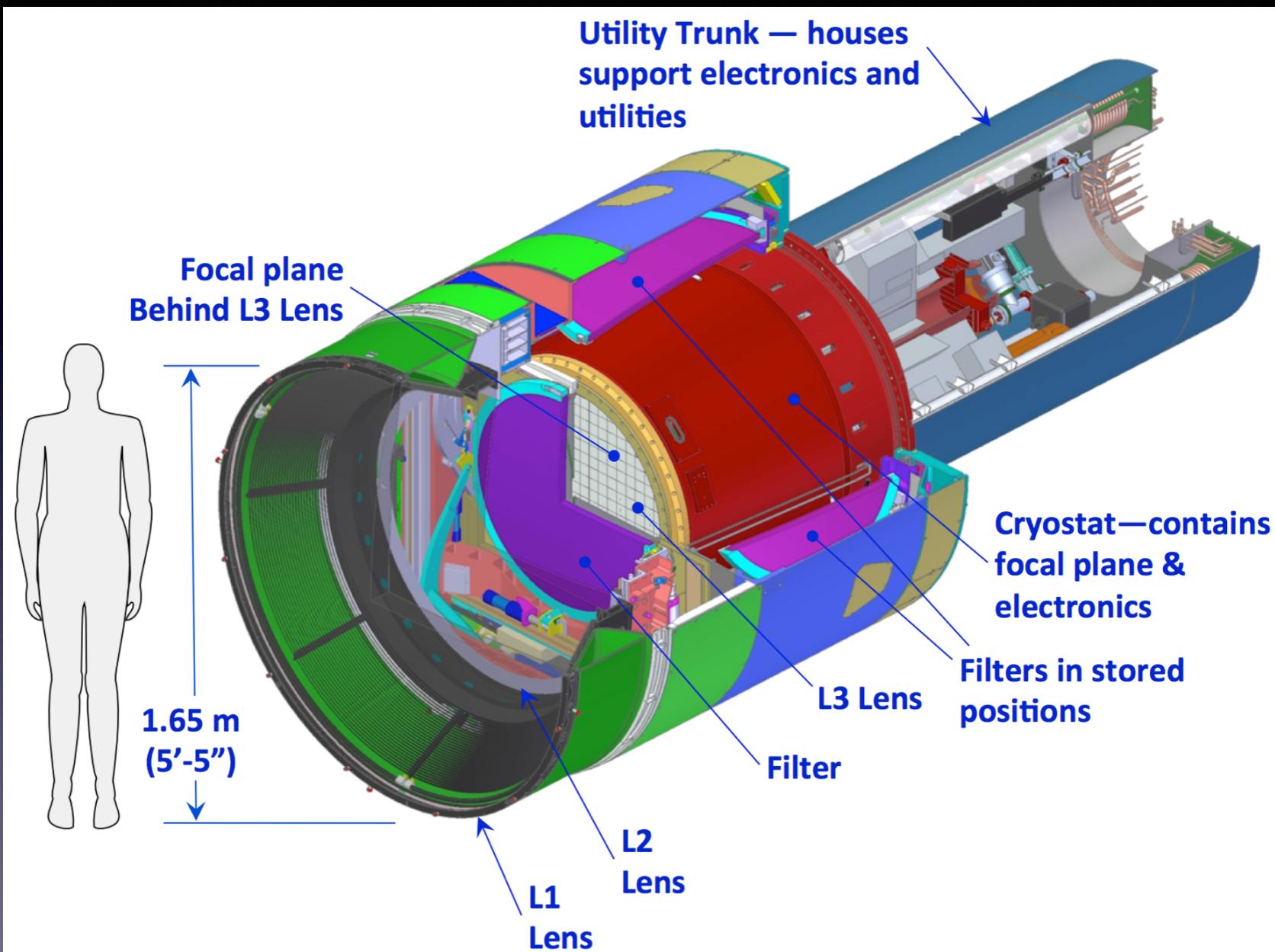






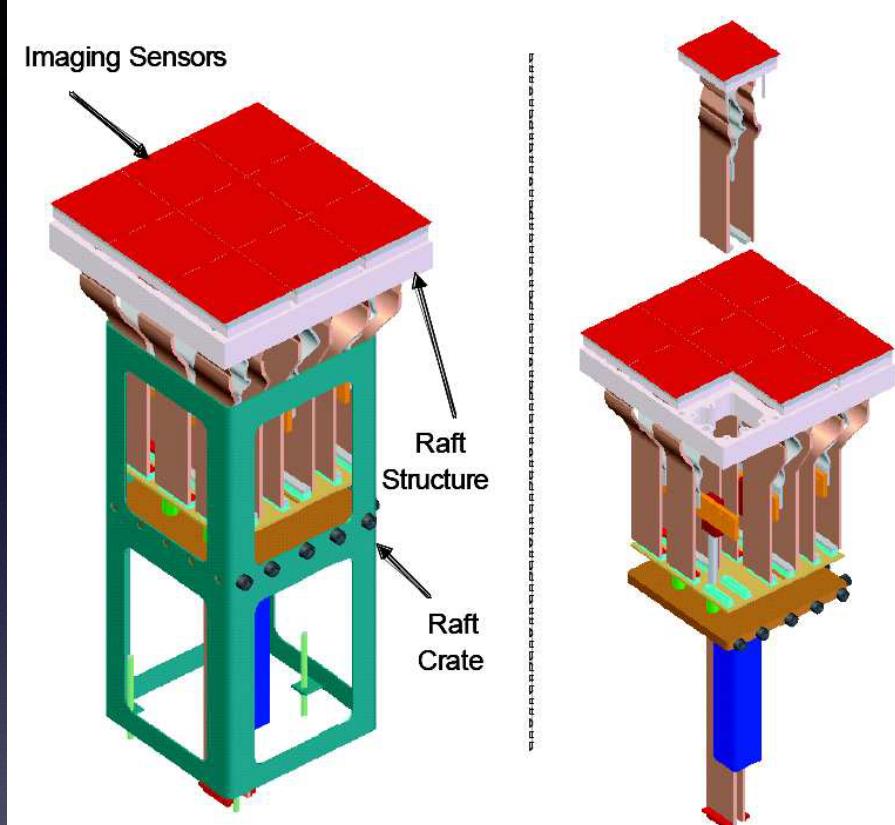
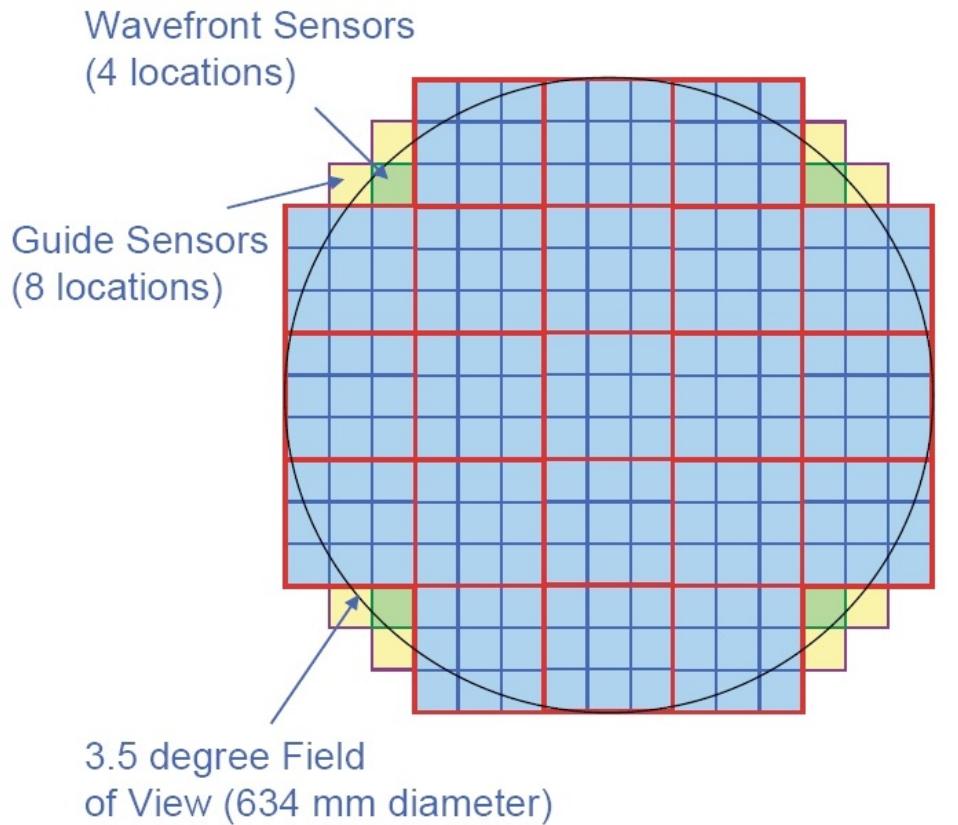
Telescope Mount Assembly before going from Spain to Chile

LSST camera



The largest astronomical camera: 2800 kg, 3200 Megapix

LSST camera



Modular design: 3200 Megapix = 189 x 16 Megapix CCD
9 CCDs share electronics: raft (=camera)

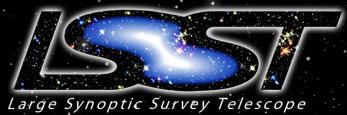
Problematic rafts can be replaced relatively easily

Mar 10, 2019



First light: 2021





LSST Operations: Sites & Data Flows

HQ Site

Science Operations
Observatory Management
Education & Public Outreach

Base Site

Base Center
Long-term storage (copy 1)
Data Access Center
Data Access & User Services



French Site

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

Archive Site

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

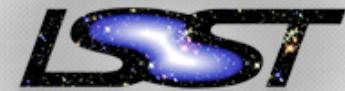
Data Access Center

Data Access and User Services

Summit Site

Telescope & Camera
Data Acquisition
Crosstalk Correction

At the highest level, LSST objectives are:



- 1) Obtain about 5.5 million images, with 189 CCDs (4k x 4k) in the focal plane; this is about **a billion 16 Megapixel images of the sky**
- 2) Calibrate these images (and provide other metadata)
- 3) Produce catalogs (“model parameters”) of detected objects (37 billion)
- 4) Serve images, catalogs and all other metadata, that is,
serve LSST data products to LSST users

The ultimate deliverable of LSST is not just the telescope, nor the camera, but the fully reduced science-ready data as well. Software!

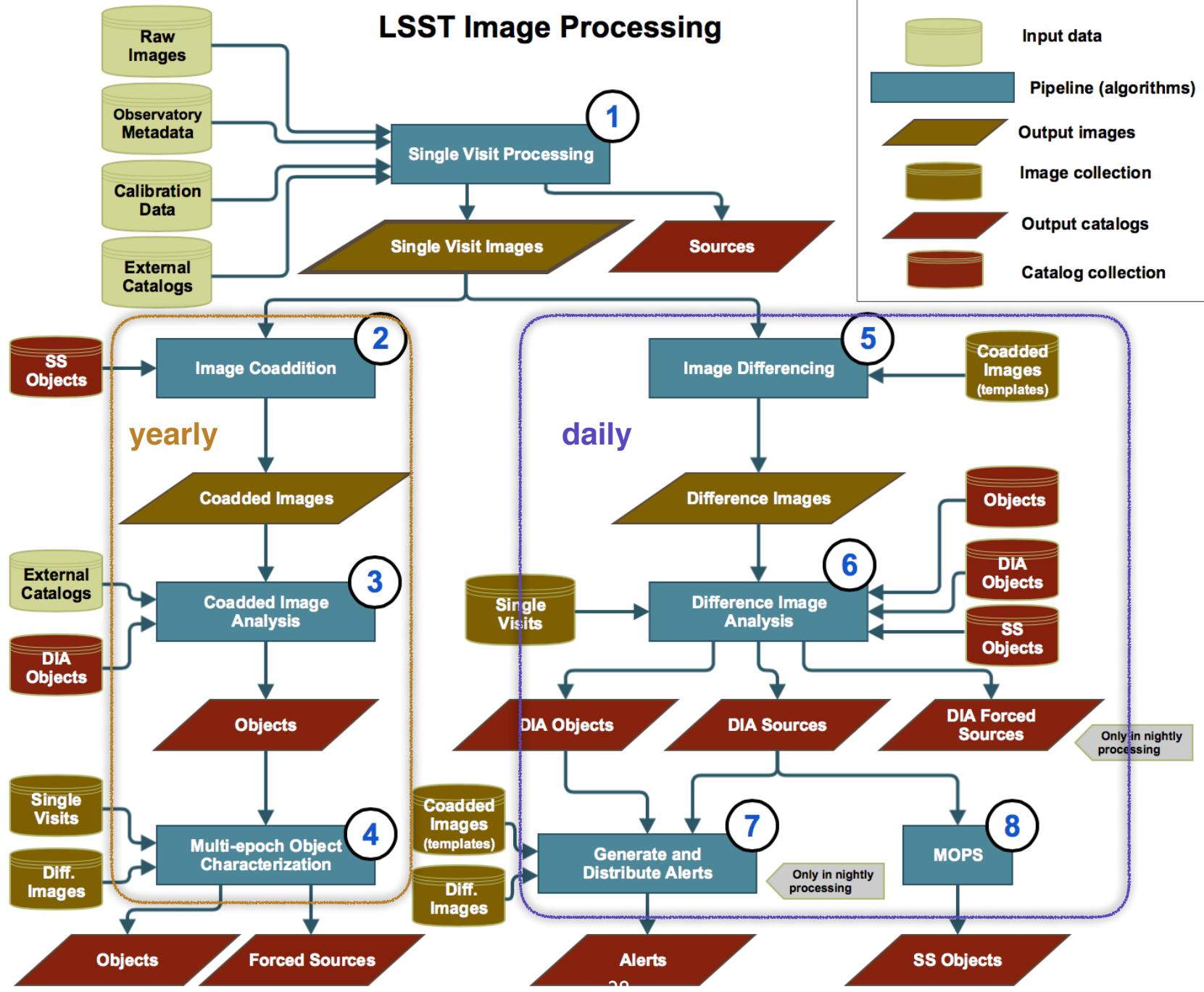
LSST Data Management System (“software”)



- 20 TB of data to process every day (~one SDSS/day)
- 1000 measurements for 40 billion objects during 10 years
- Existing tools and methods (e.g. SDSS) do not scale up to LSST data volume and rate (100 PB!)
- About 5-10 million lines of code (C++/python)



LSST Image Processing



Astronomical Image Formation (ground based optical telescopes)

Optics

+Tracking

+Diffraction

+Detector
Misalignments &
Perturbations



+Lens Misalignments

+Mirror Misalignments
Perturbations,
& Micro-roughness

+Detector

+High Altitude
Atmosphere

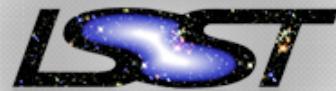
+Mid Altitude
Atmosphere

+Low Altitude
Atmosphere

+Pixelization

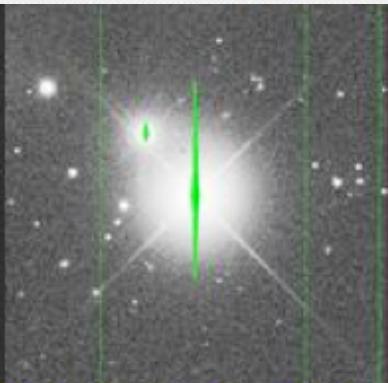
+Saturation &
Blooming

Basic steps in astronomical image processing



A raw data frame.

The difference in bias levels from the two amplifiers is visible.



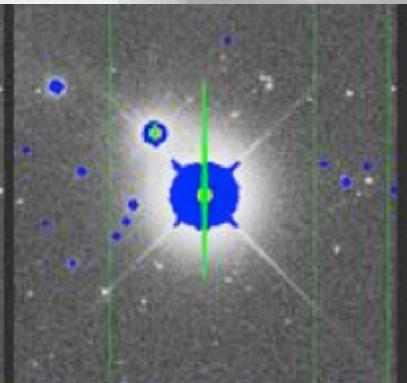
Bias-corrected frame

with saturated pixels, bad columns, and cosmic rays masked in green.



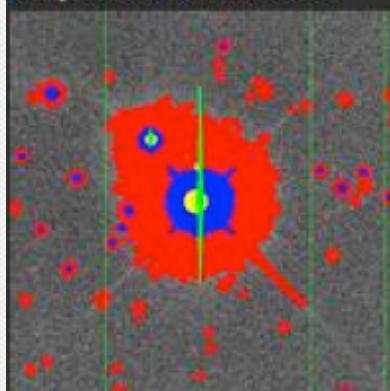
Frame corrected for

saturated pixels, bad columns, and cosmic rays.



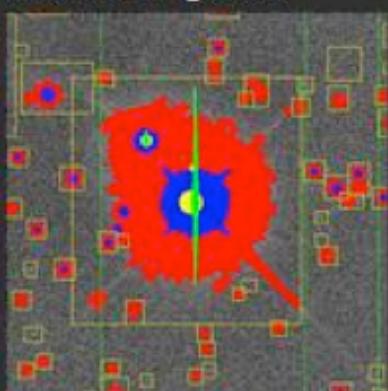
Bright object

detections marked in blue.



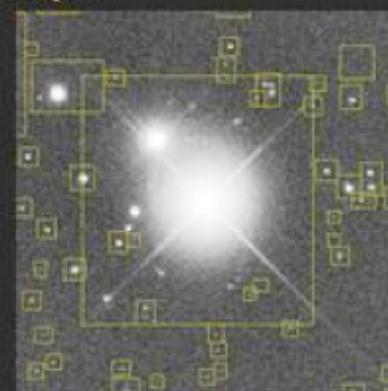
Faint object

detections marked in red.



Measured objects,

masked and enclosed in boxes. Small empty boxes are objects detected only in some other band.



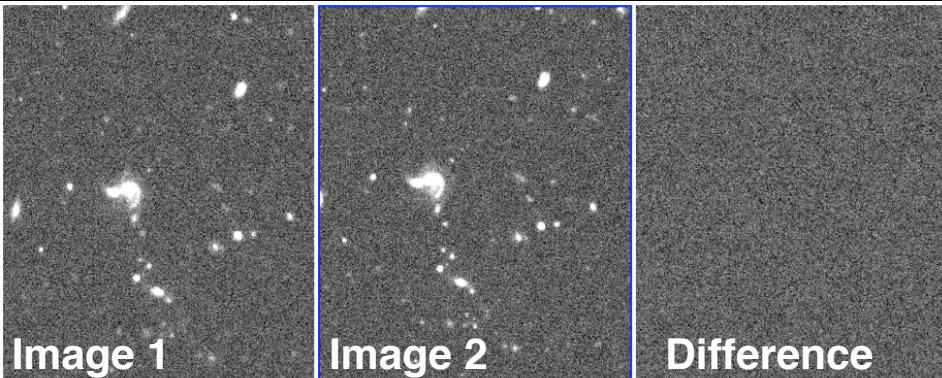
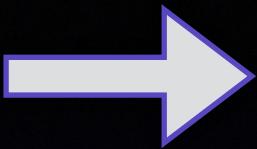
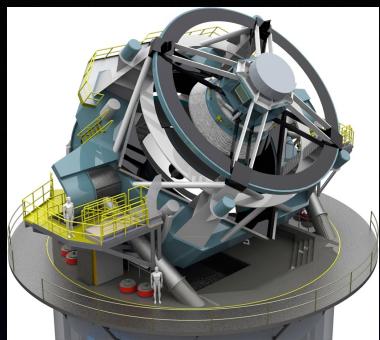
Measured objects in

the data frame.



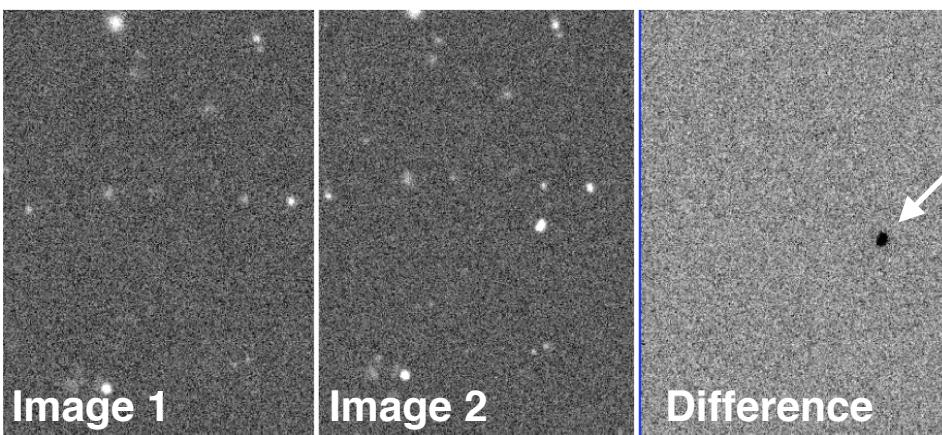
Reconstructed

image using postage stamps of individual objects and sky background from binned image.



Additional “followup” data obtained to:

- confirmation and classification
- provide better temporal resolution
- use different filters/wavelengths
- obtain spectra (distance!)
- other measurements (e.g. polarimetry)



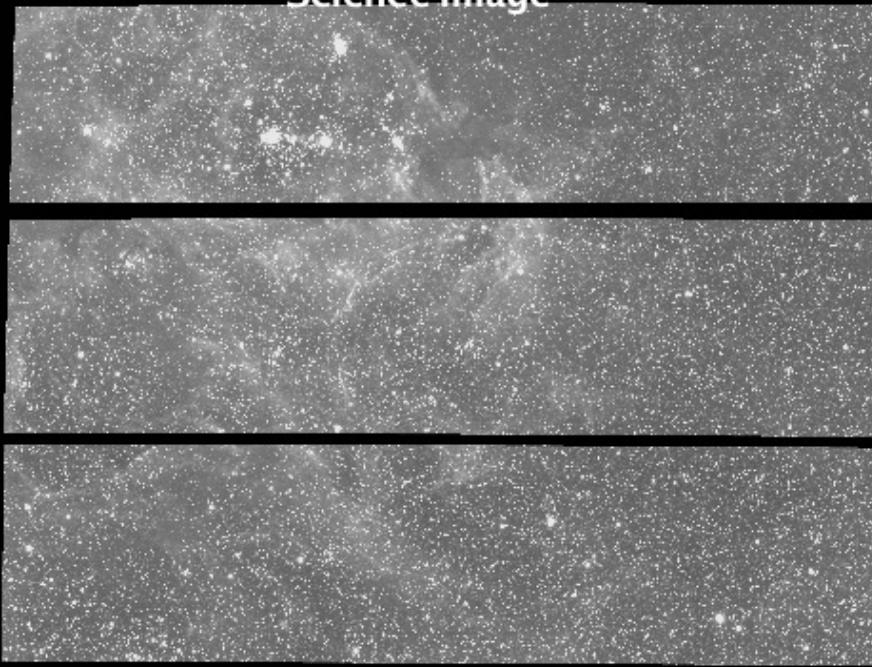
Alerts can trigger “Followup” observations:



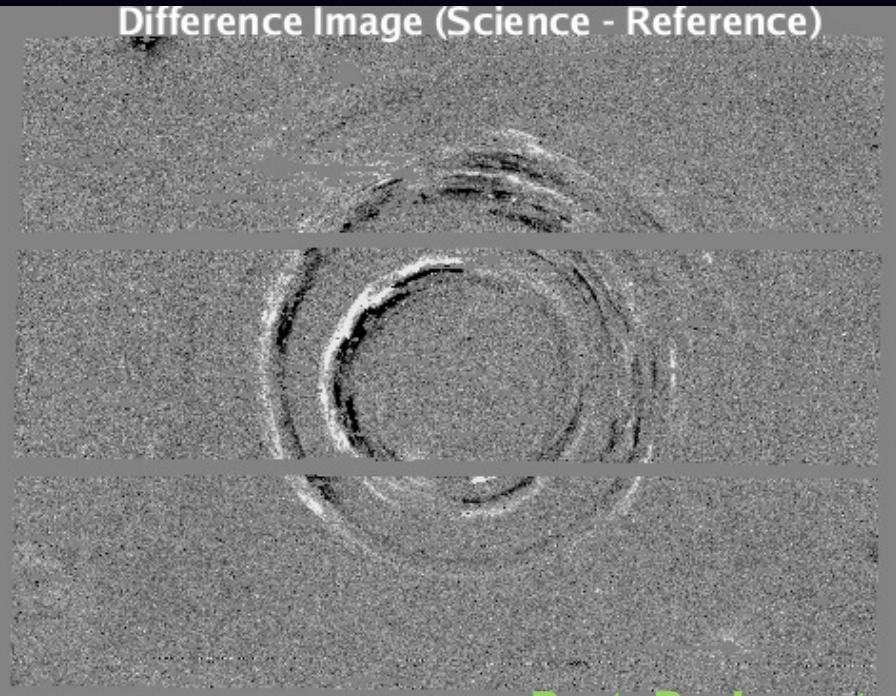
Time Domain: objects changing in time
positions: asteroids and stellar proper motions
brightness: cosmic explosions and variable stars

Not only point sources - echo of a supernova explosion:

Science Image



Difference Image (Science - Reference)



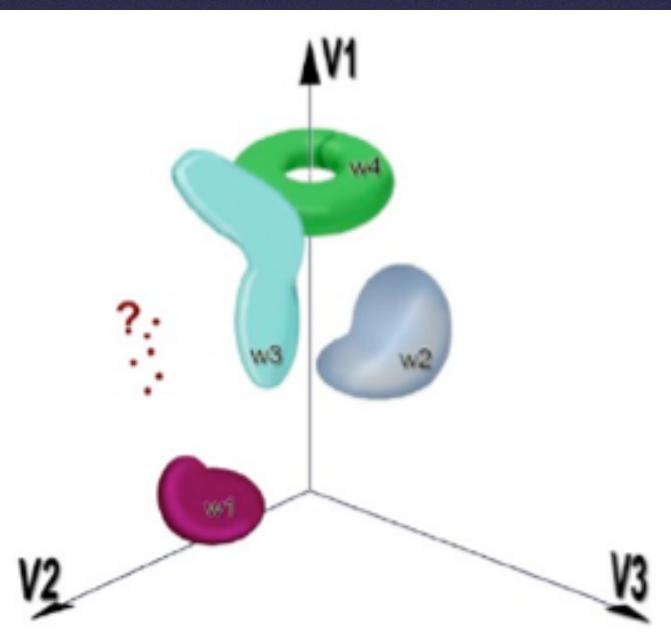
Rest, Becker, et al.

As many variable stars from LSST, as all stars from SDSS
Web stream with data for transients within 60 seconds.
Real time alerts!

Statistical analysis of a massive LSST dataset

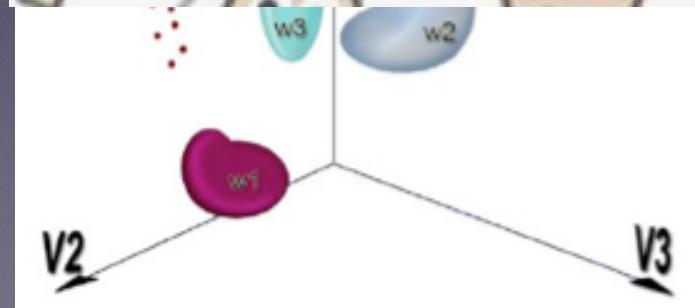
- A large (100 PB) database and sophisticated analysis tools: for each of 40 billion objects there will be about 1000 measurements (each with a few dozen measured parameters)

Data mining and knowledge discovery



- 10,000-D space with 40 billion points
 - Characterization of known objects
 - Classification of new populations
 - Discoveries of unusual objects
- Clustering, classification, outliers**

Statistical analysis of a massive LSST dataset



- Classification of new populations
 - Discoveries of unusual objects
- Clustering, classification, outliers**

1) Introduction

- astroML

News

October 2012: astroML 0.1 has been released! Get the source on [Github](#)

Our Introduction to astroML paper received the CIDU 2012 best paper award.

Links

[astroML Mailing List](#)

[GitHub Issue Tracker](#)

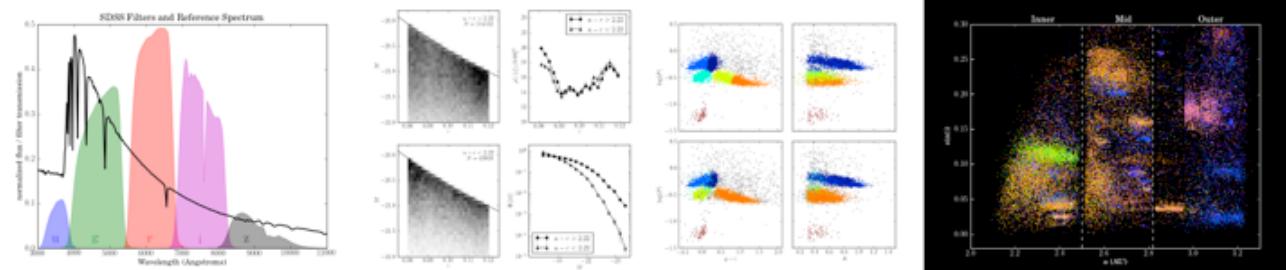
Videos

[Scipy 2012 \(15 minute talk\)](#)

Citing

If you use the software, please consider citing astroML.

AstroML: Machine Learning and Data Mining for Astronomy

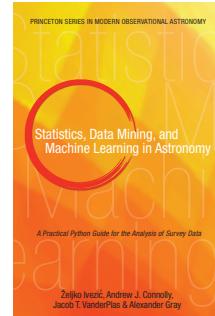


AstroML is a Python module for machine learning and data mining built on `numpy`, `scipy`, `scikit-learn`, and `matplotlib`, and distributed under the 3-clause BSD license. It contains a growing library of statistical and machine learning routines for analyzing astronomical data in python, loaders for several open astronomical datasets, and a large suite of examples of analyzing and visualizing astronomical datasets.

The goal of astroML is to provide a community repository for fast Python implementations of common tools and routines used for statistical data analysis in astronomy and astrophysics, to provide a uniform and easy-to-use interface to freely available astronomical datasets. We hope this package will be useful to researchers and students of astronomy. The astroML project was started in 2012 to accompany the book **Statistics, Data Mining, and Machine Learning in Astronomy** by Zeljko Ivezic, Andrew Connolly, Jacob VanderPlas, and Alex Gray, to be published in late 2013. The table of contents is available here: [here \(pdf\)](#).

Downloads

- Released Versions: [Python Package Index](#)
- Bleeding-edge Source: [github](#)



User Guide

1. Introduction

- 1.1. Philosophy

Open source!
www.astroml.org

Textbook Figures

This section makes available the source code used to generate every figure in the book *Statistics, Data Mining, and Machine Learning in Astronomy*. Many of the figures are fairly self-explanatory, though some will be less so without the book as a reference. The table of contents of the book can be seen [here \(pdf\)](#).

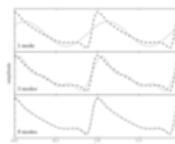
Figure Contents

Each chapter links to a page with thumbnails of the figures from the chapter.

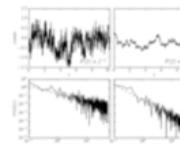
- Chapter 1: Introduction
- Chapter 2: Fast Computation and Massive Datasets
- Chapter 3: Probability and Statistical Distributions
- Chapter 4: Classical Statistical Inference
- Chapter 5: Bayesian Statistical Inference
- Chapter 6: Searching for Structure in Point Data
- Chapter 7: Dimensionality and its Reduction
- Chapter 8: Regression and Model Fitting
- Chapter 9: Classification
- Chapter 10: Time Series Analysis
- Appendix

Chapter 10: Time Series Analysis

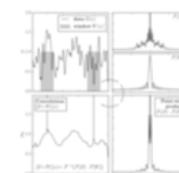
This chapter covers the analysis of both periodic and non-periodic time series, for both regularly and irregularly spaced data.



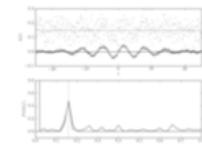
Fourier Reconstruction of
RR-Lyrae Templates



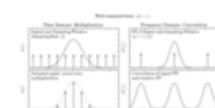
Generating Power-law
Light Curves



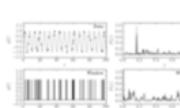
Plot a Diagram explaining
a Convolution



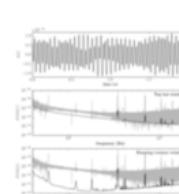
Fast Fourier Transform
Example



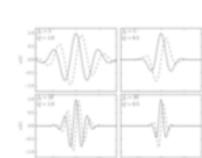
The effect of Sampling



The effect of Sampling



Plot the power spectrum of
the LIGO big dog event



Examples of Wavelets

1) Introduction

- astroML

If you haven't already, please install astroML by following instructions at:

https://www.astroml.org/user_guide/installation.html

or google for “astroML”, go to

<https://www.astroml.org>

and scroll down to the table of contents.

Disclaimer: you may need to run python 2.7 (that is, not python 3.x) as I noticed some errors I didn't understand when testing this morning! Sorry!

1) Introduction

- astroML

To test, start ipython shell and do:

```
[Macintosh-3:~ ivezic$ ipython
[TerminalIPythonApp] WARNING | Config option `ignore_old_config` not recognized
by `TerminalIPythonApp`.
Python 2.7.13 |Anaconda custom (x86_64)| (default, Dec 20 2016, 23:05:08)
Type "copyright", "credits" or "license" for more information.

IPython 5.3.0 -- An enhanced Interactive Python.
?          --> Introduction and overview of IPython's features.
%quickref --> Quick reference.
help       --> Python's own help system.
object?    --> Details about 'object', use 'object??' for extra details.

[In [1]: from astroML.datasets import fetch_sdss_spectrum

[In [2]: spec = fetch_sdss_spectrum(1615, 53166, 513)

In [3]: ]
```

If there are no error messages, you are good to go!
If there are problems (e.g. with “GMM”) go to py2.7

We need more downloads, which you can do by tomorrow:

Download dataAll.tar.gz (245 MB) as <https://ls.st/e3k> **SLOW!**

Make a directory astroML_data in your home directory and download dataAll.tar.gz to that directory.

Then unpack it:

```
> cd ~/astroML_data  
> gunzip dataAll.tar.gz (possibly done by your machine)  
> tar -xvf dataAll.tar
```

and you should see 4 *.fit files, 2 *.npy files and one *.npz file.

In addition, clone my directory with lectures (wherever):

```
> git clone git@github.com:dirac-institute/SPSAS2019.git
```

ALSO SLOW!

Test astroML and git repository installation by starting jupyter notebook with file

.../SPSAS2019/lectures/test.ipynb and executing the code there:

> jupyter notebook .../SPSAS2019/lectures/test.ipynb &

It should produce this figure:

