Joint pixel level processing of WFIRST and LSST

Enabling Cosmological Resonances Between WFIRST and LSST

Michael D. Schneider
with Will Dawson, Josh Meyers, Sam Schmidt

September 14, 2016

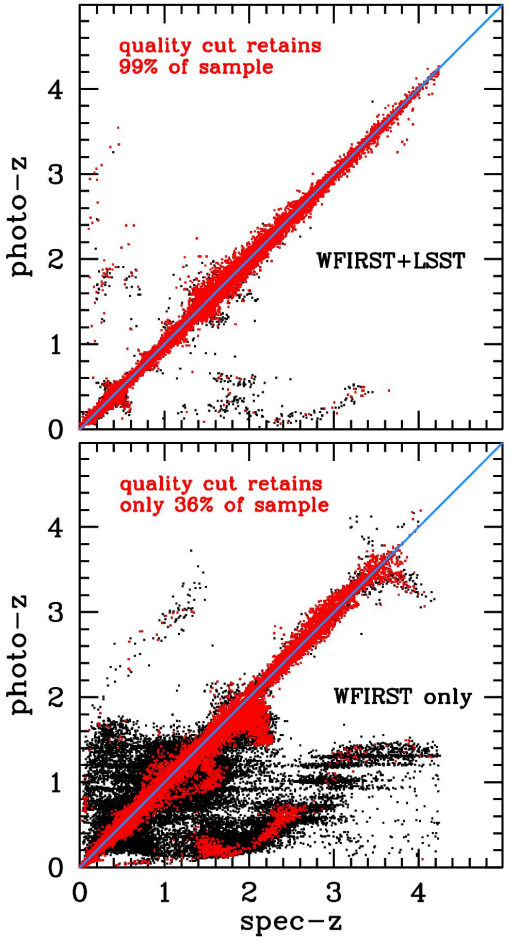
Collaborators:
D. Bard, D. Hogg, D. Lang, P. Marshall, R. Mandelbaum, K. Ng, T. Tyson

LLNL-PRES-691561
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

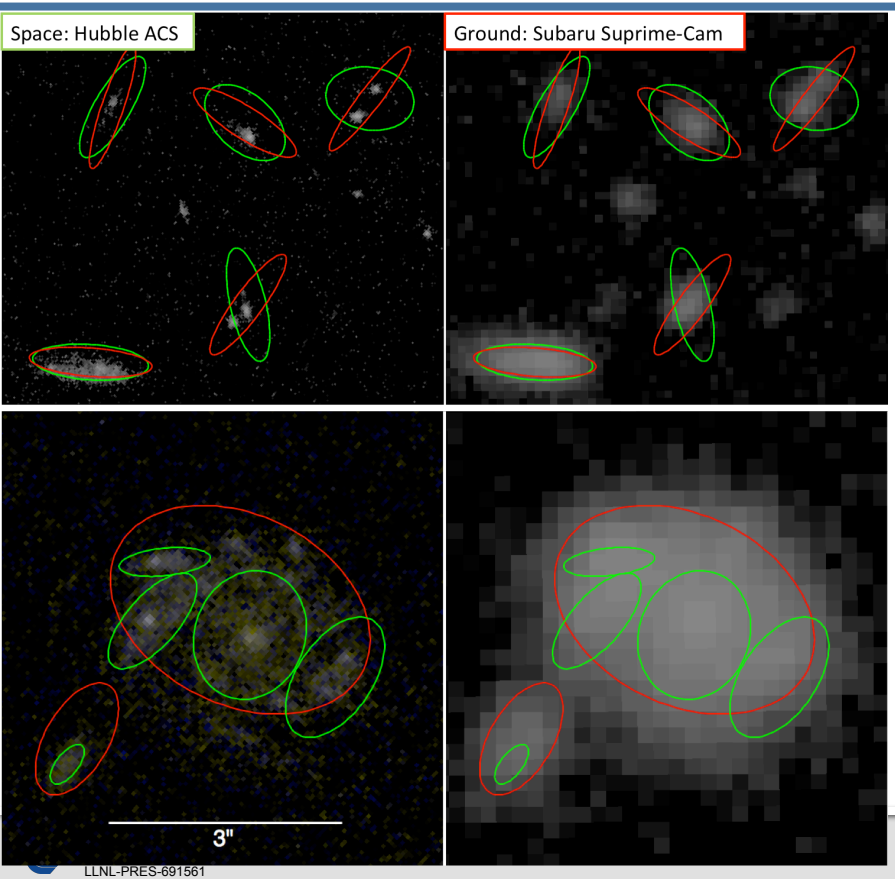
Lawrence Livermore National Laboratory

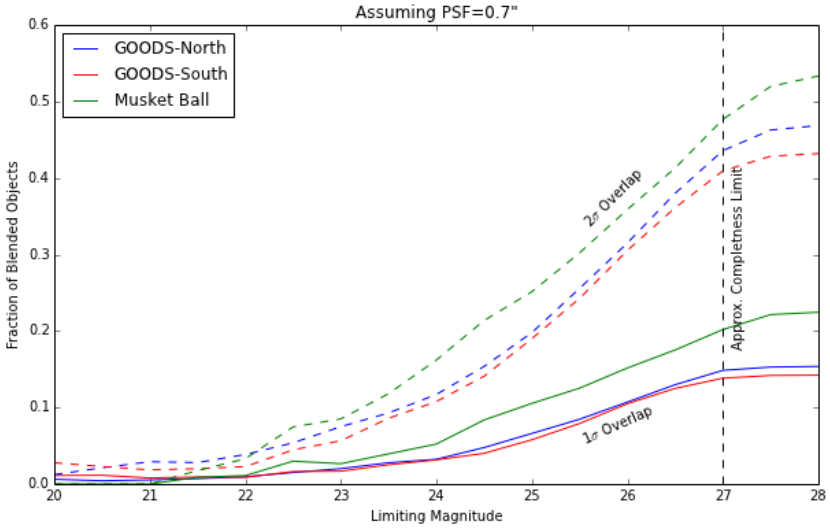
The 4 passbands in the High Latitude Survey do not yield useable photo-z's

- WFIRST-AFTA photo-z's limited by distinguishing features in galaxy SEDs at WFIRST wavelengths
  - not limited by photometric precision or spectroscopic training samples
- Combination with LSST 6 optical passbands is more than sufficient for shear, assuming a reliable cross-matching of catalogs/sources can be made.



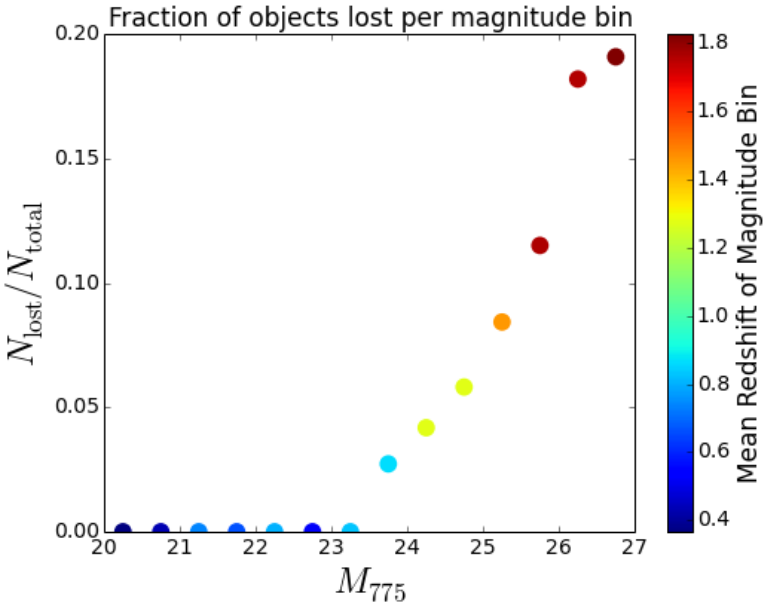
Catalog cross-matching is confused by significant object blending as seen by LSST

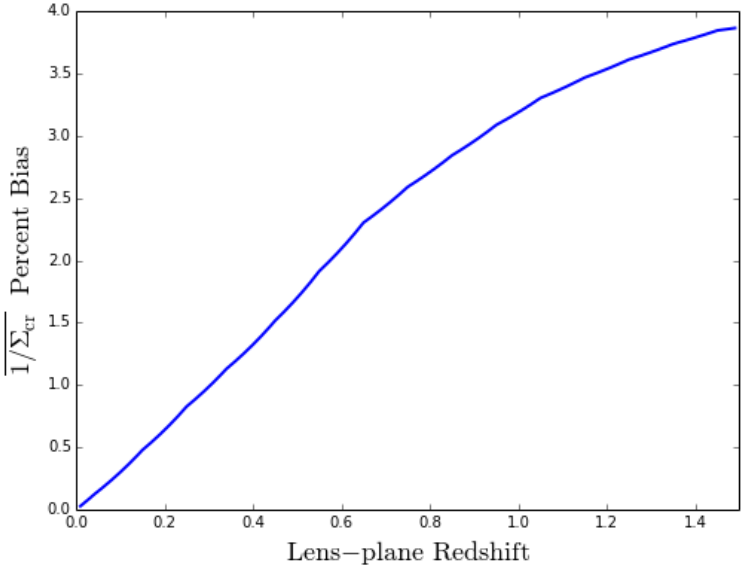




Dawson+2015

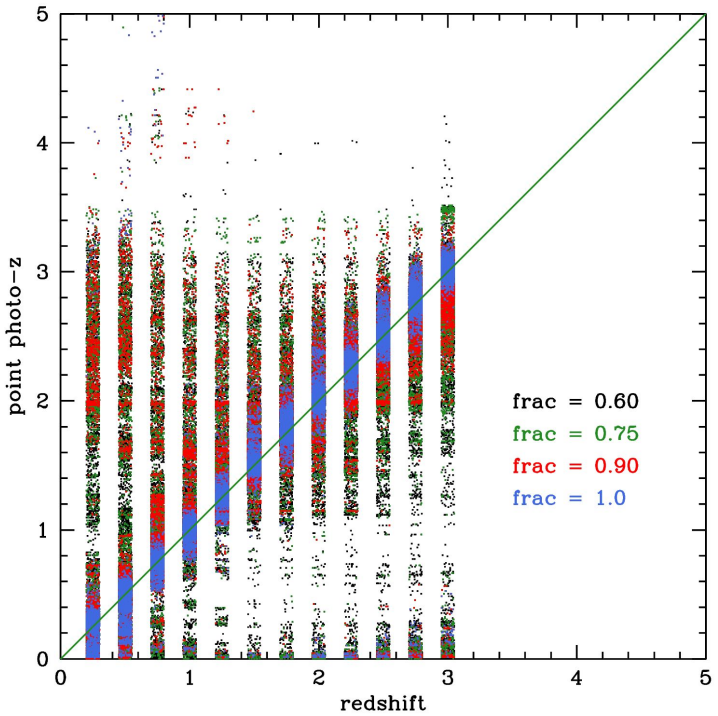
Blending and catalog cross-matching errors cause biases in the inferred redshift distribution and lensing analyses



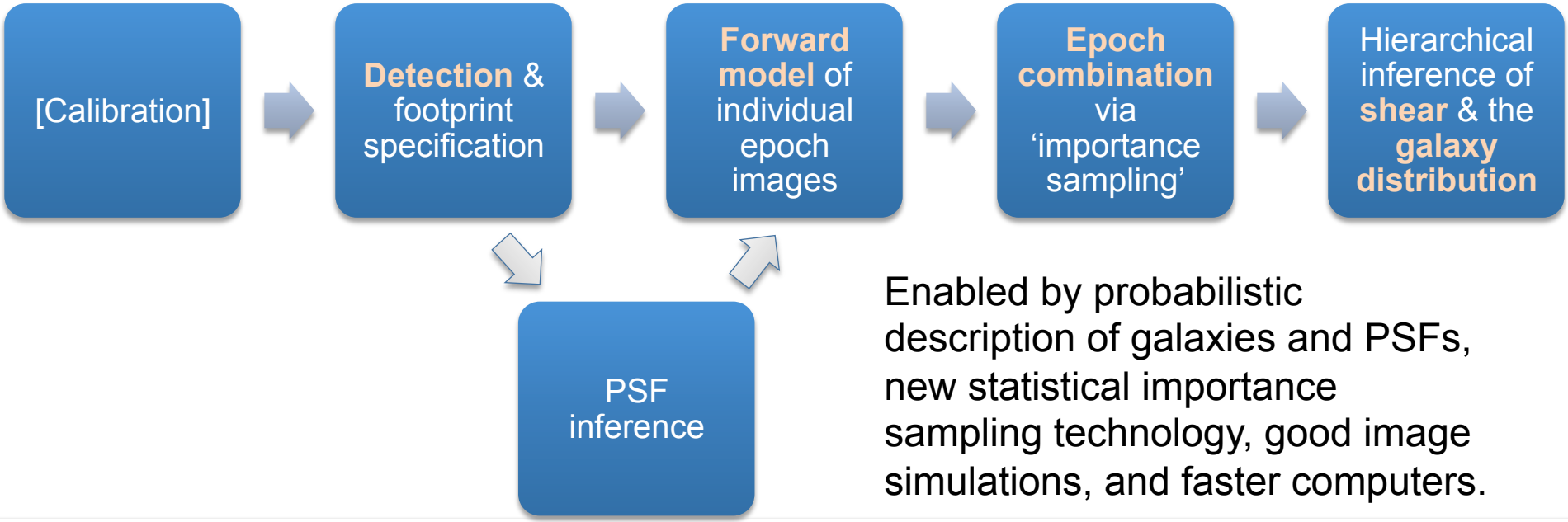


Blending of galaxies can cause spurious photo-z's for many galaxy spectral types and redshift ranges

- Galaxy pair flux fractions from 0.6 – 1.0, where 1.0 indicates a single galaxy without any blending
- Simulations on a grid of:
  - Pair redshift separation
  - SNR
  - Spectral types
- Photo-z estimates use 10 bands:
  - 6 LSST + 4 WFIRST

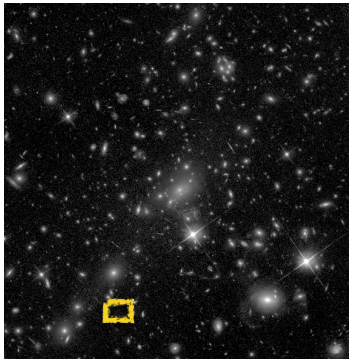


We have developed a probabilistic image reduction pipeline motivated by (a) challenges in multi-epoch/multi-telescope combinations, (b) improved shear measurements for LSST



Enabled by probabilistic description of galaxies and PSFs, new statistical importance sampling technology, good image simulations, and faster computers.

Maximum likelihood detection demonstrated on HST Frontier Fields data

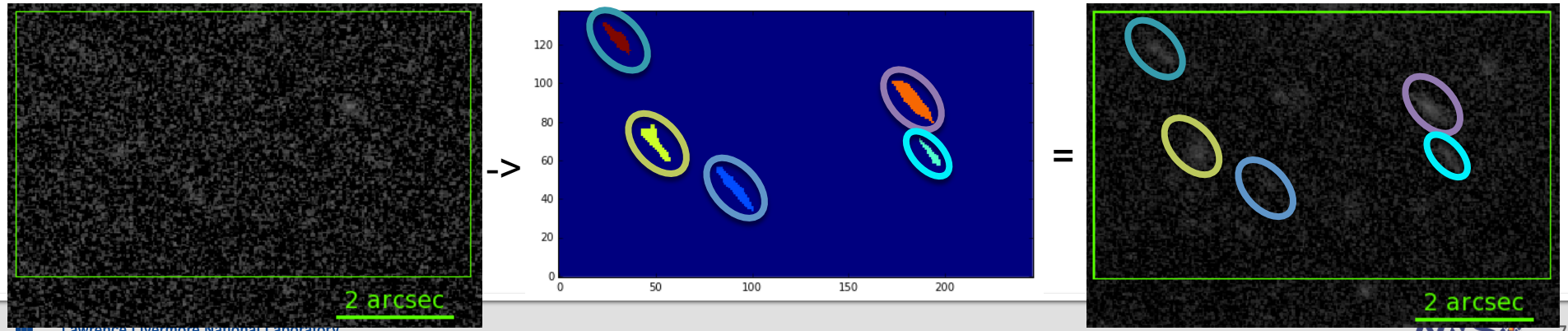


References:
Szalay, Connolly, Szokolsky (1999);
Kaiser for Pan-STARRS
Bosch for LSST

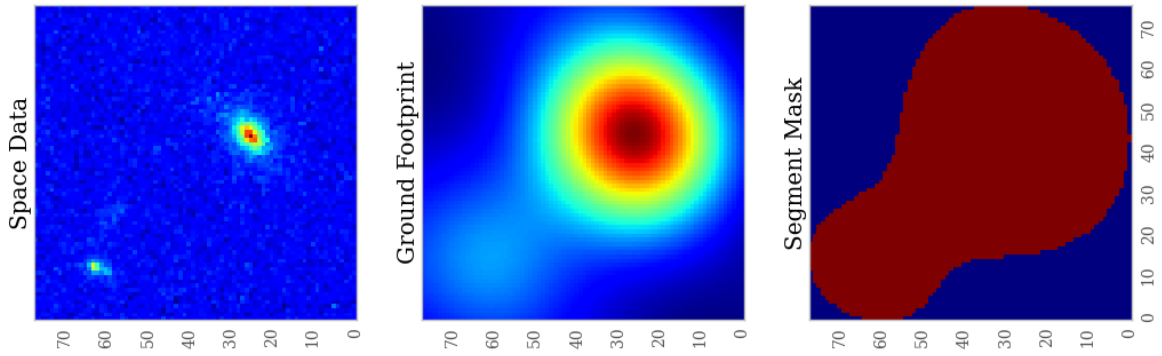
Shallow 2 orbit image.

Detections of faint high-z lensed galaxies in faint image.

Deep 15 orbit image.



Footprint specification: Define subsets of pixels such that the pixel likelihood function factors across footprints



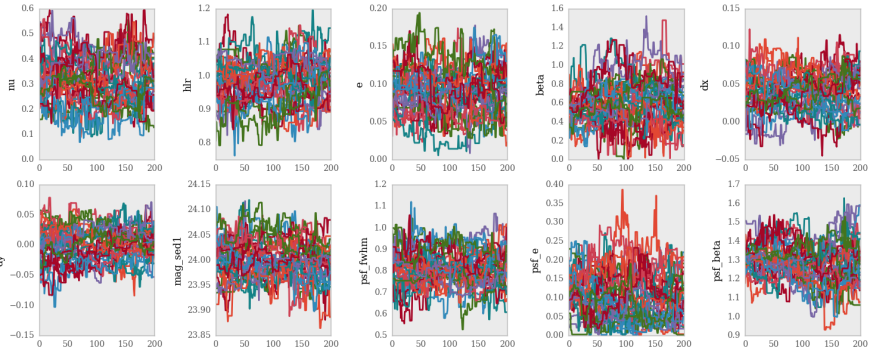
Footprint definition assumes pixel noise is uncorrelated

- The epoch with the largest PSF defines the footprint
- Challenge for updating analyses when new data is available
- Larger than necessary footprints lead to larger computing requirements

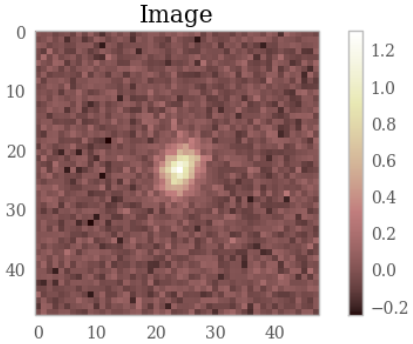
Source characterization via probabilistic image modeling

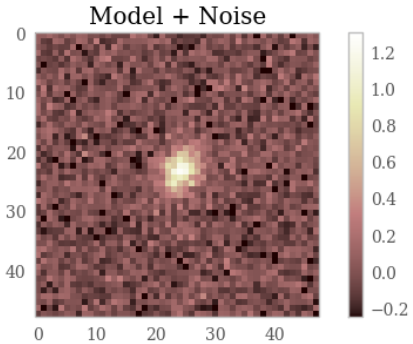
Infer image model parameters via MCMC under an interim prior distribution for the galaxy and PSF parameters.

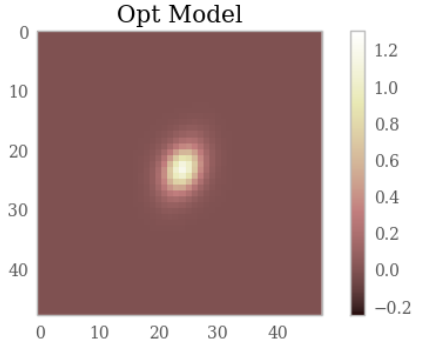
MBI GREAT3 analysis with:
The Tractor (Lang & Hogg)

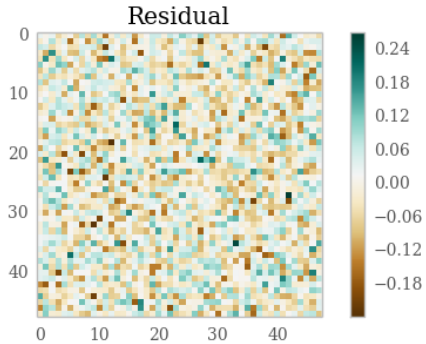


GalSim models inside an MCMC chain









Benefits of forward modeling footprint images

- Naturally handles varying pixel scales, PSFs, and (potentially) other instrument signatures
- Avoids ‘noise bias’ in galaxy shape estimates because we do not compute nonlinear transformations of the noise
- Avoids remapping noisy pixels (e.g., ‘drizzle’)
- Fits blends without noise manipulations as in ColorPro
- Optimal in principle up to ‘model fitting biases’.

Improved pipeline for forward modeling of images

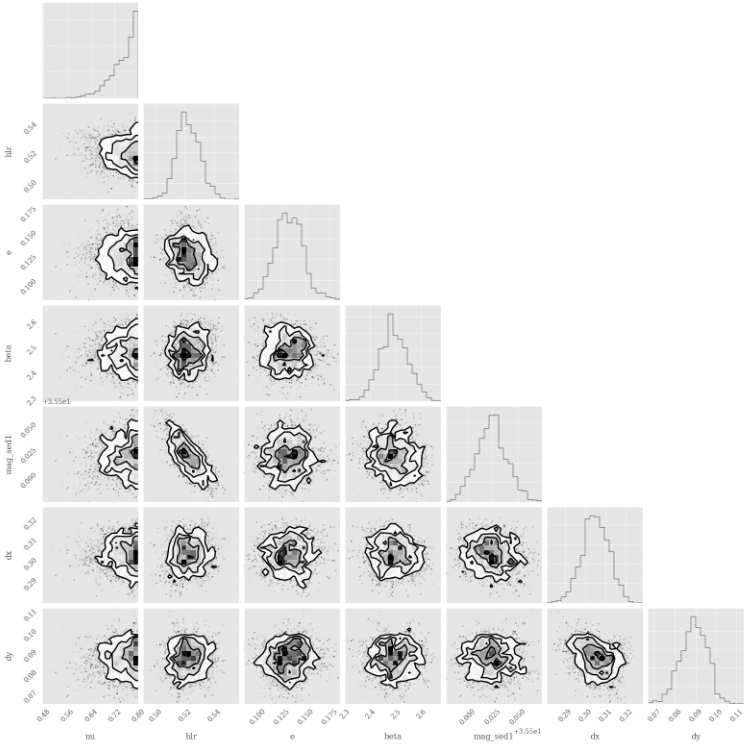
For use as a shape pipeline and deblender

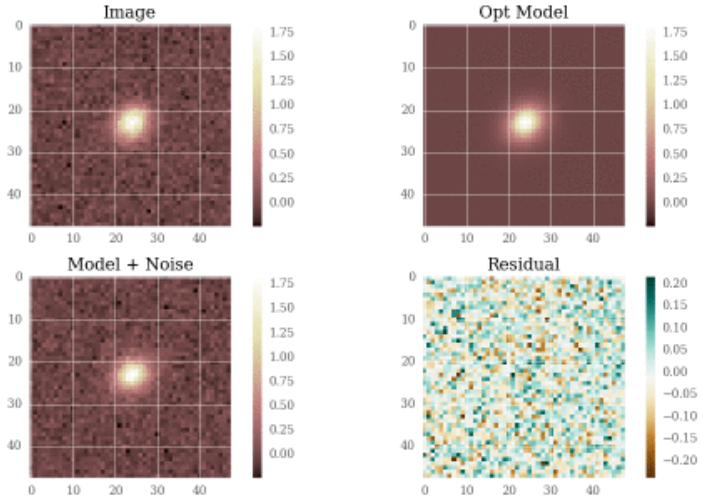
- We have a GalSim wrapper package that does image model fitting like The Tractor (Lang & Hogg)
  - Flexible model parameter selections, MCMC sampling, chromatic / achromatic models

- Recent improvements:
  - MCMC sampling optimizations,
  - Maximum likelihood optimizations (prior to MCMC),
  - Dynamic settings to improve MCMC convergence metrics,
  - Comprehensive schema for images & metadata in an HDF5 framework

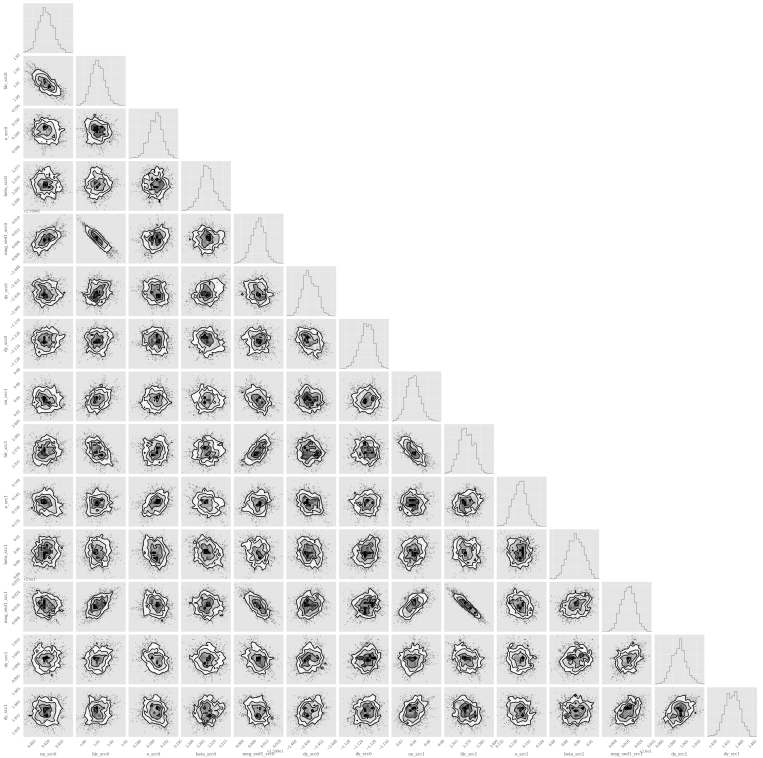
- Automated pipeline returns converged results in all cases
  - ~15 seconds per galaxy,
  - achromatic models,
  - 7 galaxy parameters – including Sersic index

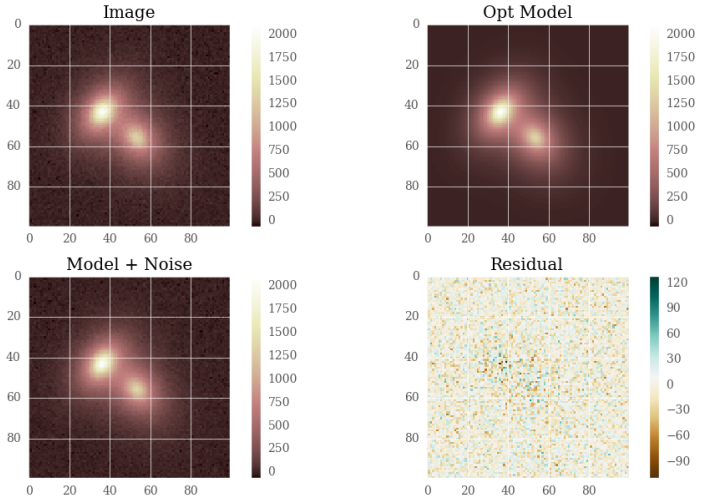
Example outputs from an automated pipeline run on GalSim simulations





Can also forward model blends by extending the parameter space





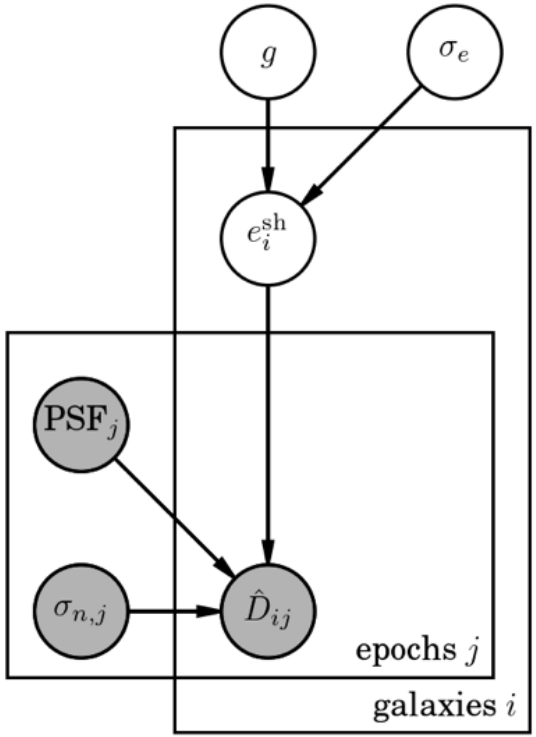
What do we do with image model posterior samples?
Think about mitigating noise bias – at least 2 strategies

1. Calibrate using simulations. (im3shape, sfit)
   - But corrections are up to 50x larger than expected sensitivity!

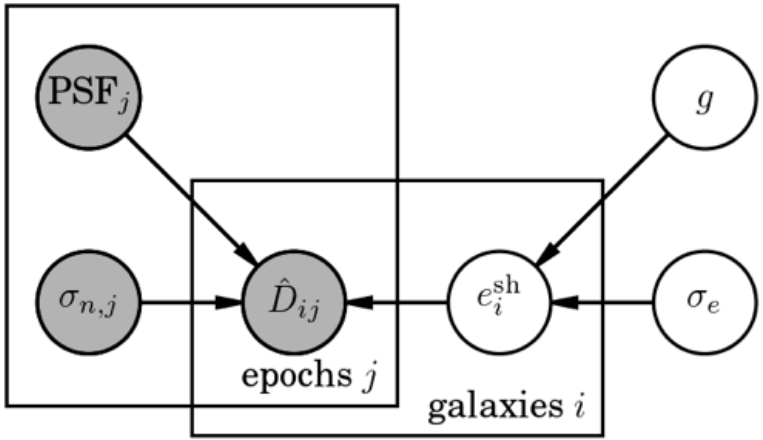
2. Propagate entire ellipticity distribution function P(ellip | data).
   - Use Bayes' theorem: P(ellip | data) $\propto$ P(data | ellip) P(ellip)
   - Measure P(ellip) in deep fields. (lensfit, ngmix, FDNT).
   - Infer simultaneously with shear in a hierarchical model. (MBI).

A hierarchical model for the galaxy distribution

- σ_e = intrinsic ellipticity dispersion
- e^{int} = galaxy intrinsic ellipticity
- g = shear
- e^{sh} = galaxy sheared ellipticity
- PSF = point spread function
- D = model image
- σ_n = pixel noise
- D = data: observed image



Our graphical model tells us how to factor the joint likelihood



- Use a probabilistic graphical model to encode the factorization of the joint probability distribution of variables in the model.
- We don't care about e^{sh} for cosmology, so integrate it out.

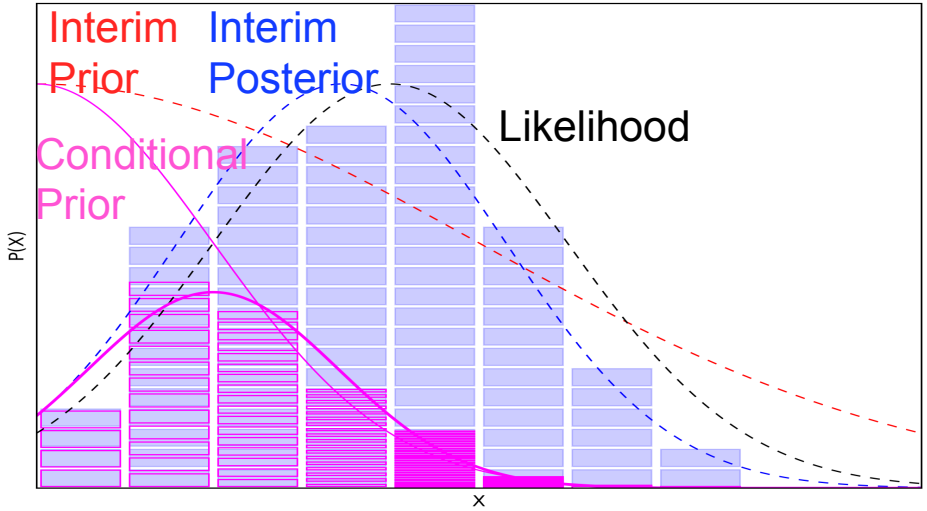
$$\Pr\left(g, \sigma_e | \left\{\mathrm{PSF}\right\}_j, \{\sigma_{n,j}, \{D_{ij}\}\}\right)$$

$$\propto \int d^{n_{\rm gal}}\left\{e_i^{\rm sh}\right\} \left[\prod_{ij} \Pr\left(D_{ij} | \mathrm{PSF}_j, \sigma_{n,j}, e_i^{\rm sh}\right)\right] \left[\prod_i \Pr\left(e_i^{\rm sh} | g, \sigma_e\right) \Pr(g) \Pr(\sigma_e)\right]$$

Huge complicated integral to compute for every posterior evaluation.

Importance Sampling to separate the fitting of individual galaxies: the pseudo-marginal likelihood

- Don’t go back to pixels for every time we sample a new g or σ_e.
- For each galaxy, draw image model parameter samples under a fixed “interim” prior. This is embarrassingly parallelizable.
- Use reweighted samples to approximate the integral via Monte Carlo.



$$\text{Draw } K \text{ samples } e_{i,k}^{\text{ab}} \sim \mathrm{P}\left(e_i^{\text{ab}} \middle| \hat{D}_i, I_0\right) \propto \mathrm{P}\left(\hat{D}_i \middle| e_i^{\text{ab}}\right) \mathrm{P}\left(e_i^{\text{ab}} \middle| I_0\right)$$

Importance sampling to separate the sampling of individual galaxies: The pseudo-marginal likelihood

Want:

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

Galaxy dist. Likelihood

Have samples from:

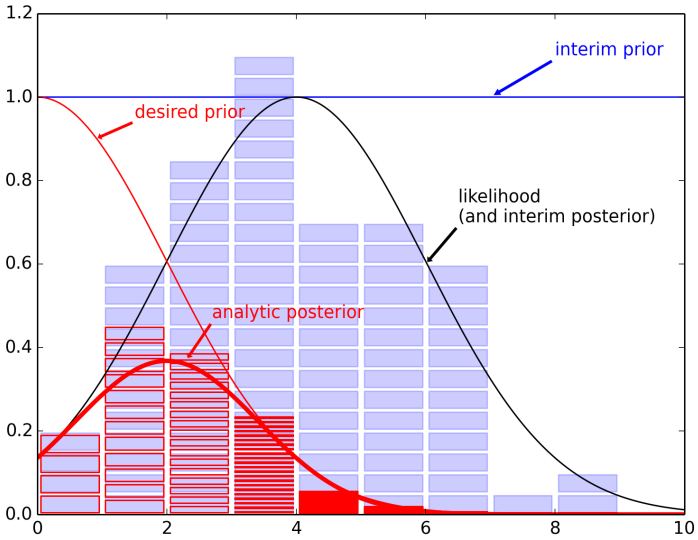
$$\Pr(\omega_n|\mathbf{d}_n, I_0)$$

'Interim prior' specification

Importance sampling:

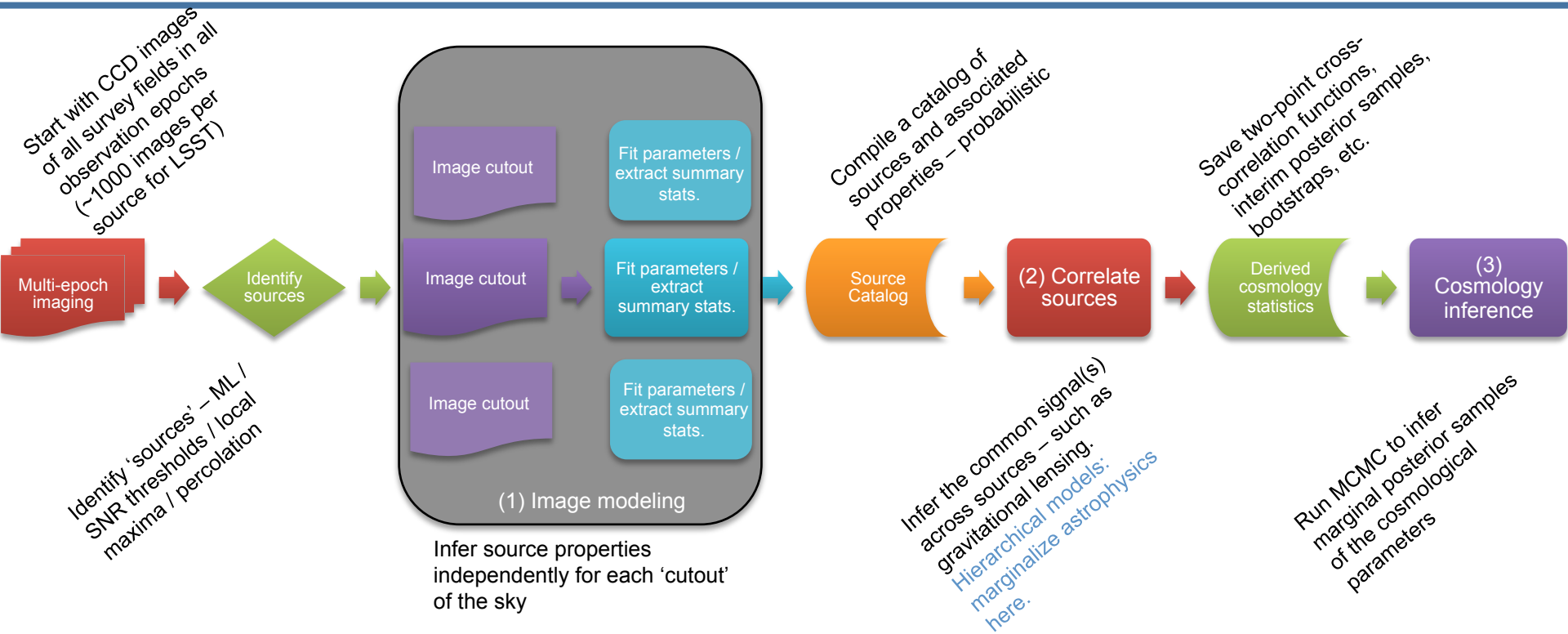
$$\Pr(\mathbf{d}_n|\alpha) \approx \frac{Z_n}{K} \sum_k \frac{\Pr(\omega_{nk}|\alpha)}{\Pr(\omega_{nk}|I_0)},$$

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



Credit: J. Meyers

The end-to-end simulation and analysis pipeline (almost): 3 levels of model inference



How do we combine multiple observations of the same galaxy?
Naïvely we must joint fit all epochs simultaneously

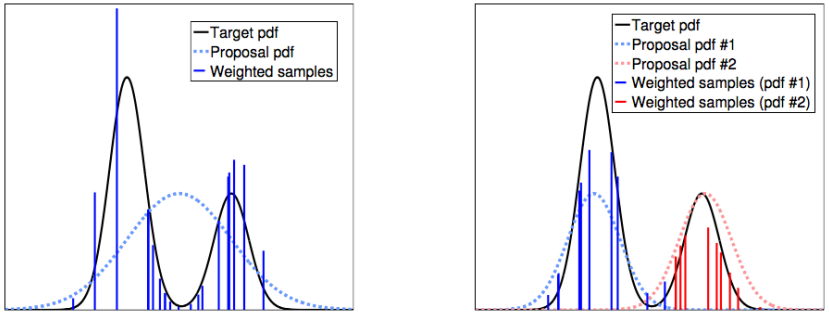
Problem: Imagine we have fit pixel data from LSST years 1 – 4.
How do we incorporate WFIRST observations without redoing (expensive) calculations?

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

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

$$q(\omega_n) = \frac{1}{n_{\rm epochs}} \sum_{i=1}^{n_{\rm epochs}} \Pr(\omega_n|\mathbf{d}_{n,i},\Pi_i,I_0)$$

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



(a) Single proposal pdf (standard IS). (b) Two proposal pdfs (MIS).

Fig. 1: Approximation of the target pdf, $\pi(\mathbf{x})$, by the random measure χ.

Generalized multiple importance sampling (MIS) weights

MIS sampling distribution: sample from the conditional posterior for each epoch individually

$$q(\omega_n) = \frac{1}{n_{\text{epochs}}} \sum_{i=1}^{n_{\text{epochs}}} \Pr(\omega_n | \mathbf{d}_{n,i}, \Pi_i, I_0)$$

MIS weights: Evaluate the ratio of the conditional posterior for each epoch i to that of the MIS sampling distribution

$$w_i = \frac{\Pr(\mathbf{d}_{n,i}|\omega_n, \Pi_i)\Pr(\omega_n|\alpha)}{\sum_{i=1}^{n_{\text{epochs}}} \Pr(\mathbf{d}_{n,i}|\omega_n, \Pi_i)\Pr(\omega_n|I_0)}$$

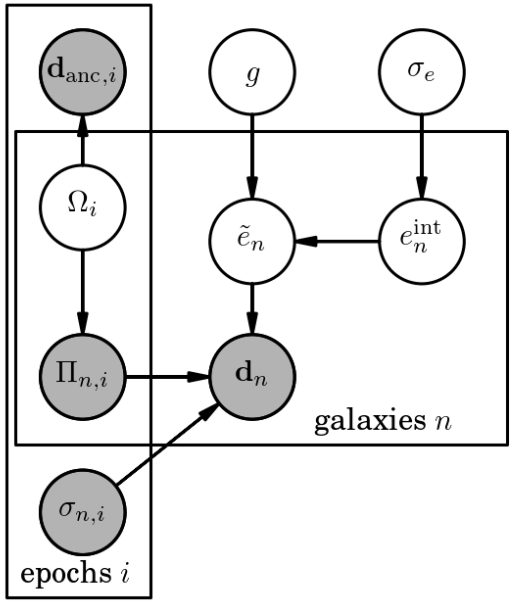
'*cross-pollination*' needed:
Evaluate the likelihood of epoch i given model parameter samples from epoch j, for all combinations of i, j.

A standard scatter / gather operation

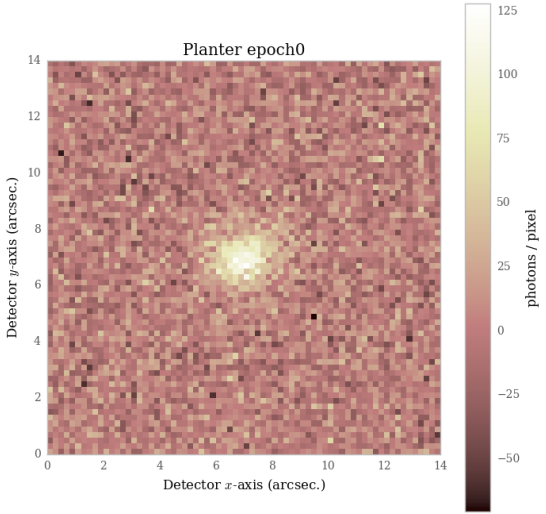
$$\Pr(\mathbf{d}_{n,i}|\omega_n^{(j)}\Pi_i)$$

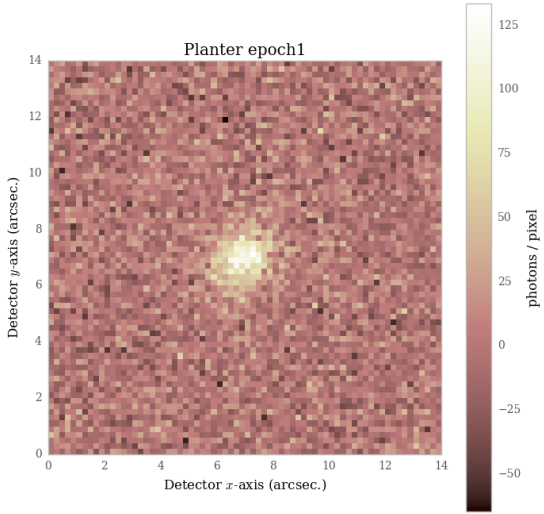
Marginalizing PSFs: MIS makes this tractable

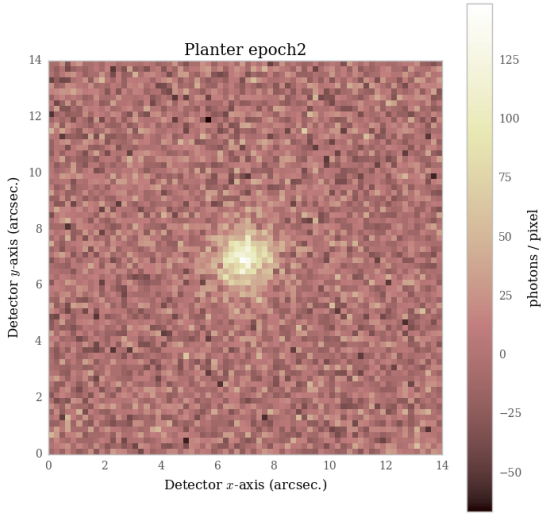
- LSST will have ~200 epochs per object per filter
  - We aim to marginalize the PSF $\prod_{n,i}$ in every epoch
  - The marginalization is constrained by:
    - Consistency of PSF realizations over the focal plane for each epoch
    - Consistency of the underlying source model across epochs
- Simplest 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!
- *Challenge*: PSF correlates inferences across sources in an image, but we want to fit images individually



Example: 1 galaxy, 3 epochs – fit the galaxy model parameters

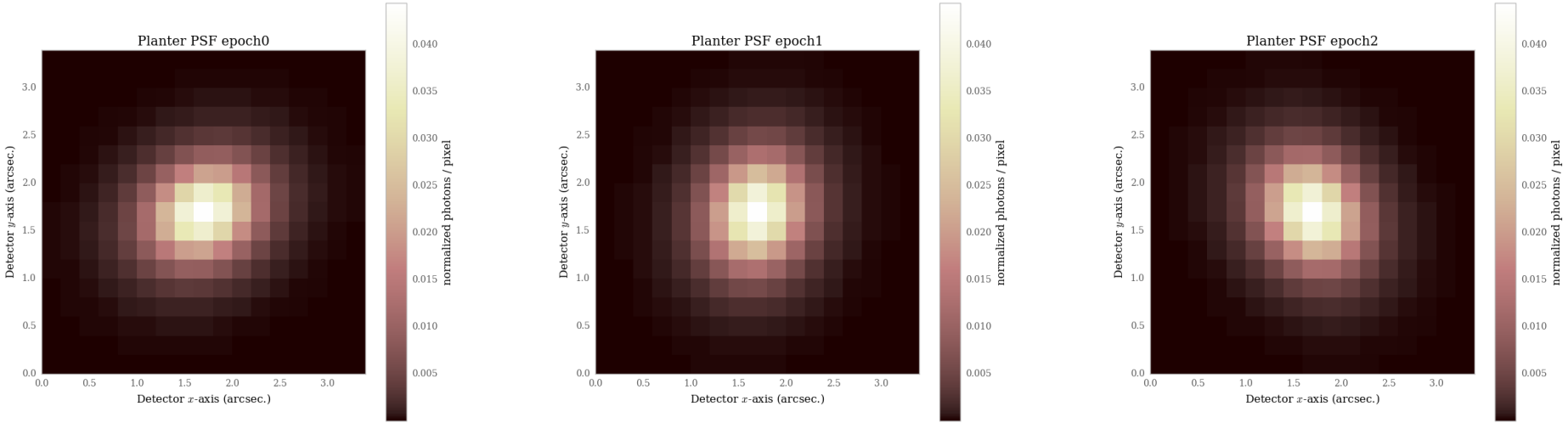






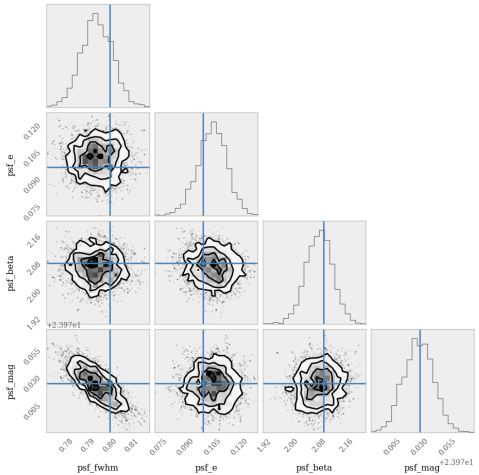
Each epoch has highly elliptical PSFs (|e| = 0.1) of same size, but different orientations

The PSF FWHM also matches the galaxy HLR making the single-epoch inferences noticeably different from each other. There is therefore a large gain of information in combining epochs.

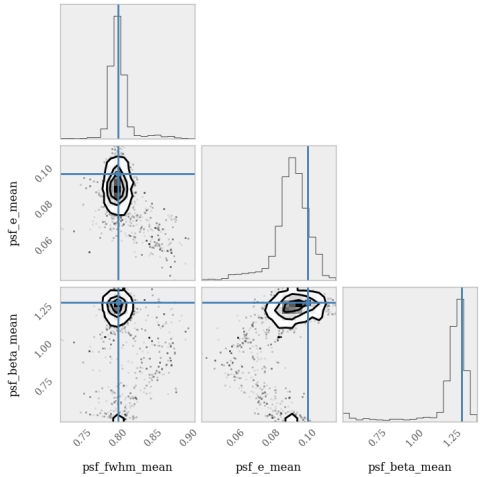


Interim posterior samples at each stage of the PSF hierarchical model

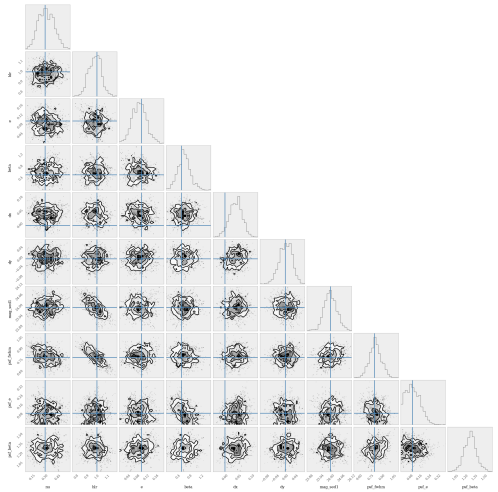
1) Fit stars



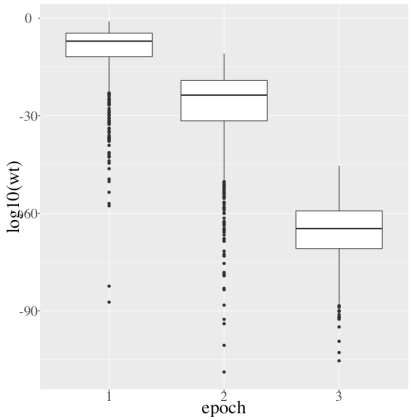
2) Constrain PSF model



3) Fit galaxies & PSFs

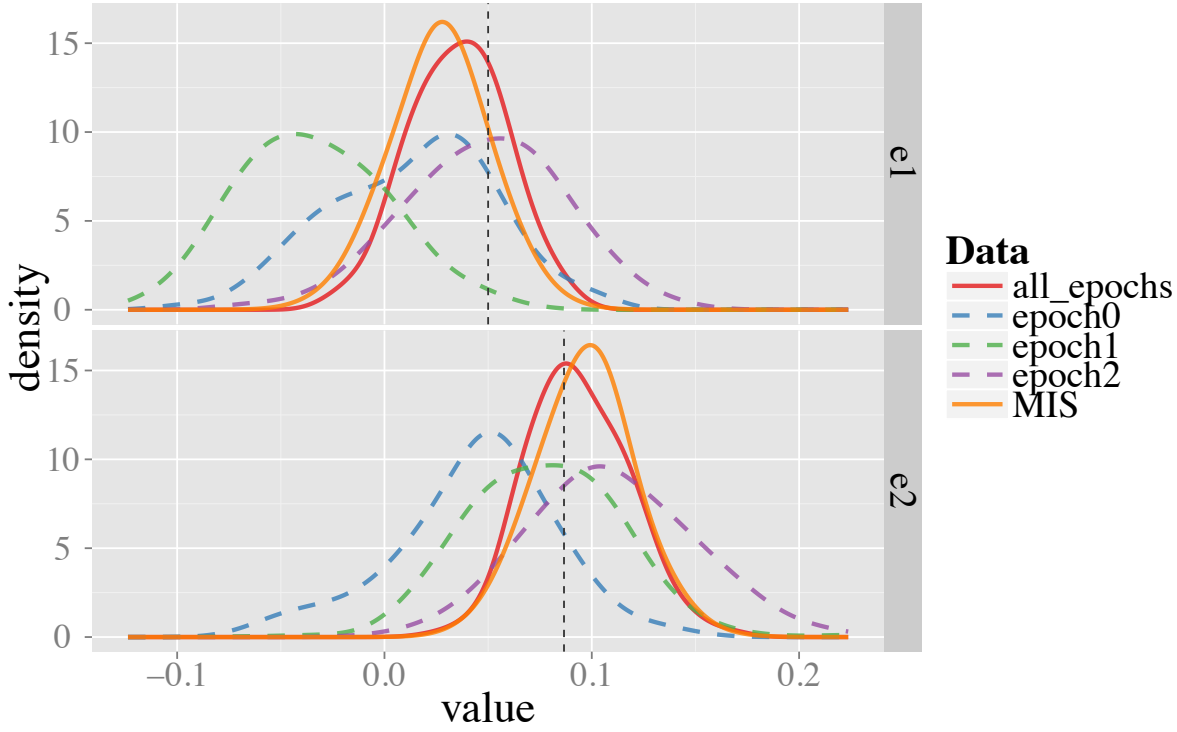


4) Calculate MIS weights to combine epochs

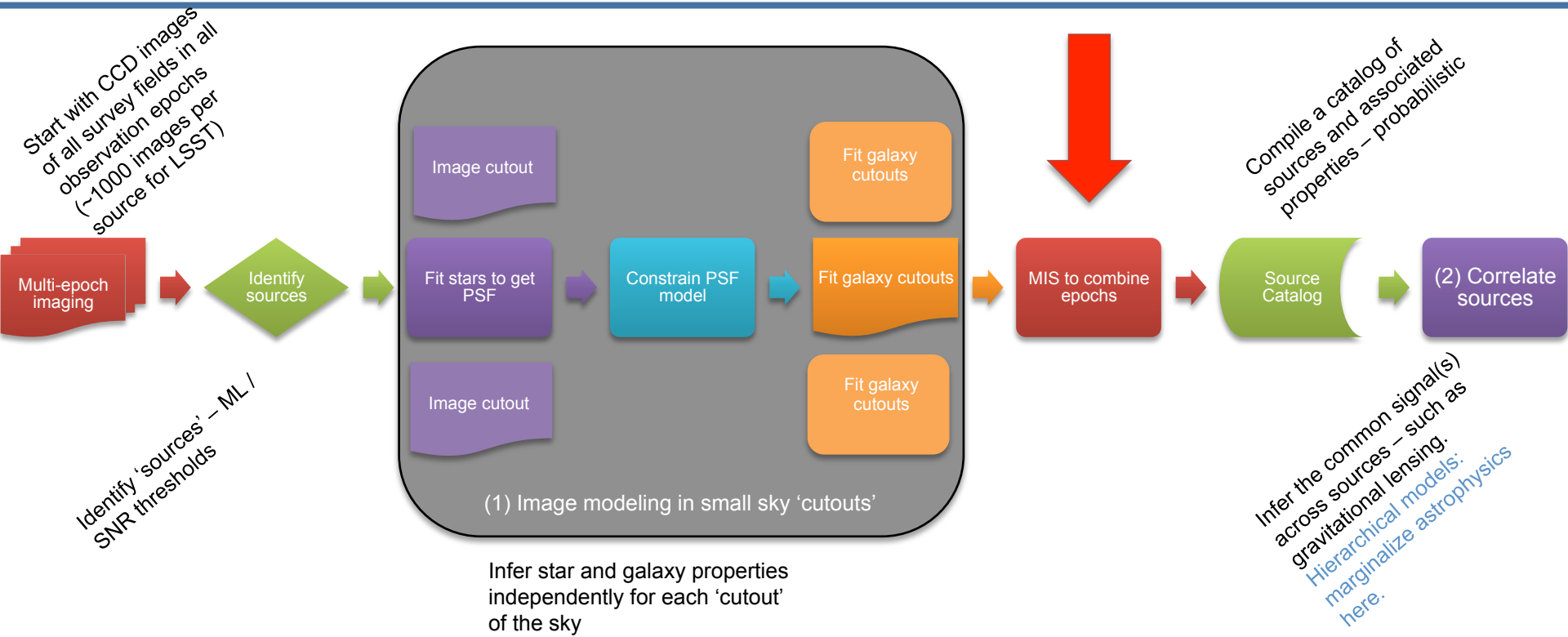


DESC *PSF Task Force*

Comparison of single-epoch and combined epochs marginal posteriors



Simulation and analysis pipeline: MIS-enabled



‘Cross-pollination’ benefits and challenges

- Individual epochs and telescope images can be fit independently
  - Maybe components of a ‘footprint’ as well?
- Allows tractable PSF marginalization
  - Use more stars for a better PSF model
  - Reduce uncertainties in combining images with very different PSFs
- Allows multi-band photometry
  - A role for the SOM & modified photo-z inference?
- *Challenge*: streaming approaches and sample sparsity

Getting color information from importance sampling fits to different passbands

$$\Pr(\omega, \{m\}_i | \{\mathbf{d}\}_i, I) \propto \left[\prod_{i=1}^{n_{\rm epochs}} \Pr(\mathbf{d}_i | \omega, m_i) \right] \Pr(\omega, \{m\}_i | I)$$

$$\Pr(\omega, C | \{\mathbf{d}\}_i, I) \propto \left[\prod_{i=1}^{n_{\rm epochs}} \int dm_i \Pr(\mathbf{d}_i | \omega, m_i) \Pr(m_i | I) \Pr(C | m_i) \right] \Pr(\omega | I)$$

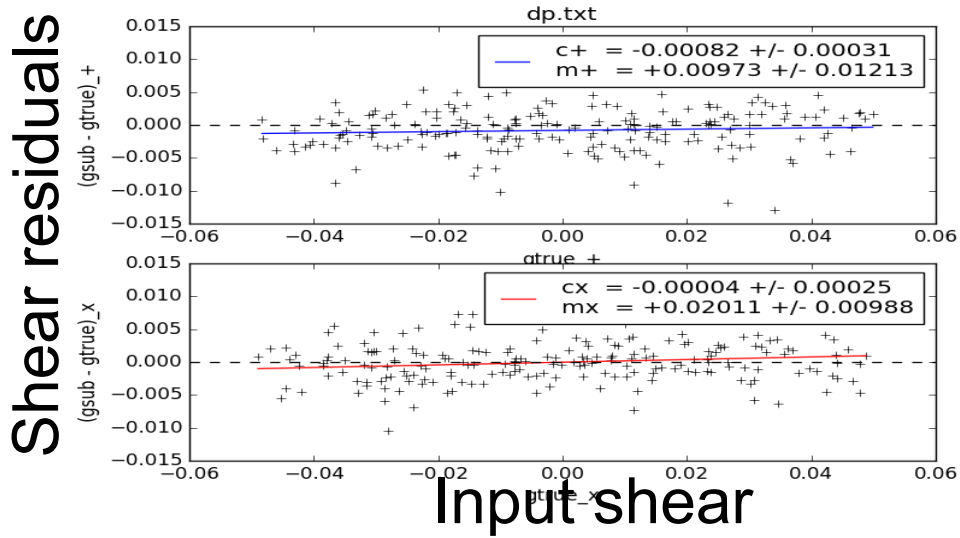
"Prior" on the colors given a single-band model amplitude. Introduces noise in principle, but likely adds needed model flexibility.

1. Draw interim samples of ω, m_i for band i
2. Repeat step 1 for all bands individually
3. For each band i and interim sample k, draw samples of multi-band colors C_k given $m_{i;k}$
4. Evaluate the MIS weights for each sample $i \times k$

Hierarchical shear inference demonstrated with GREAT3

- Tested hierarchical approach using simulations from the third GRavitational lEnsing Accuracy Test (GREAT3).
- Hierarchical inference performs significantly better than ensemble average maximum likelihood ellipticity.
- The DPMM ellipticity prior performs better than the single Gaussian ellipticity prior.

Dirichlet Process Inference



<ML> : 13% shear calibration errors

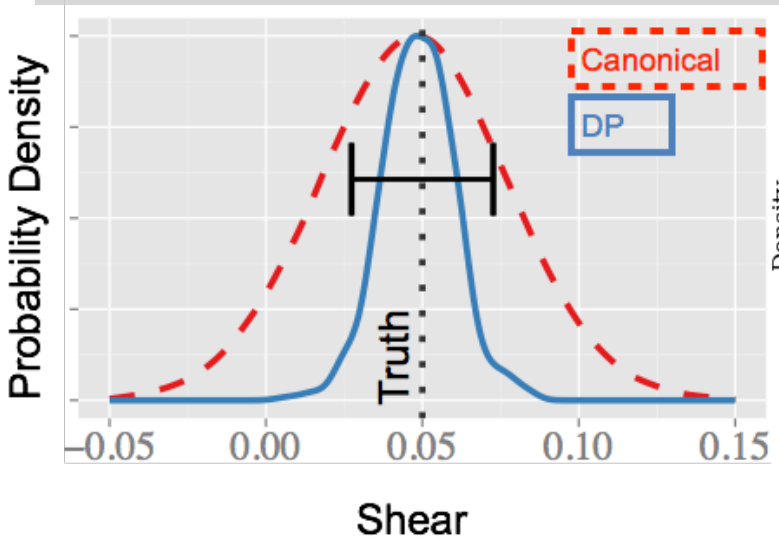
H.I. : 4% shear calibration errors

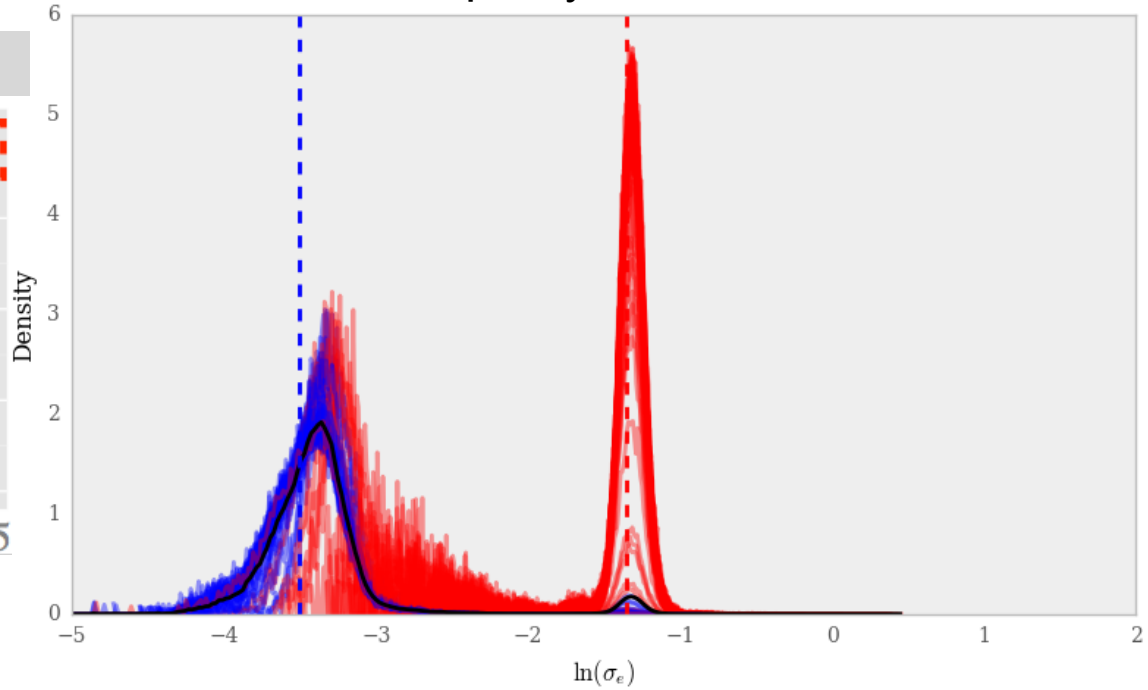
DP : 1-2% shear calibration errors

Simulation study: We can beat the traditional 'shape noise' statistical error bound by inferring latent structure in the data

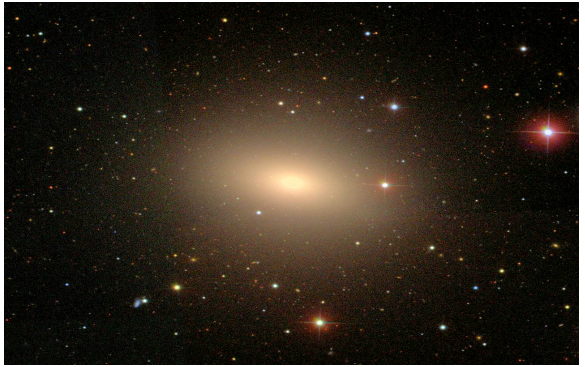
100 galaxies drawn from 1 of 2 Gaussian ellipticity distributions

3x improvement in cosmic shear precision





Multi-variate DP mixture model (in progress): “standardizable” ellipticities.



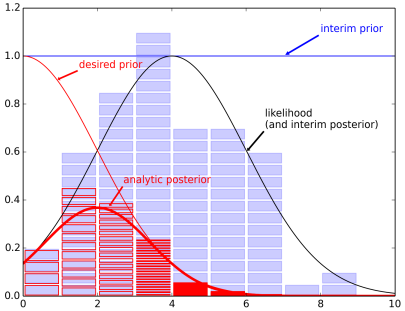
- Elliptical galaxies have a narrower intrinsic ellipticity distribution than late-type. Higher sensitivity to shear!
- Ellipticals/spirals also distinguishable by color and morphology (e.g., Sersic index, Gini coefficient, asymmetry), potentially providing additional variables with which to cluster.
- Other correlations to exploit?

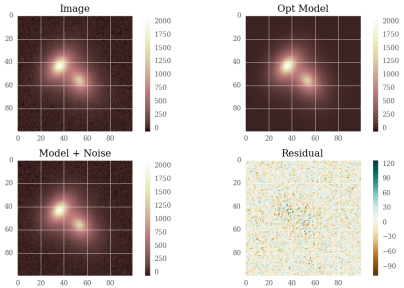
Summary

- 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 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
  - Required given the ambiguous cross-matching and different detectors and PSFs

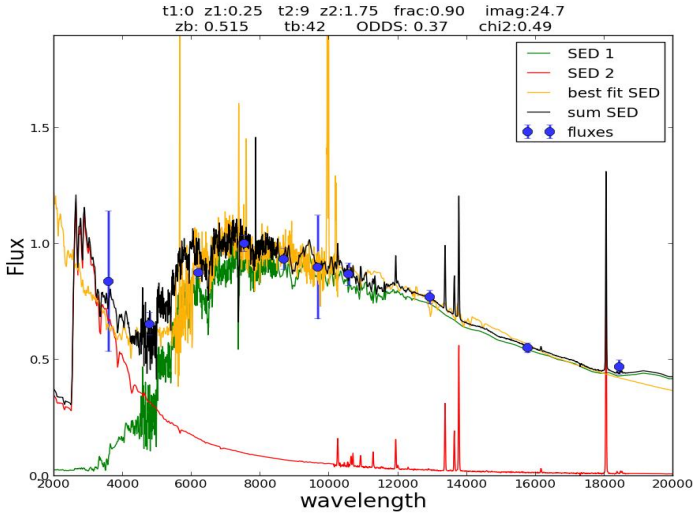
- Computation requirements: 15 sec / galaxy * 5e9 galaxies * 1e3 epochs ~ 20 billion cpu-hours (on 2015 processors).
  - Cross-pollinating & hierarchical modeling are sub-dominant in computing time

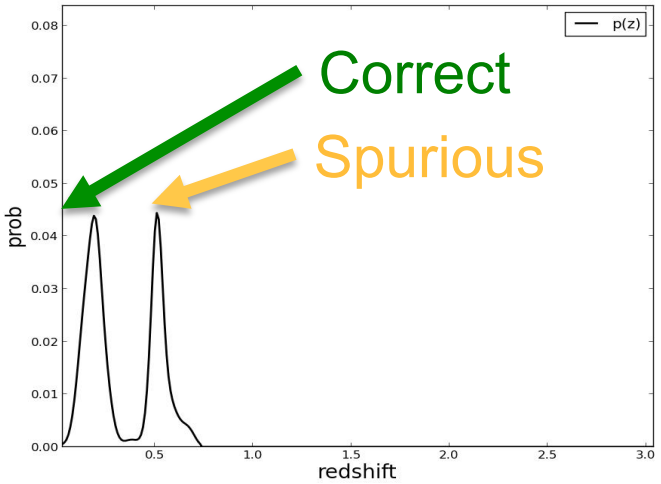





Lawrence Livermore
National Laboratory

Example: Photo-z failure for a blended pair of galaxies





i-band

90% flux: *z* = *0.25* elliptical

10% flux: *z* = *1.75* starburst

Confused with *z* ~ *0.5* spiral

Get the correct low-*z* result with ~50% probability.
The high-*z* starburst galaxy is “lost”.