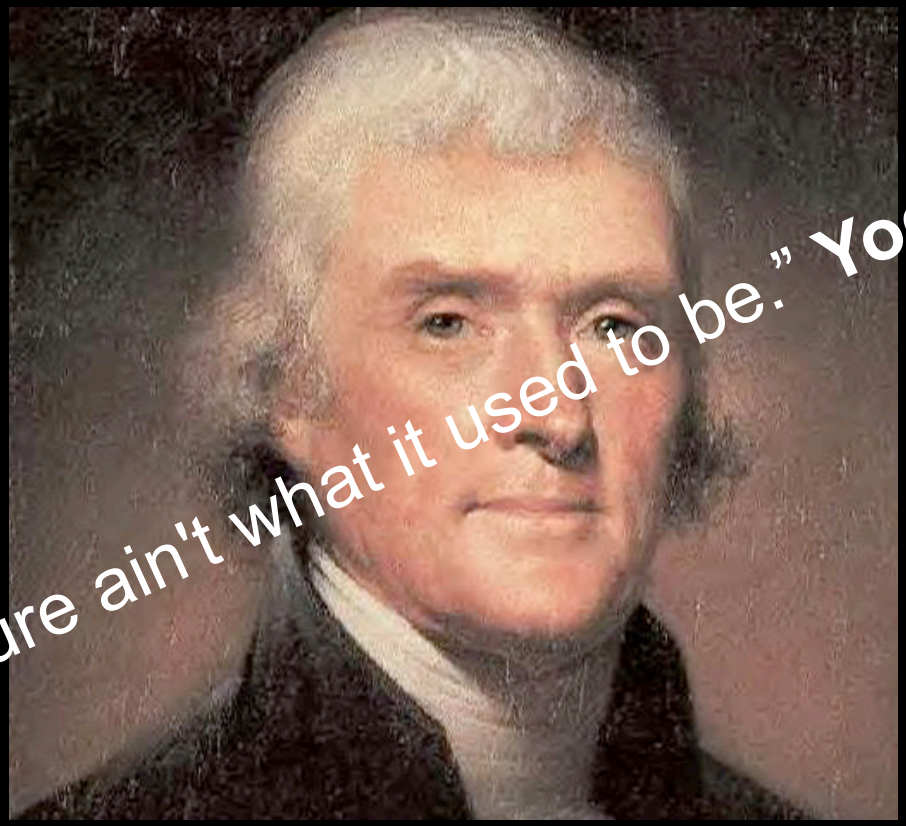
Lecture Two: The impact of a new generation of surveys



"The future ain't what it used to be." **Yogi Berra**

"I like the dreams of the future better than the history of the past." **Thomas Jefferson**

What drives survey design

- **Statistical uncertainties**
  - Shot noise: drives large n
  - Sample variance: drives large V

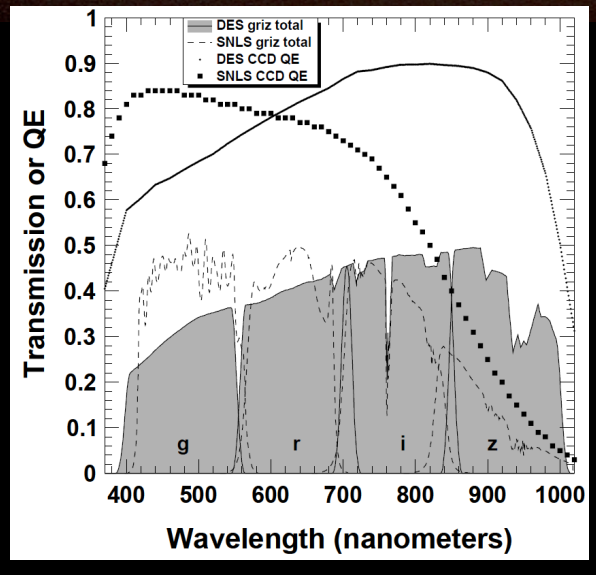
$$\frac{\Delta P(k)}{P(k)} \propto \frac{1}{\sqrt{V}}\left(1+\frac{1}{nP(k)}\right) \quad \therefore \text{ want } nP(k) > 1$$

- **Systematic errors**
  - Absolute calibration: drives offset pointings (reproducability)
  - Relative calibration: drives multiple visits (photometricity)
- **Analysis efficiency**
  - Large n: drives compute resources
  - Complexity is the real challenge
- **Theoretical considerations**
  - Simulations exceed the volumes of galaxy surveys and have all of the same computational challenges

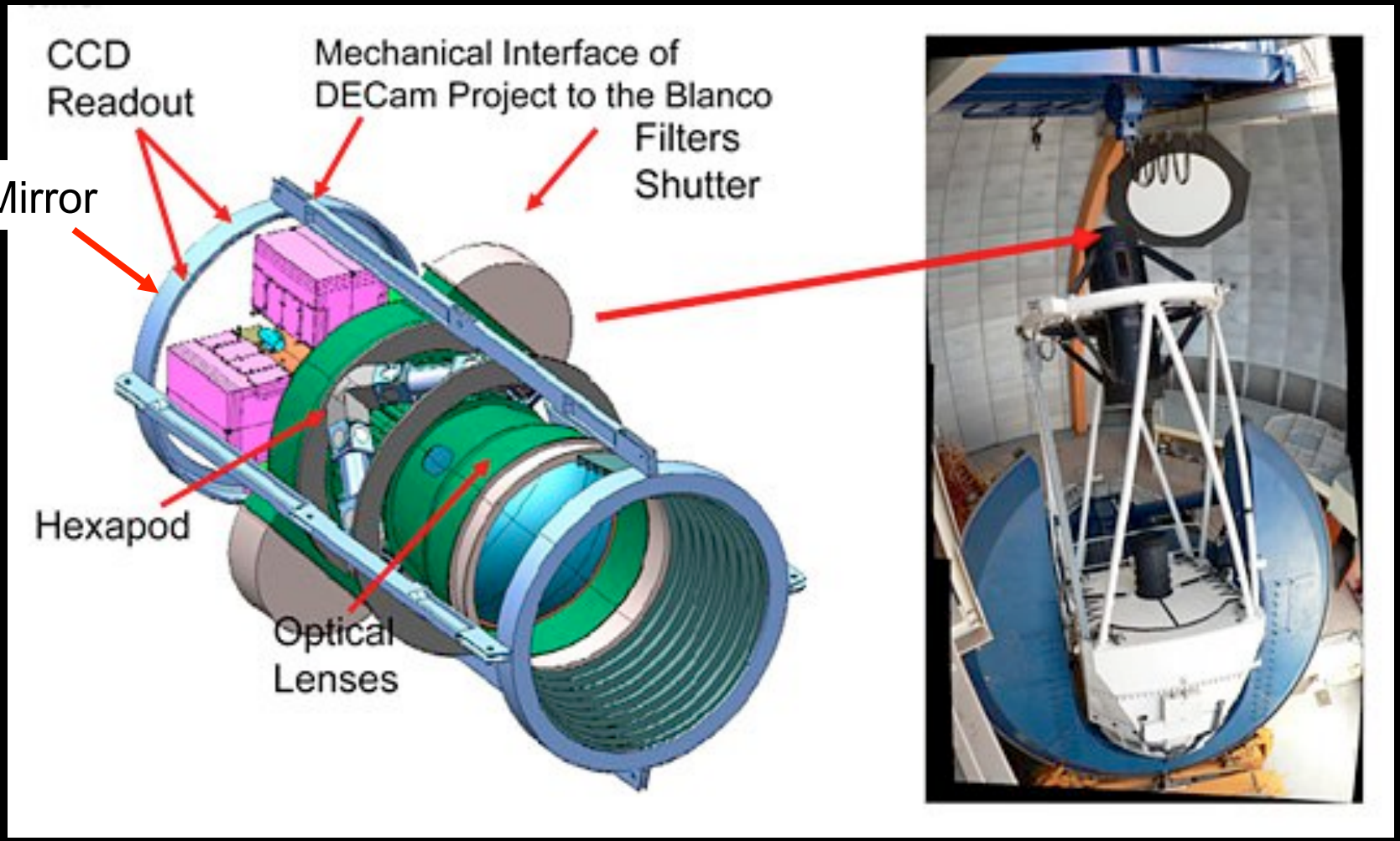
The Dark Energy Survey (2012-2017)

Blanco 4-meter at CTIO

- Wide field and deep
  - 5000 sq deg (wide)
  - 30 sq deg (deep)
  - 500+ nights (i~24th)
- Dark energy is primary science
  - Cluster counts
  - Weak lensing
  - Baryon acoustic oscillations
  - Supernovae
- CTIO Chile
  - Existing telescope (Blanco 4m)
  - 570 Mpixel camera
  - g,r,i,z,y

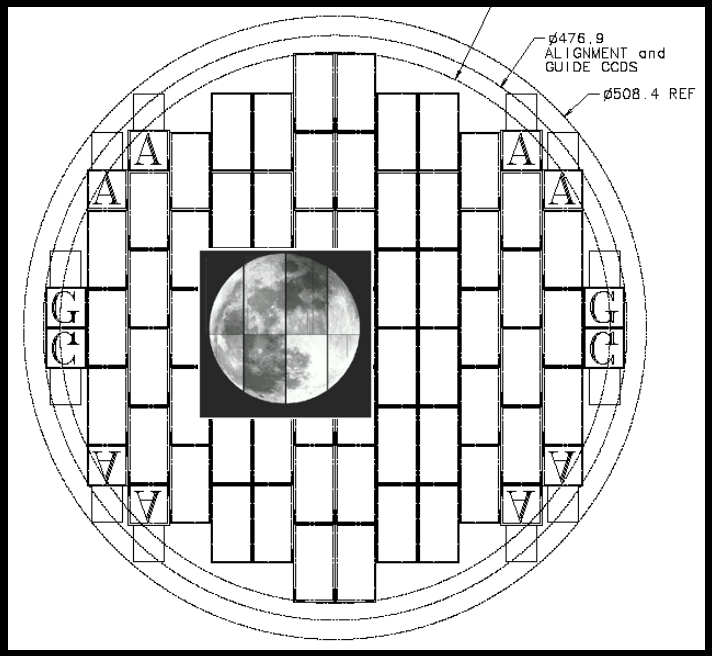


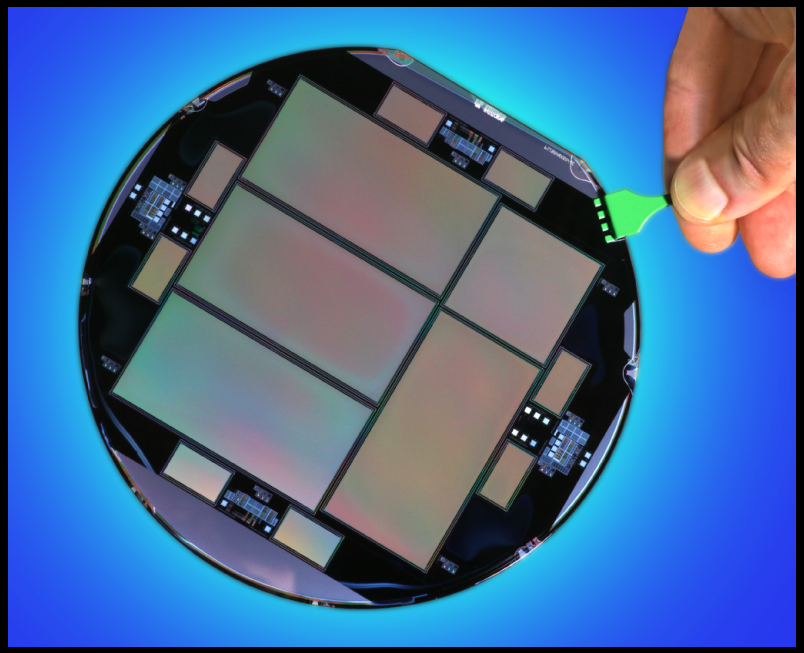
DECam: new top end on Blanco



Dark Energy Survey

Thick fully depleted CCDs





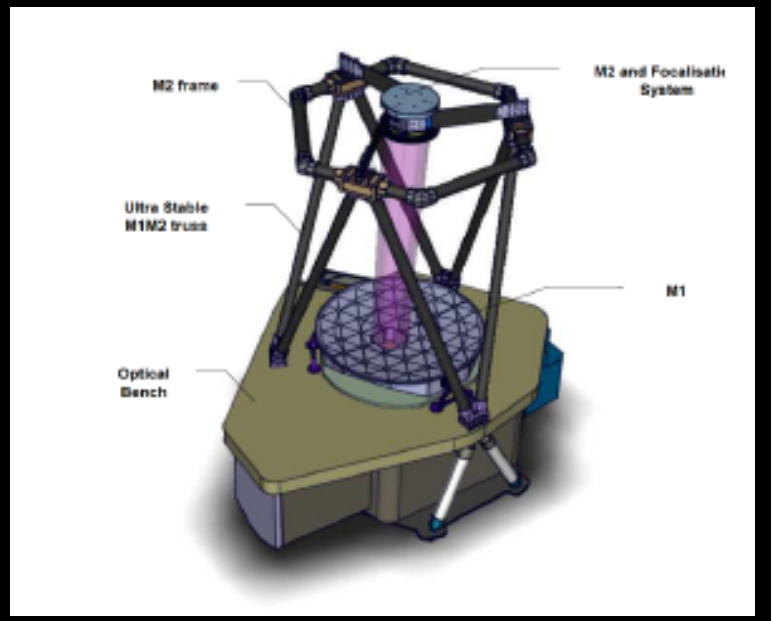
Dark Energy Survey

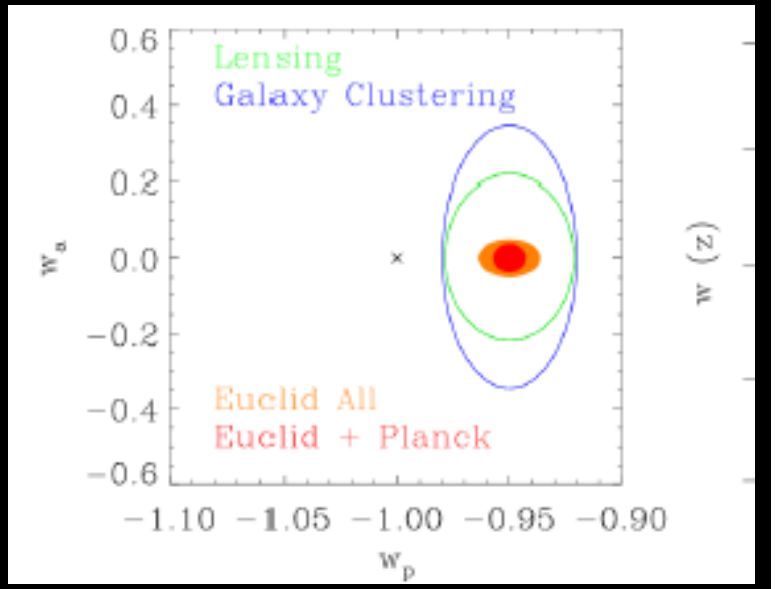
High responsivity in the red (2K x 4K devices)

62 CCDs producing a 520 Mpixel camera

EUCLID (2018-2025)

- Wide field and deep
  - 15000 sq deg (wide)
  - 40 sq deg (deep)
  - 7 years
- Dark energy is primary science
  - Weak lensing
  - Baryon acoustic oscillations
  - Supernovae
- Visible and near-ir
  - 550-2000 nm (photometry)
  - 1100-2000 nm (spectroscopy)

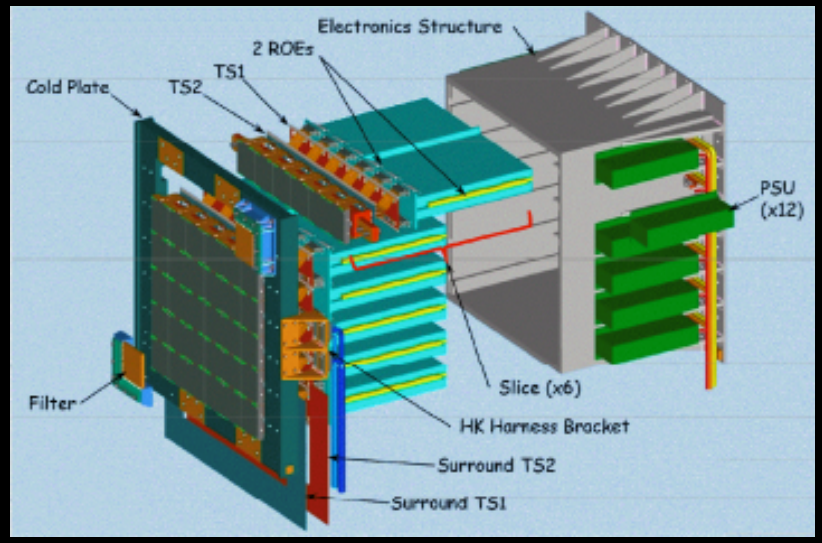


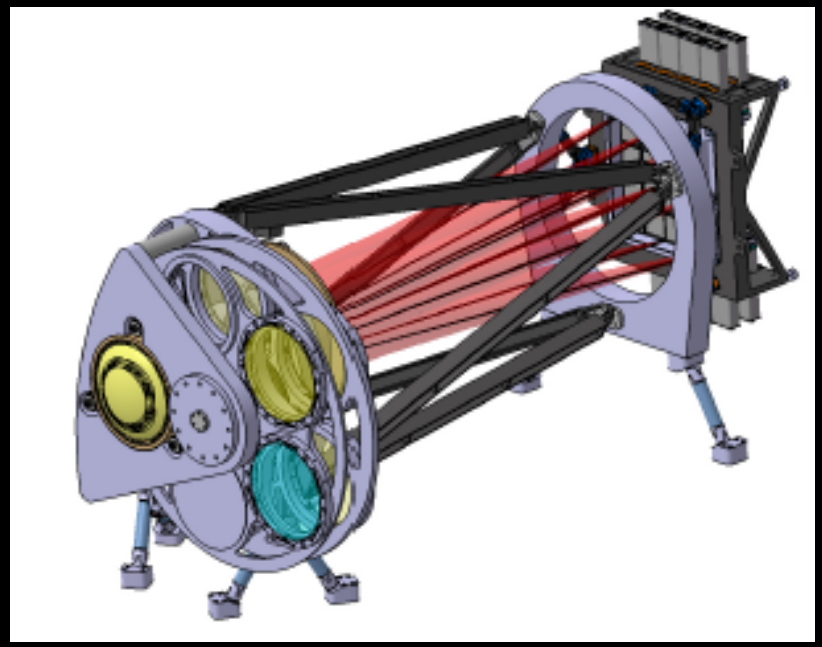


Euclid Redbook

Visible and near-ir camera

- Visible camera
  - 6x6 CCDs
  - 4Kx4K devices
- Near-infrared Camera
  - 0.55 sq degree FOV
  - 920-2000nm response (Y, J, H)
  - 16 devices (0.3 arcsec pixels)
  - 1.1-2 μm spectral range
  - 4 dithers provides coverage (in all bands) for >93% FOV

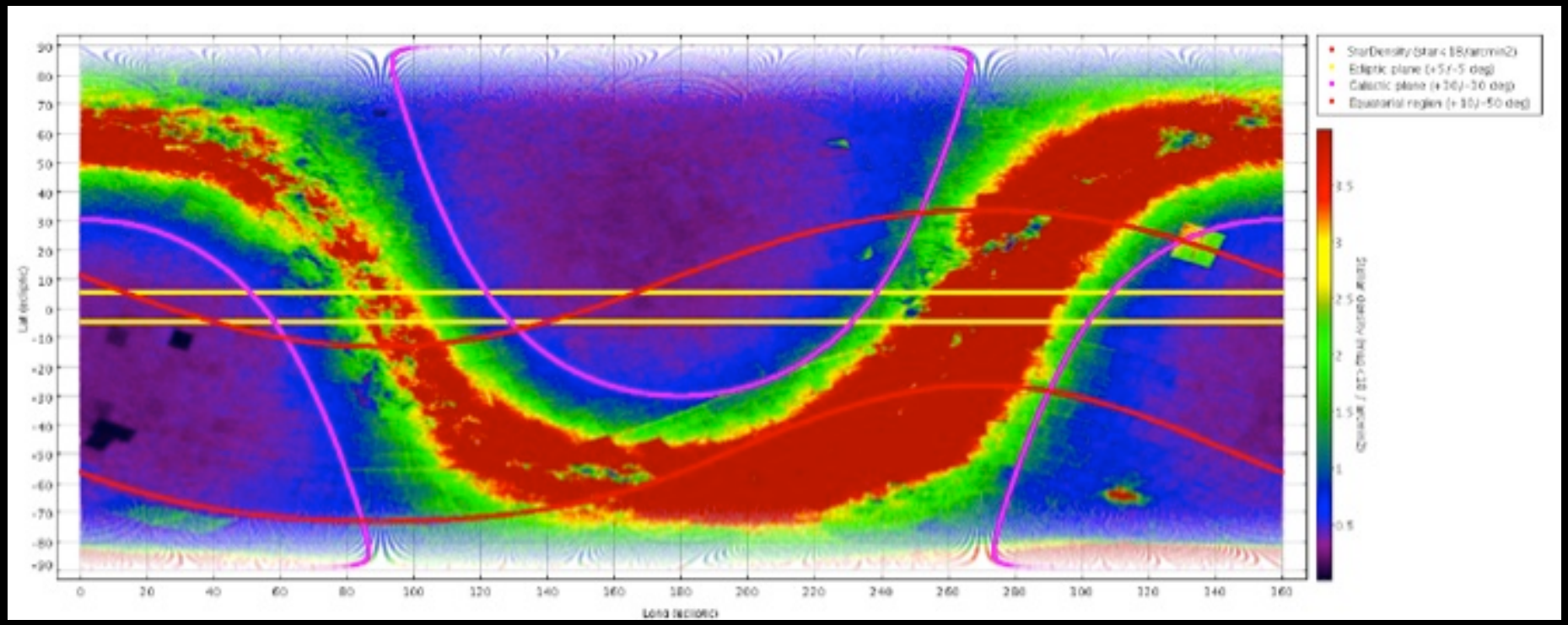




Euclid Redbook

Survey strategy

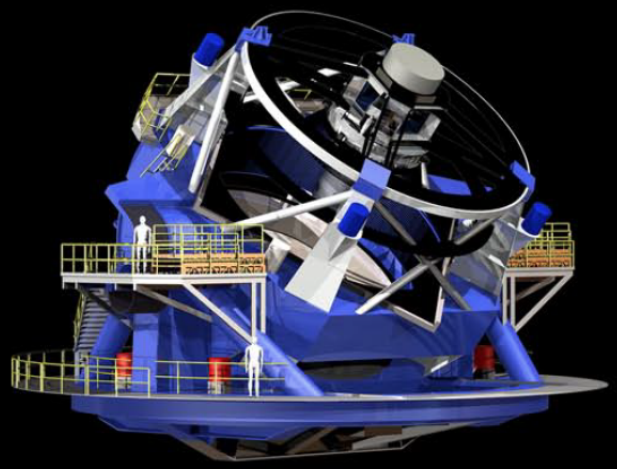
Ecliptic Coordinates

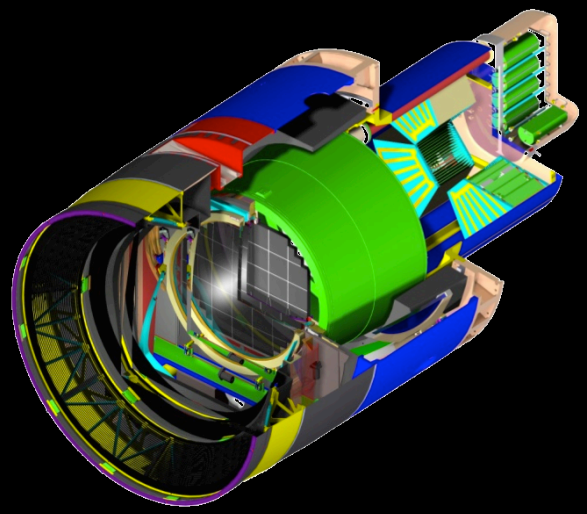


Surveys above the Galactic plane (+/- 30 degrees) from an L2 orbit, weighting high ecliptic and lower stellar density regions more. Deep field are situated at the North and South Ecliptic Poles (40 passes per field)

Large Synoptic Survey Telescope (2020-2030)

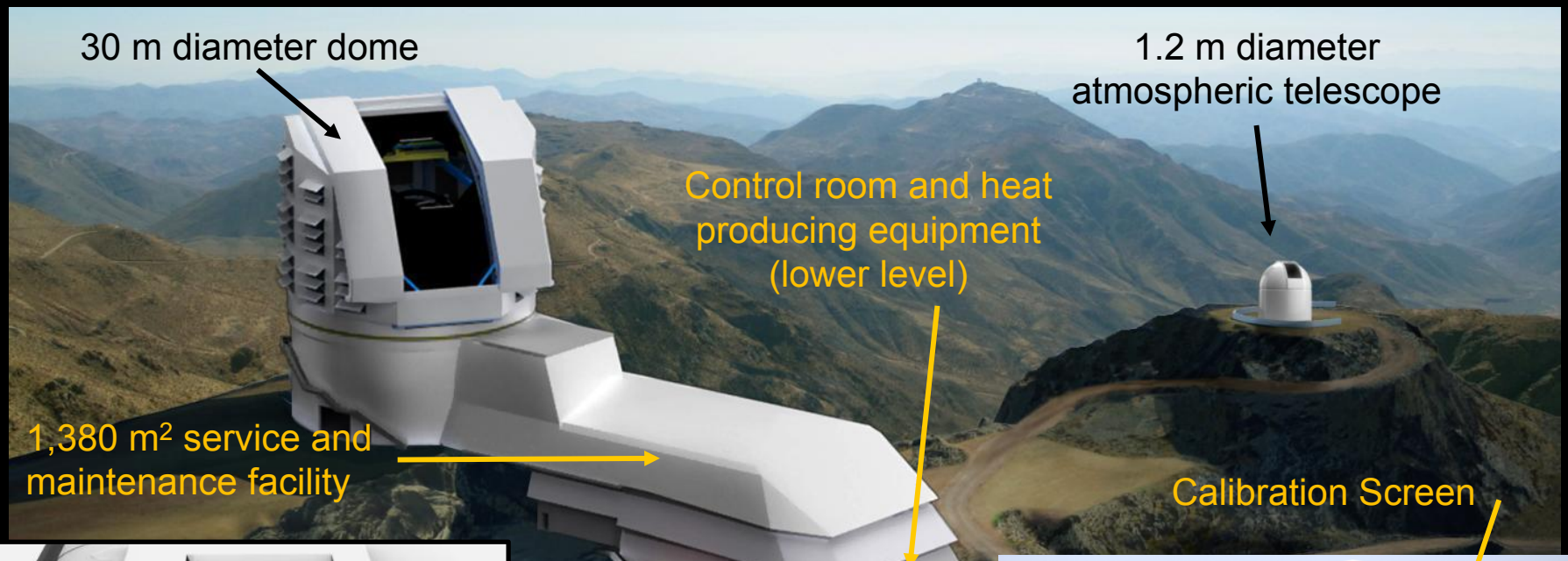
- Wide field and deep
  - 27000 sq deg (wide)
  - 100 - 200 sq deg (deep)
  - 10 years
- Broad range of science
  - Dark energy
  - Galactic structure
  - Census of the Solar system
  - Transient universe
- 3.2 Gpixel camera
  - 9.6 sq degree FOV
  - ugrizy filters

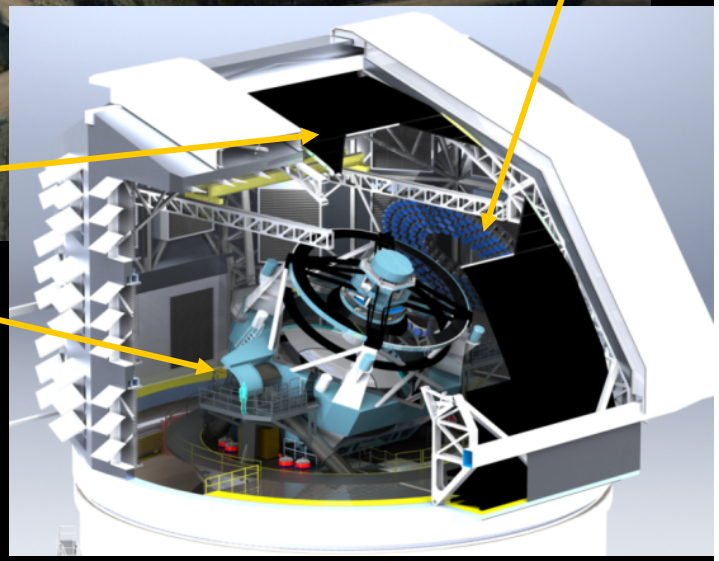




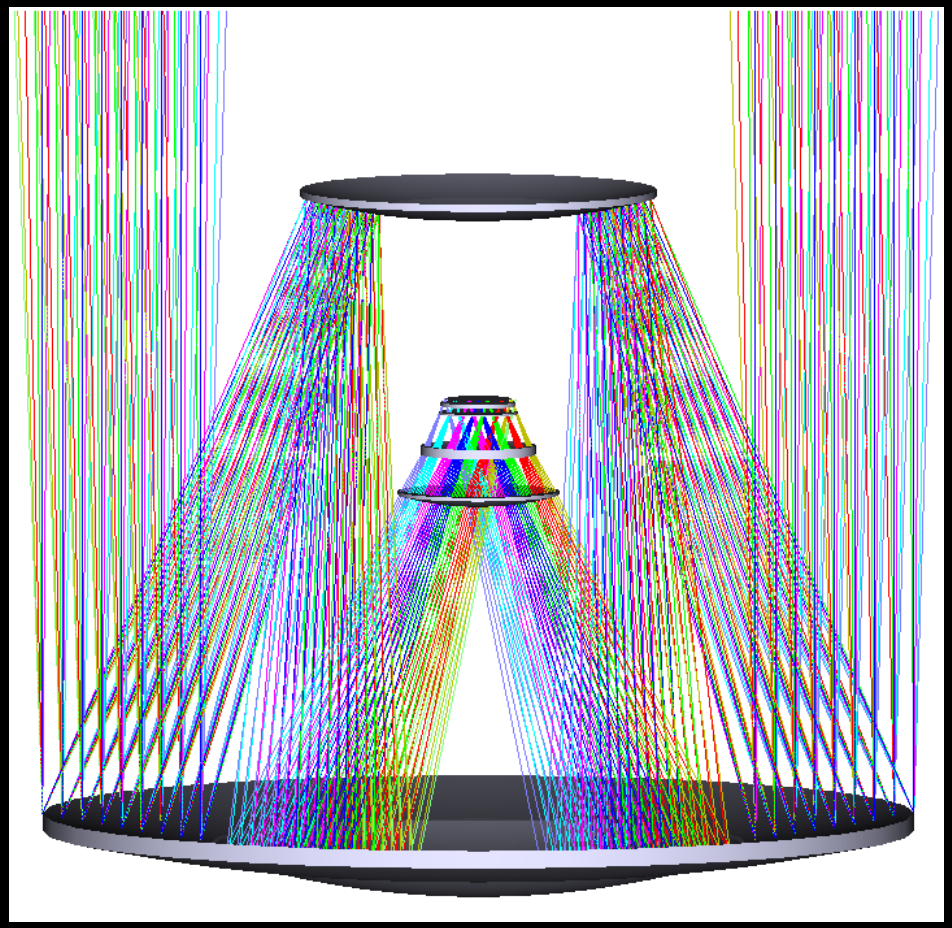
The LSST Site and Base Facilities in Chile

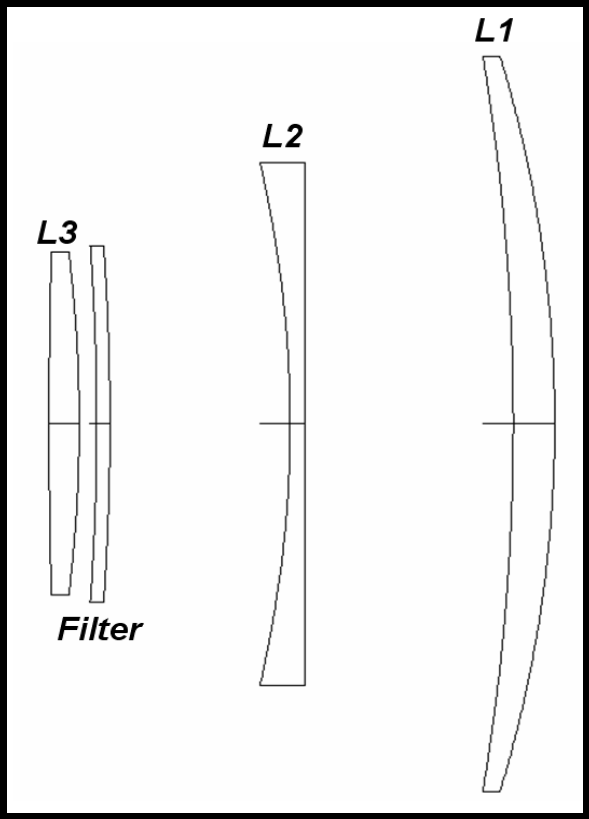
8.4m survey telescope and 1.2m atmospheric telescope





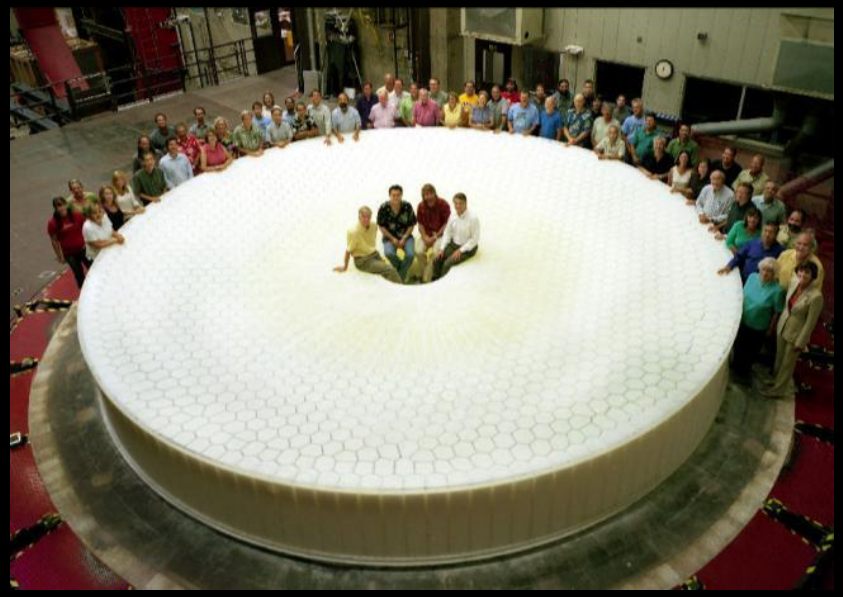
Optical design of the LSST





3 Mirror Modified Paul-Baker Design for achieving a low focal ratio over such a large field. The camera optics correct for chromatic aberration.

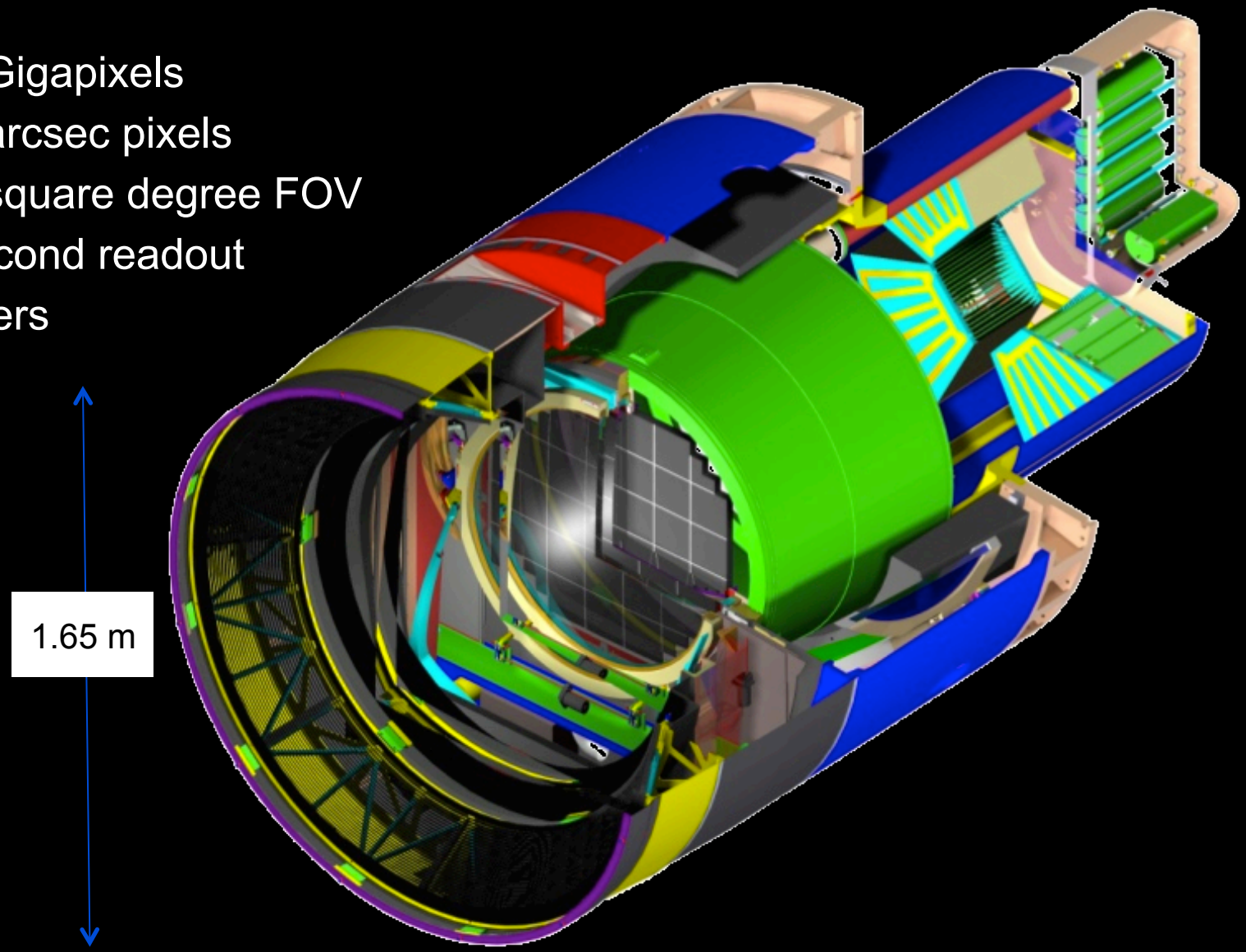
Primary/Tertiary cast from a single borosilicate blank.



- Primary-Tertiary was cast in the spring of 2008.
- Secondary fabricated by Corning in 2009.

A 3.2 Gigapixel camera

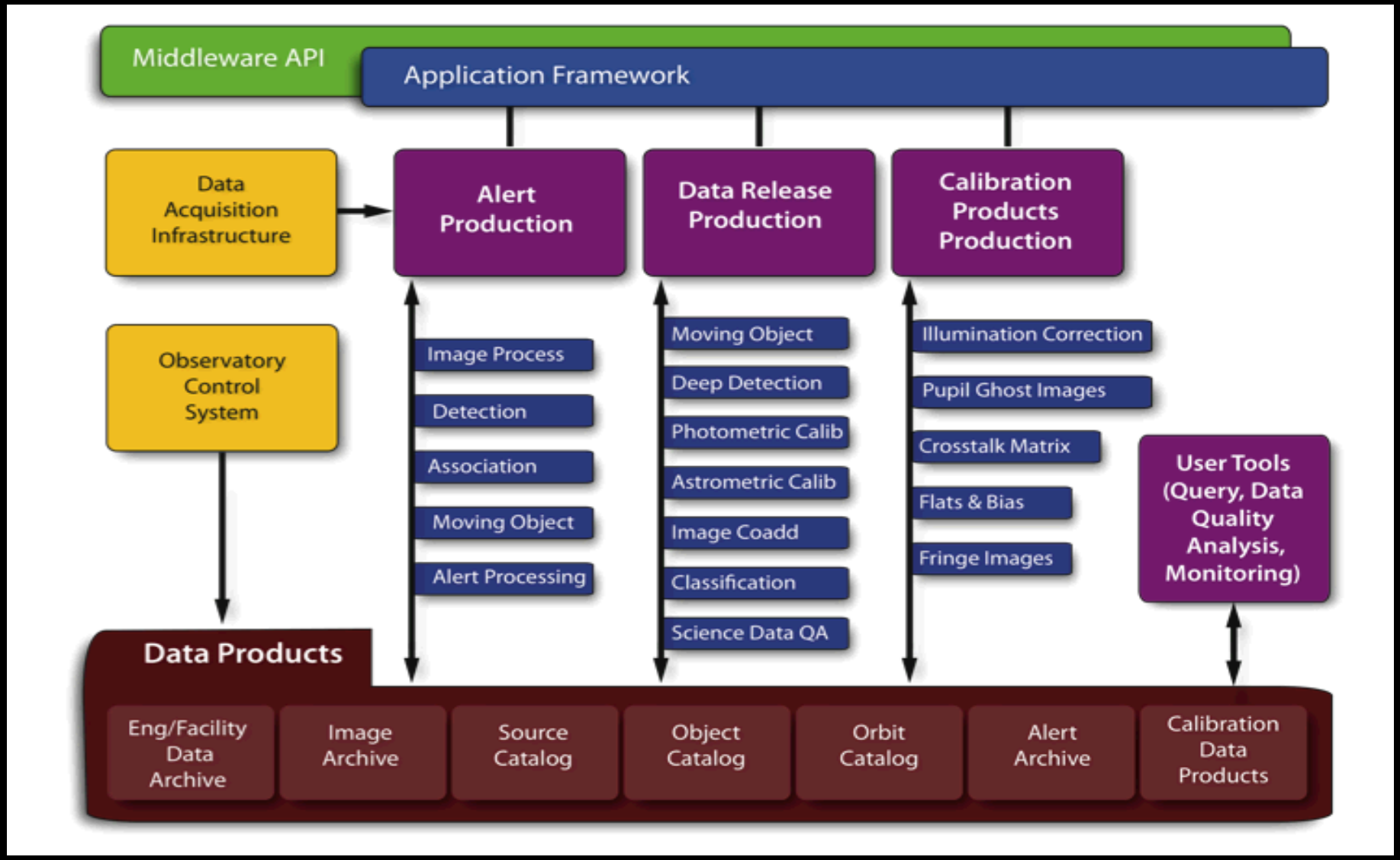
- 3.2 Gigapixels
- 0.2 arcsec pixels
- 9.6 square degree FOV
- 2 second readout
- 6 filters



Processing the data flow from the LSST

- Each "Visit" comprises a pair of back-to-back exposures
  - 2x15 sec exposure; duration = 34 seconds with readout
- The data volume associated with this cadence is unprecedented
  - one 6-gigabyte image every 17 seconds
  - 15 terabytes of raw scientific image data / night
  - 100-petabyte final image data archive
  - 20-petabyte final database catalog
  - 2 million real time events per night every night for 10 years

LSST data processing pipelines are designed, prototyped and tested in data challenges



Computational Challenges

- Indexing of massive data streams
  - 200TB today 1PB a year from now 100PB in 10 years
  - Extract small subsets of data (typical use case)
- What happens when we cant store the data?
  - Square Kilometer Array (400 GB/s to 2 PB/s)
- Architectures increasingly CPU-heavy, IO-poor
  - Designed for high performance simulations
  - Fast sequential read performance is critical!!!
- Data are distributed globally
  - Heterogeneous sets of archives
- Moving 100TB of data
  - 1Gbps 10 days
  - 10 Gbps 1 day (but need to share backbone)
  - Move the analysis to the data

What Balance Do We Need: Amdahl's Laws

Gene Amdahl (1965): one of the fathers of the mainframe

Laws for a balanced system

i. Parallelism: max speedup is S/(S+P)

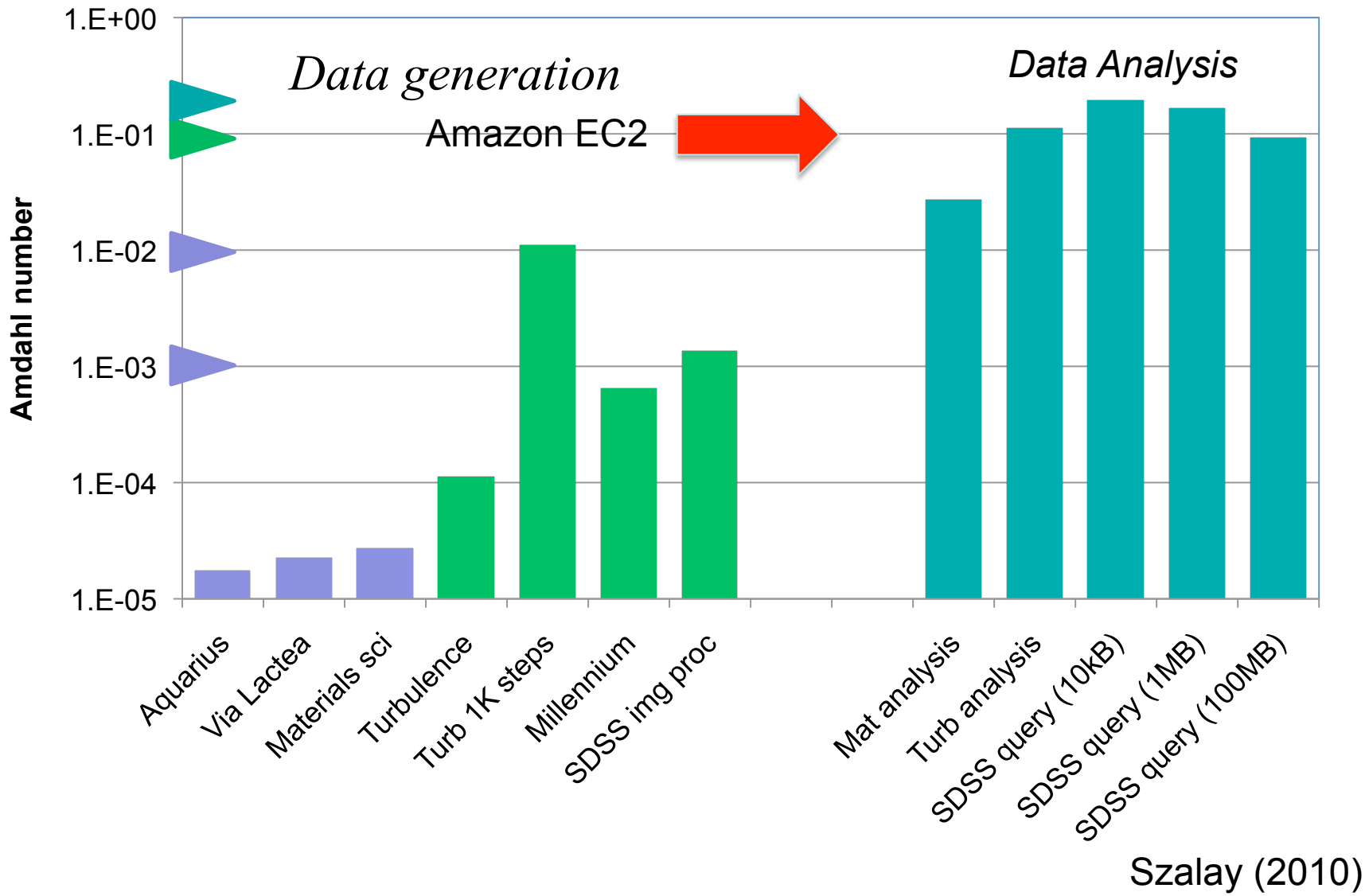
ii. One bit of IO/sec per instruction/sec (BW)

iii. One byte of memory per one instruction/sec (MEM)

Goal is Amdahl number ~1

Modern multi-core systems move farther away from Amdahl's Laws (Bell, Gray and Szalay 2006)

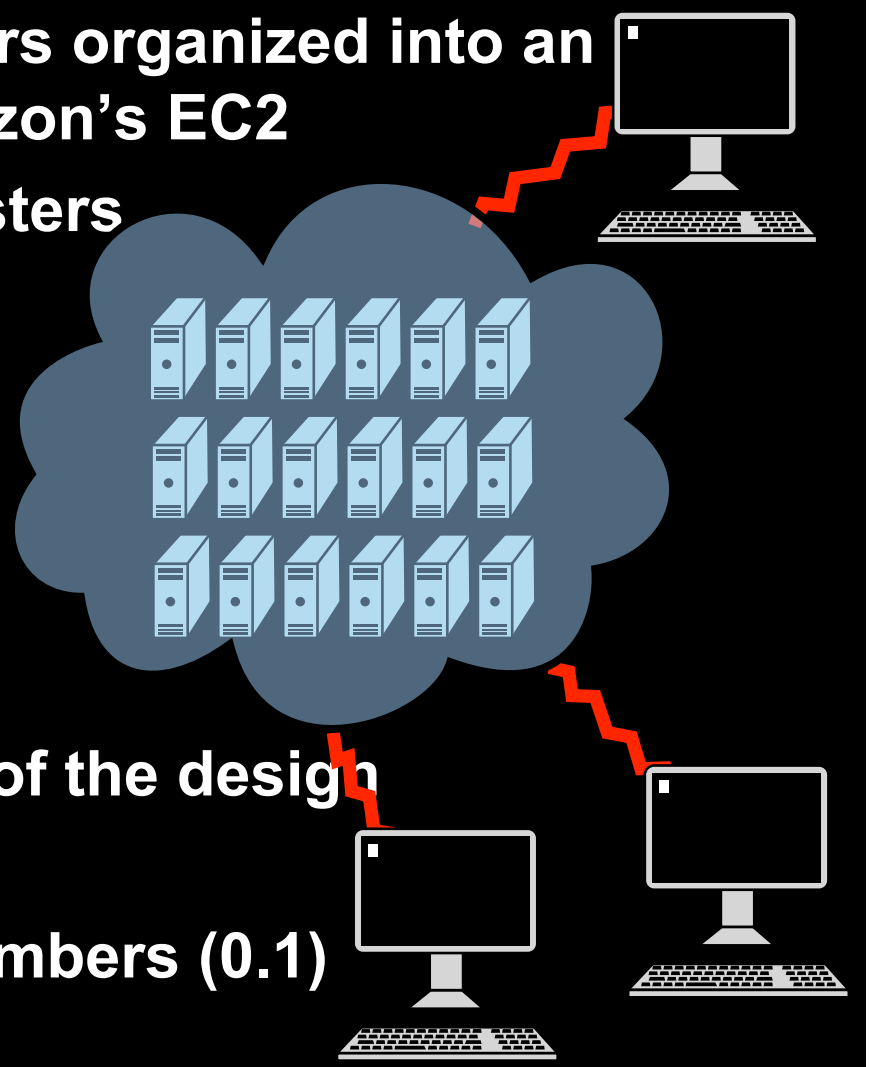
Amdahl Numbers for Data Sets



Szalay (2010)

Cloud Computing – a commercial solution

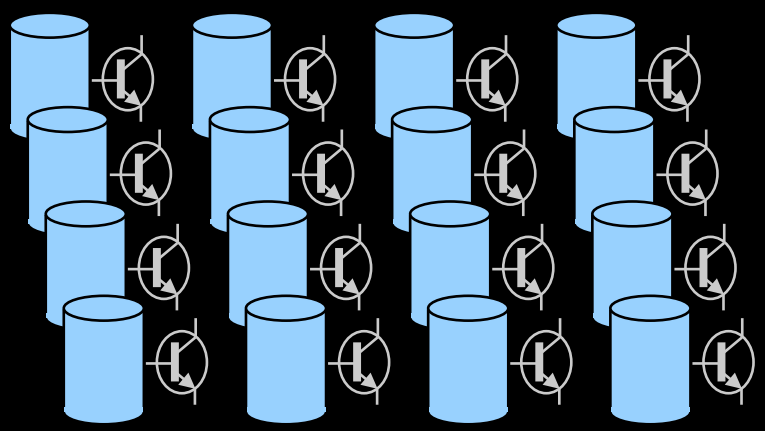
- 1000s of commodity computers organized into an on-demand cluster, e.g., Amazon's EC2
- Cheaper than specialized clusters
- Cluster is accessed from anywhere via the internet
- Networking logistics handled automatically
  - Users need very little network computing experience
- Robust to node failures; part of the design
- Nodes easily/rapidly added.
- Approaching high Amdahl numbers (0.1)



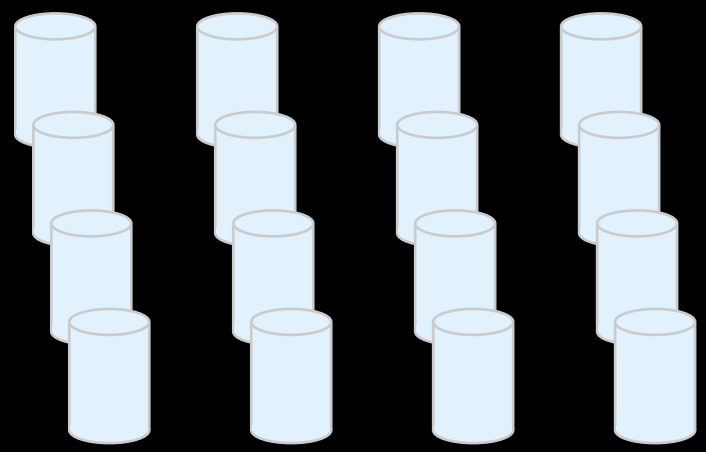
MapReduce and Hadoop

MapReduce
A massively parallel database-processing framework

In one sense:
A parallel database



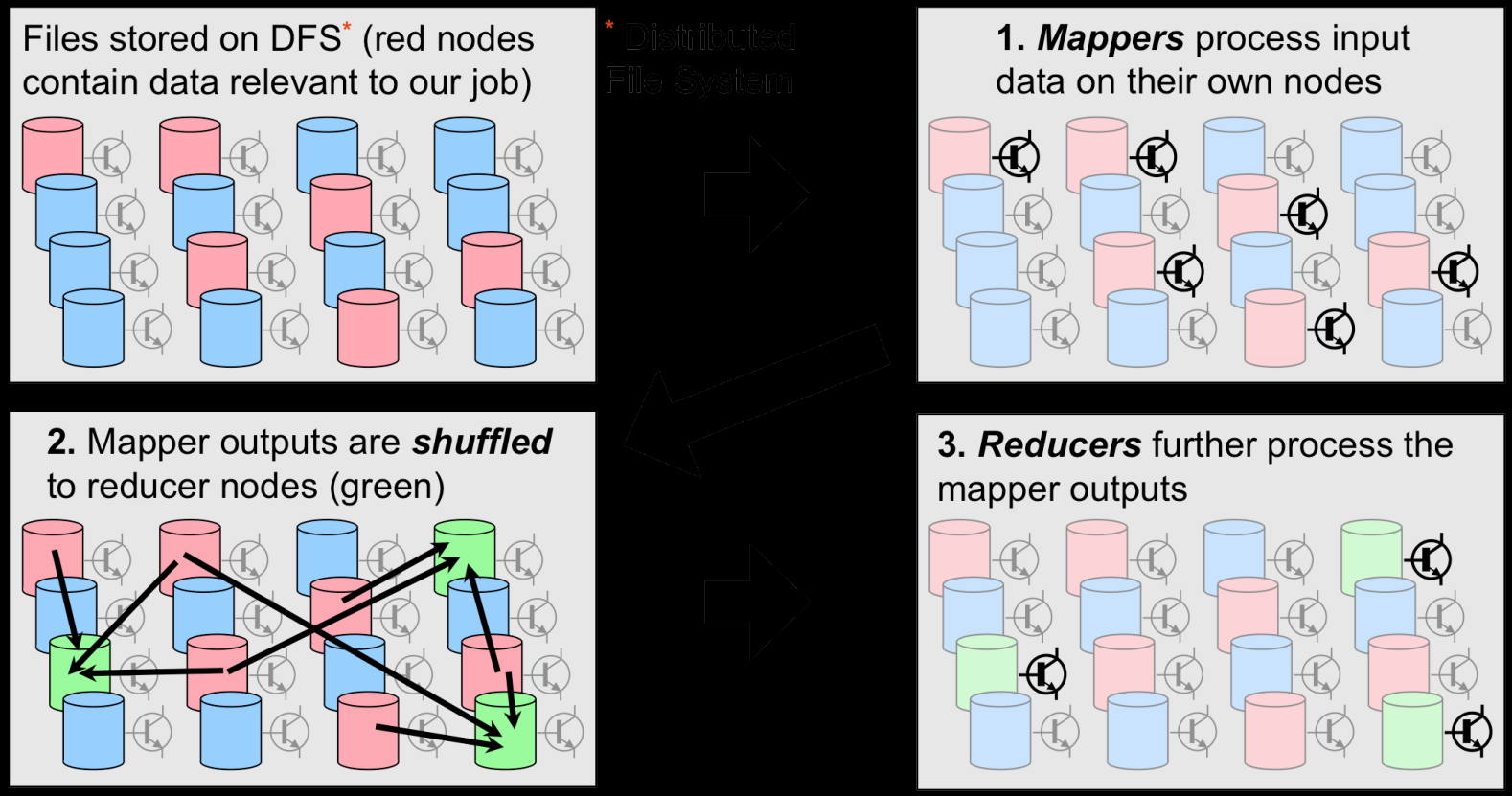
In another sense:
A parallel computing cluster



It's both!

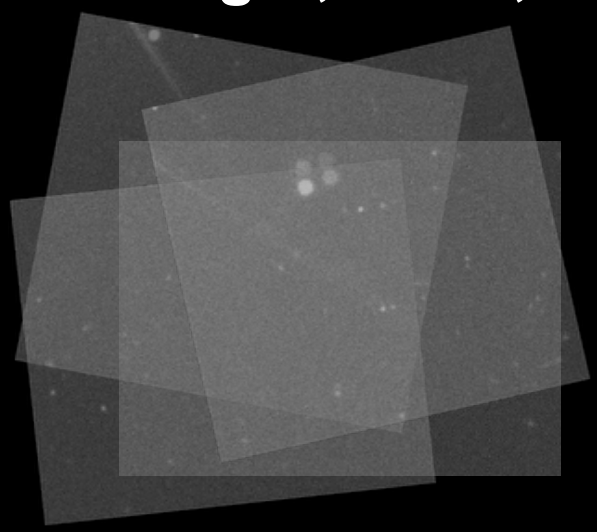
The whole process

1. *Mappers* process local data to an intermediate state.
2. Mapper outputs are *shuffled* to reducers.
3. *Reducers* further process the data, producing final output.



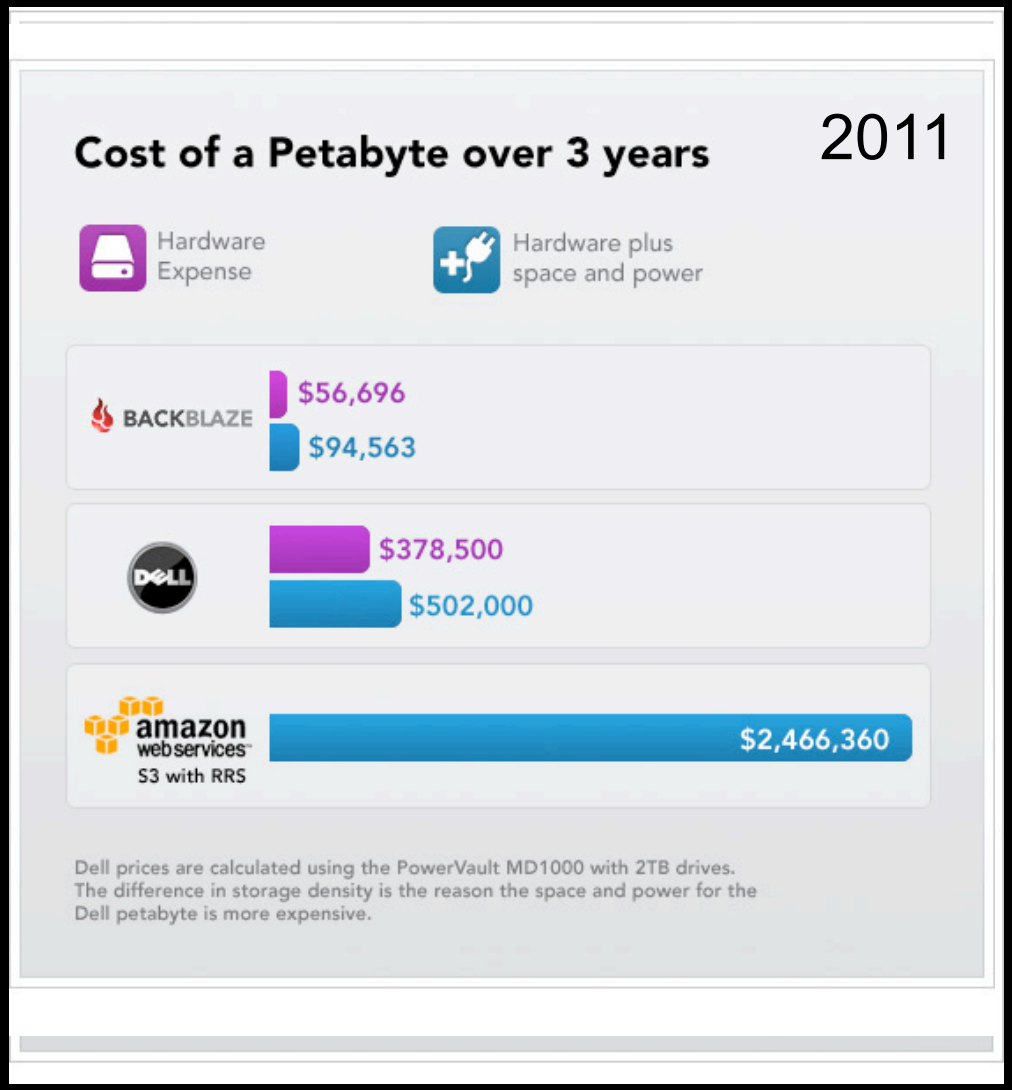
Seeding the cloud: operating on 4 million images

- Given multiple partially overlapping images and a *query* (color and sky bounds):
- Find images' intersections with the query bounds.
- Background-subtract, project coordinate system & interpolate (warp), and PSF-match intersections.
- Weight, stack, and mosaic into a final product.

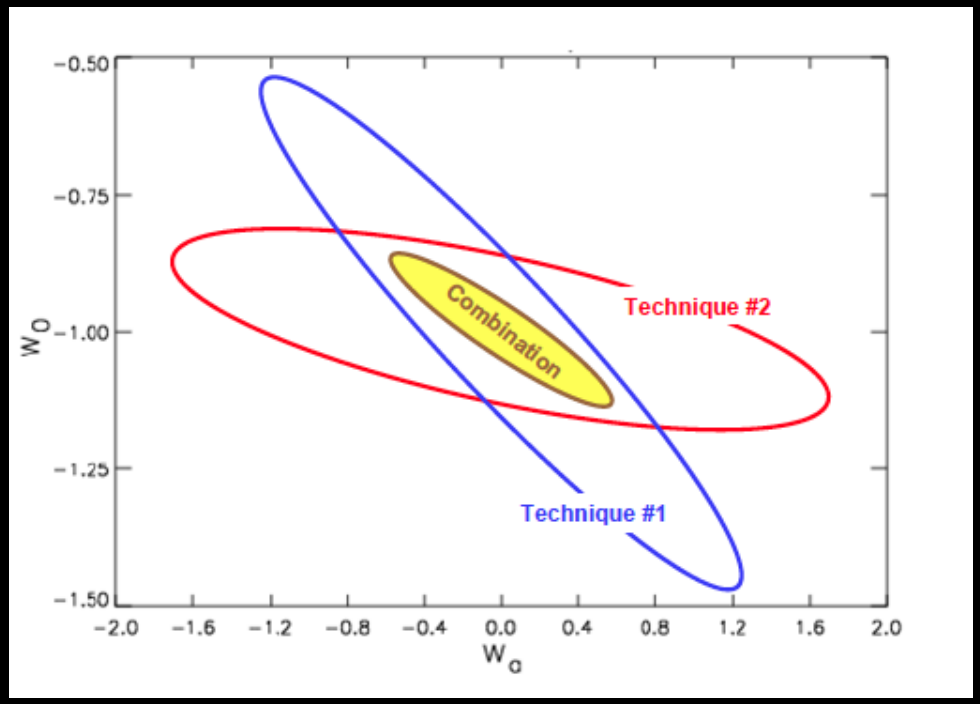


Wiley et al 2010

Why should you care?



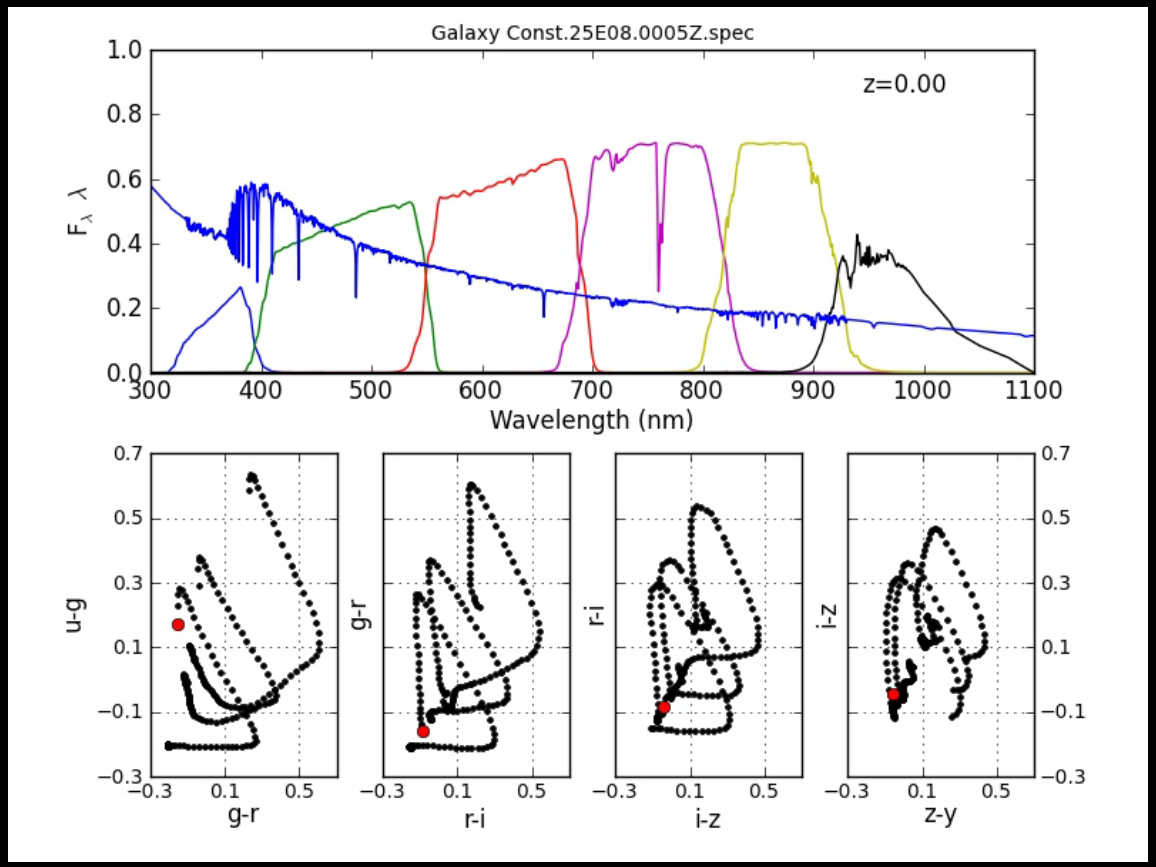
Science challenges: defining a figure of merit



$$w(a) = w_0 + w_a(1-a)$$

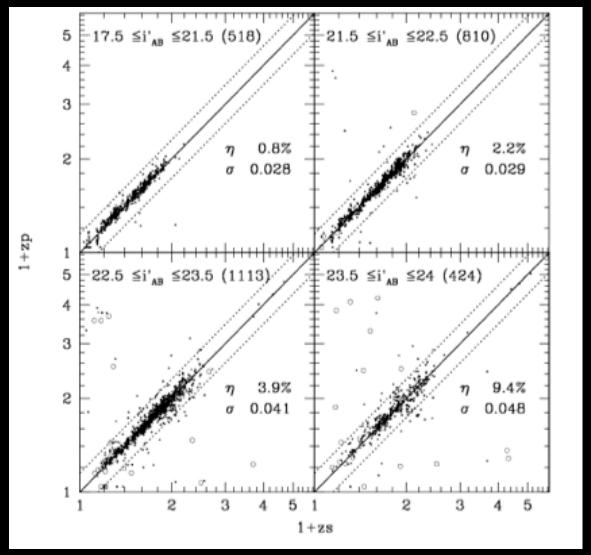
The DETF figure of merit is the reciprocal of the area of the error ellipse enclosing the 95% confidence limit in the w_0–w_a plane. Larger figure of merit indicates greater accuracy.

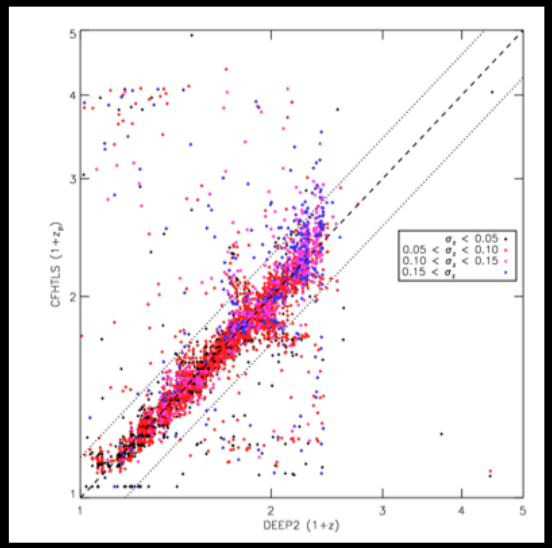
Photometric redshifts



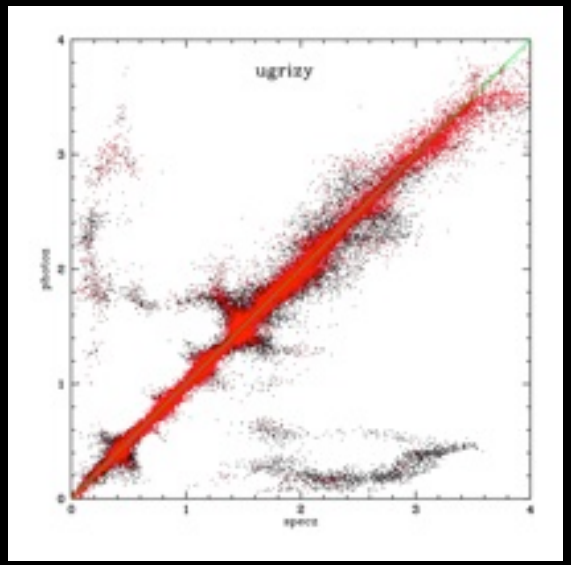
Current limits of redshift estimation

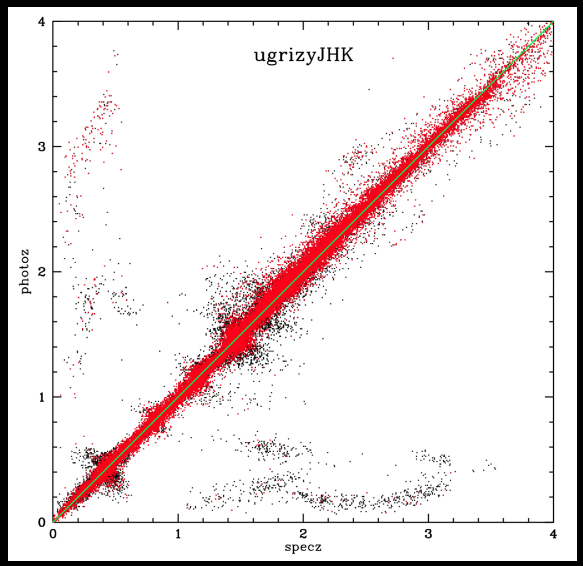
Ilbert et al



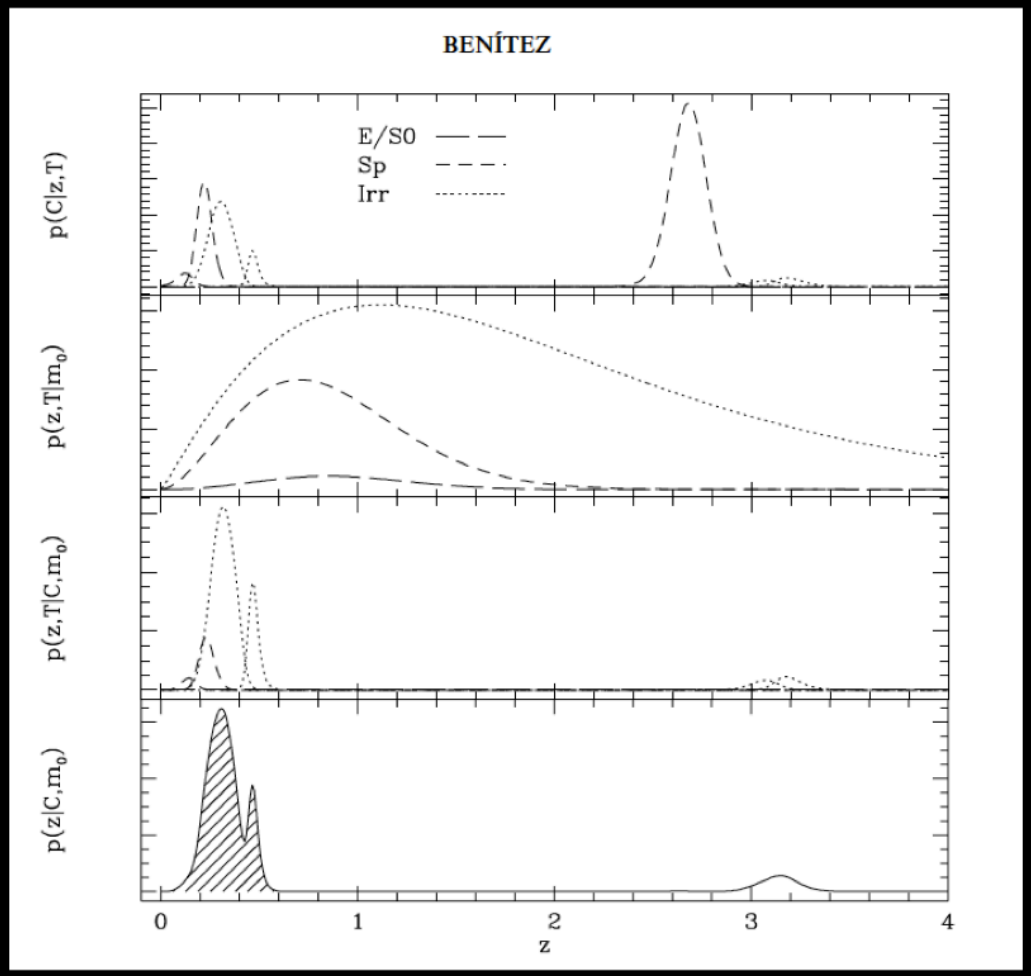


Newman et al





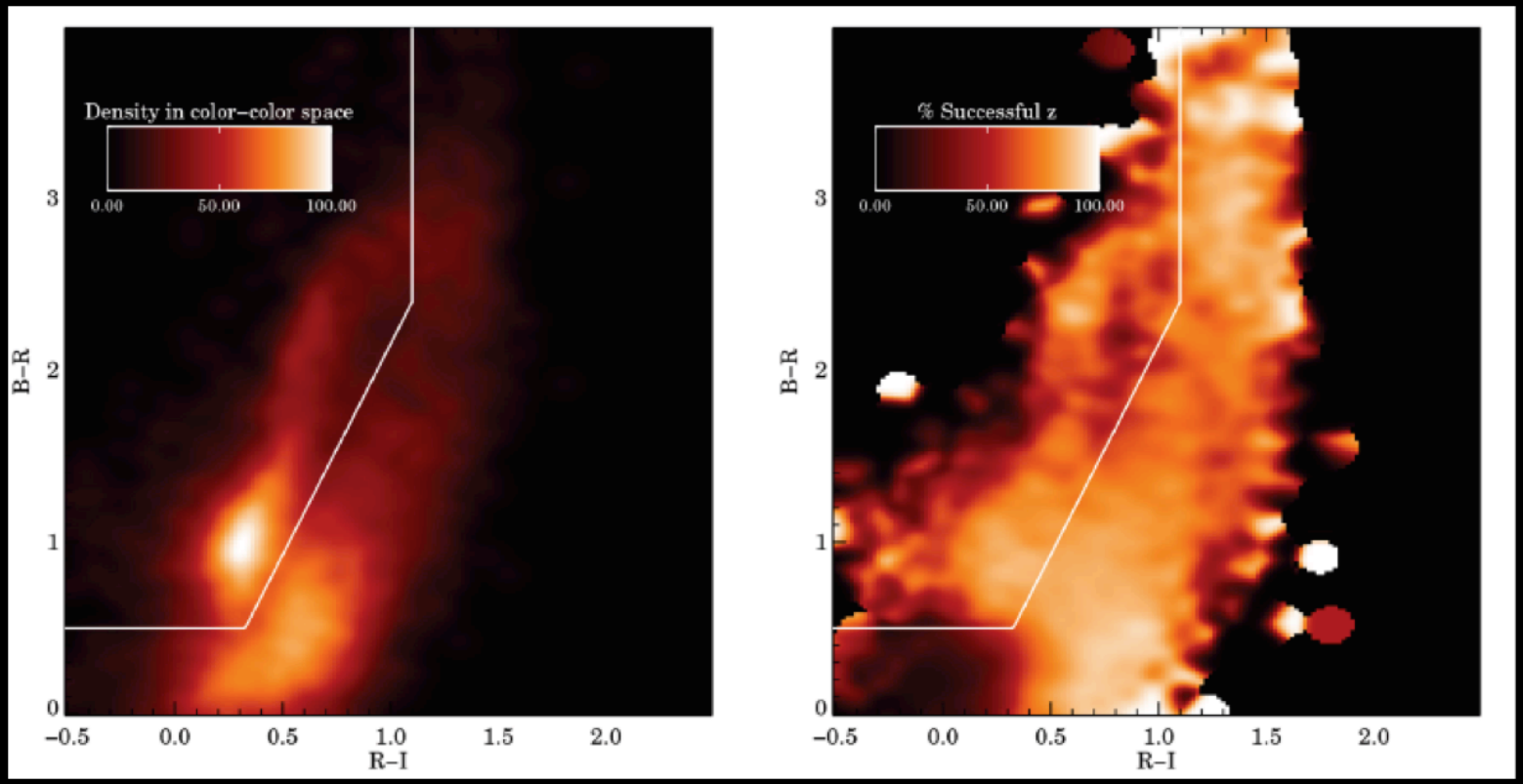
P(z): photometric redshifts



Open questions for photometric redshifts

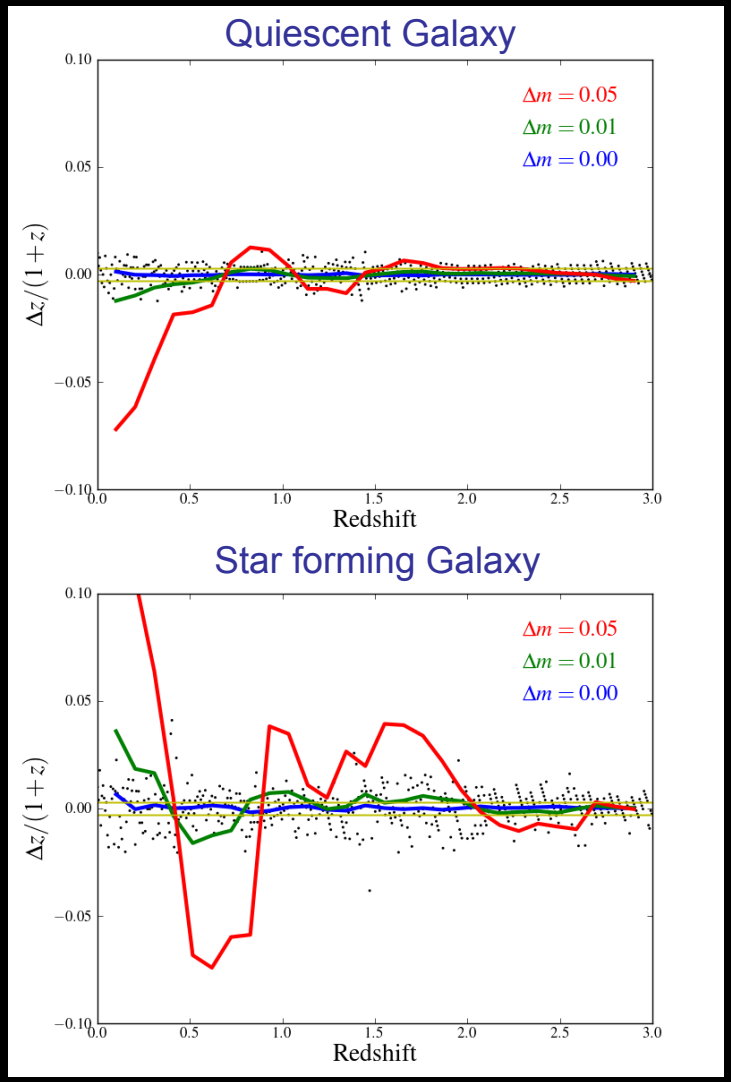
- Filter set
  - What is the optimal filter set
- Training samples
  - Samples of 25-30K required (from the variance in SEDs)
  - High redshift is a challenge but achievable (but how to draw them)
- Characterization of photometric errors
  - Photometry of sources at the limits of photometric surveys is not a solve problem (deblending, optimal aperture definition, characterization of variance and covariance)
  - Shape and inclination are detectable in photometric redshift residuals
- Characterization of the P(z)
  - Exclusion of bad photo-z's from the likelihood function
  - Propagating P(z) to an analysis is expensive

Training set inefficiencies

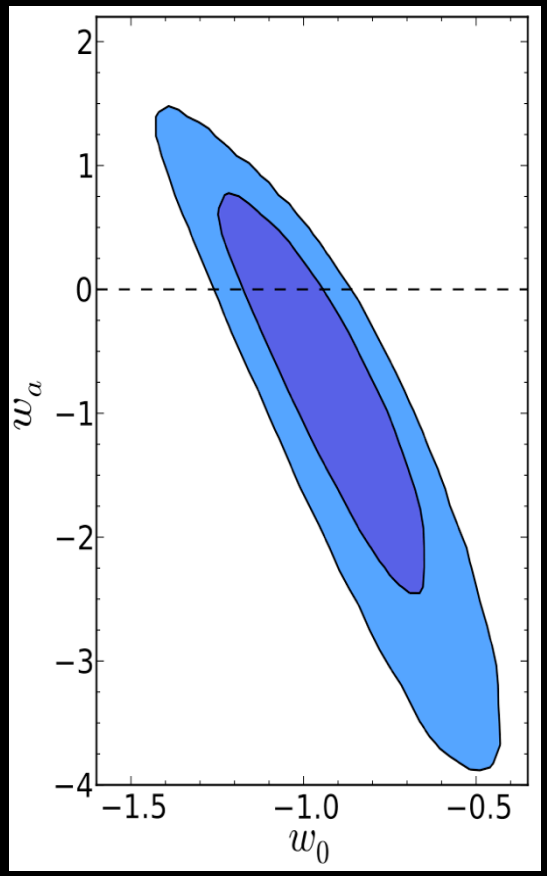


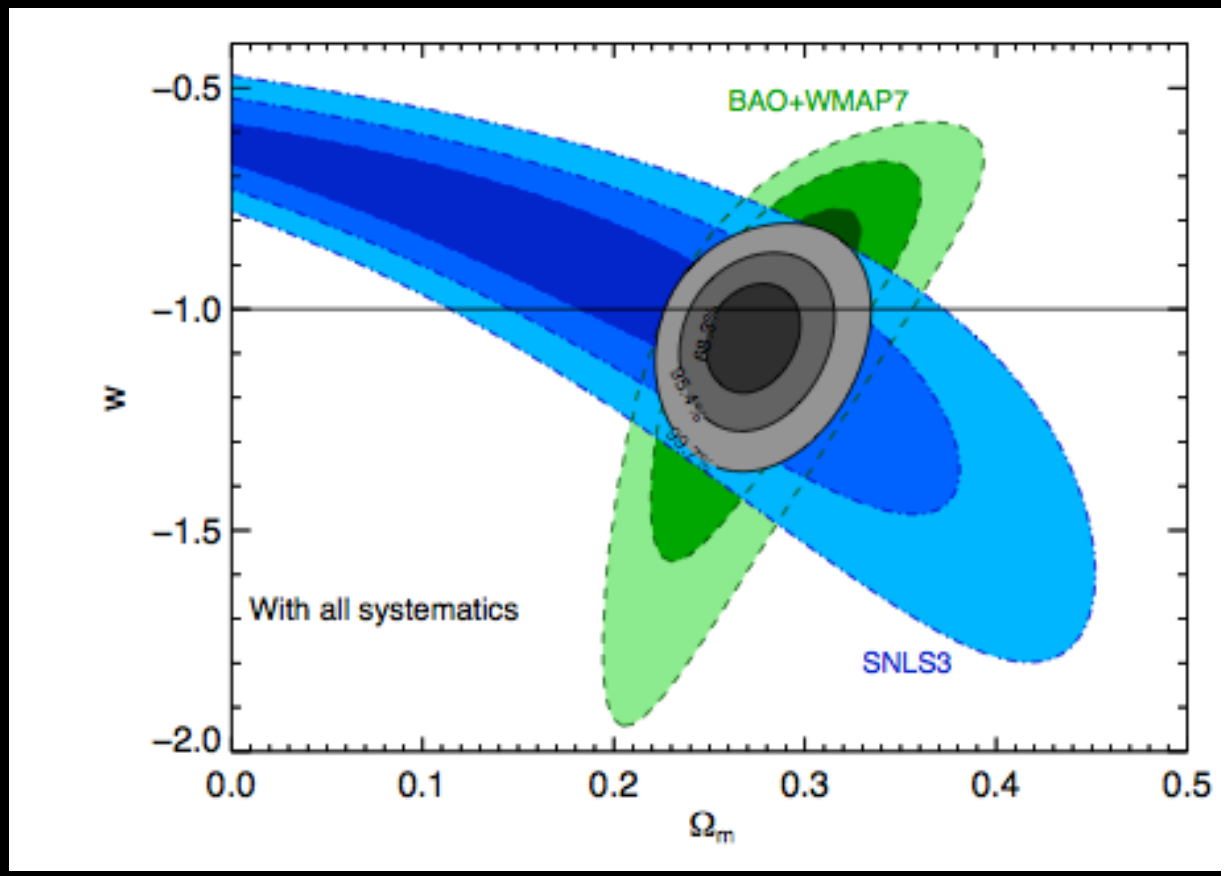
Newman et al 2011

Sensitivity to photometric error



Cosmology from supernovae





Sullivan et al. 2011

State of supernovae science: FOM=11

Calibration becomes the primary systematic

Photometric calibration is the primary systematic

Primary calibration of a photometric system and definition of the relative throughputs

Table 3. Detailed summary of systematic uncertainties

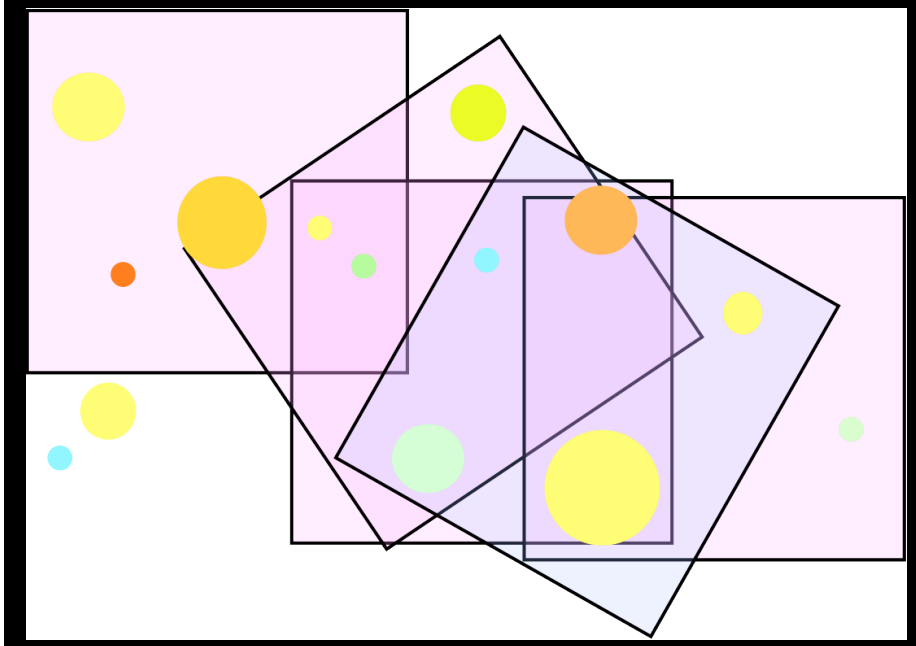
Source	Ω_m	w	Relative area[a]
Statistical only	$0.2763^{+0.0163}_{-0.0132}$	$-1.0430^{+0.0543}_{-0.0546}$	1.0
All systematics	$0.2736^{+0.0186}_{-0.0145}$	$-1.0676^{+0.0799}_{-0.0821}$	1.693
All systematics, except calibration	$0.2756^{+0.0164}_{-0.0133}$	$-1.0481^{+0.0573}_{-0.0580}$	1.068
All systematics, except host term	$0.2738^{+0.0186}_{-0.0145}$	$-1.0644^{+0.0790}_{-0.0809}$	1.677
All systematics, fixing α, β[b]	$0.2656^{+0.0179}_{-0.0144}$	$-1.1168^{+0.0807}_{-0.0824}$	1.641
Contribution of different systematics:			
Calibration	$0.2750^{+0.0185}_{-0.0150}$	$-1.0581^{+0.0774}_{-0.0791}$	1.614
SN Ia model	$0.2767^{+0.0163}_{-0.0132}$	$-1.0403^{+0.0543}_{-0.0547}$	1.013
Peculiar velocities	$0.2761^{+0.0163}_{-0.0132}$	$-1.0452^{+0.0544}_{-0.0548}$	1.002
Malmquist bias	$0.2758^{+0.0163}_{-0.0132}$	$-1.0474^{+0.0548}_{-0.0553}$	1.014
Non SN Ia contamination	$0.2763^{+0.0163}_{-0.0132}$	$-1.0430^{+0.0543}_{-0.0546}$	1.000
Milky Way extinction	$0.2762^{+0.0164}_{-0.0133}$	$-1.0441^{+0.0553}_{-0.0557}$	1.023
SN redshift evolution	$0.2763^{+0.0163}_{-0.0132}$	$-1.0408^{+0.0544}_{-0.0547}$	1.017
Host galaxy term	$0.2762^{+0.0163}_{-0.0132}$	$-1.0453^{+0.0556}_{-0.0562}$	1.029
Calibration:			
Colors of BD 17° 4708	$0.2719^{+0.0170}_{-0.0137}$	$-1.0720^{+0.0639}_{-0.0639}$	1.239
SED of BD 17° 4708	$0.2771^{+0.0170}_{-0.0138}$	$-1.0390^{+0.0623}_{-0.0630}$	1.205
SNLS zeropoints	$0.2767^{+0.0168}_{-0.0136}$	$-1.0421^{+0.0603}_{-0.0609}$	1.166
Low-z zeropoints	$0.2753^{+0.0164}_{-0.0133}$	$-1.0527^{+0.0578}_{-0.0586}$	1.078
SDSS zeropoints	$0.2767^{+0.0164}_{-0.0133}$	$-1.0411^{+0.0544}_{-0.0548}$	1.015
SNLS filters	$0.2789^{+0.0170}_{-0.0138}$	$-1.0330^{+0.0585}_{-0.0586}$	1.136
Lowz filters	$0.2766^{+0.0163}_{-0.0132}$	$-1.0402^{+0.0547}_{-0.0550}$	1.010
SDSS filters	$0.2770^{+0.0164}_{-0.0133}$	$-1.0396^{+0.0544}_{-0.0548}$	1.007
HST zeropoints	$0.2769^{+0.0164}_{-0.0133}$	$-1.0412^{+0.0544}_{-0.0548}$	1.007
NICMOS nonlinearity	$0.2767^{+0.0164}_{-0.0133}$	$-1.0418^{+0.0545}_{-0.0548}$	1.009

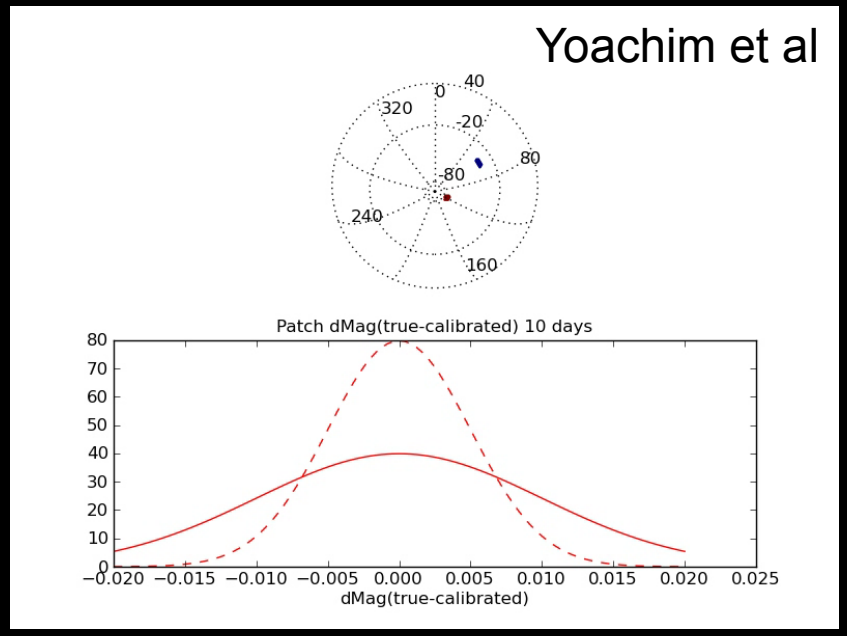
Sullivan et al. 2011

Open questions for SNe cosmology

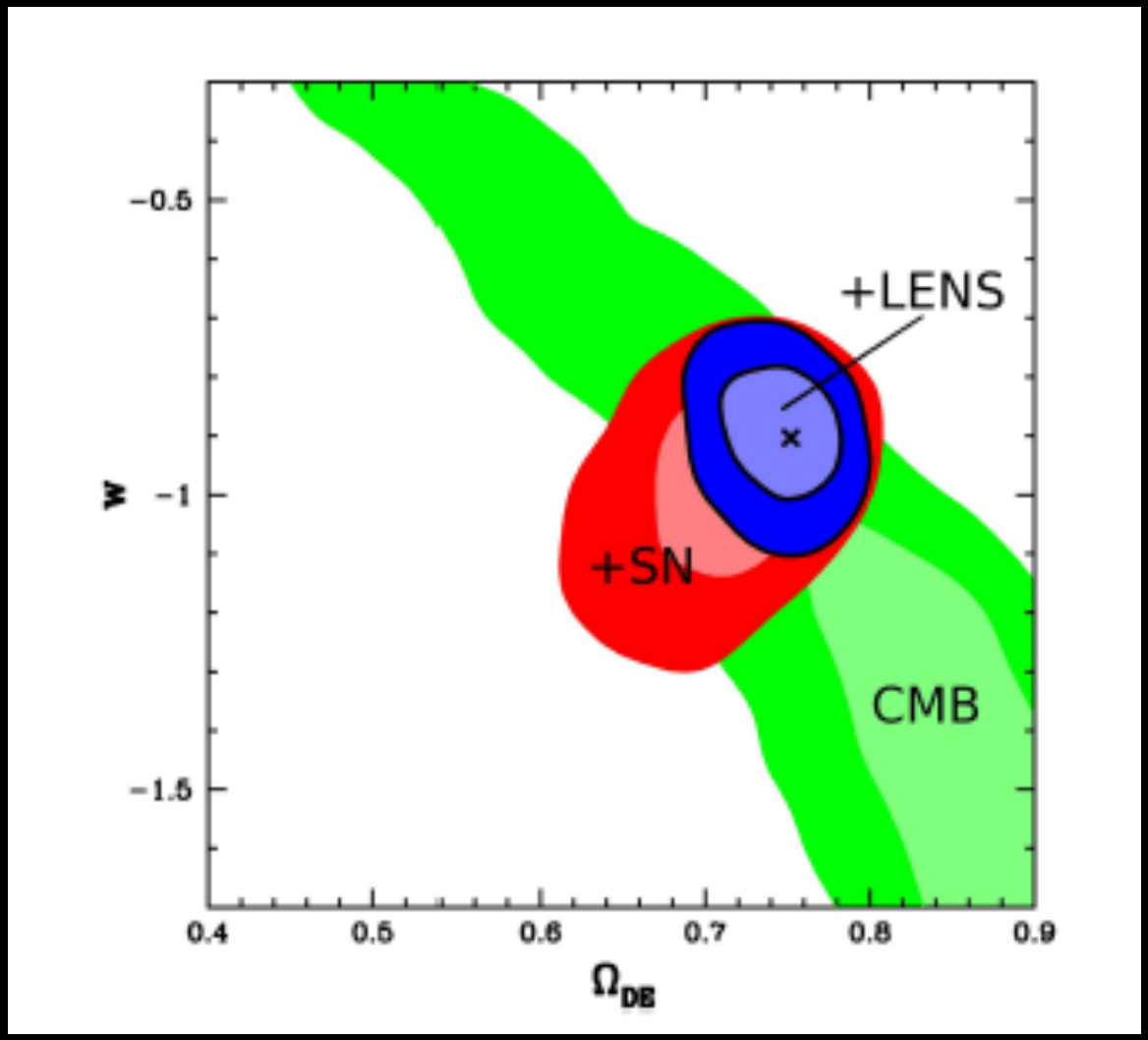
- Systematics (calibration and understanding the photometric system)
  - Point source repeatability: 0.005 mag rms
  - All-sky photometric zeropoint uniformity: 0.01 mag rms
  - Extinction
- Sampling techniques for light curves
- Subdivision of samples to study stability of results
- Designing surveys to minimize the systematics (or to enable the detection and characterization of these systematics)

Global calibration of ground-based data



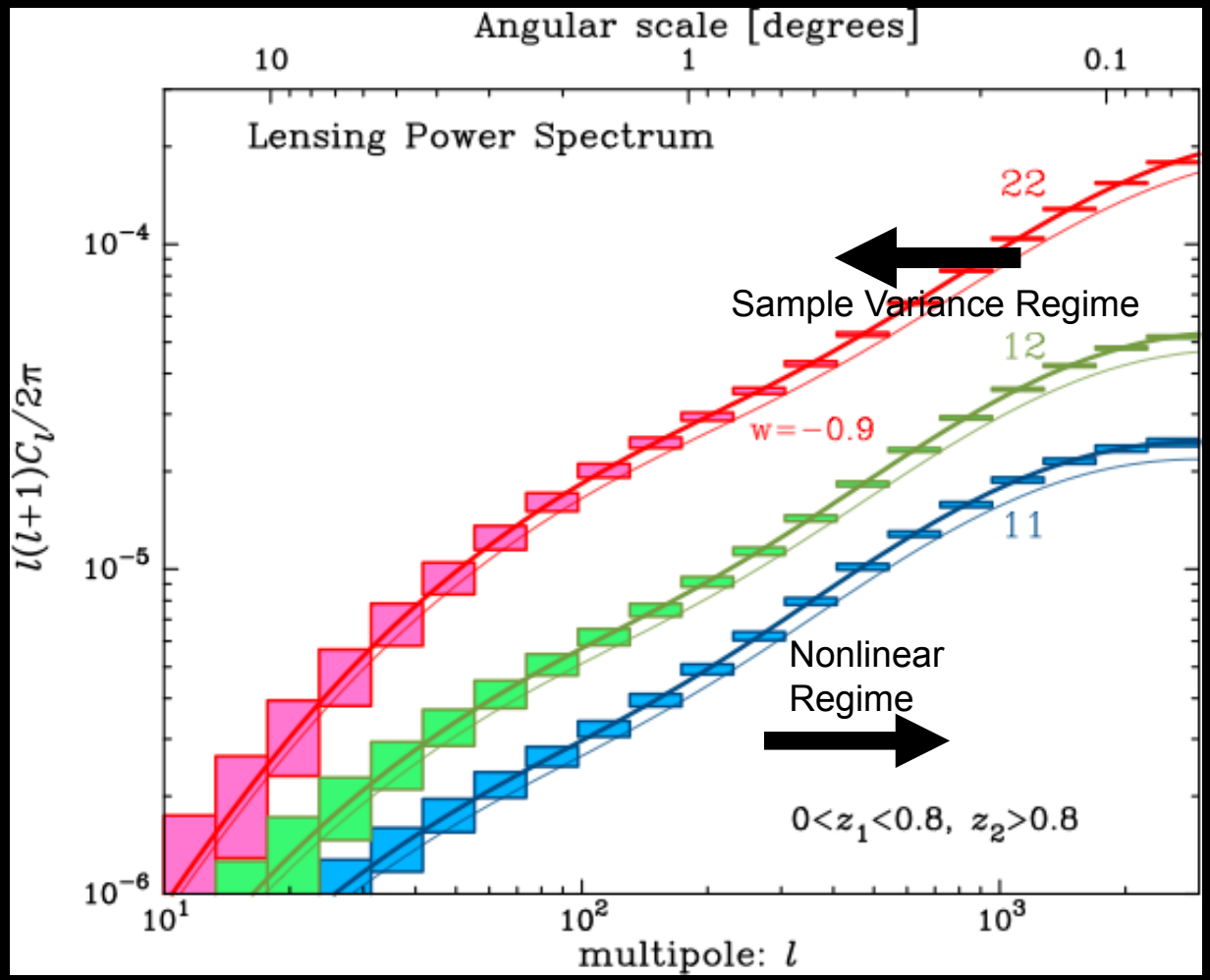


Gravitational Lensing



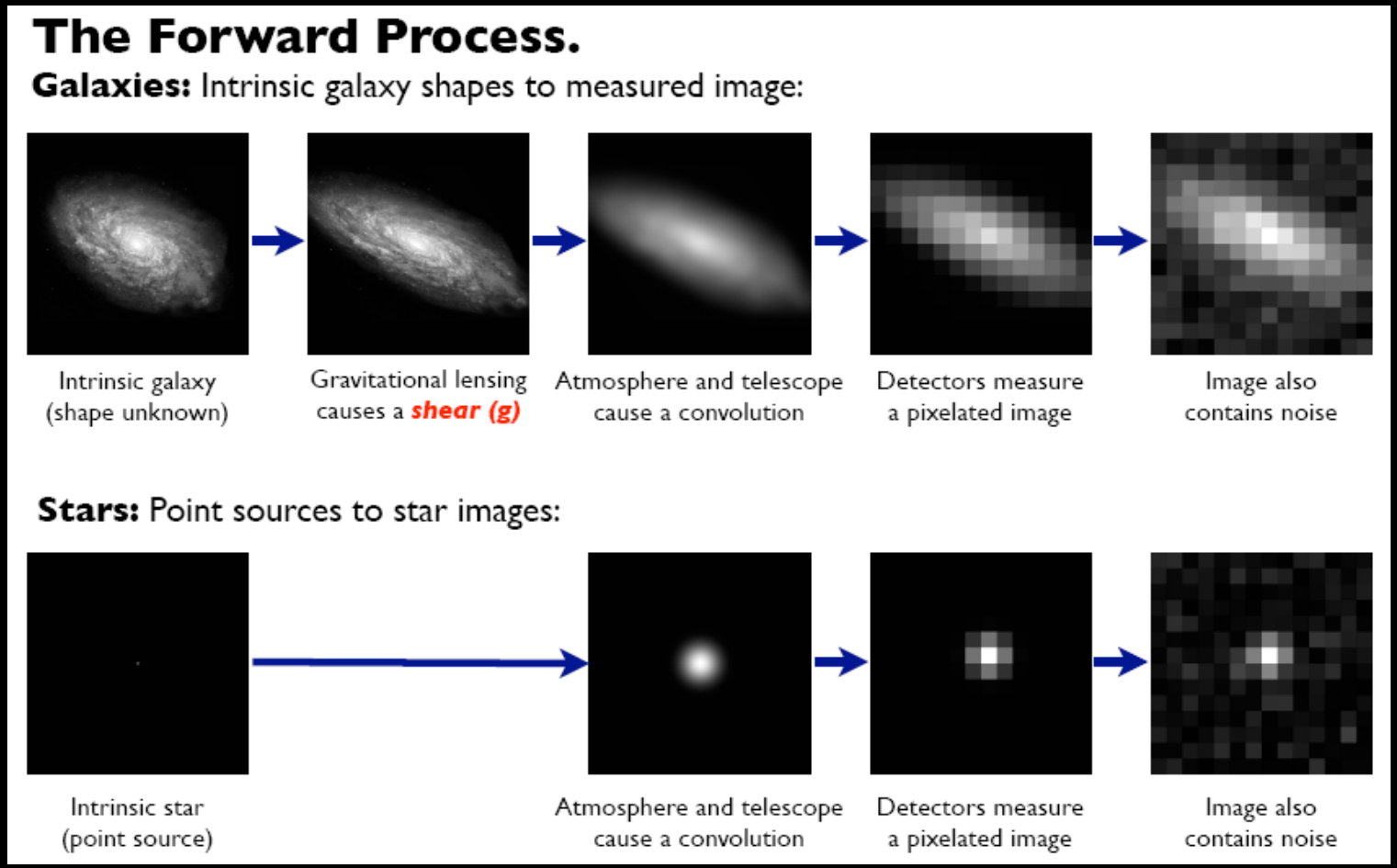
Jarvis et al 2006

Gravitational lensing



LSST Science Book

Measuring shapes



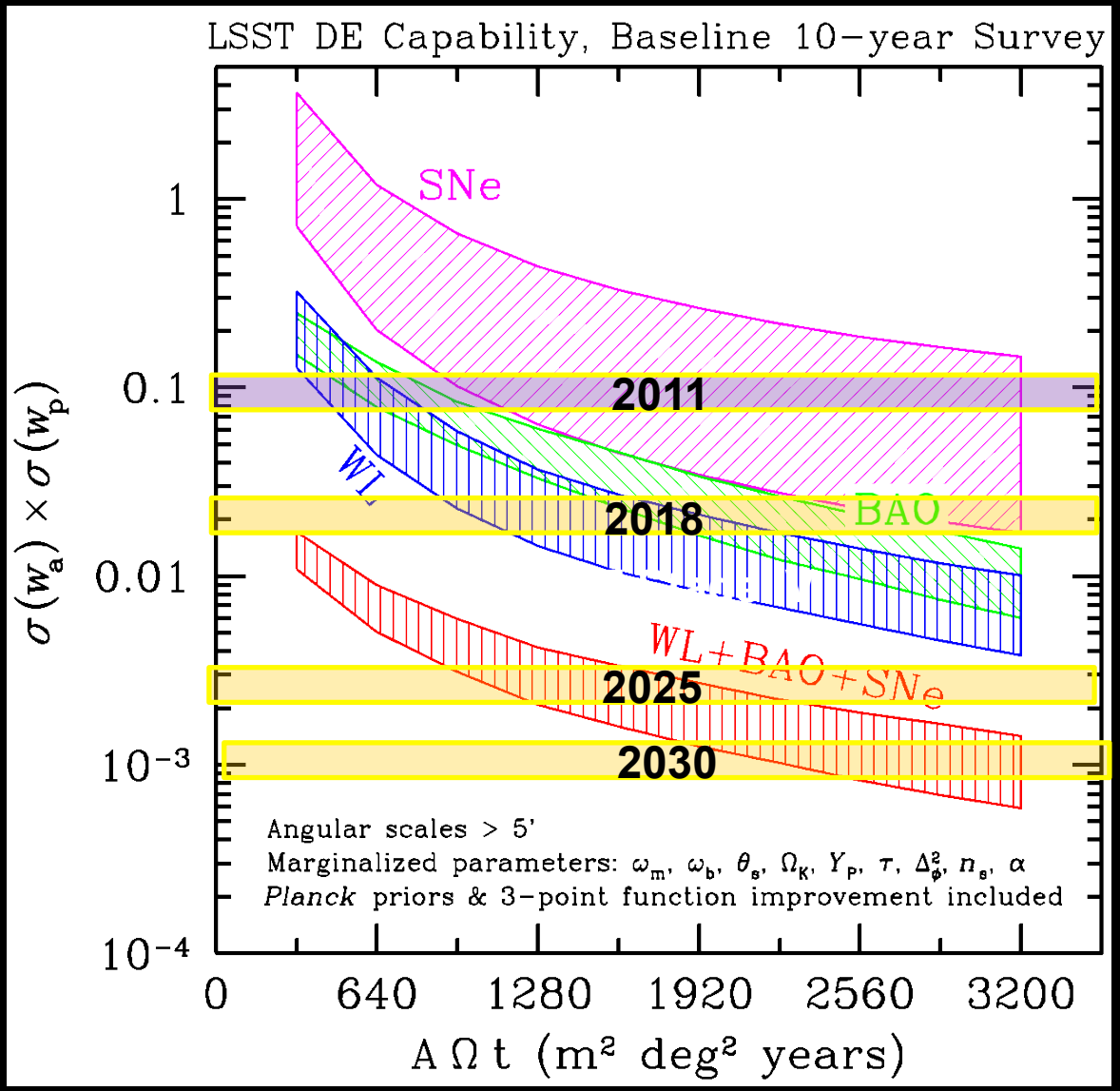
S. Bridle

Open questions for weak lensing cosmology

- Intrinsic alignment errors
- Photometric redshifts
  - Accuracy is critical in reducing the bias in dark energy parameters and in controlling intrinsic alignment errors. Photo-z's are characterized by scatter and bias. Bias required to be controlled at the ~0.003 level for LSST.
- Shear calibration and PSFs
  - Lensing shape measurements are very challenging (require ~0.002 level). With simulated images can get sub-percent performance in high S/N regime. Multiple exposures help reduce PSF from stochastic sources - atmosphere etc.
  - PCA based interpolation corrects non-stochastic sources and improves with survey size

Courtesy of Bhuv Jain

Evolution of the FOM



FoM = 11

FoM = 50

FoM = 500

FoM > 800