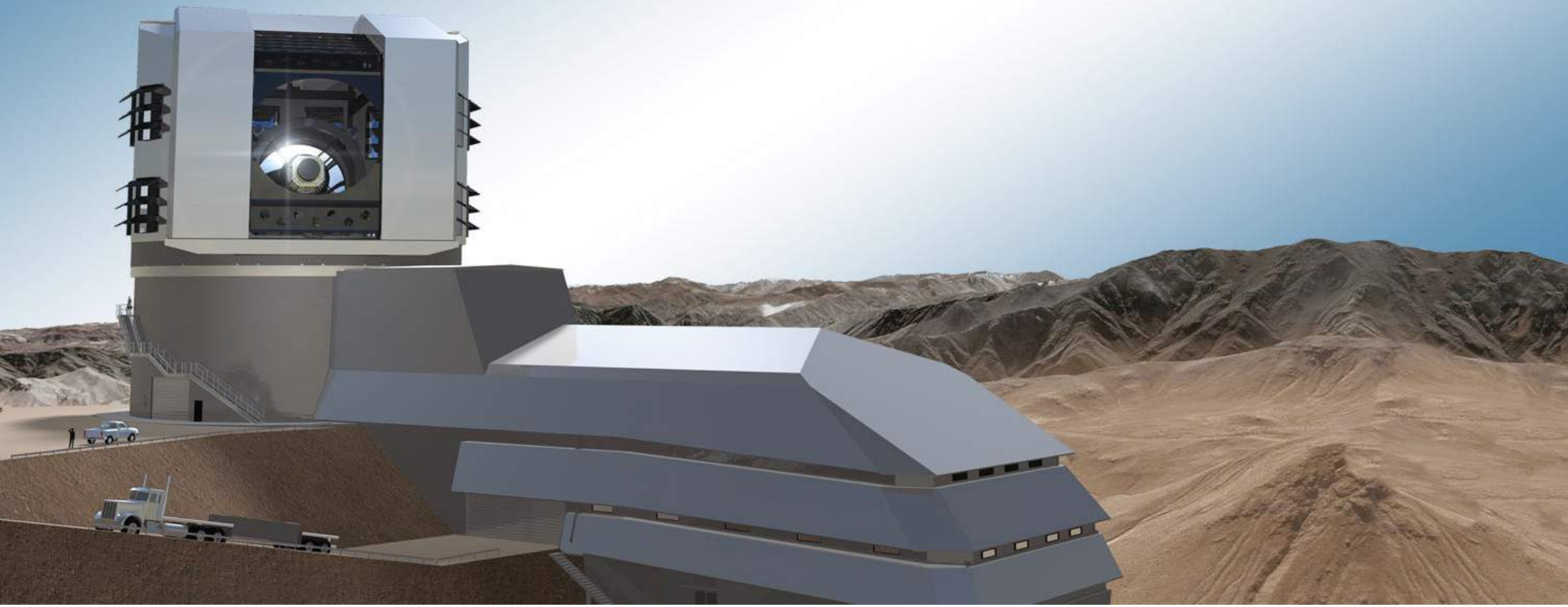
Photometric Redshifts for LSST

Jeffrey Newman, U. Pittsburgh / PITT-PACC

Follow-up Task Force Co-convener, LSST Dark Energy Science Collaboration



Outline

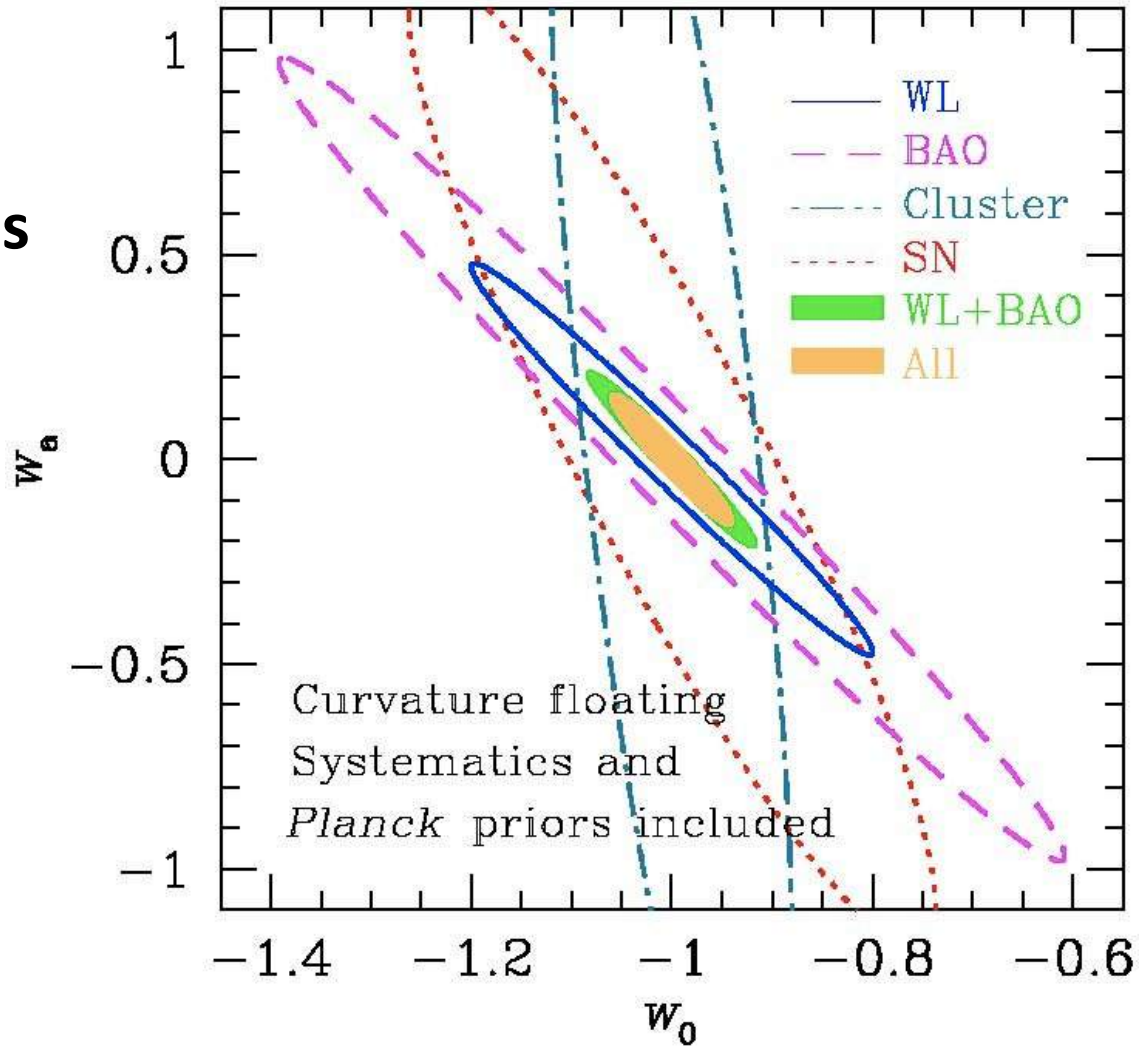
- Overview of photometric redshifts
  - Template methods
  - Training-based methods
- Requirements and resources for training and calibrating photometric redshifts
- Some open issues
  - Spectroscopic incompleteness
  - Robust training
  - p(z) coverage
  - Combining results from multiple codes
  - p(z,α) storage
  - Defining ideal LSST algorithm
  - Optimizing spectroscopic samples
- Some examples of problems with current codes

LSST constrains dark energy in many ways… all will rely on redshift information

- LSST will constrain dark energy via 4 major probes:
  - Weak gravitational lensing
  - Baryon Acoustic Oscillations
  - Type Ia supernovae
  - Cluster counts

  (Plus strong lensing, etc.)

- For all of these, as well as much galaxy and AGN science, we want to measure some observable as a function of redshift



LSST Science Book

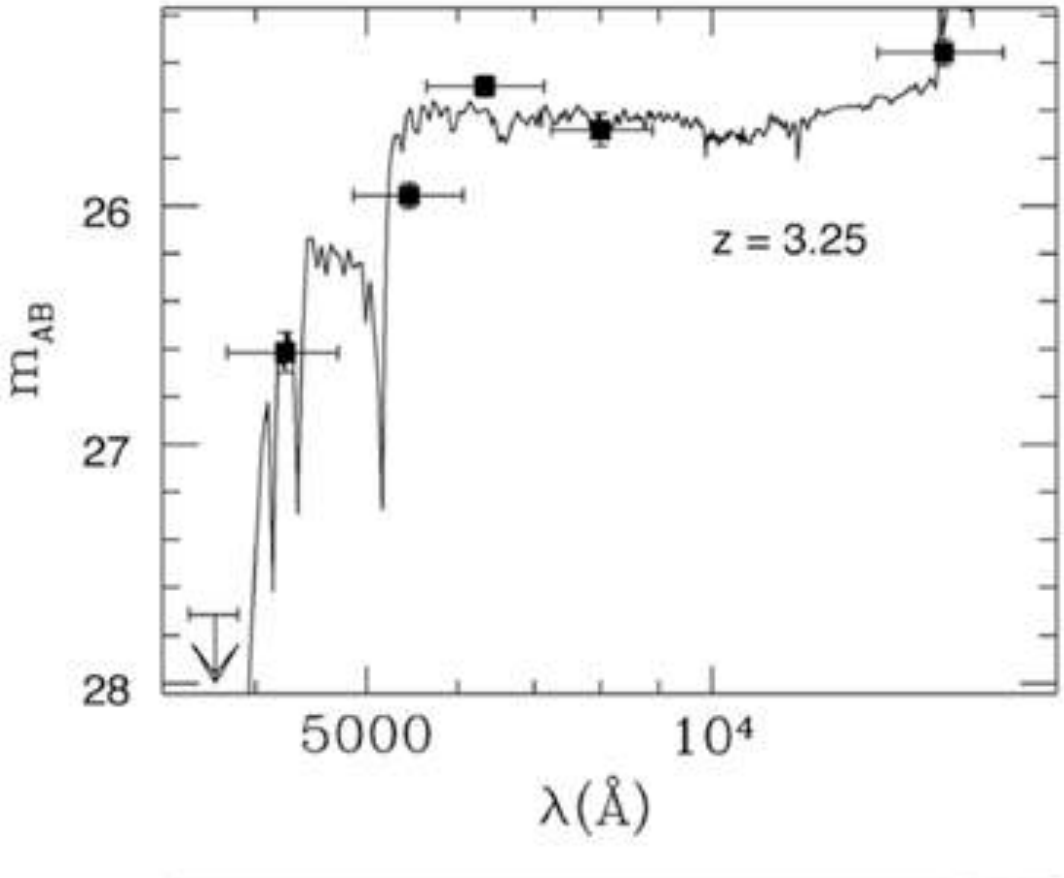
Spectroscopy provides ideal redshift measurements – but is infeasible for large samples

- Redshift ('z') measurements allow us to determine how far back in Universe's history we are looking for an object
- Study galaxy evolution, cosmology, etc. by measuring properties as a function of redshift
- To determine: measure spectrum of light from object with spectrograph; compare observed wavelengths of spectral features to rest frame values to get z
- At LSST "gold sample" (i<25.3) depths, ~100 hours on a 10m telescope to determine a redshift (75% of time) spectroscopically
- With a next-generation, 5000-fiber spectrograph on a 10m telescope, still >50,000 telescope-years to measure redshifts for LSST "gold" weak lensing sample (4 billion galaxies)!

Credit: ESO

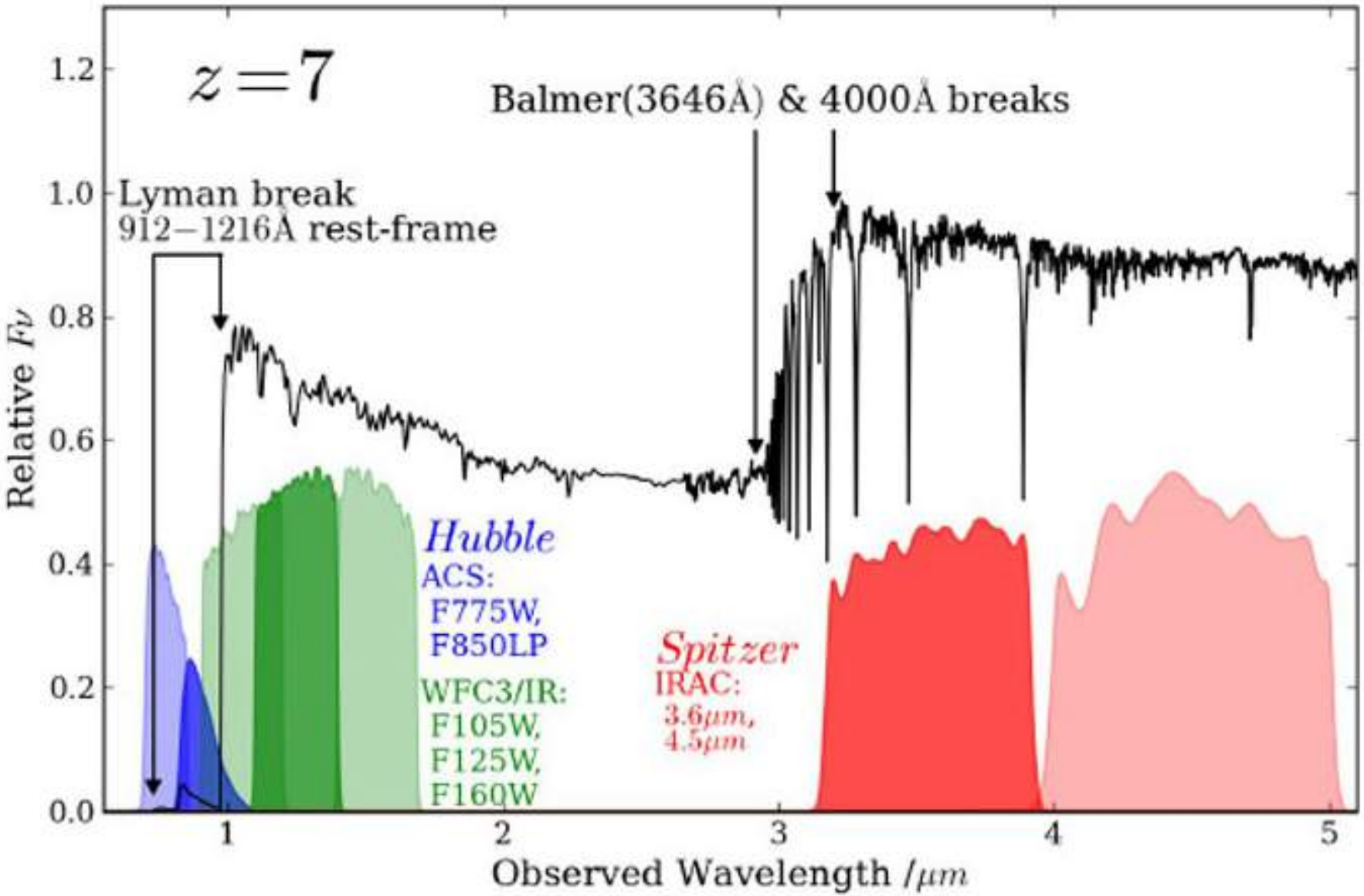
Spectroscopy provides ideal redshift measurements – but is infeasible for large samples

- Alternative: use broad spectral features to determine *z* : a photometric redshift or photo-z
- Advantage: high multiplexing
- Disadvantages: lower precision, calibration uncertainties



Credit: ESO

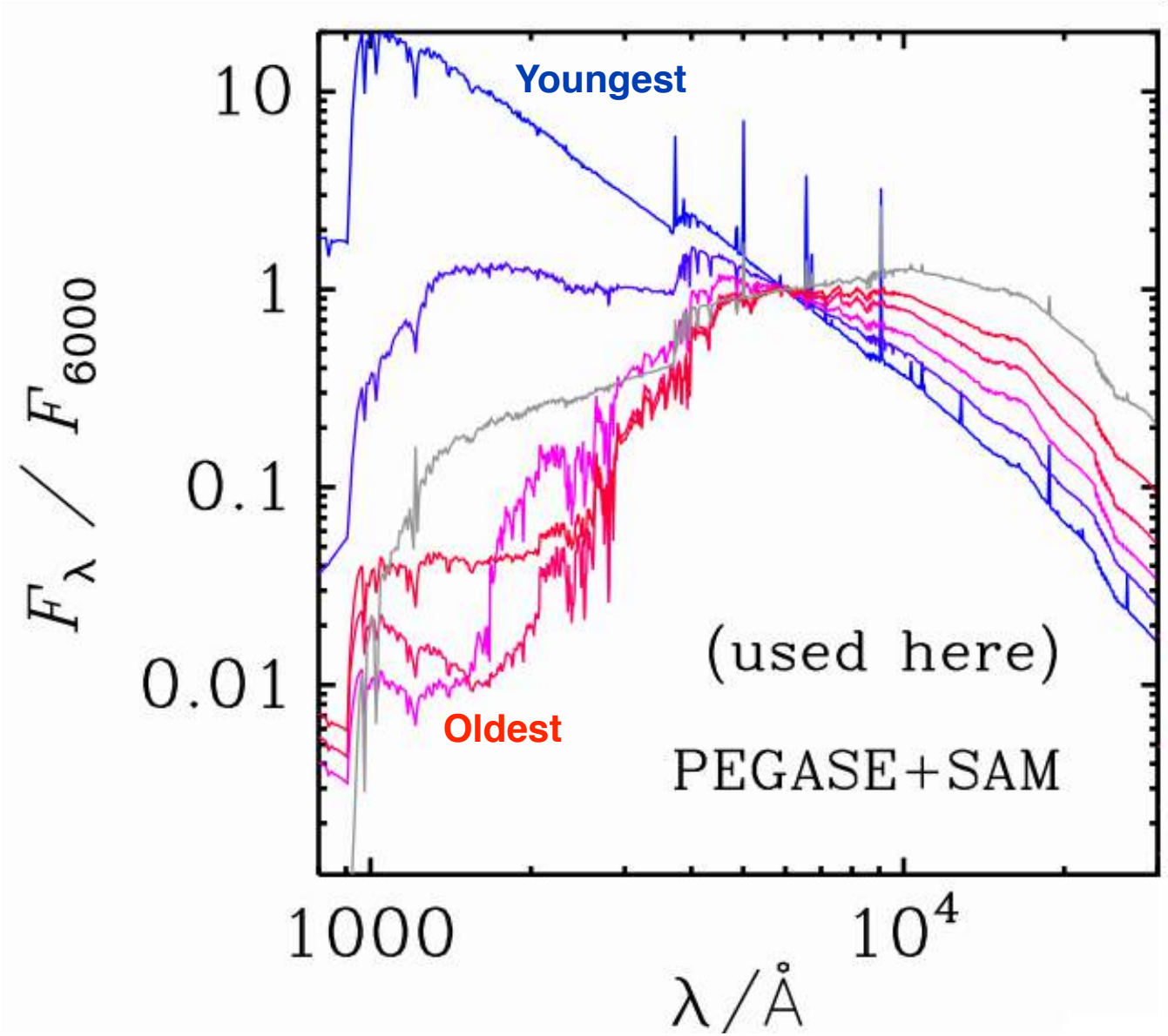
Photometric redshifts rely on the existence of broad spectral features in galaxy spectra...



Dunlop 2012

but those features are stronger in some galaxies than others

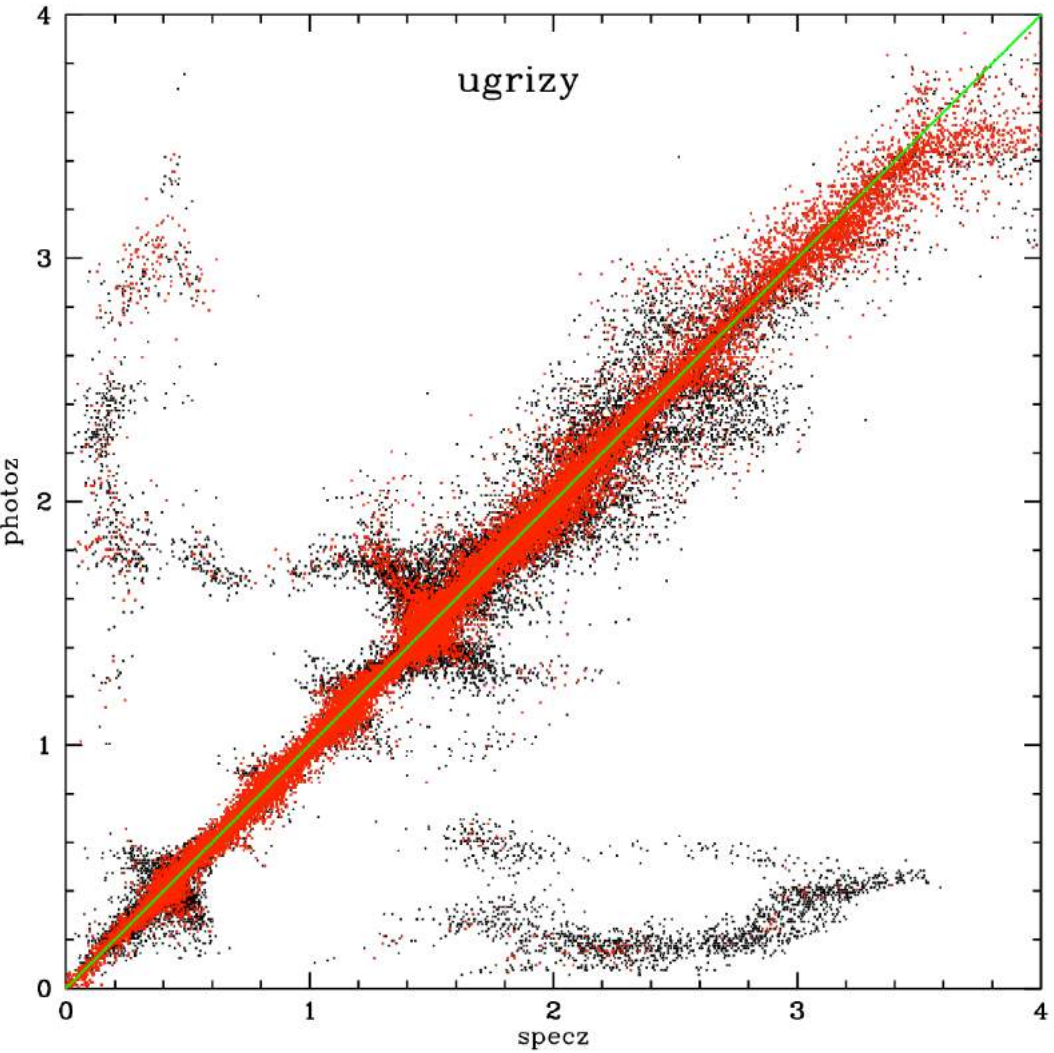
- Galaxies with older stellar populations exhibit stronger 'breaks'
- As a result, photo-z's can be more precise for redder galaxies
- At higher redshifts, blue galaxies with young stellar populations dominate - photo-z problem gets harder



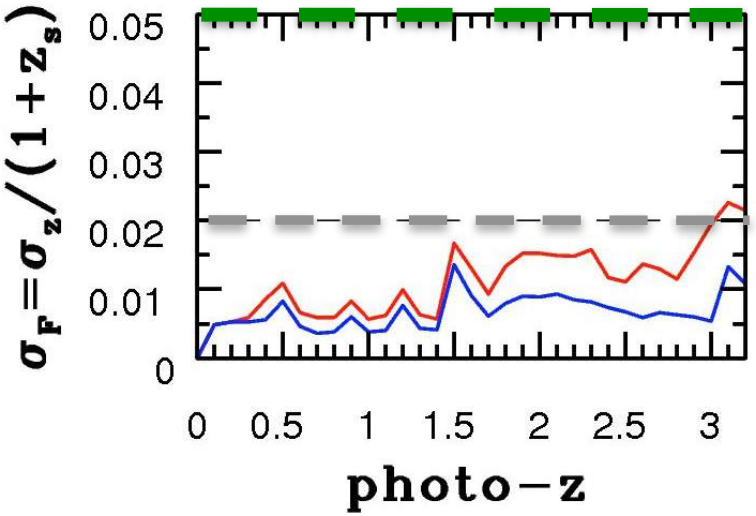
Brammer et al. 2008

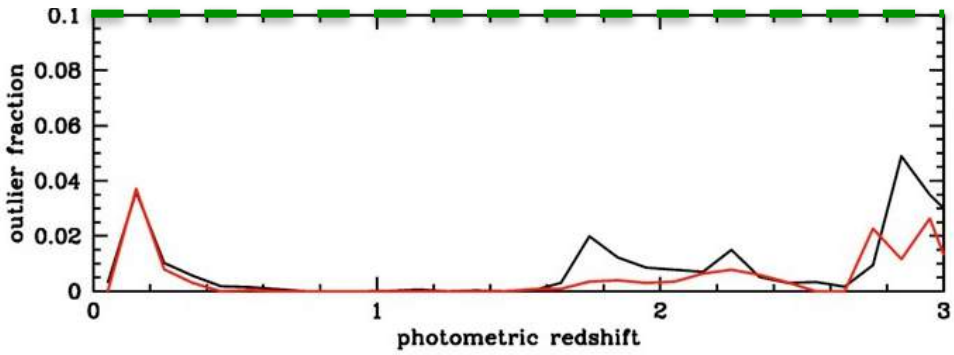
Example: expected photo-z performance for LSST

ugrizy



Green: Requirements on actual performance; grey: requirements on performance <u>with perfect template knowledge</u> (as in these sims)



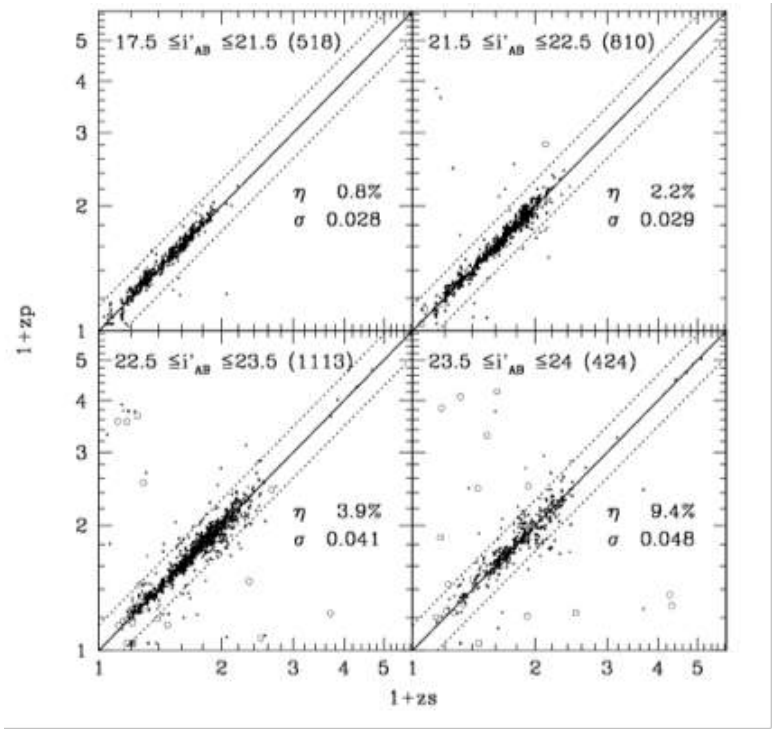


S. Schmidt

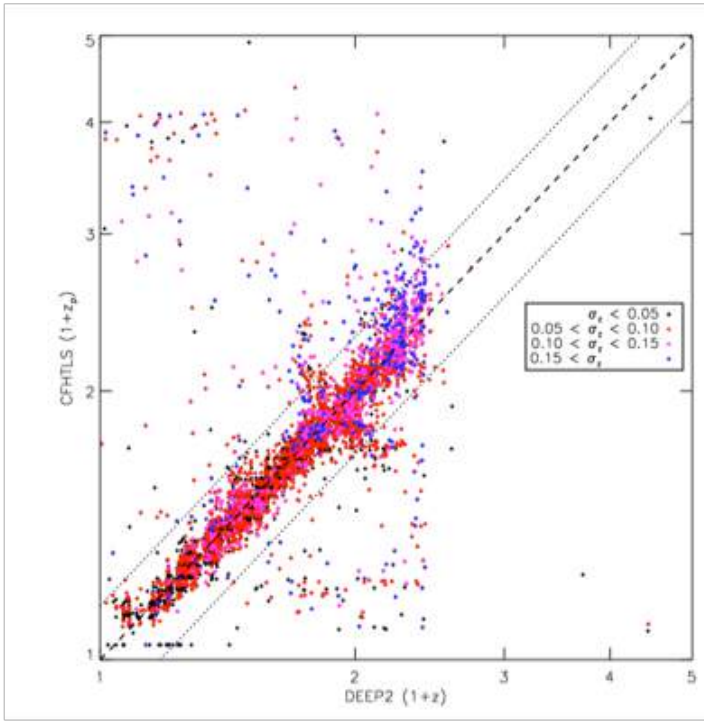
Basic methods: Template fitting photo-z's

- Use galaxies with known *z* to calibrate set of underlying galaxy spectral energy distributions (SEDs) and photometric band-passes
  - Determine posterior probability distribution for *z* | *ugrizy*
  - Also provides info on galaxy properties from template fit

Needs spectra of galaxies spanning full range of possible properties to tune templates, establish priors, etc.



Ilbert et al. 2006

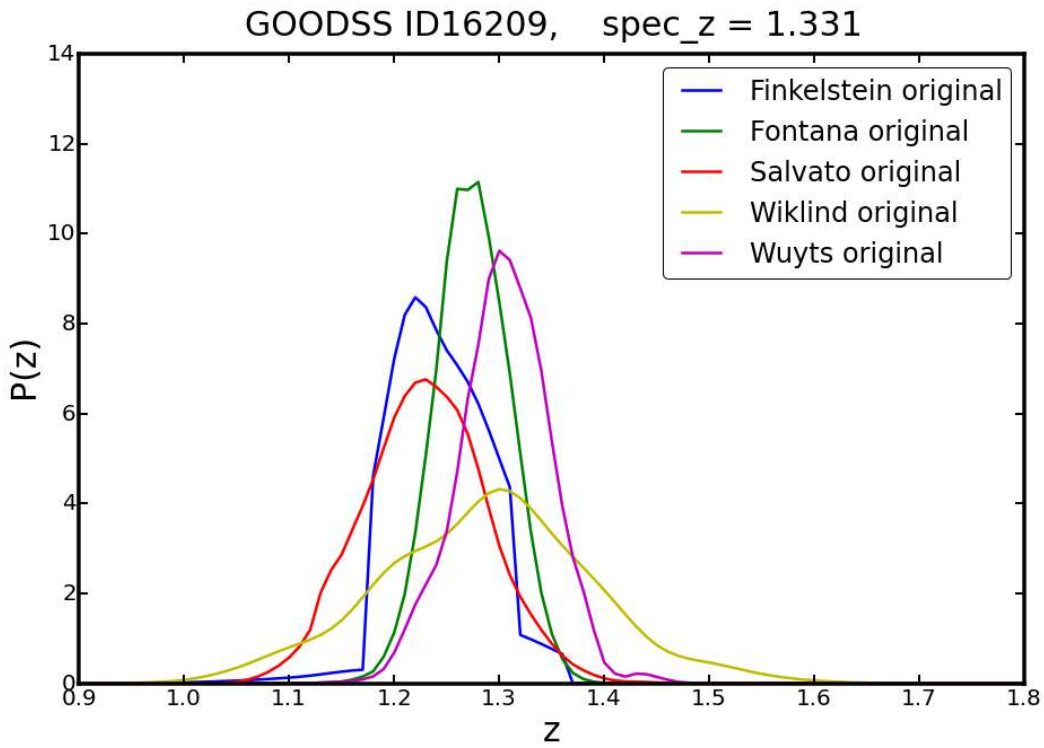


Ilbert photo-z's vs. DEEP2 *z*

Plot by Ben Weiner

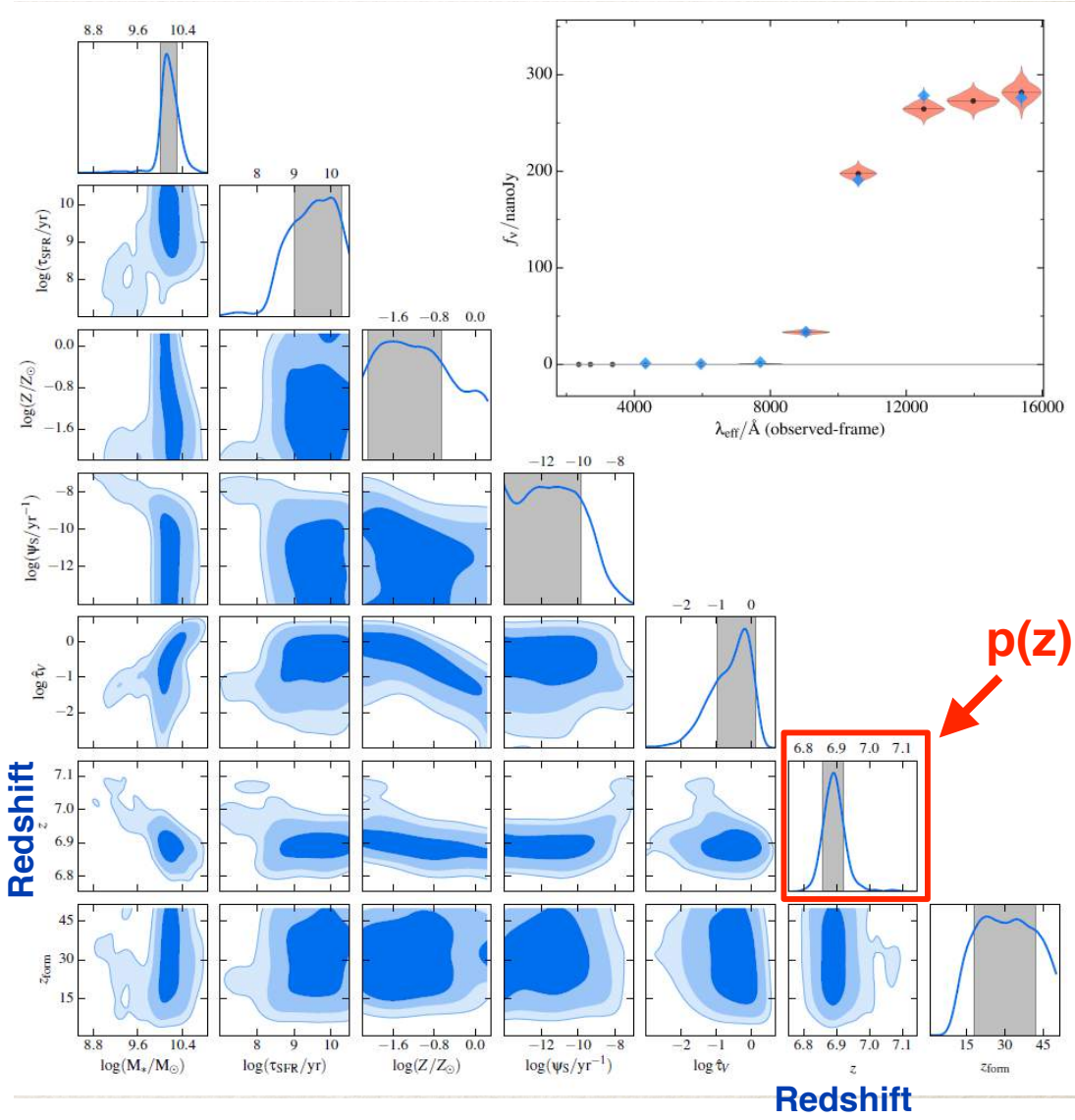
Describing photometric redshift measurements

- Some codes simply output the best-fit *z* with errors
- Generally better to use the posterior probability distribution for *z* | fluxes: *p(z)*
- probability that a < *z* < b = $\int_a^b p(z)\,dz$
  - $\int_0^\infty p(z)\,dz = 1$
- Various definitions for 'point'/ single estimate from *p(z)*:
  - Peak of distribution
  - Expectation value of *z* (possibly only calculating using highest peak)



Describing photometric redshift measurements

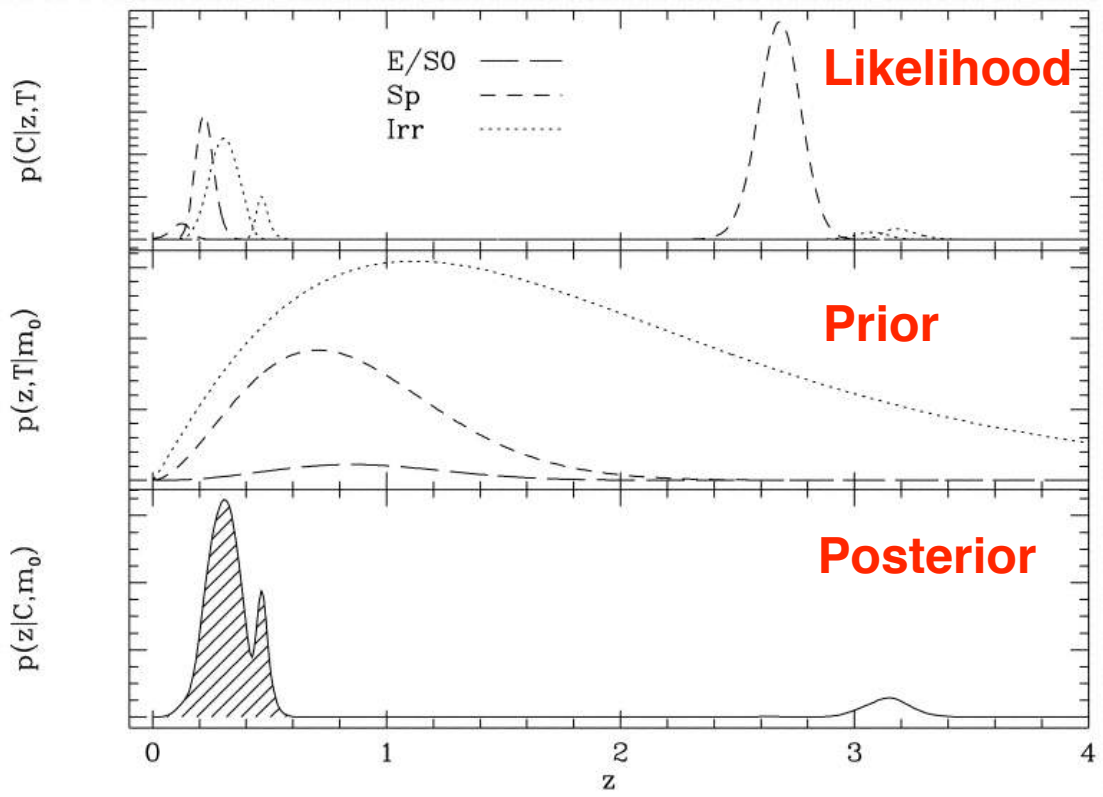
- Can also provide info on galaxy properties from template fit
- E.g., template index *T* or galaxy parameters α_i such as stellar mass, star formation rate, etc.): **p(z,α)**



Chevallard & Charlot 2016

Basic methods: Template fitting photo-z's

- Typical algorithms:
  - Determine likelihood of colors (=ratios of fluxes between bands) as a function of *z* and template
    - Often via $\chi^2(z,T)$ or min($\{\chi^2(z|T)\}$); some algorithms use linear combinations of templates



Benitez 2000

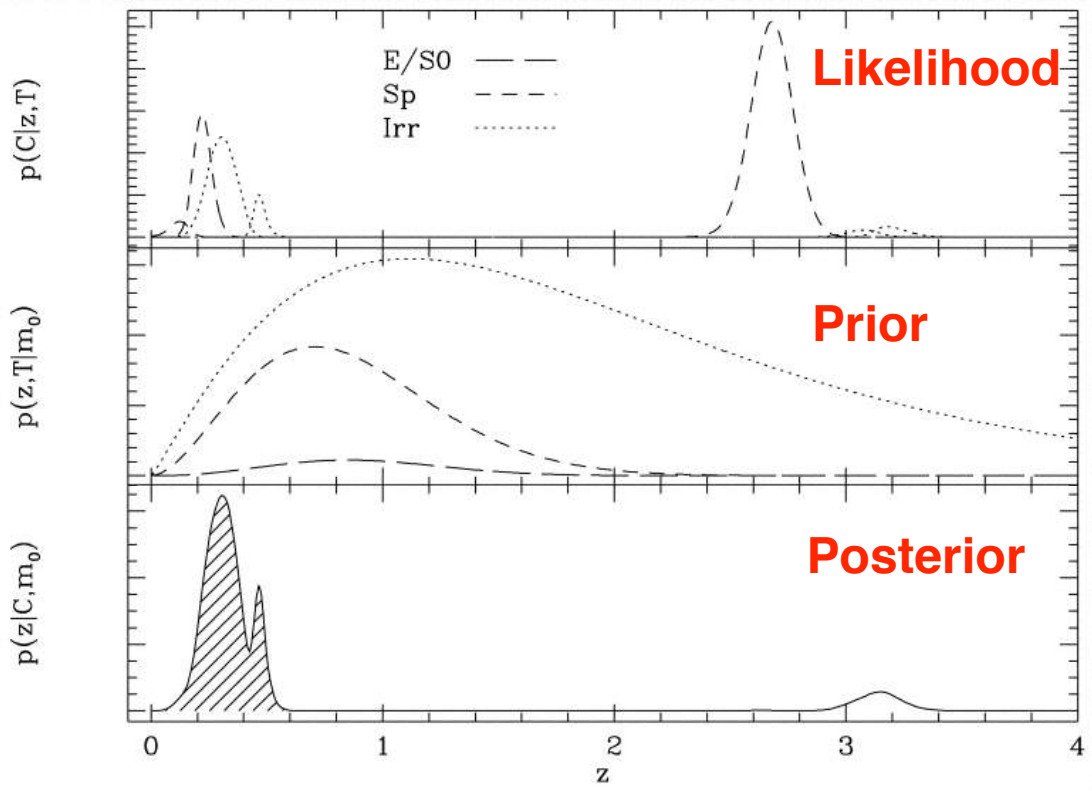
  - Typically utilize prior for redshift or redshift & type based on magnitude (sometimes size/morphology as well)
  - Then multiply to get posterior. . .

Use spectra of galaxies spanning full range of possible properties to tune templates/filter systems, establish priors, etc.

Basic methods: Template fitting photo-z's

- Note that this is standard Bayesian language:
- *p(fluxes | z)* = likelihood
- *p(z)* = prior
- *p(z | fluxes)* = posterior
- By Bayes' theorem,

$$p(z|\text{fluxes}) = \frac{p(\text{fluxes}|z)\, p(z)}{p(\text{fluxes})}$$

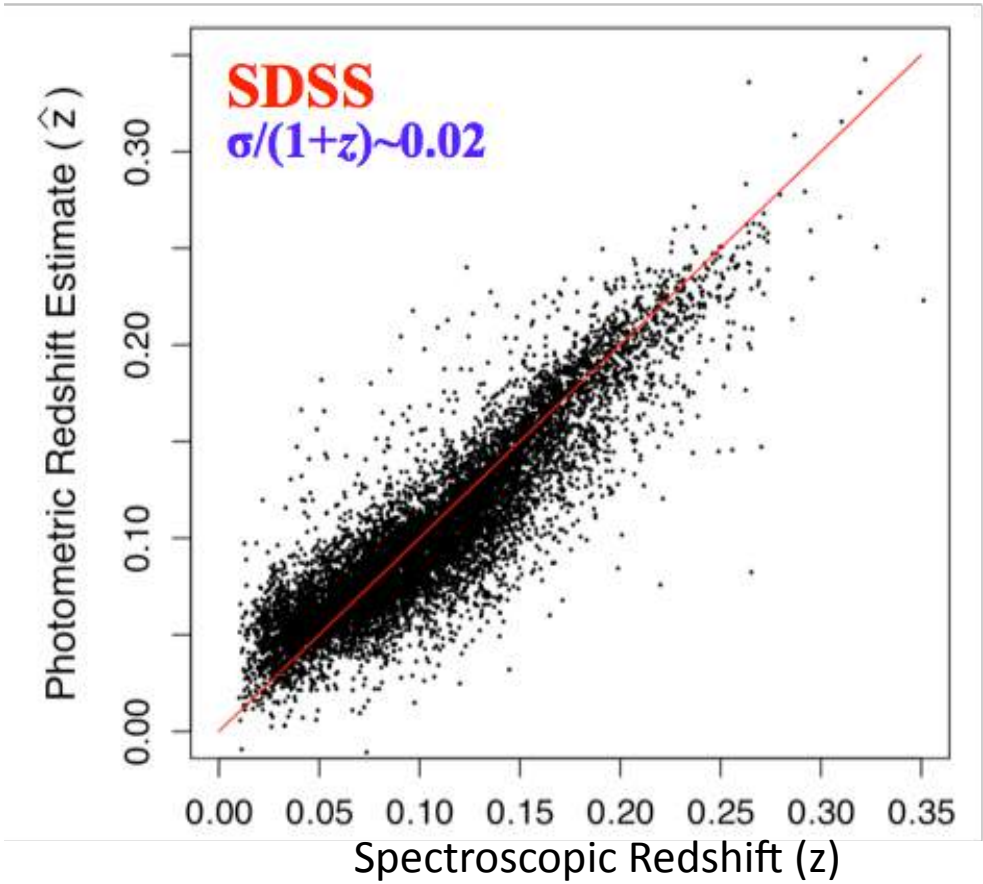


Benitez 2000

- We often just normalize the integral of the posterior to be 1 rather than calculating *p*(fluxes)

Basic methods: Training-based photo-z's

- Use galaxies with known redshift and uniform/well-understood sampling to determine relationship between *z* and colors/fluxes
- Can take advantage of progress in machine learning & stats, but generally extrapolate poorly; Training set MUST span full range of properties & z of galaxies

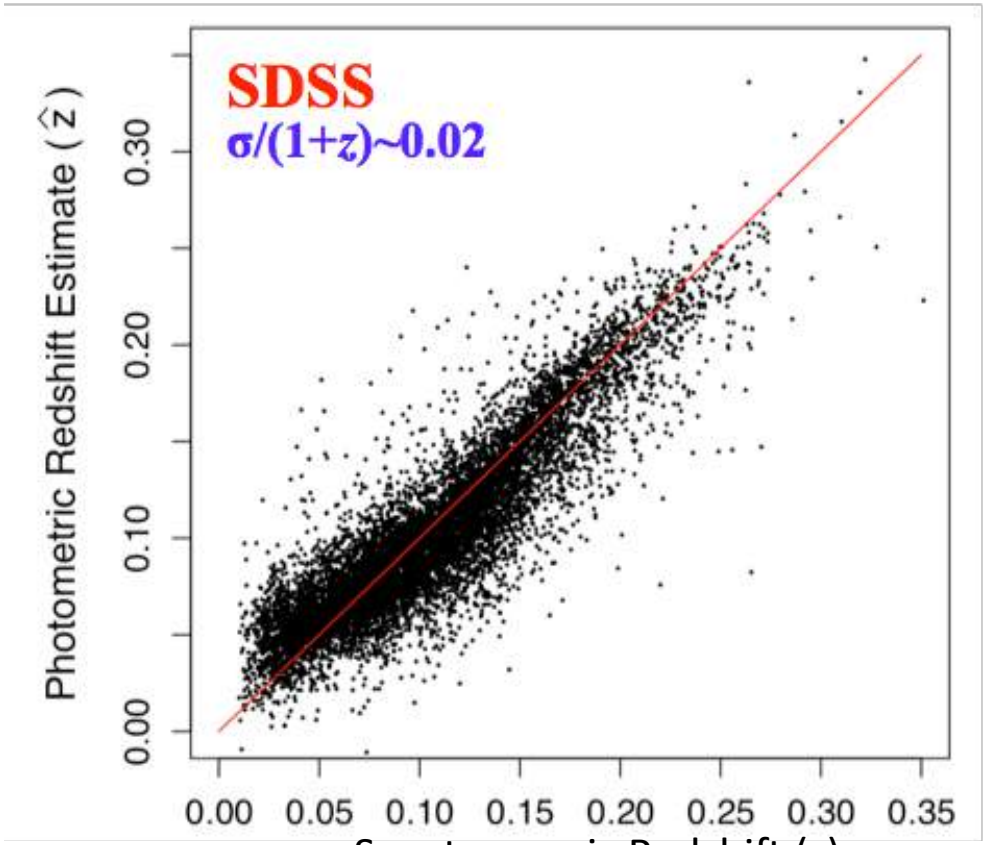


Freeman, JN et al. 2009

Basic methods: Training-based photo-z's

- Many algorithms: e.g.
  - Neural networks
  - Boosted Decision Trees
  - Random Forest regression
  - k-Nearest Neighbor
  - Diffusion map + regression

- For bright, nearby galaxies, training sets are ~complete and both template-based & training-set-based algorithms perform extremely similarly

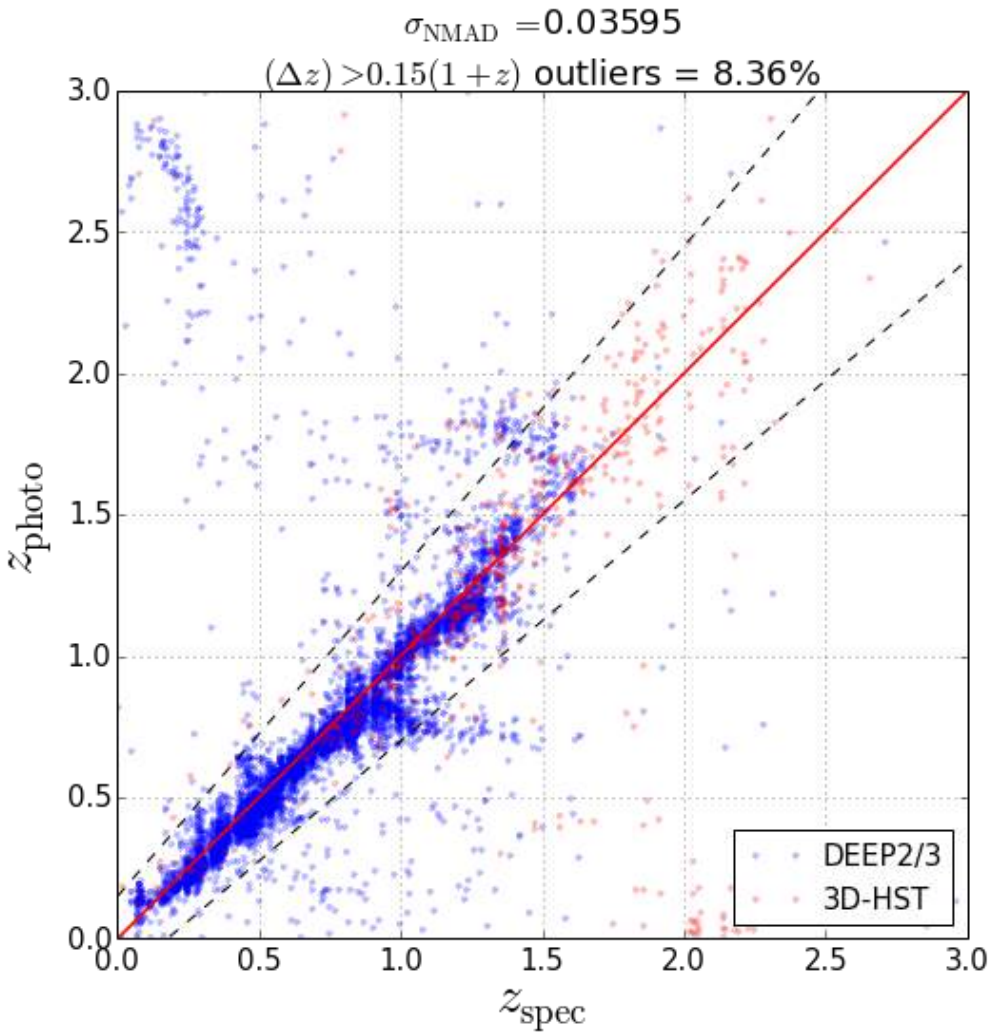


Freeman, JN et al. 2009

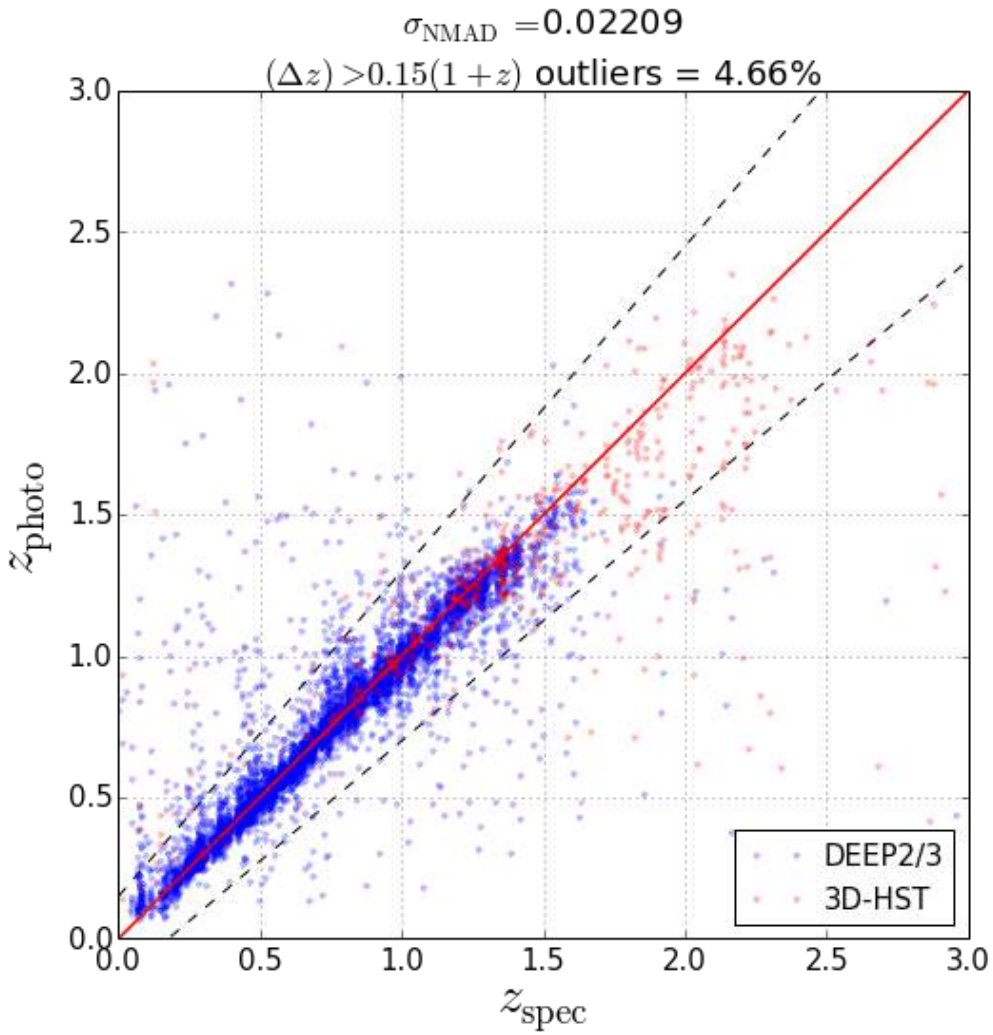
At higher redshifts, the photo-z problem is more difficult

- **Zhou et al. 2018 (in prep.): empirical, LSST-like dataset: CFHT LS *ugriz* + Subaru *y* + DEEP2/DEEP3/3D-HST redshifts**

EAZY (template code, untuned)



Random Forest Regression



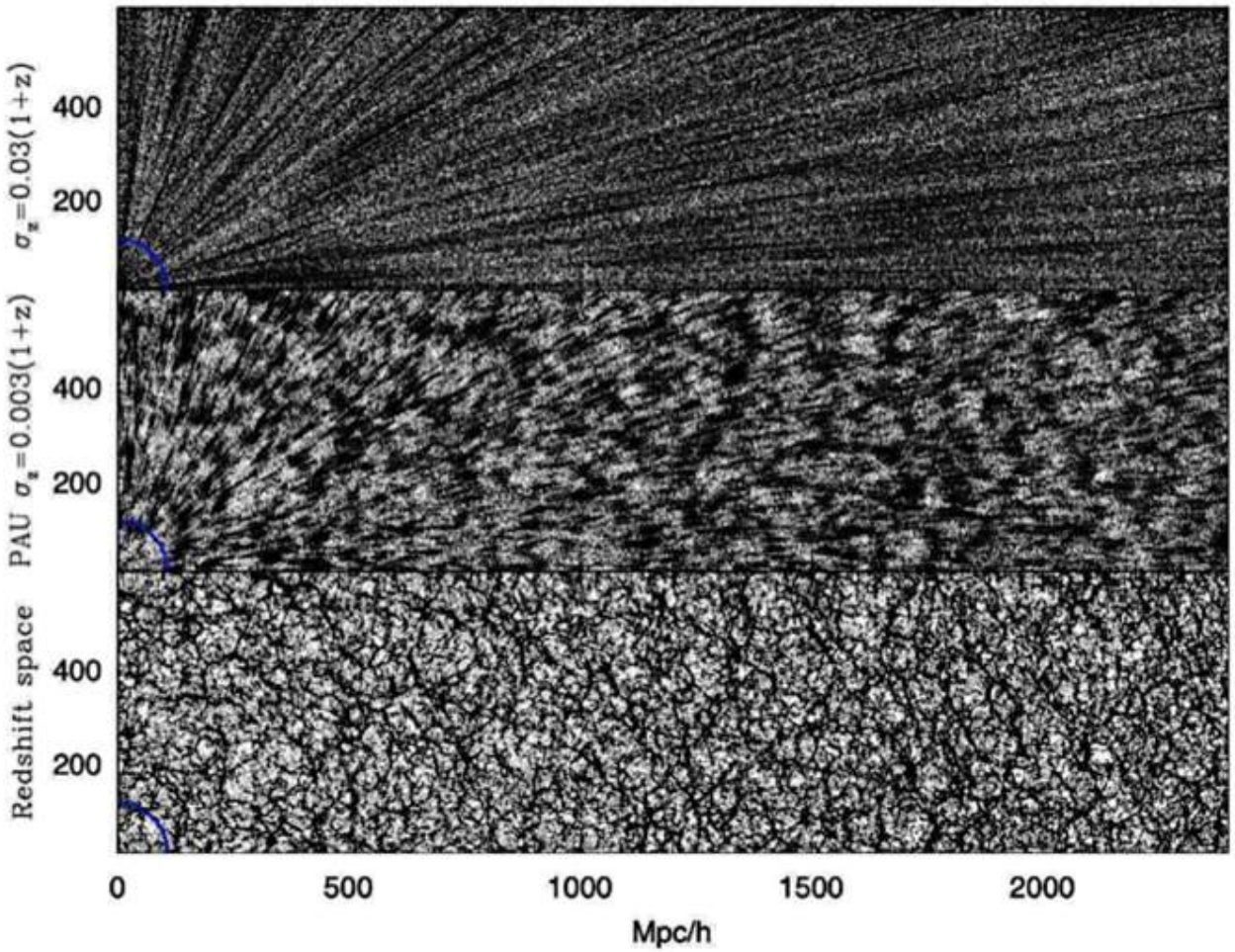
Zhou, JN et al. 2018, in prep.

Outline

- Overview of photometric redshifts
  - Template methods
  - Training-based methods
- Requirements and resources for training and calibrating photometric redshifts
- Some open issues
  - Spectroscopic incompleteness
  - Robust training
  - p(z) coverage
  - Combining results from multiple codes
  - p(z,α) storage
  - Defining ideal LSST algorithm
  - Optimizing spectroscopic samples
- Some examples of problems with current codes

Two spectroscopic needs for photo-z work: training and calibration

- Better training (optimization of algorithms using sets of objects with spectroscopic redshift measurements) shrinks photo-z errors for individual objects

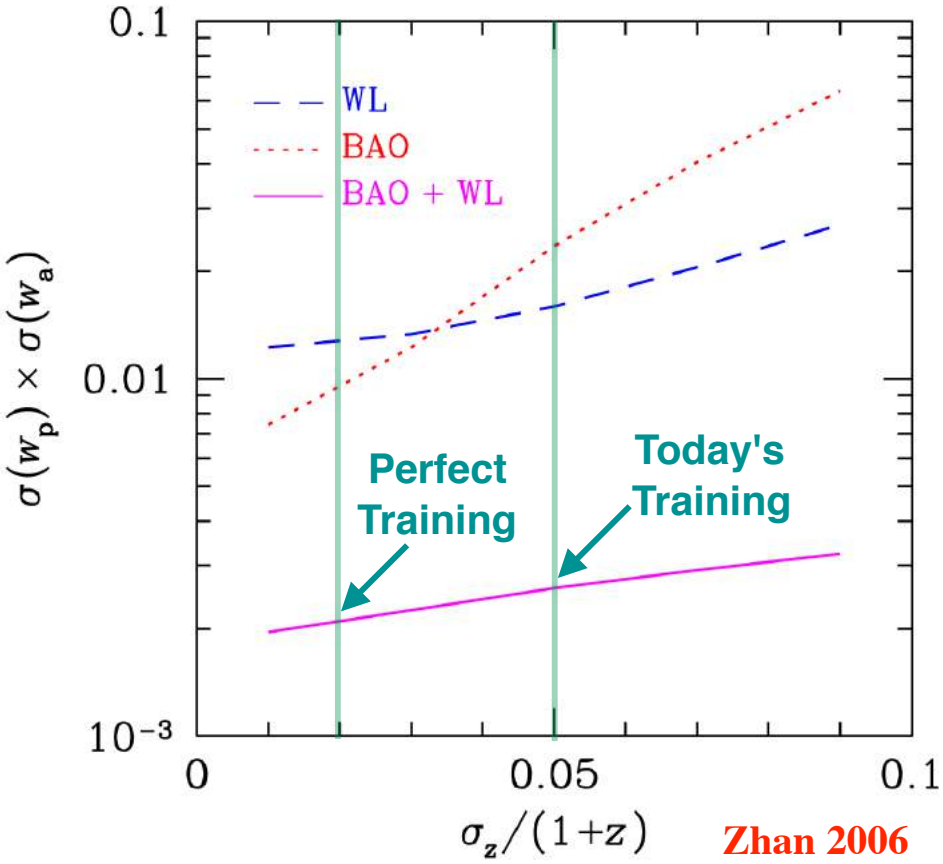


Benitez et al. 2009

– Training datasets will contribute to calibration of photo-z's. ~Perfect training sets can solve calibration needs.

Improved photometric redshift *training* would greatly increase the science gains from LSST

- E.g.: all LSST probes of dark energy will rely on measuring observables as a function of photometric redshift
- Smaller photo-z errors from better-trained algorithms can improve dark energy constraints, especially for BAO and clusters

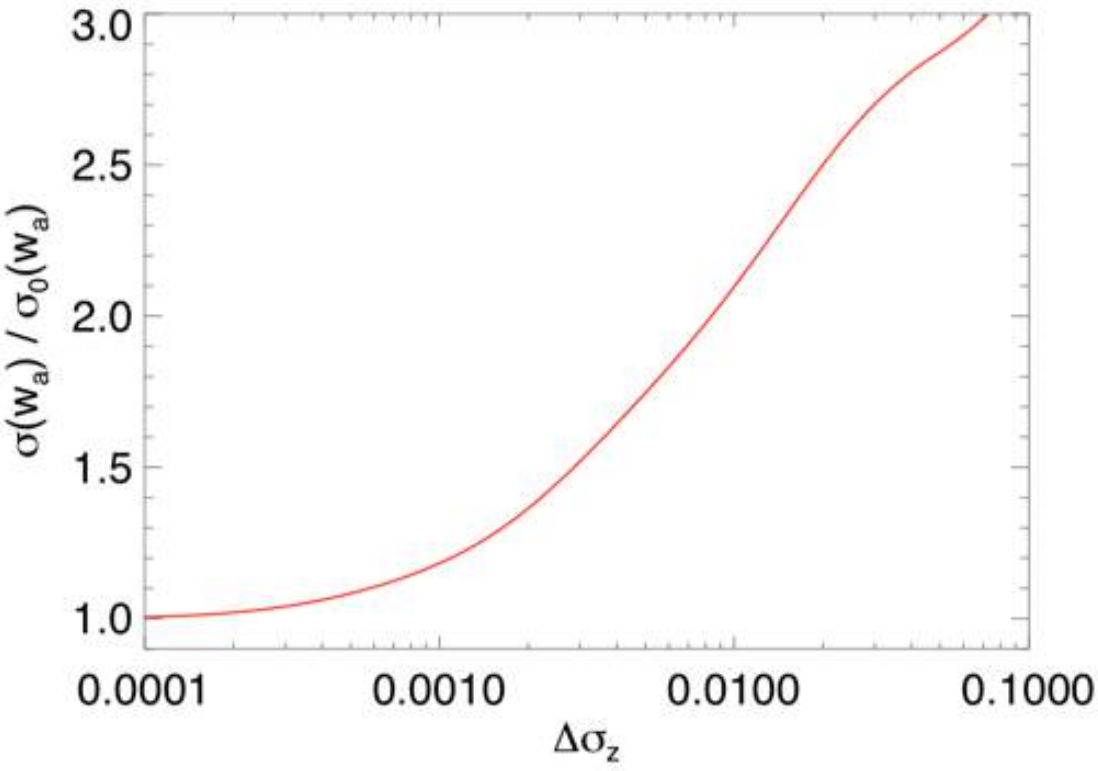


Zhan 2006

- LSST system-limited photo-z accuracy is σ_z~0.02-0.025(1+z) (vs. σ_z~0.05(1+z) in similar samples today): difference is knowledge of templates/intrinsic galaxy spectra
- Perfect training set would increase LSST DETF FoM by at least 40%

Excellent calibration of photo-z's is needed or else dark energy inference will be wrong

- For weak lensing and supernovae, individual-object photo-*z*'s do not need high precision, but the calibration must be accurate - i.e., bias and errors need to be *extremely* well-understood or dark energy constraints will be off

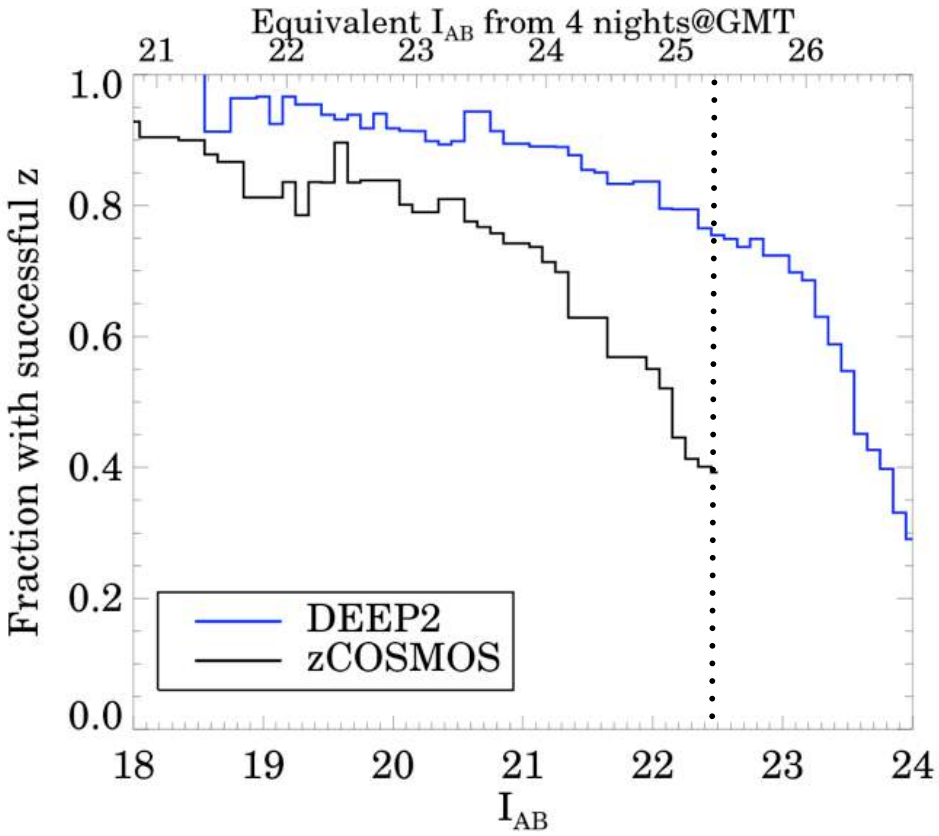


Newman et al. 2015

– *uncertainty in* bias, $\sigma(\delta_z) = \sigma(<z_p - z_s>)$, and in scatter, $\sigma(\sigma_z) = \sigma(\mathrm{RMS}(z_p - z_s))$, must both be $<\sim 0.002(1+z)$ in each bin for Stage IV surveys. Calibration may be done via cross-correlation methods using DESI/4MOST redshifts (Newman 2008)

Requirements for photometric redshift training for LSST

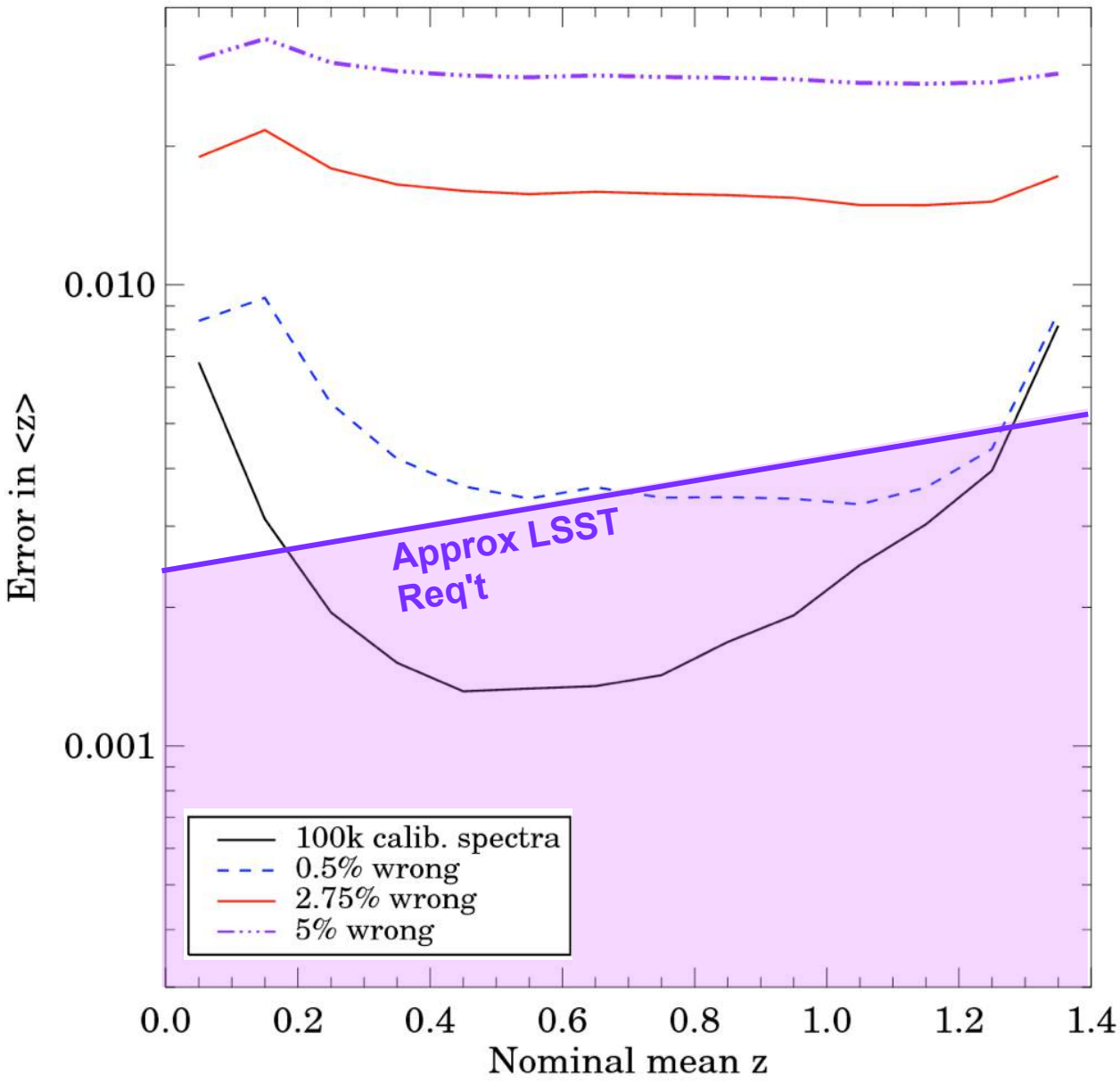
- Need highly-secure spectroscopic redshifts for 20k-30k galaxies sampling full range of galaxy colors, magnitudes, and redshifts
- Newman et al. 2015, *Spectroscopic Needs for Imaging Dark Energy Experiments,* presents a baseline scenario:
  - >30,000 galaxies down to LSST weak lensing limiting magnitude (*i*~25.3)
  - 15 widely-separated fields at least 20 arcmin diameter to allow sample/cosmic variance to be mitigated & quantified
    - Equal cosmic variance to Euclid C3R2 plan but much lower sky area
  - Long exposure times are needed to ensure >75% redshift success rates: >100 hours at Keck to achieve DEEP2-like S/N at *i*=25.3
  - See http://adsabs.harvard.edu/abs/2015APh....63...81N



Newman et al. 2015

Note: even for 100% complete samples, current false-z rates would be a problem

- Only the highest-confidence redshifts should be useful for precision calibration: lowers spectroscopic completeness further when restrict to only the best
- A major reason why getting highly secure redshifts is important



Based on simulated redshift distributions for ANNz-defined DES bins in mock catalog from Huan Lin, UCL & U Chicago, provided by Jim Annis

Summary of (some!) potential instruments for photo-z training

Instrument / Telescope	Collecting Area (sq. m)	Field area (sq. deg.)	Multiplex
4MOST	10.7	4.000	1,400
Mayall 4m / DESI	11.4	7.083	5,000
WHT / WEAVE	13.0	3.139	1,000
Magellan LASSI	32.4	1.766	5,000
Subaru / PFS	53.0	1.250	2,400
VLT / MOONS	58.2	0.139	500
Keck / DEIMOS	76.0	0.015	150
FOBOS	76.0	0.087	500
ESO SpecTel	87.9	4.9	3,333
MSE	97.6	1.766	3,249
GMT/MANIFEST + GMACS v. A	368	0.087	760
GMT/MANIFEST + GMACS v. B	368	0.087	420
TMT / WFOS	655	0.011	100
~~Fiber WFOS-pessimistic~~	~~655~~	~~0.022~~	~~1,000~~
~~Fiber WFOS-optimistic~~	~~655~~	~~0.056~~	~~2,000~~
E-ELT / Mosaic Optical	978	0.009	200
E-ELT / MOSAIC NIR	978	0.009	100

Updated from Newman et al. 2015, *Spectroscopic Needs for Imaging Dark Energy Experiments*

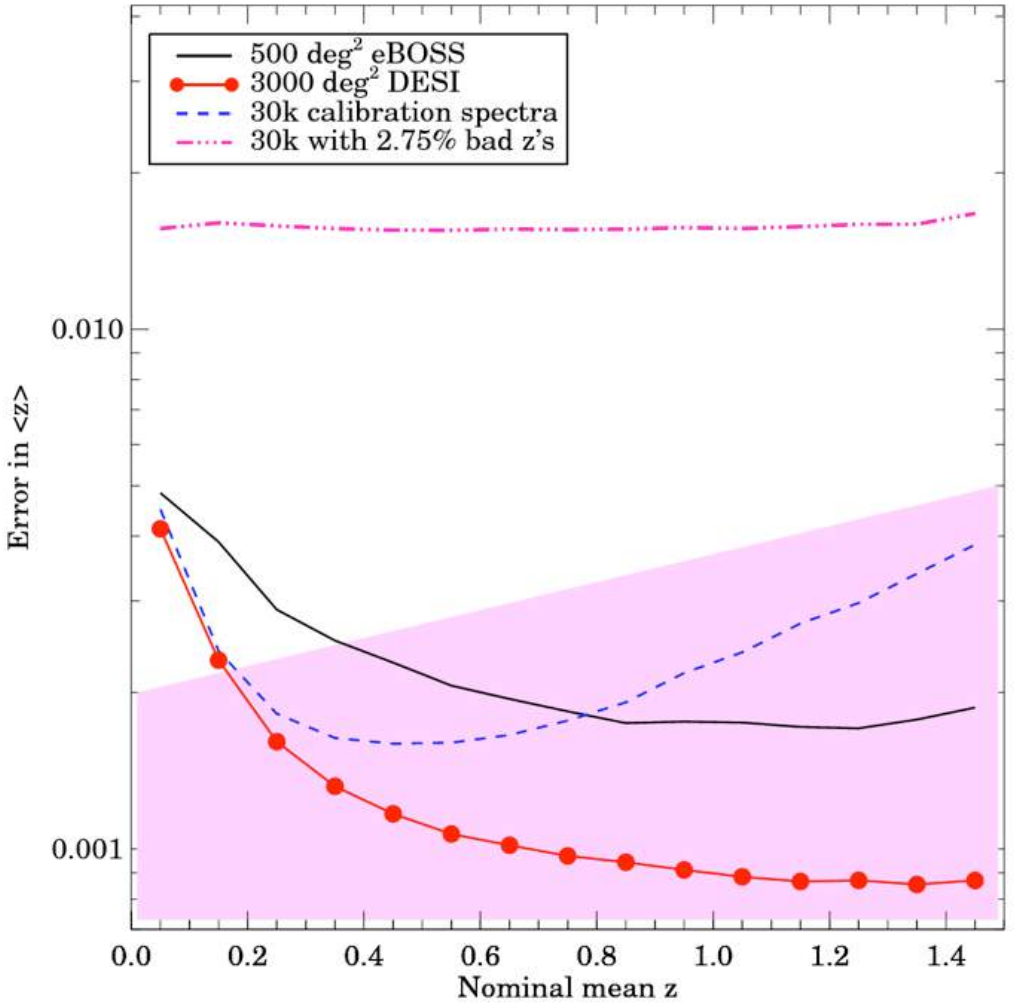
Dark time (with 1/3 losses for weather + overheads) required for each instrument

Instrument / Telescope	Total time (years), >75% complete LSST sample	Total time (years), >90% complete LSST sample
4MOST	7.7	48.4
Mayall 4m / DESI	5.1	31.9
WHT / WEAVE	9.0	56.0
Magellan LASSI	1.8	11.2
Subaru/PFS	1.1	6.9
VLT/MOONS	4.0	25.0
Keck/Deimos	10.2	63.9
Keck/FOBOS	4.4	27.5
ESO SpecTel	0.66	4.1
MSE	0.60	3.7
GMT/MANIFEST + GMACS v. A	0.42	2.6
GMT/MANIFEST + GMACS v. B	0.75	4.7
TMT / WFOS	1.8	11.1
~~Fiber WFOS-pessimistic~~	~~0.36~~	~~2.2~~
~~Fiber WFOS-optimistic~~	~~0.14~~	~~0.87~~
E-ELT / MOSAIC Optical	0.60	3.7
+	+	+
E-ELT / MOSAIC NIR	1.2	7.4

Updated from Newman et al. 2015, *Spectroscopic Needs for Imaging Dark Energy Experiments*

If spectroscopy proves incomplete, calibration will probably need to come from cross-correlation methods

- Galaxies of all types cluster together: trace same dark matter distribution
- Enables reconstruction of *z* distributions via spectroscopic/ photometric cross-correlations (Newman 2008)
- For LSST calibration, require >100k objects over >100 deg², spanning full z range
- >500 degrees of overlap with DESI-like survey would meet LSST science requirements (>4000 sq deg of overlap expected)



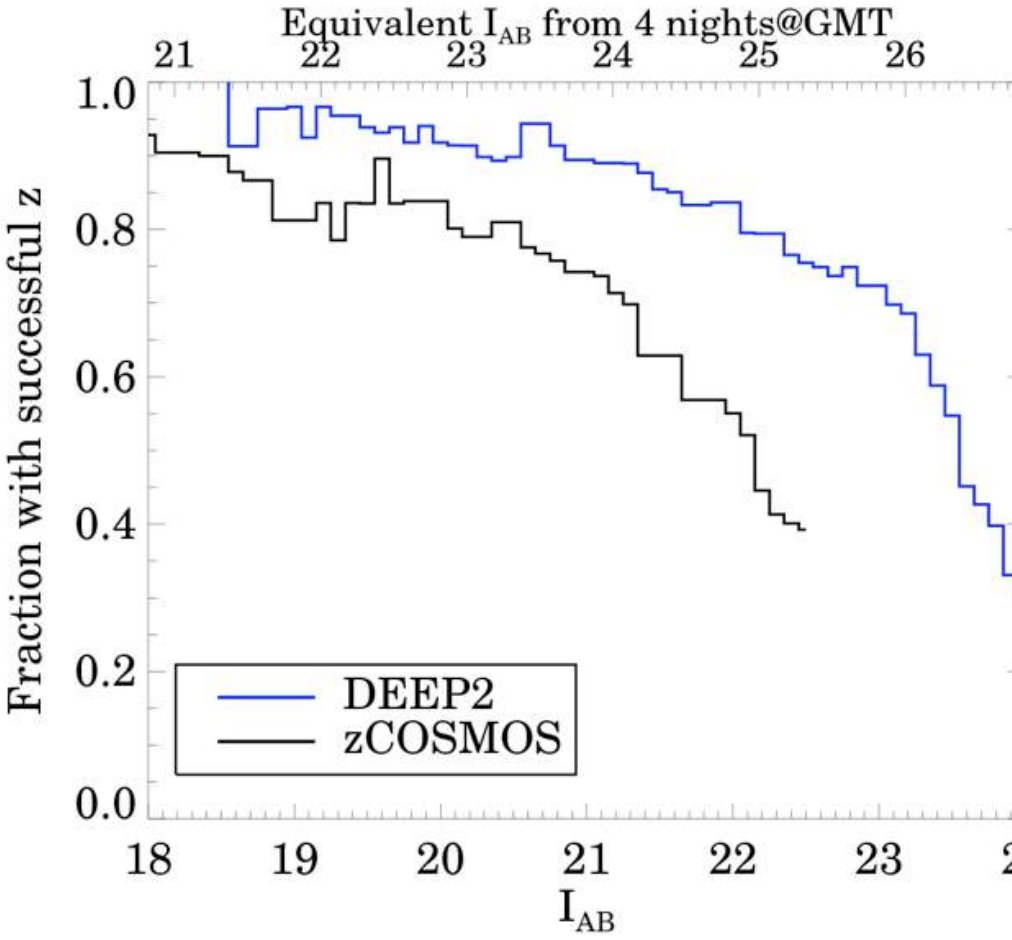
Snowmass white paper: *Spectroscopic Needs for Imaging DE Experiments* (Newman et al. 2015, http://arxiv.org/abs/1309.5388)

Outline

- Overview of photometric redshifts
  - Template methods
  - Training-based methods
- Requirements and resources for training and calibrating photometric redshifts
- Some open issues
  - Spectroscopic incompleteness
  - Robust training
  - p(z) coverage
  - Combining results from multiple codes
  - p(z,α) storage
  - Defining ideal LSST algorithm
  - Optimizing spectroscopic samples
- Some examples of problems with current codes

Open issues: dealing with incompleteness in training/ calibration datasets

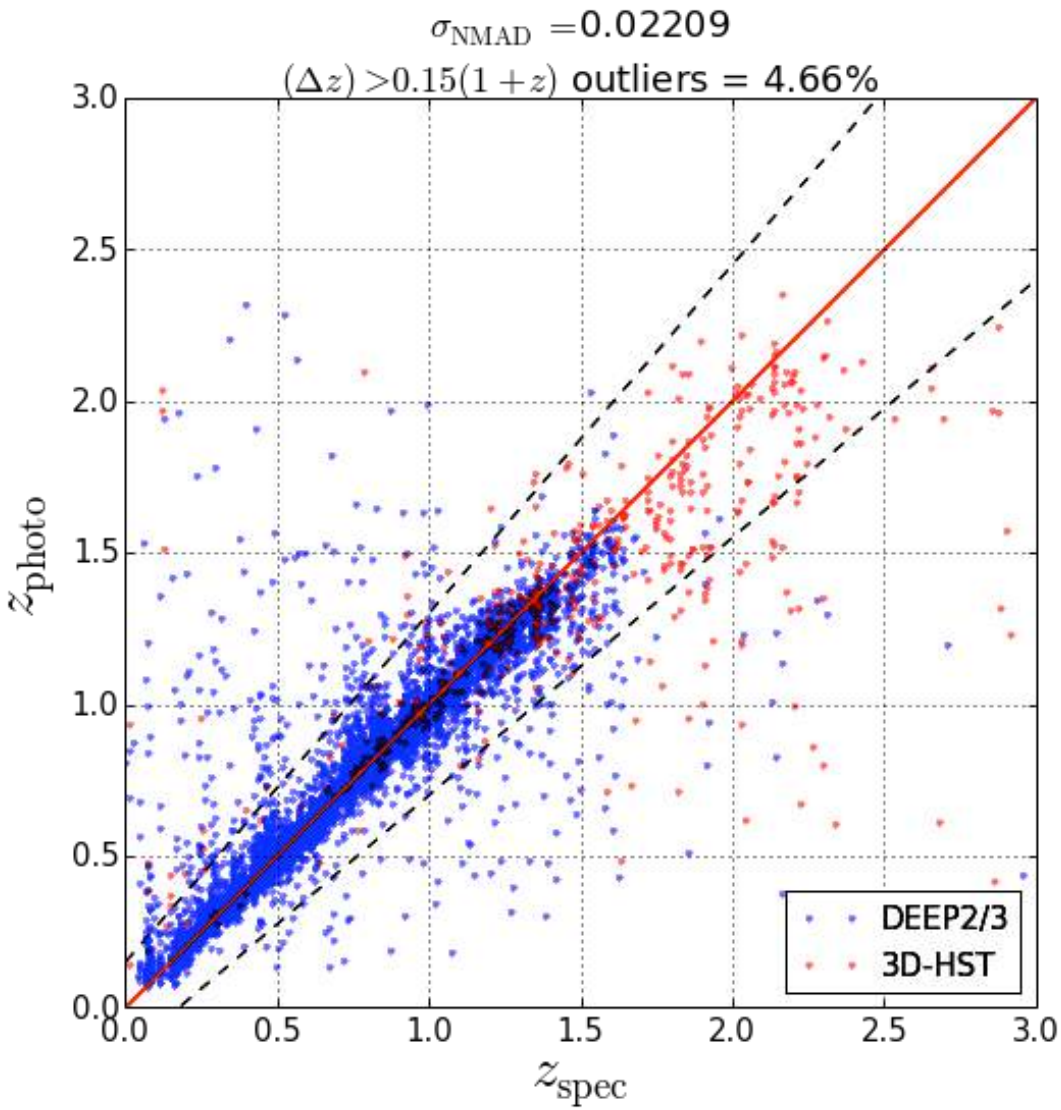
- In current deep spectroscopic surveys, 25-60% of targets fail to yield secure redshifts
- *z* success rate depends on galaxy properties
- Estimated need 99-99.9% completeness to prevent systematic errors in calibration, unless apply other methods (e.g., cross-correlations)
- Major issue for training-set techniques



2013) and zCOSMOS (Lilly et al. 2009)

Open issues: Robust training methods

- 1% incorrect-redshift rate is sufficient to bias photo-z's beyond tolerances
- Depending on survey, up to 5% of 'secure' redshifts are incorrect
- If can train algorithms in a manner robust to outlier/ wrong redshifts, could use the broader set of less-secure spectroscopic redshifts
- ML methods that extrapolate well would also be interesting



Zhou, JN et al. 2018, in prep.

Open issues: Making posteriors great again

- **CANDELS code comparison: Dahlen et al. 2013**
- **11 code/template combinations were tested using ~600 redshifts in GOODS-S (trained with a separate set of 600 redshifts)**
- **Generally χ^2 minimization, generally with some sort of prior.**
- **Codes with p(z)'s available are marked by ★**

Code		Code ID	Template set	bias_z^a	OLF^b	σ_F^c	σ_O^d	
Rainbow		A	PEGASE[b]	-0.010	0.092	0.167	0.041	
GOODZ		B	CWW[c], Kinney[d]	-0.007	0.036	0.099	0.035	
EAZY	★	C	EAZY[e]+BX418[f]	-0.009	0.051	0.114	0.044	
SPOC		D	BC03[g]	-0.030	0.147	0.197	0.073	
zphot	★	E	PEGASEv2.0[b]	-0.007	0.041	0.104	0.037	
EAZY		C	EAZY[e]	-0.009	0.053	0.121	0.037	
SATMC		F	BC03[g]	-0.008	0.093	0.272	0.064	
HyperZ		G	Maraston05[h]	0.013	0.078	0.189	0.050	
LePhare	★	H	BC03[g]+Polletta07[i]	-0.008	0.048	0.132	0.038	
WikZ	★	I	BC03[g]	-0.023	0.046	0.153	0.049	
EAZY	★	C	EAZY[e]	-0.005	0.039	0.127	0.034	
				-0.008	0.029	0.088	0.031	median(all)
				-0.009	0.031	0.079	0.029	median(5)

Dahlen et al. 2013

Open issues: Making posteriors great again

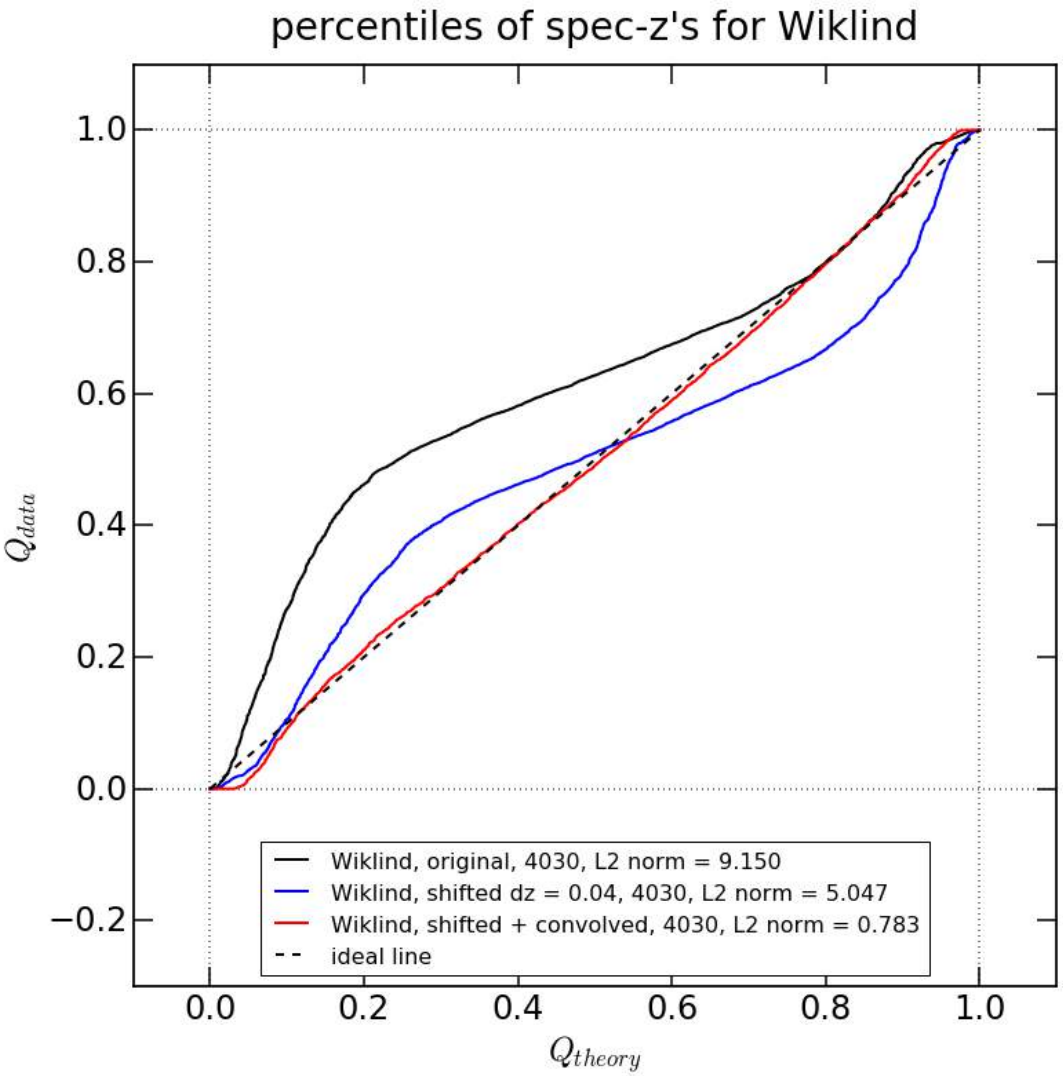
- Many analyses assume that photo-z codes are providing posterior PDFs with proper coverage (and assuming that they can add PDFs to get N(z); talk to Alex Malz if you want to learn about the right way to do that...)
- Dahlen et al. 2013 tested the fraction of spectroscopic redshifts that are in the inner 68% or inner 95% of their PDFs
- Coverage is all over the place; no codes were good at both 68% and 95% points

Code		WFC3 H-selected	
conf. int:		68.3%	95.4%
2A		46.1	
3B		81.6	92.8
4C	★	64.0	88.2
5D		2.5	4.2
6E	★	52.0	84.7
7C		65.0	87.3
8F		15.3	15.6
9G		16.3	44.1
11H	★	35.2	54.0^a
12I	★	88.7	96.7
13C	★	52.0	72.7

Dahlen et al. 2013

Open issues: Making posteriors great again

- LSST Dark Energy Science Collaboration has done a controlled exploration of this problem... more in a minute
- Meanwhile, kludge in Kodra et al. 2019: modify *p(z)*'s for CANDELS HST survey
  - Shift p(z) by constant in *z* direction; convolve with Gaussian kernel; and take to a power (equivalent to rescaling errors in χ^2 calculation)
- Optimize parameters by minimizing total L2 norm of deviation in quantile-quantile plot from expected line
- Quantile-quantile shows the fraction of objects whose true redshift is below quantile Q_{theory} in the object's photo-z PDF: ideally, unity line

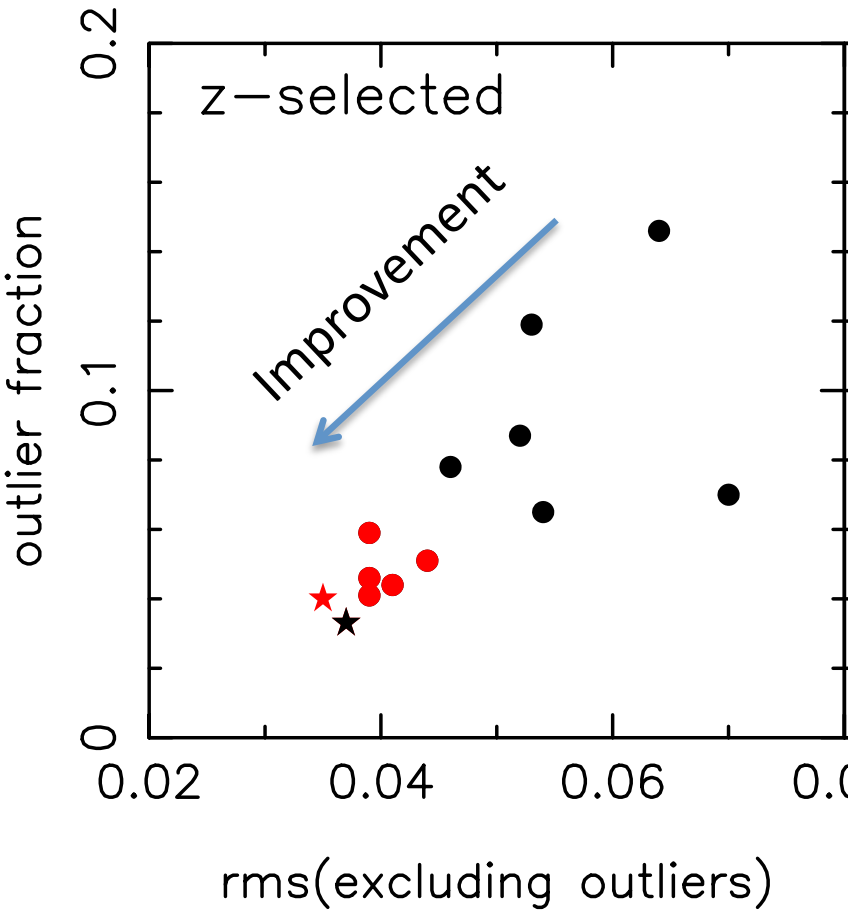


Kodra, JN et al. 2019, in prep.

Open issues: Combining PDF results from multiple codes

- **Dahlen et al. found that medians of point estimates from multiple codes (★'s) have smaller scatter (relative to spec-z) than any individual code**

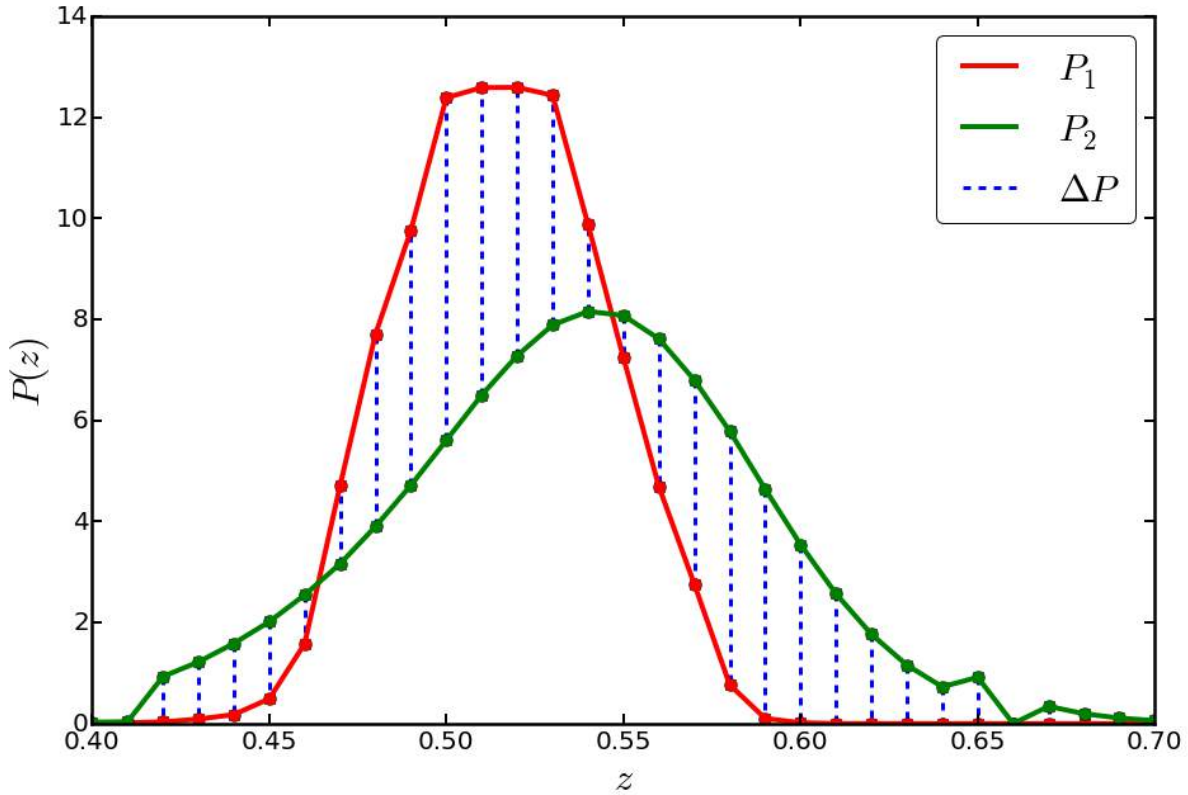
- **All codes are run on the same data! Current codes do not make optimal use of available information...**



Dahlen et al. 2013

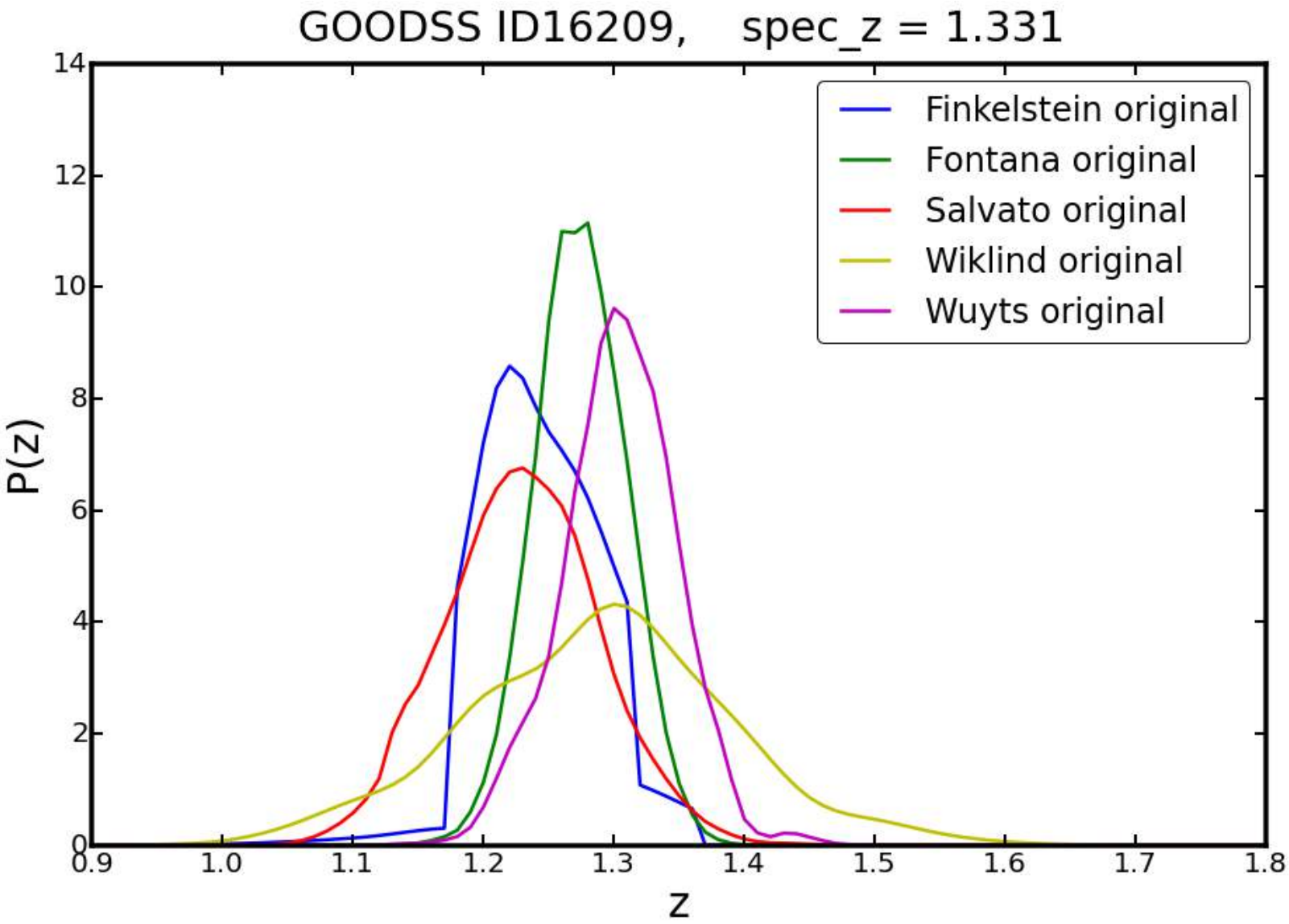
Open issues: Combining PDF results from multiple codes

- Dahlen et al. presented a hierarchical Bayesian combination method (cf. Press & Kochanek, Lang & Hogg, etc.)
- Izbicki & Lee 2016 use weighted combinations of codes
- Kodra et al. (in prep) investigates using PDF that minimizes total Fréchet distance to remaining PDFs: analogous to median



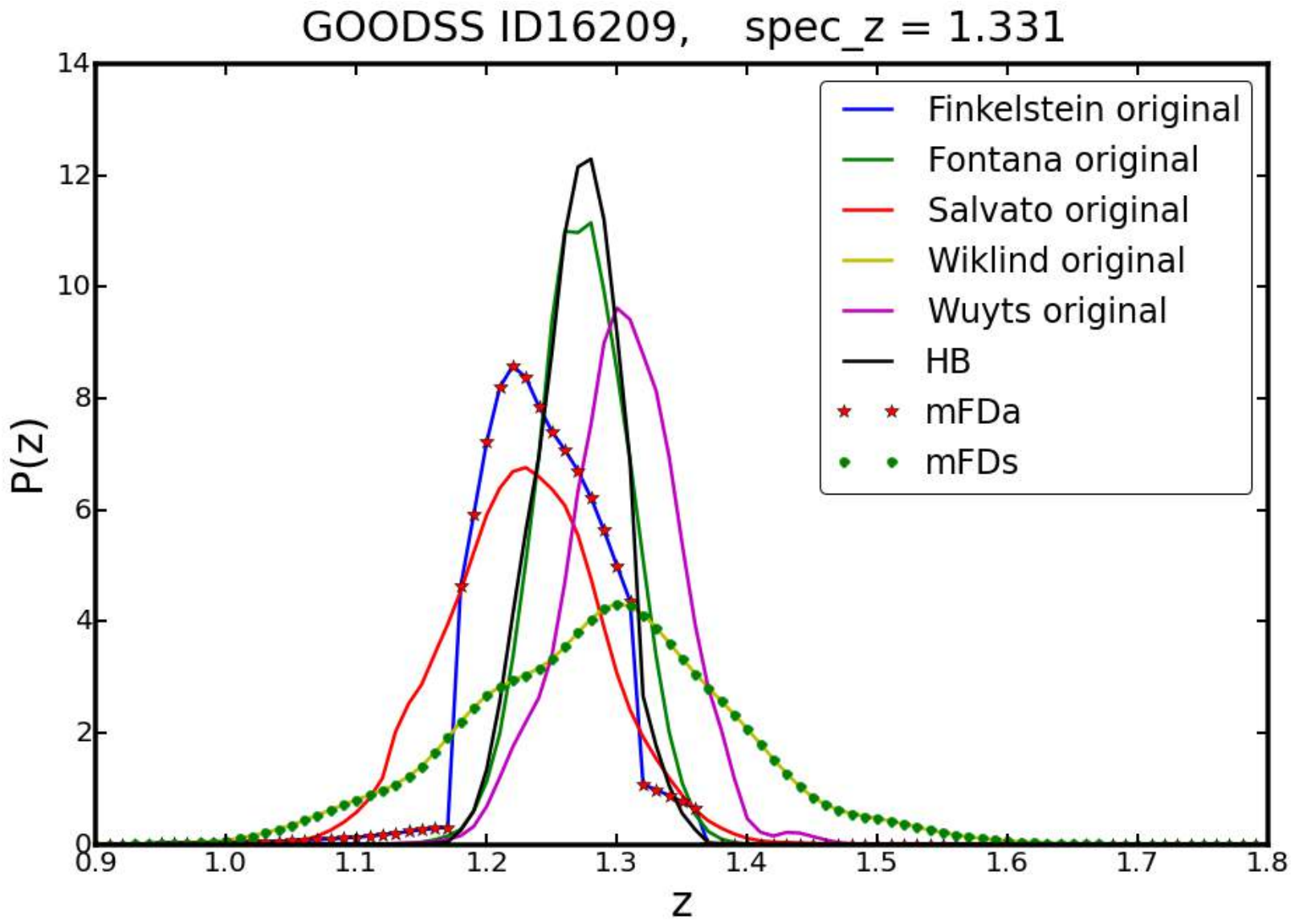
D. Kodra

Open issues: Combining PDF results from multiple codes



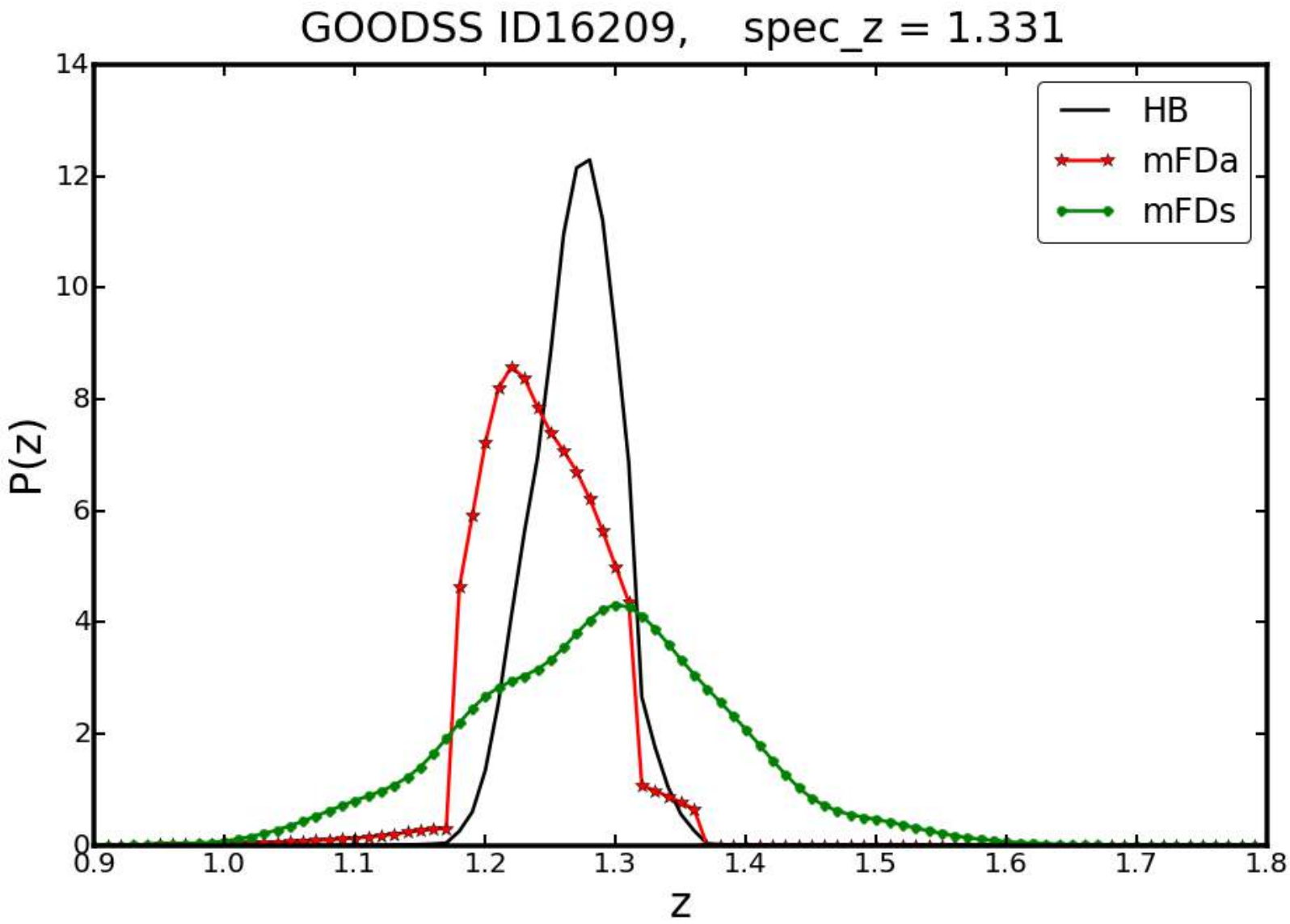
D. Kodra

Open issues: Combining PDF results from multiple codes



D. Kodra

Open issues: Combining PDF results from multiple codes

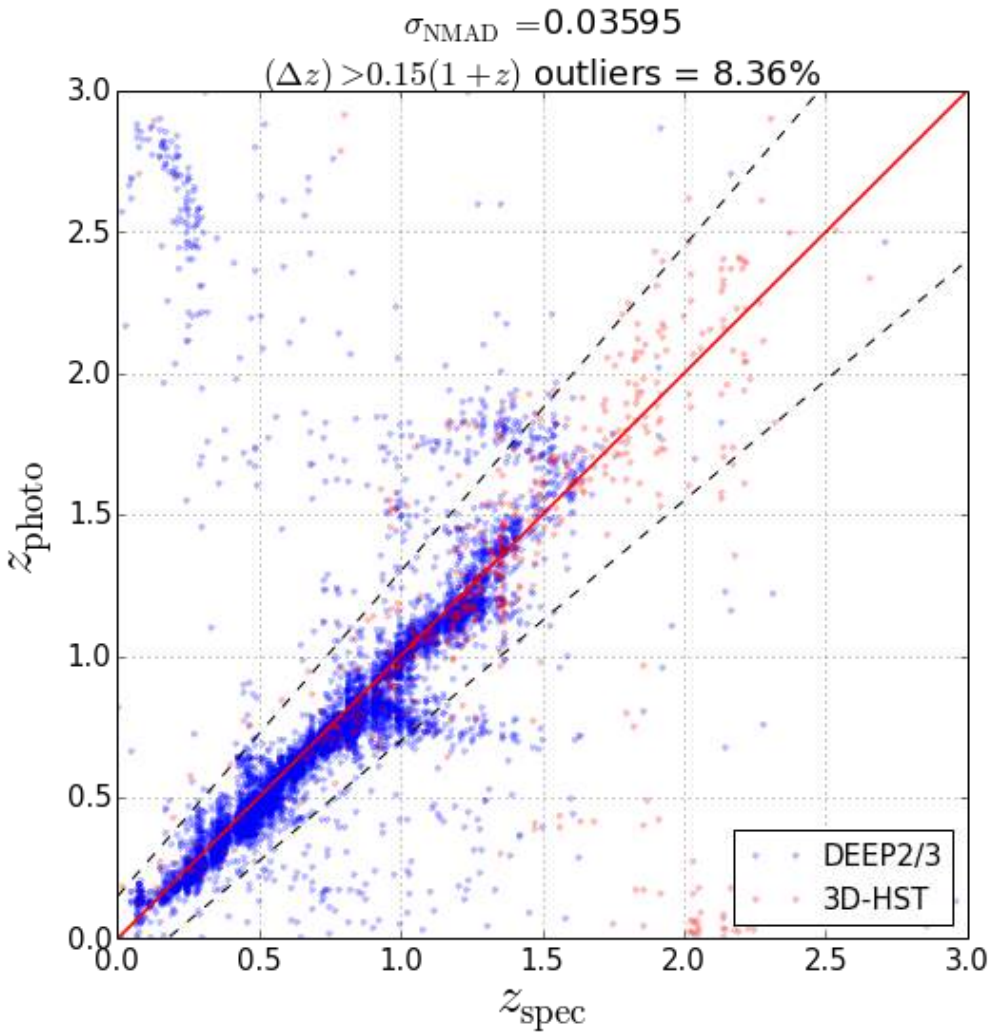


D. Kodra

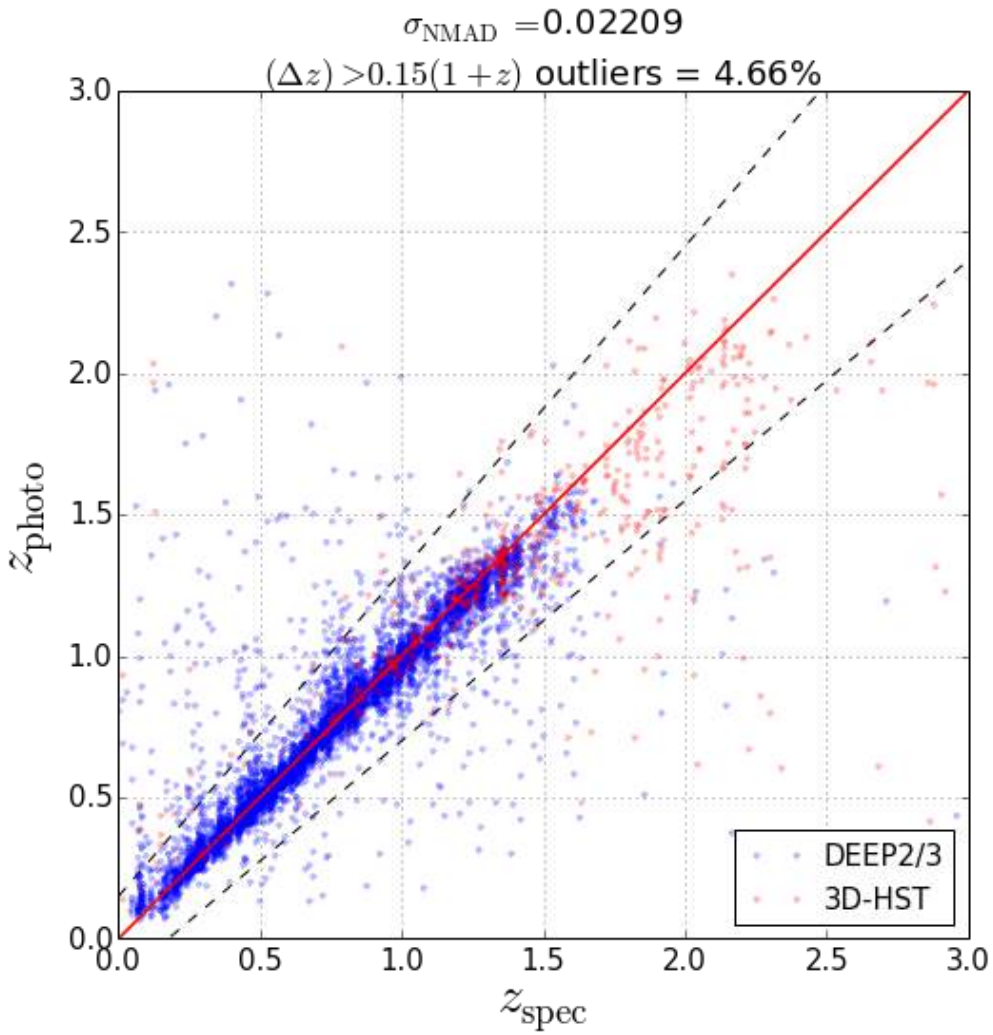
A related case: template-based and training-based methods have different failure modes - how best to combine?

- **Identify potential outliers from discrepant results?**

EAZY (template code, untuned)



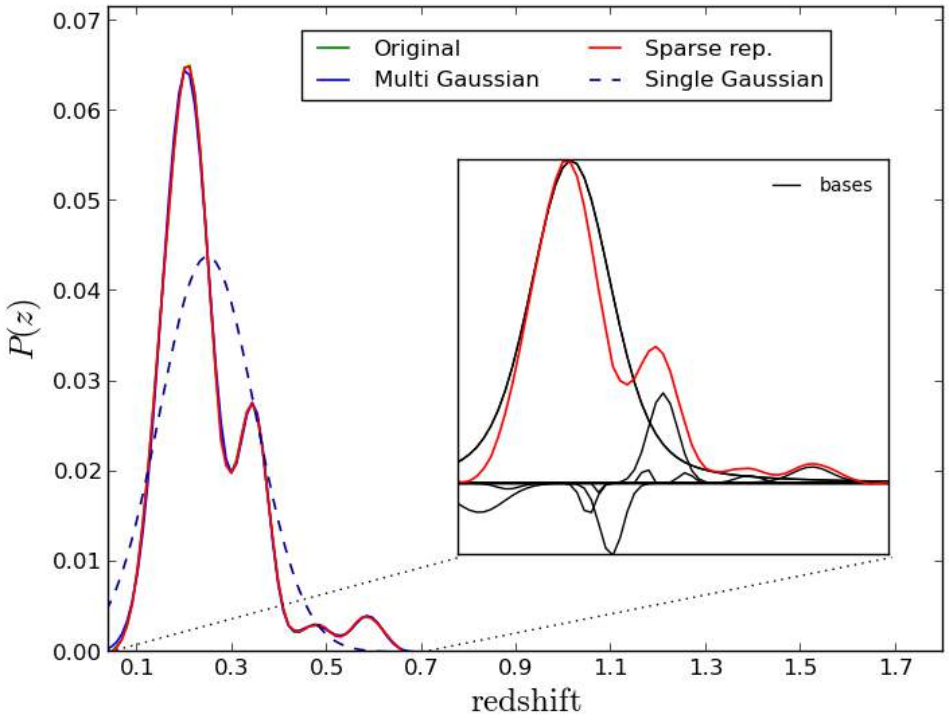
Random Forest Regression

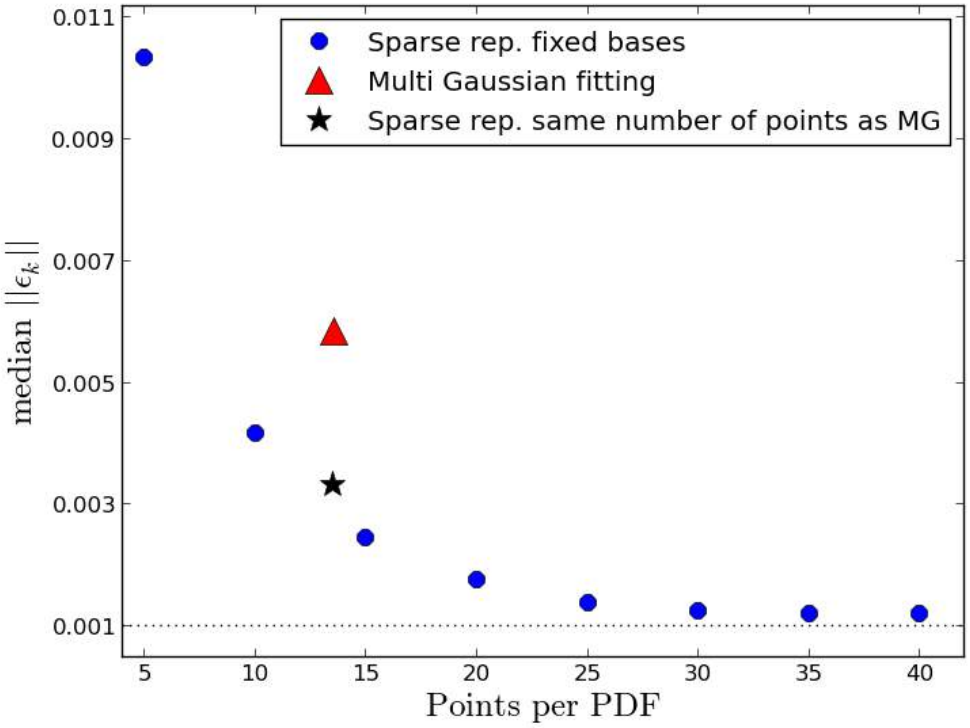


Zhou, JN et al. 2016, in prep.

Open issues: Storing p(*z*,*α*)

- Carrasco-Kind & Brunner 2014 achieved strong compression of photo-z PDFs using sparse representation and well-chosen basis set
- For many LSST applications, want 2+-dimensional PDFs
- Can suitably sparse (<few hundred #s) representations be achieved?
- Are samples from PDFs OK for all science cases?

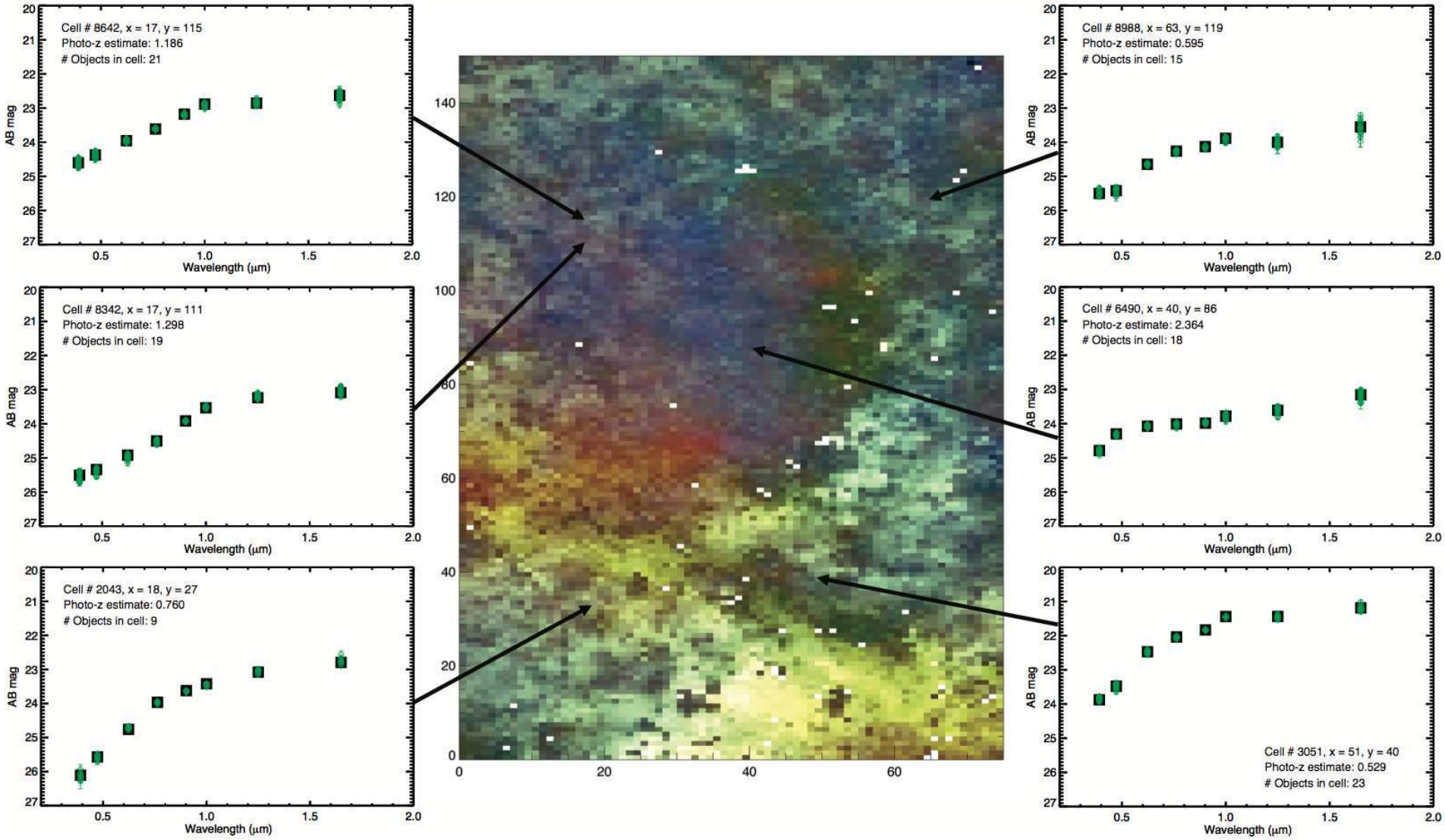




Carrasco-Kind & Brunner 2014

Open issues: Optimizing spectroscopic targeting

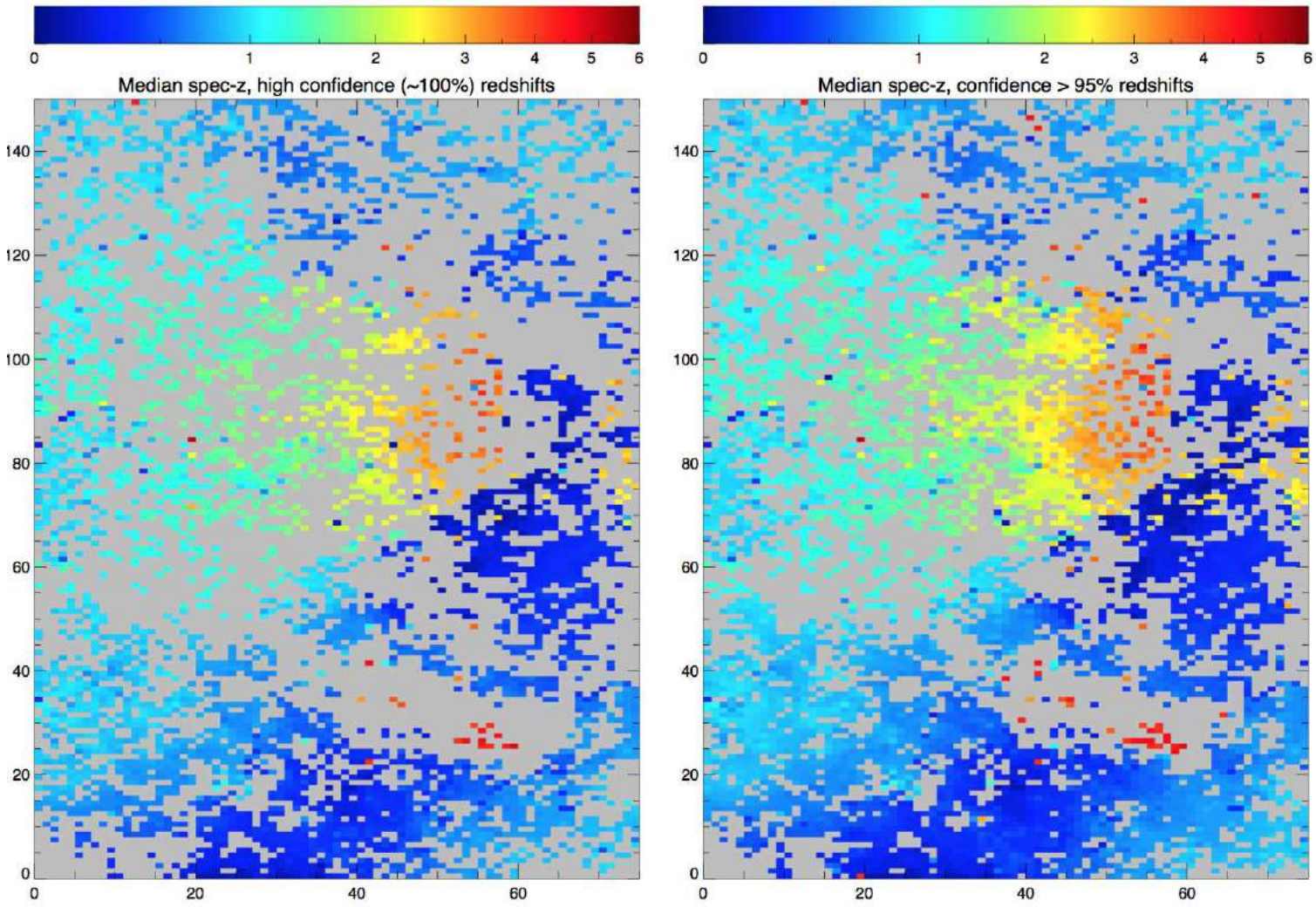
- **Current state of the art: Masters et al. 2015**
- **Self-organized map of galaxy colors**



Masters et al. 2015

Open issues: Optimizing spectroscopic targeting

- **Prioritize cells with few redshifts for spectroscopic follow-up**
- **Are there better ways to do this?**



Masters et al. 2015

Open issues: Ideal photo-z code?

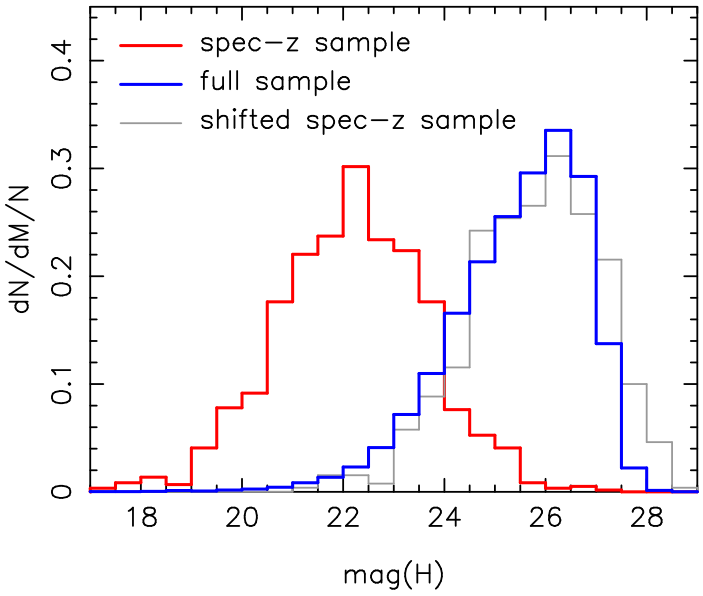
- What might an ideal LSST photo-z algorithm look like?
  - Trained with >30,000 spectra spanning range of spectra
  - Develops priors & tweaks templates via hierarchical Bayesian hyperparameters
  - Incorporates variations in effective filter wavelengths due to observational conditions: requires applying algorithm to O(1000) measurements instead of O(6)
  - Incorporates AGN classification and AGN photo-z determination: colors are not constant with time for many objects!
  - Want algorithms to be fast: create ML-based emulators for template photo-z's?
  - For bright objects, may also be useful to compare to ML techniques to identify potential outliers

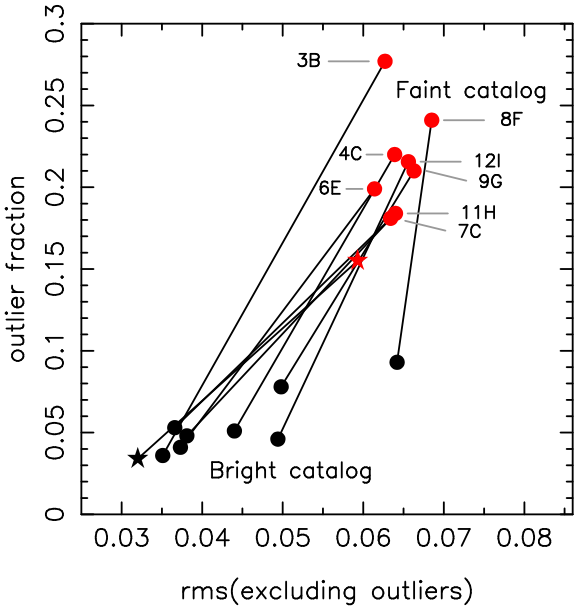
Outline

- Overview of photometric redshifts
  - Template methods
  - Training-based methods
- Requirements and resources for training and calibrating photometric redshifts
- Some open issues
  - Spectroscopic incompleteness
  - Robust training
  - p(z) coverage
  - Combining results from multiple codes
  - p(z,α) storage
  - Defining ideal LSST algorithm
  - Optimizing spectroscopic samples
- Some examples of problems with current codes

CANDELS code comparison: Dahlen et al. 2013

- **Many tests of photo-z algorithms with deep, high-redshift dataset. Examples:**
  - **Test photo-z performance as degrade photometry (using same test spectroscopic data)**
  - **Dependence of errors on redshift, magnitude, & color**
  - **Investigation of (lack of) consistency between photometric zero point shifts from different codes**
  - **Empirical test of photo-z errors using Δz between close pairs**



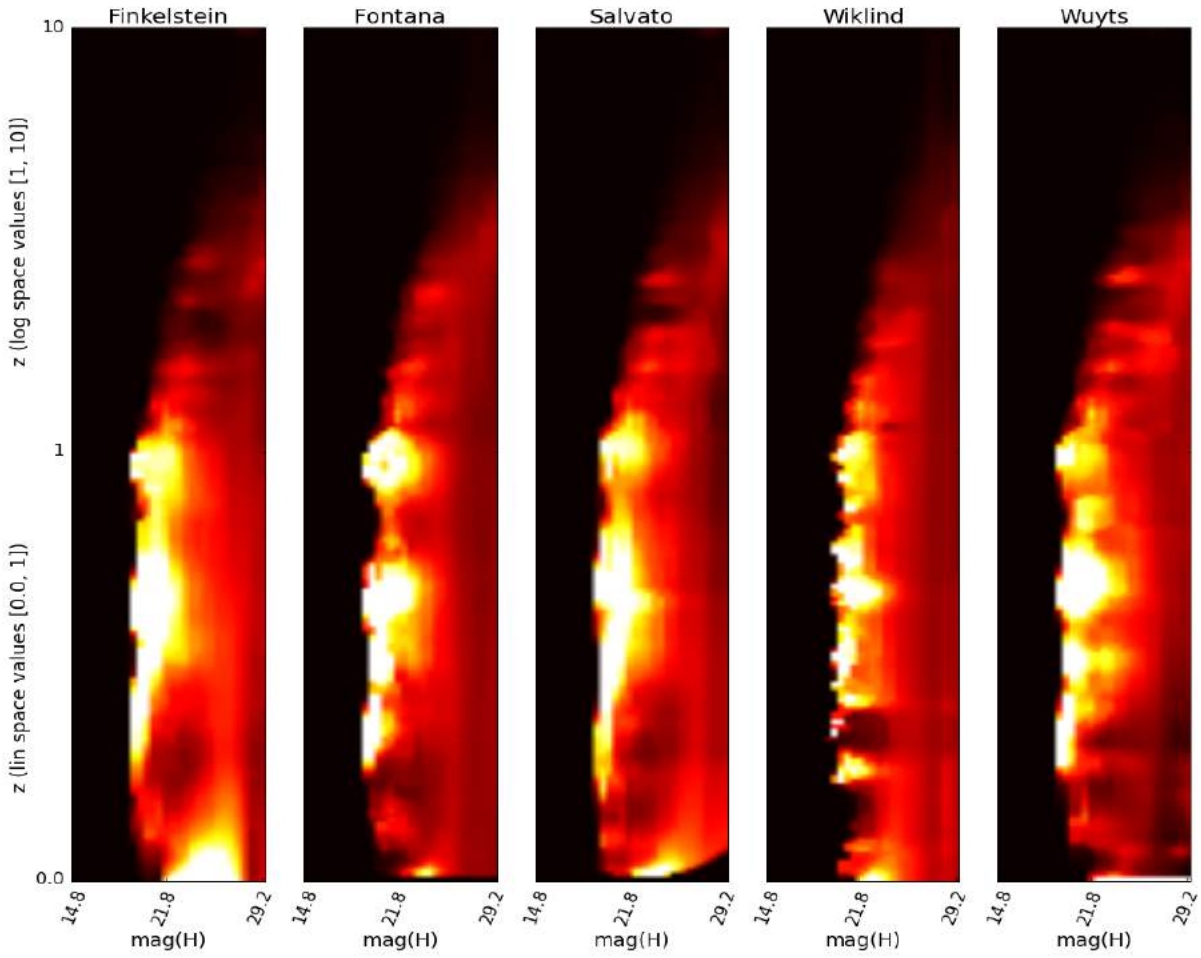


Dahlen et al. 2013

New work: Kodra et al. 2019

- Compare predictions of codes in space of p(z | H)

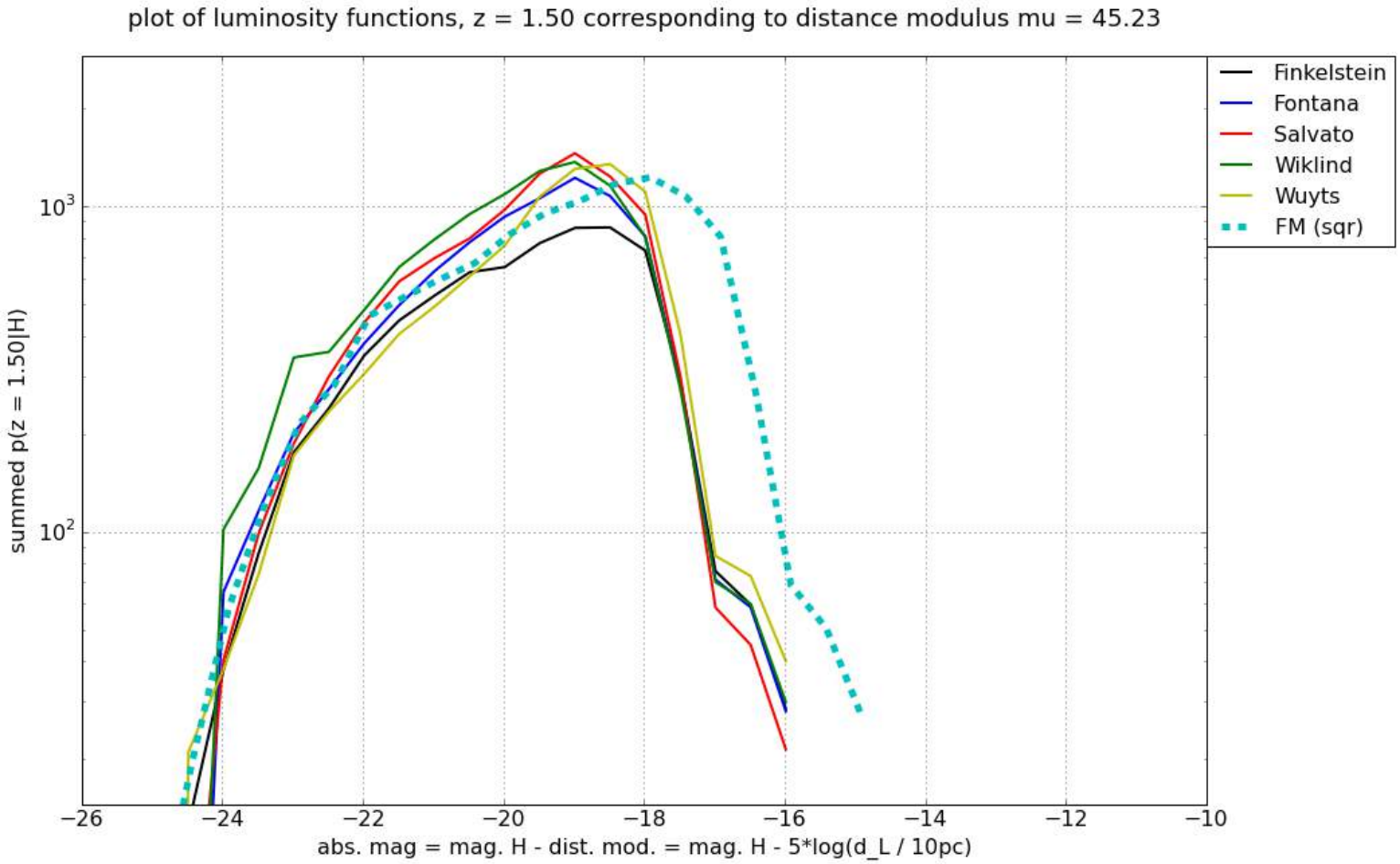
- Disagreement on where there are redshift spikes
- Priors have huge effect at low z (non-monotonic behavior)
- Different effective smoothings
- The performance of these codes for z_{peak} isn't all that different. . .



D. Kodra

New work: Kodra et al. 2019

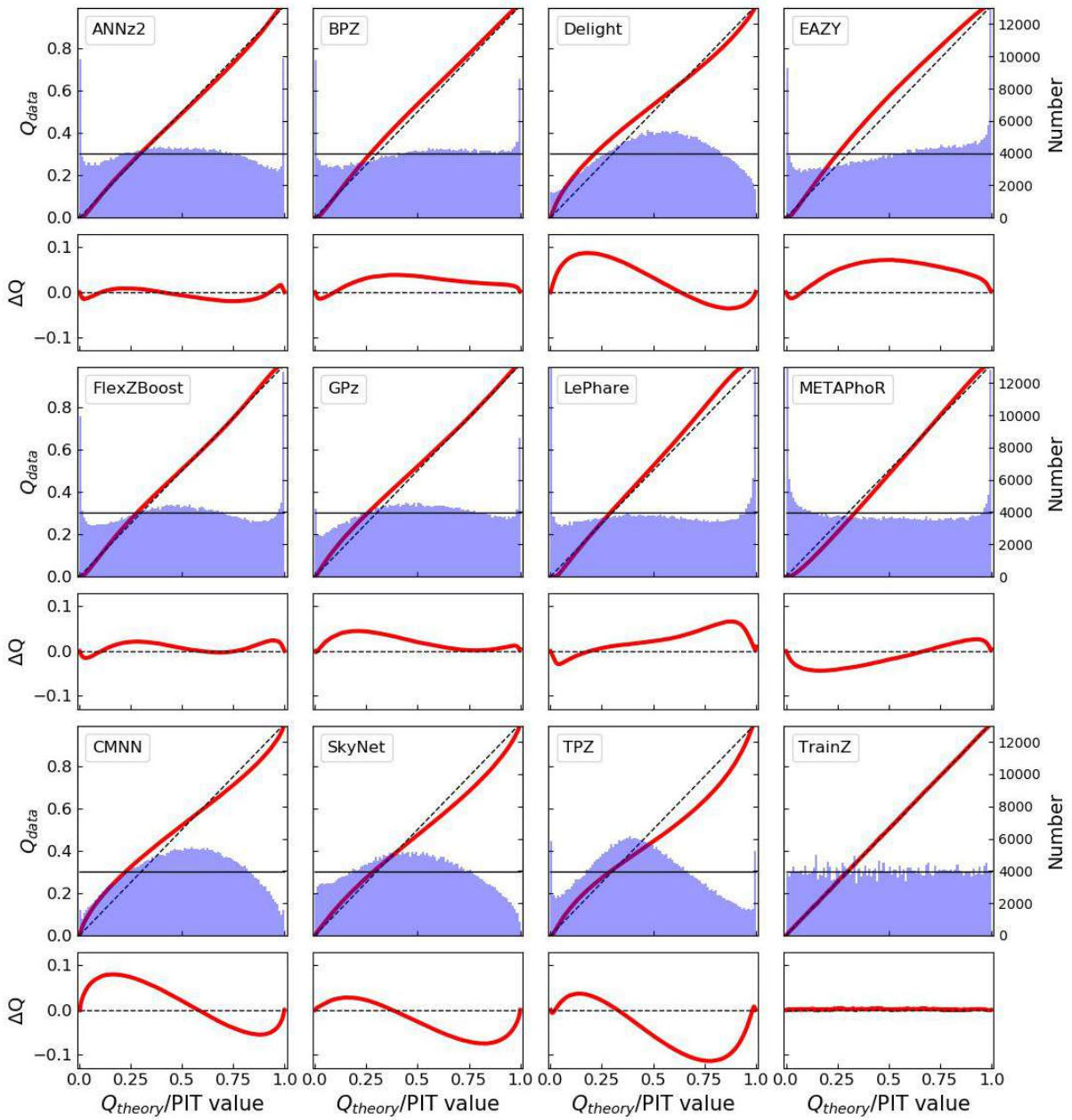
- **This can have large (factor of few) effects on the inferred number of objects at a given redshift**



D. Kodra

New work: Schmidt, Malz et al. 2019

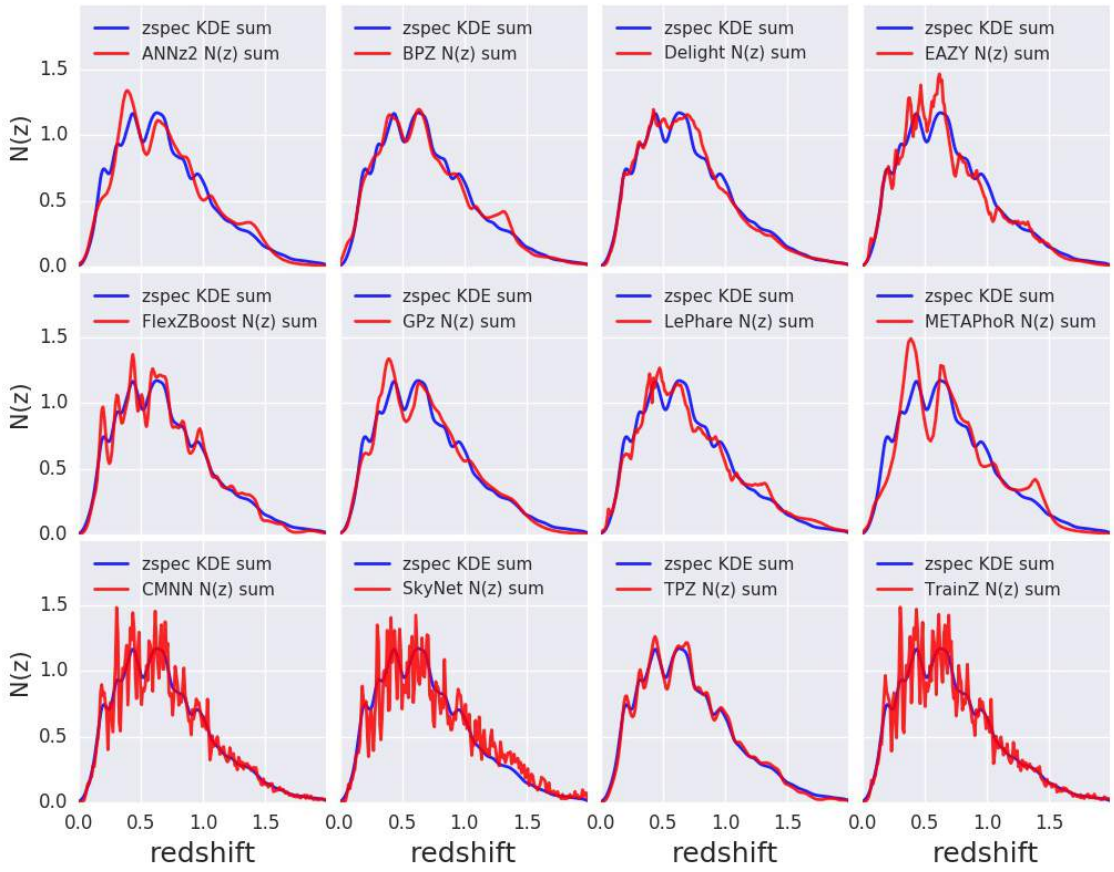
- Testing a dozen photo-z codes with large, representative training sets, and full template knowledge and priors passed to algorithms
- Still fail to yield *p(z)* which meet the statistical definition of a probability distribution (assessed via Q-Q statistics and Probability Integral Transform [PIT])



S. Schmidt

New work: Schmidt, Malz et al. 2019

- **Substantial variation in stacked *p(z)* among algorithms**



S. Schmidt

Conclusions

- Training-based methods are easier to get good results from than template-based methods, but don't extrapolate well
- Key issues for LSST are where to get deep training sets, and inability to get complete training sets
- A variety of interesting problems to work on in the near future
- Current codes appear sufficient to meet LSST requirements, but are clearly suboptimal. Better photo-z's will greatly increase the value of LSST - e.g. 40% increase in Dark Energy Figure of Merit

Spectroscopic training set requirements

- Goal: make δ_z and $\sigma(\sigma_z)$ so small that systematics are subdominant
- Many estimates of training set requirements (Ma et al. 2006, Bernstein & Huterer 2009, Hearin et al. 2010, LSST Science Book, etc.)
- General consensus that roughly 20k-30k extremely faint galaxy spectra are required to characterize:
  - Typical z_{spec}-z_{phot} error distribution
  - Accurate catastrophic failure rates for all objects with $z_{phot} < 2.5$
  - Characterize all outlier islands in z_{spec}-z_{phot} plane via targeted campaign (core errors easier to determine)

What qualities do we desire in our training sets?

- **Sensitive spectroscopy of faint objects (to *i*=25.3)**

  - Need a combination of large aperture and long exposure times from the ground; >20 Keck-nights (=4 GMT-nights) equivalent per target, minimum

- **High multiplexing**

  - Obtaining large numbers of spectra is infeasible without it

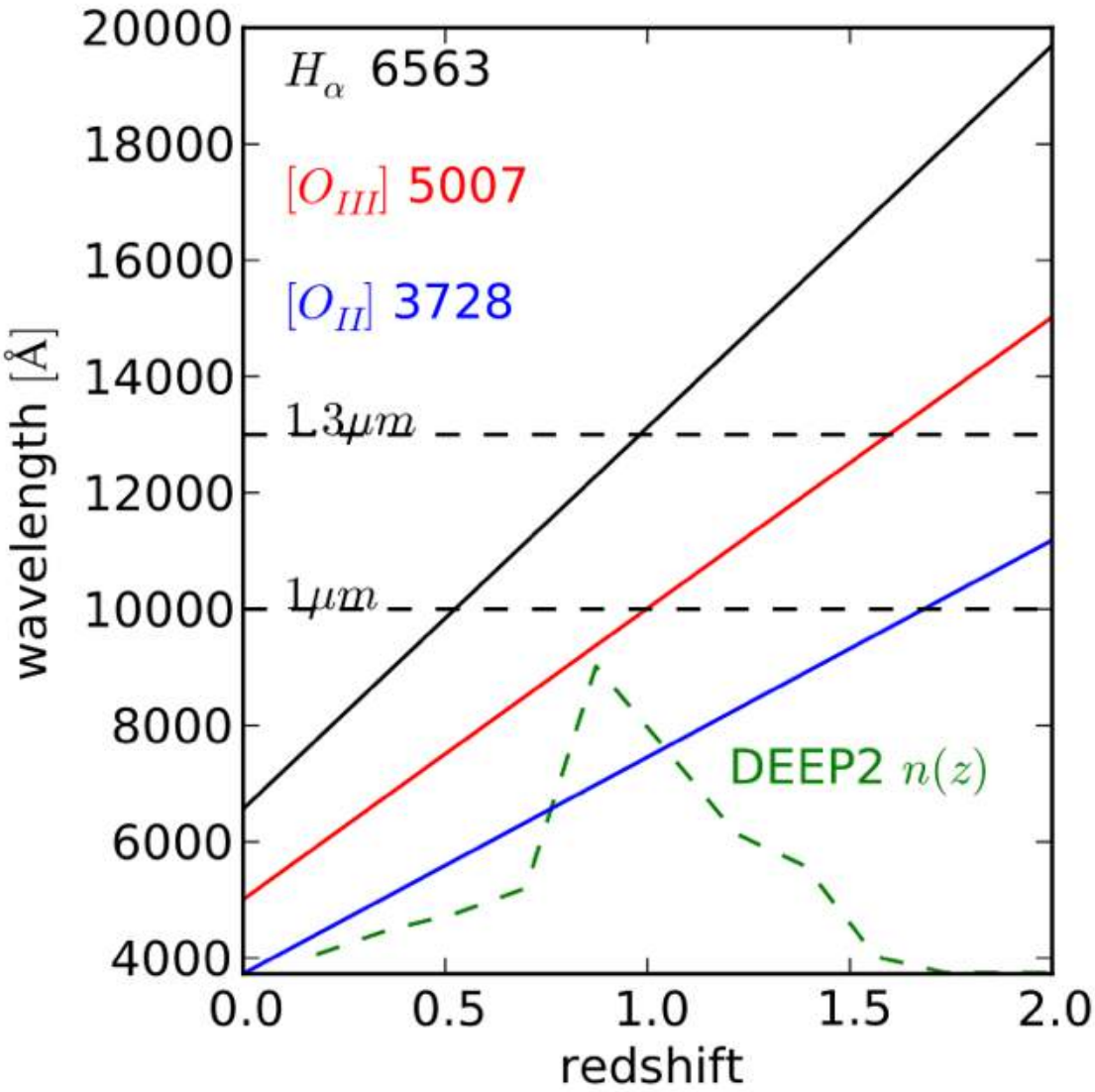
See Newman et al. 2015, *Spectroscopic Needs for Imaging Dark Energy* Experiments, for details

What qualities do we desire in our training sets?

- **Coverage of full optical window if working from the ground**

  - Ideally, from below 4000 Å to ~1.5µm

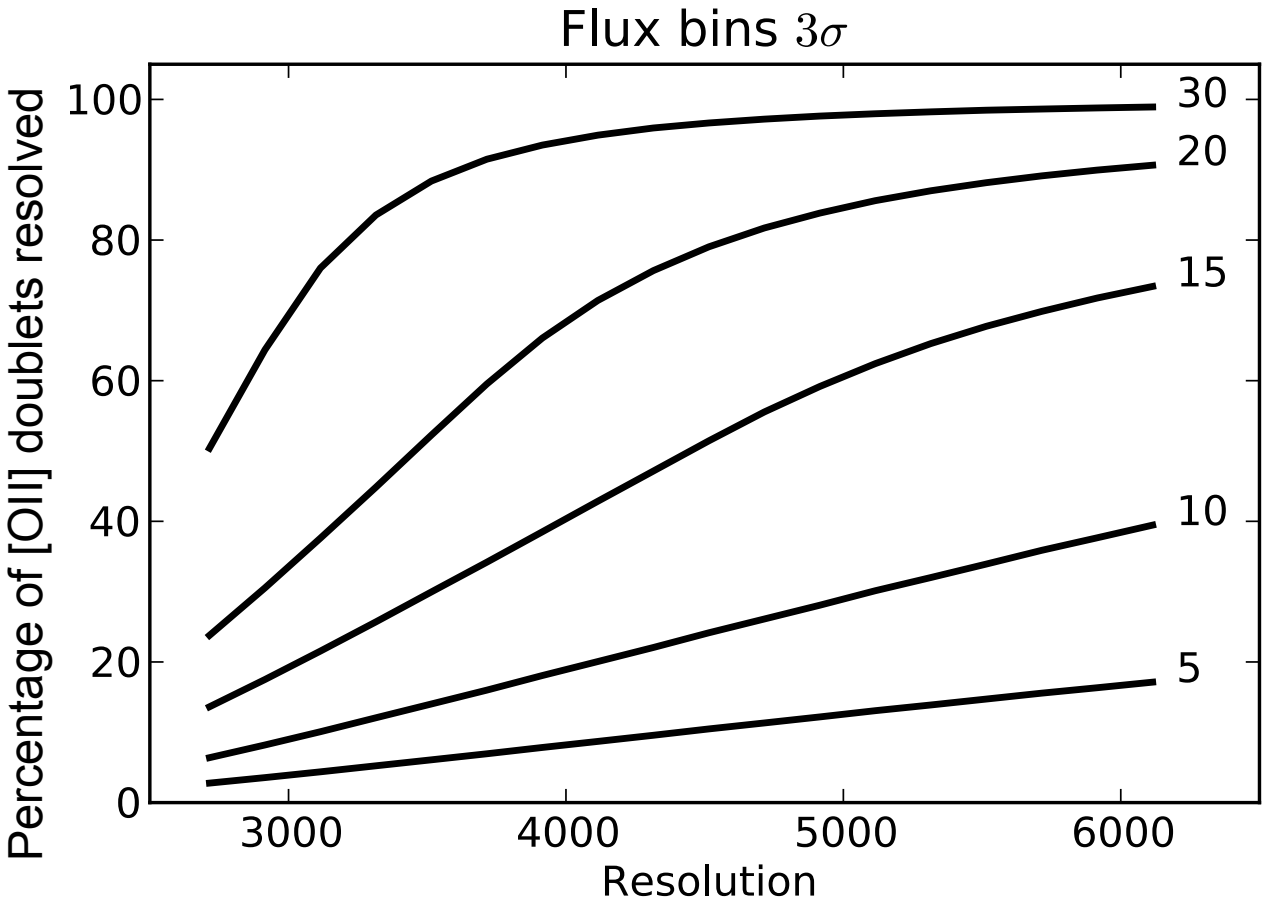
    - Require multiple features for secure redshift



Comparat et al. 2013, submitted

What qualities do we desire in our training sets?

- **Significant resolution (R>~4000) at red end if working from the ground**
  - **Allows redshifts from [OII] 3727 Å doublet alone, key at *z*>1**
  - **Not necessary if get multiple features from deep IR coverage**



Comparat et al. 2013

What qualities do we desire in our training sets?

- **Field diameters > ~20 arcmin**

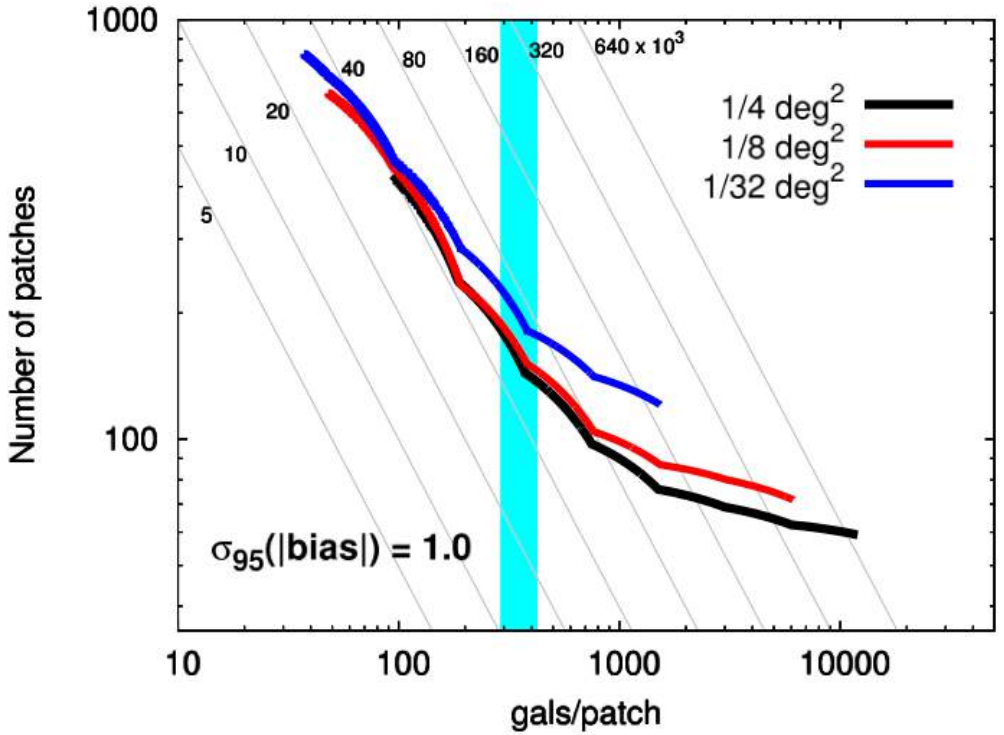
  - Need to span several correlation lengths for accurate clustering measurements (key for galaxy evolution science and cross-correlation techniques)

  - r_0 ~ 5 h^{-1} Mpc comoving corresponds to ~7.5 arcmin at z=1, 13 arcmin at z=0.5

- **Many fields**

  - Minimizes impact of sample/ cosmic variance.

    - e.g., Cunha et al. (2012) estimate that 40-150 ~0.1 deg² fields are needed for DES for sample variance not to impact errors (unless we get clever)



Cunha et al. 2012

How much time would be required to complete surveys from the Najita et al. Kavli/NOAO/ LSST report on different platforms?

- This is an attempt to take the largest surveys proposed in the Kavli report and work out how long would be needed to do them
- Common set of assumptions: one-third loss to instrumental effects, weather and overheads; 4m = Mayall/DESI; 8m = Subaru/PFS; all instrumental efficiencies identical; equivalent # of photons will yield equal noise; ignoring differences in seeing/image quality and fiber/slitlet size. Only medium-resolution fibers included. Assuming full spectral range can be covered simultaneously (likely not true for E-ELT).
- See report (available at http://arxiv.org/abs/1610.01661) for details of these surveys
- Will give time in years on each platform; note that this is generally dark time (very faint targets!)
- Costs based on TSIP + inflation: $1k/m^2/night