What I worry about when I worry about photo-z's

Jeffrey Newman, U. Pittsburgh / PITT-PACC

Many people assume photo-z codes provide a statistical PDF for the redshift of each object... that is not currently the case

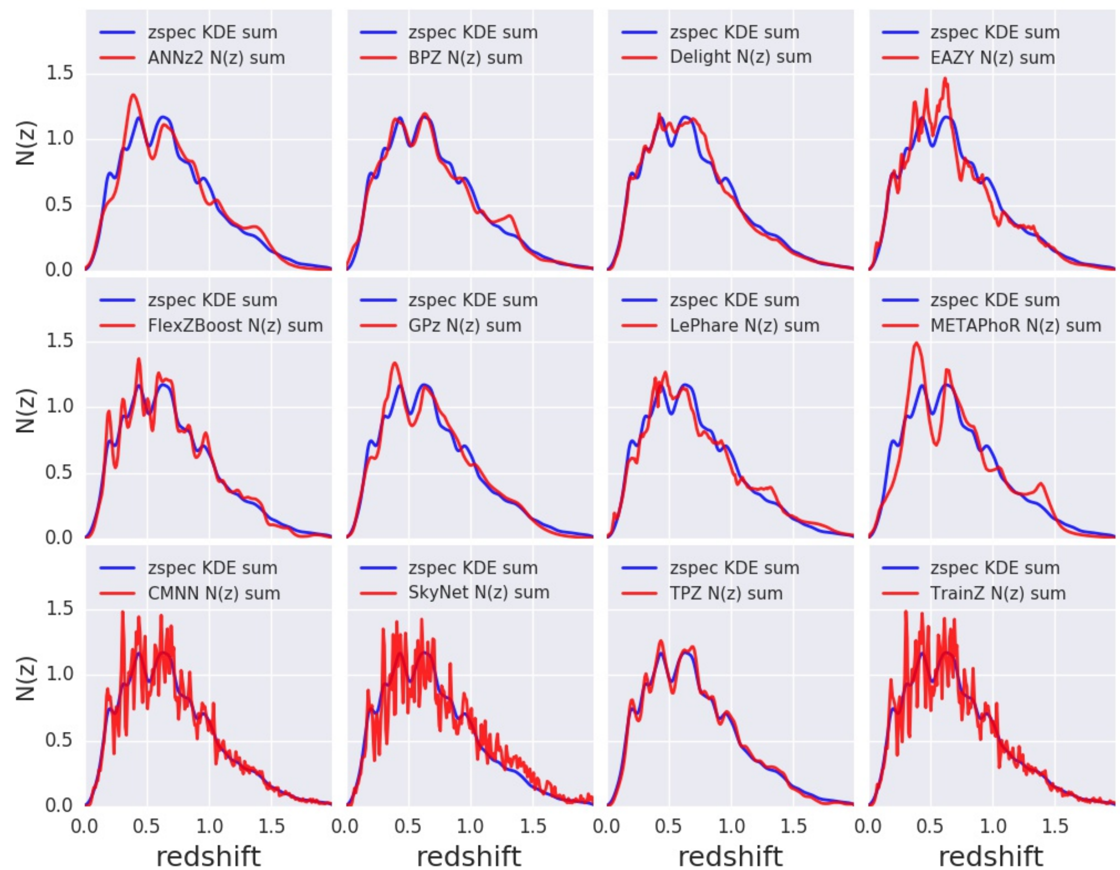
- **Dahlen et al. 2013 tested the fraction of spectroscopic redshifts that are in the inner 68% or inner 95% of their PDFs for CANDELS photo-z's**
- **Coverage is all over the place; no codes were good at both 68% and 95% points**

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

Many people assume photo-z codes provide a statistical PDF for the redshift of each object... that is not currently the case

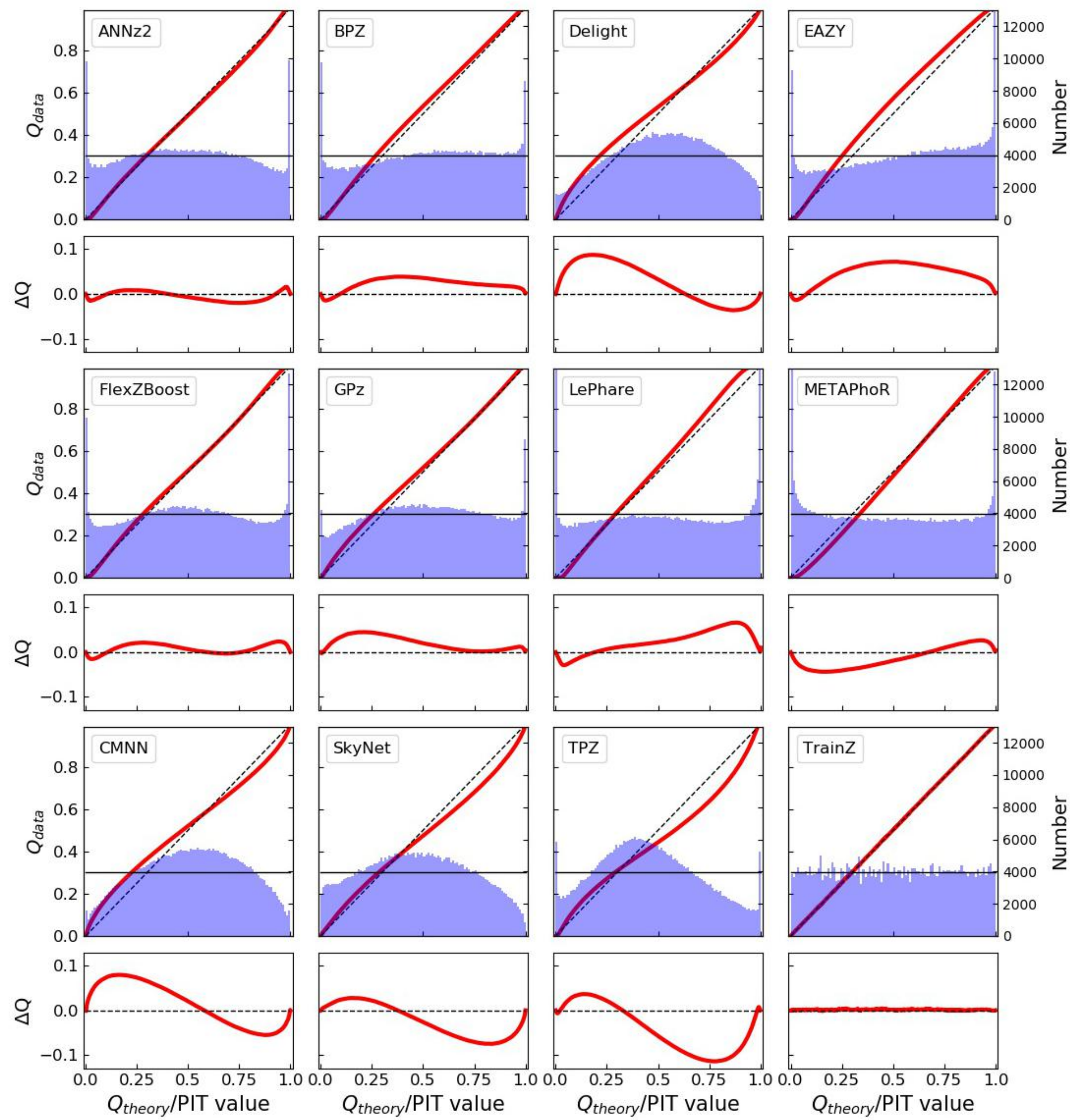
- **Many dark energy probes use per-object redshift probability distribution function (*p(z)*) information**
- **Schmidt, Malz et al. 2019: Testing a dozen photo-z codes with large, representative training sets, and full template knowledge and priors passed to template-based algorithms**
- **Substantial variation in stacked *p(z)* among algorithms (though talk to Alex Malz about why you shouldn't do that for science!)**



S. Schmidt

Many people assume photo-z codes provide a statistical PDF for the redshift of each object... that is not currently the case

- **Even when given perfect training sets and template knowledge, codes 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])**
  - **Except for degenerate 'TrainZ' algorithm that just uses input *z* distribution as *p(z)*: gives bad predictions for individual objects**

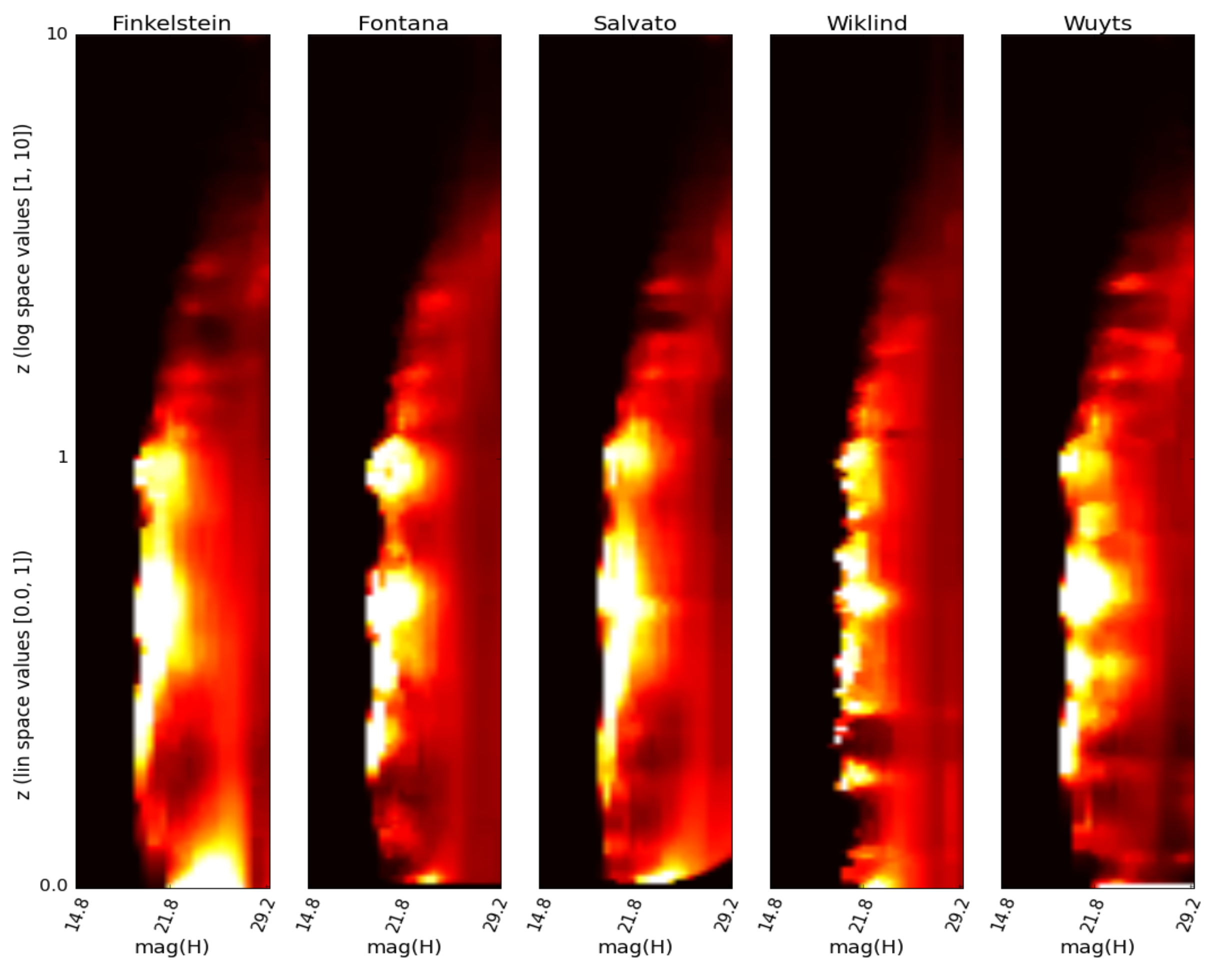


S. Schmidt

Codes that have good performance when assessed by spectroscopic redshifts can disagree greatly even when applied to the same data

- **Kodra et al. 2019: compares predictions of CANDELS codes in space of *p*(z | H): a test that requires no spectroscopy**

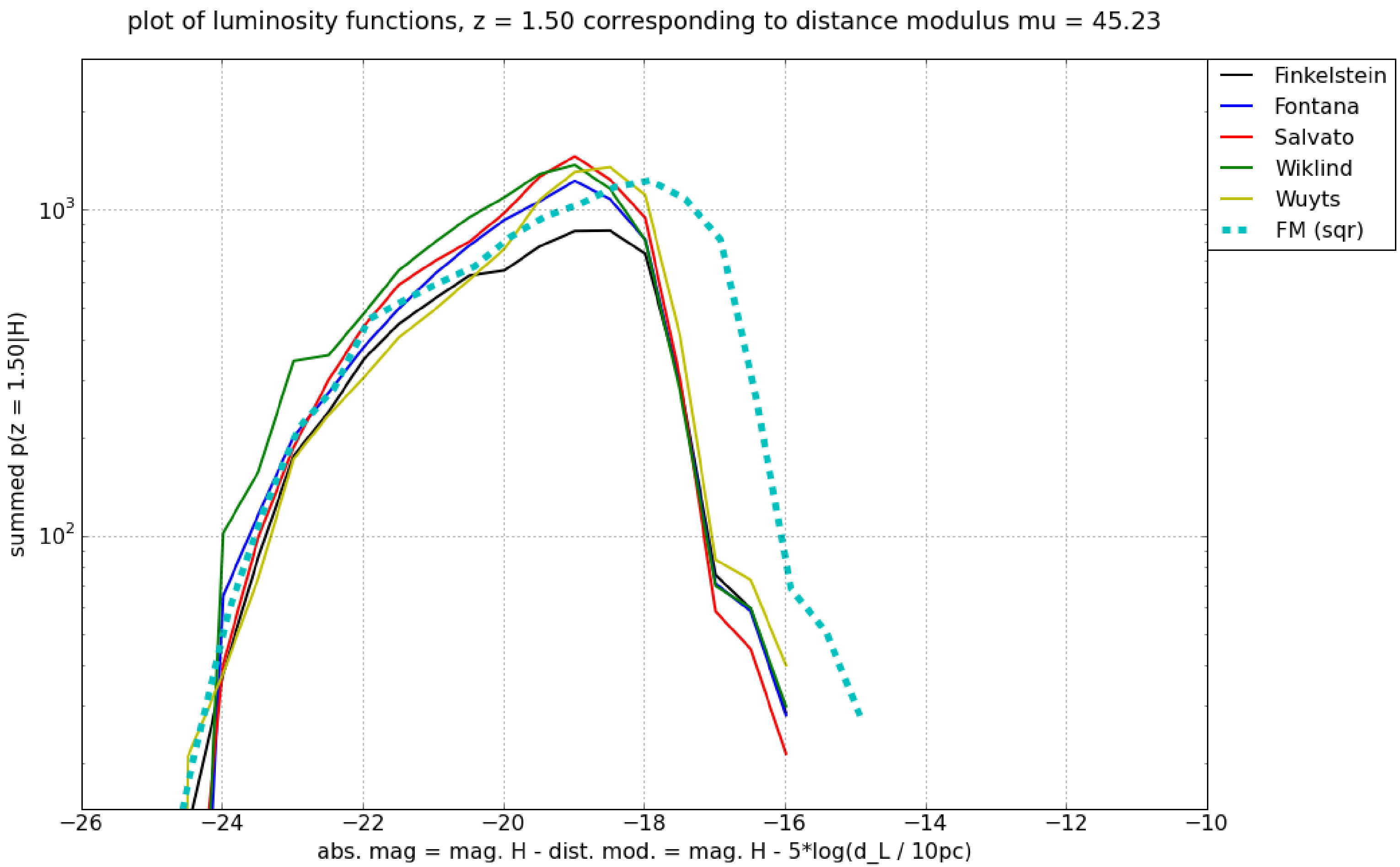
- **Disagreement on where there are redshift spikes**
- **Priors have huge effect at low z (non-monotonic behavior)**
- **Different effective smoothings**
- **The performance of these codes for z_{peak} isn't all that different. . .**



D. Kodra

Codes that have good performance when assessed by spectroscopic redshifts can disagree greatly even when applied to the same data

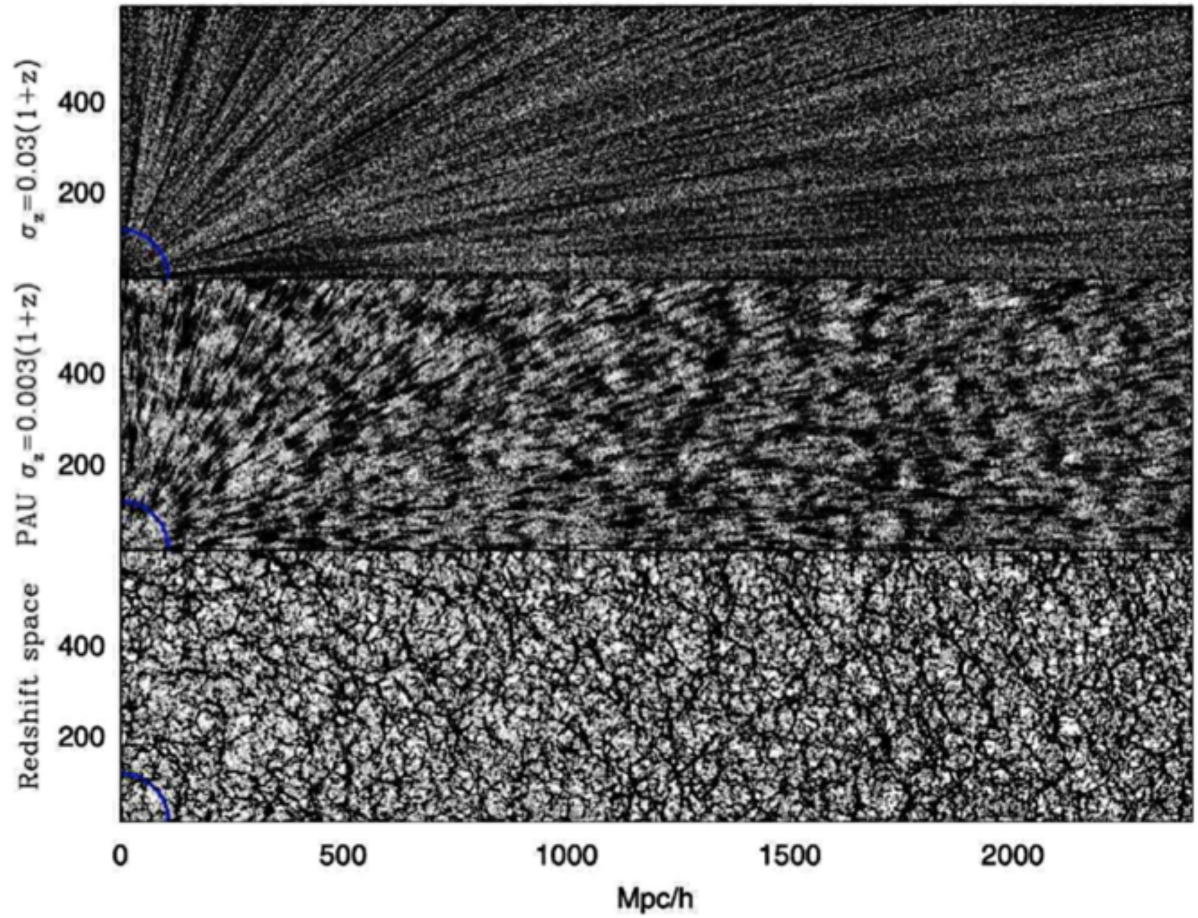
- **This can have large (factor of few) effects on the inferred number of objects at a given redshift**



D. Kodra

Spectroscopic samples can be used for training photo-z algorithms, making them better

- Training: optimization of algorithms using sets of objects with spectroscopic redshift measurements
- Basis of all machine learning algorithms (including SOM), but useful for template methods too
- Better training shrinks photo-z errors for individual objects: training *improves* photo-z's, *makes them better*

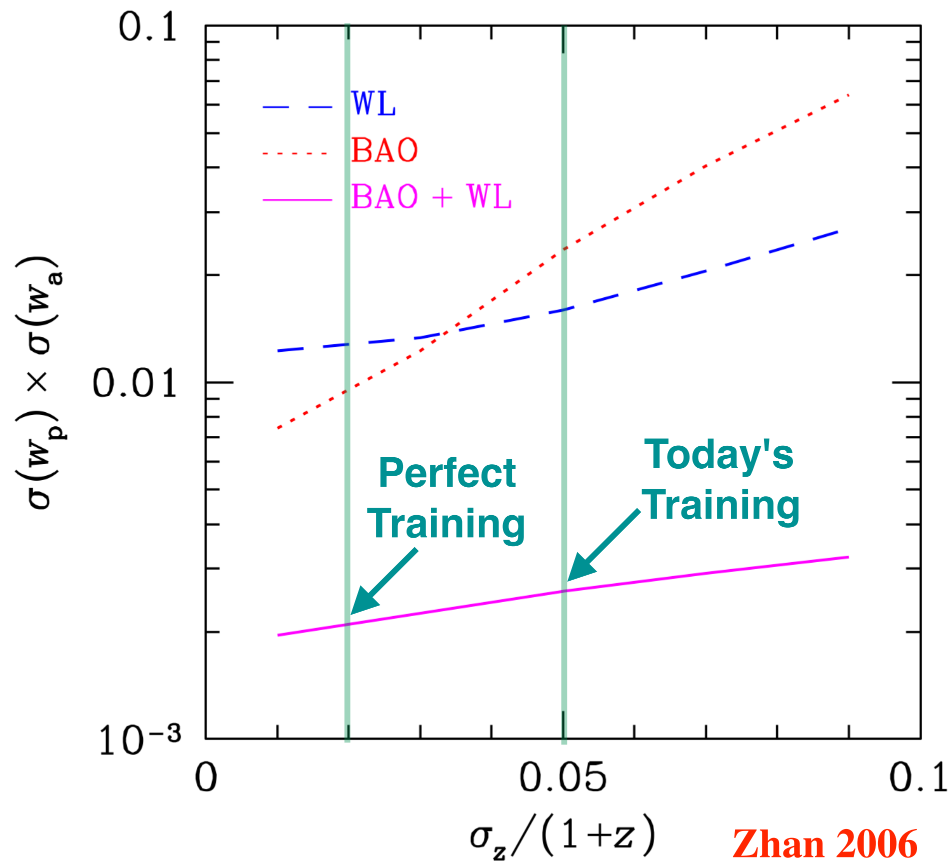


Benitez et al. 2009

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

Improved photometric redshift training can increase the science from imaging experiments like LSST

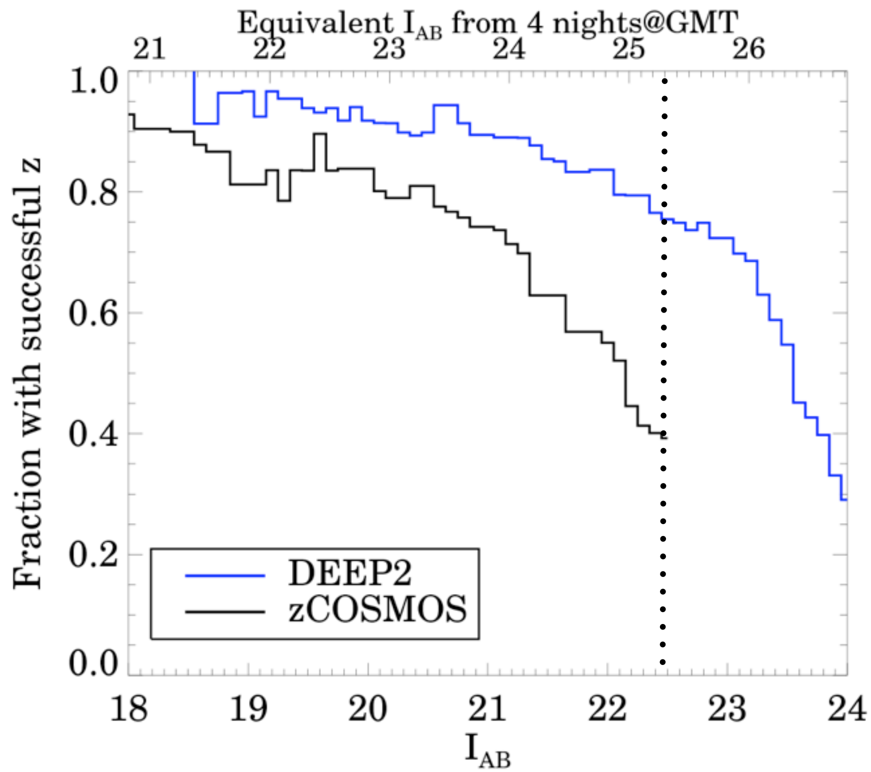
- Smaller photo-z errors from better-trained algorithms using representative samples of galaxies with spectroscopic redshifts can improve dark energy constraints, especially for BAO and clusters



- 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%

Based on past experience, our training sets may be systematically incomplete

- **In existing deep samples, a significant fraction (>20%) of faint galaxies fail to yield secure spectroscopic redshifts**
- **Spectral features must be outside wavelength range covered or be weak**
- **Broader wavelength coverage from new instruments should help, but how much?**
- **If we want to use training redshifts for calibration (e.g. KIDS 'Direct' method), need >99% - >99.9% completeness**
  - **Long exposure times are needed to ensure even >75% redshift success rates for upcoming projects: ~180 hours at Keck to achieve DEEP2-like S/N at *i*=25.3 LSST lensing limit**
  - See http://adsabs.harvard.edu/abs/2015APh....63...81N



Newman et al. 2015

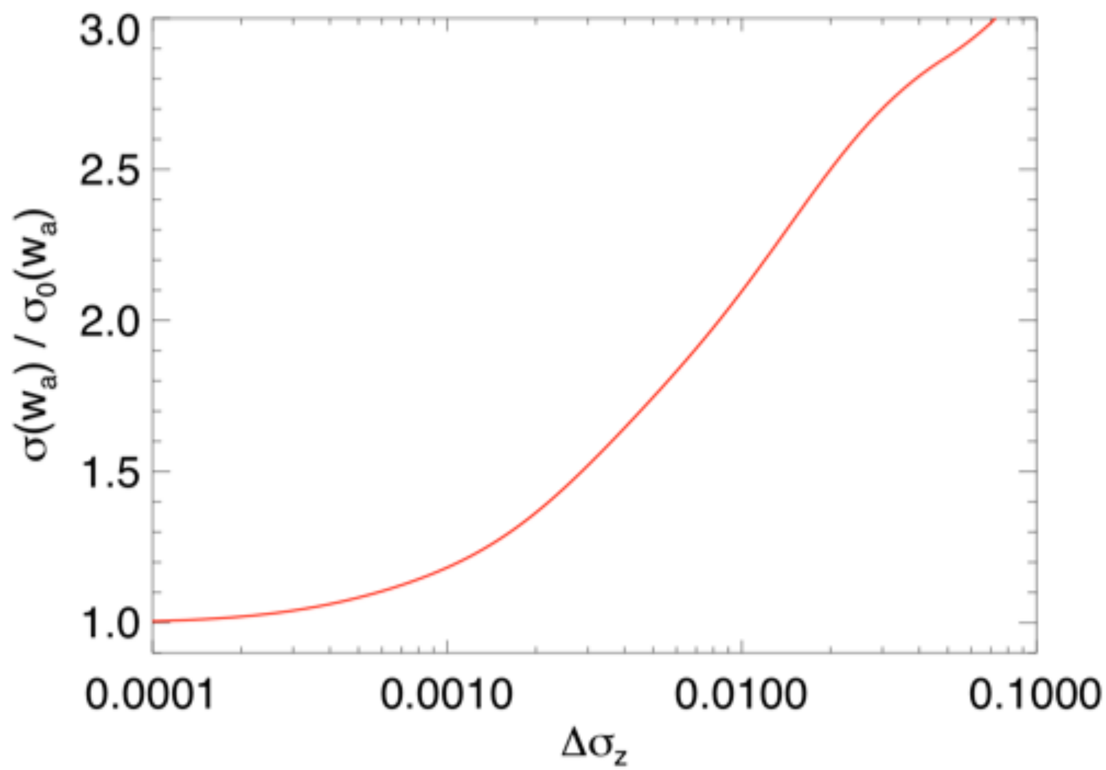
Biggest concern about training photo-z's: how will we get the telescope access for the faint samples LSST + WFIRST need?

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
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*

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
- Poor training causes increased random errors; poor calibration causes systematic errors
  - *uncertainty in* bias, $\sigma(\delta_z)= \sigma(<z_p - z_s>)$, and in scatter, $\sigma(\sigma_z)= \sigma(RMS(z_p - z_s))$, must both be <~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)



Newman et al. 2015

For direct calibration, even with 100% complete samples, current false-z rates can compromise calibration accuracy

- Only the highest-confidence redshifts should be useful for precision calibration: lowers spectroscopic completeness further when restrict to only the best
- Estimates of width of distribution are particularly sensitive to outliers:
  - For a σ=0.1 sample, **one Δ*z*=1 outlier in a thousand redshifts biases recovered σ by 0.005!** (0.001 effect on mean *z*)

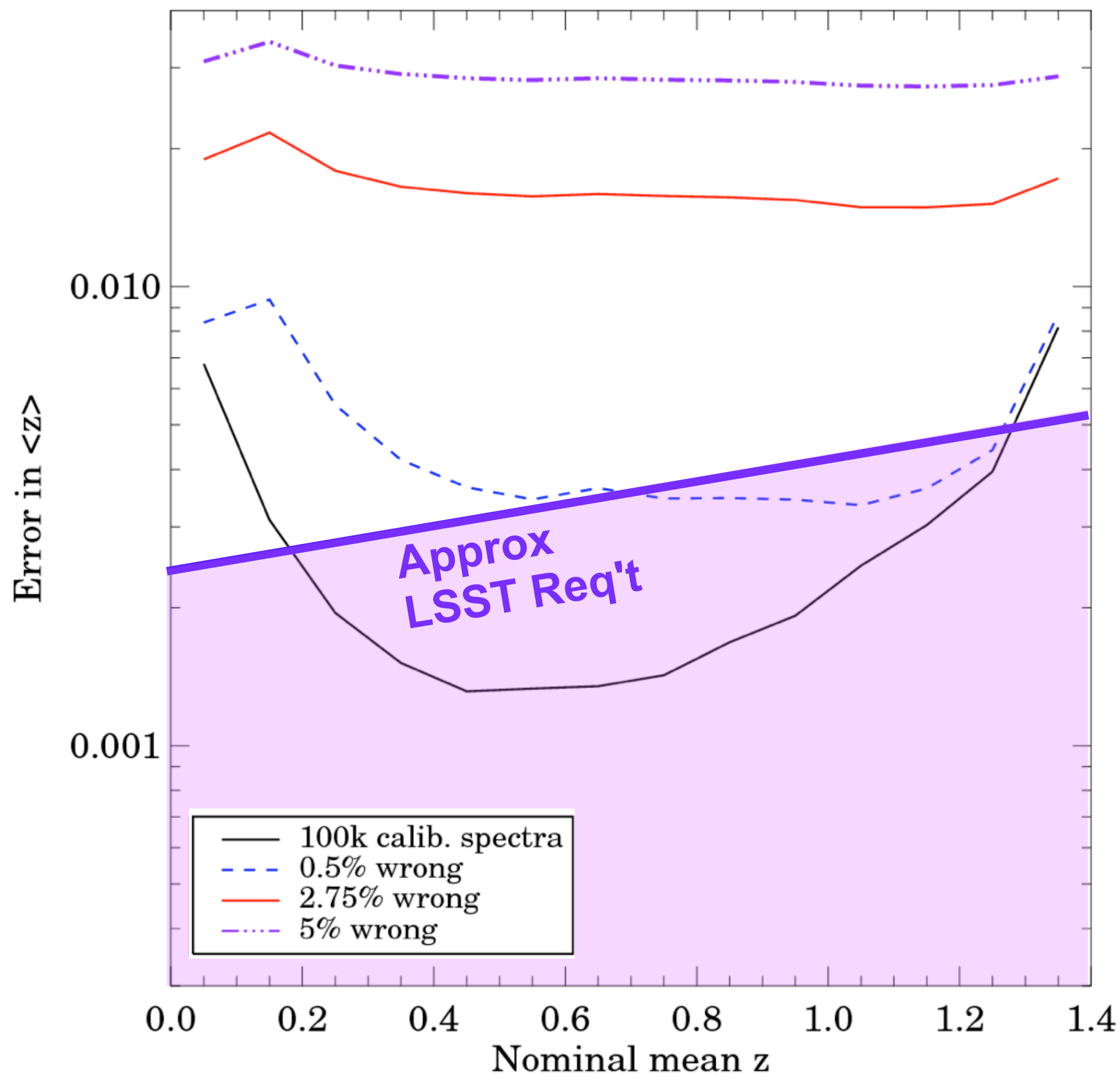
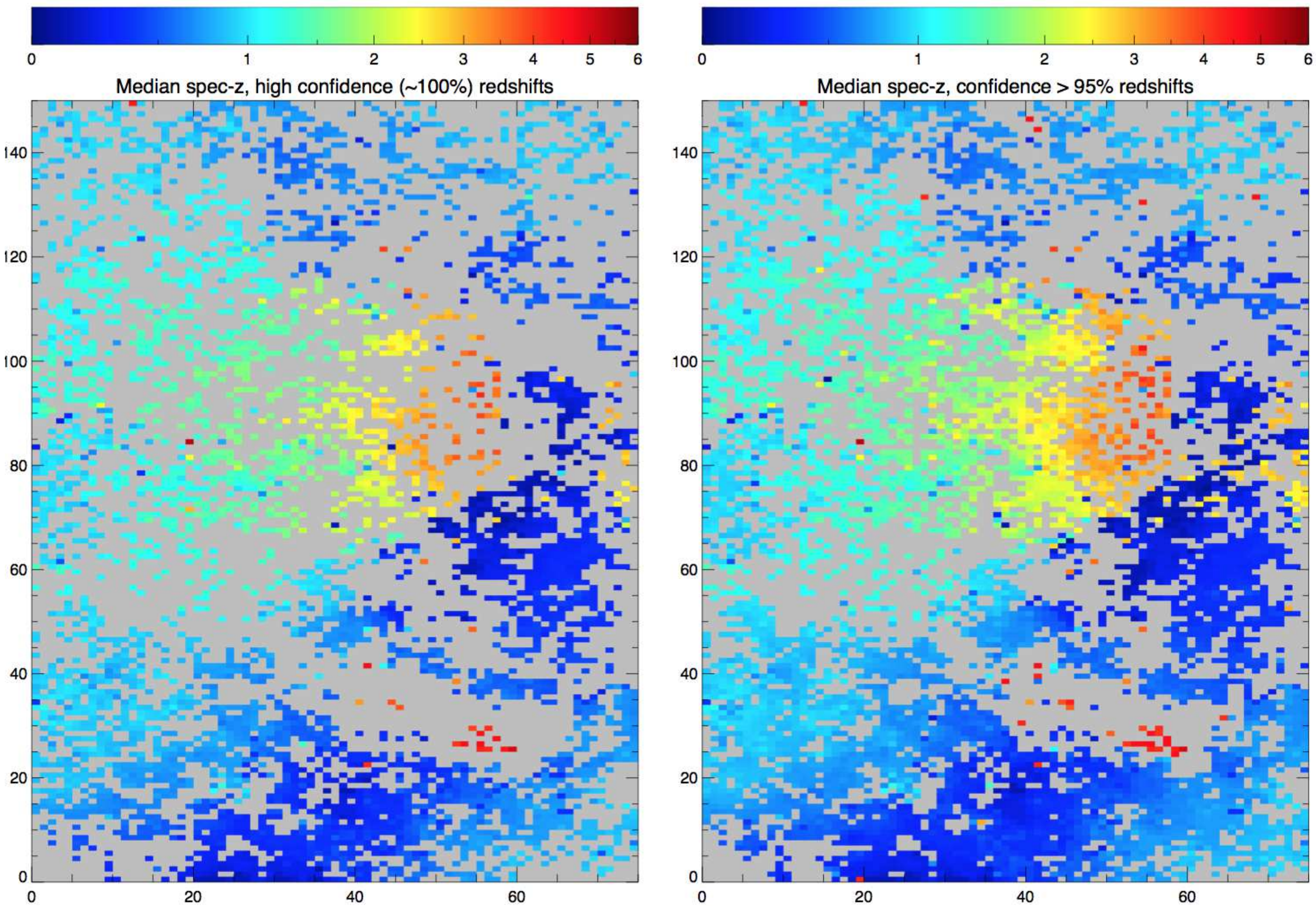


Figure based on simulated redshift distributions for ANNz-defined DES bins in mock catalog from Huan Lin, UCL & U Chicago, provided by Jim Annis

If we restrict to the highest-confidence redshifts, much more of color space is untrained

- **Grey regions: cells in self-organized maps of galaxy color space that are not constrained by spectroscopic redshifts**

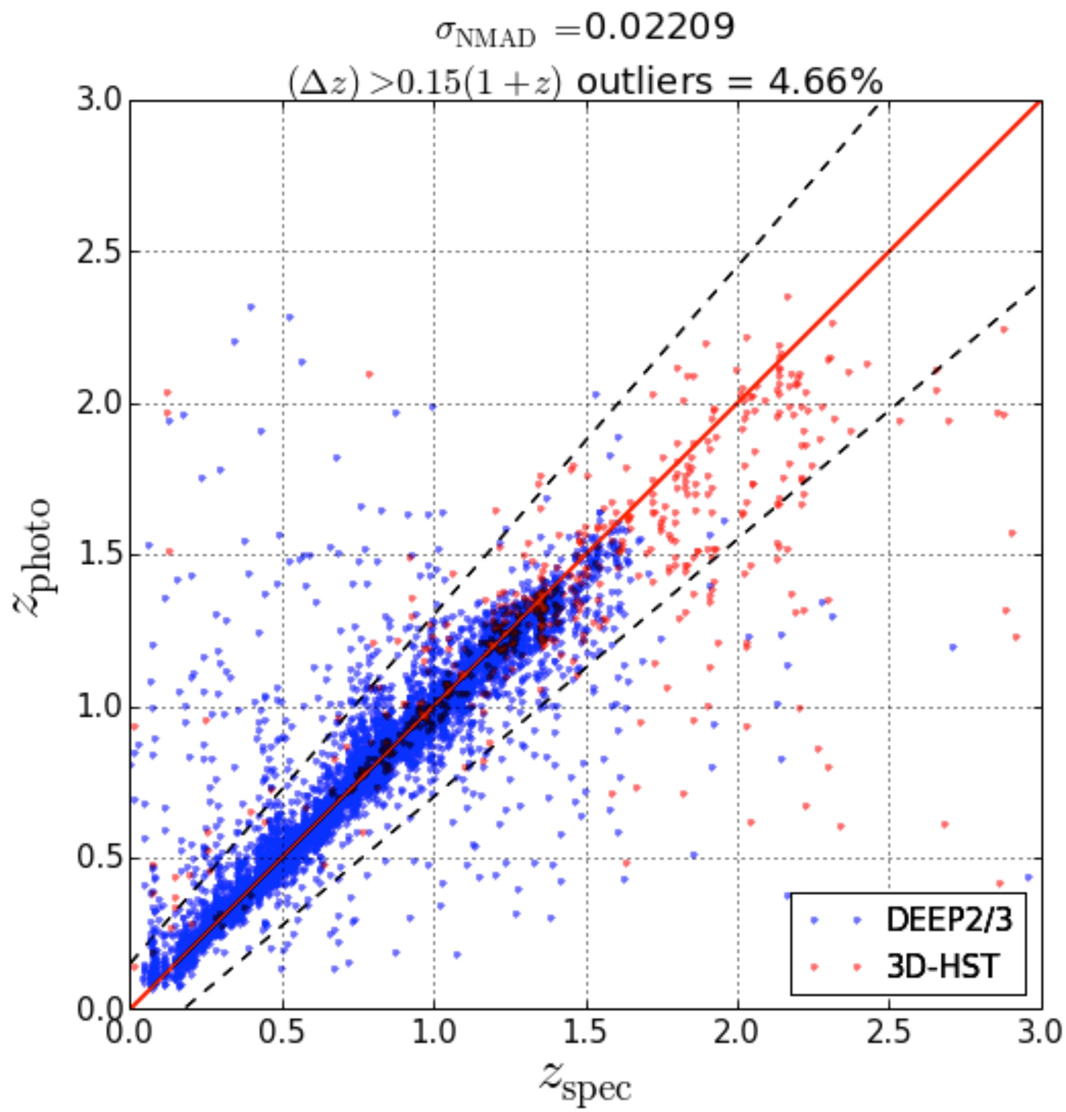


cells with <1% failure rate z's　　with <5% failure rate z's

Masters et al. 2015

An additional issue: some photo-z/spec-*z* outliers are physical

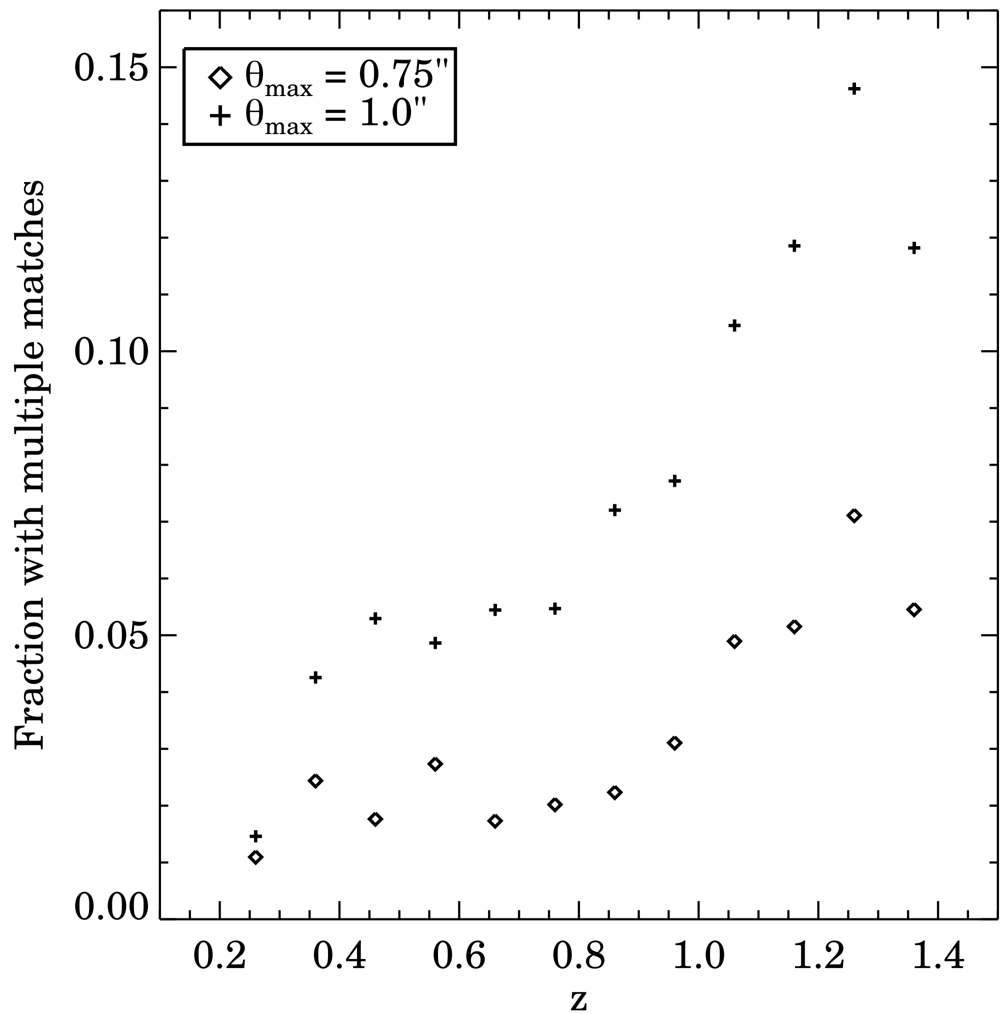
- **A few percent of DEEP2 spectroscopic targets correspond to multiple galaxies when you look at HST catalogs**
- **1% of DEEP2 objects show spectral features from multiple redshifts**
- **Can identify many but NOT all of these blends with space-based imaging**



Zhou, Cooper, JN et al. 2019, in prep.

An additional issue: some photo-z/spec-*z* outliers are physical

- A few percent of DEEP2 spectroscopic targets correspond to multiple galaxies when you look at HST catalogs
- 1% of DEEP2 objects show spectral features from multiple redshifts
- Can identify many but NOT all of these blends with space-based imaging



Newman et al. 2013

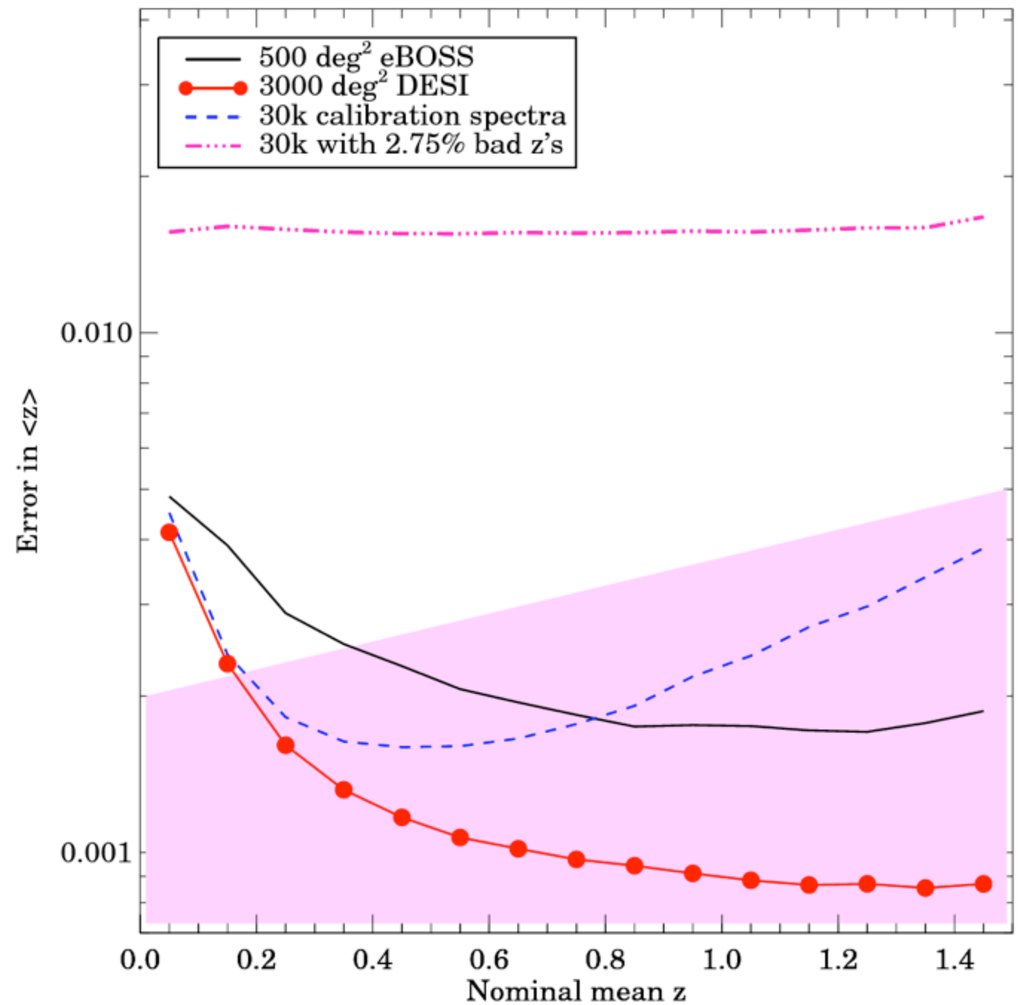
If spectroscopy proves incomplete, calibration will probably need to come from cross-correlation methods...

- 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, >500 degrees of overlap with DESI-like survey would meet LSST science requirements (>4000 sq deg of overlap expected)

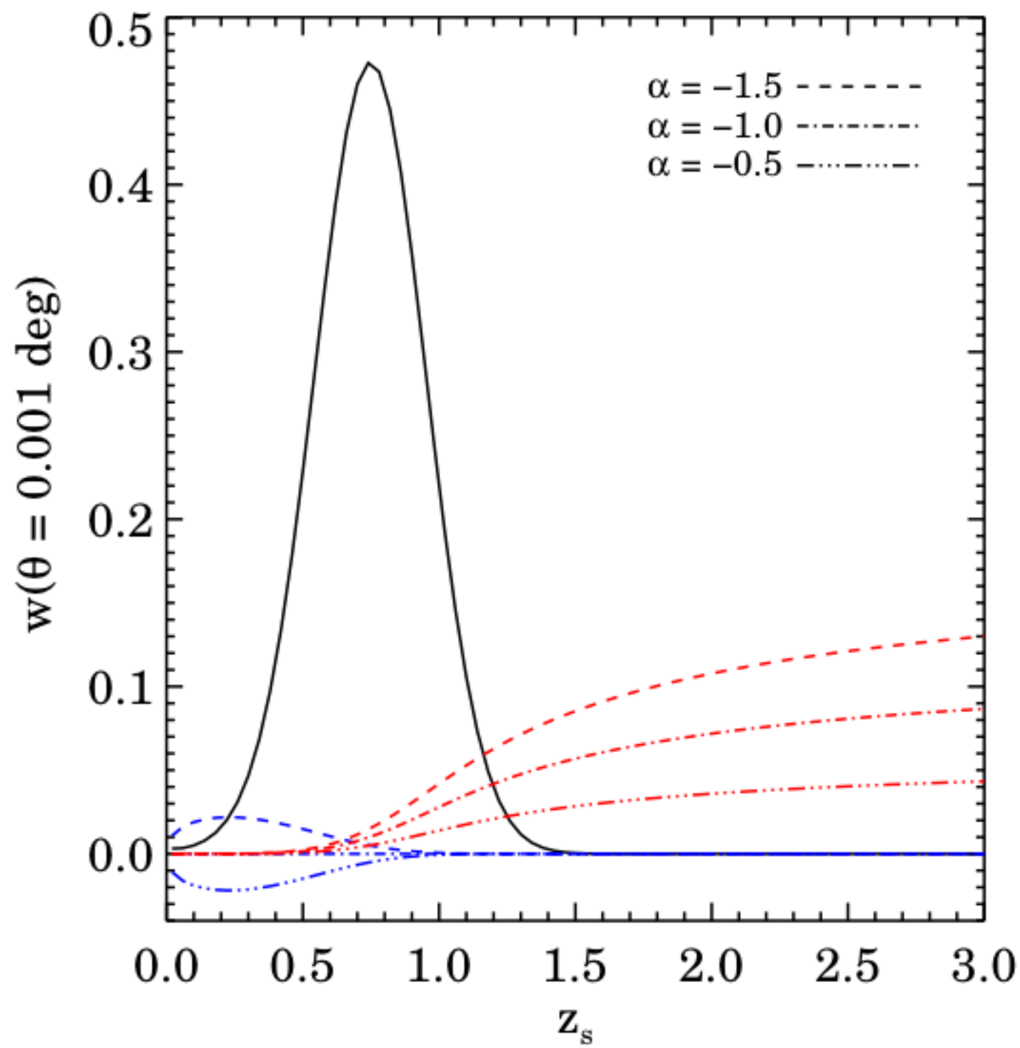
    - ... IF LSST data is uniform (after calibration), as DESI is in North



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

Biggest concern: disentangling cross-correlations from clustering and lensing magnification

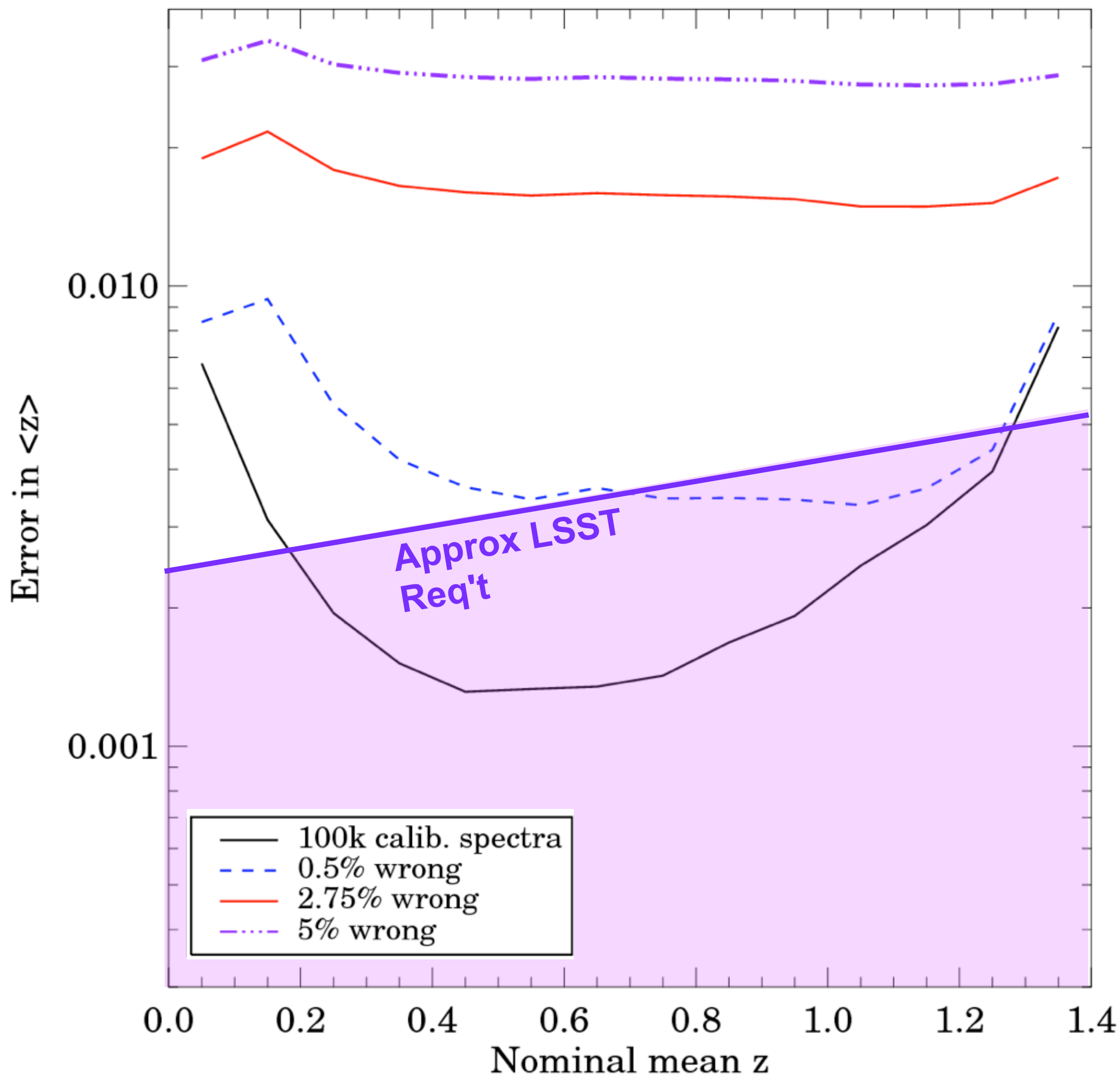
- ***Black***: cross-correlations between photo-z objects (z=0.75 Gaussian) and spectroscopic sample as a function of *z*
- ***Blue***: observed cross-correlation due to spectroscopic objects lensing photometric ones
- ***Red***: observed cross-correlation due to photometric objects lensing spectroscopic ones
- Weak/CMB lensing could help us predict the red curves



Daniel Matthews Ph. D. thesis, 2014

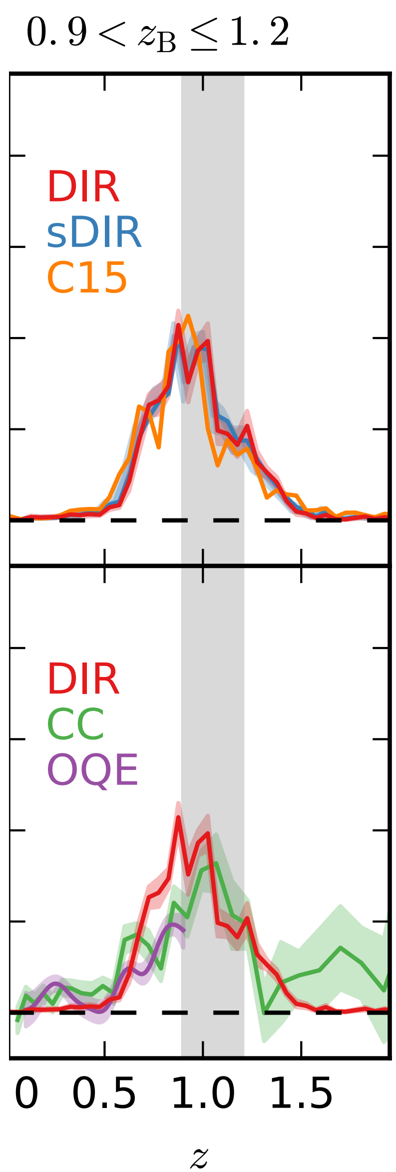
Note: even for 100% complete samples, current false-z rates would be a problem

- 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

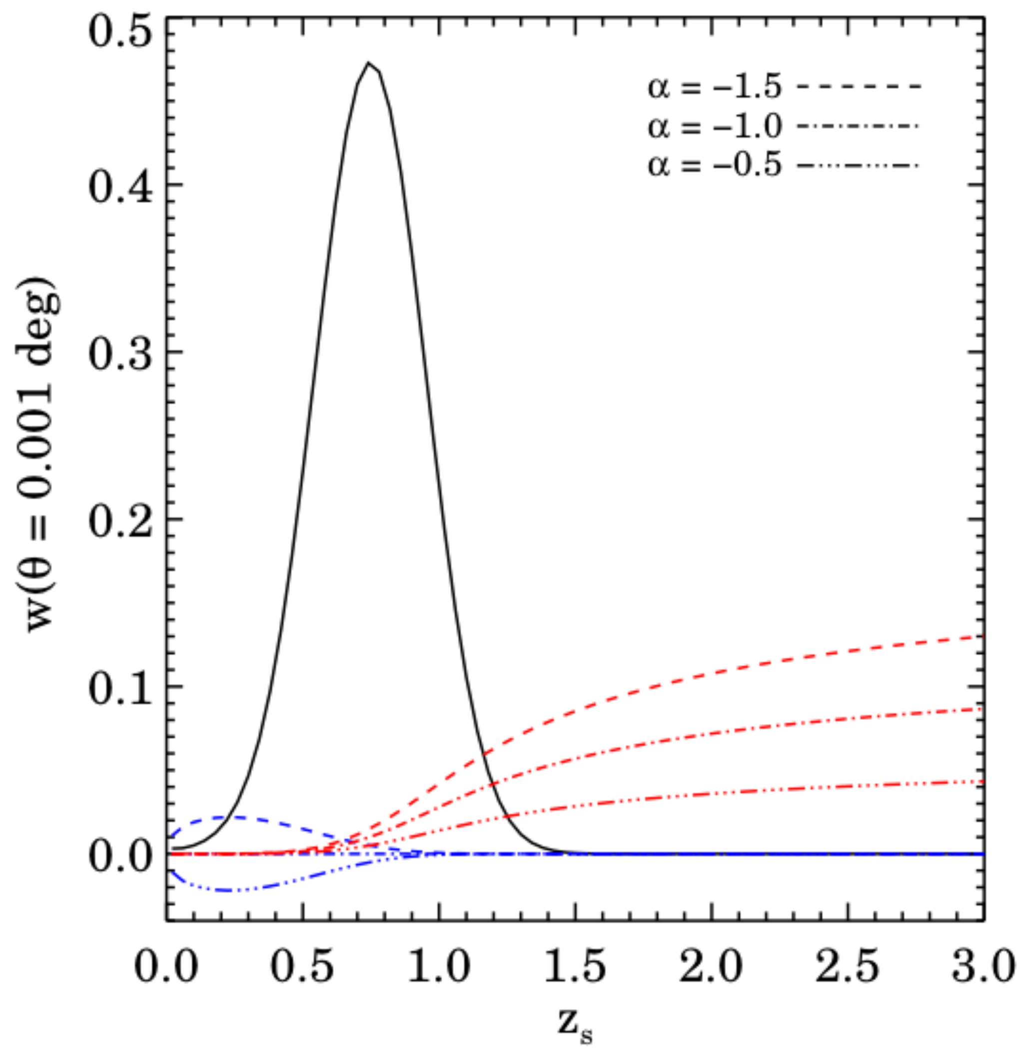


Based on simulated redshift distributions for ANNz-defined DES bins in mock catalog from Huan Lin, UCL & U Chicago, provided by Jim Annis

Biggest concern: disentangling cross-correlations from clustering and lensing magnification



Hildebrandt et al. 2018

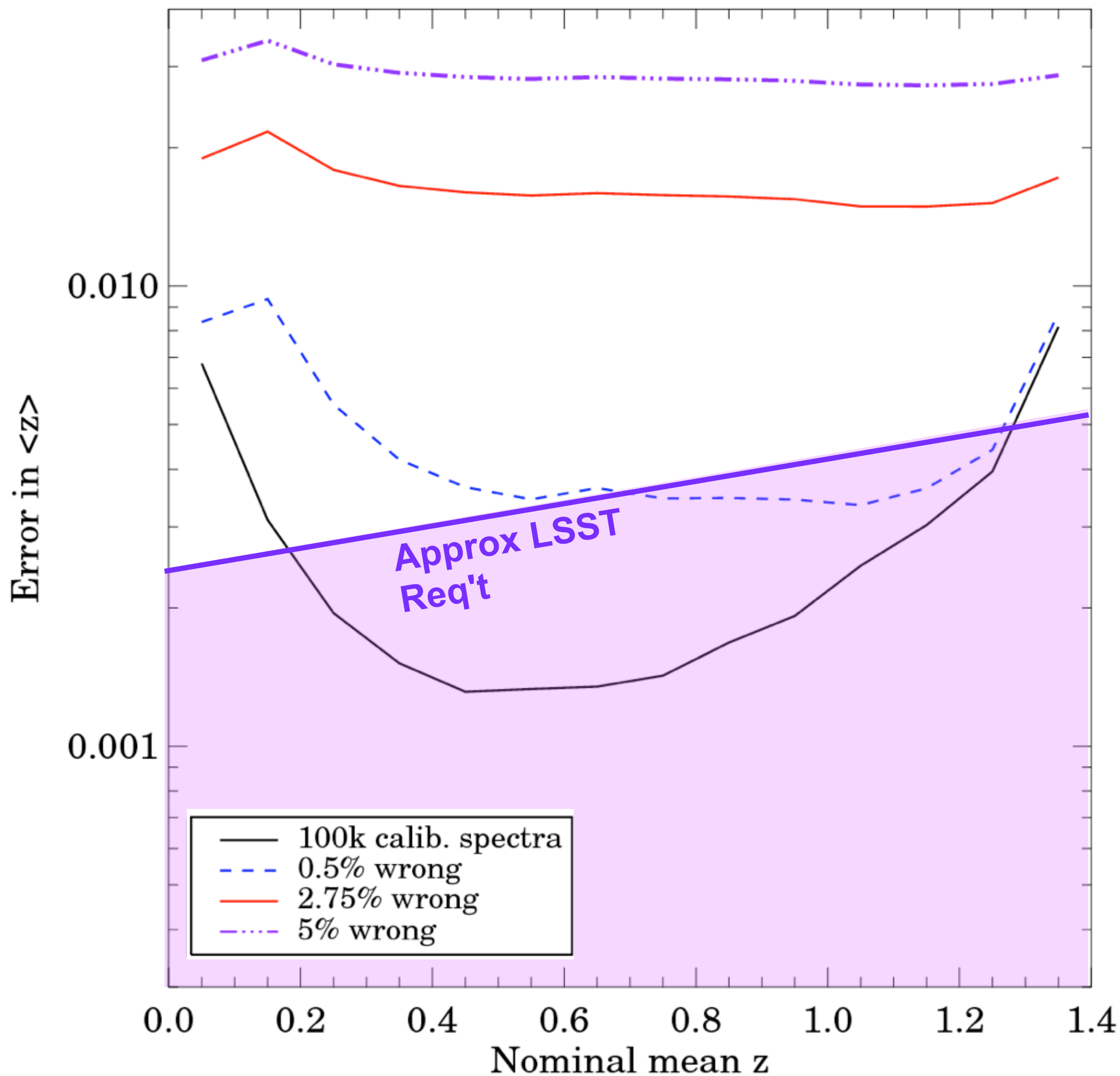


Daniel Matthews Ph. D. thesis, 2014

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

What might an ideal photo-z algorithm look like?

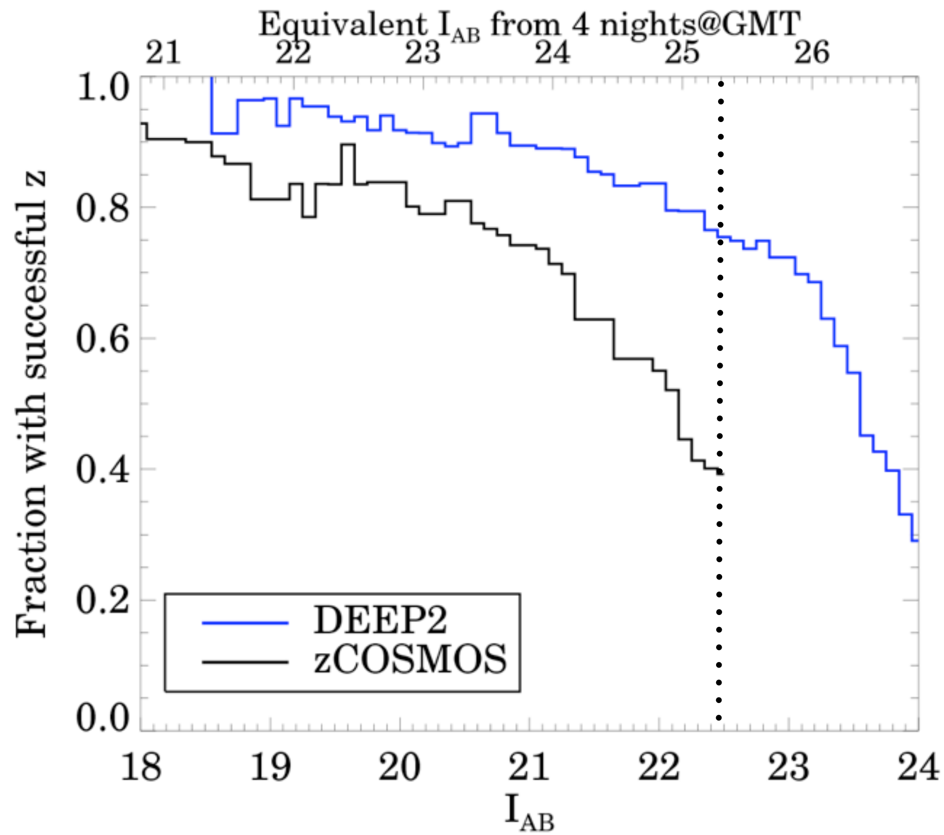
- What might an ideal LSST photo-z algorithm for the next decade look like?
  - Trained with >30,000 spectra spanning range of photometric objects
  - Develops priors & tweaks templates via hierarchical Bayesian hyperparameters
  - Incorporates variations in effective filter wavelengths due to observational 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 template to ML techniques to identify potential outliers (different failure modes)

Conclusions

- Current codes appear sufficient to meet LSST requirements, but are not optimal. Better photo-z's would increase the value of LSST.
- Don't assume that photo-z algorithms will give you PDFs that meet the statistical definition
- Don't assume that we will get LSST/WFIRST depth photo-z training sets without broad community support to make that happen
- Don't assume that those training samples will definitely be complete enough to use for calibration
- Don't assume that all your spectroscopic redshifts will be correct
  - Showing false-z rates are low enough for calibration is expensive... can't use the same redshifts to select good regions of color space and to demonstrate that failure rates are small
- Don't assume that you can ignore magnification signal in cross-correlation photo-z calibration (remove iteratively?)

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

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

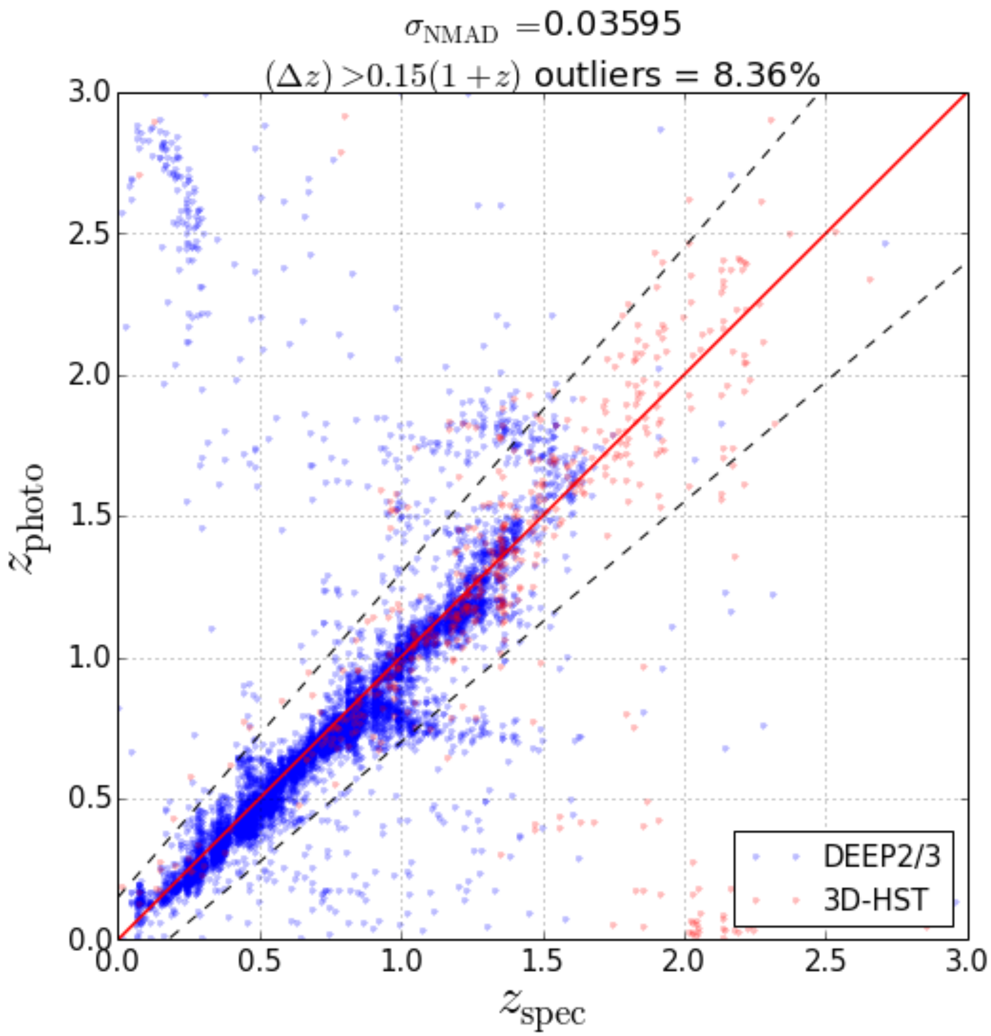
Instrument / Telescope	Total time (years), >75% complete LSST sample	Total time (years), >90% complete LSST sample
4MOST	7.7	48.4
Mayall 4m / DESI	5.1	31.9
WHT / WEAVE	9.0	56.0
Magellan LASSI	1.8	11.2
Subaru/PFS	1.1	6.9
VLT/MOONS	4.0	25.0
Keck/Deimos	10.2	63.9
Keck/FOBOS	4.4	27.5
ESO SpecTel	0.66	4.1
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*

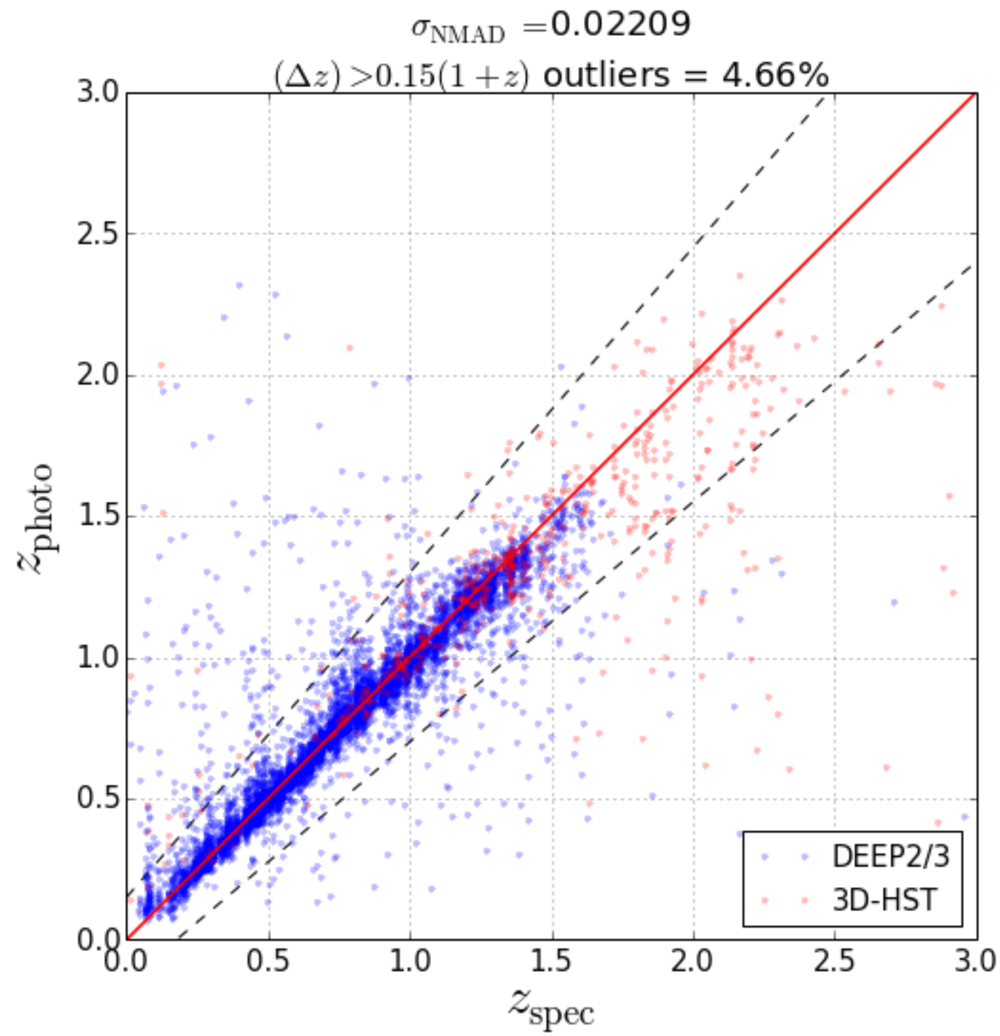
Open issues: template-based and training-based methods have different failure modes - how best to combine?

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

EAZY (template code, untuned)



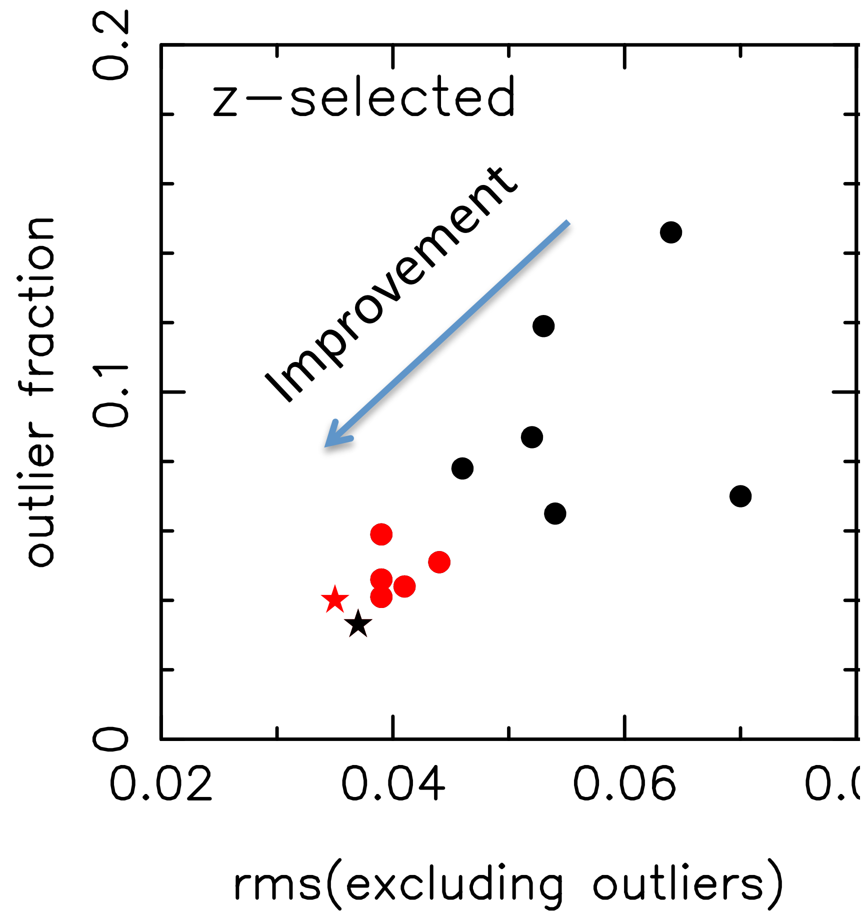
Random Forest Regression



Zhou, JN et al. 2016, in prep.

Open issues: Combining PDF results from multiple codes

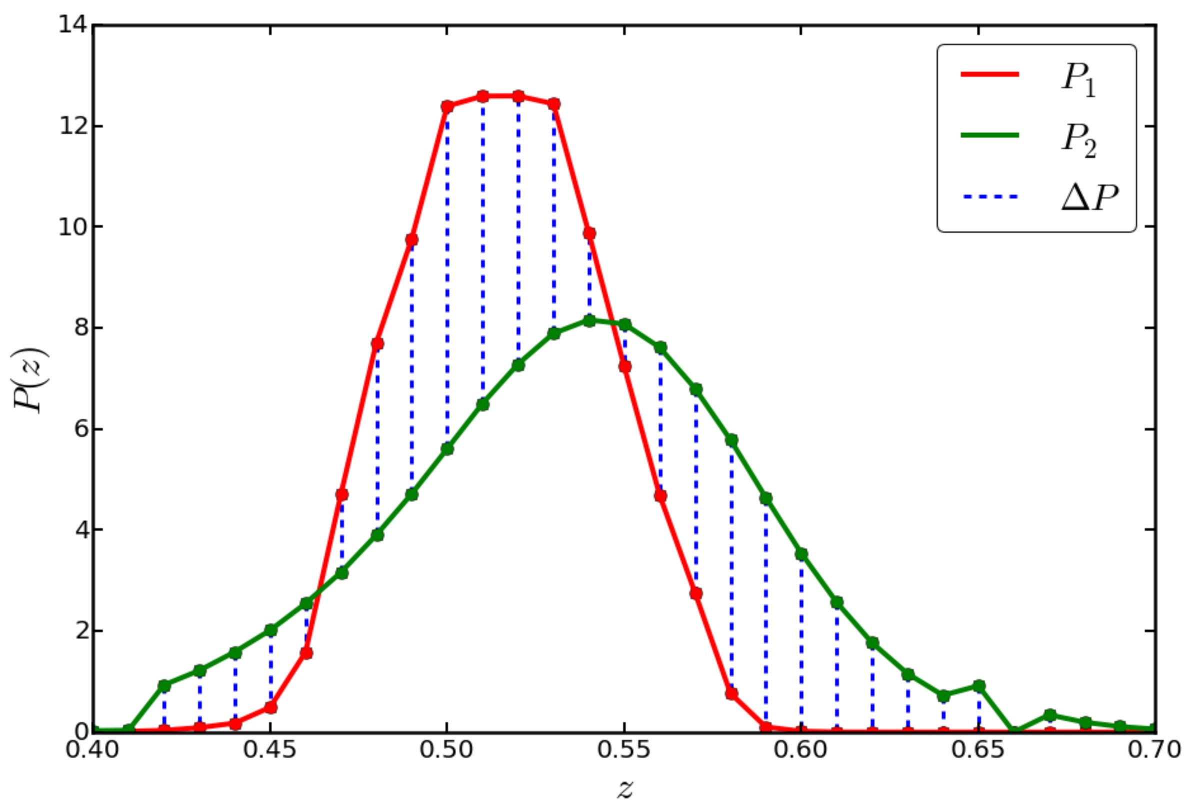
- **Dahlen et al. found that medians of point estimates from multiple codes (★'s) have smaller scatter (relative to spec-z) than any individual code**
- **All codes are run on the same data! Current codes do not make optimal use of available information...**



Dahlen et al. 2013

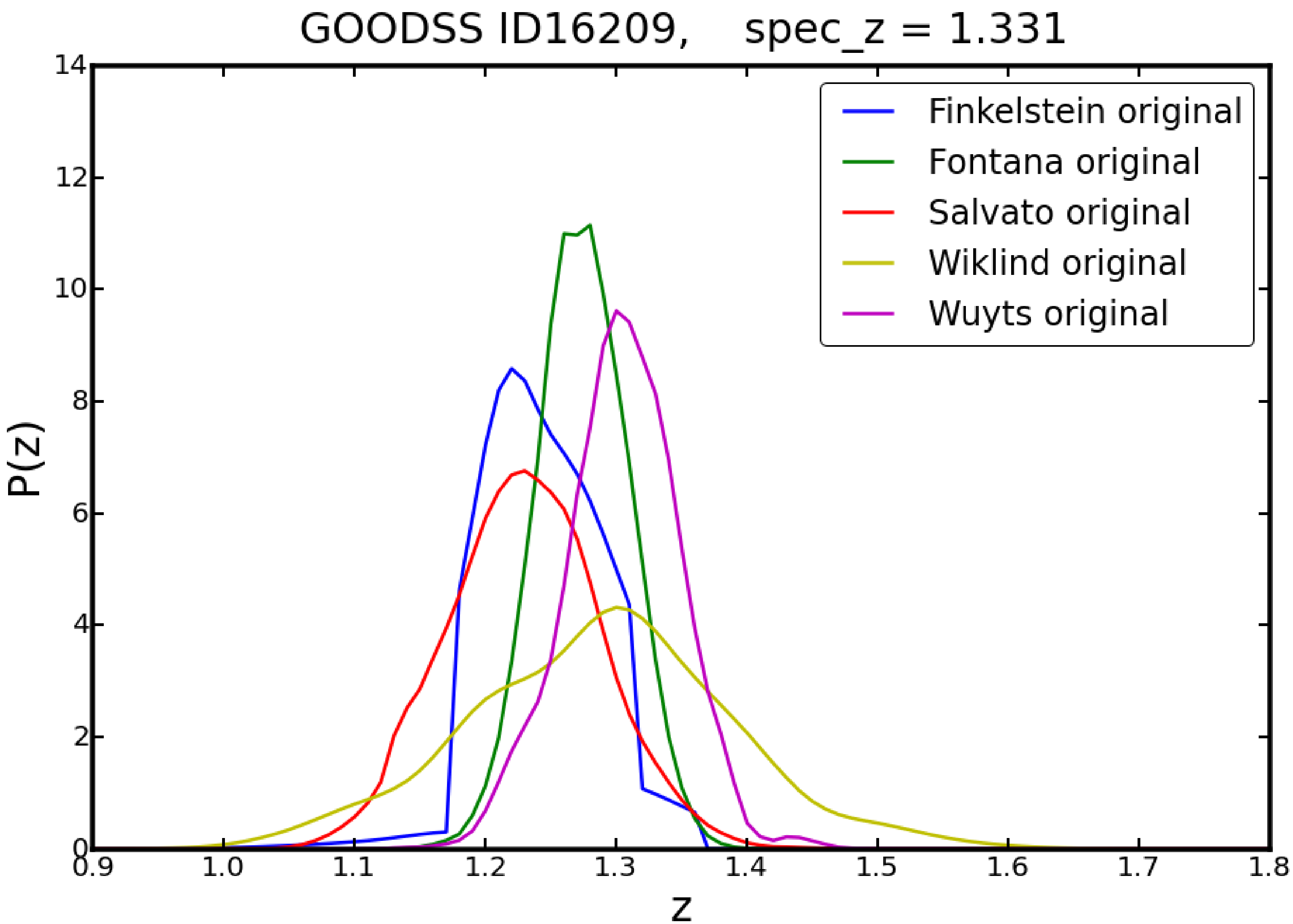
Open issues: Combining PDF results from multiple codes

- 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



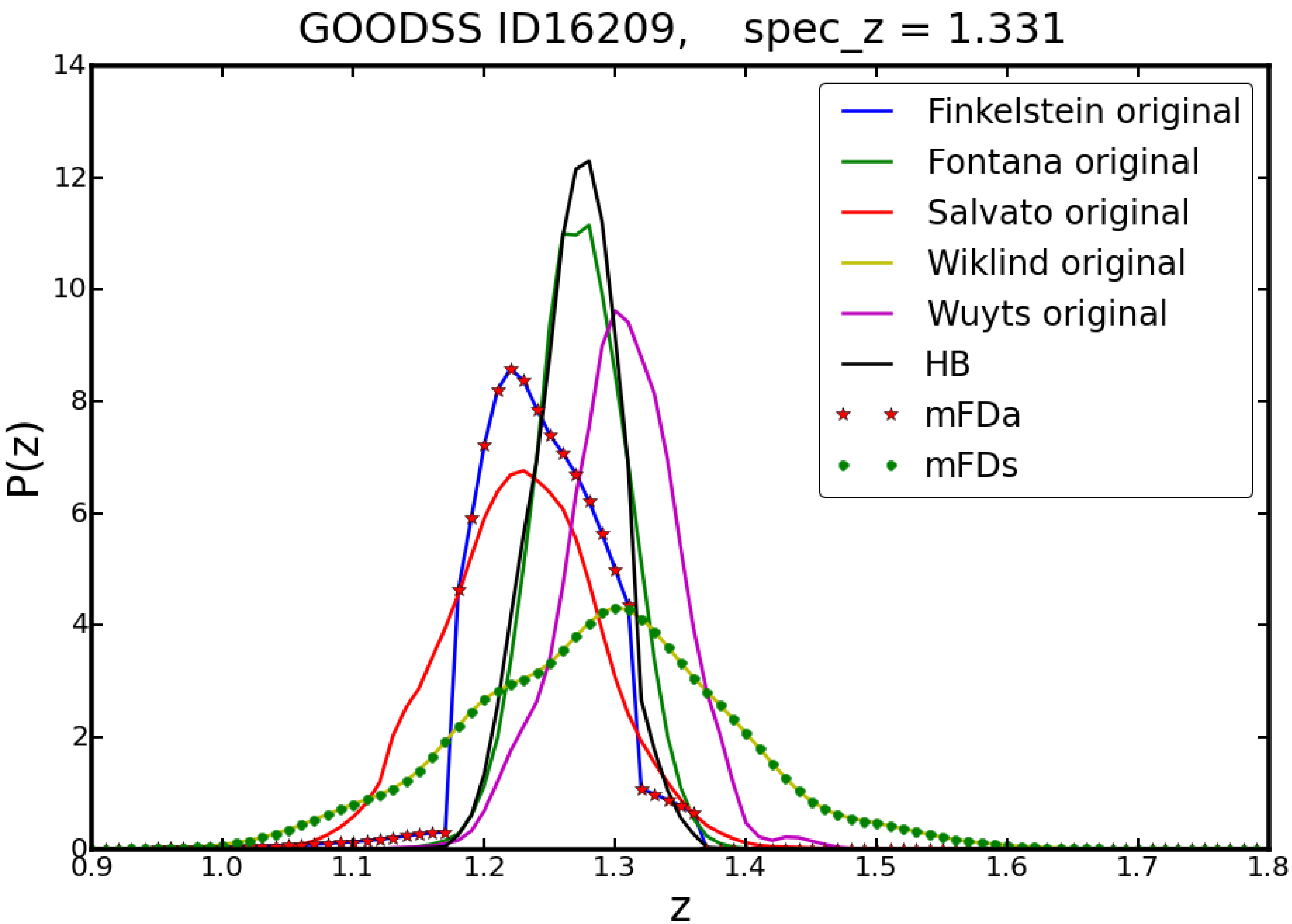
D. Kodra

Open issues: Combining PDF results from multiple codes



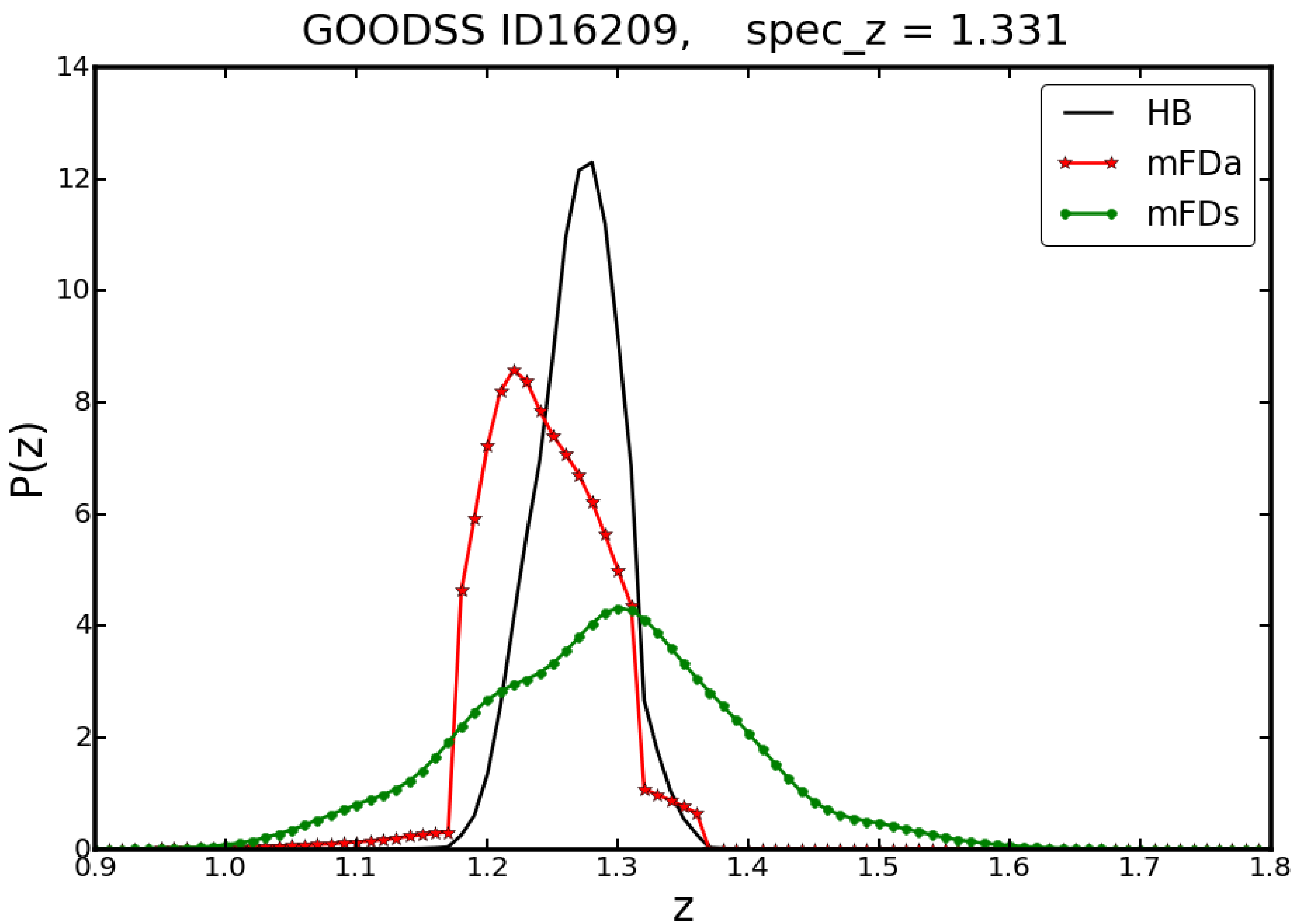
D. Kodra

Open issues: Combining PDF results from multiple codes



D. Kodra

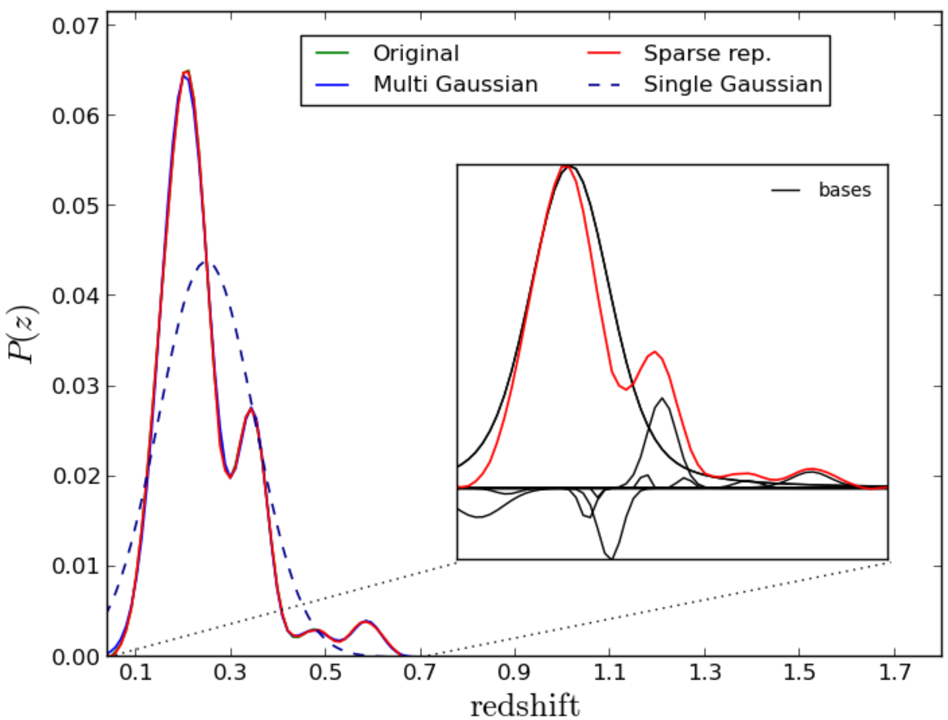
Open issues: Combining PDF results from multiple codes

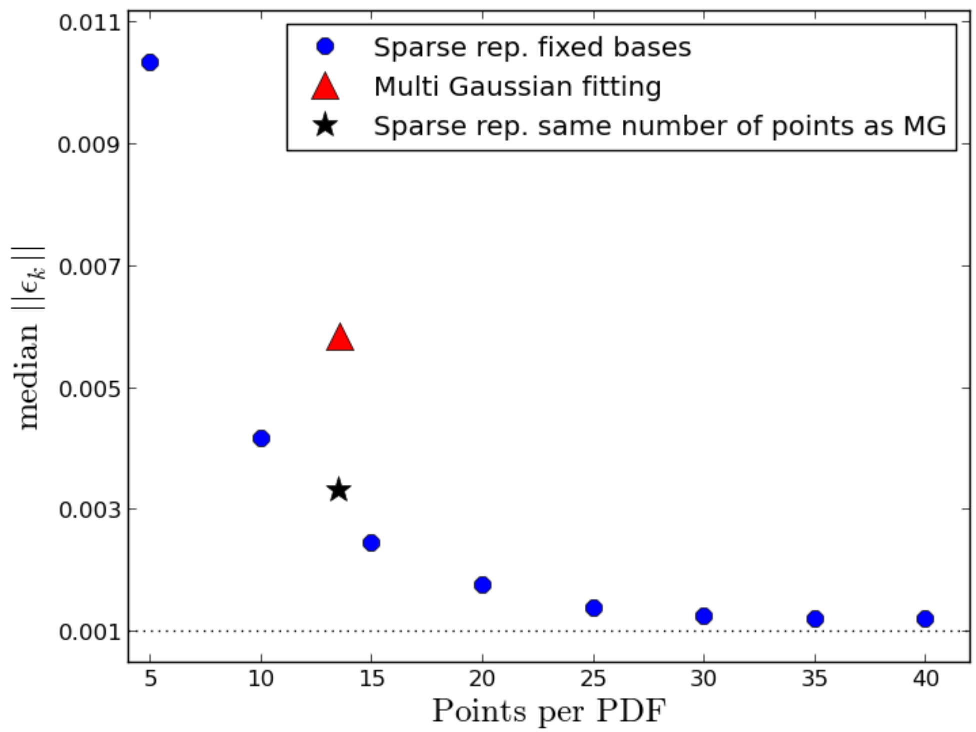


D. Kodra

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?

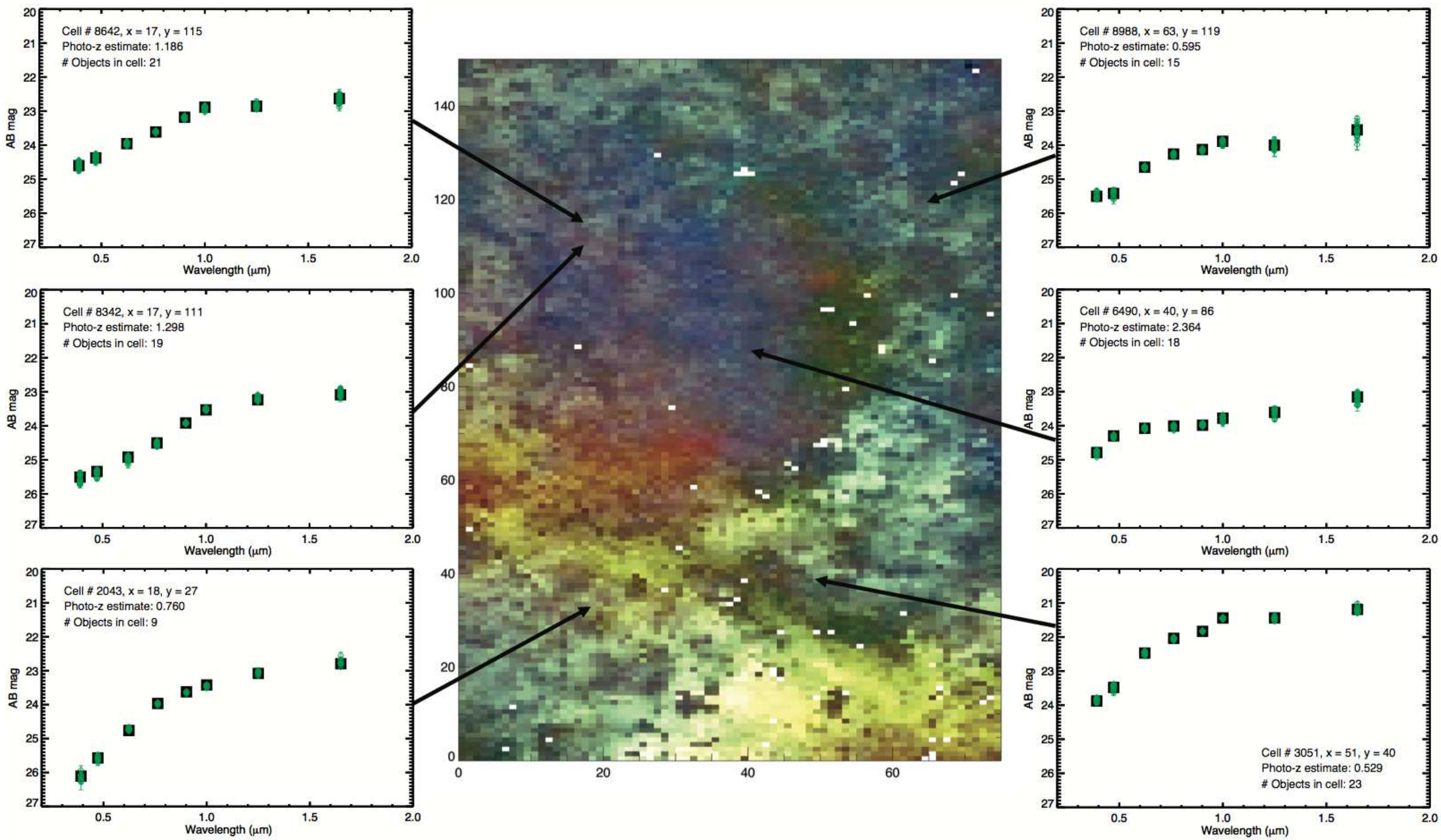




Carrasco-Kind & Brunner 2014

Open issues: Optimizing spectroscopic targeting

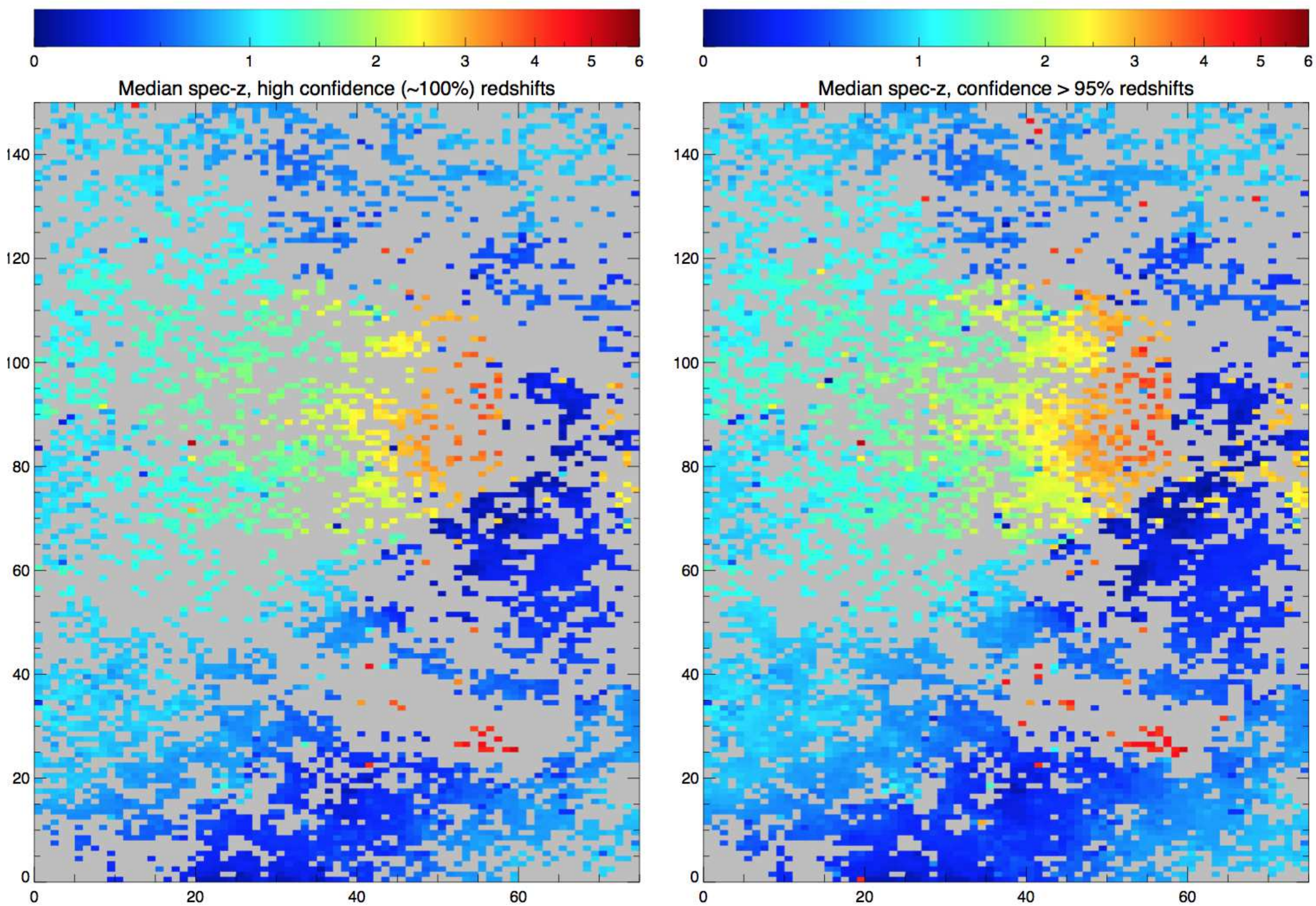
- **Current state of the art: Masters et al. 2015**
- **Self-organized map of galaxy colors**



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

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

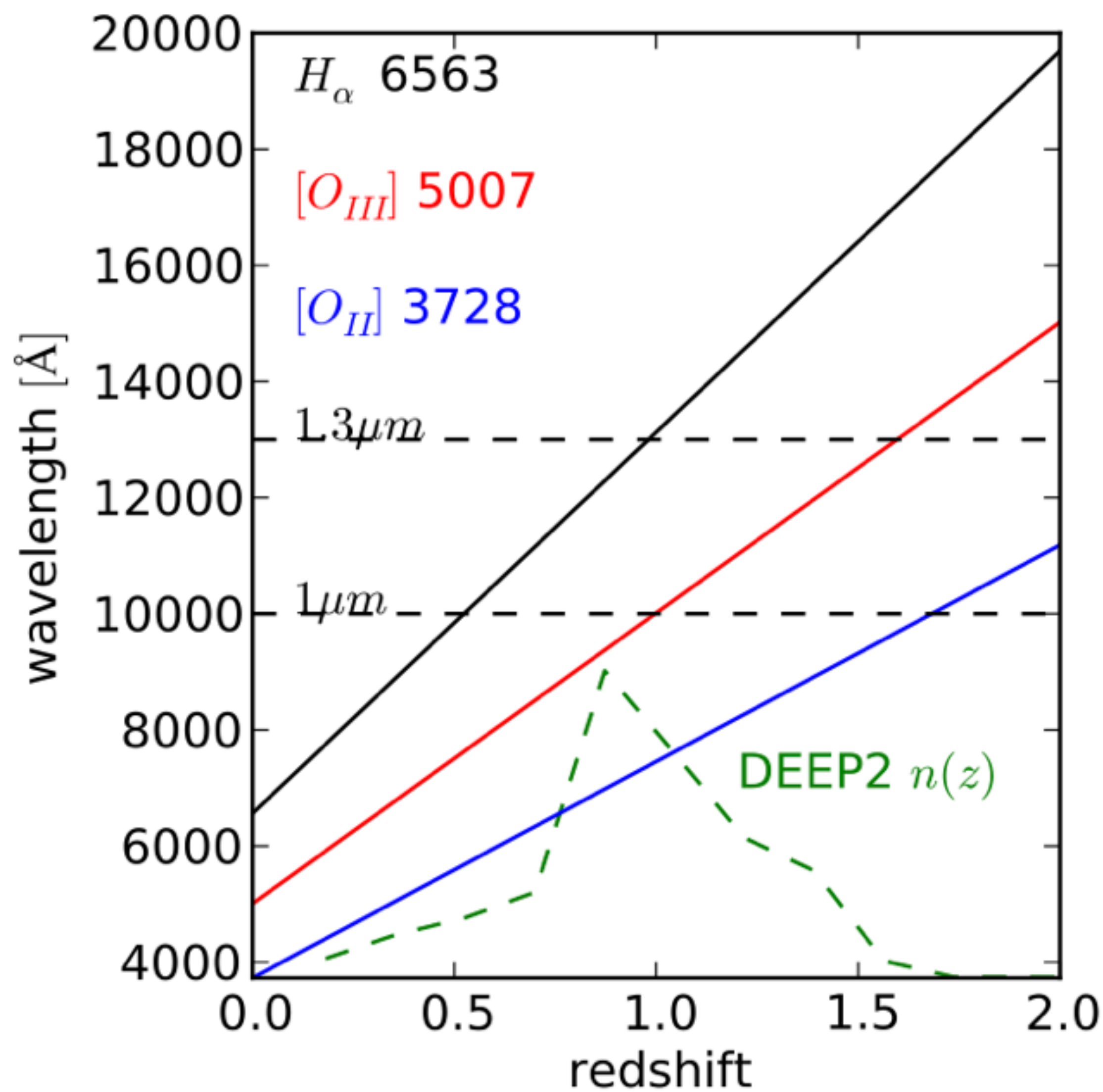
See Newman et al. 2015, *Spectroscopic Needs for Imaging Dark Energy* Experiments, for details

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

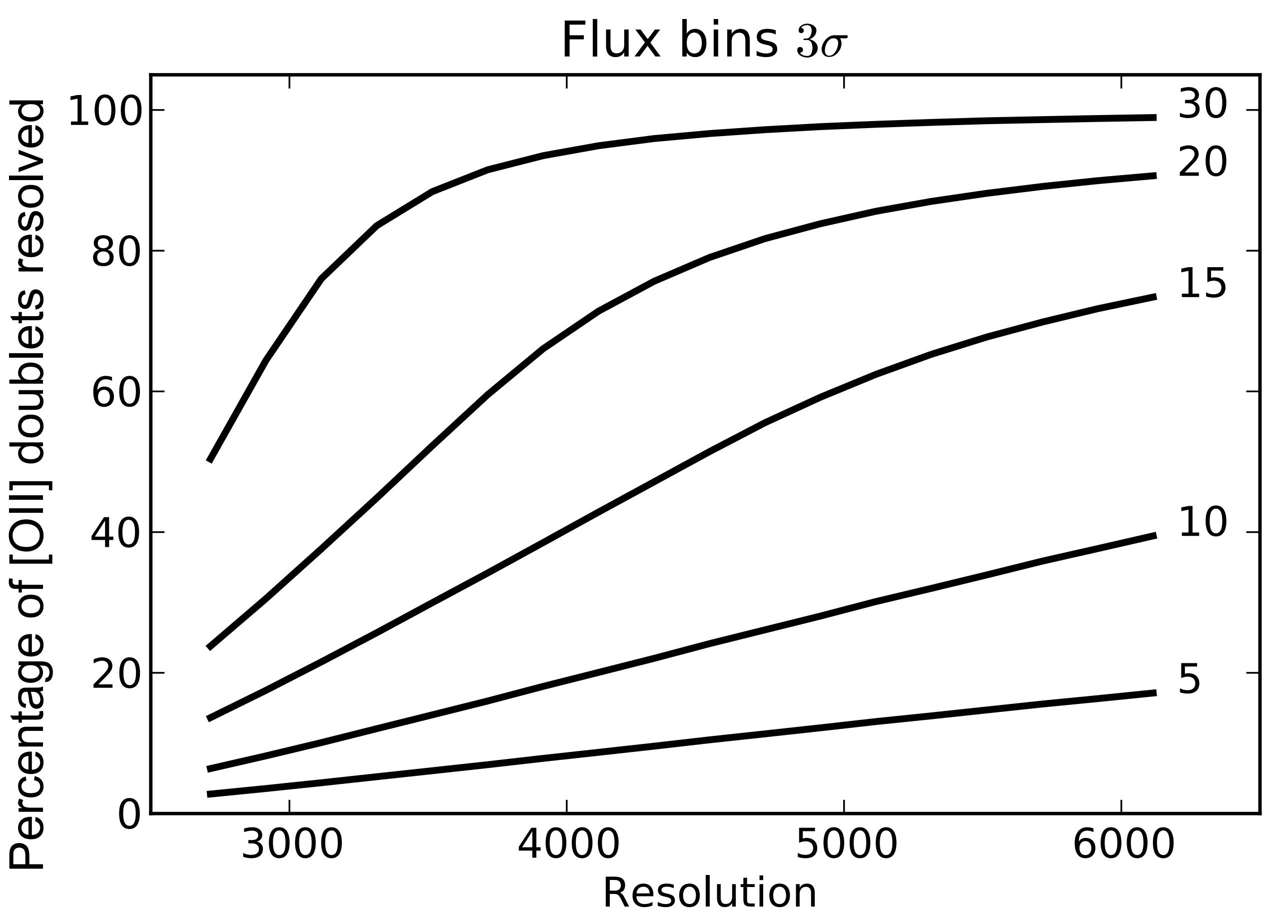
    - Require multiple features for secure redshift



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**

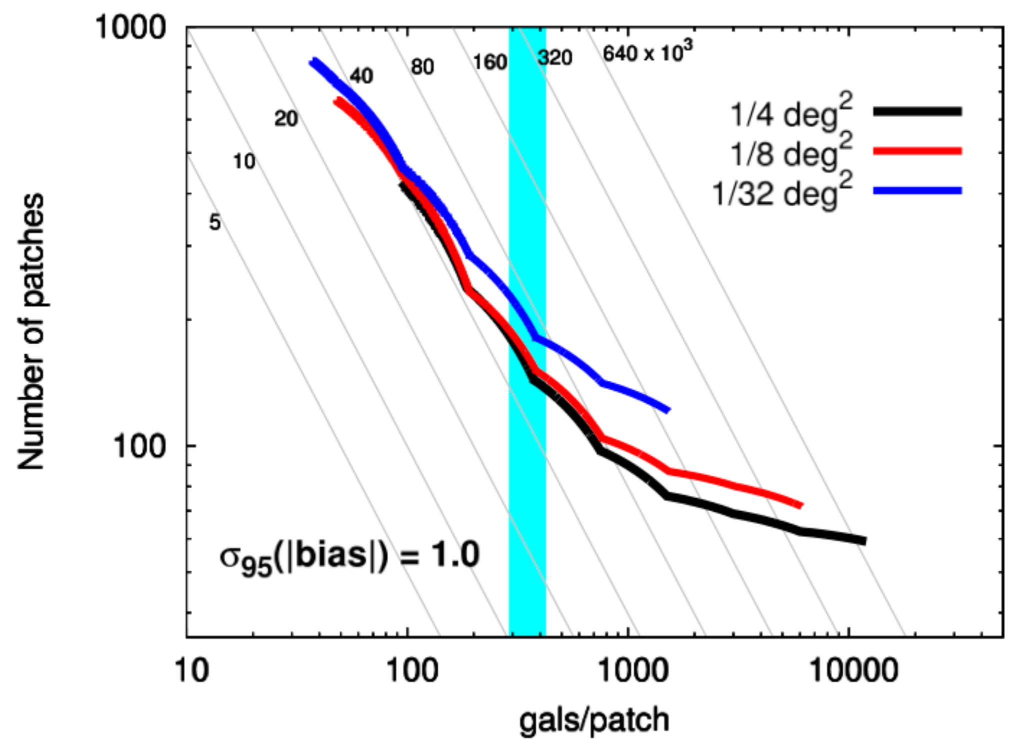


Comparat et al. 2013

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) estimated 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