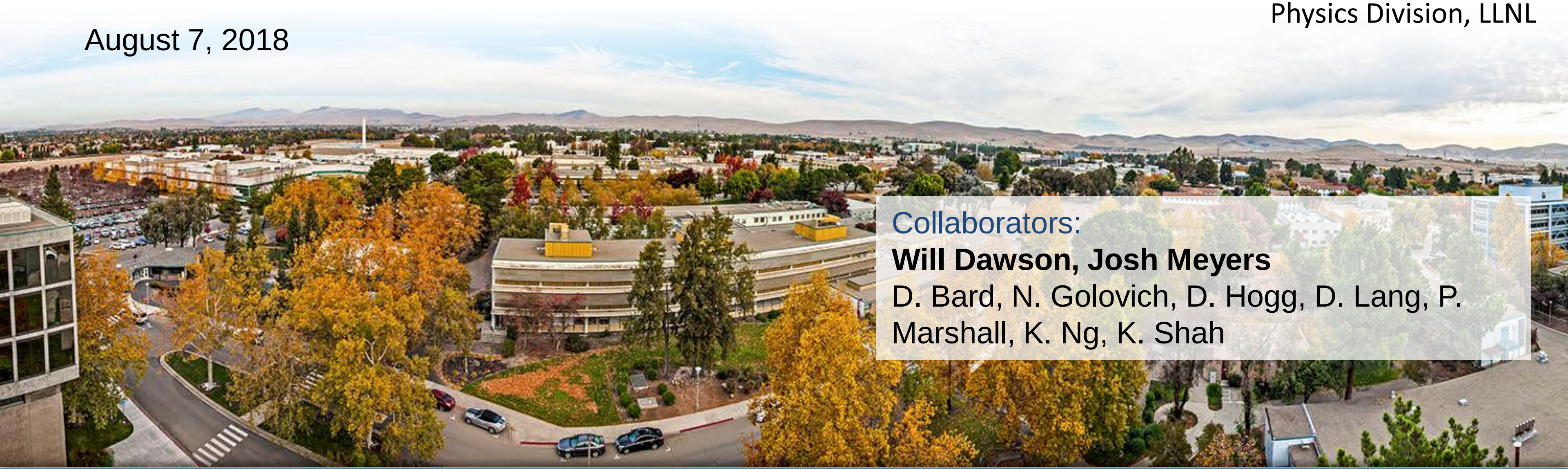
# Probabilistic Modeling to Measure Dark Energy

Data Science Workshop 2018

Michael D. Schneider
Physics Division, LLNL

August 7, 2018



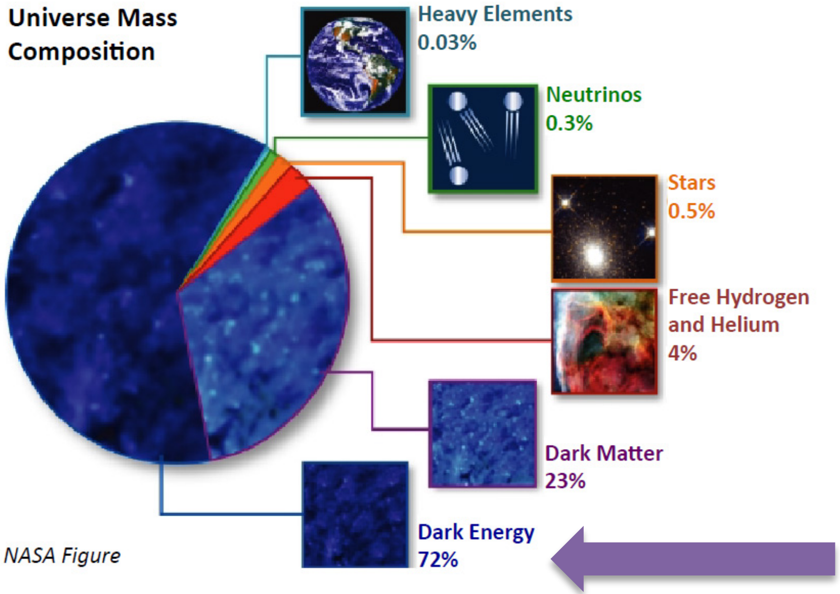
Collaborators:
**Will Dawson, Josh Meyers**
D. Bard, N. Golovich, D. Hogg, D. Lang, P. Marshall, K. Ng, K. Shah

LLNL-PRES-733055

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE-AC52-07NA27344. Lawrence Livermore National Security, LLC

Lawrence Livermore National Laboratory

# *Science driver:* The expansion of the universe is accelerating, but we know almost nothing about the mechanism – "Dark Energy"



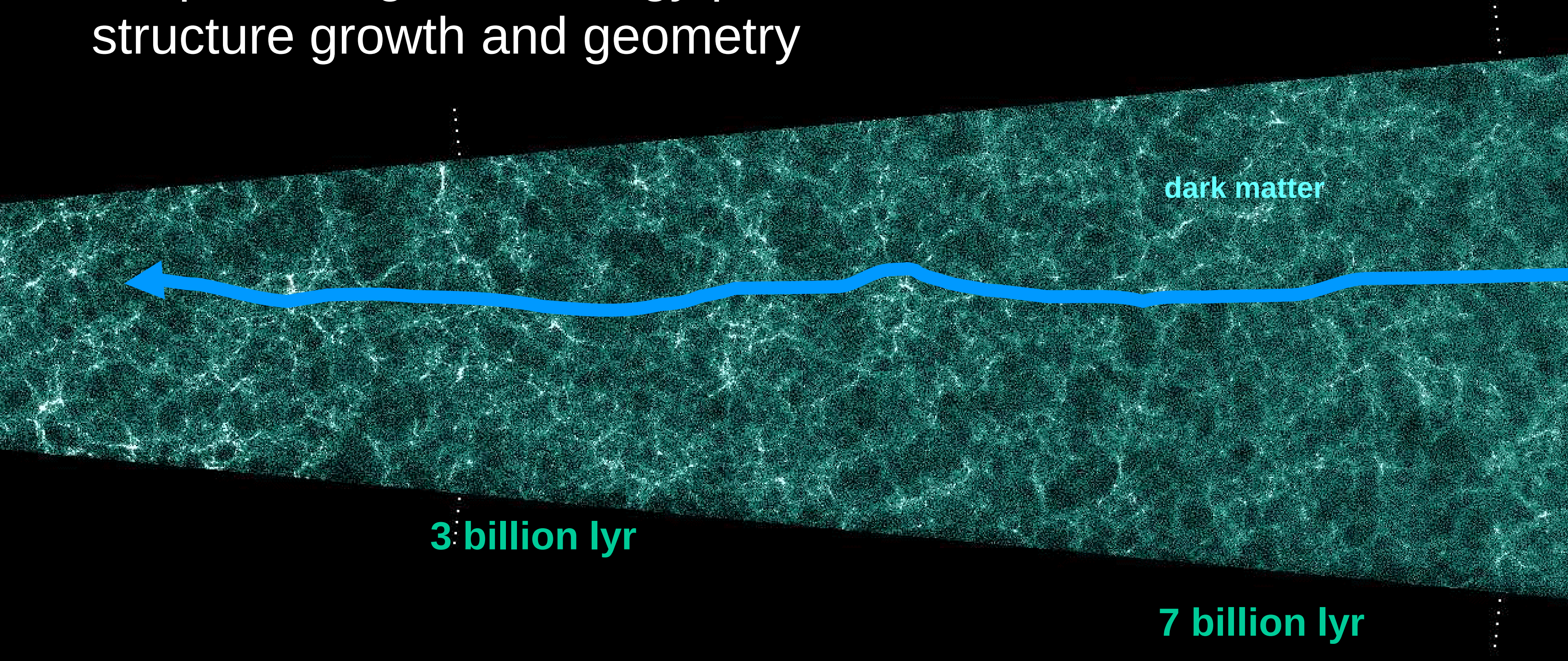

What causes cosmic acceleration?

**Three possibilities:**

1. The universe is filled with a negative-pressure component that gives rise to 'gravitational repulsion': **Dark Energy**
2. The theory of General Relativity (gravity) is wrong on cosmic distance scales
3. The universe is inhomogeneous and only apparently accelerating, due to large-scale structure (unlikely given current data)

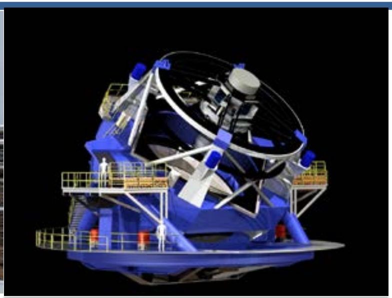
Path to the greatest prize in physics:
Reconciliation of gravity and quantum mechanics.

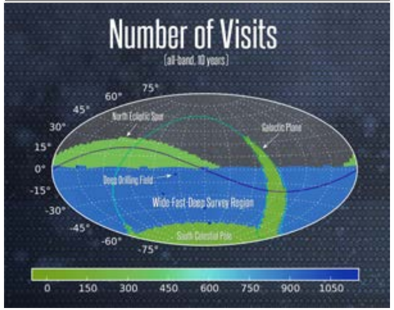
# Gravitational lensing traces mass structure vs cosmic time – A promising dark energy probe that is sensitive to both structure growth and geometry



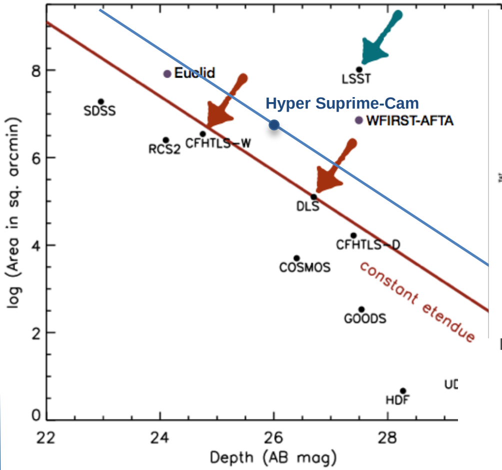
# The Large Synoptic Survey Telescope (LSST) will be a primary dark energy instrument in the next decade

## The LSST gravitational lensing measurement of dark energy is systematics limited






- DOE/NSF project
- 3.2Gpix camera
- 2π steradians
- 10 sq. deg. FOV
- 30 sec visits every night for 10 years
- 15Tb / night
- Commissioning start: FY20
- 15-years of LLNL contributions



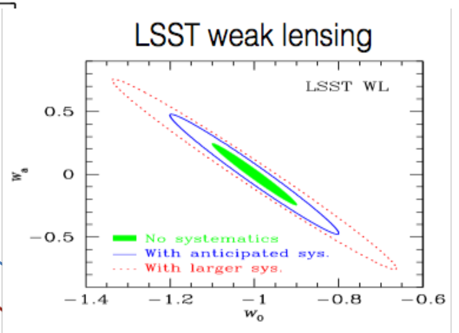
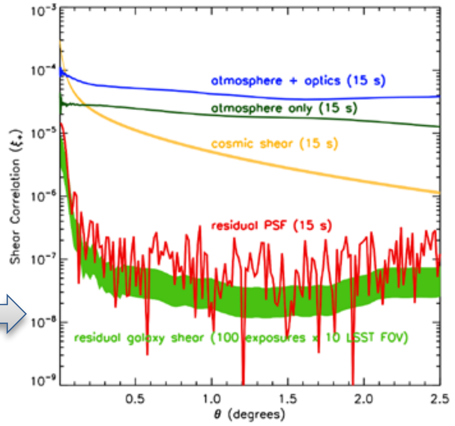



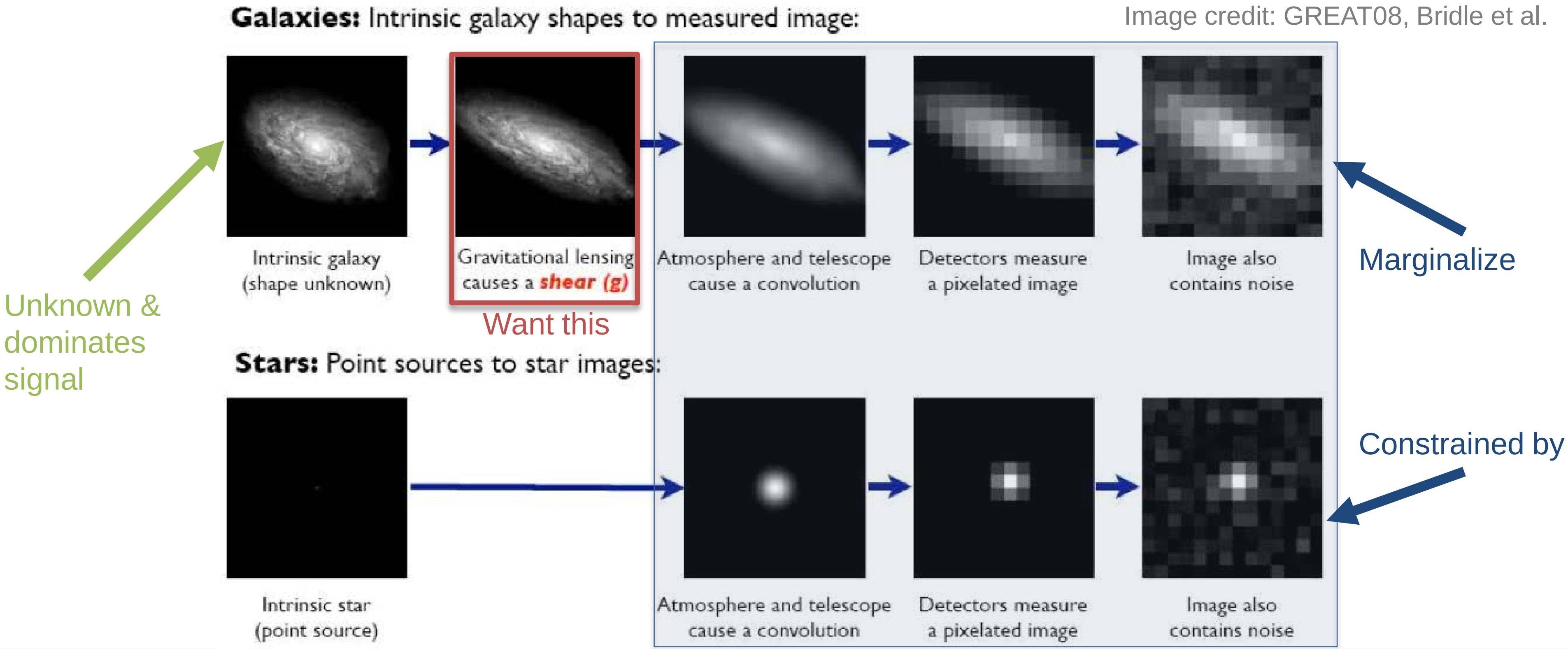

Image systematics must be controlled at this level




4 billion galaxy images

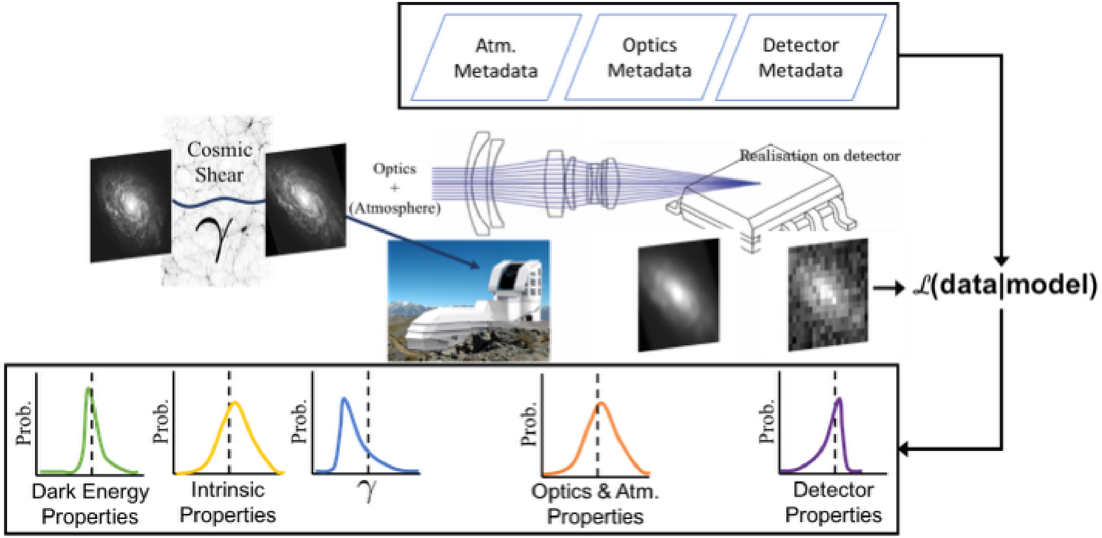
# Weak lensing of galaxies: the forward model

Image credit: GREAT08, Bridle et al.

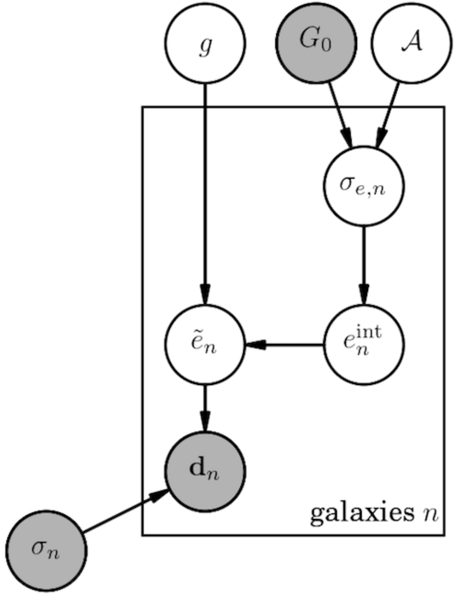


# *Our solution:* Hierarchical Bayesian forward models of the image data

The forward model of our galaxy image data




A simplified probabilistic graphical model for galaxy image data with cosmic shear, *g*

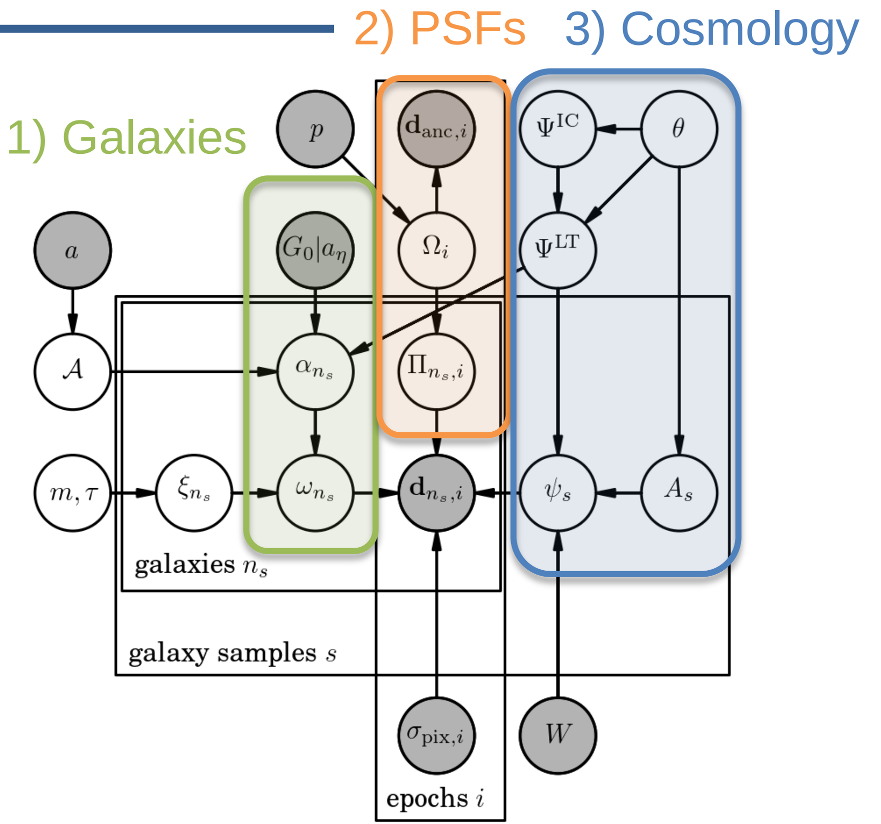


# The complete statistical model for cosmic shear

arXiv:1411.2608

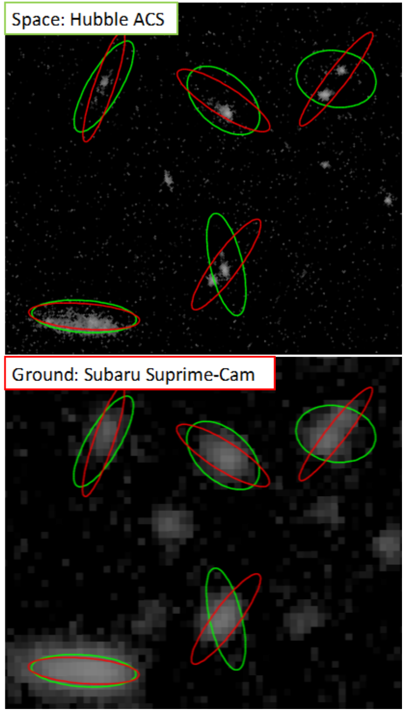
## Precision control of systematics via forward modeling and marginalization

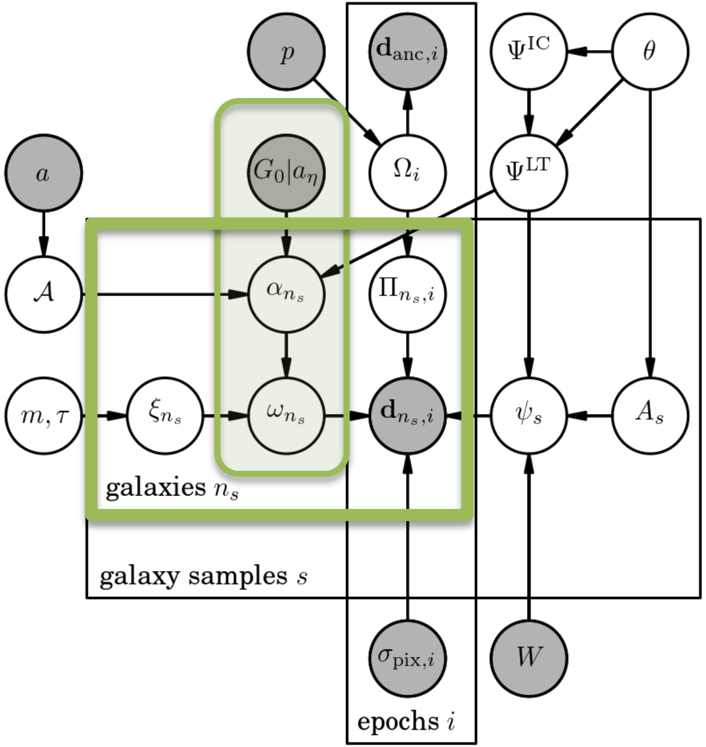
| Parameter | Description |
|---|---|
| $\boldsymbol{\theta}$ | Cosmological parameters |
| $\Psi^{\rm IC}$ | Initial conditons for the 3D gravitational potential |
| $\Psi^{\rm LT}$ | Late-time 3D gravitational potential |
| $\psi_s$ | 2D lens potential (given source photo-$z$ bin $s$) |
| $A_s$ | Parameters for the line-of-sight source distribution |
| $\Pi_{n_s,i}$ | PSF for galaxy $n_s$ observed in epoch $i$ |
| $\Omega_i$ | Observing conditions in epoch $i$ |
| $\{\omega_{n_s}\}$ | Galaxy model parameters; $n_s = 1, \ldots, n_{{\rm gal},s}$ |
| $\{\alpha_{n_s}\}$ | Parameters for the distribution of $\{\omega_{n_s}\}$ |
| $\{\xi_{n_s}\}$ | Scaling parameters for $\{\omega_{n_s}\}$ |
| $m, \tau$ | Hyperprior parameters for $\{\xi_{n_s}\}$ |
| $\mathcal{A}$ | Hyperparameter for $\{\alpha_{n_s}\}$ classifications |
| $\{\mathbf{d}_{n_s,i}\}$ | Pixel data for galaxies $n_s = 1, \ldots, n_{{\rm gal},s}$ in epoch $i$ |
| $G_0 \vert a_\eta$ | Prior specification for $\{\alpha_{n_s}\}$ |
| $s$ | Source sample (e.g., photo-$z$ bin) |
| $W$ | Survey window function |
| $\mathbf{d}_{{\rm anc},i}$ | Ancillary data for PSF in epoch $i$ |
| $p$ | Prior params. for observing conditions |
| $a$ | Prior params. for $\mathcal{A}$ |
| $\sigma_{{\rm pix},i}$ | Pixel noise r.m.s. in epoch $i$ |
| $I$ | Model selection assumptions |



# Naïve approach is intractable

1. Galaxy models be joint fitted to all available epochs $i$

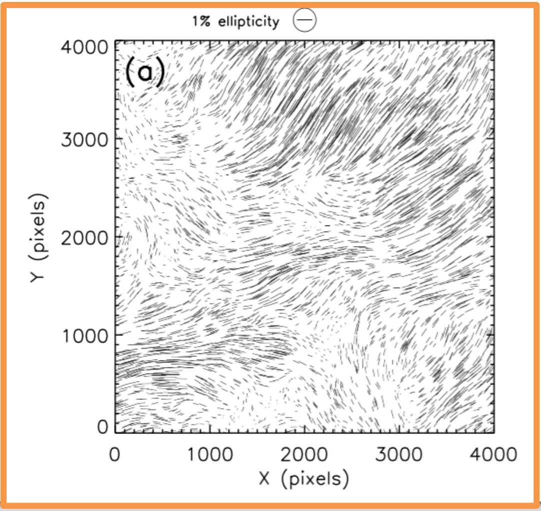


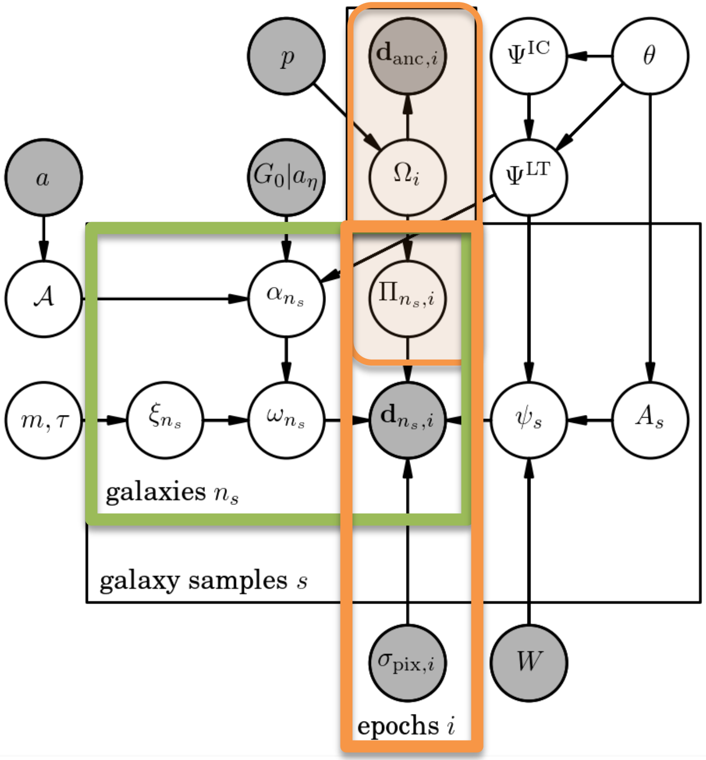



# Naïve approach is intractable

1. Galaxy models be joint fitted to all available epochs $i$

2. PSF models must be joint fitted to all galaxies in an exposure $n_s$

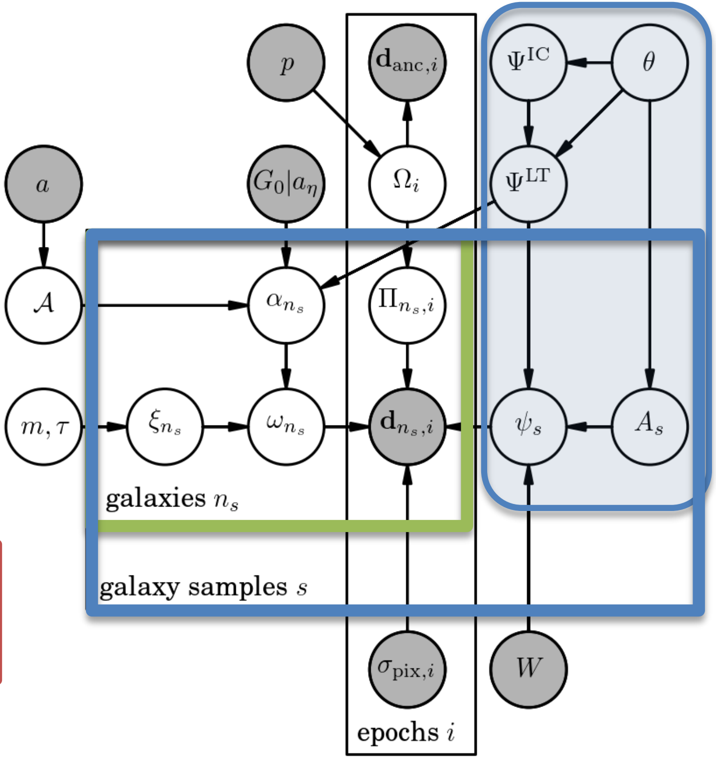


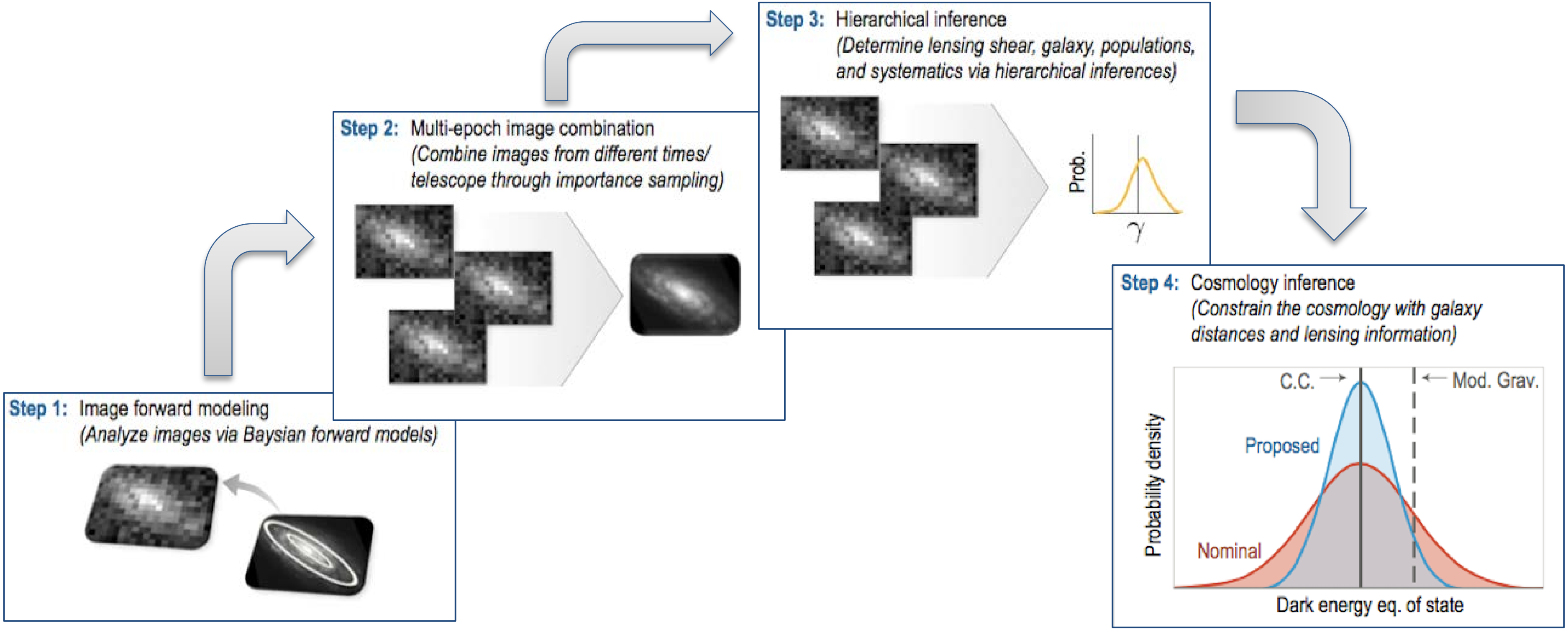
# Naïve approach is intractable

1. Galaxy models be joint fitted to all available epochs $i$
2. PSF models must be joint fitted to all galaxies in an exposure $n_s$
3. Cosmology must be joint fitted to all galaxy samples & epochs

The principled inference requires fitting all pixels of all surveys simultaneously



# We have developed a tractable ‘divide & conquer’ computational approach for the complete statistical model for multiple imaging surveys



# Importance Sampling allows tractable divide & compute :
## The pseudo-marginal likelihood

**Want:**

Galaxy dist. — Likelihood

$$\Pr(\mathbf{d}|\alpha) \propto \prod_{n=1}^{n_{\rm gal}} \int d\omega_n \Pr(\omega_n|\alpha) \Pr(\mathbf{d}_{n,i}|\omega_n)$$

**Have samples from:**

$$\Pr\left(\omega_n|\mathbf{d}_n, I_0\right)$$

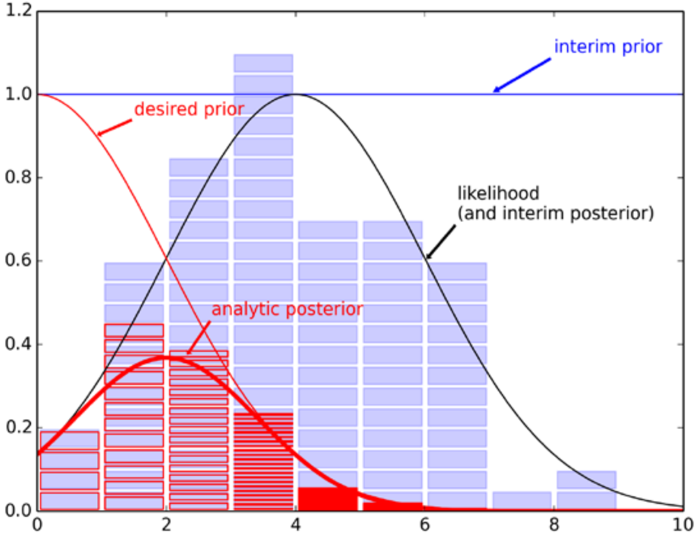
'Interim prior' specification

**Importance sampling:**

$$\Pr(\mathbf{d}_n|\alpha) \approx \frac{Z_n}{K} \sum_k \frac{\Pr(\omega_{nk}|\alpha)}{\Pr(\omega_{nk}|I_0)},$$

$$\Pr(\mathbf{d}|\alpha) = \prod_{n=1}^{n_{\rm gal}} \Pr(\mathbf{d}_n|\alpha).$$

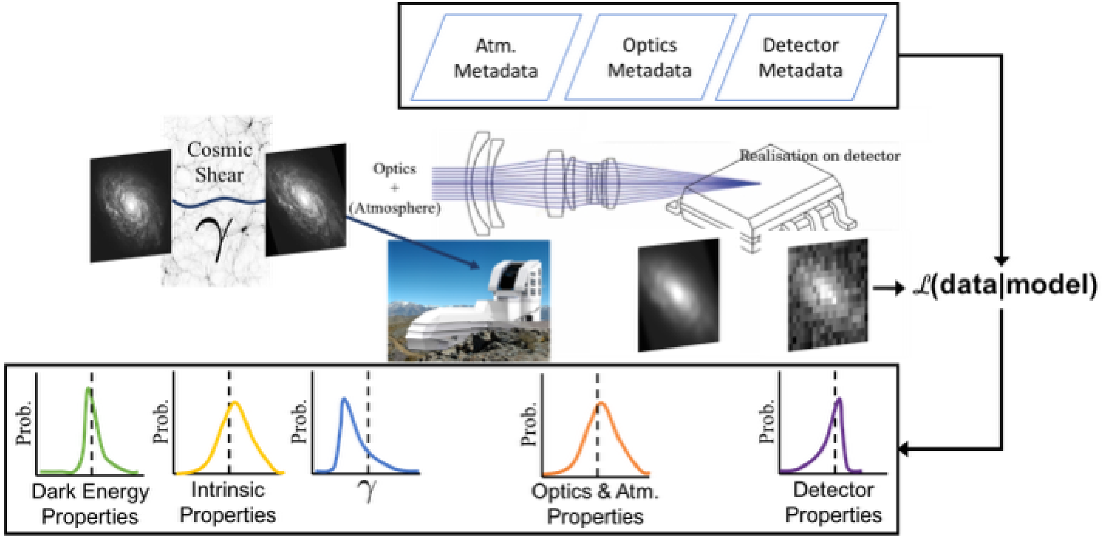
*Ongoing research question:*
How many interim samples are needed?



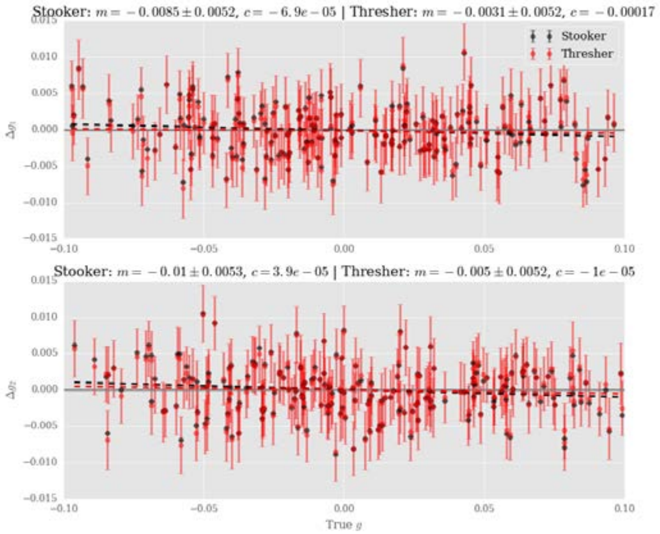

Credit: J. Meyers

# Our hierarchical Bayesian forward models can meet LSST systematics tolerances for galaxy shear when the model is accurate enough

## The forward model of our galaxy image data



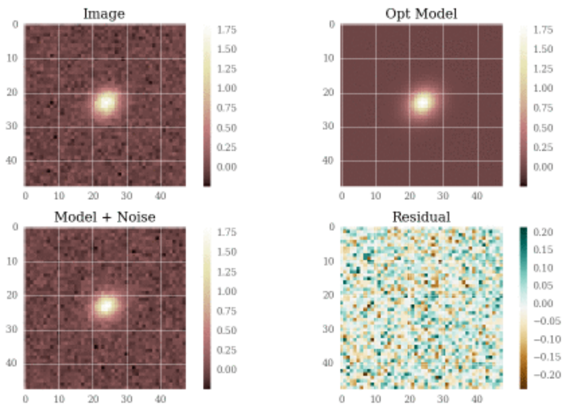

## Our approach works:

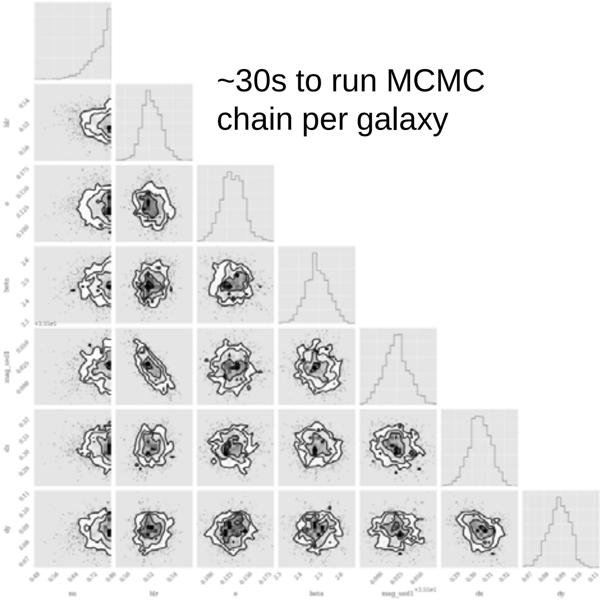



Sensitivity analyses with (simplified) galaxy simulation suites show biases well below LSST tolerances

# Accurate modeling in an MCMC framework is more computationally demanding than traditional approaches – but still tractable

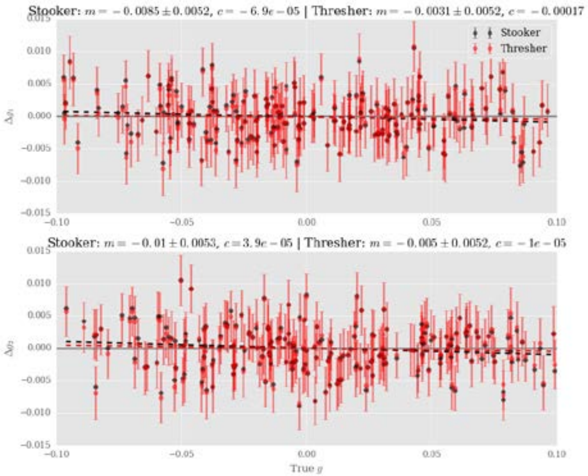
Every data point on the right is inferred from thousands of galaxy image model posterior inferences such as these:



~30s to run MCMC chain per galaxy

Our approach works:

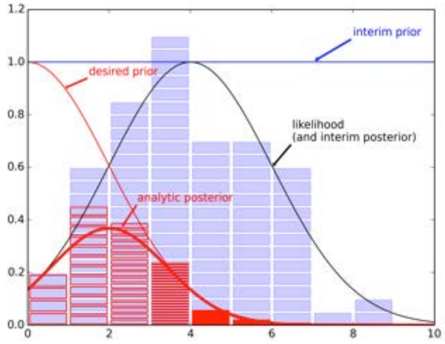


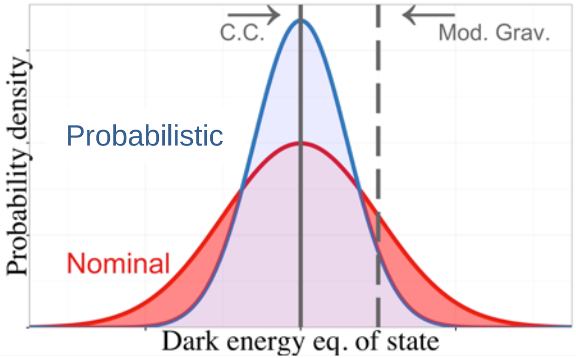

Sensitivity analyses with (simplified) galaxy simulation suites show biases well below LSST tolerances

LSST data volume: 4 billion galaxies, each seen 1000 times

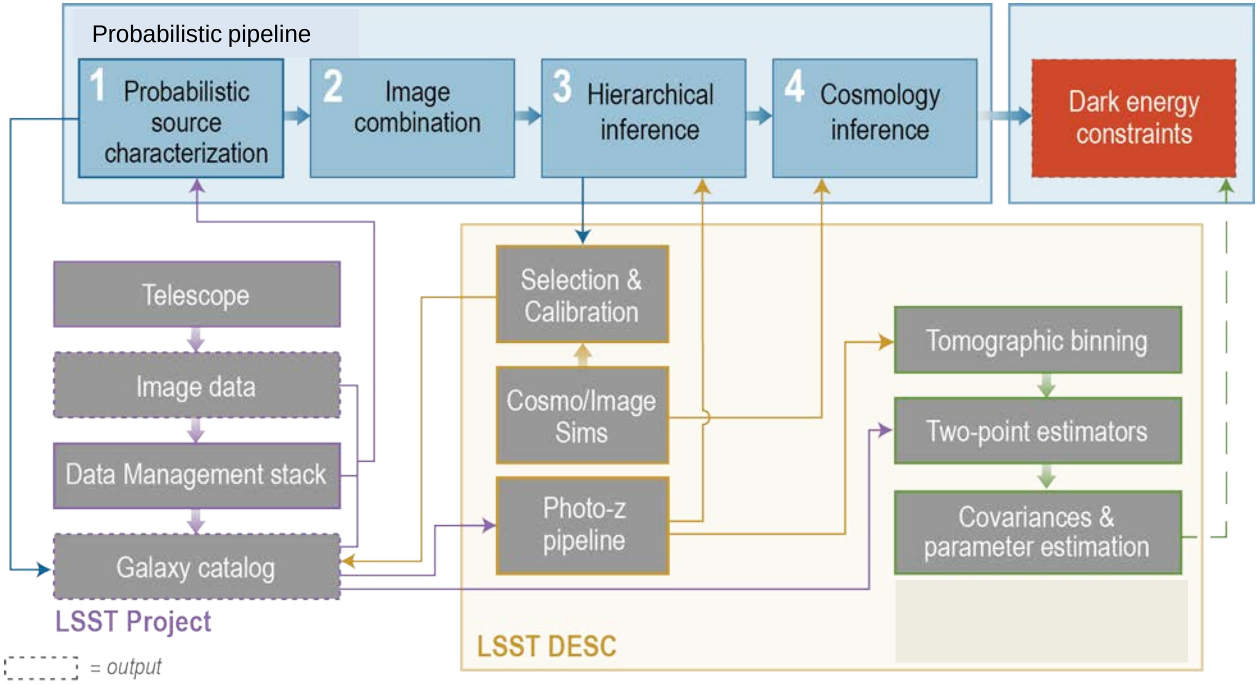
# Summary

- Cosmic shear is systematics limited & signal is dominated by PSF and astrophysics
  - A probabilistic approach is warranted to infer a small signal and mitigate biases
- A hierarchical probabilistic model for cosmic shear can trade bias for variance, but also can increase precision by learning latent structure in the galaxy distribution.
- Importance sampling methods allow tractable approaches to a probabilistic forward model of LSST & WFIRST imaging
  - With billions of galaxies and hundreds of epochs per galaxy modeling LSST or WFIRST imaging requires an approach to separating analyses of data subsets, even though statistically correlated
- We are able to sample from a probabilistic model with multiple hierarchies to marginalize both correlated image systematics and astrophysical properties of galaxies.



# The probabilistic weak lensing workflow plan for LSST



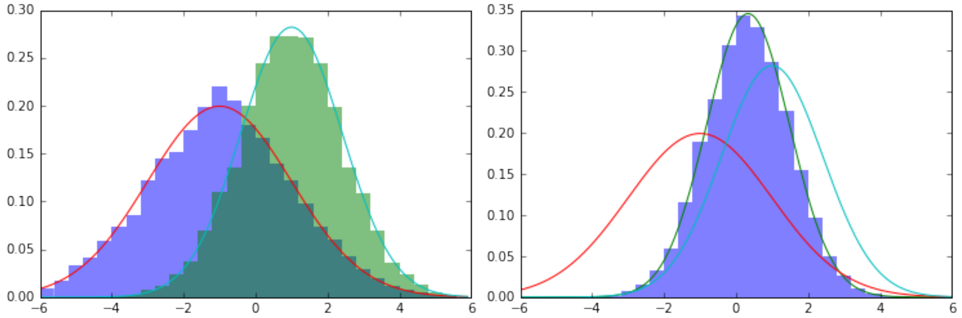
# How do we combine multiple observations of the same galaxy?

## Naïvely we must joint fit all epochs simultaneously

*Problem:* Imagine we have fit pixel data from LSST year 1.
How do we incorporate year 2 observations without redoing (expensive) calculations?

*Solution:* Consider single-epoch samples as draws from a multi-modal importance sampling distribution:

arXiv:1511.03095
Generalized Multiple Importance Sampling
Elvira, Martino, Luengo, & Bugallo



# Multiple importance sampling (MIS) via weighted pseudo-marginals

1. Sample from the conditional posterior for each epoch individually
2. Evaluate the ratio of the conditional posterior for each epoch $i$ to that of the MIS sampling distribution

'*cross-pollination*' needed:

Evaluate the likelihood of epoch $i$ given model parameter samples from epoch $j$, for all combinations of $i, j$.

A standard scatter / gather operation

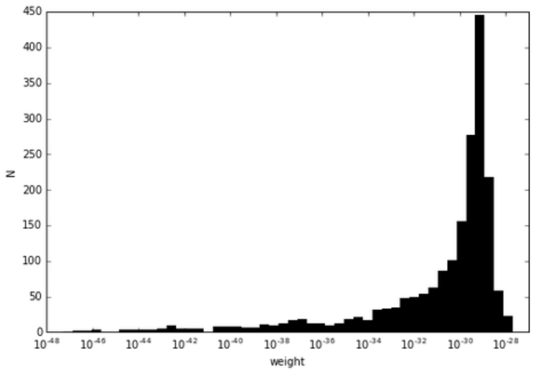
# Multiple importance sampling enables streaming data analysis

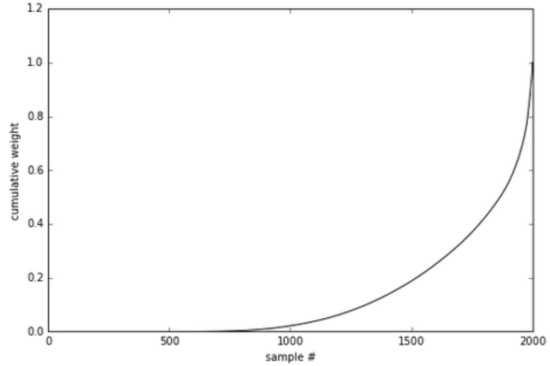
## Efficiency is significantly enhanced by using old data as a sampling 'prior'

- Draw parameter samples from first epoch under a nominal interim prior

- Draw samples from subsequent epochs with a prior informed by previous epoch samples

- Simulation studies show:
  - ~10% of samples have significant weight when combining 200 epochs in streaming fashion



arXiv:1610.06673

# Probabilistic cosmological mass mapping

*Objective:* infer the 3D gravitational potential of the initial conditions

Interpolate the unobserved lensing potential with GP

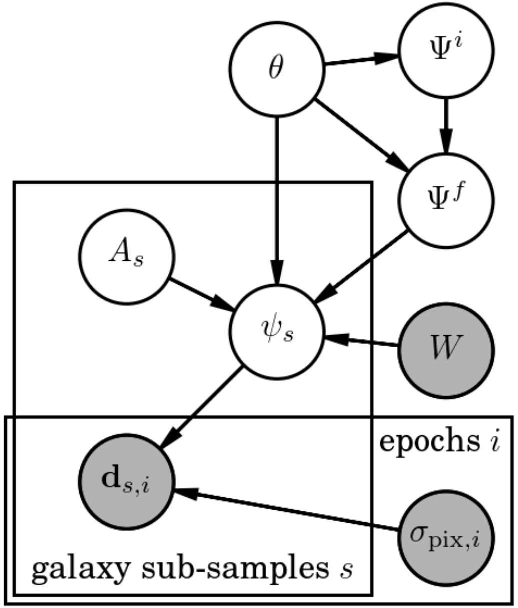
$$\psi_s \sim GP(0, \Sigma),$$

$\kappa, \gamma_1, \gamma_2$ are the second (spatial) derivatives of $\psi_s$

$$\mathrm{Cov}(\psi_{,ij}(\vec{x}), \psi_{,k\ell}(\vec{y})) = \Sigma_{,x_i x_j y_k y_\ell}(\vec{x}, \vec{y}).$$

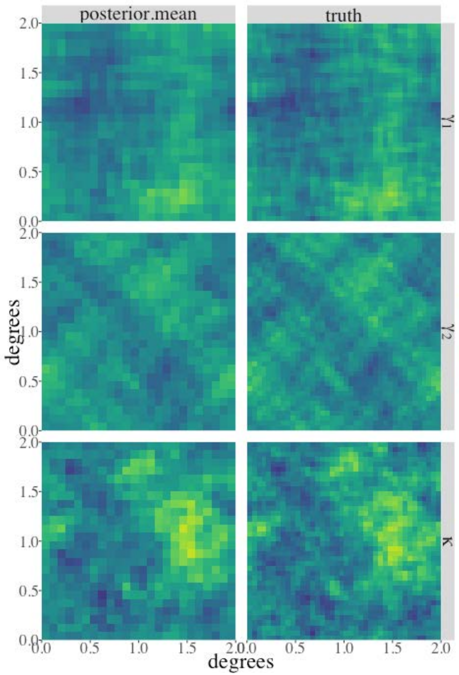
GP kernels of $\kappa, \gamma_1, \gamma_2$ are linear combinations of the 4th (spatial) derivatives of the kernel of $\psi_s$

Zero E/B mode mixing by construction

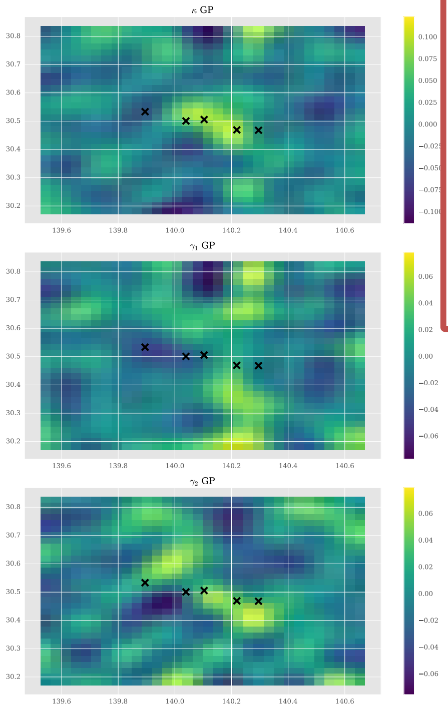


# Hierarchical inference of cosmological lensing mass distributions

Validation with simulations

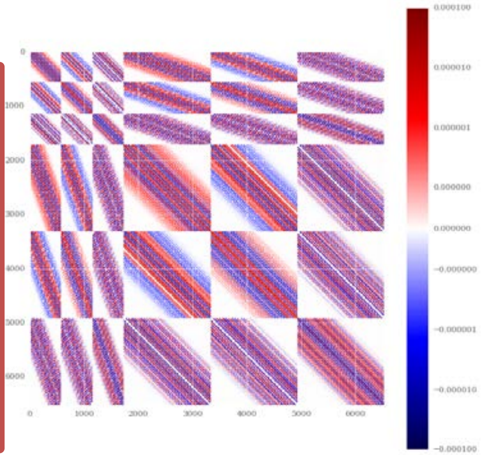


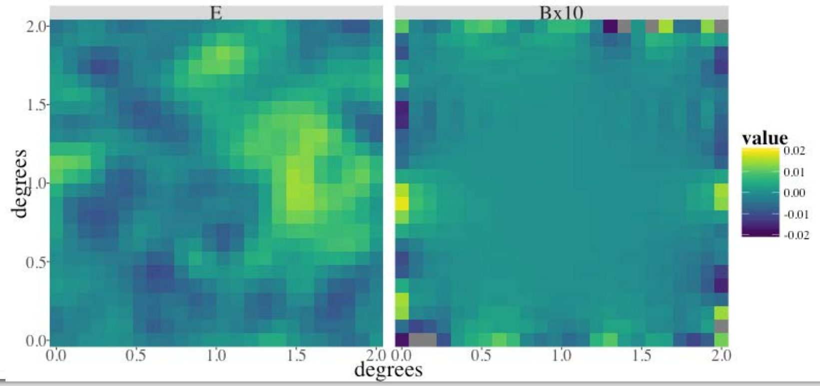

A real merging galaxy cluster




**New:**

- Linear and nonlinear scales reconstructed in one framework
- No E/B mode mixing by construction








Schneider+2017, ApJ

# Marginalizing PSFs: MIS makes this tractable

- LSST will have ~200 epochs per object per filter
  - We aim to marginalize the PSF $\prod_{n,i}$ in every epoch
  - The marginalization is constrained by:
    - Consistency of PSF realizations over the focal plane for each epoch
    - Consistency of the underlying source model across epochs
- Simplest approach (statistically, not computationally): Infer galaxy models given all epoch imaging simultaneously
  - "Interim" samples are of size: ~10 galaxy params + 200 * ~4 PSF params = ~1k parameters!