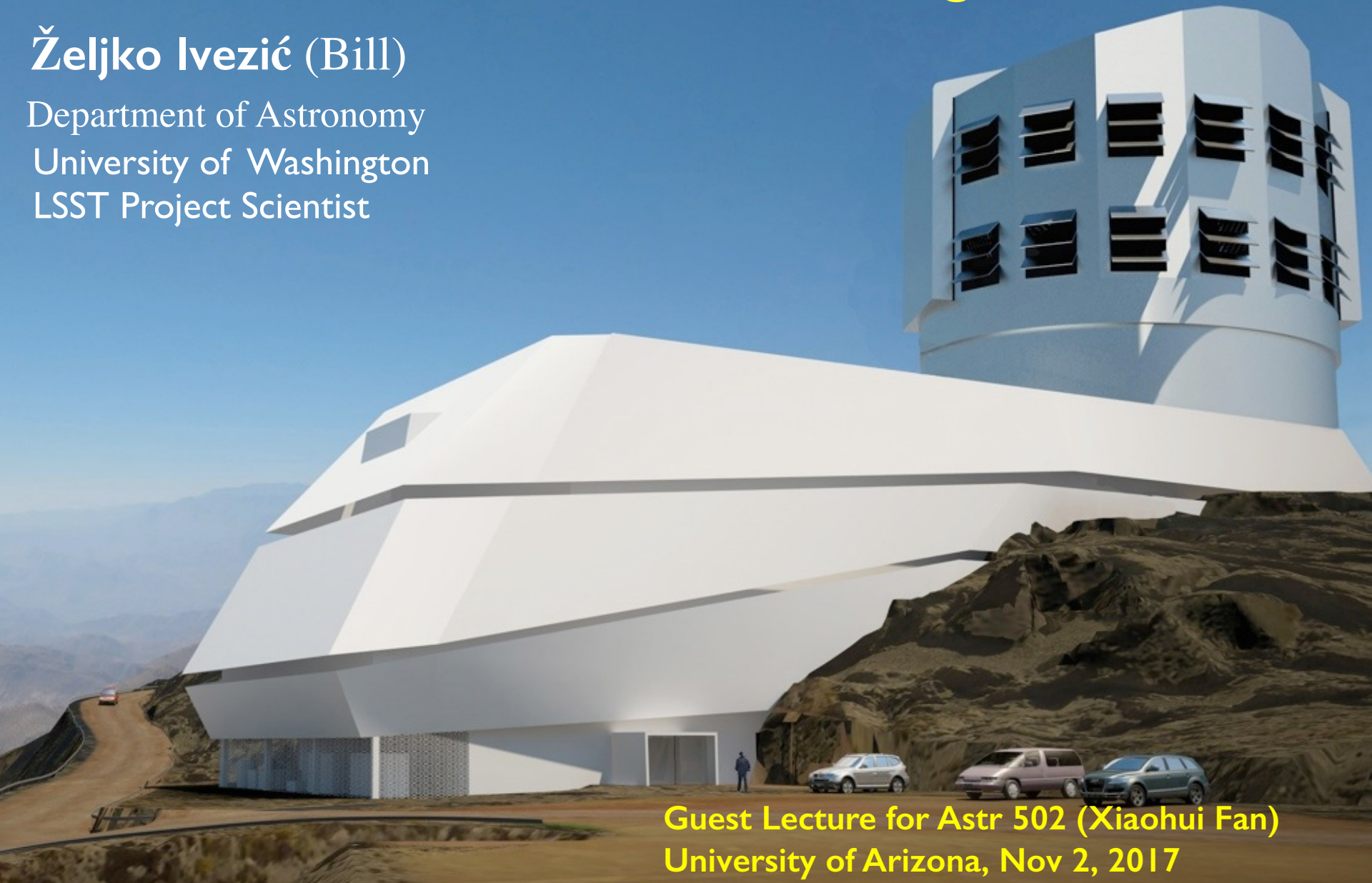
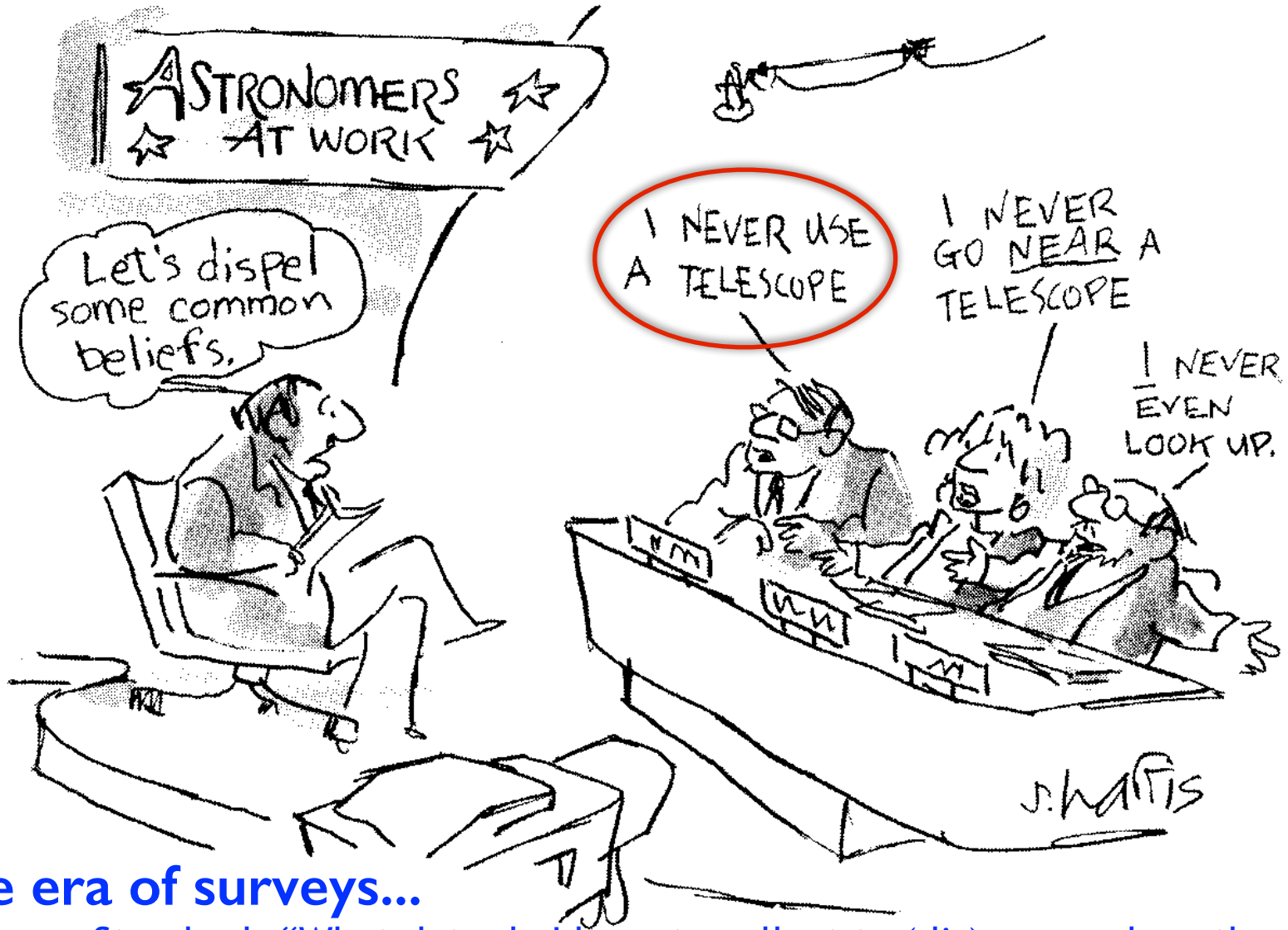
LSST: the greatest ever movie of the Universe and associated astro-statistical challenges

Željko Ivezić (Bill)
Department of Astronomy
University of Washington
LSST Project Scientist



Guest Lecture for Astr 502 (Xiaohui Fan)
University of Arizona, Nov 2, 2017

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



The era of surveys…

- Standard: "What data do I have to collect to (dis)prove a hypothesis"?
- Data-driven: "What theories can I test given the data I already have?"

Outline

- Brief overview of LSST
  – science drivers
  – system design
- Some data analysis challenges
  - multi-color time-resolved faint sky map
  - 20 billion galaxies (median redshift ~1)
  - 20 billion stars (to the edge of the Milky Way)
  - “millions and millions” of supernovae, quasars, asteroids…
  - “Everything I’d like to do with LSST data, but don’t know (yet) how”
- Two vignettes about the Central Limit Theorem

3.6×10^{-31} erg/s/cm^2/Hz
36 nJy
100x fainter than SDSS

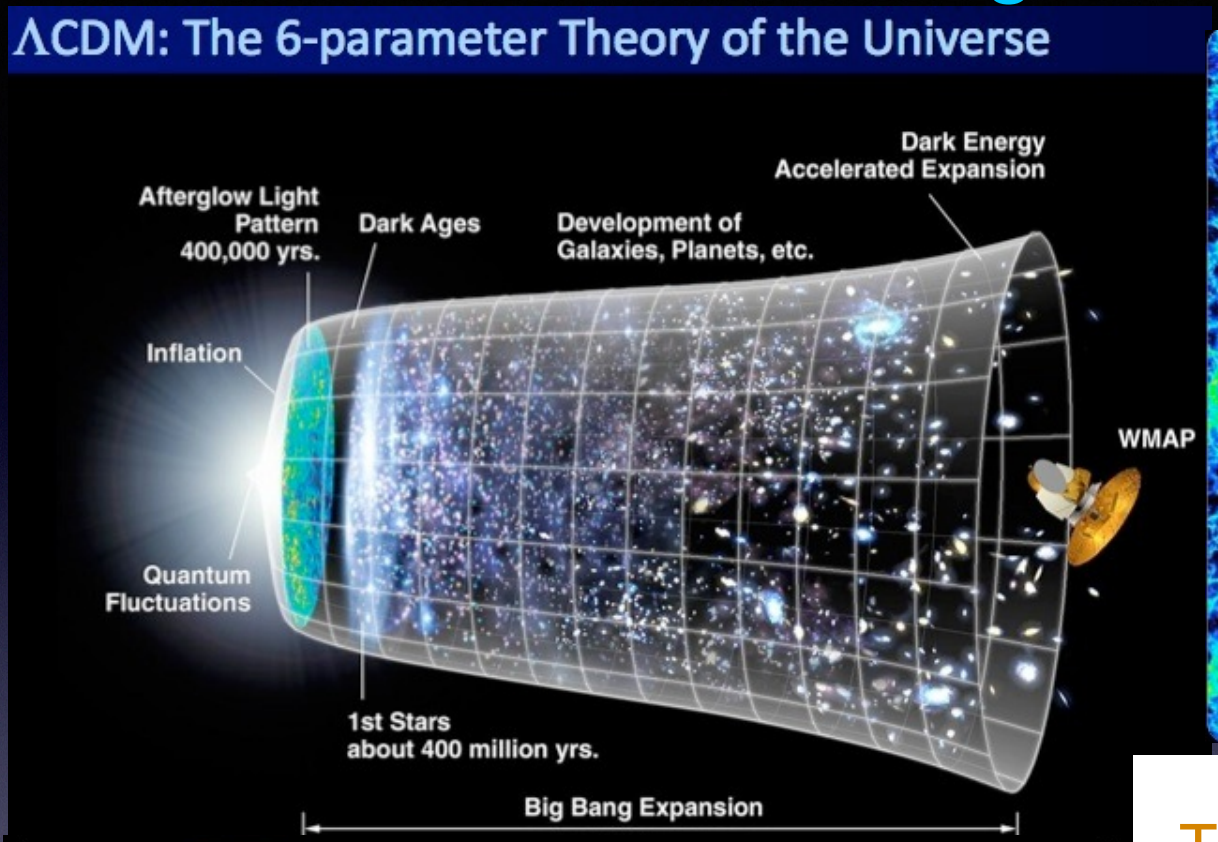
LSST in one sentence:

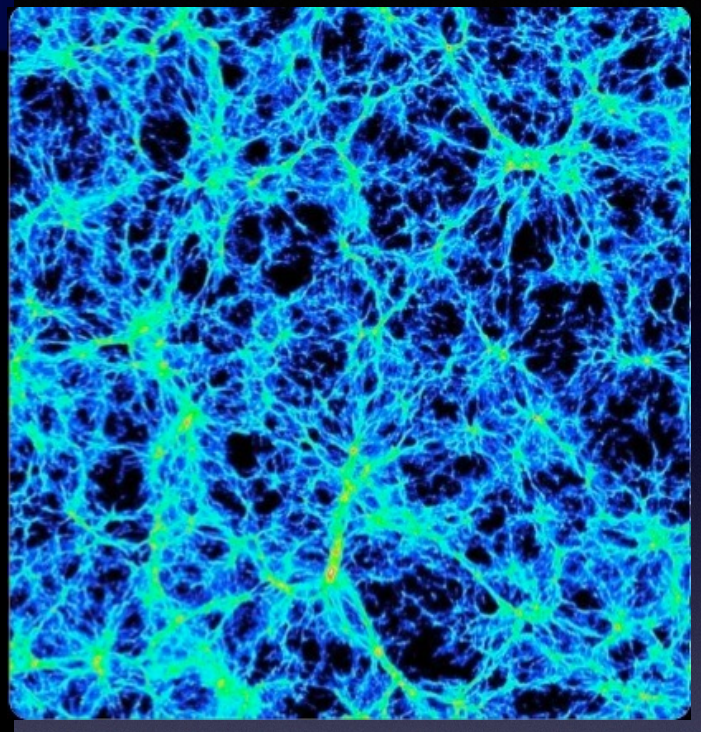
An optical/near-IR survey of half the sky in ugrizy bands to r~27.5 based on ~1000 visits over a 10-year period:

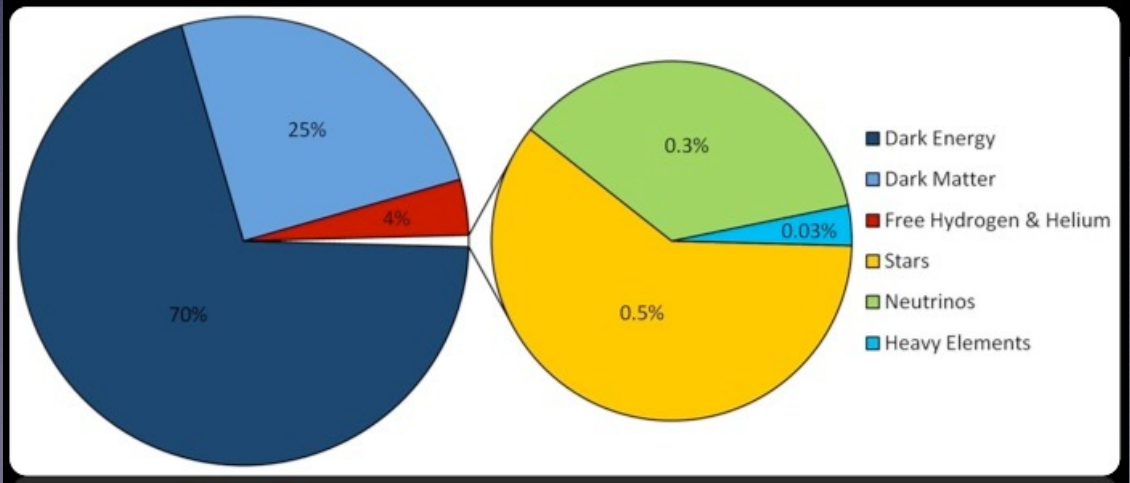
More information at www.lsst.org and arXiv:0805.2366

A catalog of 20 billion stars and 20 billion galaxies with exquisite photometry, astrometry and image quality!

New Cosmological Puzzles







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

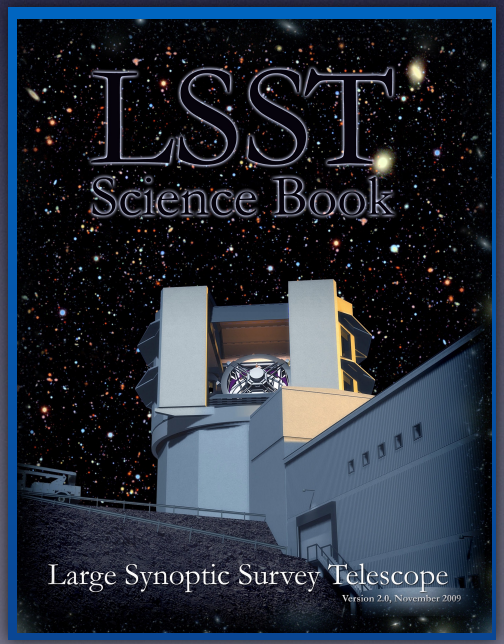
LSST Science Themes

- 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

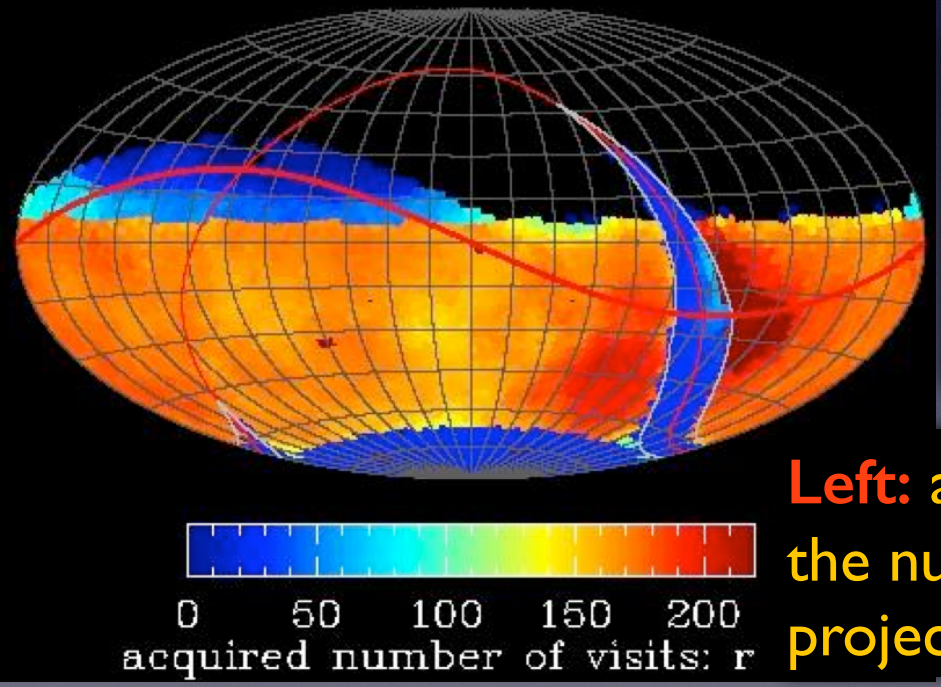
Summarizes LSST hardware, software, and observing plans, science enabled by LSST, and educational and outreach opportunities

245 authors, 15 chapters, 600 pages



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)

SDSS
gri
3.5'x3.5'
r~22.5

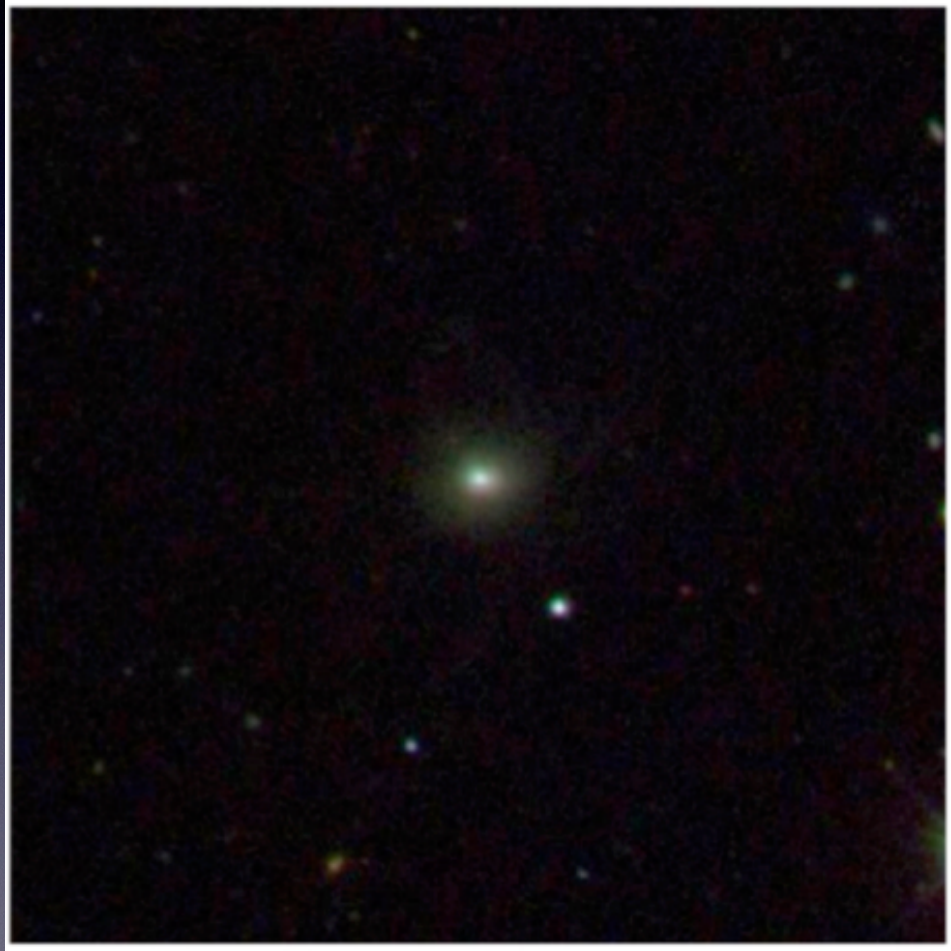
HSC
gri
3.5'x3.5'
r~27

Thanks to Robert Lupton

Not just point source depth: faint surface brightness limit

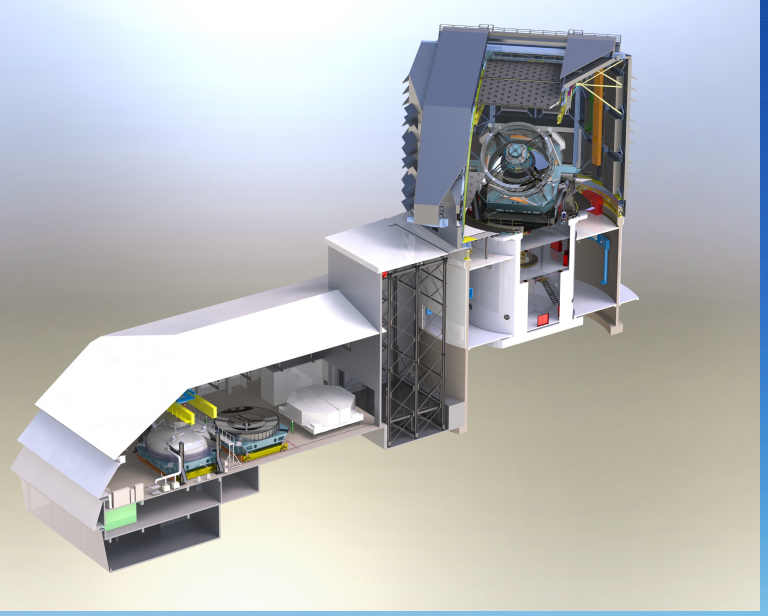
SDSS

3x3 arcmin, gri



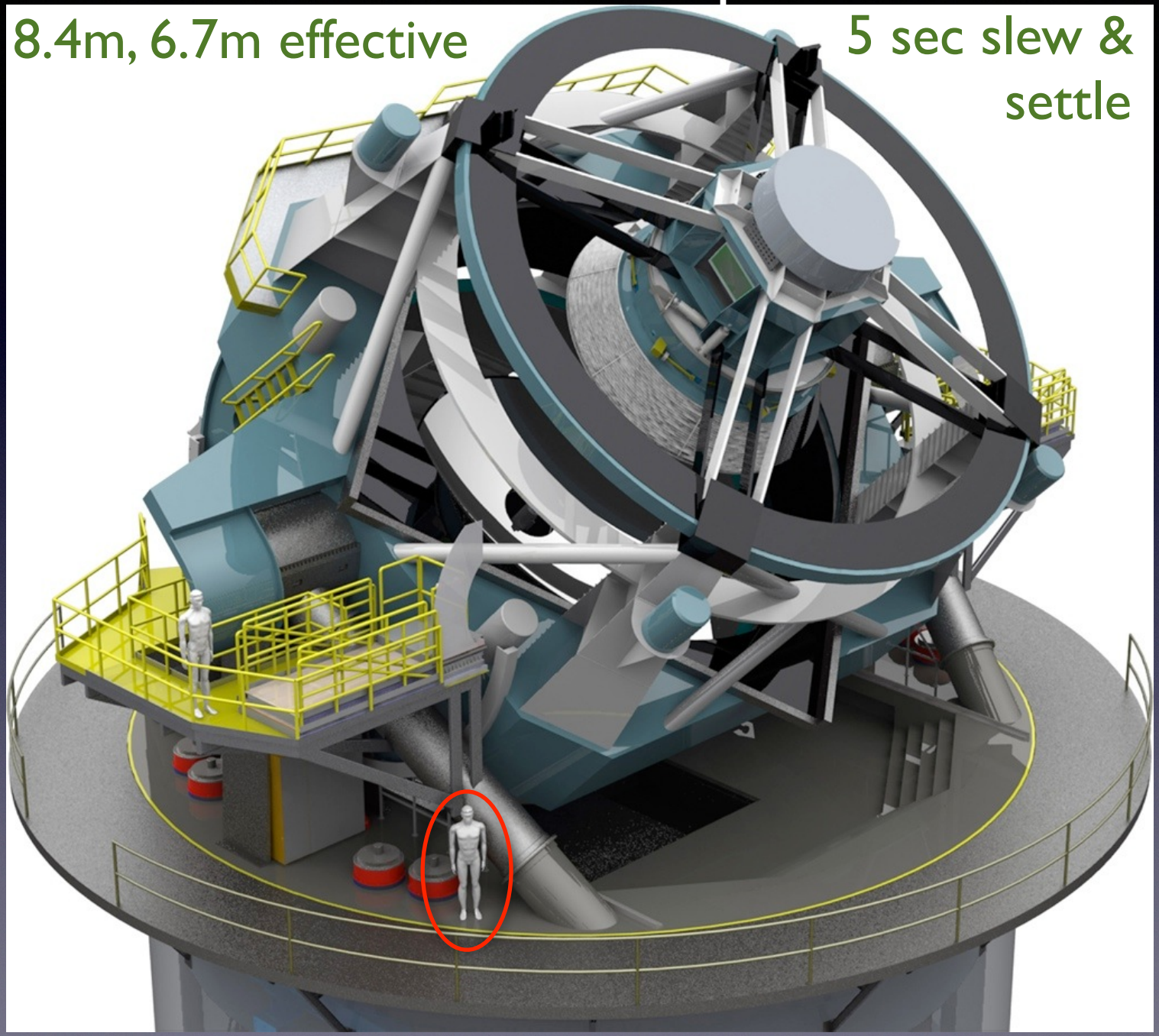
MUSYC r ~26

Let’s build LSST!




Oct 17, 2017
First light: 2020

LSST Telescope

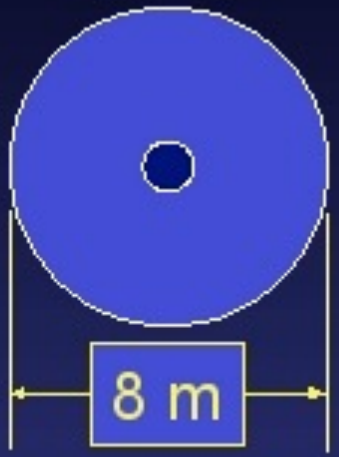


The field-of-view comparison: Gemini vs. LSST

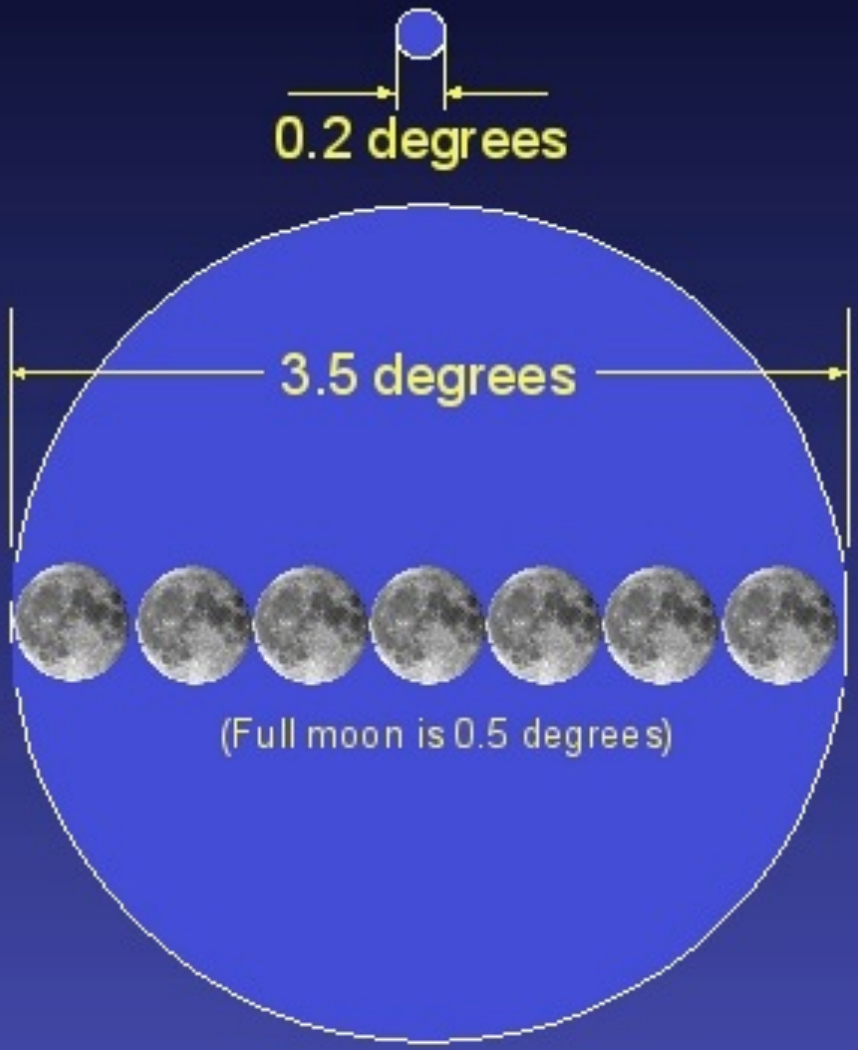
Gemini South Telescope

LSST

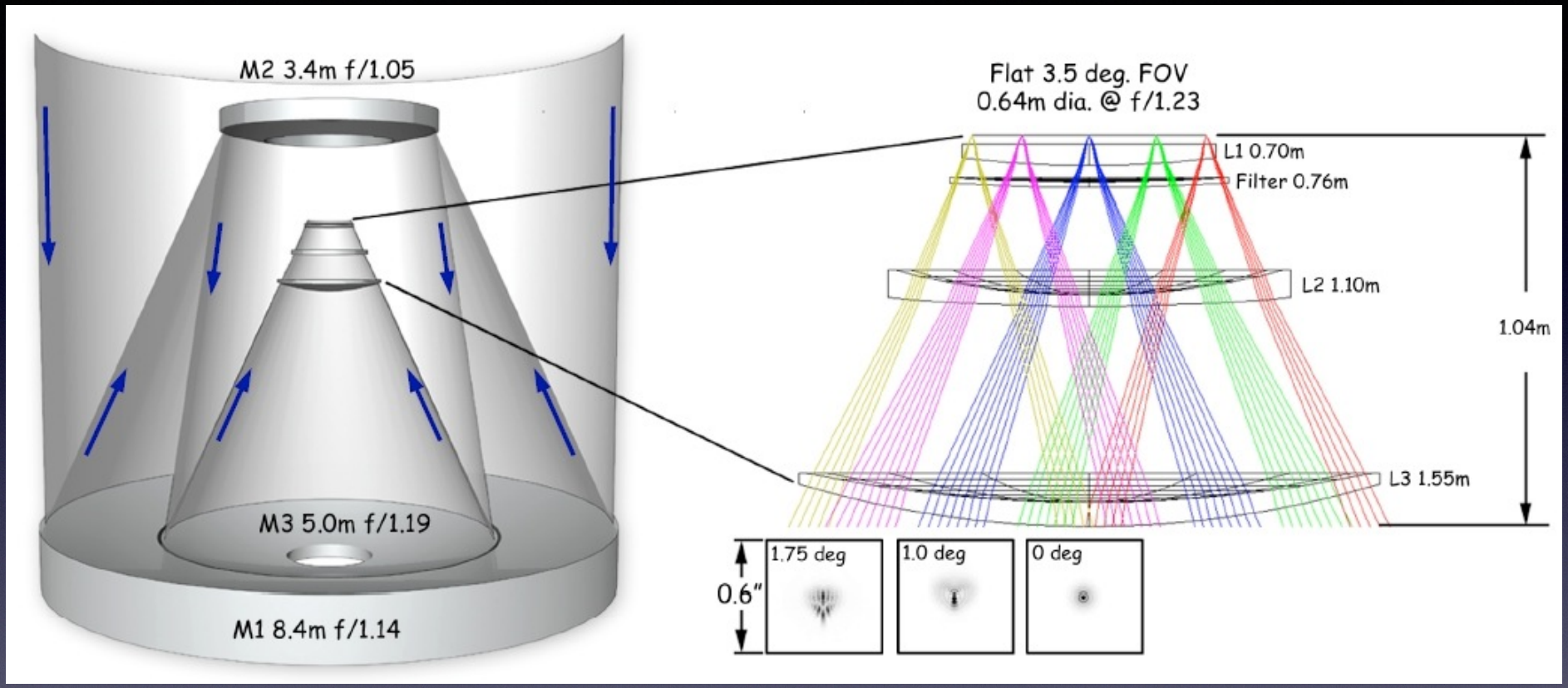
Primary Mirror Diameter



Field of View

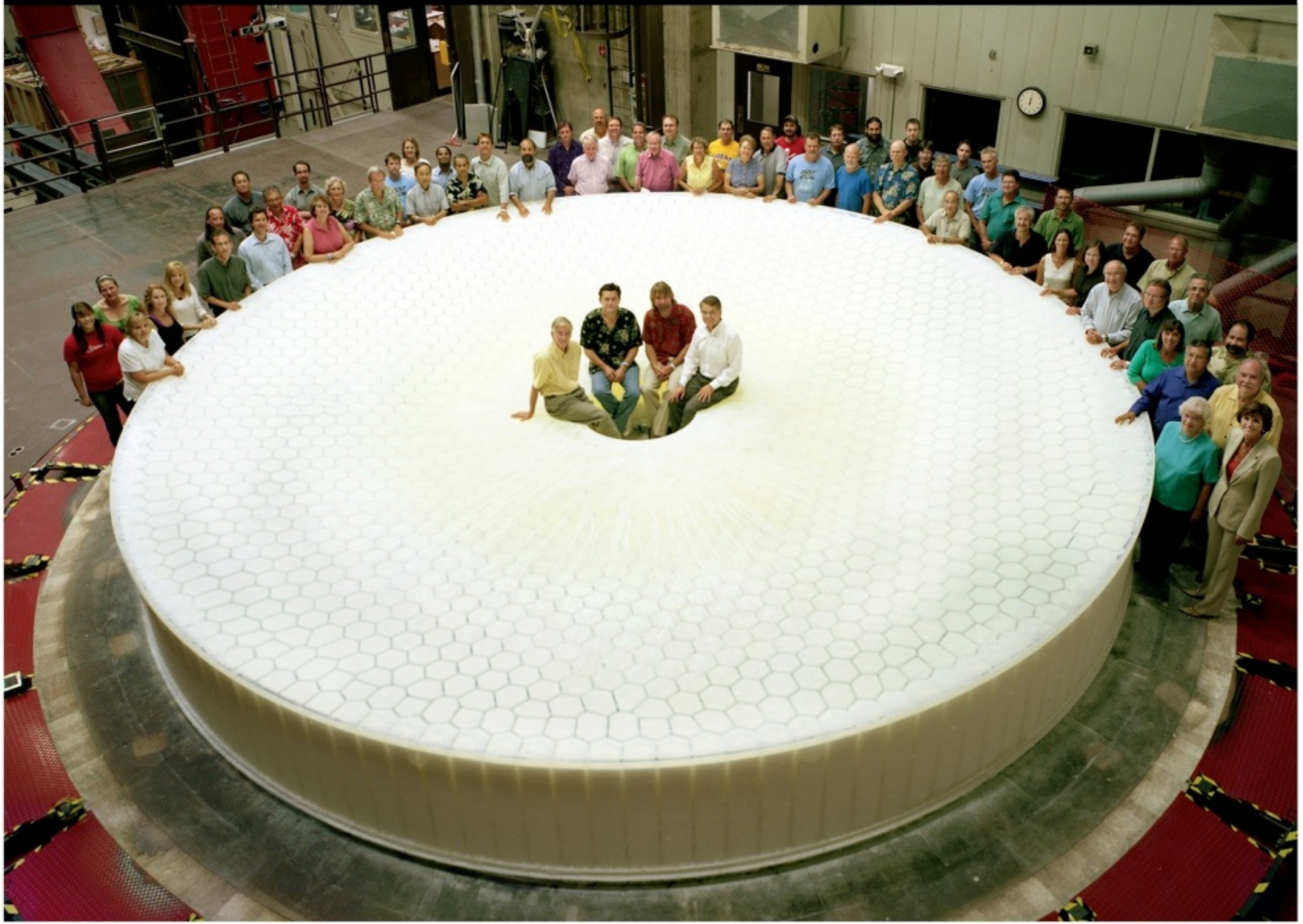


Optical Design for LSST



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



LSST Primary/Tertiary Mirror Blank

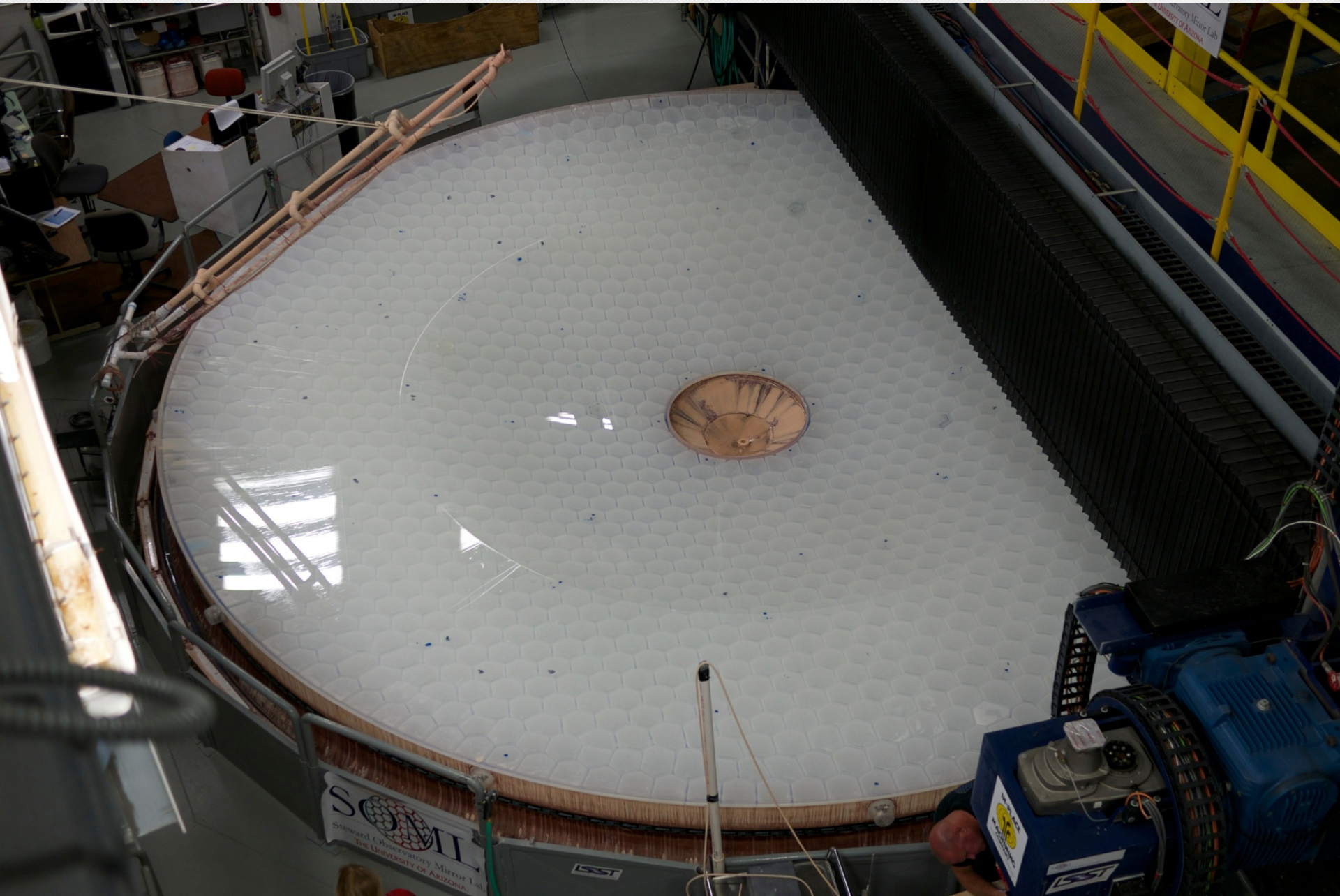
August 11, 2008, Steward Observatory Mirror Lab, Tucson, Arizona

The University of Arizona, Tucson Arizona

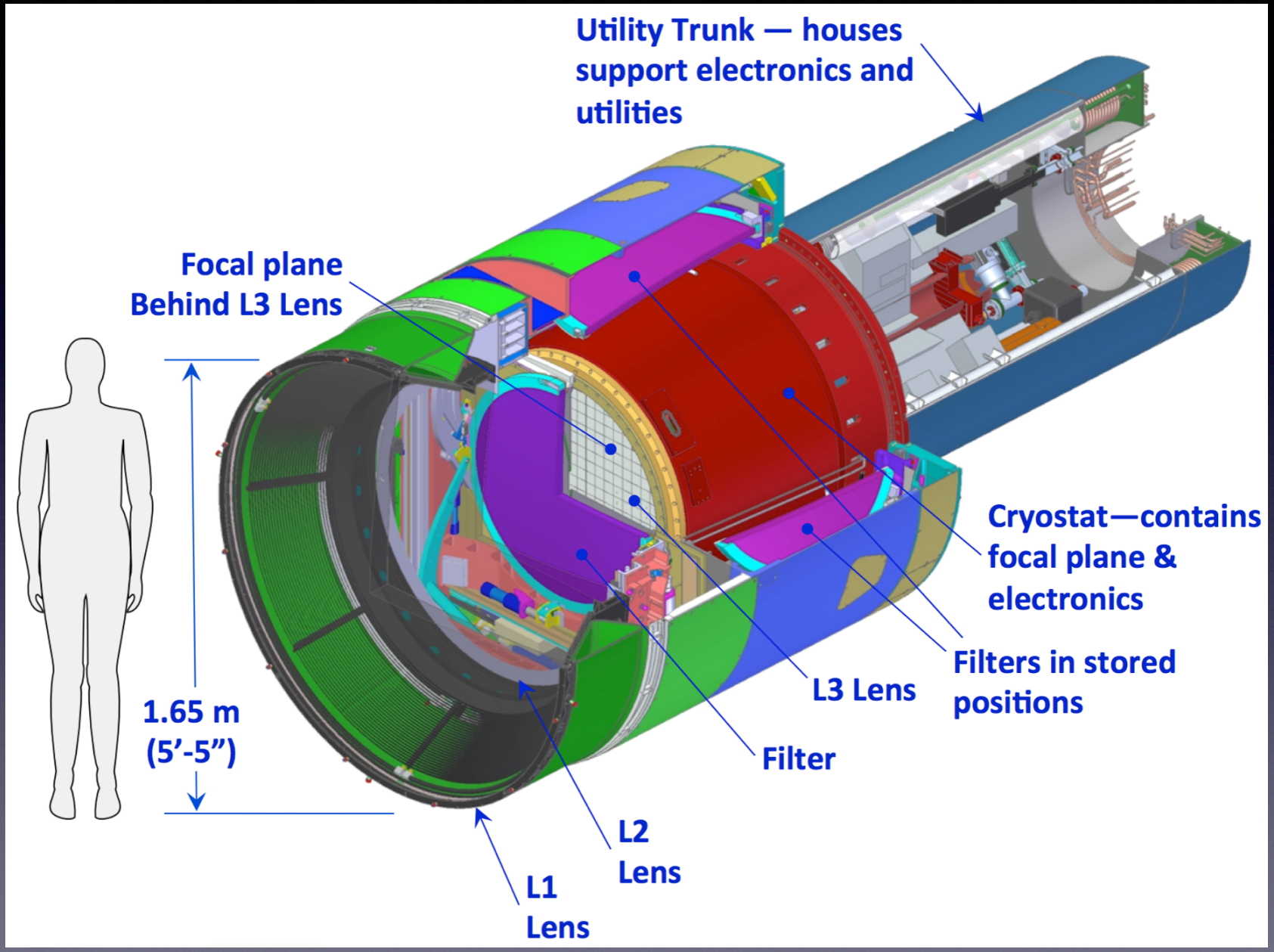
SOML

Large Synoptic Survey Telescope

Done!

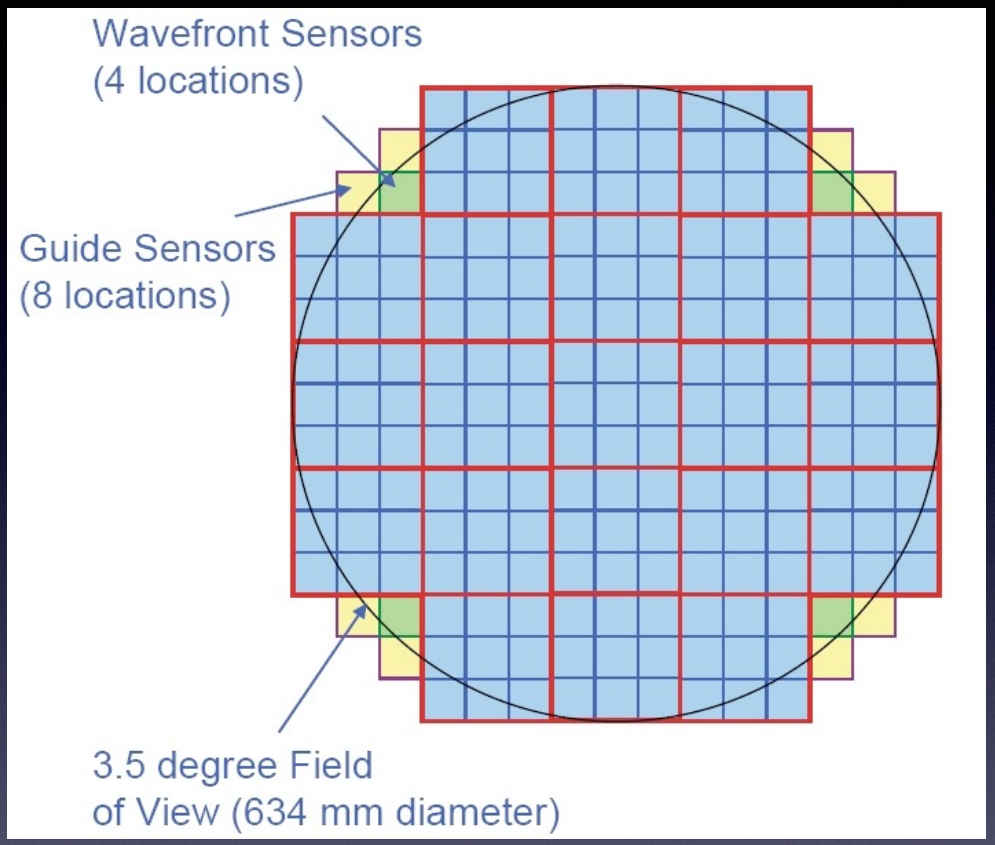

SOML
Steward Observatory Mirror Lab
The University of Arizona

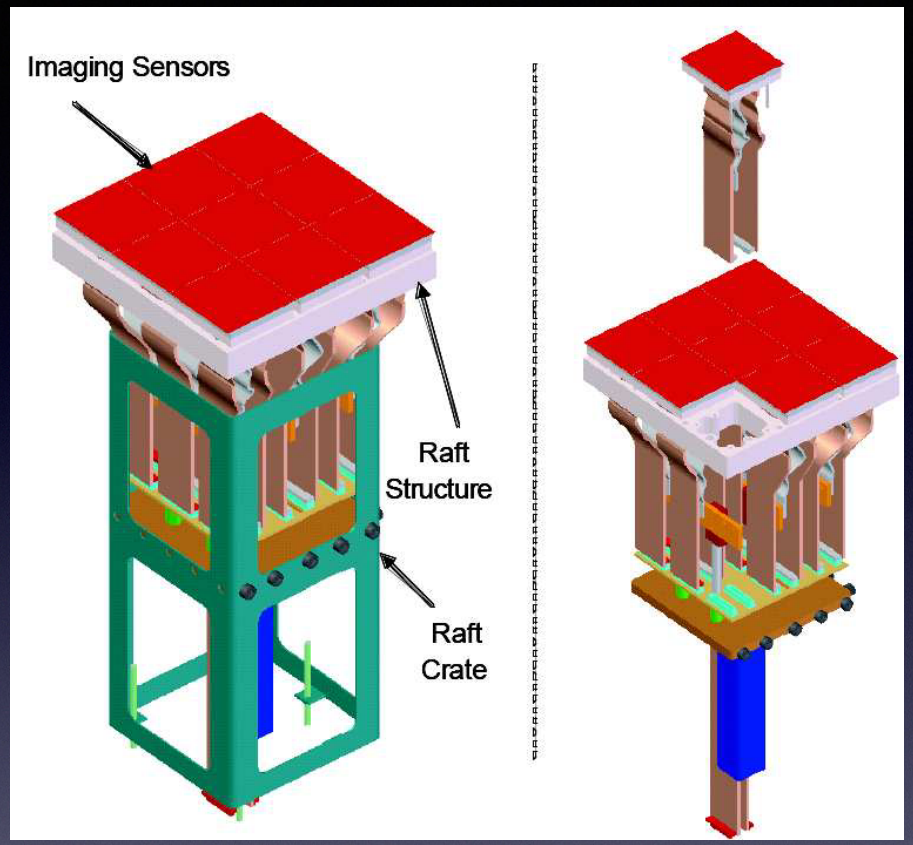
LSST camera



The largest astronomical camera: 2800 kg, 3.2 Gpix

LSST camera





Modular design: 3200 Megapix = 189 x16 Megapix CCD

9 CCDs share electronics: raft (=camera)

Problematic rafts can be replaced relatively easily

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)

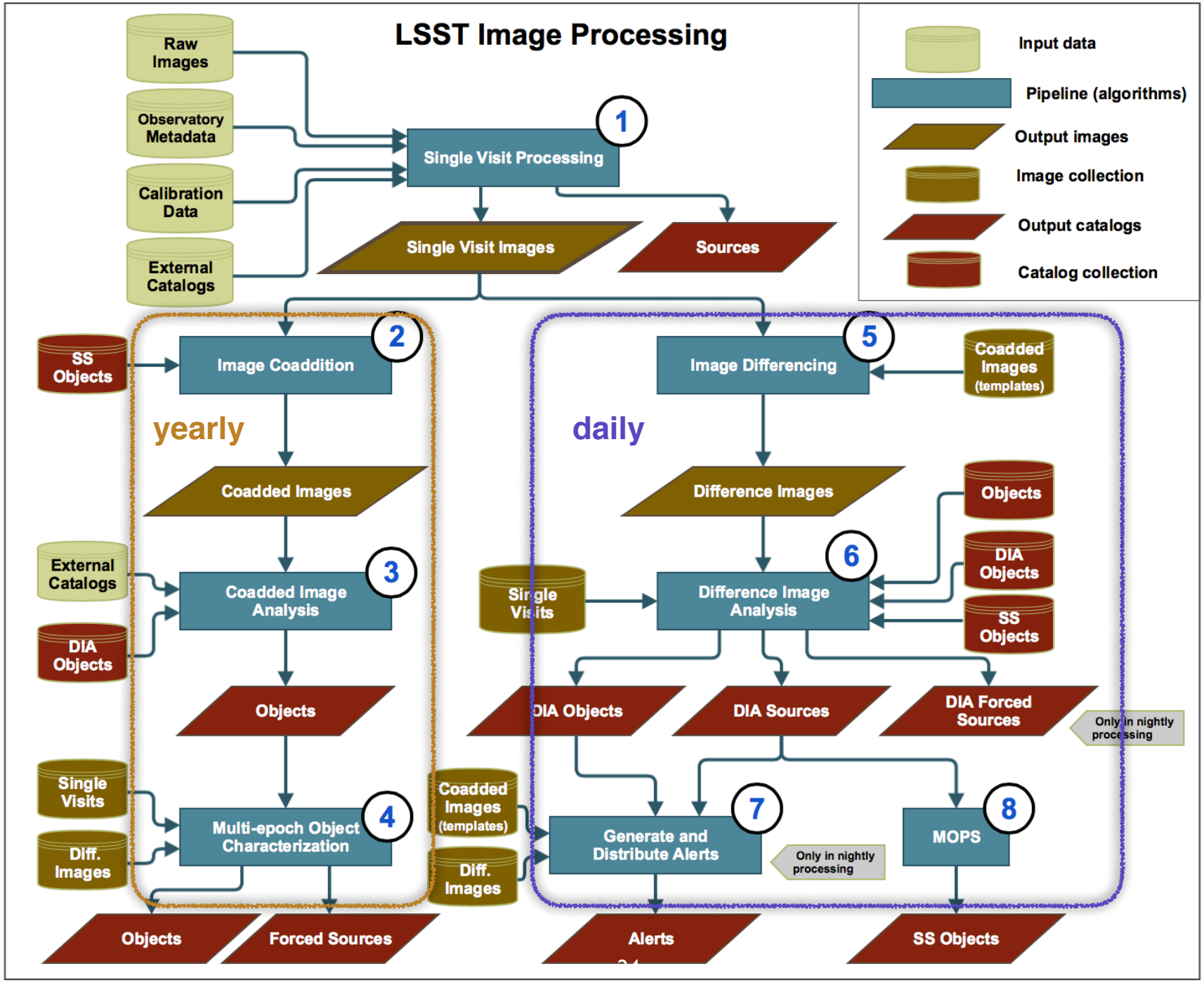
3) Produce catalogs ("model parameters") of detected objects (37 billion)

4) **Serve** images, catalogs and all other metadata, that is, **LSST data products to LSST users**

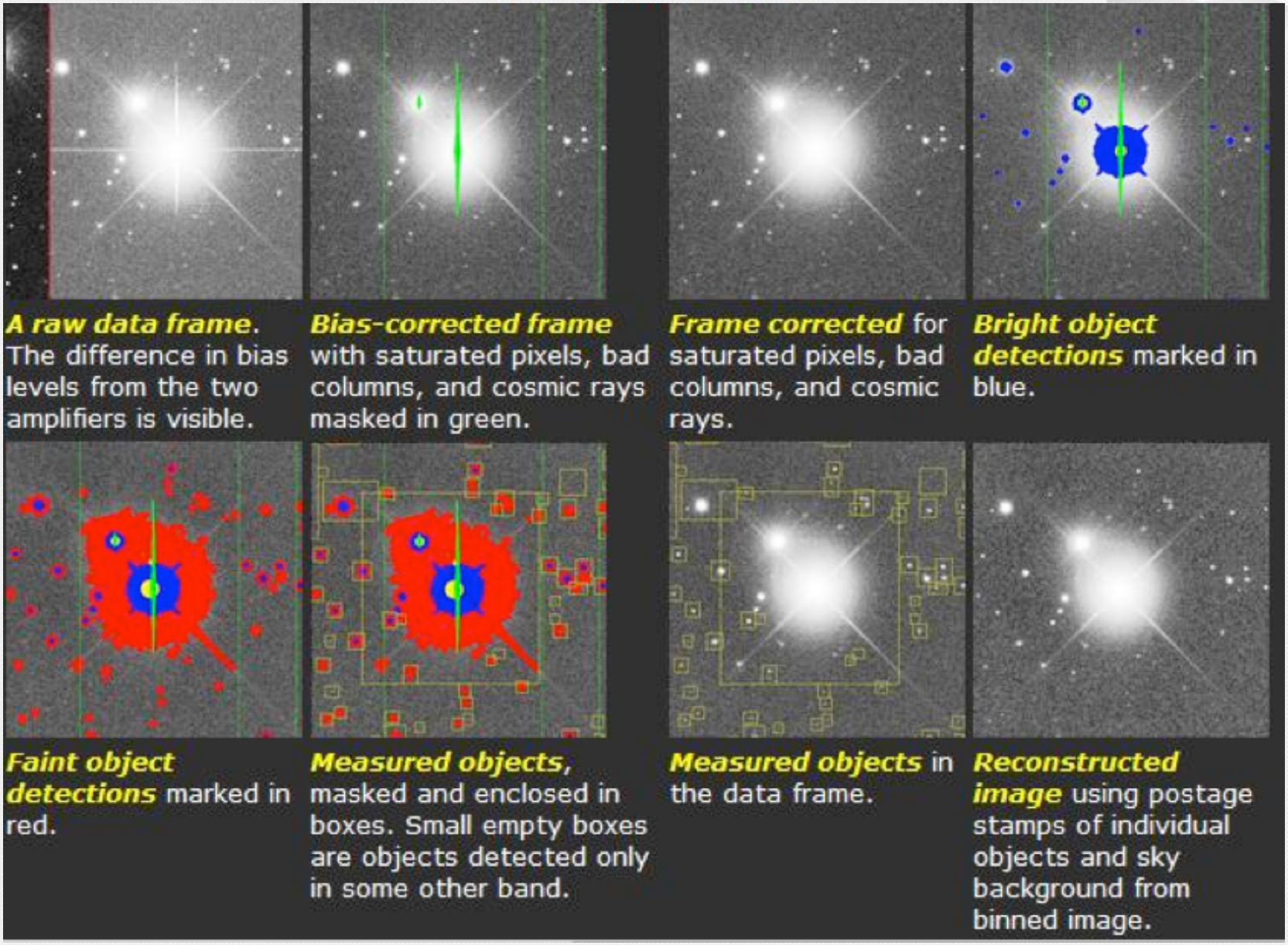
The ultimate deliverable of LSST is not just the telescope, nor the camera, but **the fully reduced science-ready data** as well. **Software!**

Software: the subsystem with the highest risk

- 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 new code (C++/python)


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

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.

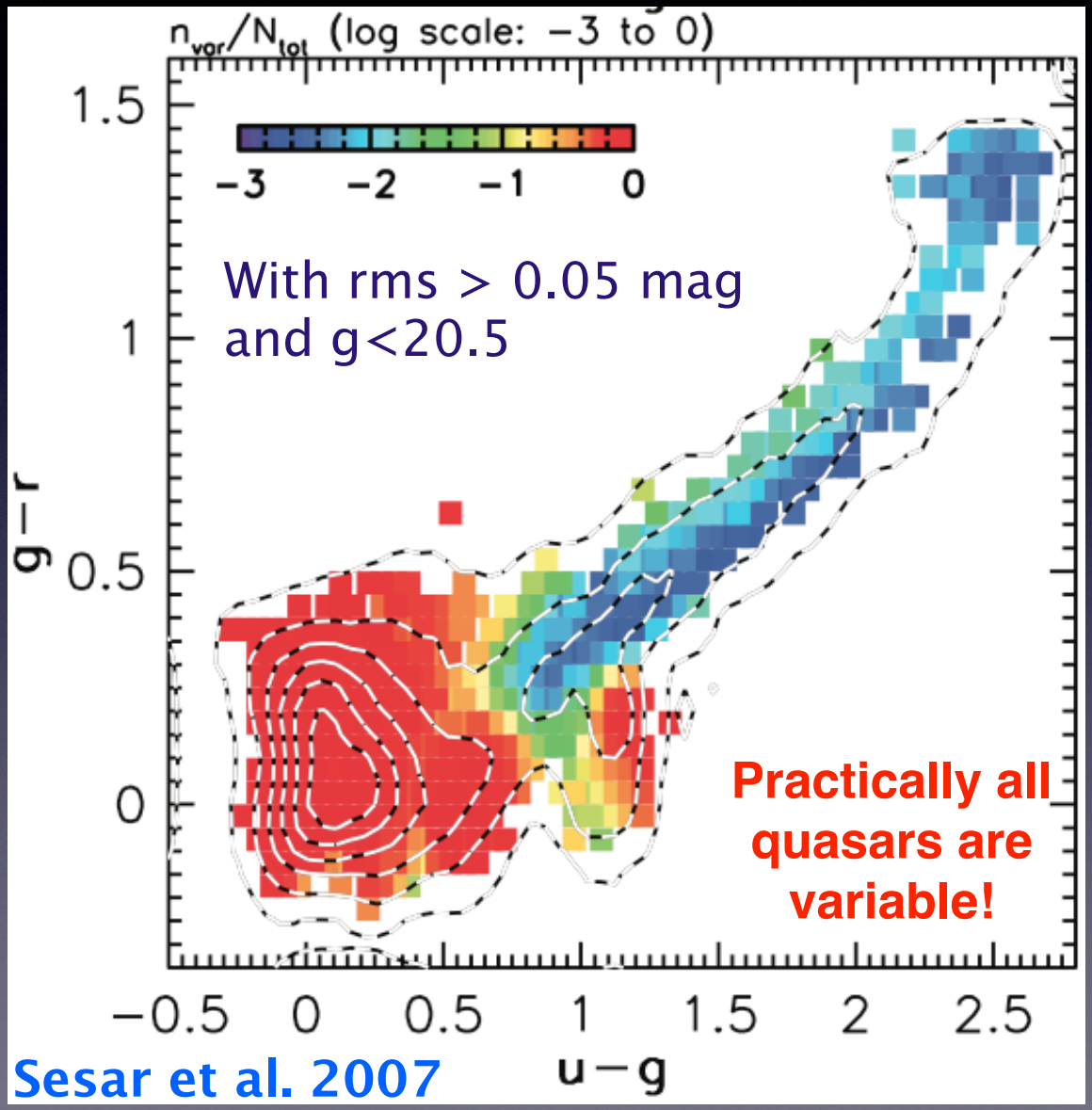
Measured objects, masked and enclosed in boxes. Small empty boxes are objects detected only in some other band.

Measured objects in the data frame.

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

An example of analysis of time domain data

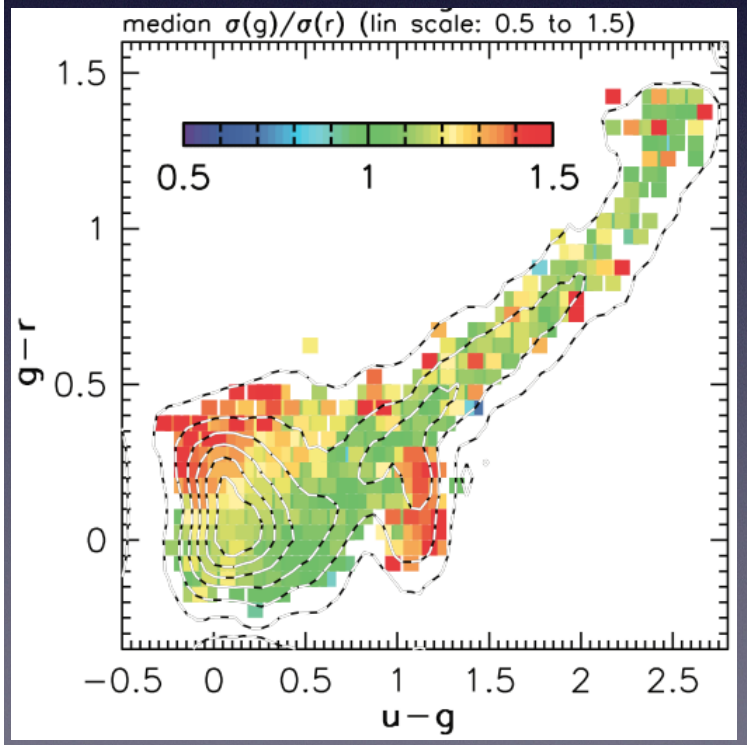
The fraction of variable objects in SDSS Stripe 82:



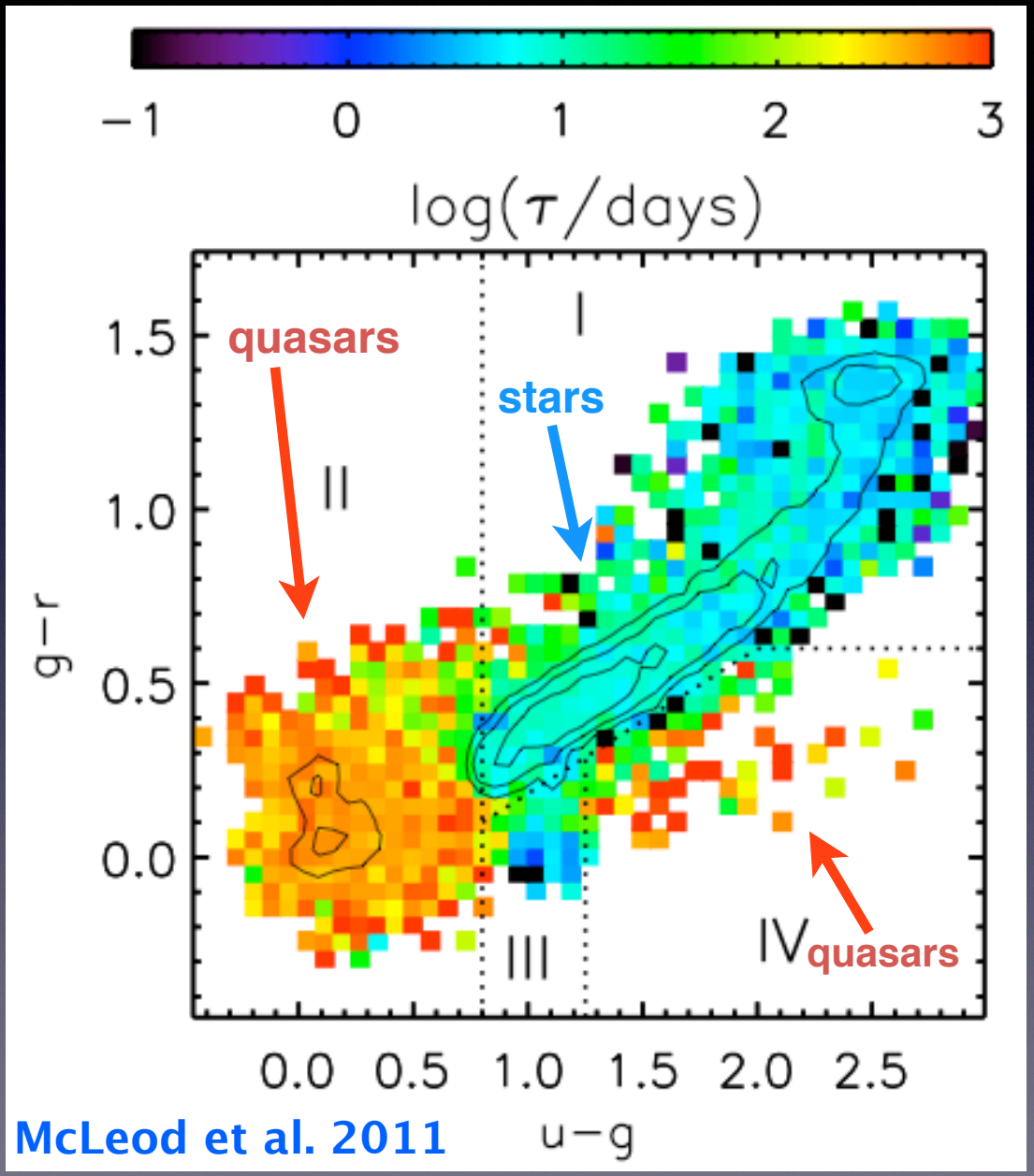
Sesar et al. 2007

The sample is dominated by quasars and RR Lyrae.

Quasars and RR Lyrae have different variability properties: rms(g)/rms(r)



The variability time scales



Time scale τ is defined via **damped random walk** (because not all variable sources are periodic)

Quasars are easily distinguished from stars by their long time scales.

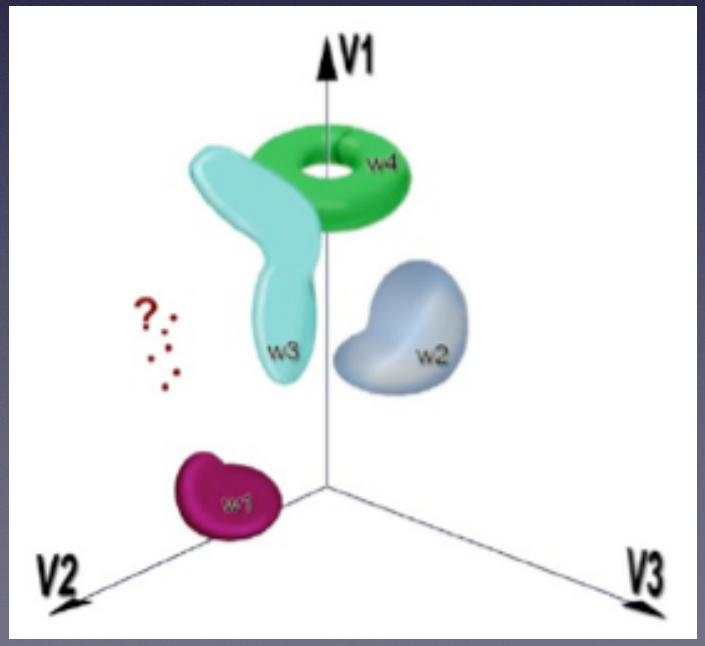
Variability is even better than color selection!

Case study: light curve data and proper motion data for over 1 million sources from SDSS Stripe 82 (all are publicly available)

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

Data analysis challenges in the era of Big Data

1) Large data volume (petabytes)

2) Large numbers of objects (billions)

3) Highly multi-dimensional spaces (thousands)

4) Unknown statistical distributions

5) Time-series data (irregular sampling)

6) Heteroscedastic errors, truncated, censored and missing data

7) Unreliable quantities (e.g. unknown systematics and random errors)

The bottleneck will not be (is not any more?) data availability but instead our ability to extract useful and reliable information from data.

Everything I'd like to do with LSST data, but don't know (yet) how (arXiv:1612.04772)

1) Interpretation of spectral energy distributions (SEDs)

2) Spatial correlations

3) Moving objects

4) Variable objects

5) Systematic measurement uncertainties

6) Astrophysical simulations and astrophysical systematics

7) LSST System Enhancements

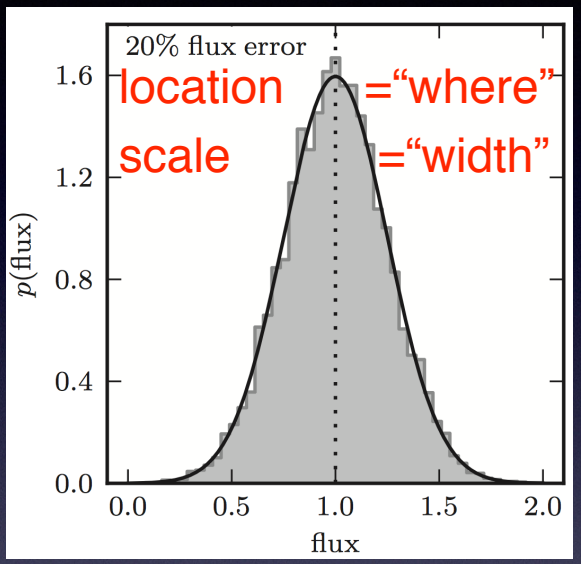
8) New algorithms in LSST

Two vignettes about the CLT

- How to estimate the location and scale of an underlying pdf from which data were drawn?
  - Central Limit Theorem
- What to do when CLT doesn't work?
- When CLT should not be used even if it works?

• How to compute an "average" value?

o First, "average" can mean different quantities, most often the mean and median (often, and erroneously, "average" and "mean" are considered synonymous)



o Given a list of numbers, x_i, i=1…N, their mean is

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

o "Everyone" knows that! We often take the mean of many measurements to improve the accuracy of the final result.

o What is not known by everyone is why and when this "averaging" works, and how to estimate how good it is. These answers are provided by the Central Limit Theorem.

- **Why do we estimate location using the mean?**
  - Because of the Central Limit Theorem:

Given an *arbitrary* distribution $h(x)$, characterized by its location μ and scale σ, the mean of N values x_i drawn from that distribution will approximately follow a Gaussian distribution with $N(\mu,\sigma/\sqrt{N})$, with the approximation accuracy improving with N.

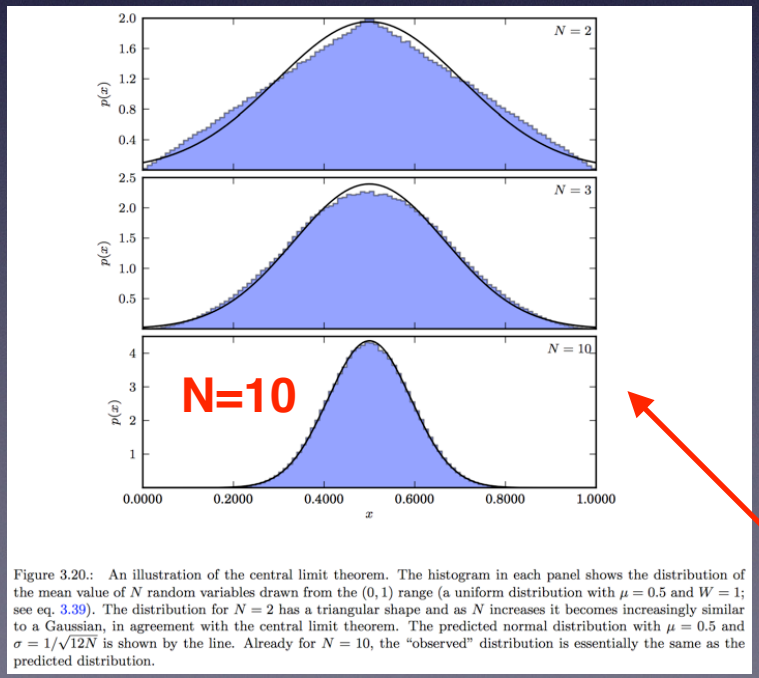
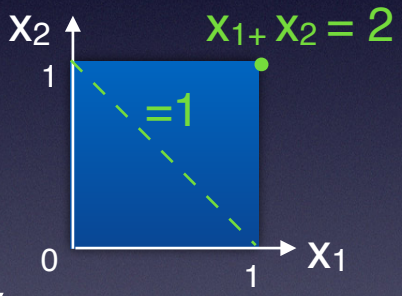


Figure 3.20.: An illustration of the central limit theorem. The histogram in each panel shows the distribution of the mean value of N random variables drawn from the $(0, 1)$ range (a uniform distribution with $\mu = 0.5$ and $W = 1$; see eq. 3.39). The distribution for $N = 2$ has a triangular shape and as N increases it becomes increasingly similar to a Gaussian, in agreement with the central limit theorem. The predicted normal distribution with $\mu = 0.5$ and $\sigma = 1/\sqrt{12N}$ is shown by the line. Already for $N = 10$, the "observed" distribution is essentially the same as the predicted distribution.



The CLT can be easily proven using standard tools from statistics, such as characteristic functions and convolutions. Here is an example of CLT in action based on a uniform distribution.

- **Why do we estimate location using the mean?**
  - Because of the Central Limit Theorem!
  - This is a remarkable result since the details of the distribution $h(x)$ are not specified - we can "average" our measurements (i.e., compute their mean value) and expect the $1/\sqrt{N}$ improvement in accuracy *regardless of details in our measuring apparatus!*

But: it was implicitly assumed that h(x) has finite σ - not always true!
The Cauchy (Lorentzian) distribution: **σ is undefined**

$$p(x|\mu,\gamma) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{\gamma^2+(x-\mu)^2}\right)$$

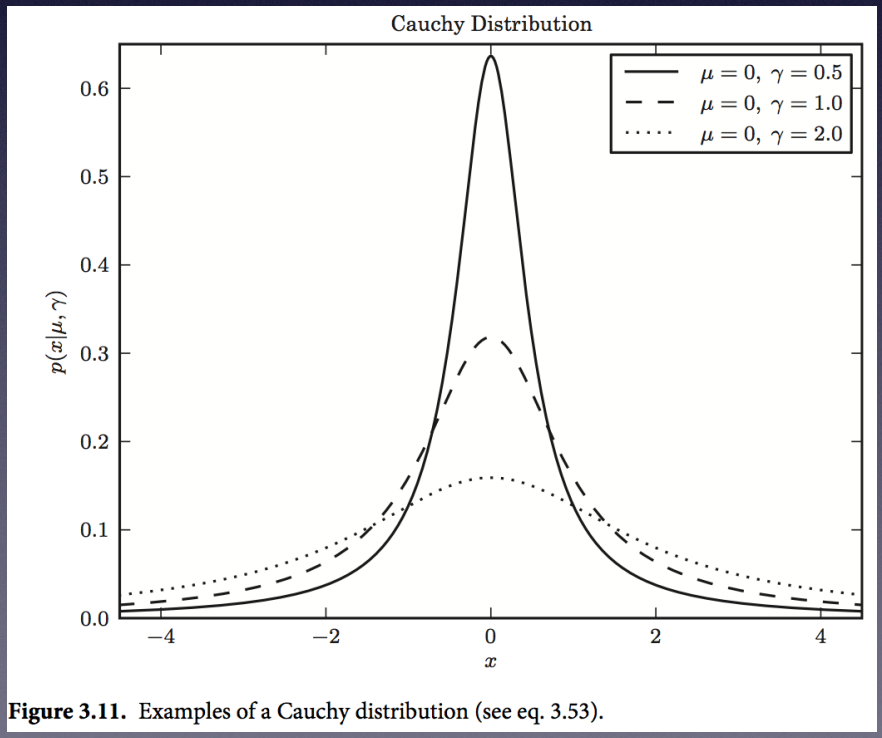


Figure 3.11. Examples of a Cauchy distribution (see eq. 3.53).

- **How do we estimate location for the Cauchy distribution?**

$$p(x|\mu,\gamma) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{\gamma^2+(x-\mu)^2}\right)$$

Task: given measurements x_i, i=1…N, drawn from the Cauchy distribution, find the best estimate of μ, let's call it μ^0, and its uncertainty, σ_μ

In this case, using the mean value is a very bad idea!

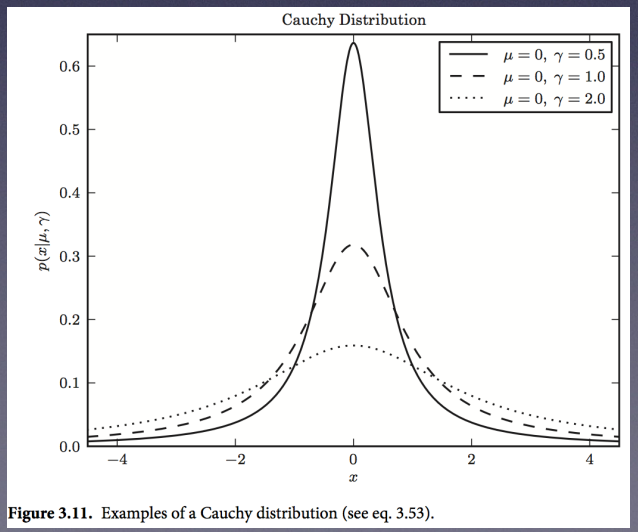
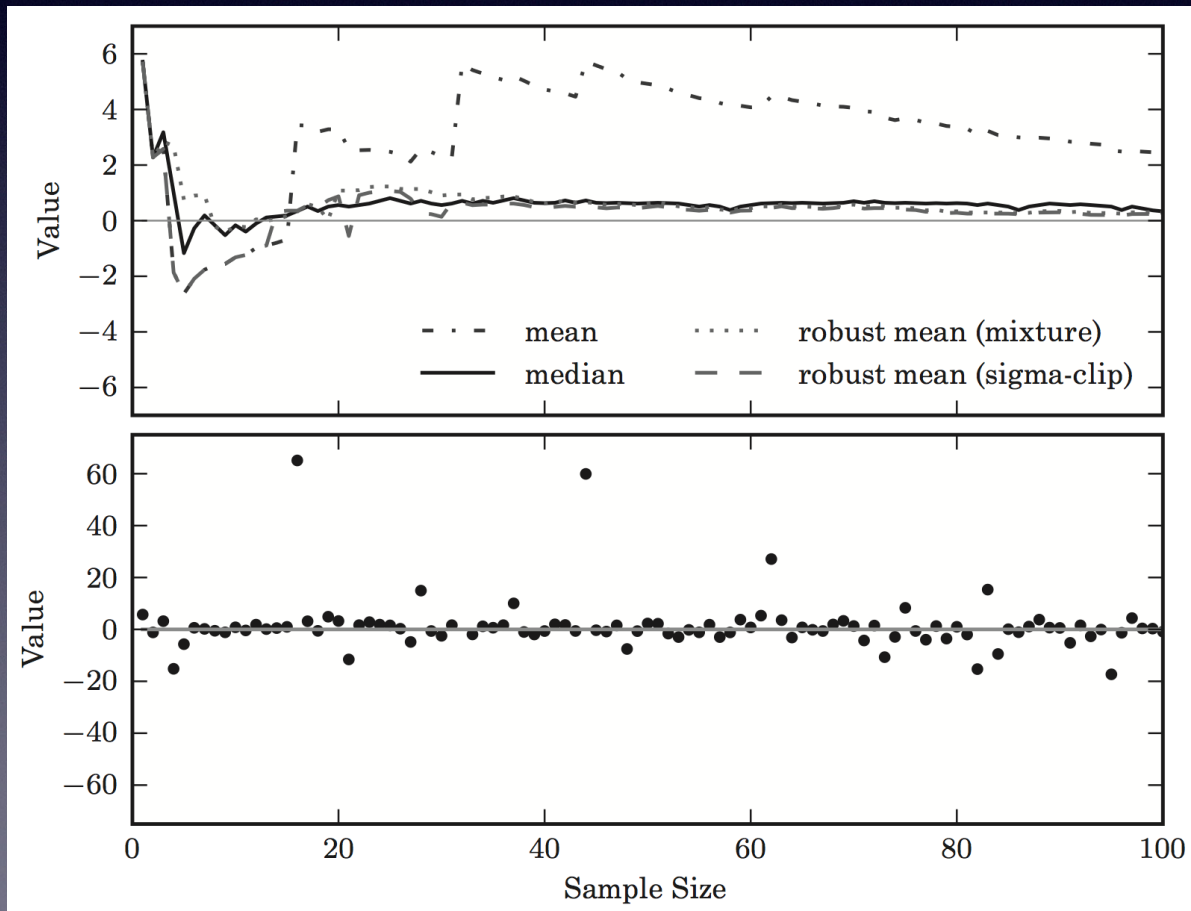


Figure 3.11. Examples of a Cauchy distribution (see eq. 3.53).



Vignette #1

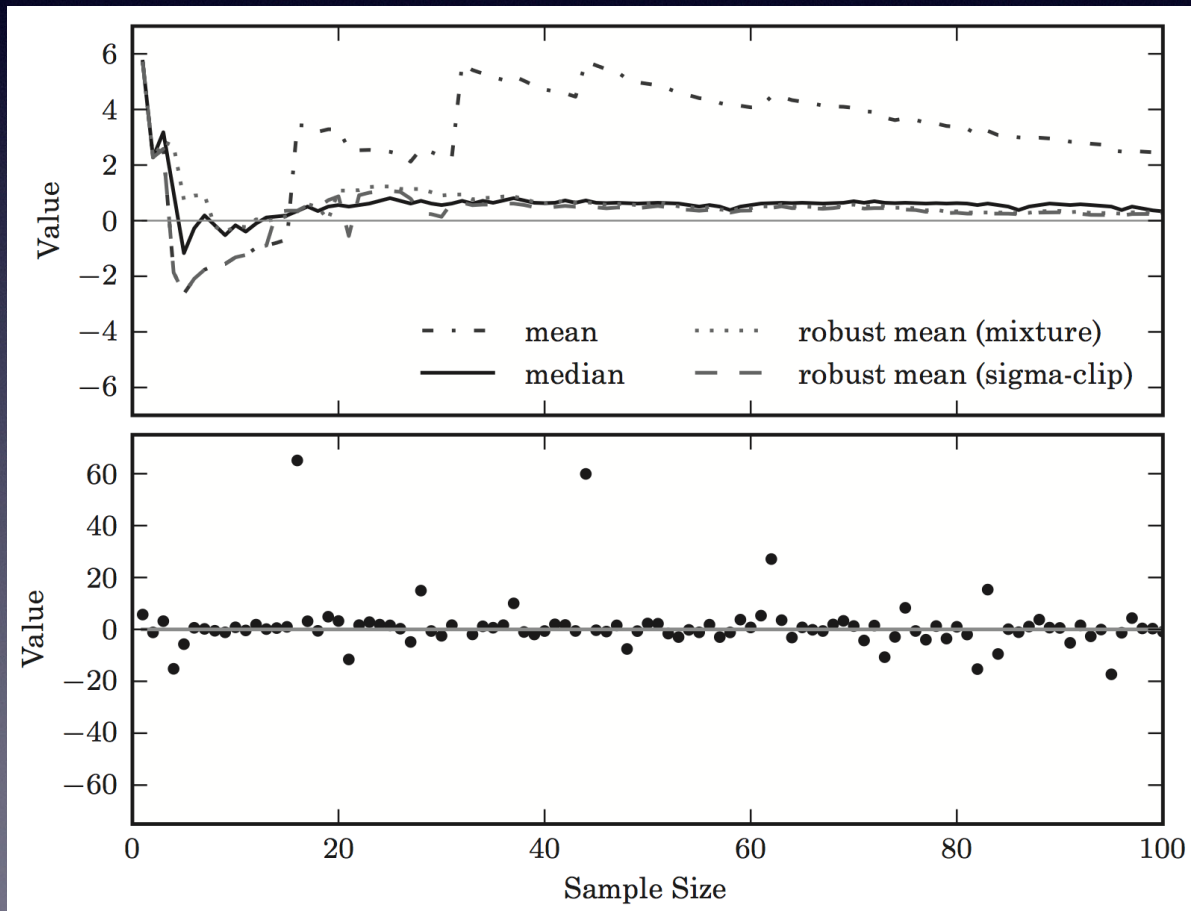
$$p(x|\mu,\gamma) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{\gamma^2+(x-\mu)^2}\right)$$

Task: given measurements x_i, i=1…N, drawn from the Cauchy distribution, find the best estimate of μ, let's call it μ^0, and its uncertainty, σ_μ

In this case, using the mean value is a very bad idea!

Bayes to the rescue!

$$p(M,\boldsymbol{\theta}|D,I) = \frac{p(D|M,\boldsymbol{\theta},I)\,p(M,\boldsymbol{\theta}|I)}{p(D|I)}$$

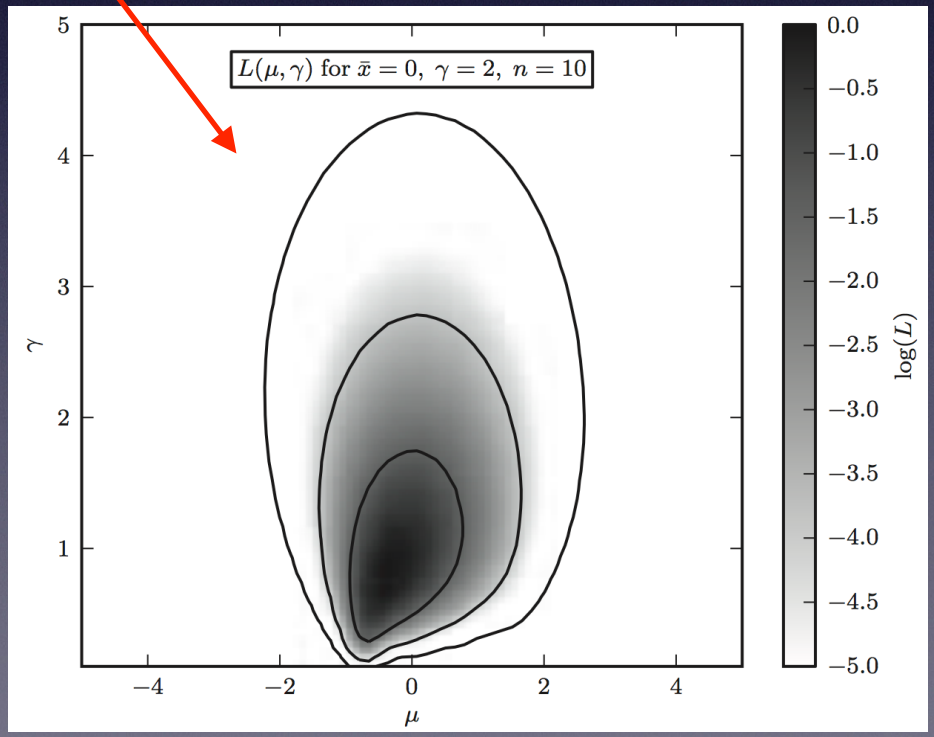


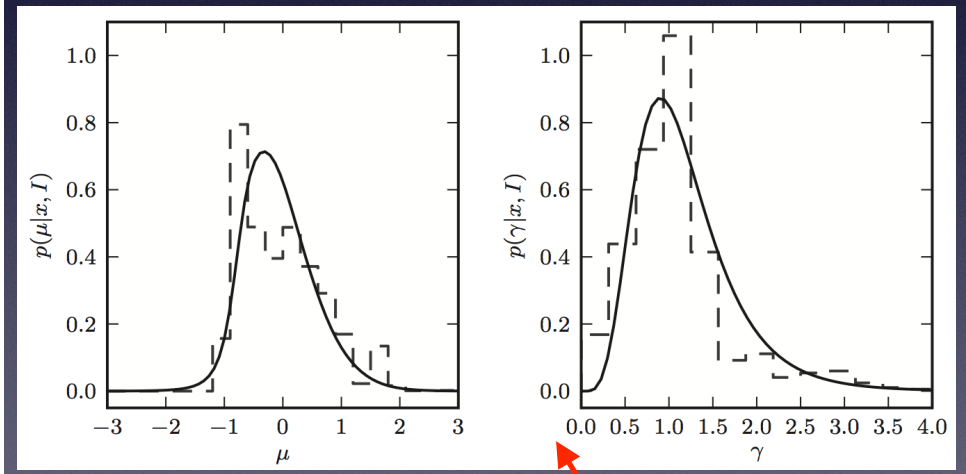
$$p(x|\mu,\gamma) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{\gamma^2+(x-\mu)^2}\right)$$

$$p(\{x_i\}|\mu,\gamma,I) = \prod_{i=1}^{N}\frac{1}{\pi}\left(\frac{\gamma}{\gamma^2+(x_i-\mu)^2}\right)$$

$$L_p \equiv \ln\left[p(\mu,\gamma|\{x_i\},I)\right] = \text{constant} + (N-1)\ln\gamma - \sum_{i=1}^{N}\ln\left[\gamma^2+(x_i-\mu)^2\right]$$

Bayes to the rescue!





Solid: exact (Lp)
Dashed: approximate (robust)

• Robust statistics

Task: given measurements x_i, i=1…N, drawn from the Cauchy distribution, find the best estimate of μ, let's call it μ^0,and its uncertainty, σ_μ

Use the median value of x_i as an estimate of location, μ^0
The scale parameter ("width") can be estimated from the interquartile range (note: q_{50} is the median):

$$\sigma_G = 0.7413\,(q_{75} - q_{25})$$

In the case of Gaussian, σ_G is equal to standard deviation (σ)

The uncertainty of μ^0 (i.e. of the median) can be estimated from

$$\sigma_\mu = \sqrt{\frac{\pi}{2N}}\,\sigma_G$$

- **Robust statistics**

Median and σ_G are good estimators of location and scale parameters also in cases when outliers are present (e.g. “real data”)

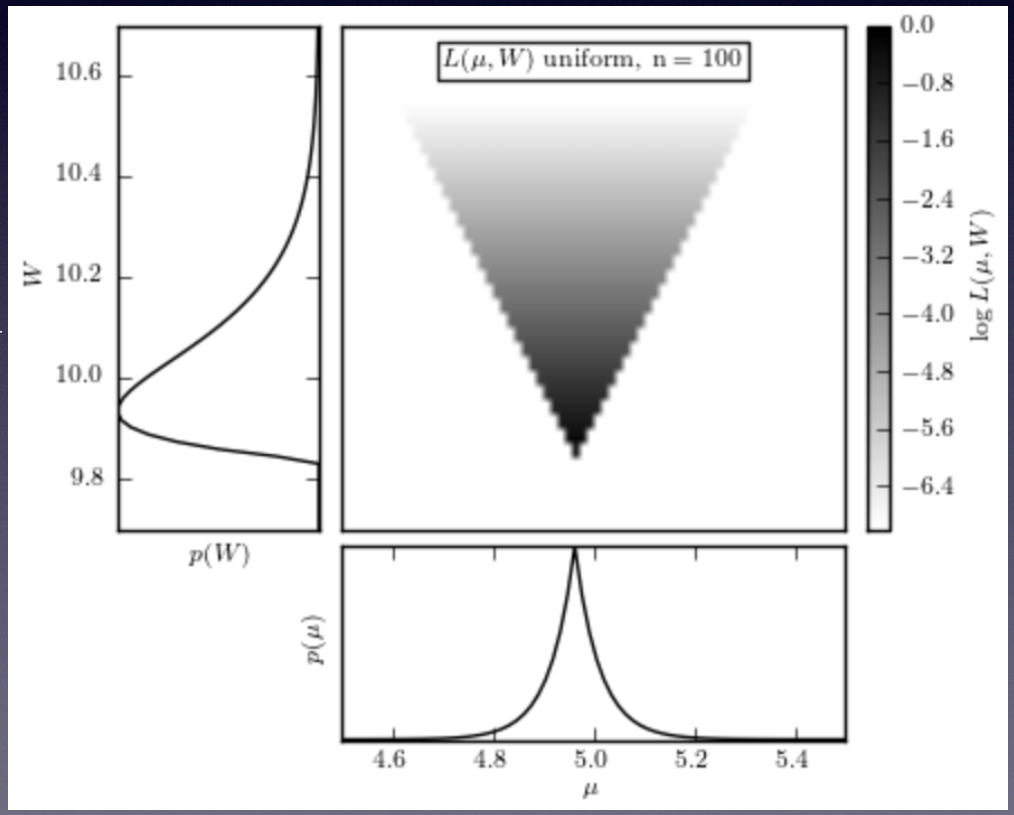
The price we pay for using the median instead of the mean is 25% larger uncertainty for the former than for the latter (assuming nearly Gaussian distributions). This is often good price to pay to avoid catastrophic failures!

Vignette #2

$$p(x_i|\mu, W) = \frac{1}{W} \text{ for } |x_i - \mu| \leq \frac{W}{2}$$

Task: given measurements x_i, i=1…N, drawn from the Uniform distribution, find the best estimate of μ, let's call it μ^0, and its uncertainty, σ_μ

In this case, using the mean value results in much larger uncertainty than proper Bayesian treatment!



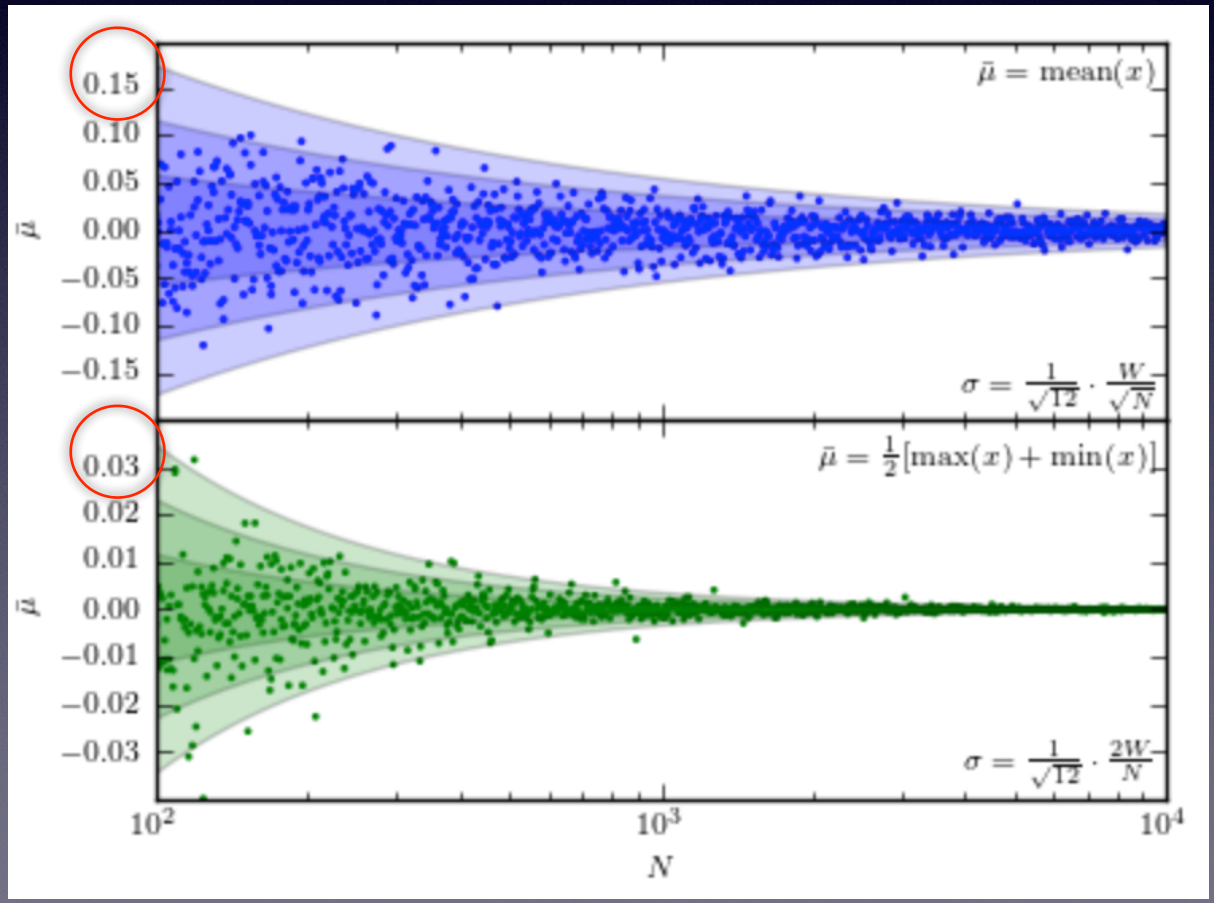
$$p(x_i|\mu, W) = \frac{1}{W} \quad \text{for } |x_i - \mu| \leq \frac{W}{2}$$

Task: given measurements x_i, i=1…N, drawn from the Uniform distribution, find the best estimate of μ, let's call it μ^0, and its uncertainty, σ_μ

Bayesian treatment yields illuminating analytic results:

$$\tilde{\mu} = \frac{\min(x_i) + \max(x_i)}{2}$$

Uncertainty of this estimate decreases as 1/N !!!



$$p(x_i|\mu, W) = \frac{1}{W} \text{ for } |x_i - \mu| \leq \frac{W}{2}$$

Task: given measurements x_i, i=1…N, drawn from the Uniform distribution, find the best estimate of μ, let's call it μ^0, and its uncertainty, σ_μ

$$\tilde{\mu} = \frac{\min(x_i) + \max(x_i)}{2}$$

By considering the distribution of extreme values of x_i, it can be shown that the expectation values are $E[\min(x_i)] = (\mu - W/2 + W/N)$ and $E[\max(x_i)] = (\mu + W/2 - W/N)$. These results can be easily understood: if N values x_i are uniformly scattered within a box of width W, then the two extreme points will be on average $\sim W/N$ away from the box edges. Therefore, the width of the allowed range for μ is $R = 2W/N$, and $\tilde{\mu}$ is an unbiased estimator of μ with a standard deviation of

$$\sigma_{\tilde{\mu}} = \frac{2W}{\sqrt{12}\,N}. \qquad (3.70)$$

While the mean value of x_i is also an unbiased estimator of μ, $\tilde{\mu}$ is a much more *efficient* estimator: the ratio of the two uncertainties is $2/\sqrt{N}$ and $\tilde{\mu}$ wins for $N > 2$.

$$p(x_i|\mu, W) = \frac{1}{W} \text{ for } |x_i - \mu| \leq \frac{W}{2}$$

Task: given measurements x_i, i=1…N, drawn from the Uniform distribution, find the best estimate of μ, let's call it μ^0, and its uncertainty, σ_μ

Given that the Uniform distribution satisfies the CLT assumptions (in particular, it has finite σ), the mean will behave as the CLT says. However, the mean is not the most efficient estimator of the location parameter in this case (nor the CLT ever claimed that…). Use instead

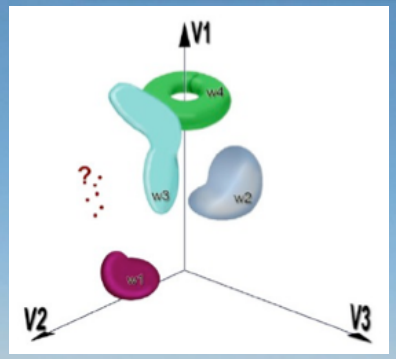
$$\tilde{\mu} = \frac{\min(x_i) + \max(x_i)}{2}$$

Or Bayes in other cases!

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

LSST data analysis, astro-statistics and astro-informatics will be closely related, and will open numerous opportunities for people with "Big Data" skills

If You Liked SDSS, You will Love LSST!



Additional slidies for an extended version

1) Interpretation of spectral energy distributions (SEDs)

- efficient and robust interpretation of time-resolved multi-band photometry for “billions and billions” of objects

Because of integration over broad bandpasses, forward modeling using a trial SED is superior to “correcting data" (fluxes, positions, sizes):

a) photo-z algorithms: observed SED depends on the redshift of an intrinsic SED (expansion of the universe, source evolution, intergalactic extinction)

b) photometric parallax for stars (will greatly benefit from Gaia parallaxes!)

c) photometric metallicity for stars (trained using spectroscopic metallicities)

d) interstellar extinction along the line of sight for stars in the Milky Way disk

e) astrometric effects due to atmosphere (point-spread-function effects, image differencing, finding quasars)

1) Interpretation of spectral energy distributions (SEDs)

- efficient and robust interpretation of time-resolved multi-band photometry for “billions and billions” of objects

Open/interesting issues:

- machine learning vs. SED templates for galaxies vs. cross-correlation of samples
- the impact of heteroscedastic noise, priors, truncated and censored data
- how much would per visit processing help (due to varying effective bandpasses)?
- posterior pdfs vs. likelihoods, optimal compression, etc
- covariances (also for pretty much everywhere else below)

2) Spatial correlations

examples:

a) large-scale distribution of galaxies (e.g. distinguish GR vs. modified gravity)

b) matched filters for dwarf galaxies and streams/tidal tails

c) LMC/SMC, Virgo overdensity, great circle streams morphologies

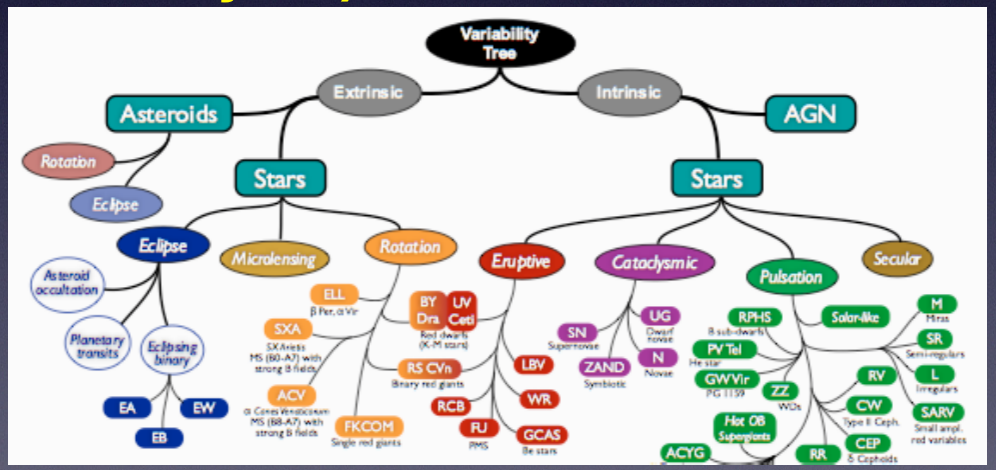
Open/interesting issues
- how to scale up to billions of points across the sky?
- can this be done directly in db?

3) Moving objects

- cadence optimization: do we really need two visits per night? Or perhaps as many as four!
- how robust and efficient would be full Bayesian approach?
- how hard is it to do shift-and-coadd for KBOs and more distant objects on scale of LSST?

4) Variable objects (time series analysis)

a) regular vs. irregular
b) short vs. long timescales
c) robust detection
d) classification



Eyer & Mowlavi 2007

Open/interesting issues:

- machine learning vs. light curve templates
- heteroscedastic noise, priors, truncated/censored data
- can this be done directly in db?

5) Systematic measurement uncertainties

e.g. across the sky, vs. flux, seeing, sky brightness, etc.

a) astrometry
b) photometry
c) galaxy shapes (cosmic shear: as small as 10^{-6})

Open/interesting issues:
- unknown SEDs!
- the impact of atmosphere (variable seeing and transmissivity, DCR, stochastic)
- multiplicative and additive errors in galaxy shear
- systematics in photo-z
- billions of objects measured a thousand times: does sqrt(N) still work in this regime?

6) Astrophysical simulations and astrophysical systematics

e.g.

- growth of cosmic structure
- formation and evolution of galaxies
- formation and evolution of Solar System
- the ISM in the Milky Way

Open/interesting issues:

- is GR correct?
- what are feedback mechanisms in galaxy formation?
- nonlinear galaxy bias
- intrinsic alignments of galaxy shapes with the density field
- baryonic effects on dark matter halo profiles
- gravity vs. hydro simulations issues (gas dynamics, star formation, feedback, etc)

7) LSST System Enhancements

observing strategy (e.g. angular and temporal sampling functions, dithering patterns)
other filters?
improved algorithms (for image differencing, calibration, etc)
Level 3 (a.k.a. “specialized processing”, as well as “everything we didn’t think of”)

Open/interesting issues:
- cadence optimization
- cadence evolution
- shift-n-add for arbitrary space-time trajectories
- complex galaxy models (e.g. tidal tails)
- transient classification

8) New algorithms in LSST

some mature, some not even started…

- Multifit (forward modeling on per visit basis)
- psf depends on time, position, instrument state, and color (more precisely on in-band SED shape)
- image differencing
- crowded field processing
- SED is unknown, especially for transients…
- incorporating other surveys in data processing, e.g. how to best benefit from Gaia, Euclid, WFIRST, etc.
- how much can you do in 60 seconds?

Data Release Processing:

assume 1000 cores and 8 months run time, so 2e10 core-seconds; with 20 billion objects: 1 sec/object

** What can you compute in about 1 second? **

8) New algorithms in LSST - and similar datasets

- how to automate tradeoffs between false positive rate (contamination) and the false negative rate (completeness)

- robust detection of extremely rare events and real-time intelligent event filtering for follow-up efforts

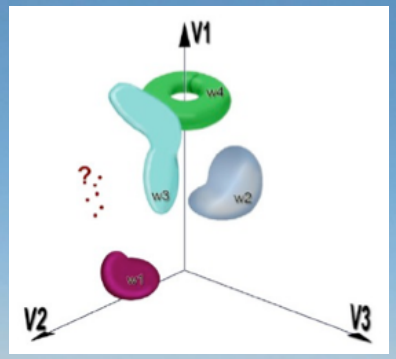
- unsupervised, semi-supervised and supervised clustering/ classification on massive datasets in real time (in db?)

- joint morphological and color-magnitude based probabilistic object classification (e.g. “galaxy”, with its type and its photo-z vs. “star” and its photometric parallax and metallicity)

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

LSST data analysis, astro-statistics and astro-informatics will be closely related, and will open numerous opportunities for people with "Big Data" skills

If You Liked SDSS, You will Love LSST!



It's going to be fun!