Improving Photometric Redshifts for Hyper Suprime-Cam (HSC) with Hierarchical Bayes and Machine Learning

Josh Speagle[1,2,3,†,*], Alexie Leauthaud[4,2], Daniel Eisenstein[1], Kevin Bundy[4,5,2], Peter Capak[6], Boris Leistedt[7], Daniel Masters[6], Daniel Mortlock[8], Hiranya Peiris[9,10], and the HSC photo-z team

[1]Harvard University, [2]Kavli IPMU, [3]CREST, JST, [4]UC Santa Cruz, [5]UC Observatories, [6]IPAC, [7]NYU, [8]Imperial College London, [9]University College London, [10]Oskar Klein Centre for Cosmoparticle Physics

[†]NSF Graduate Research Fellow, [*]jspeagle@cfa.harvard.edu

Our code can be found online at github.com/joshspeagle/frankenz.

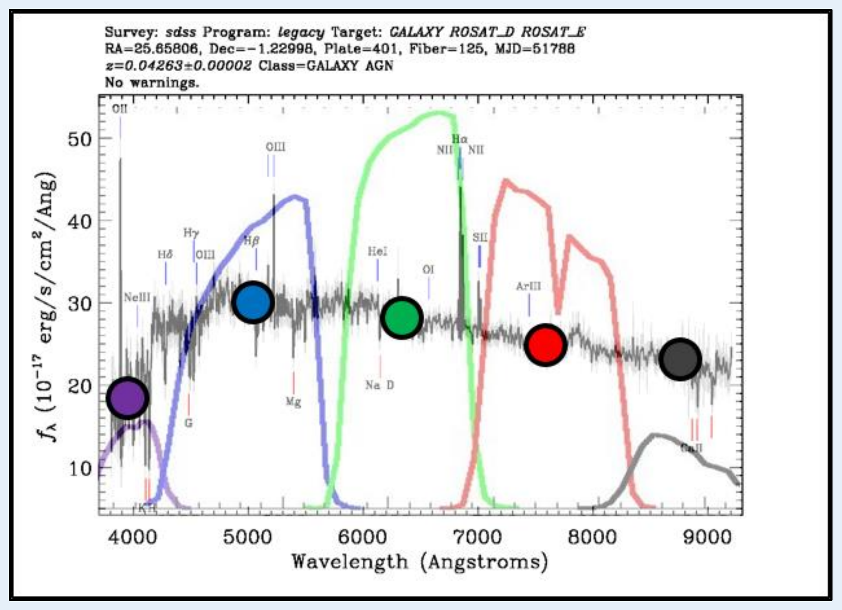
Overview

We combine Bayesian inference with machine learning to derive photo-z PDFs in a robust, non-parametric, data-driven way. This involves:

- using training data to construct a set of **empirical models**,
- exploiting **machine learning** to generate rapid likelihoods,
- improving posterior calculations to better incorporate **observational errors** and **selection effects**, and
- incorporating **hierarchical modeling** based on Leistedt et al. (2016) [arxiv:1602.05960]

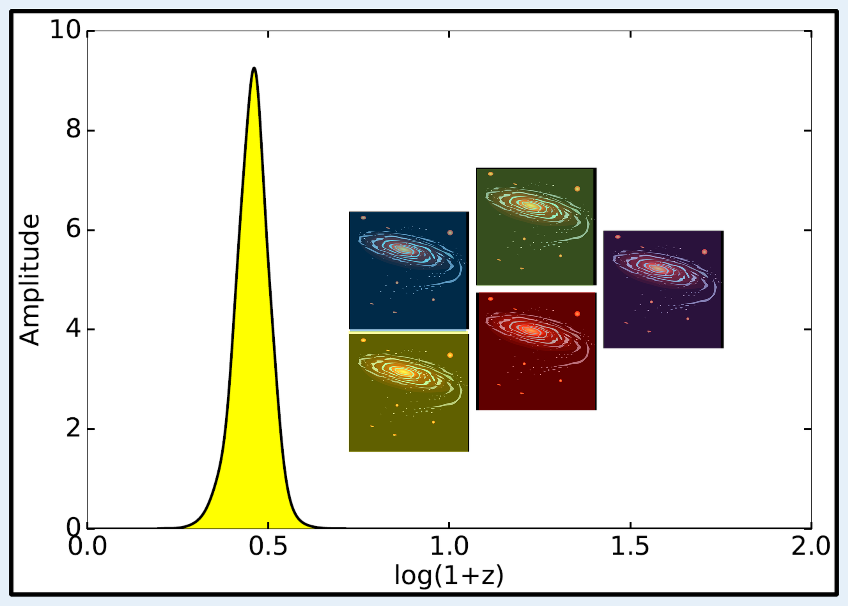
What are photometric redshifts?

Large surveys observe galaxies in several **photometric bands**, giving a rough estimate of its underlying **spectral energy distribution (SED)**.

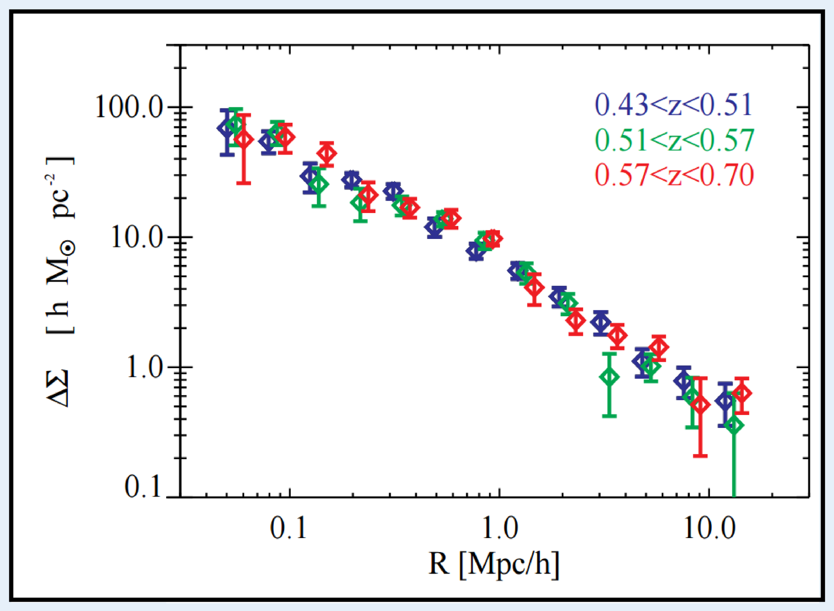


Sample SDSS spectrum from skyserver.sdss3.org with corresponding filters taken from astro.caltech.edu/~capak/.

We want to use this information to derive a galaxy's **photometric redshift (photo-z) probability density function (PDF)**.



Using these photo-z PDFs, we can do lots of cool science with large galaxy datasets!



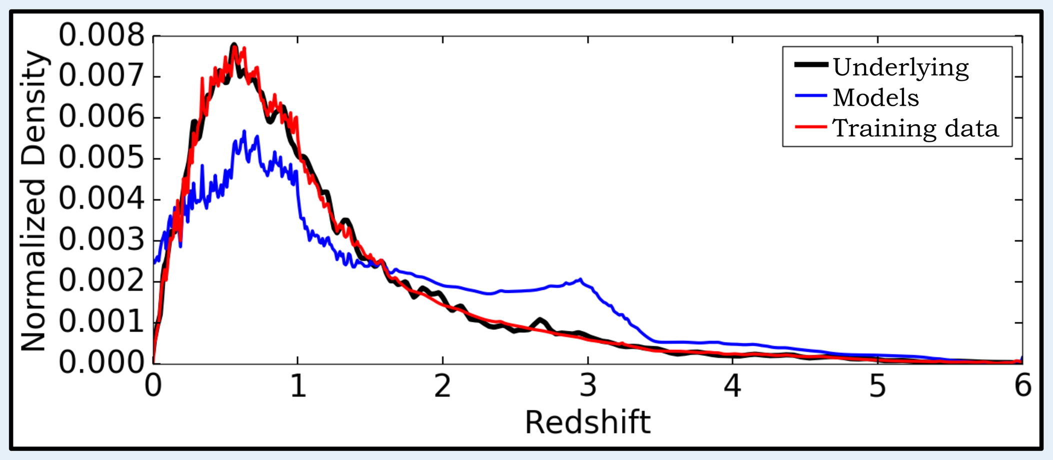
BOSS CMASS galaxy-galaxy lensing measurements from Leauthaud et al. (2016) [arxiv:1611.08606].

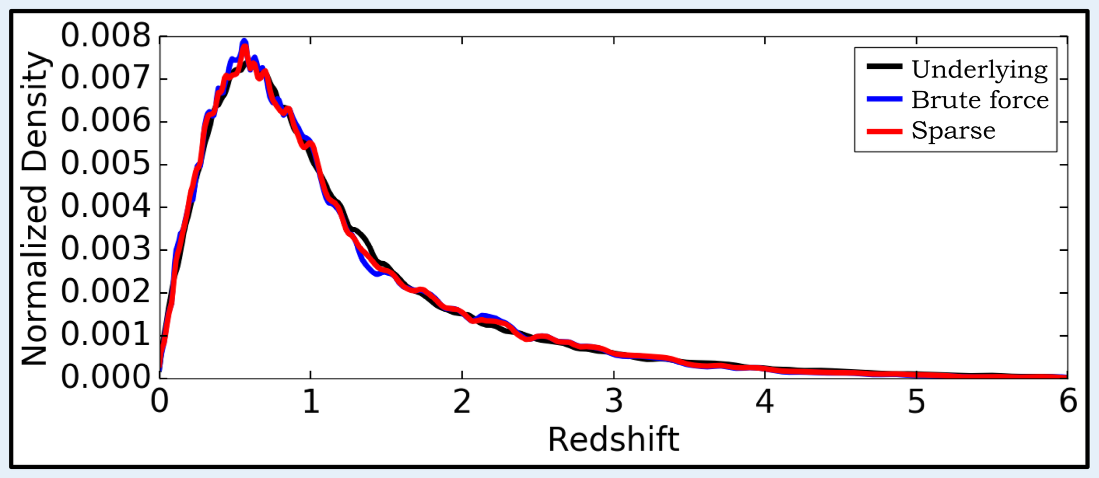
Fast, Probabilistic, Data-Driven Redshifts

Empirical models

We use **training data** instead of spectral templates as the basis for our predictions.

Right: Performance on **mock HSC data** (black) from **models** (blue) and **training data** (red). Using training data better captures the complex ways galaxies evolve over time.





Fast predictions

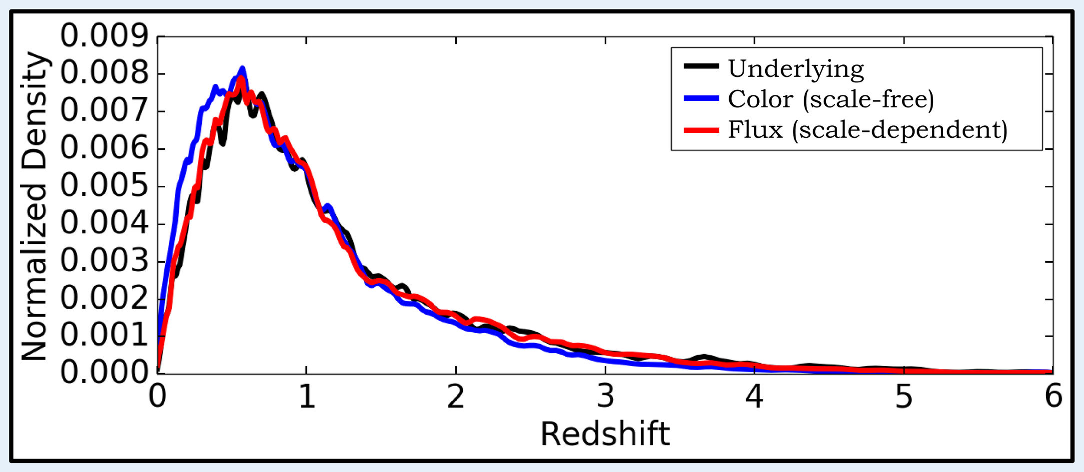
Using **machine learning**, we can quickly generate **sparse approximations** to the likelihood to take advantage of large training datasets.

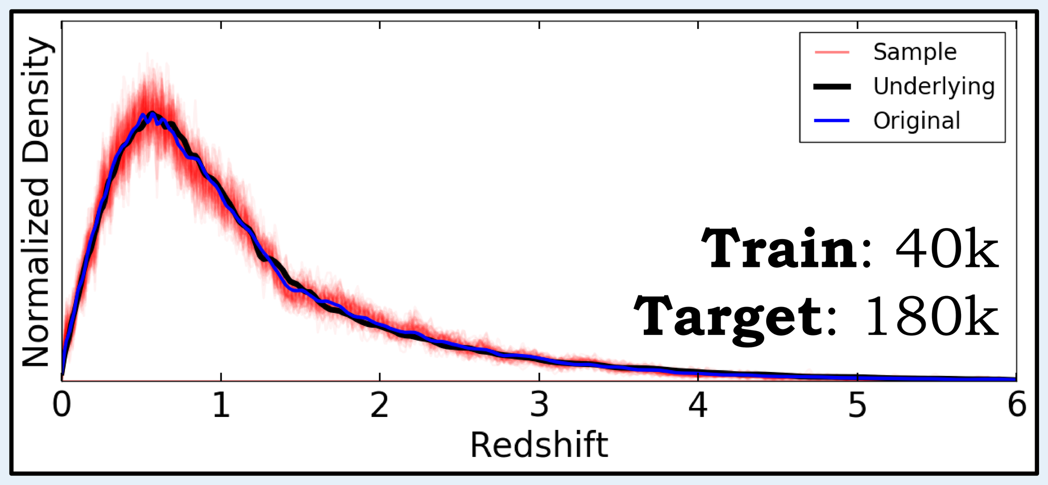
Left: As above, but comparing likelihoods computed using **all** training objects (blue) vs a **small (~1%) subset** (red). The differences between the two are negligible.

Improved likelihoods

By taking advantage of large amounts of training data, we can compute our likelihoods using **flux** rather than **color**, which better incorporates **measurement errors** and **complex priors**.

Right: As above, but for likelihoods from **color** (blue) vs **flux** (red). Fluxes give less biased redshift estimates.





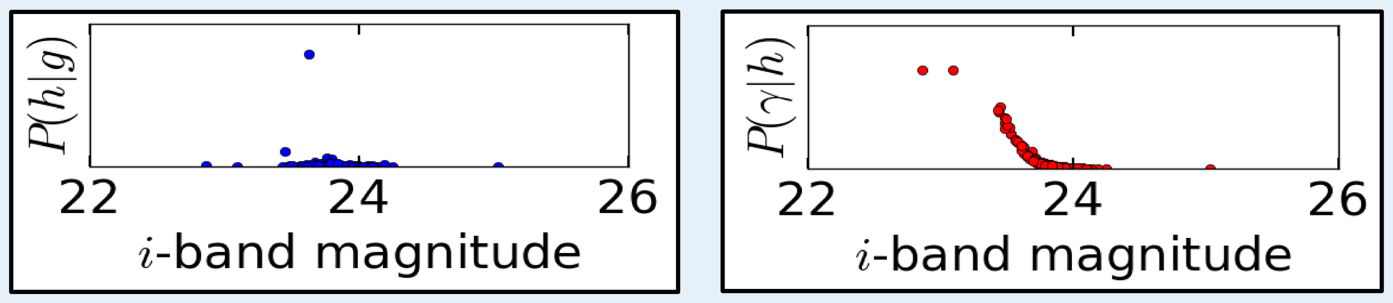
Hierarchical inference

We explore the joint distribution between individual objects and the overall population using **hierarchical Bayesian modeling**.

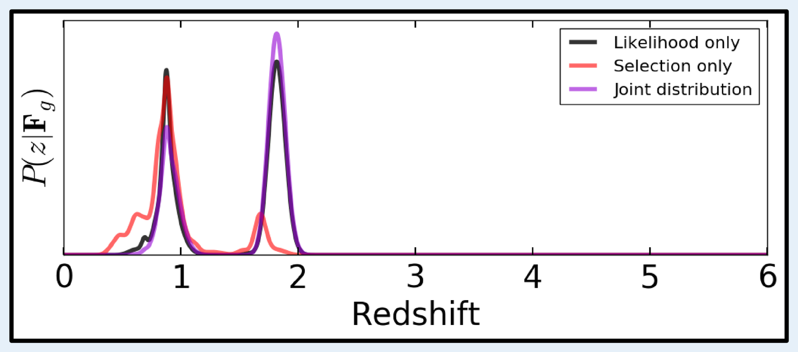
Left: As above, but for our **hierarchical** (red) vs **original** redshift estimates (blue). Our hierarchical model better captures uncertainties in our overall redshift estimates.

Selection Effects

Our Bayesian formalism allows us to incorporate the **selection function** directly into our posterior predictions for individual objects.



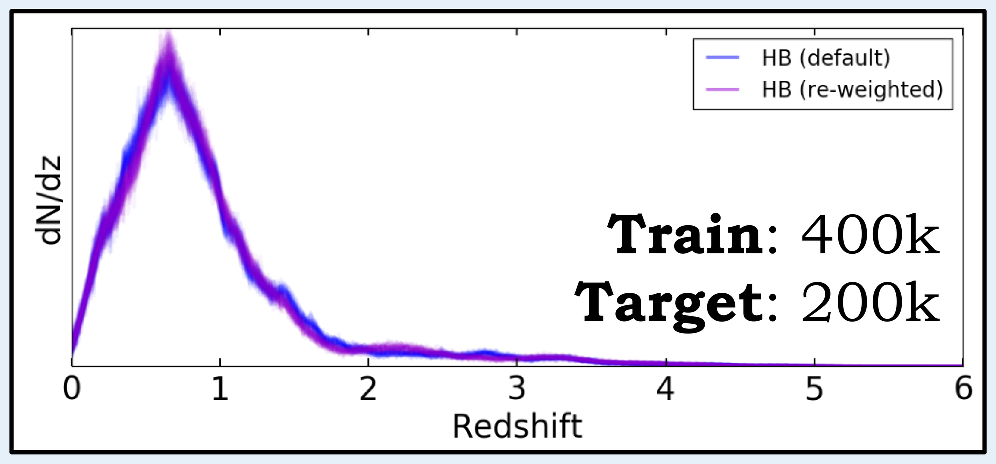
Original **likelihood** (left) and **selection probability** (right) for an object from the HSC Wide Survey as a function of magnitude.



PDFs computed using the **likelihood** (black), **selection function** (red), and **posterior** (purple).

Preliminary Results

Using **~400k training galaxies** from 11 different surveys, we computed redshifts for **~200k weak lensing-selected HSC galaxies** for **uniform** (blue) and **type-dependent** (purple) population priors.



Our redshift distributions are largely insensitive to our choice of prior.