ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

# Photo-$z$ PDF multi-technique estimation, storage and applications

Matías Carrasco Kind
Robert J. Brunner

Department of Astronomy
University of Illinois

LSST-DESC meeting Dec 4-6, 2013

# Motivation

- Photo-$z$ PDF important in cosmology
- Several methods / codes to compute photo-$z$
- Need for a meta-algorithm that combines multiple techniques
- PDF are good **but** for large datasets, storage and I/O is an issue

# Motivation

- Photo-$z$ PDF important in cosmology
- Several methods / codes to compute photo-$z$
- Need for a meta-algorithm that combines multiple techniques
- PDF are good **but** for large datasets, storage and I/O is an issue

# Motivation

- Photo-$z$ PDF important in cosmology
- Several methods/codes to compute photo-$z$
- Need for a meta-algorithm that combines multiple techniques
- PDF are good **but** for large datasets, storage and I/O is an issue

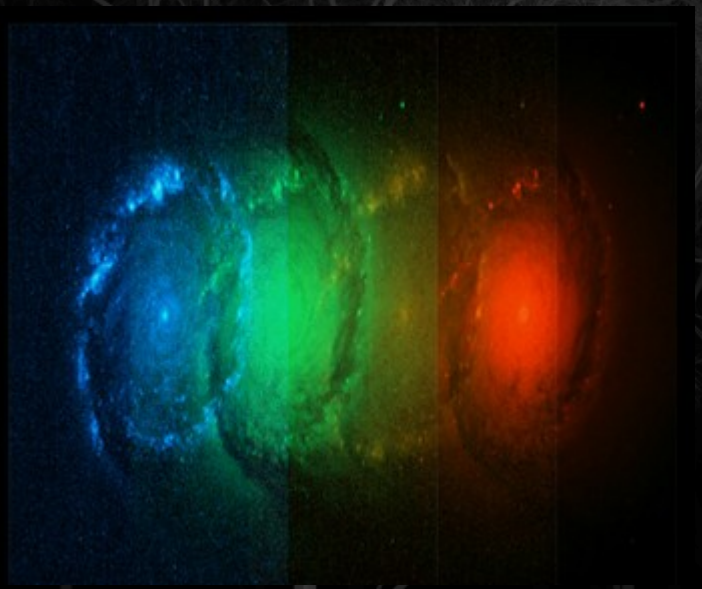
# Motivation

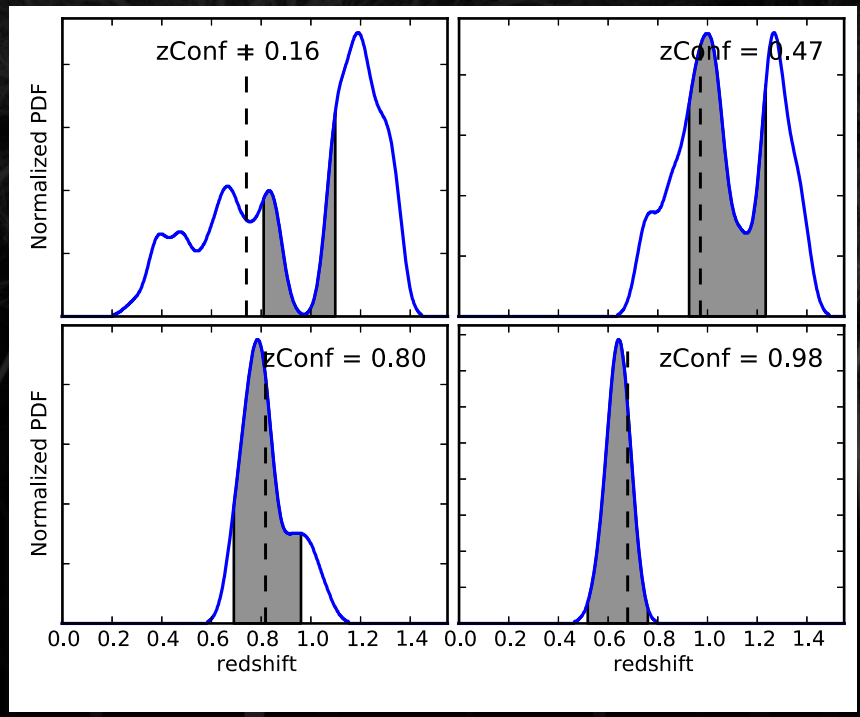
- Photo-$z$ PDF important in cosmology
- Several methods/codes to compute photo-$z$
- Need for a meta-algorithm that combines multiple techniques
- PDF are good **but** for large datasets, storage and I/O is an issue

# Motivation

- Photo-$z$ PDF important in cosmology
- Several methods/codes to compute photo-$z$
- Need for a meta-algorithm that combines multiple techniques
- PDF are good **but** for large datasets, storage and I/O is an issue

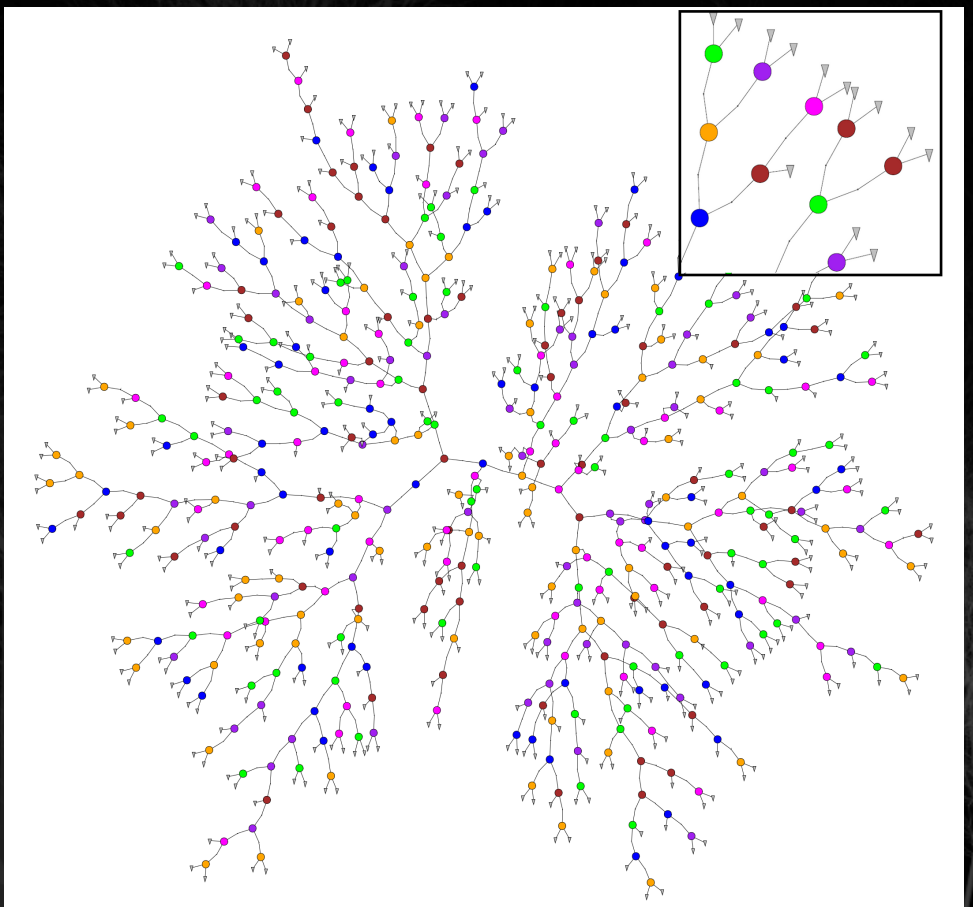
# Photo-$z$ PDF estimation





# Photo-$z$ PDF estimation: TPZ

- TPZ (Trees for Photo-Z) is a supervised machine learning code
- Prediction trees and random forest
- Incorporate measurements errors and deals with missing values
- Ancillary information: expected errors, attribute ranking and others
- Application to the S/G

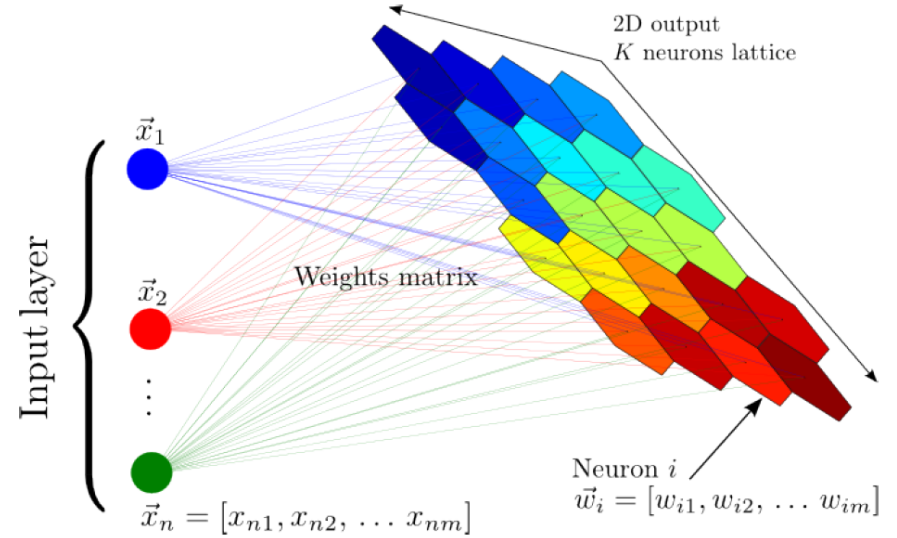


**Carrasco Kind & Brunner 2013a**

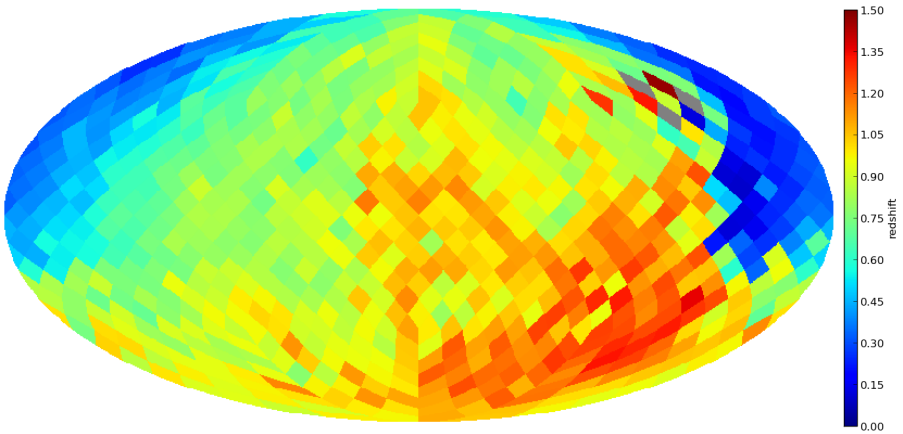
http://lcdm.astro.illinois.edu/research/TPZ.html

# Photo-$z$ PDF estimation: SOM

- SOM (Self Organized Map) is a unsupervised machine learning algorithm
- Competitive learning to represent data conserving topology
- 2D maps and *Random Atlas*
- Framework inherited from TPZ
- Application to the S/G



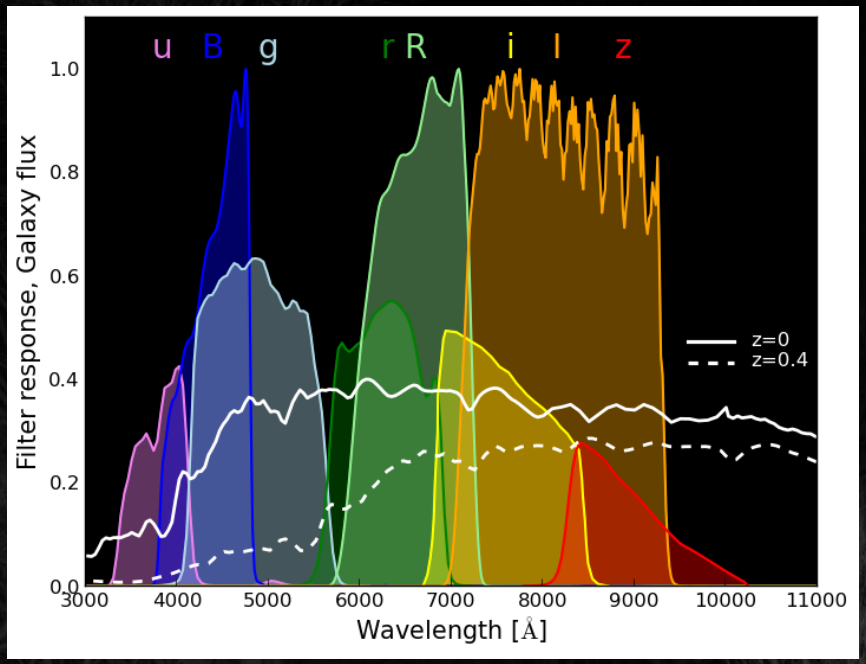

**Carrasco Kind & Brunner, MNRAS submitted**

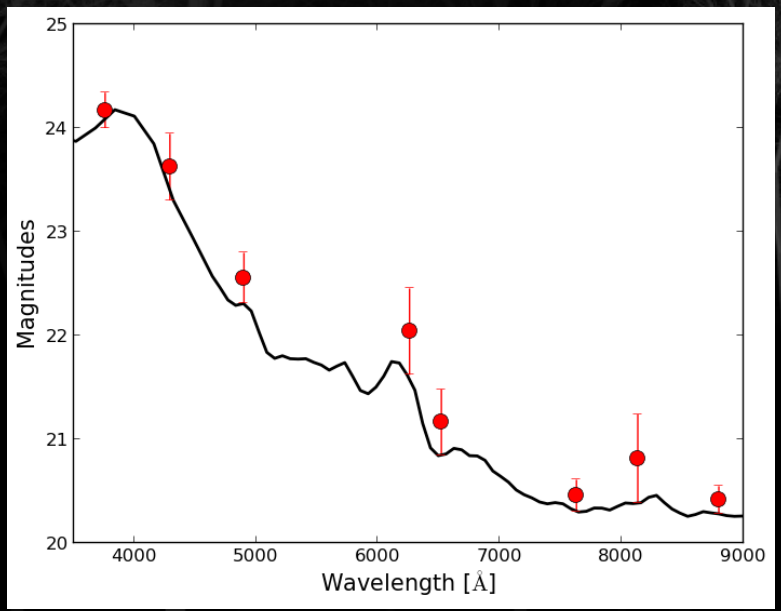



**Carrasco Kind & Brunner, MNRAS submitted**

# Photo-$z$ PDF estimation: BPZ

- BPZ (Benitez, 2000) is a Bayesian template fitting method to obtain PDFs
- Set of calibrated SED and filters
- Doesn't need training data
- Priors can be included

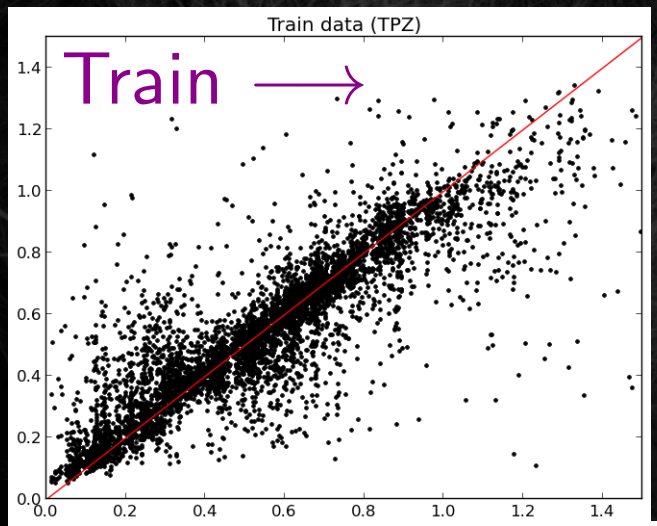


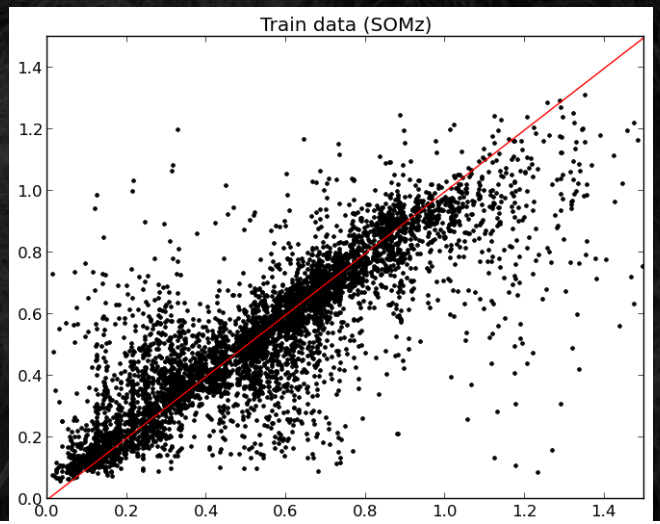



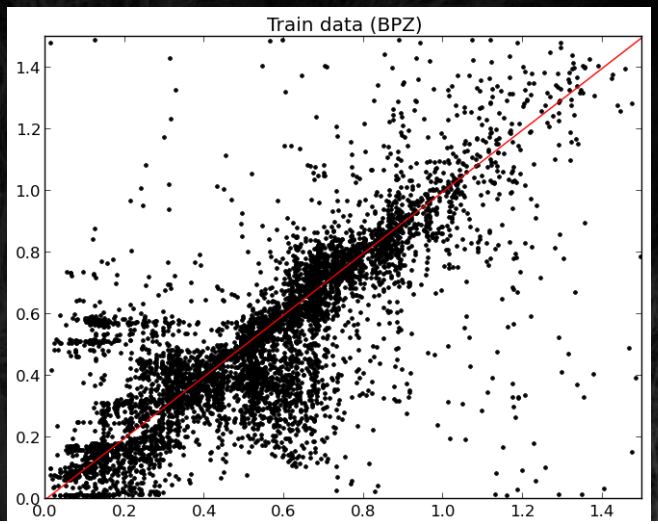
# Photo-$z$ PDF estimation: Error and validation

Out of Bag data used to validate trees/maps

Changes for every tree/map and is not used during training



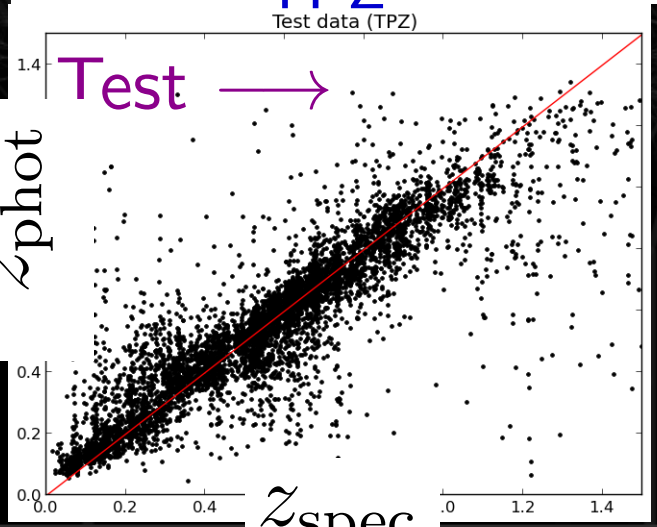



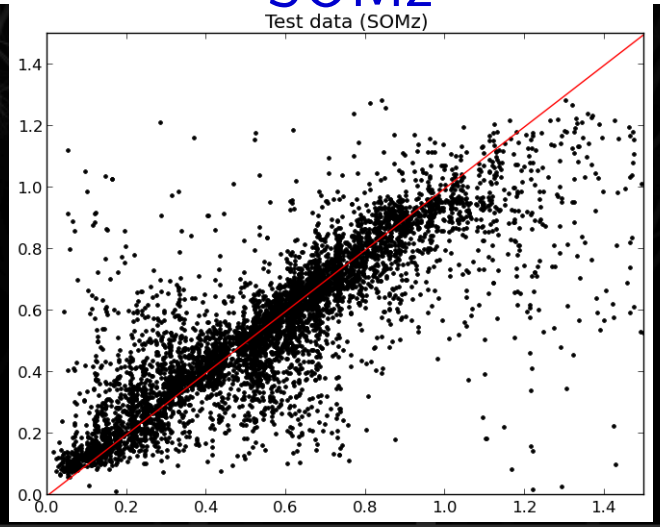



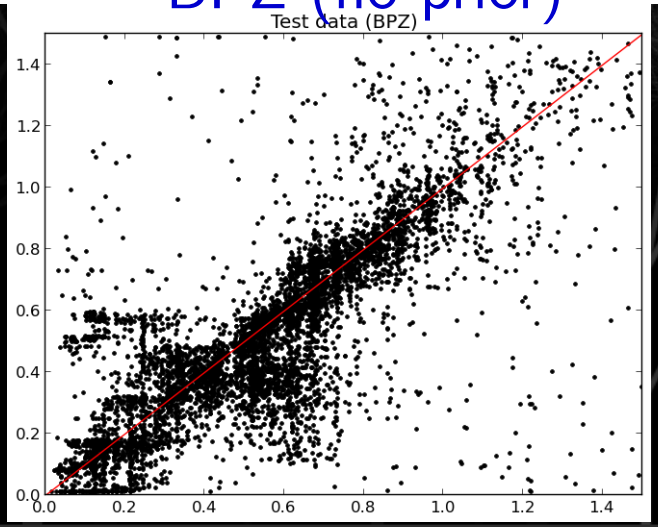

TPZ

SOMz

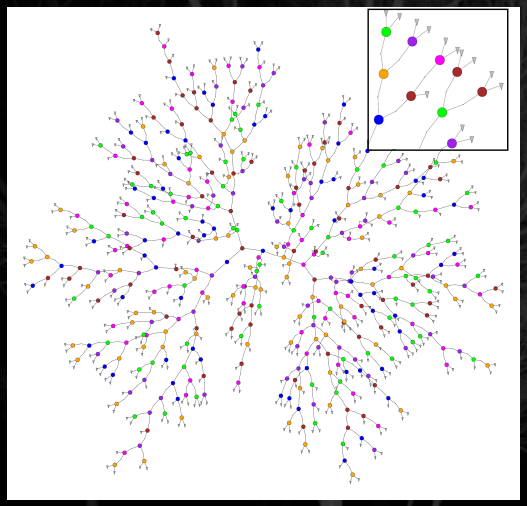
BPZ (no prior)



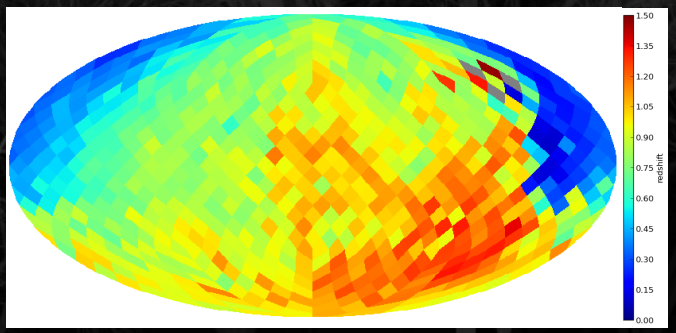



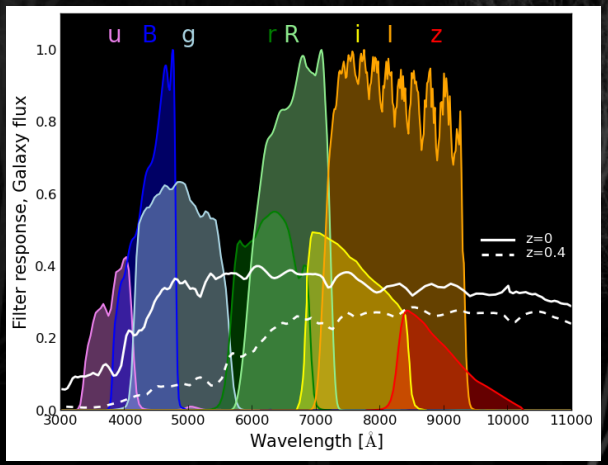
# Photo-$z$ PDF combination



+




+



# Photo-$z$ PDF combination: Bayesian framework

- Use random naïve bayes model to compute individual priors (Carrasco Kind & Brunner, 2013b)
- Currently exploring different models such as: (Carrasco Kind & Brunner, in prep.)
  - Hierarchical Bayes model (Dahlen et al., 2013)
  - Bayesian model averaging
  - MCMC parameter estimation
  - Use machine learning to learn from outliers and errors

# Photo-$z$ PDF combination: Bayesian framework

- Use random naïve bayes model to compute individual priors (Carrasco Kind & Brunner, 2013b)

- Currently exploring different models such as: (Carrasco Kind & Brunner, in prep.)

  - Hierarchical Bayes model (Dahlen et al., 2013)
  - Bayesian model averaging
  - MCMC parameter estimation
  - Use machine learning to learn from outliers and errors

# Photo-$z$ PDF combination: Bayesian framework

- Use random naïve bayes model to compute individual priors (Carrasco Kind & Brunner, 2013b)

- Currently exploring different models such as: (Carrasco Kind & Brunner, in prep.)

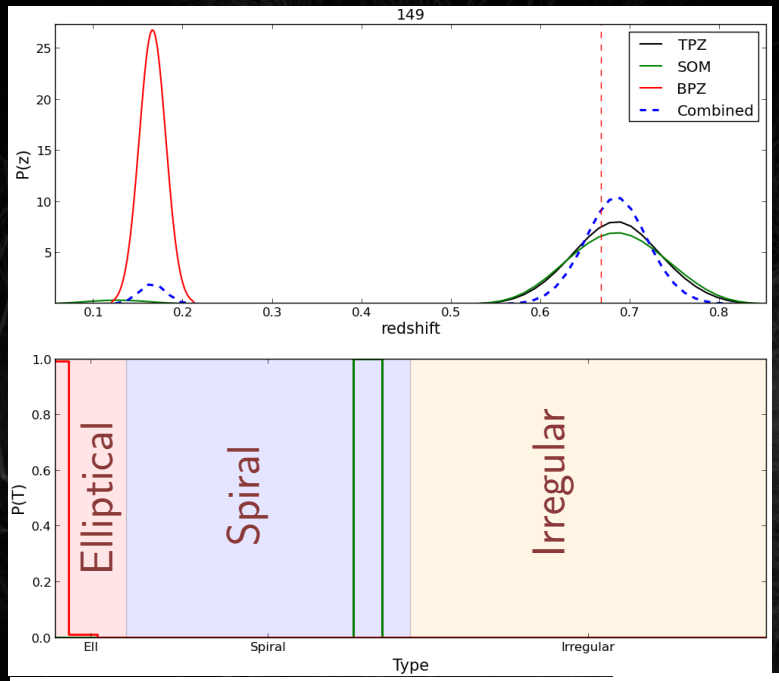
- Hierarchical Bayes model (Dahlen et al., 2013)

- Bayesian model averaging

- MCMC parameter estimation

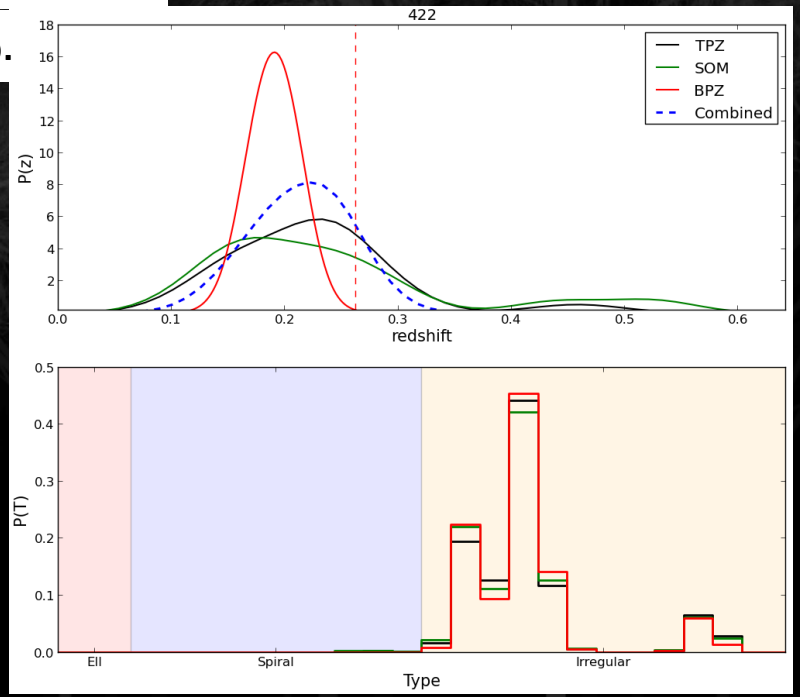
- Use machine learning to learn from outliers and errors

# Photo-$z$ PDF combination: Bayesian framework




**Carrasco Kind & Brunner, in prep.**

Examples




## Our approach

Supervised method

+

Unsupervised method

+

Template fitting

+

Weigthing scheme

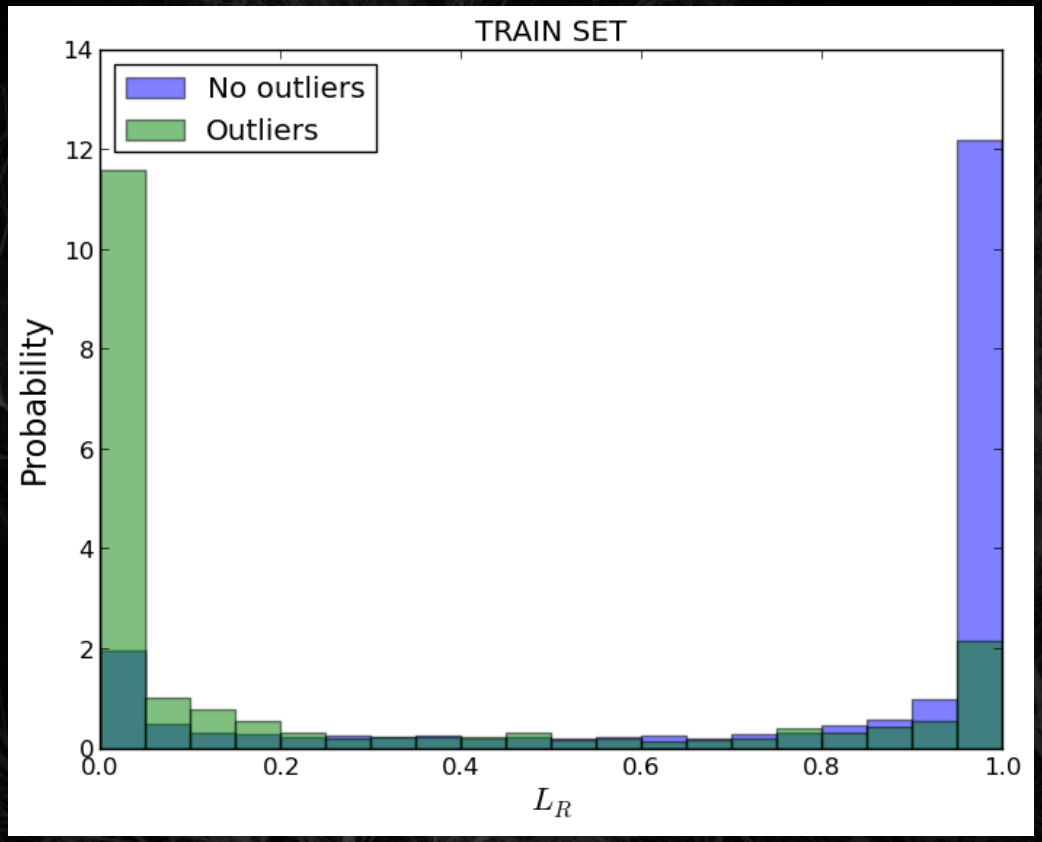
⇓

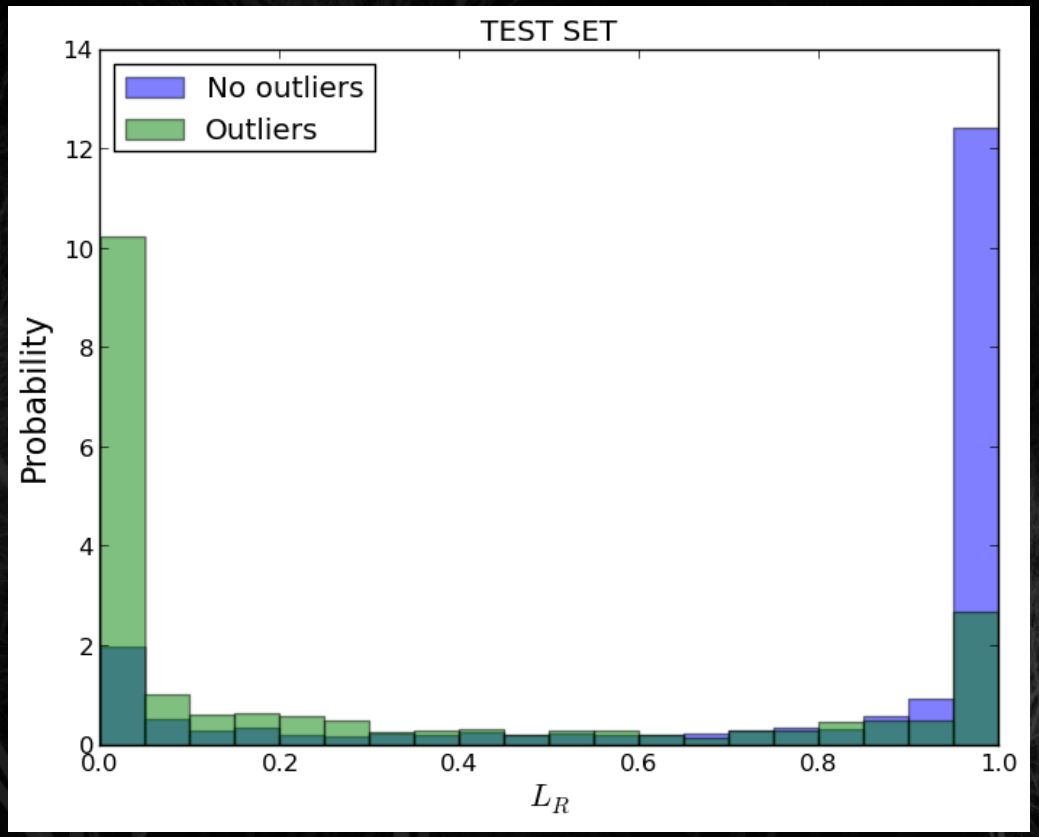
photo-$z$ PDF

+

Morphological type

# Photo-$z$ PDF combination: Outliers



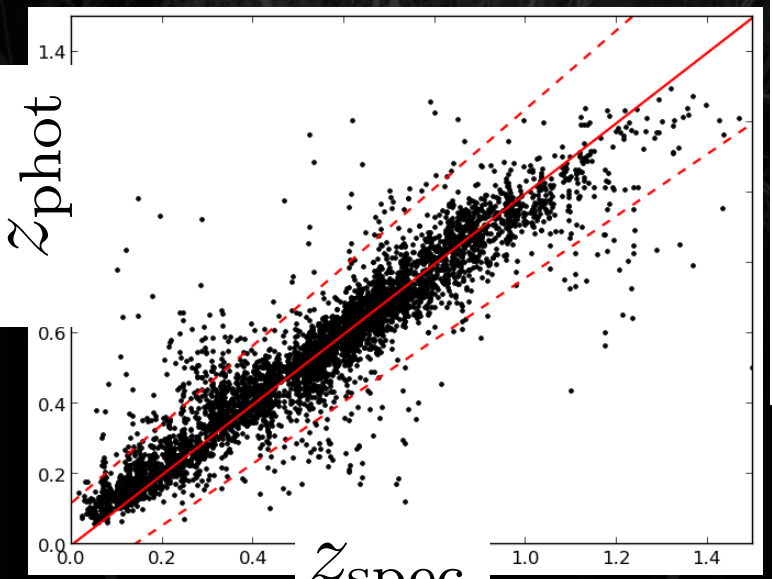



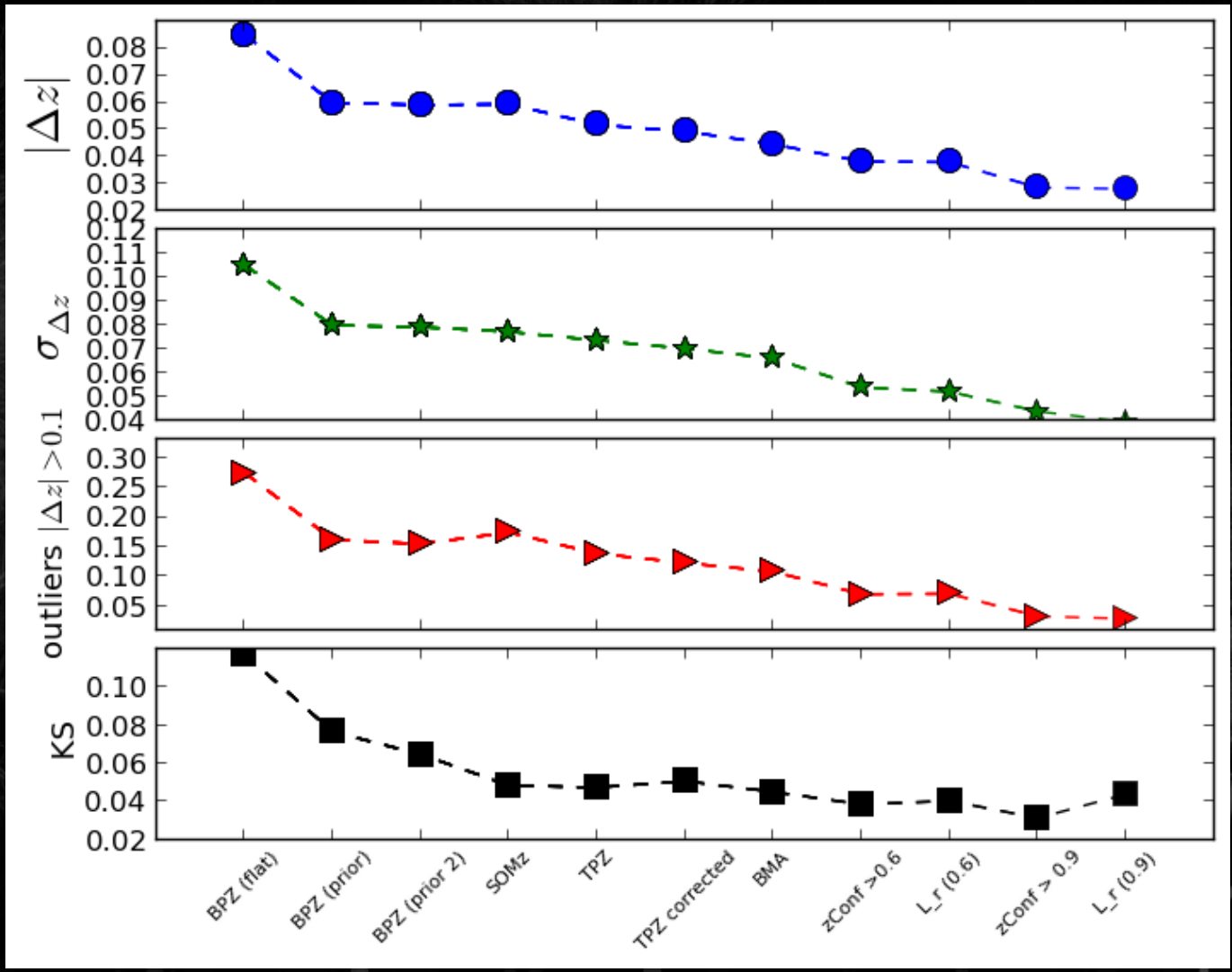

Likelihood ratio for outliers using features from all three techniques similar to Gorecki A., et al. 2013

# Photo-$z$ PDF combination: Results

Averaged metrics for all test galaxies

$$\Delta z = \frac{|z_{\rm spec} - z_{\rm phot}|}{1 + z_{\rm spec}}$$

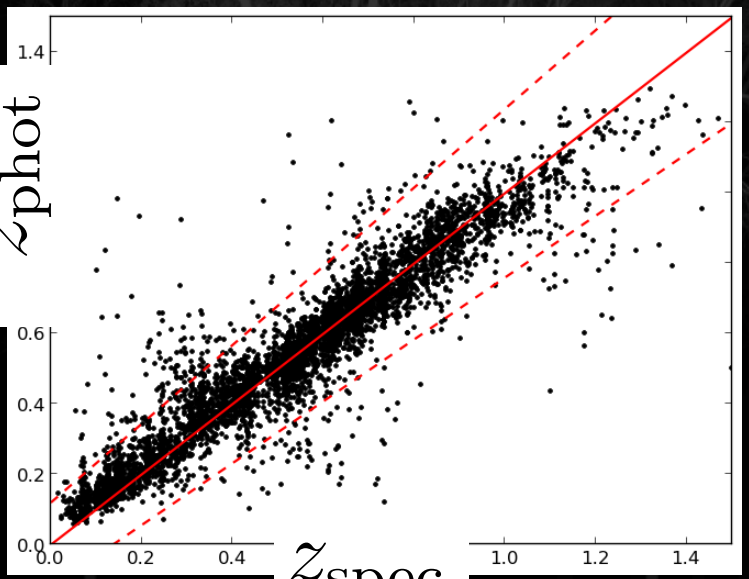


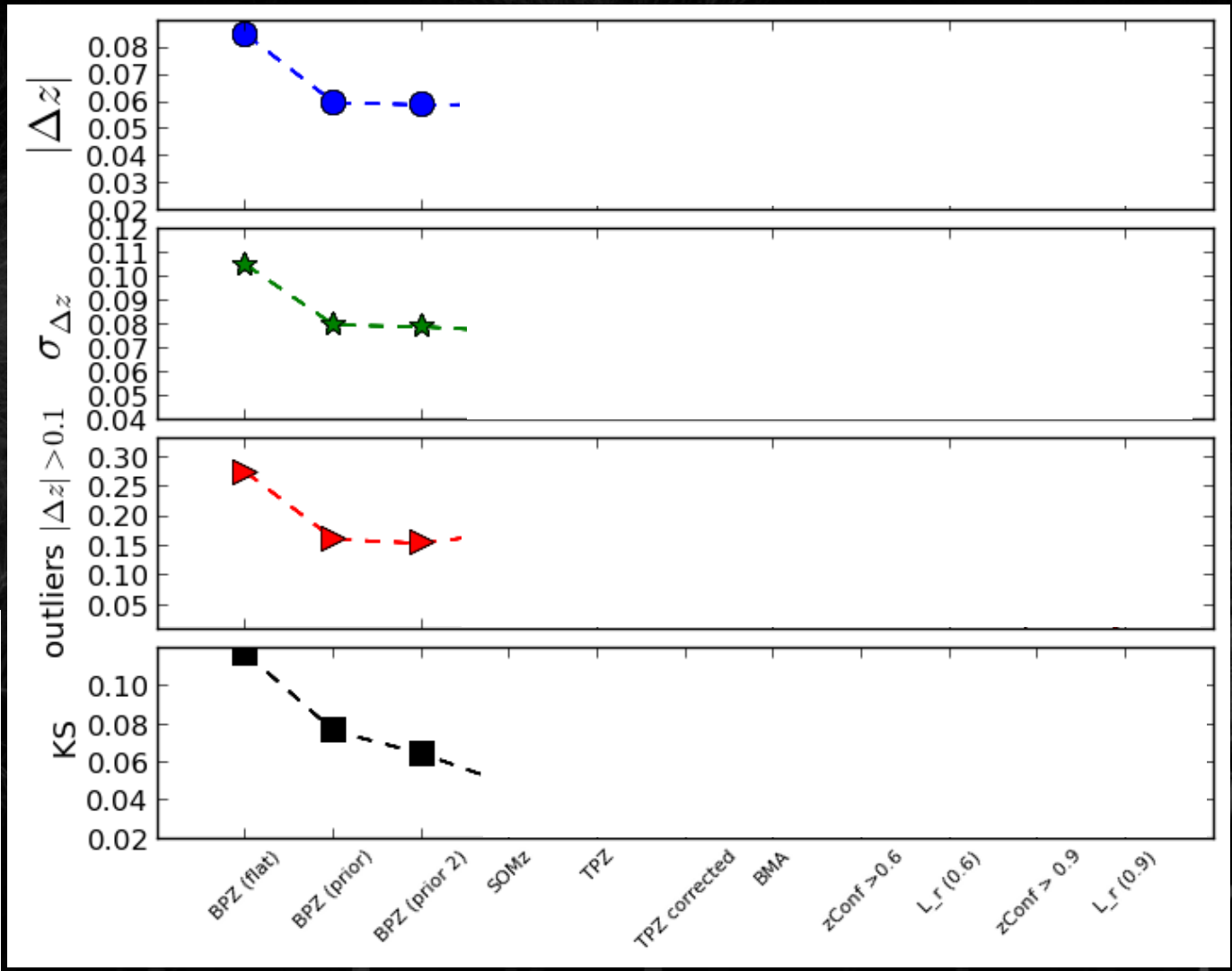



# Photo-$z$ PDF combination: Results

Averaged metrics for all test galaxies

$$\Delta z = \frac{|z_{\rm spec} - z_{\rm phot}|}{1 + z_{\rm spec}}$$

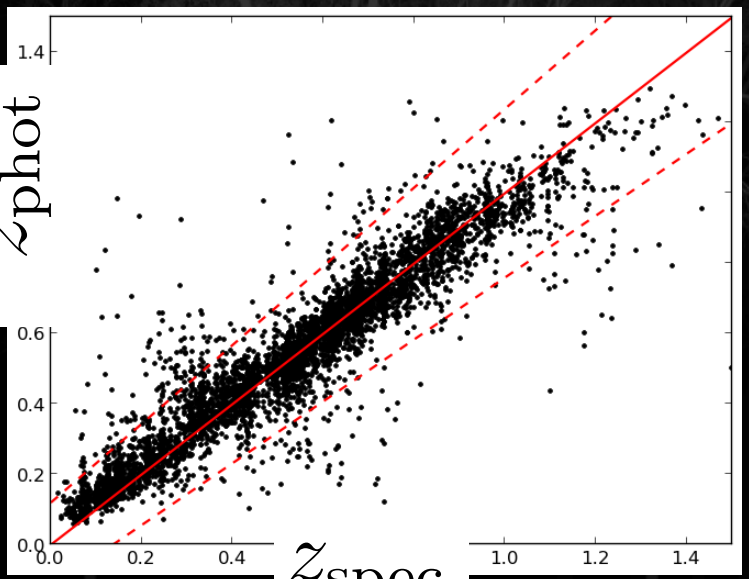


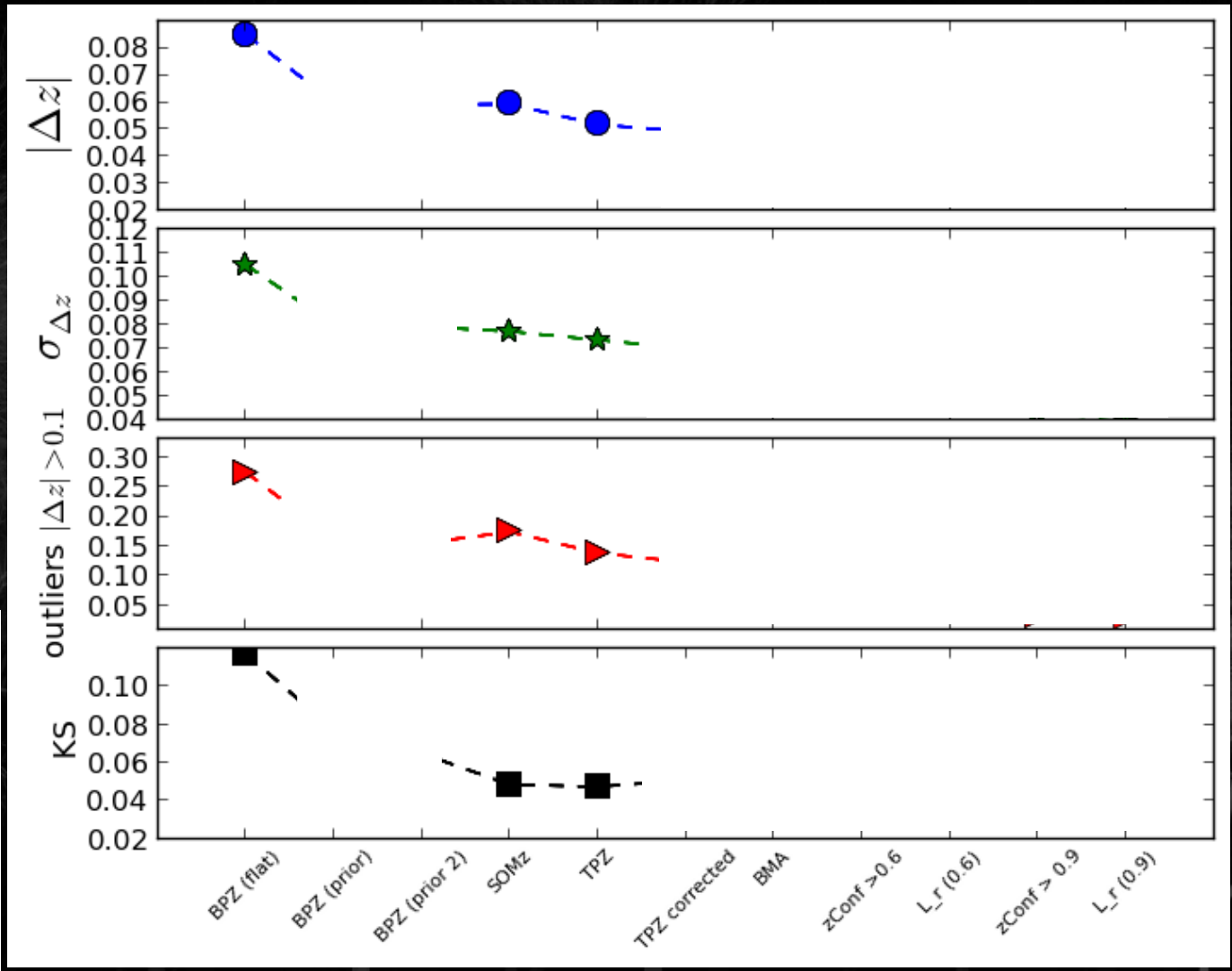



# Photo-$z$ PDF combination: Results

Averaged metrics for all test galaxies

$$\Delta z = \frac{|z_{\rm spec} - z_{\rm phot}|}{1 + z_{\rm spec}}$$

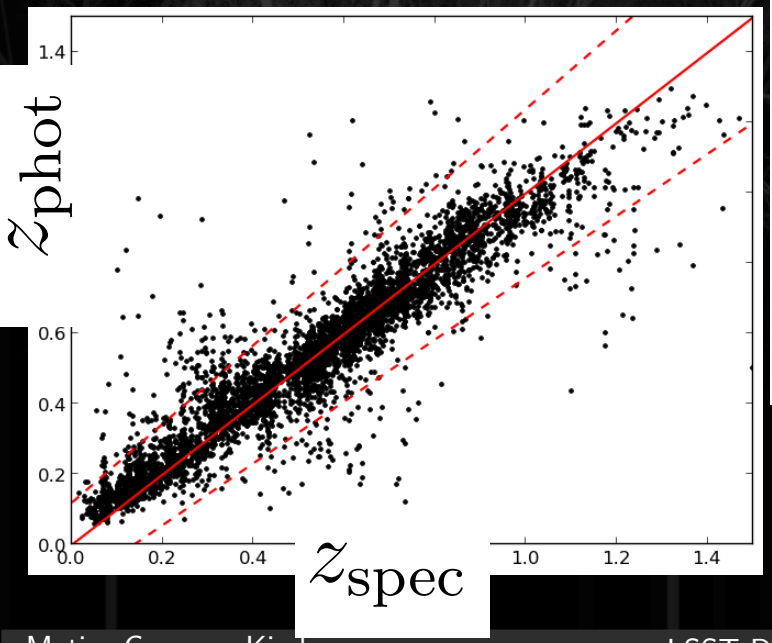


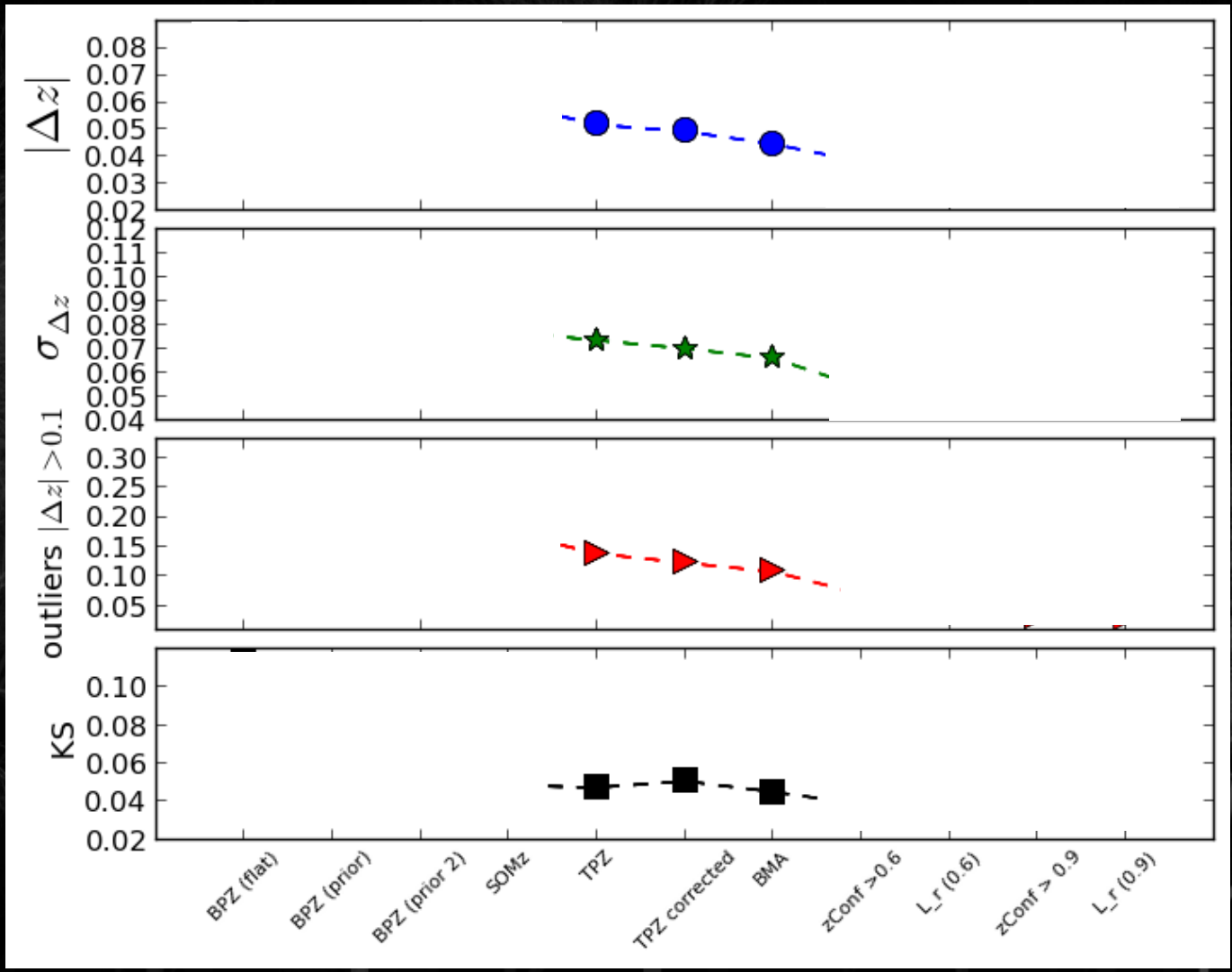



# Photo-$z$ PDF combination: Results

Averaged metrics for all test galaxies

$$\Delta z = \frac{|z_{\rm spec} - z_{\rm phot}|}{1 + z_{\rm spec}}$$

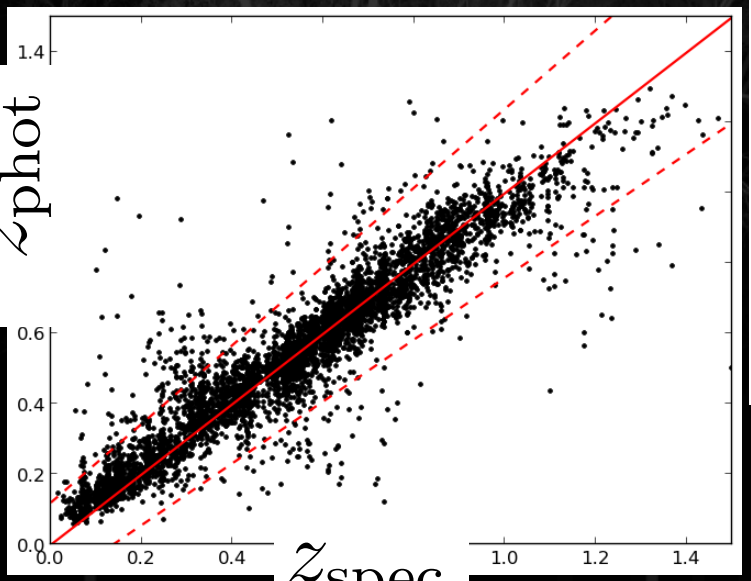


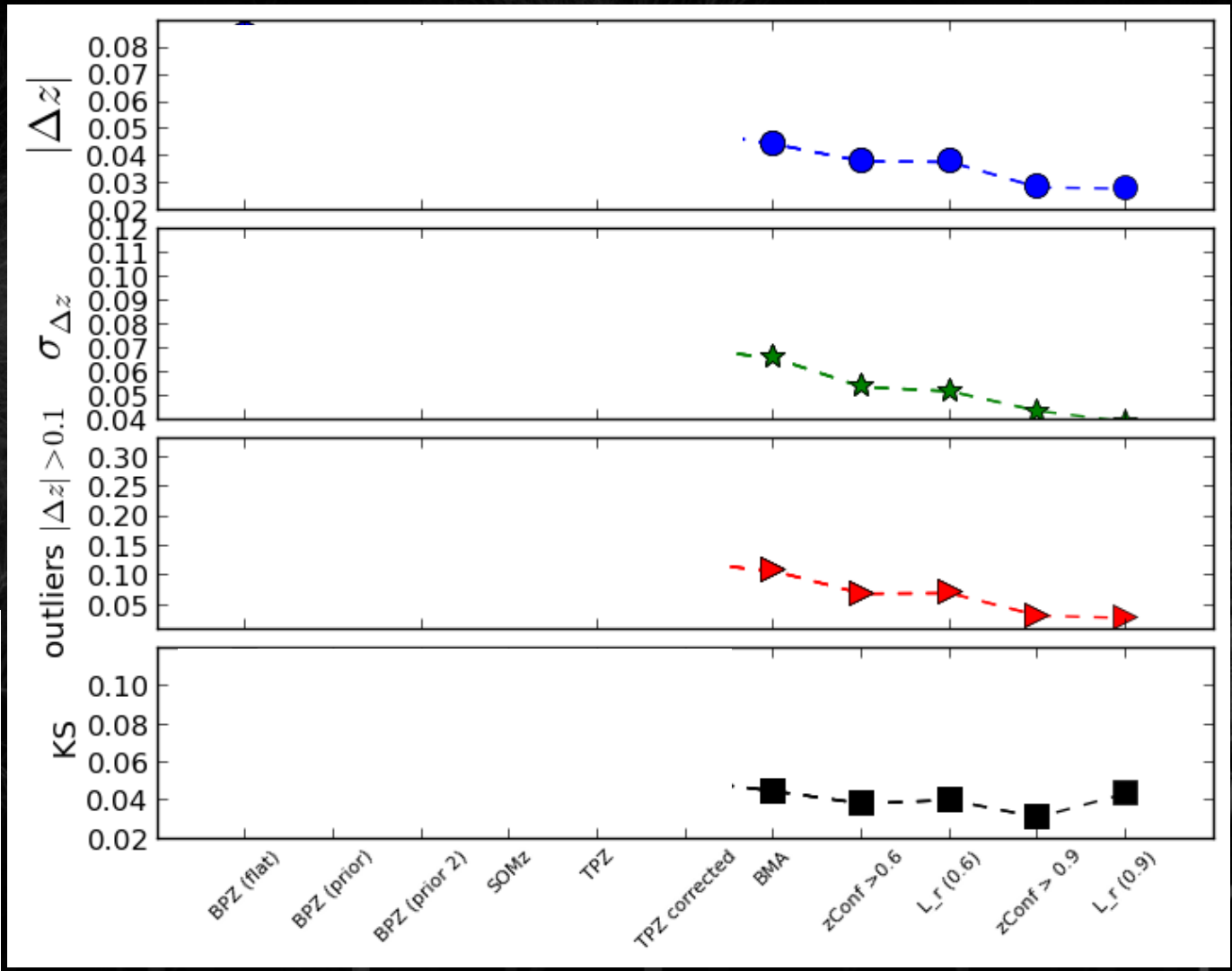



# Photo-$z$ PDF combination: Results

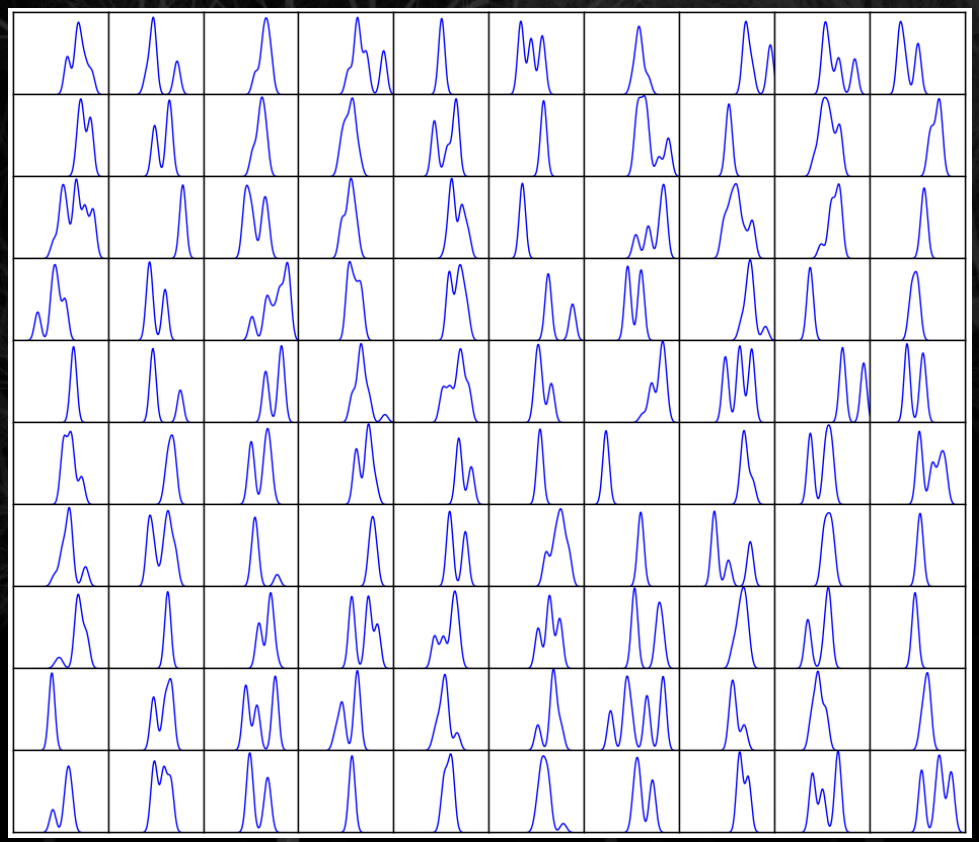
Averaged metrics for all test galaxies

$$\Delta z = \frac{|z_{\rm spec} - z_{\rm phot}|}{1 + z_{\rm spec}}$$



# Photo-$z$ PDF storage



# Photo-$z$ PDF storage: Strategies

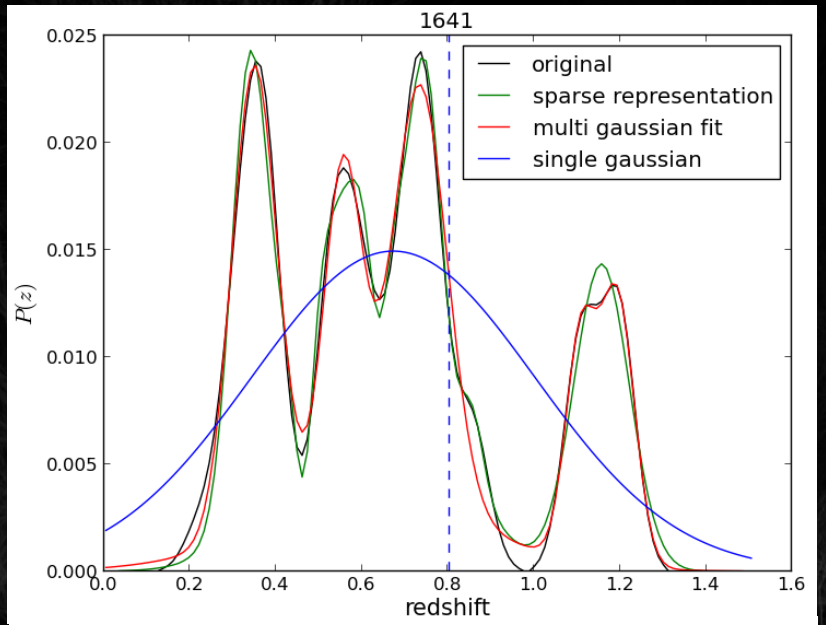
Interpolation

Fixed Gaussian fit

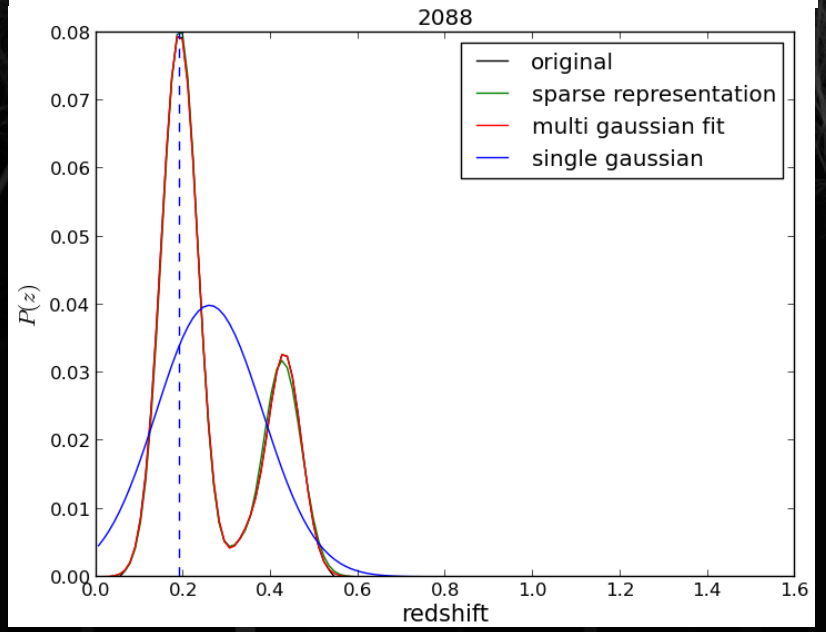
Multi-Gaussan fit

Sparse representation techniques

(Carrasco Kind, Brunner & Ching, in prep.)

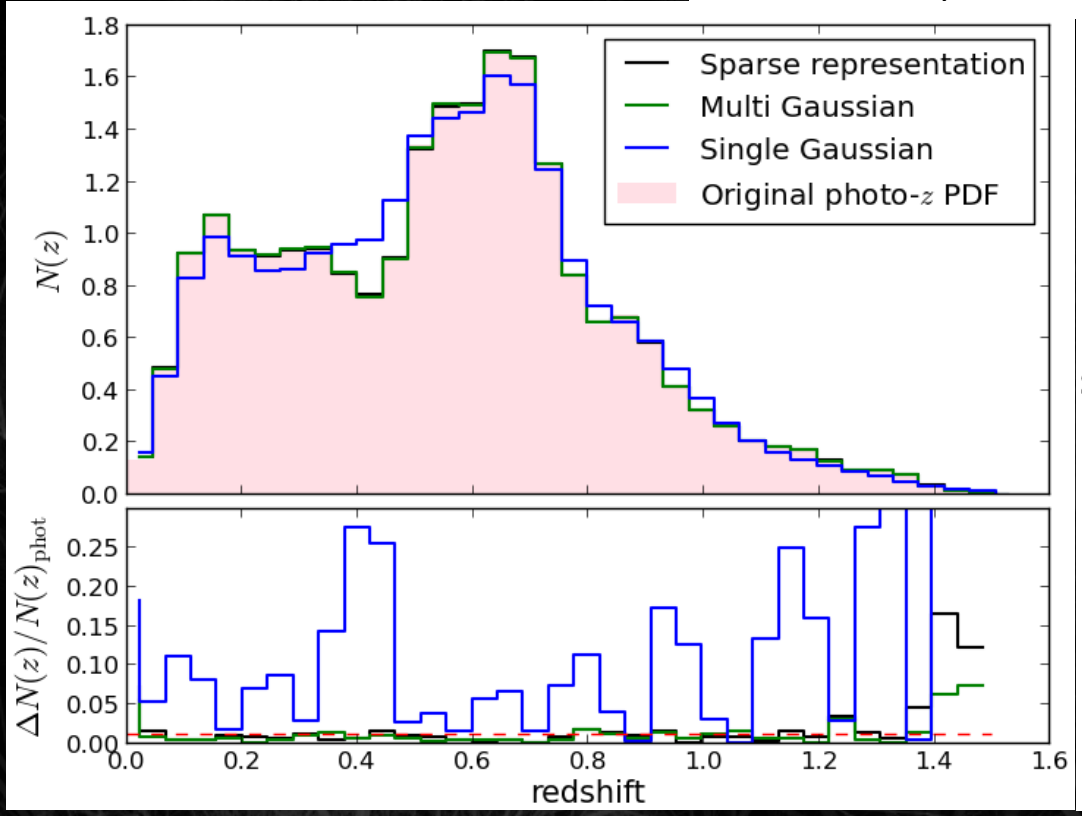


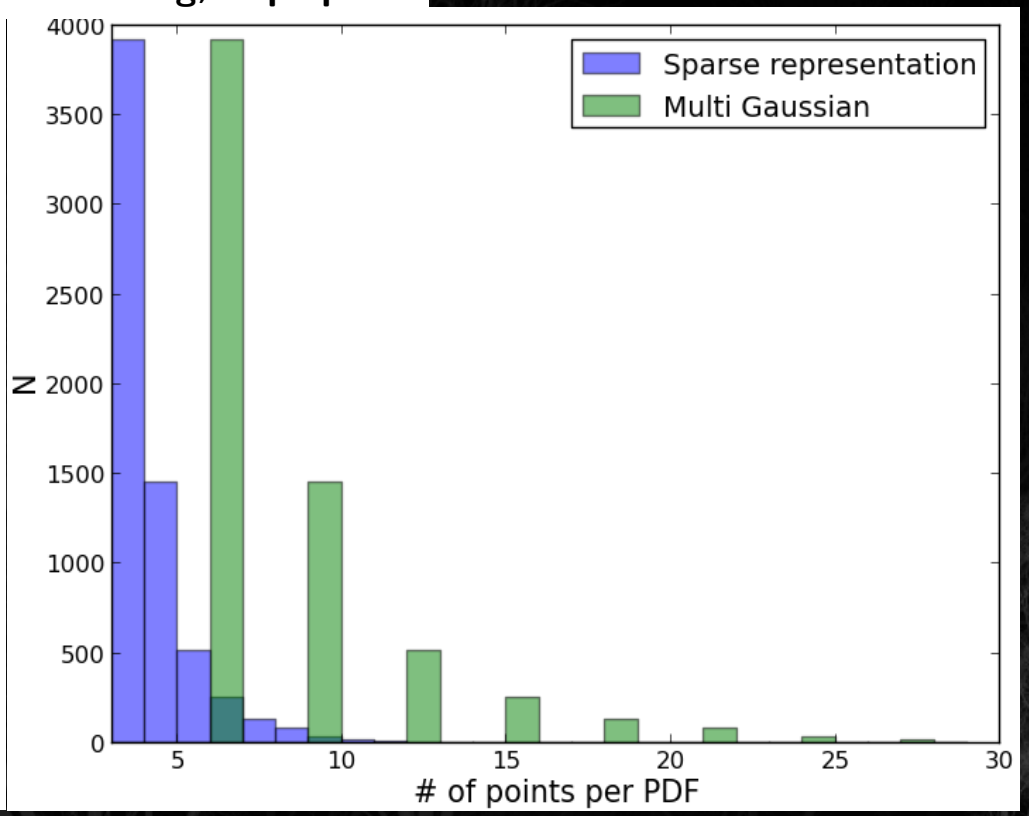

**Carrasco Kind, Brunner & Ching, in prep.**



# Photo-$z$ PDF storage: Results

**Carrasco Kind, Brunner & Ching, in prep.**

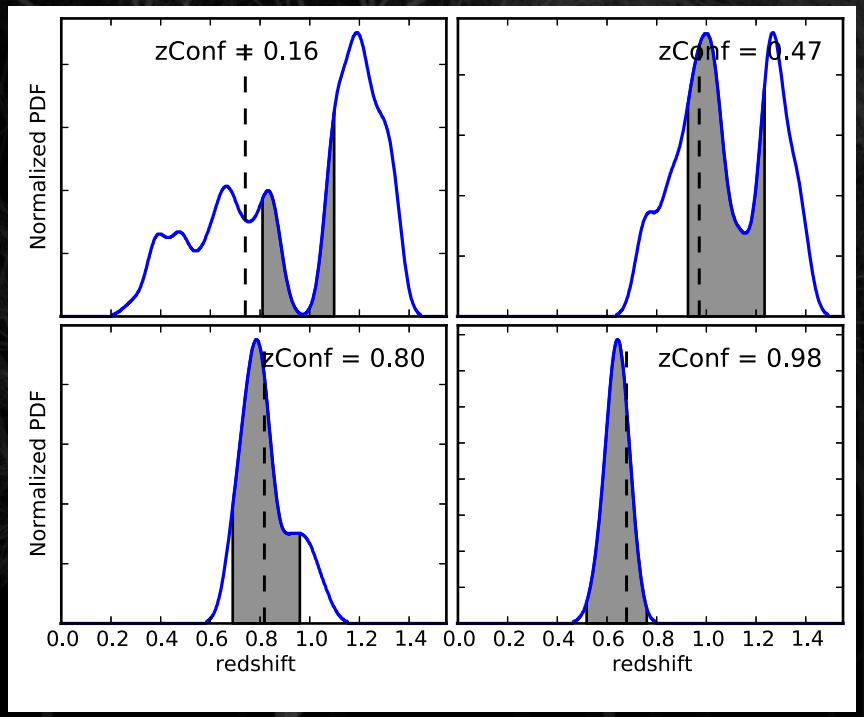


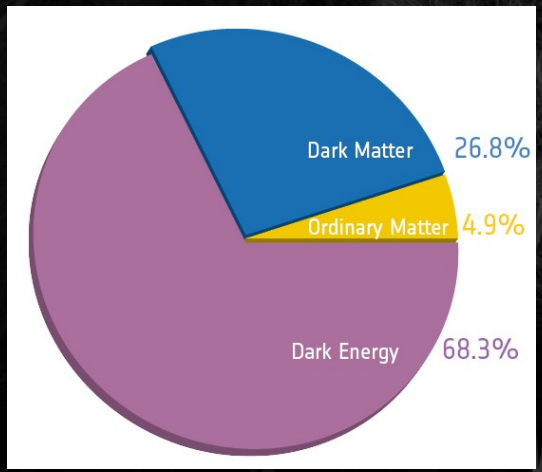




Differences less than 1% using Multi Gaussian or sparse representation

Sparse representation saves $\sim$ 50% of disk space!

# Photo-$z$ PDF applications



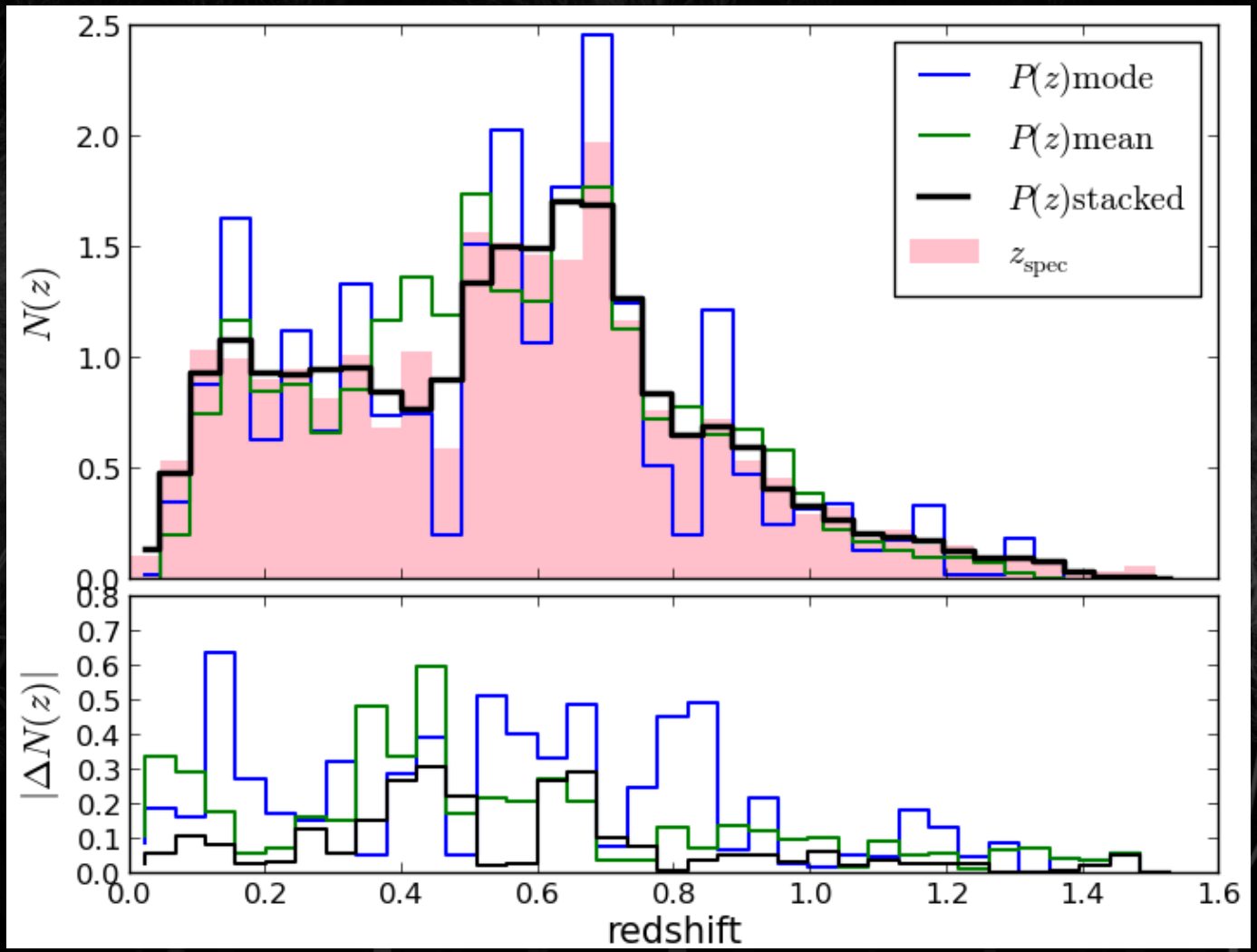



# Photo-$z$ PDF application: $N(z)$

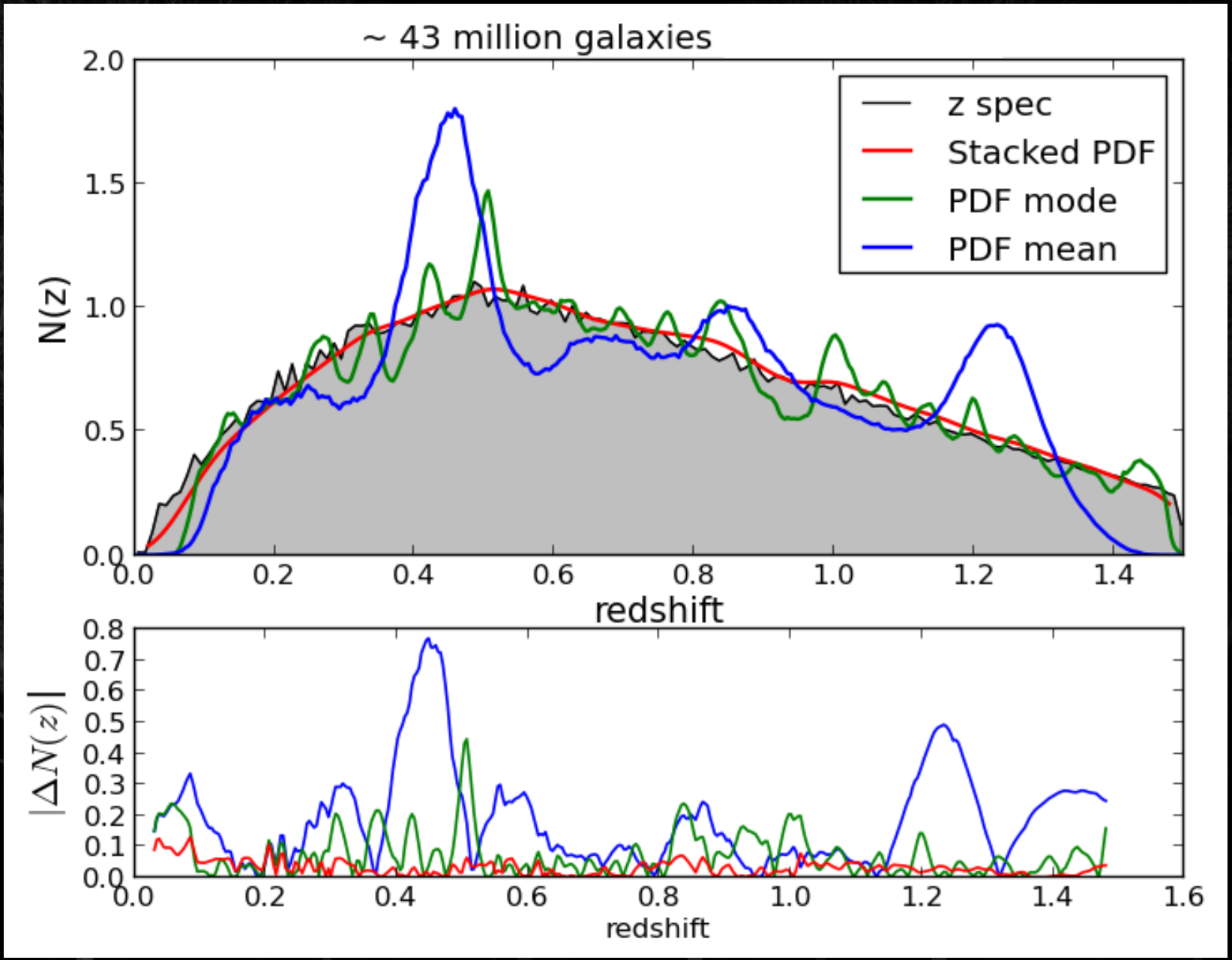
$N(z)$ distribution of galaxies, simple yet important feature

Stacked PDF produces better distribution than taken the mean of the PDF

Very important for clustering and weak lensing studies



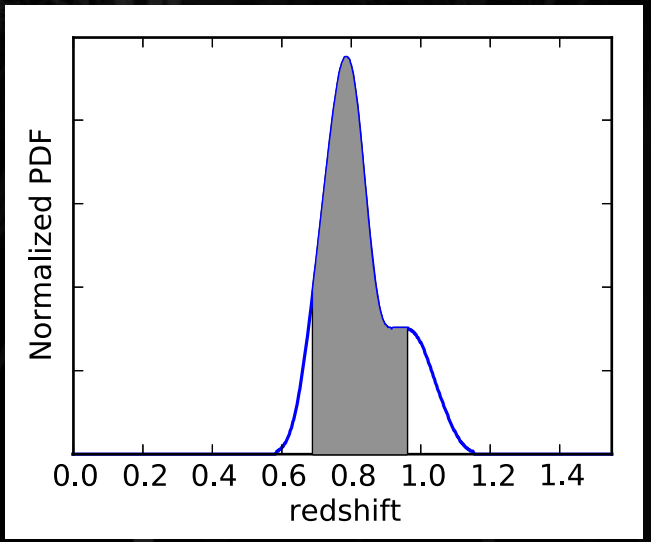
# Photo-$z$ PDF application: $N(z)$



# Photo-$z$ PDF application: Angular Power Spectrum

- The angular power spectrum (APS) contains important information about the matter density field
- 2D projection of $P(k)$ using $N(z)$in the kernel
- Constrains cosmological models. Could be used to resolve BAOs
- Use photo-$z$ PDF in overdensities

$$\delta_i = \frac{\Omega_{survey} \sum_j^{N_{in}} \int_{z_1}^{z_2} P_{ij}(z) dz}{\Omega_i \sum_j^{N_{tot}} \int_{z_1}^{z_2} P_j(z) dz} - 1$$



# Photo-$z$ PDF application: $C_\ell$ and $\omega(\theta)$

Limber approximation with no redshift-space distortions and scale-independent bias $b$:

$$C_\ell = \frac{\ell(\ell+1)}{2\pi} b^2 \int dz \phi^2(z) \frac{H(z)}{r^2(z)} P\left(\frac{\ell+1/2}{r(z)}, z\right)$$

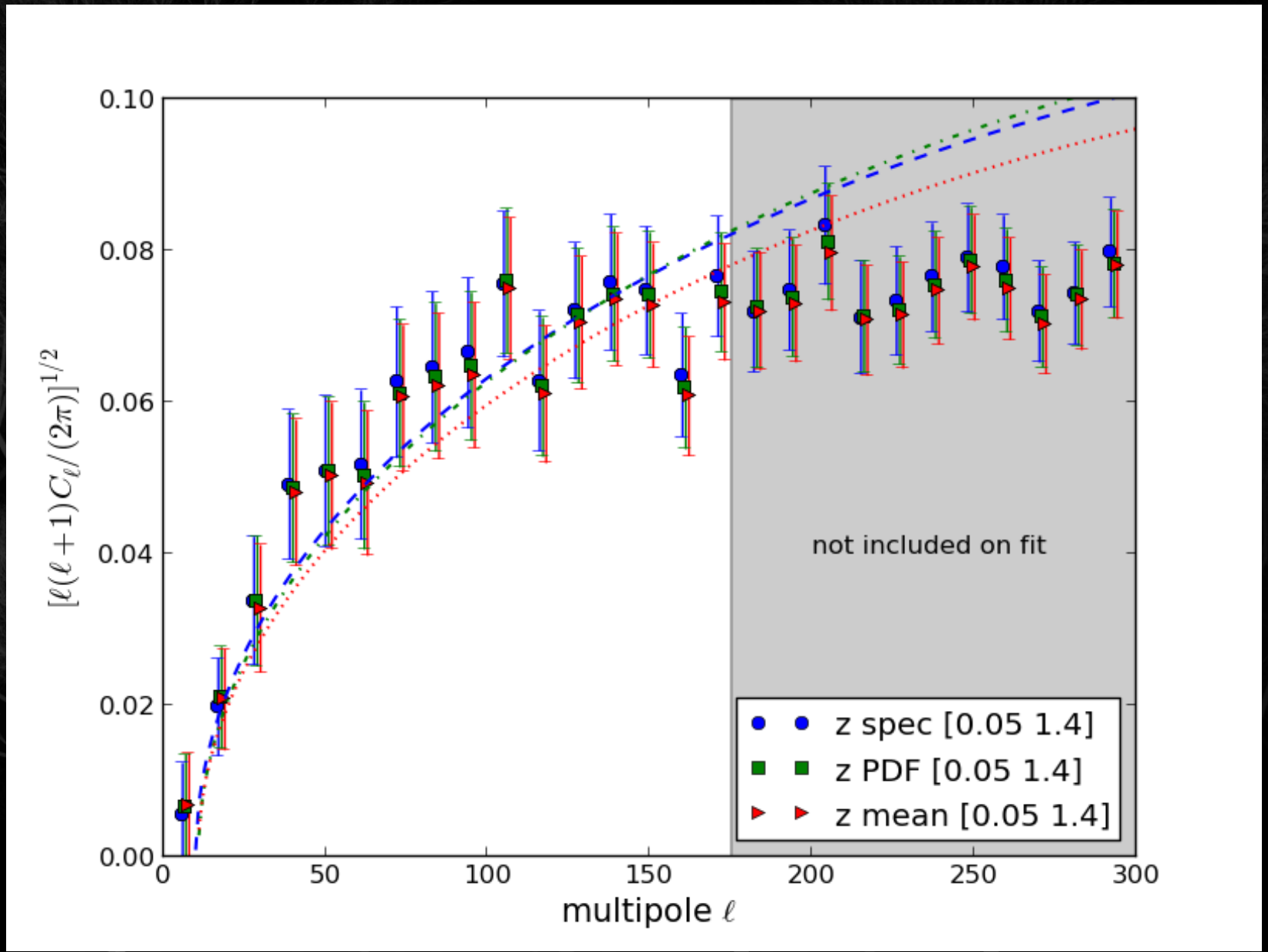
CAMB and HALOFIT for non linear $P(k,z)$

$\phi(z)$ is the galaxy distribution $N(z)$

Fitting using Monte Carlo Markov Chain methods

$$\chi^2(a_p) = \sum_{bb'} (\ln C_b - \ln C_b^T) C_b F_{bb'} C_{b'} (\ln C_{b'} - \ln C_{b'}^T)$$

# Photo-$z$ PDF application: $C_\ell$ and $\omega(\theta)$



# Conclusions

* Individual techniques: good information

* Combination technique: more and better information

* Sparse representation saves 50% in PDF storage without lossing accuracy

* Sparse representation can be incorporate in theoretical framework

* Photo-$z$ PDF in cosmological analysis to enhance signal

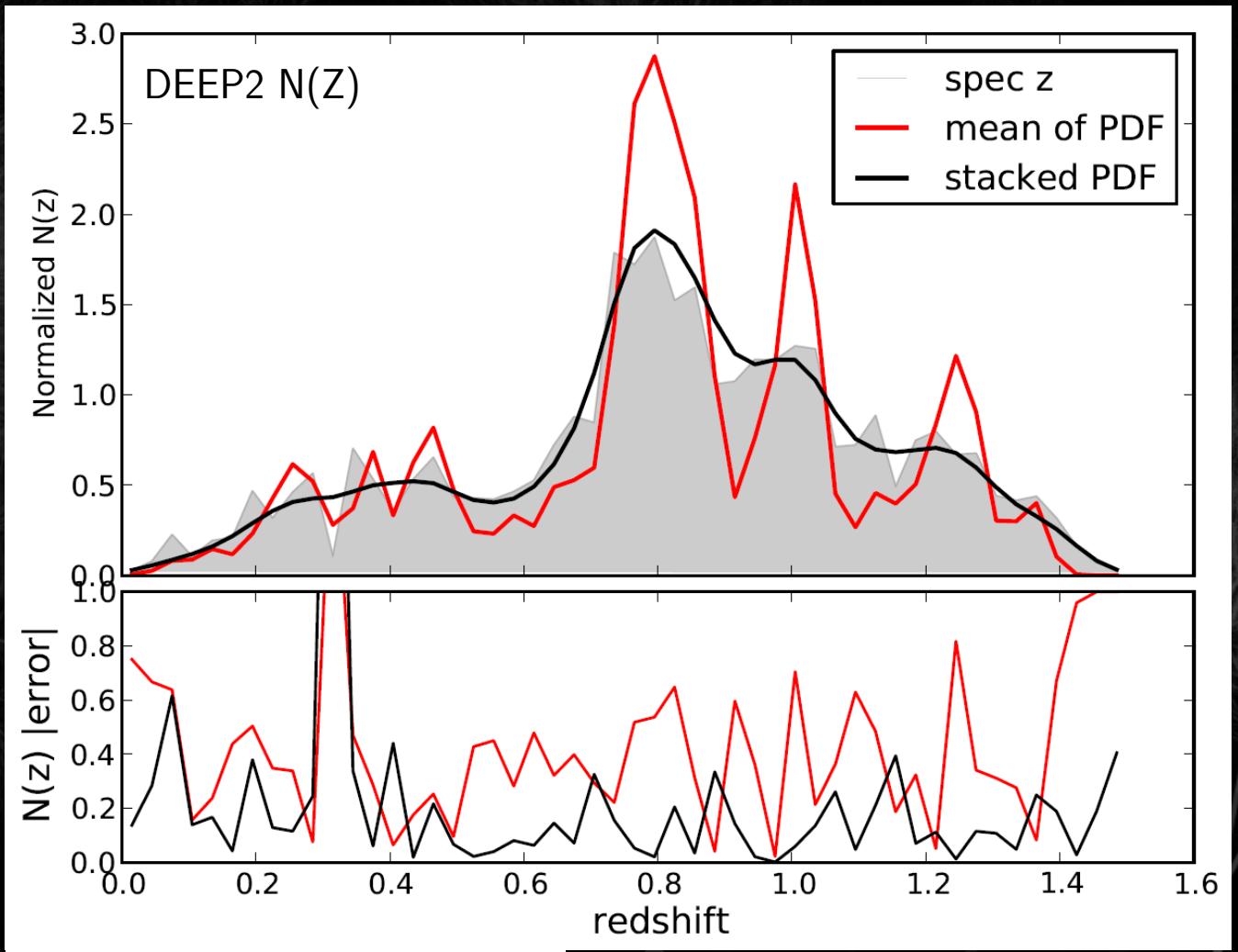
# EXTRA SLIDES

# Using photo-$z$ PDF in cosmological anaysis

$N(z)$ distribution of galaxies, simple yet important feature

Stacked PDF produces better distribution than taken the mean of the PDF

Very important for clustering and weak lensing studies



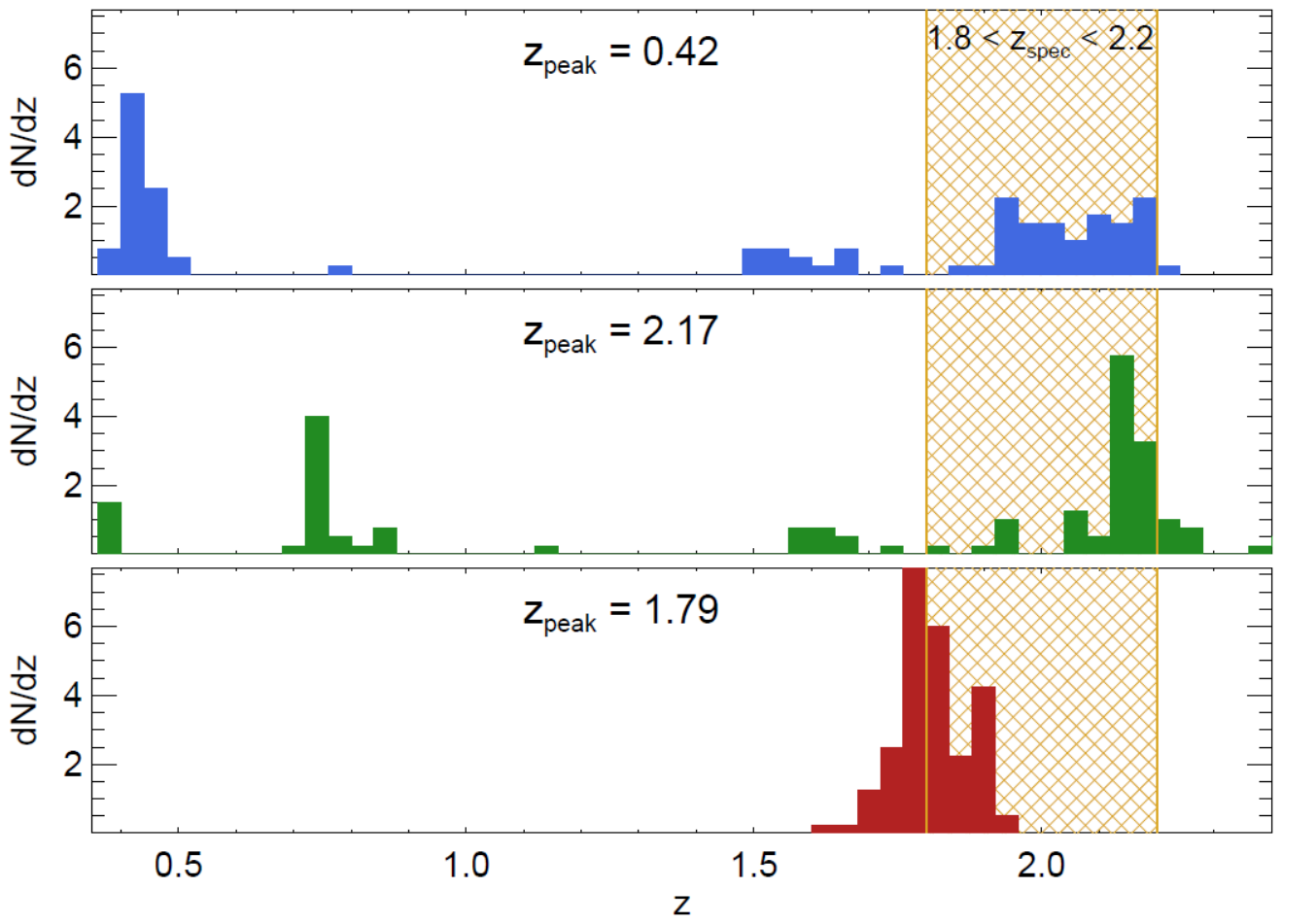

**Carrasco Kind & Brunner 2013a**

# Example application of photo-$z$ PDF

Incorporating PDF on clustering measurements

Problems of using mode of photo-$z$ PDF

Extend to other measurements




**Myers, White & Ball 2009**

# Photometric redshift PDFs using TPZ

We use TPZ to generate photo-$z$ for all galaxies.

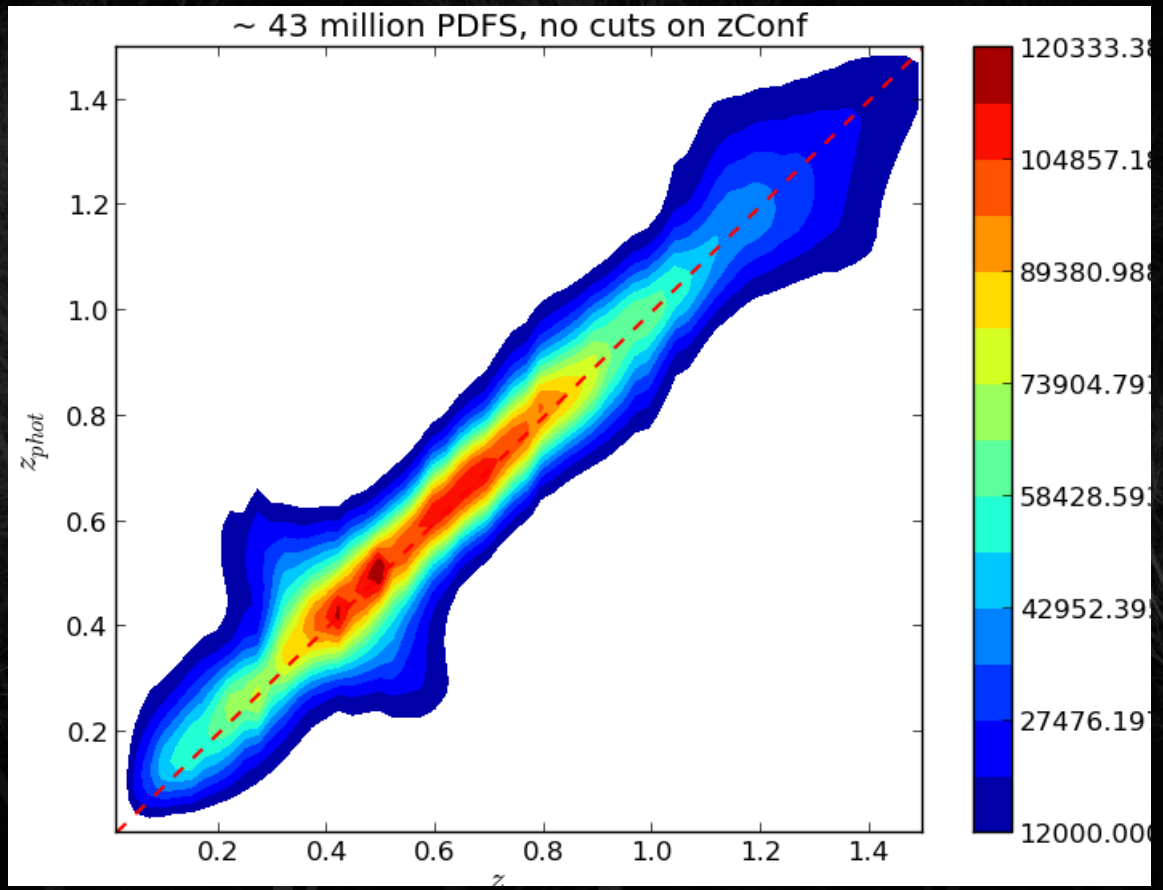
100,00 for training

5 magnitudes only

$\sim$ 0.17 sec per PDF

Store 43 million PDFs for analysis

No outlier removal



# Photometric redshift PDFs using TPZ

## Metrics

$(\Delta z = z_{phot} - z_{spec})$

$< \Delta z >= 0.0088$

$< |\Delta z| >= 0.089$

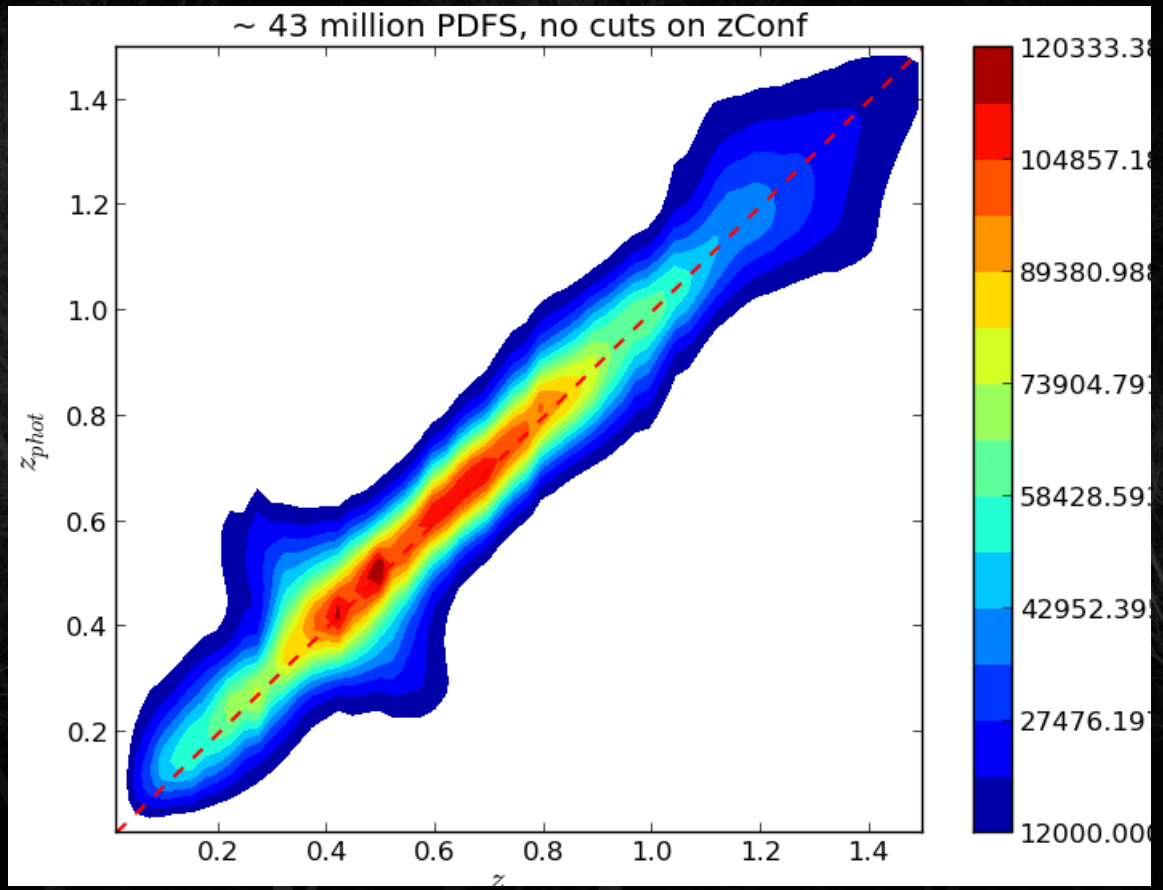
$\sigma_{\Delta z} = 0.1421$

$\sigma_{|\Delta z|} = 0.1109$

$\sigma_{68} = 0.0885$

$frac > 2\sigma = 0.0531$

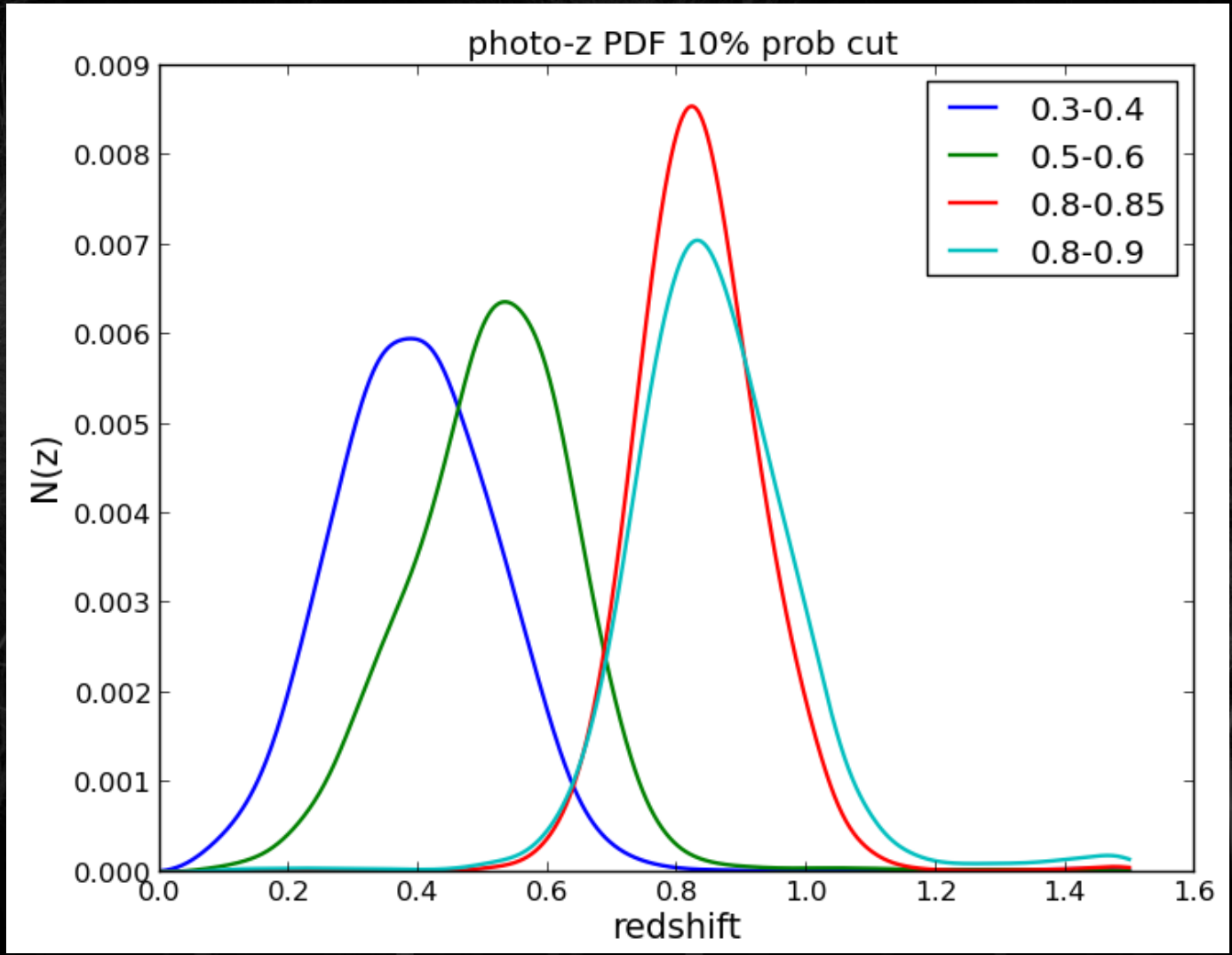
$frac > 3\sigma = 0.0207$



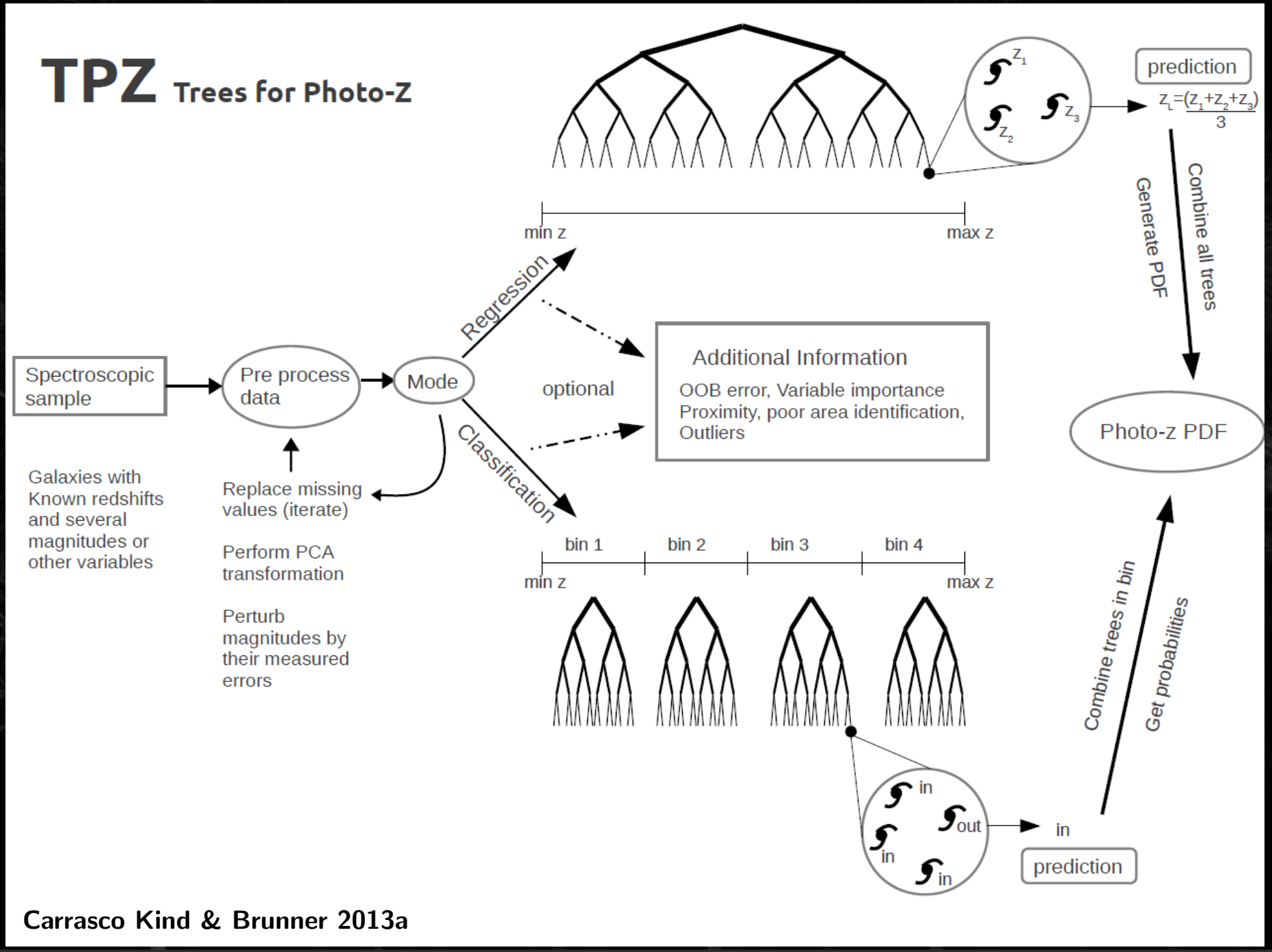
# Also in redshift shells

We consider only PDF with at least 10% of its area inside redshift shell

N(z) and overdensities from stacked PDFs



# TPZ : Scheme




**Carrasco Kind & Brunner 2013a**

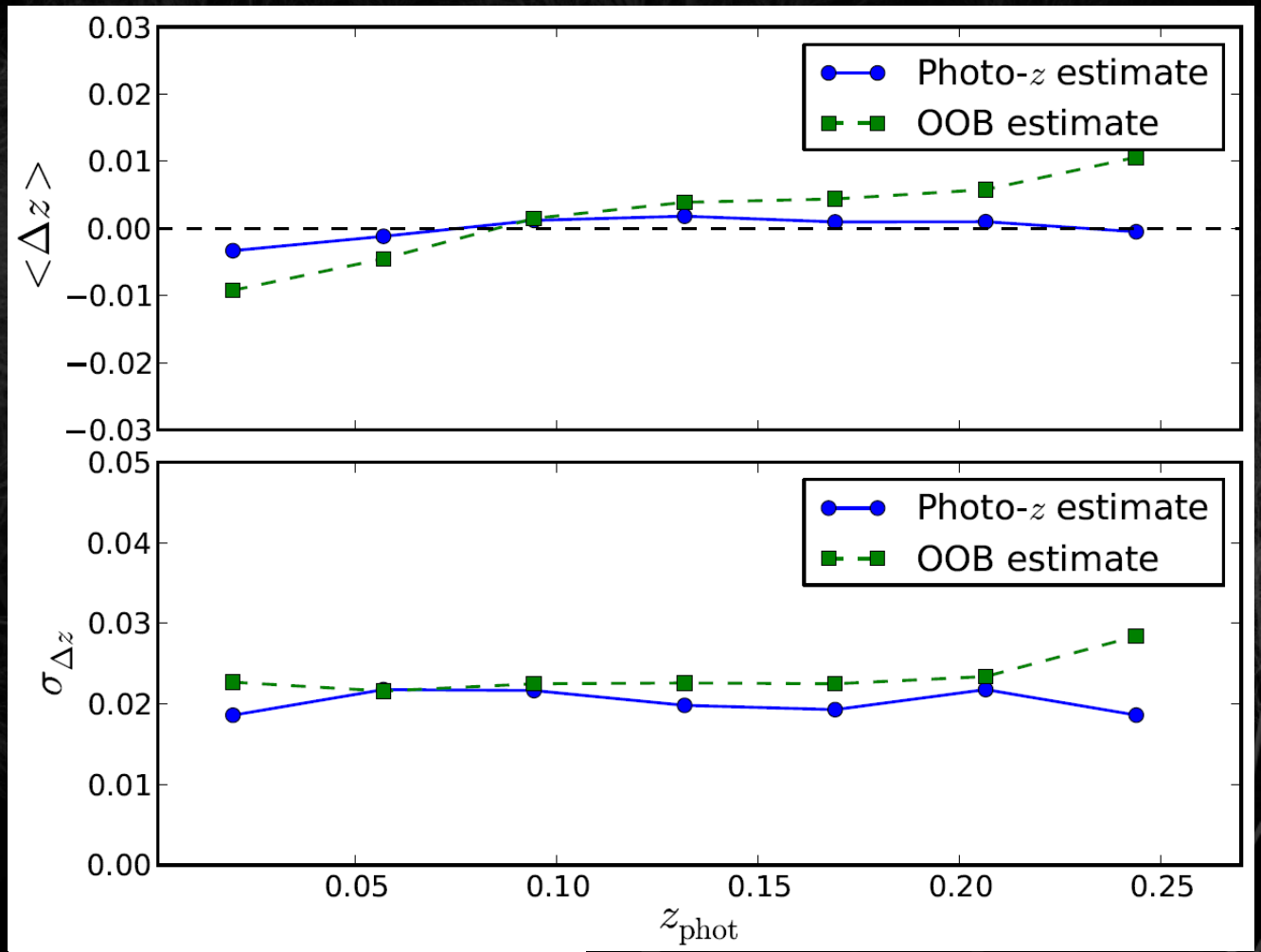
# TPZ: Ancillary information - *prior error* -

Using *Out-of-Bag* data TPZ provides useful extra information

No need of a validation set, use full training set.

Example application on SDSS MGS, 40,000 test and 15,000 training galaxies

A prior unbiased estimations of errors!



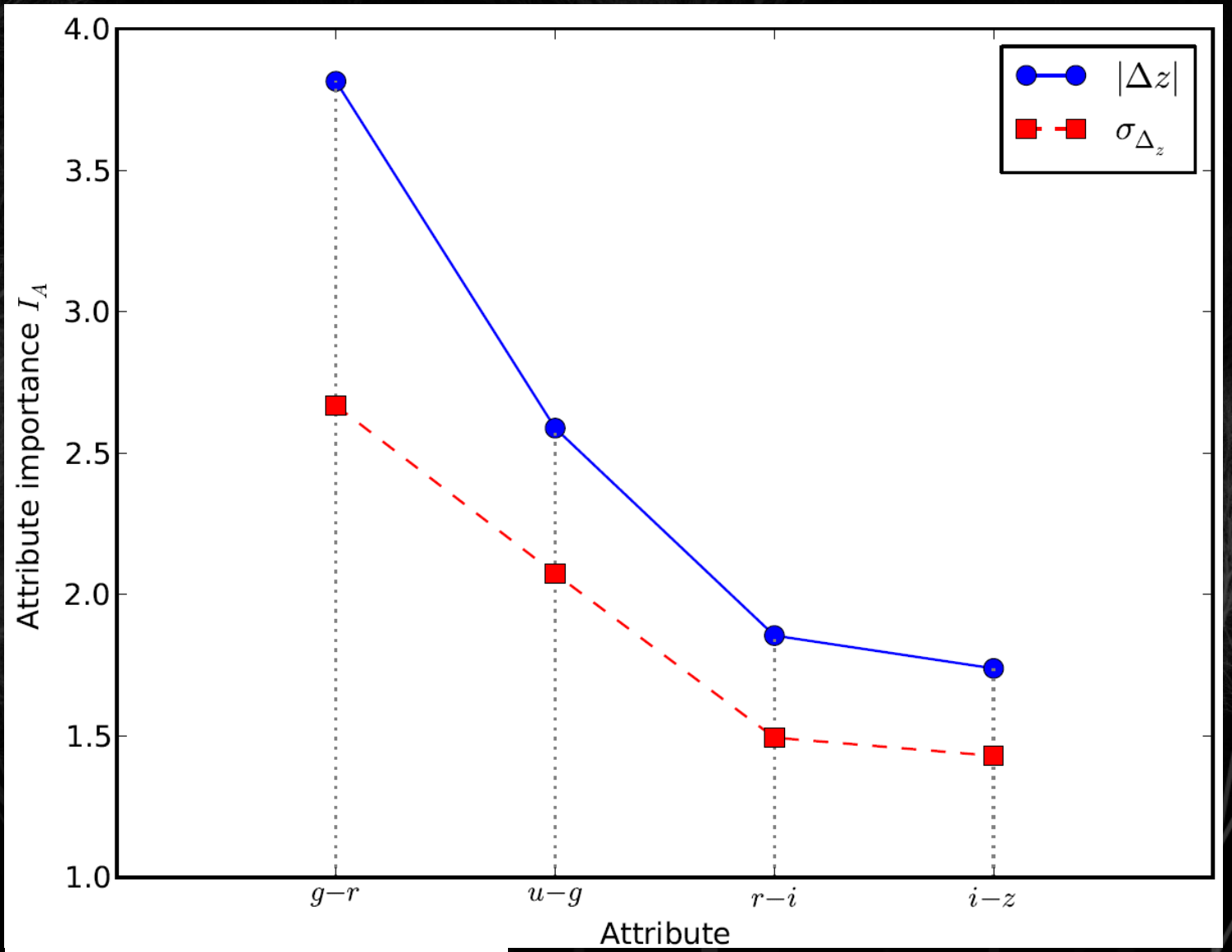

**Carrasco Kind & Brunner 2013a**

# TPZ: Ancillary information - *Attribute importance* -

Ranking statistical only

Useful for removing unimportant variables reducing the noise

Most important attributes to construct impotance map

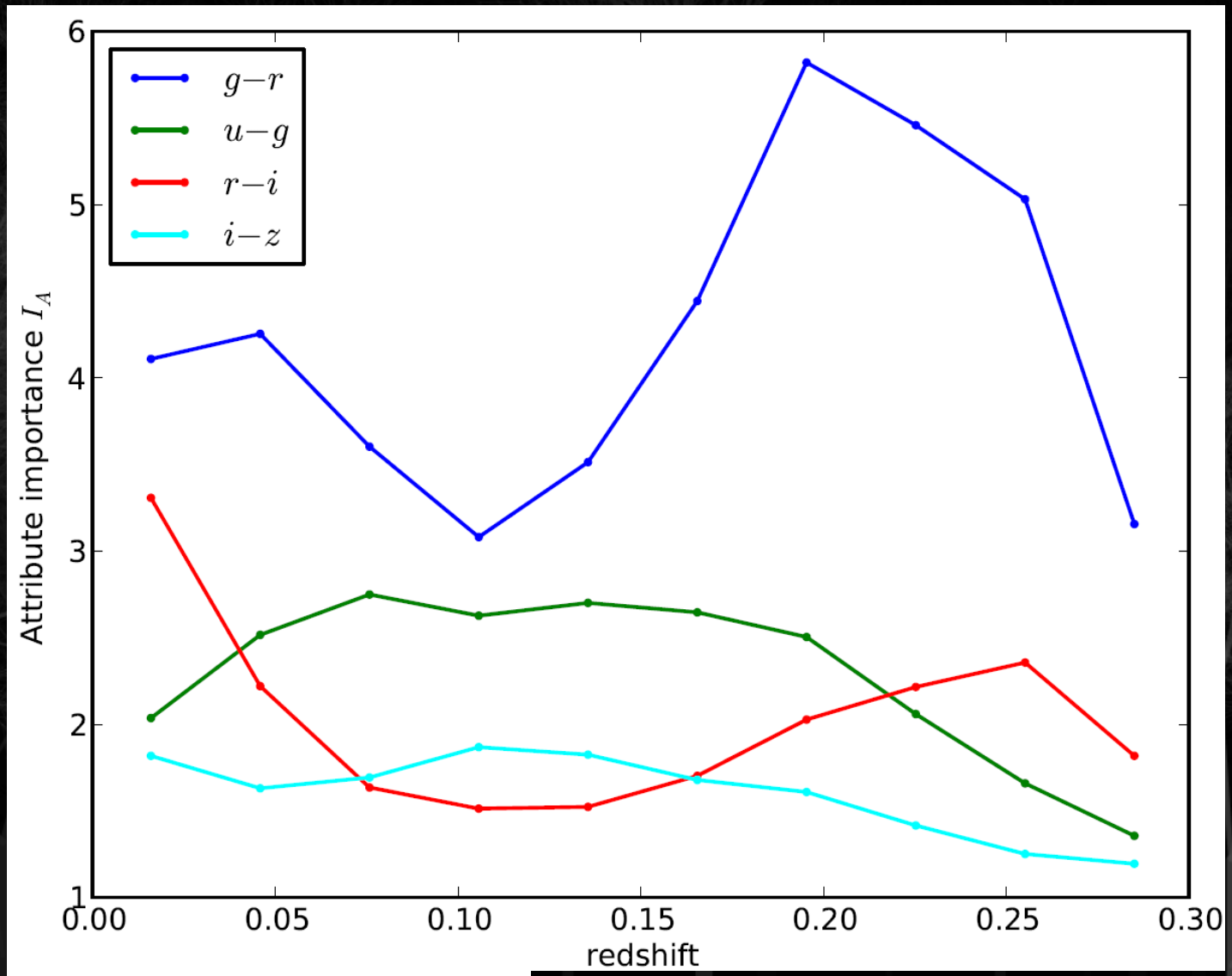



**Carrasco Kind & Brunner 2013a**

# TPZ: Ancillary information - *Attribute importance* -

How much the metrics change as we permute the attributes one at a time

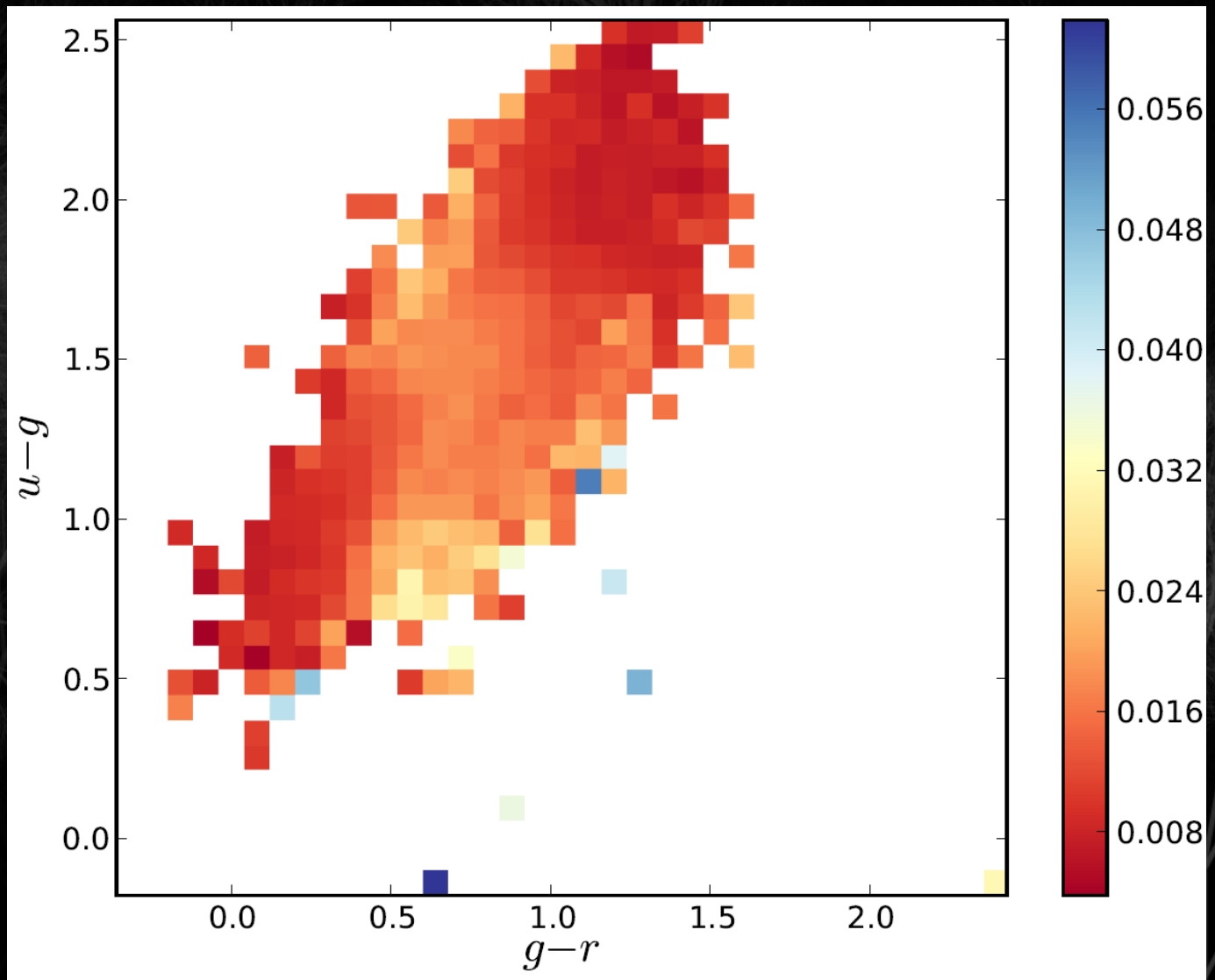
For SDSS the $g-r$ color is the most important attribute




**Carrasco Kind & Brunner 2013a**

# TPZ: Ancillary information - *Poor area identification* -

- Map of performance using two most important colors
- The redder the better
- Bimodality of SDSS galaxies
- Narrow follow up observations




**Carrasco Kind & Brunner 2013a**