ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Machine Learning, Systematics, and LSS

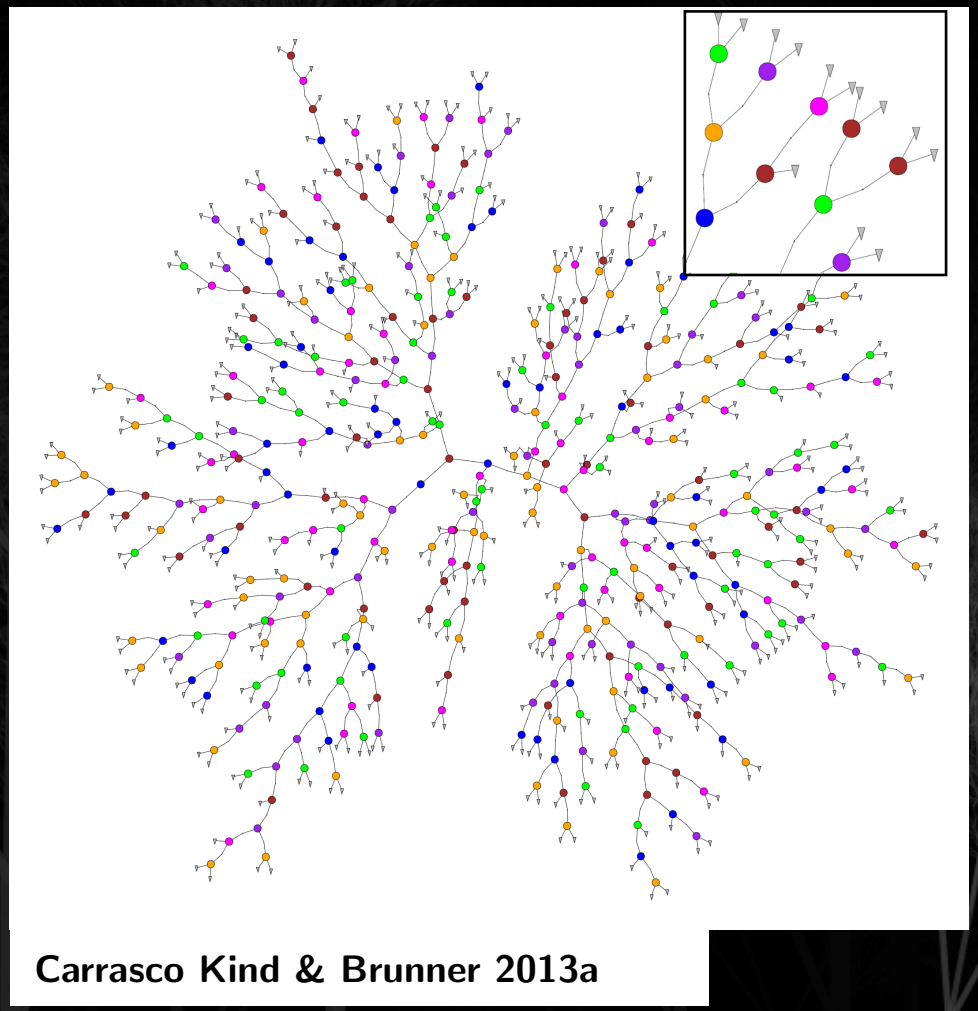
Matías Carrasco Kind
Robert J. Brunner

Department of Astronomy
University of Illinois

LSST-DESC meeting Dec 4-6, 2013

Machine Learning

Machine Learning: TPZ



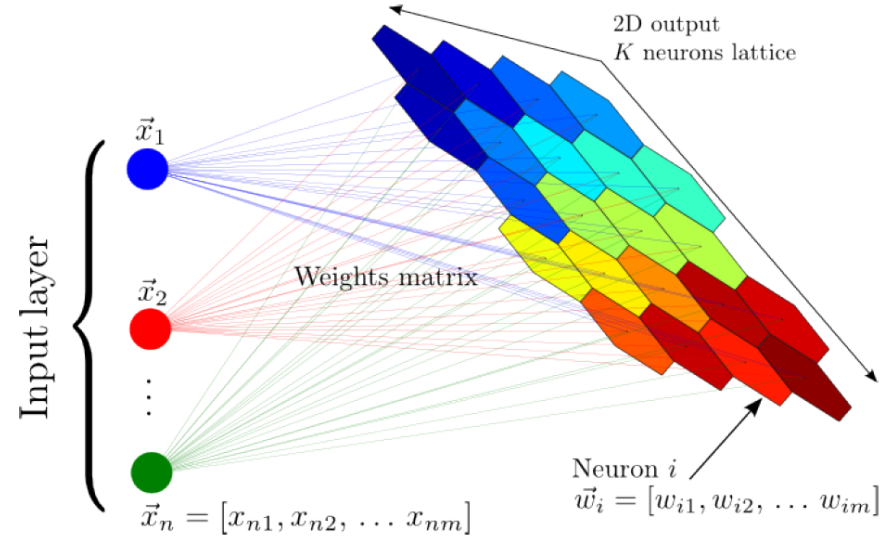
Carrasco Kind & Brunner 2013a

- 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

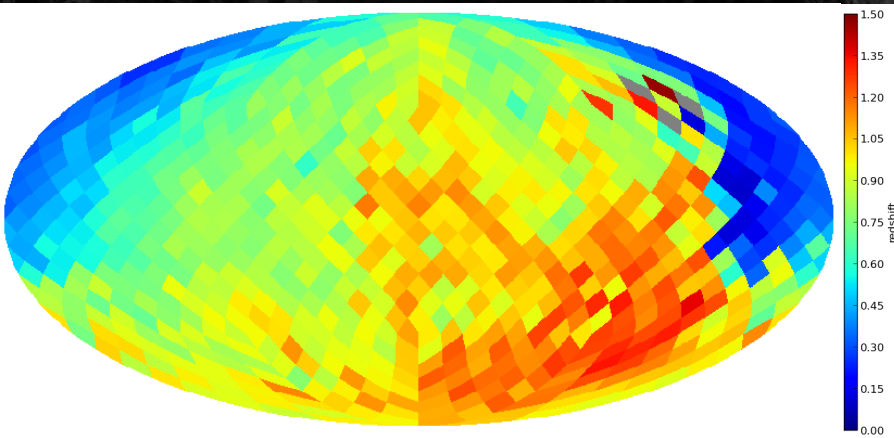
http://lcdm.astro.illinois.edu/research/TPZ.html

Machine Learning: 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



Carrasco Kind & Brunner, MNRAS submitted



Carrasco Kind & Brunner, MNRAS submitted

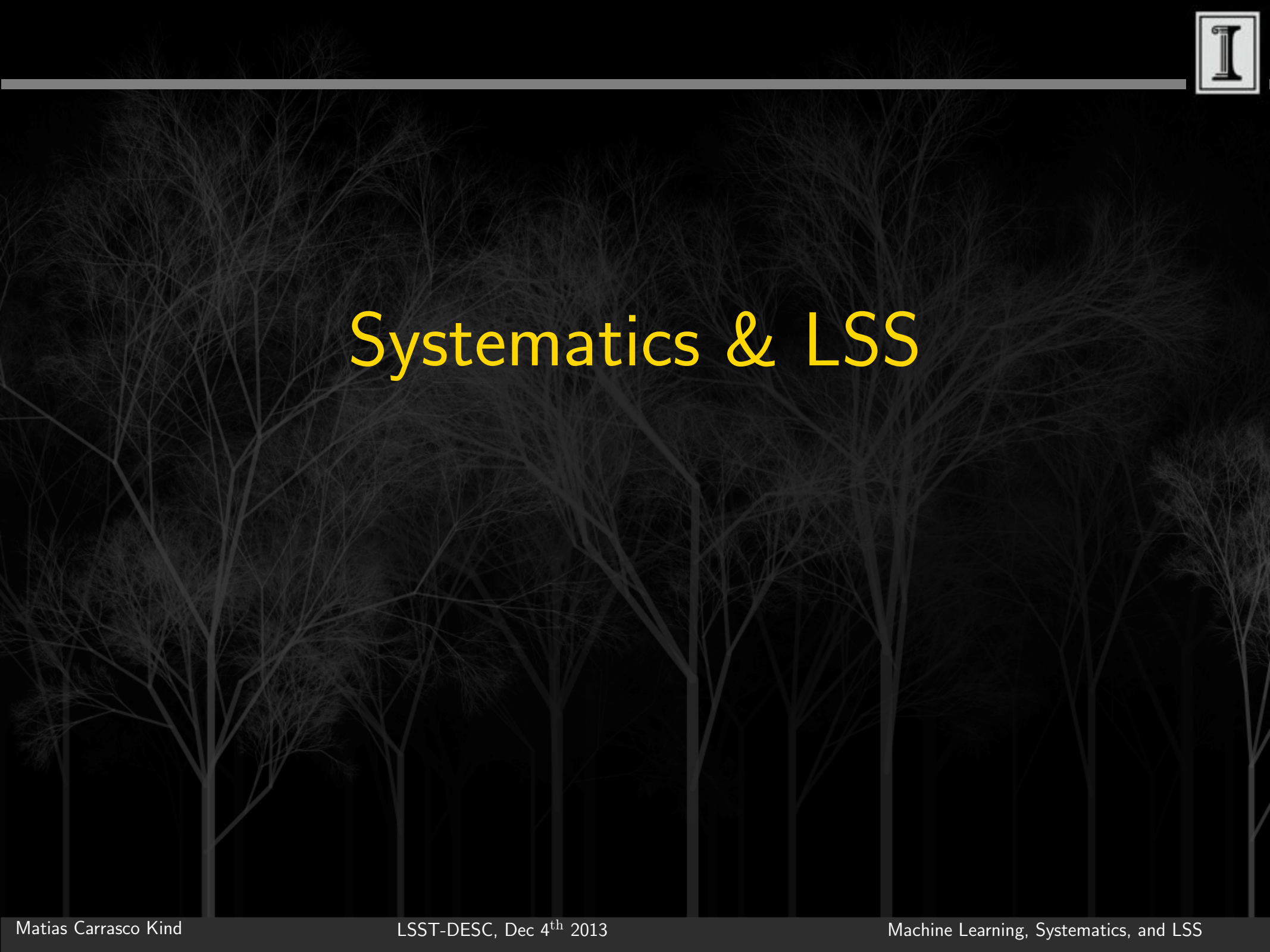
Machine Learning: Others UIUC efforts

- Random Naïve Bayes (used for spam filter) to produce photo-z priors (Carrasco Kind & Brunner, 2013b)

- Sparse representation and dictionary learning for PDF storage (Carrasco Kind, Brunner & Ching, in prep.)

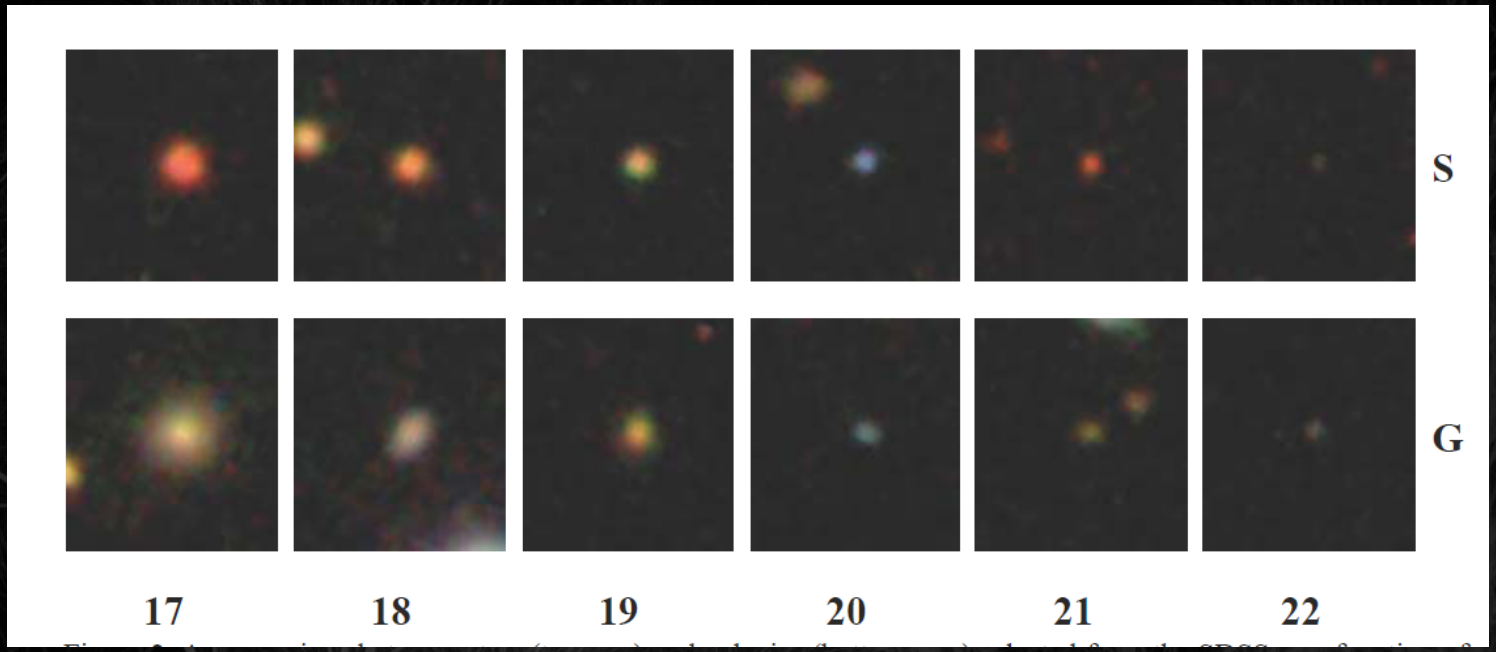
- Ensemble learning and Bayesian network for photo-z estimation and outlier rejection (Carrasco Kind & Brunner, in prep.)

- Machine Learning for Strong Lensing identification

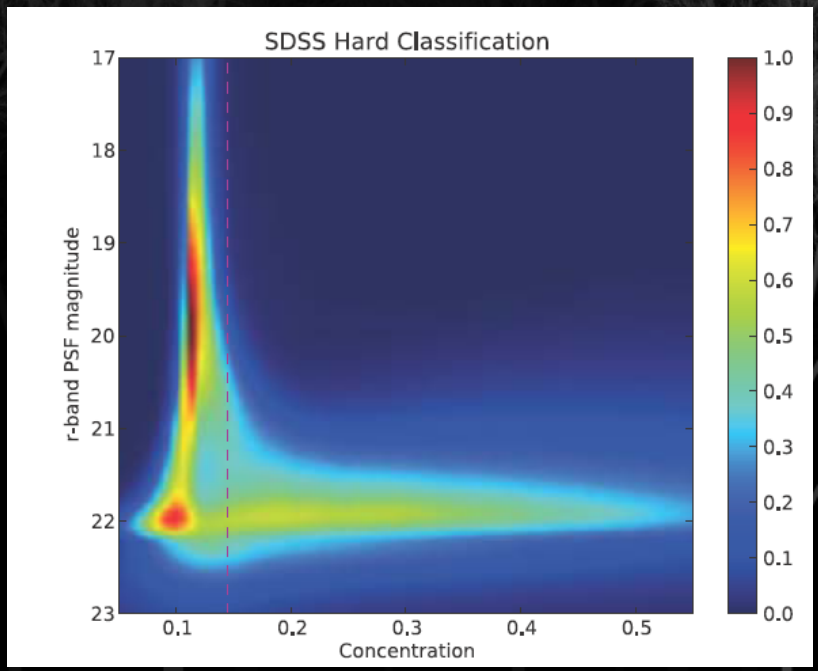


Systematics & LSS

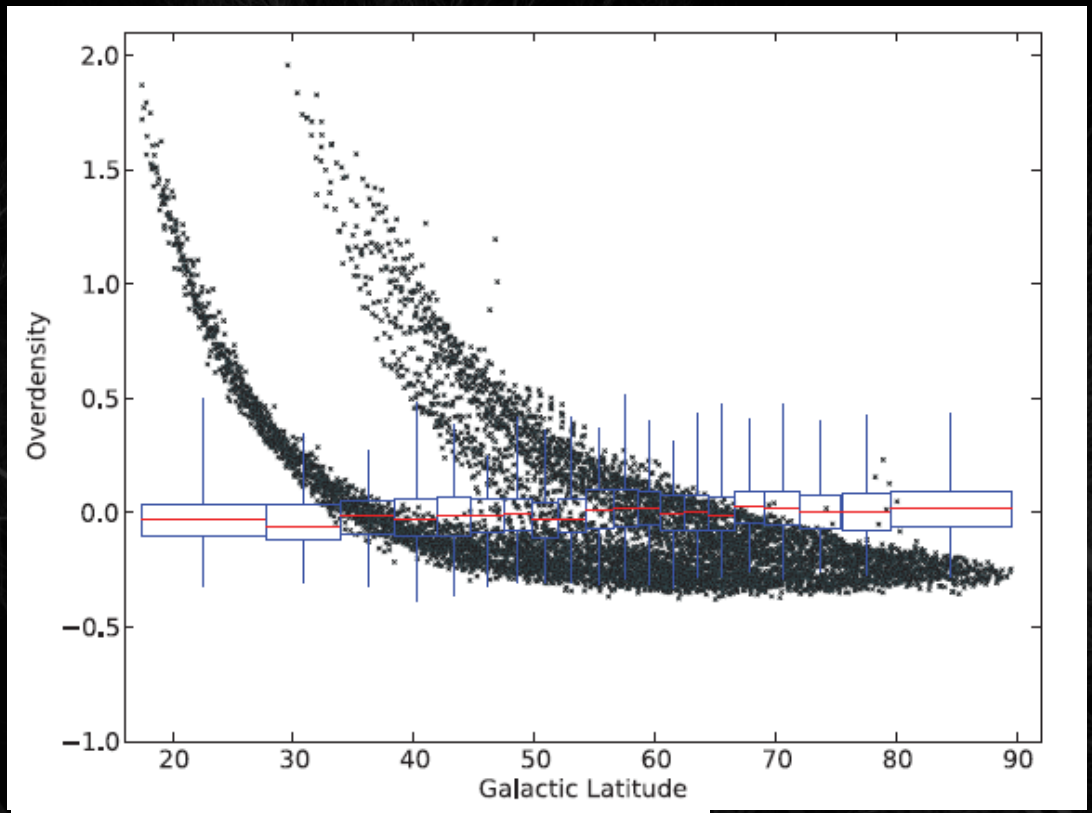
Systematics & LSS: Star/Galaxy separation



Challenging for fainter magnitudes



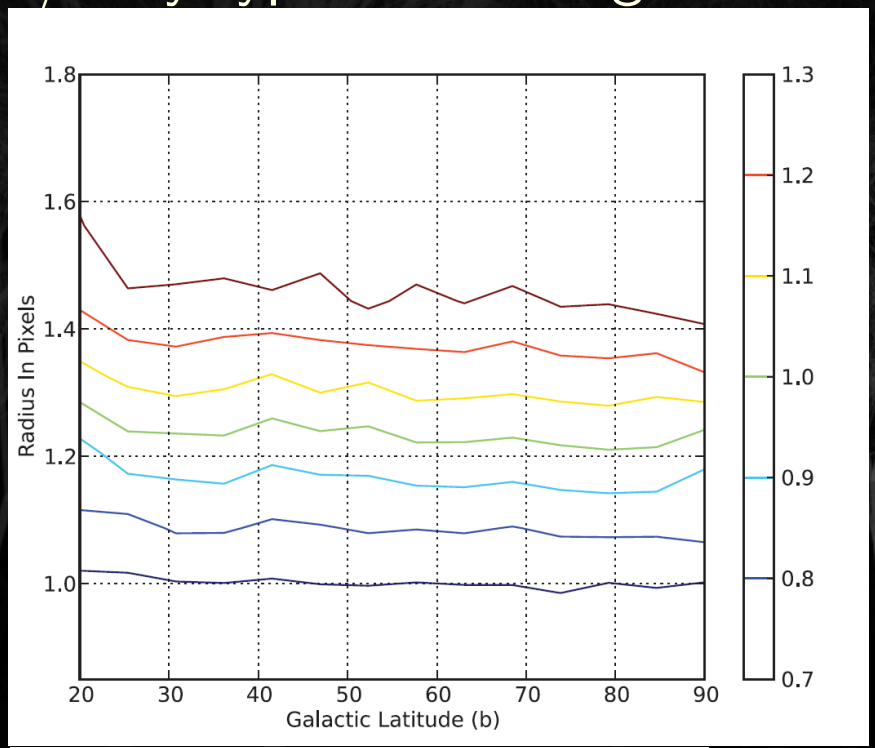
Systematics & LSS: Star/Galaxy separation in APS



Hayes, Brunner & Ross, 2012

Contamination of starts and gal. latiude

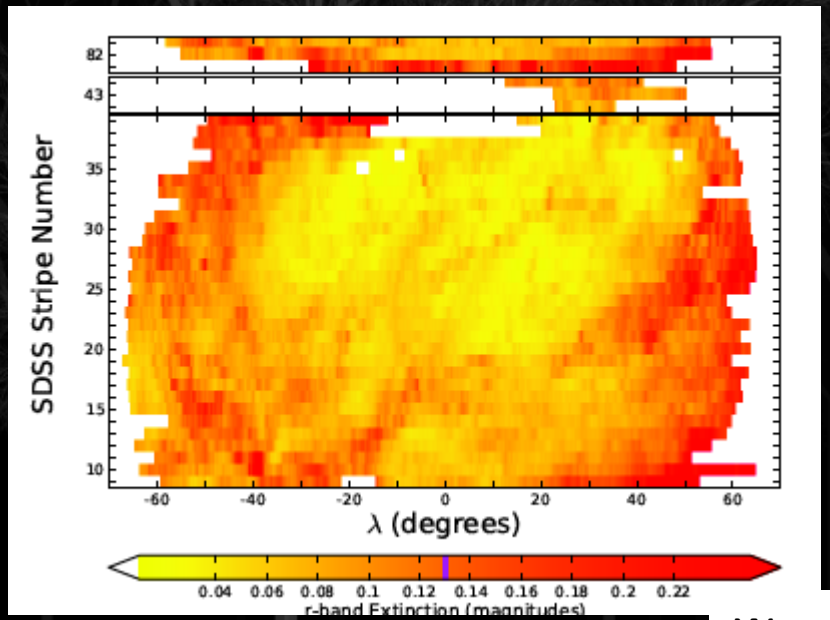
Late/early type ratio and gal. latiude

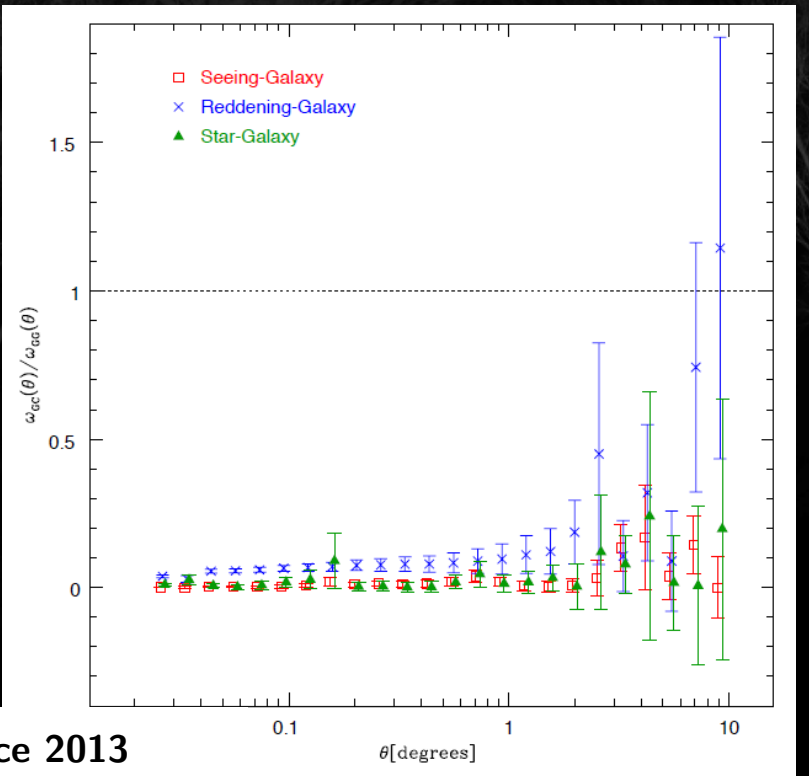


Hayes & Brunner, 2013

Systematics & LSS: ACF case

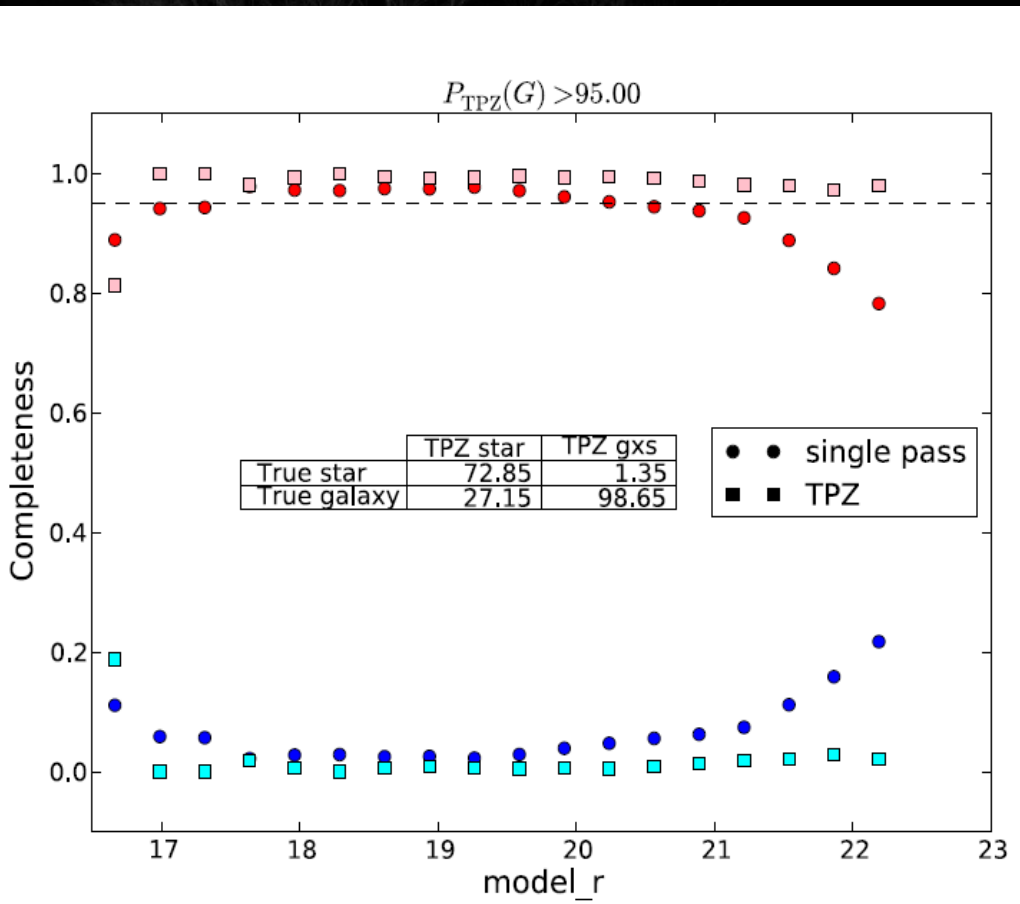
- S/G separation
- Pixelisation
- Density fluctuations in stripes
- Seeing variation
- Reddening variation
- Flag variation

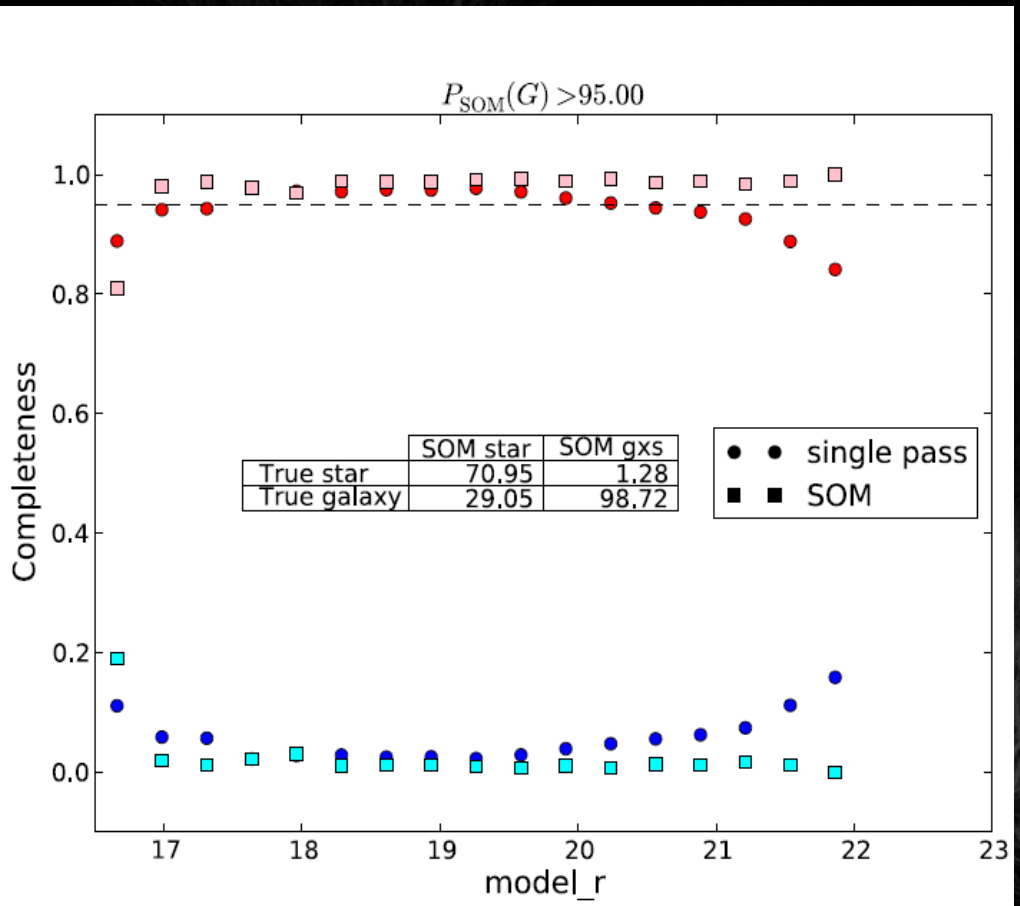




Wang, Brunner & Dolence 2013

Systematics & LSS: Star/Galaxy separation using ML

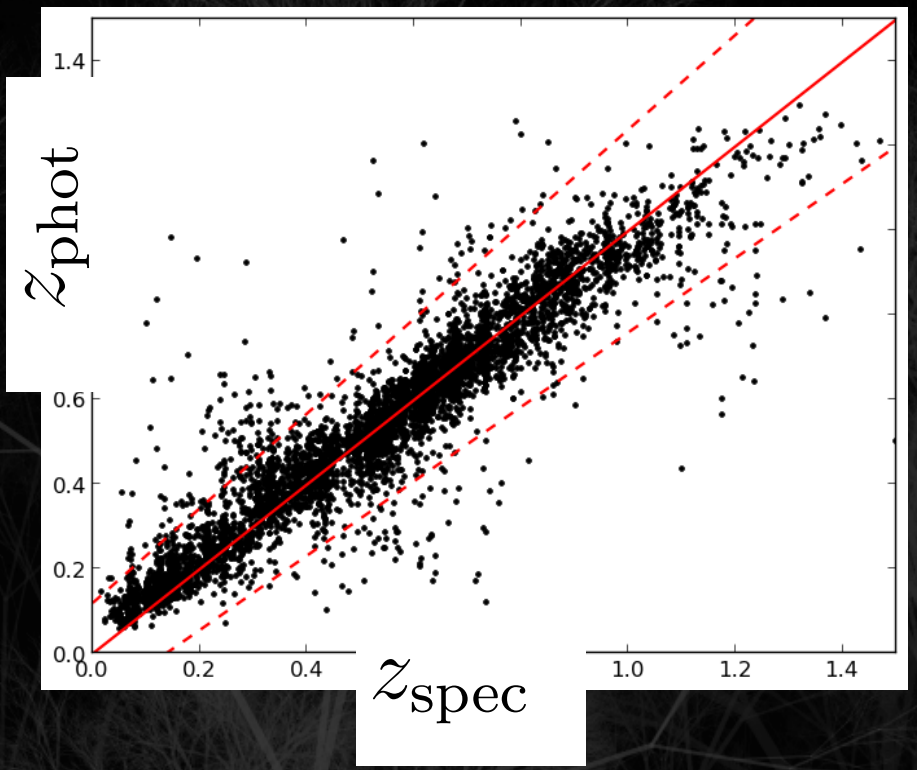


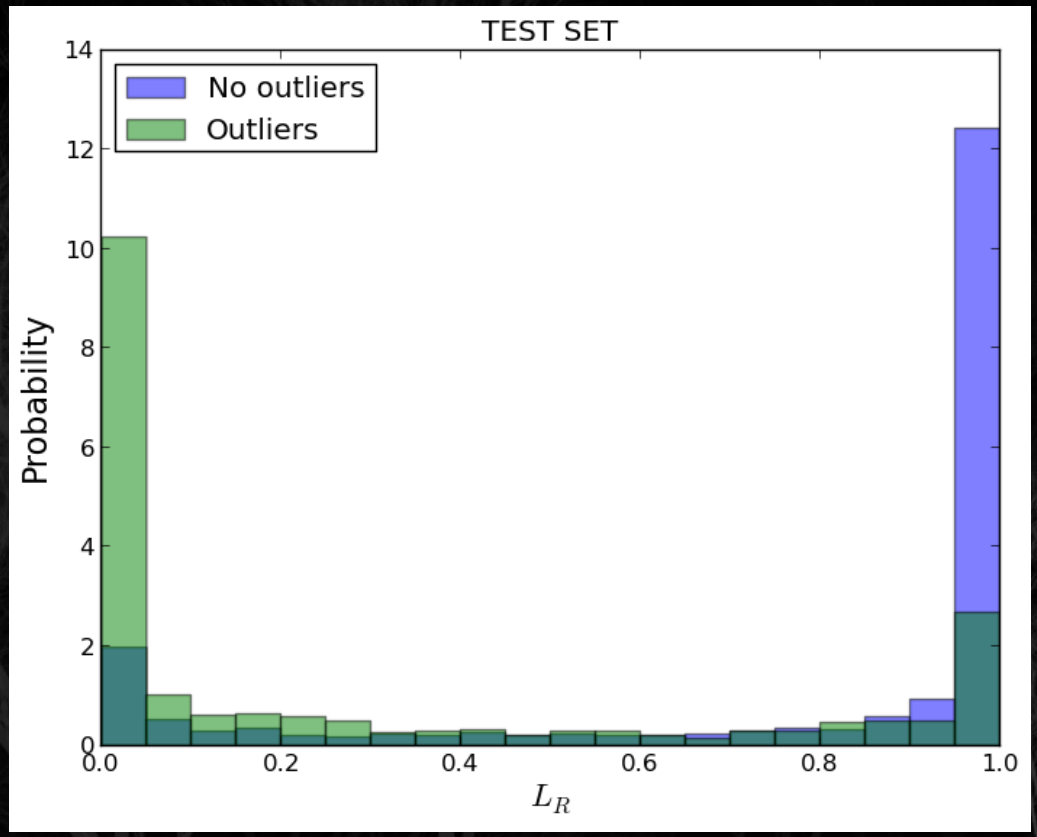


TPZ and SOM provide probability for being a galaxy

Compare with coadd stripe 82 classification

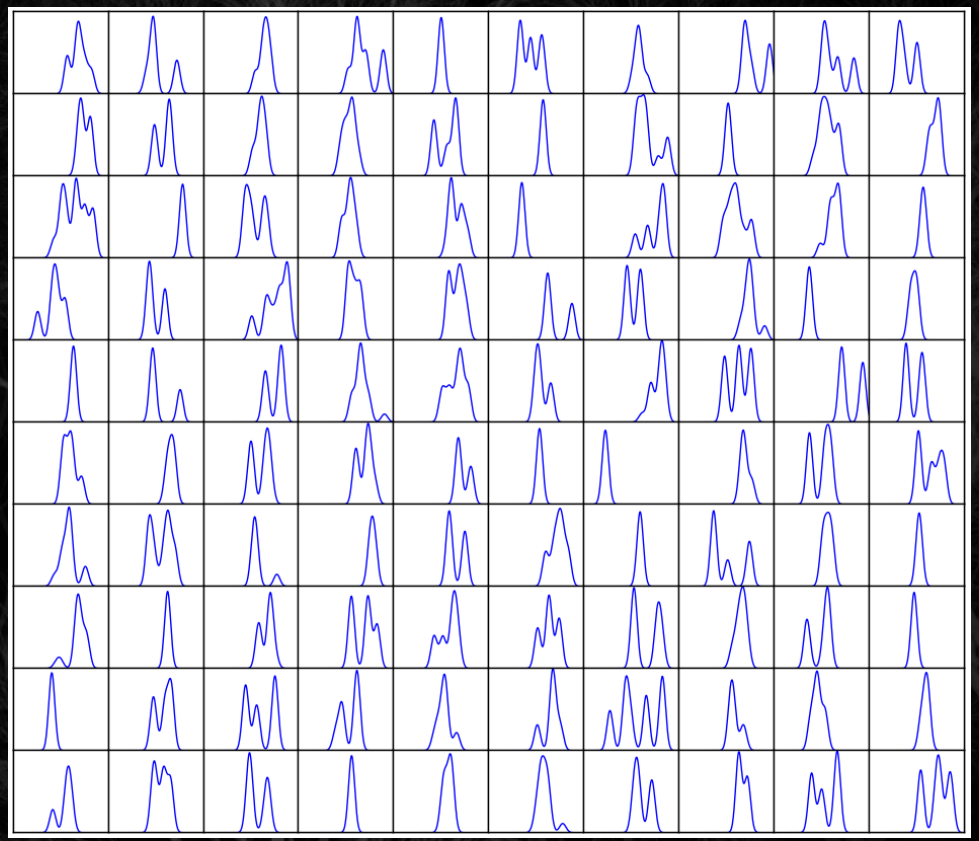
Systematics & LSS: Photo-z outliers





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

Systematics: Photo-z PDF storage

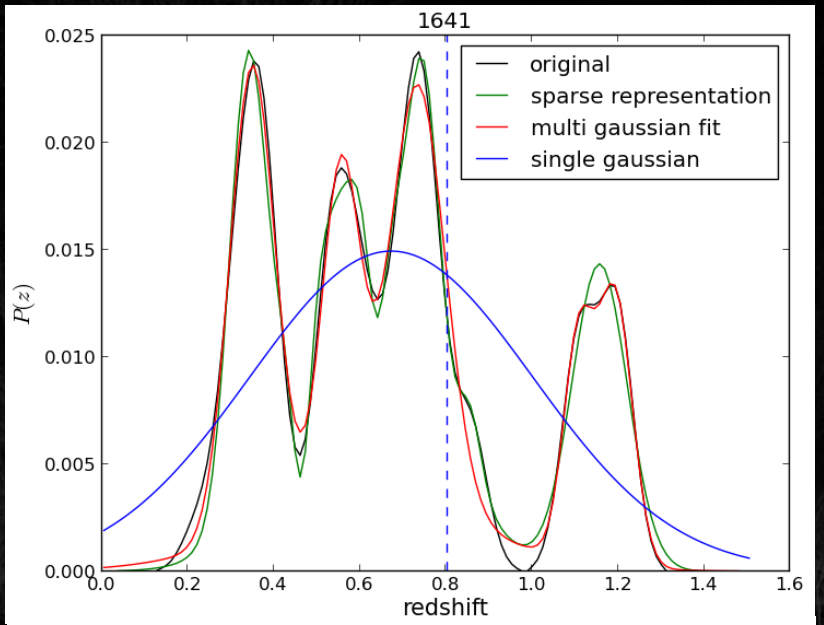


Systematics: Photo-z PDF storage

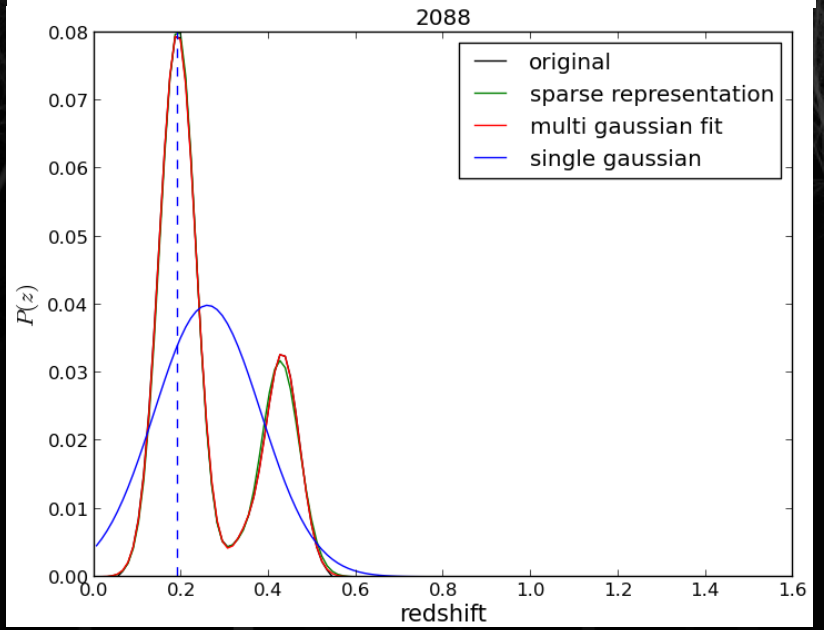
Multi-Gaussan fit

Sparse representation techniques

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

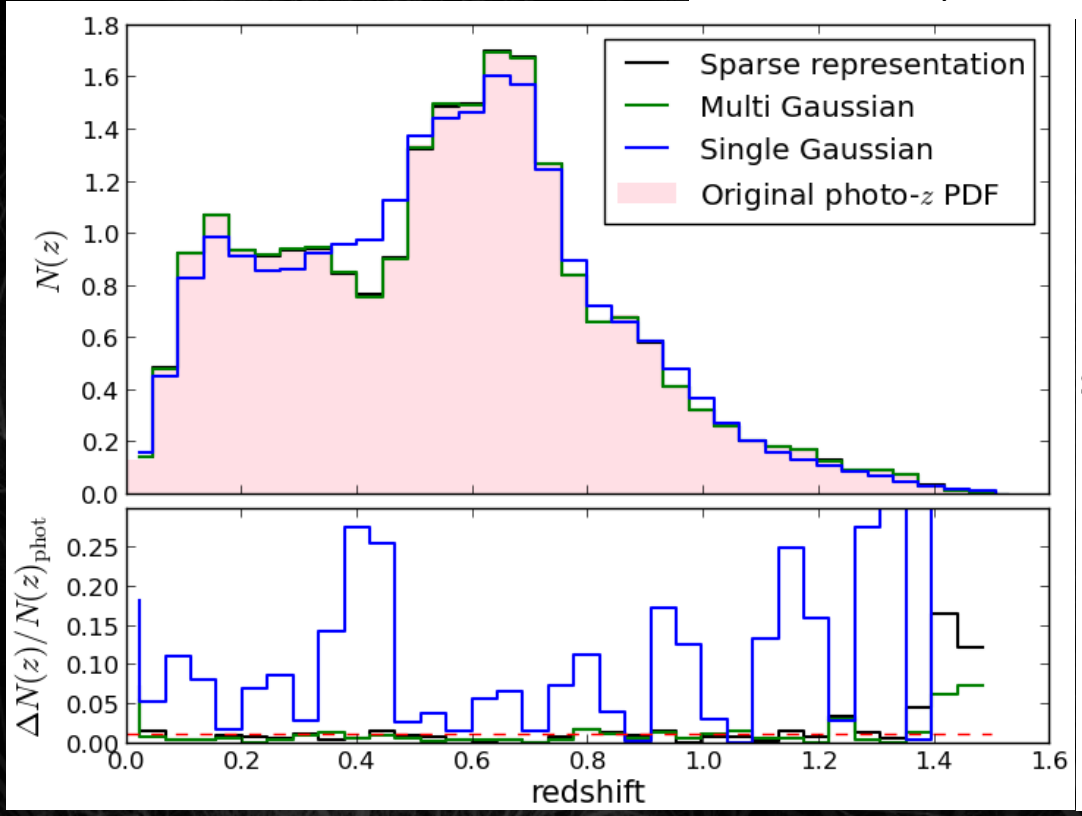


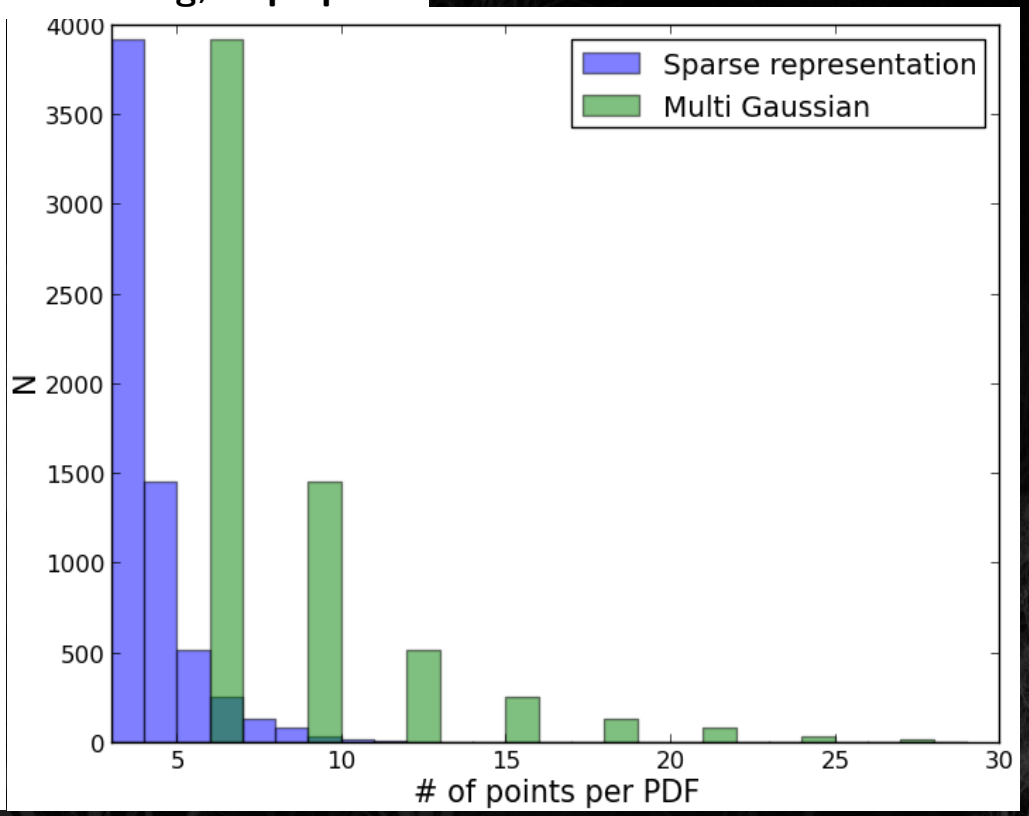
Carrasco Kind, Brunner & Ching, in prep.



Systematics: Photo-z PDF storage

Carrasco Kind, Brunner & Ching, in prep.





Differences less than 1% using Multi Gaussian or sparse representation

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

Conclusions

* Machine leraning powerful tool

*Ensemble and deep learning even better

* Systematics are important but can be addressed

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

EXTRA SLIDES

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

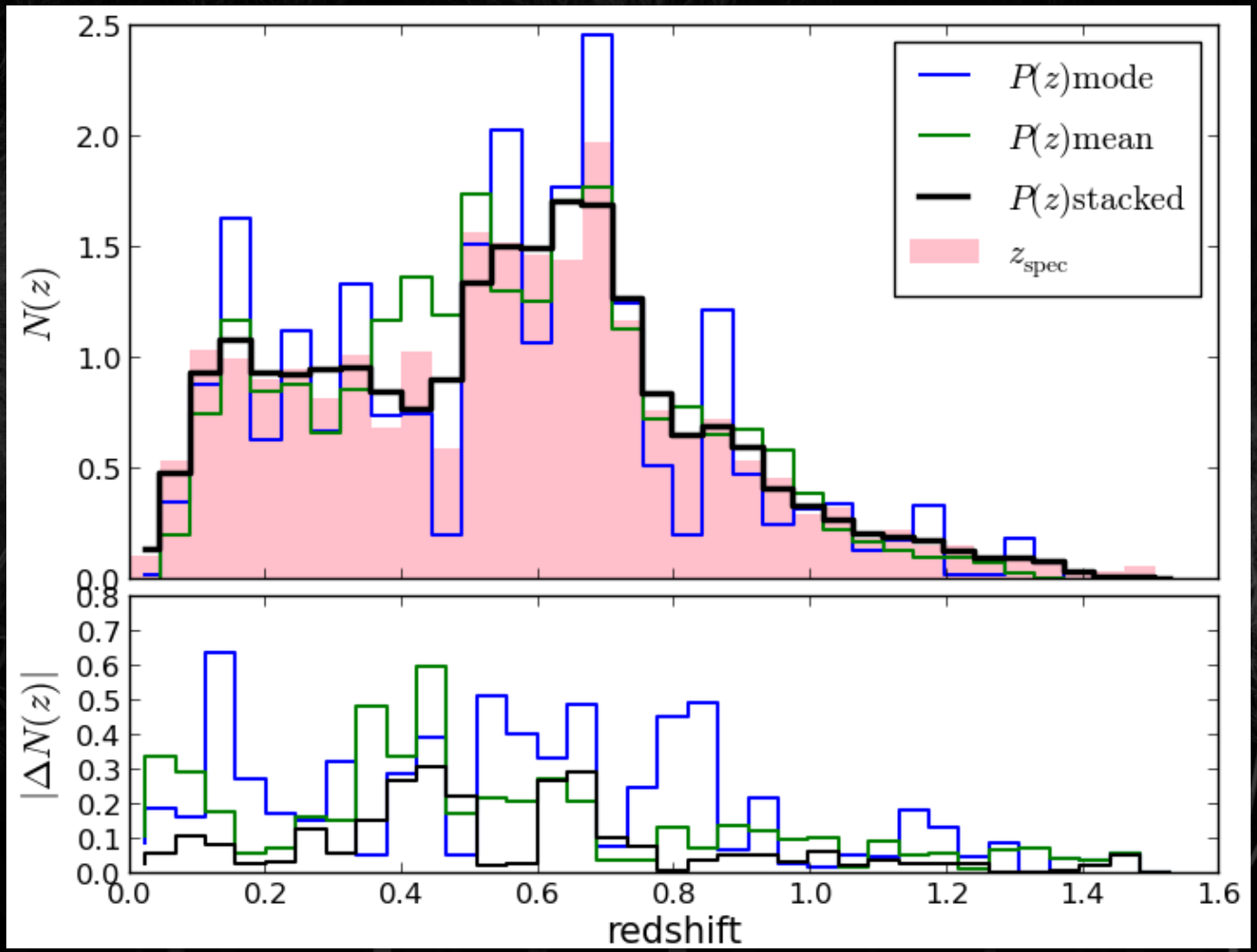


Photo-z PDF application: $N(z)$

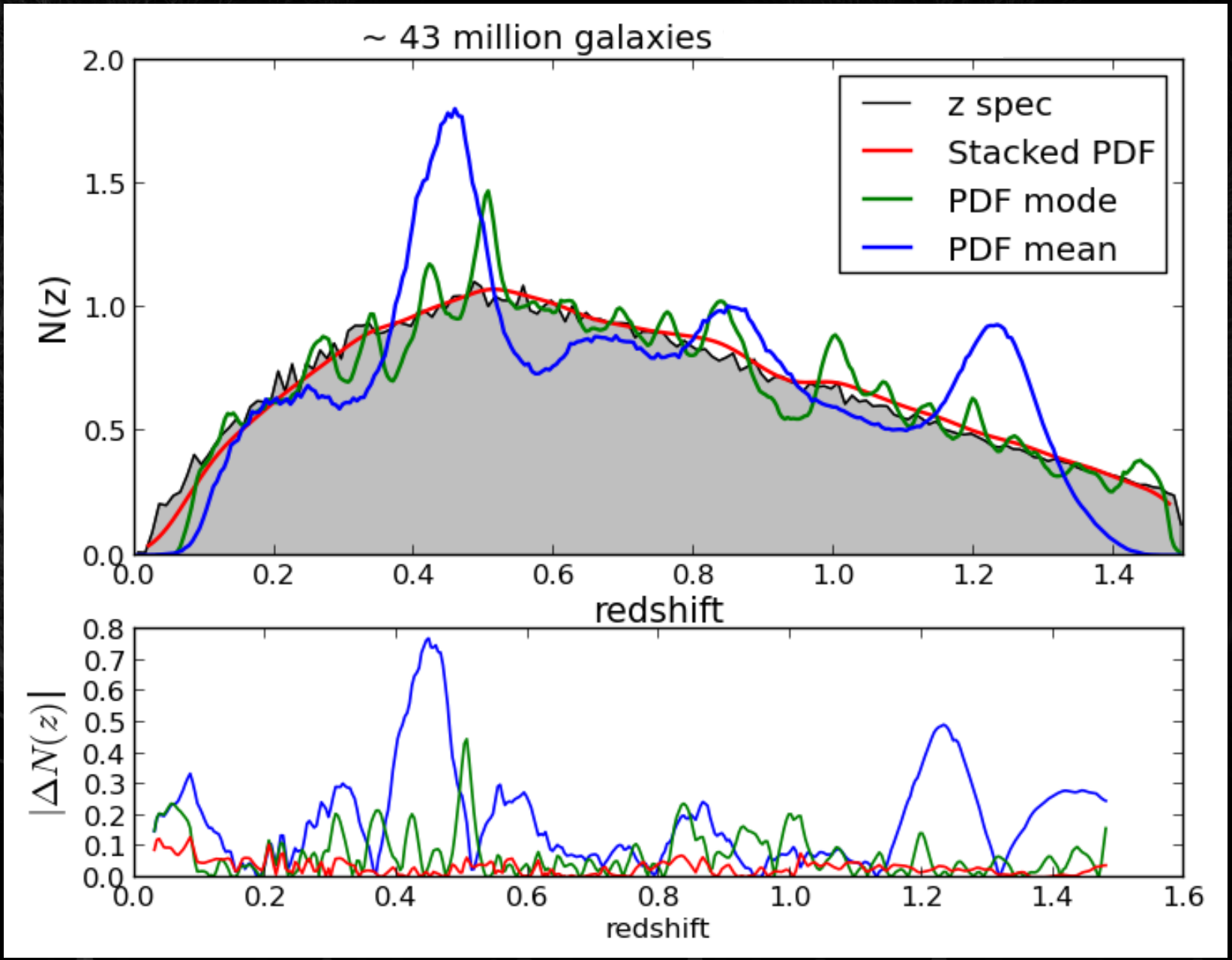


Photo-z PDF application: Angular Power Spectrum

- 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

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

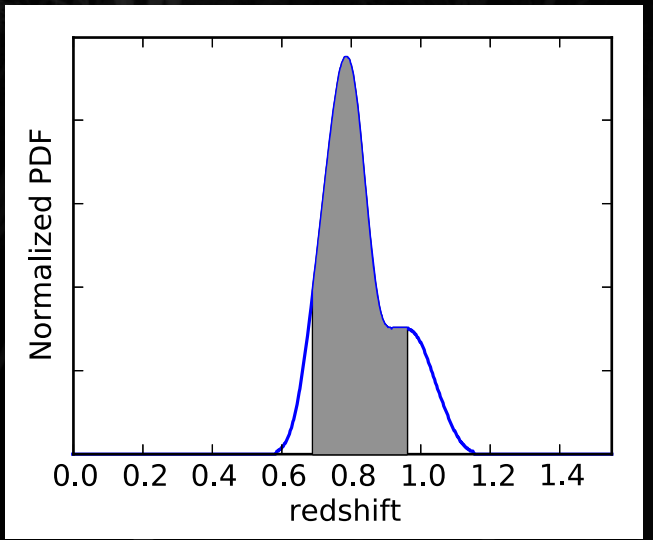


Photo-z PDF application: C_ℓ and $\omega(\theta)$

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

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

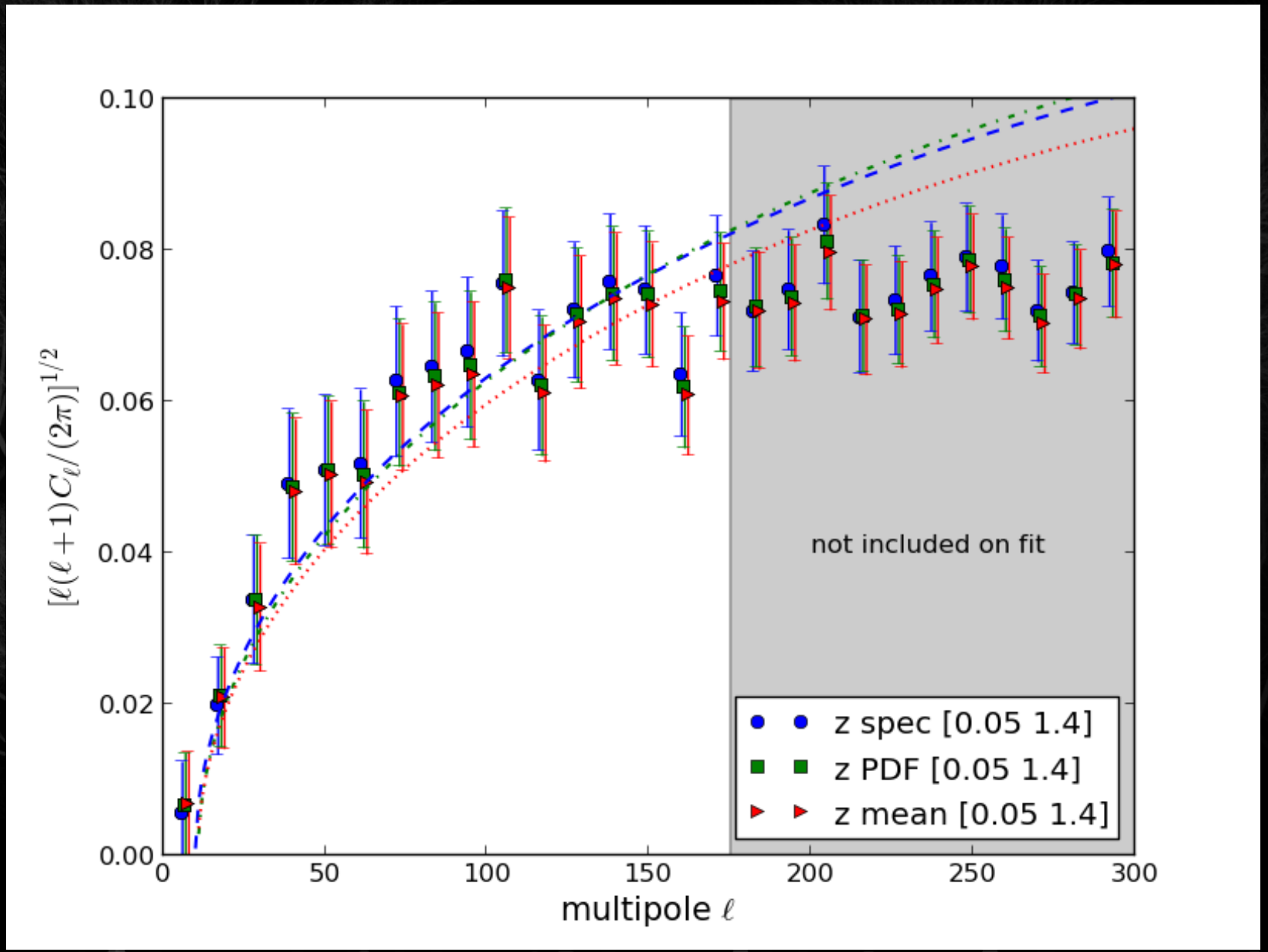
CAMB and HALOFIT for non linear $P(k,z)$

$\phi(z)$ 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_ℓ and $\omega(\theta)$

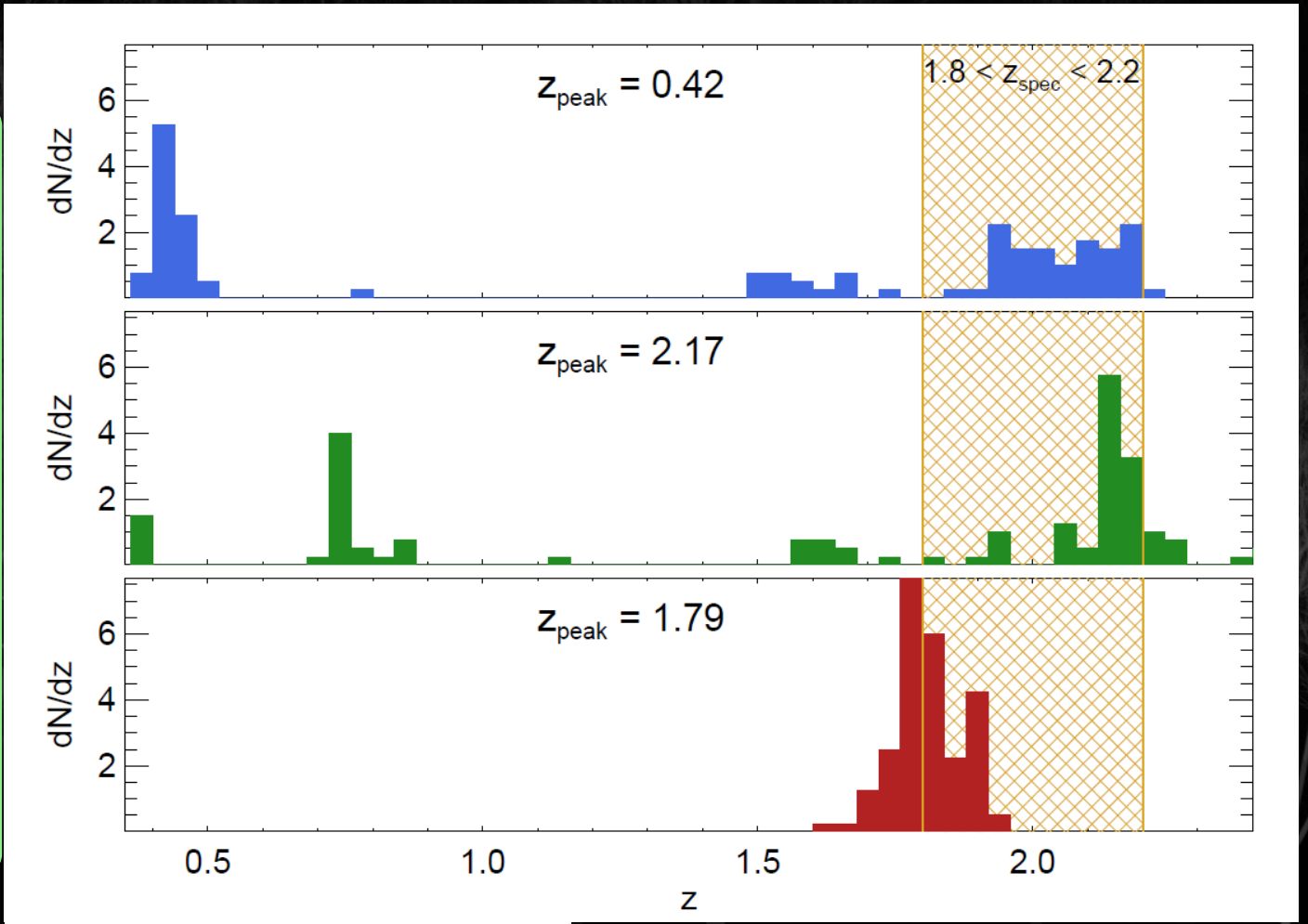


Example application of photo-z PDF

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.

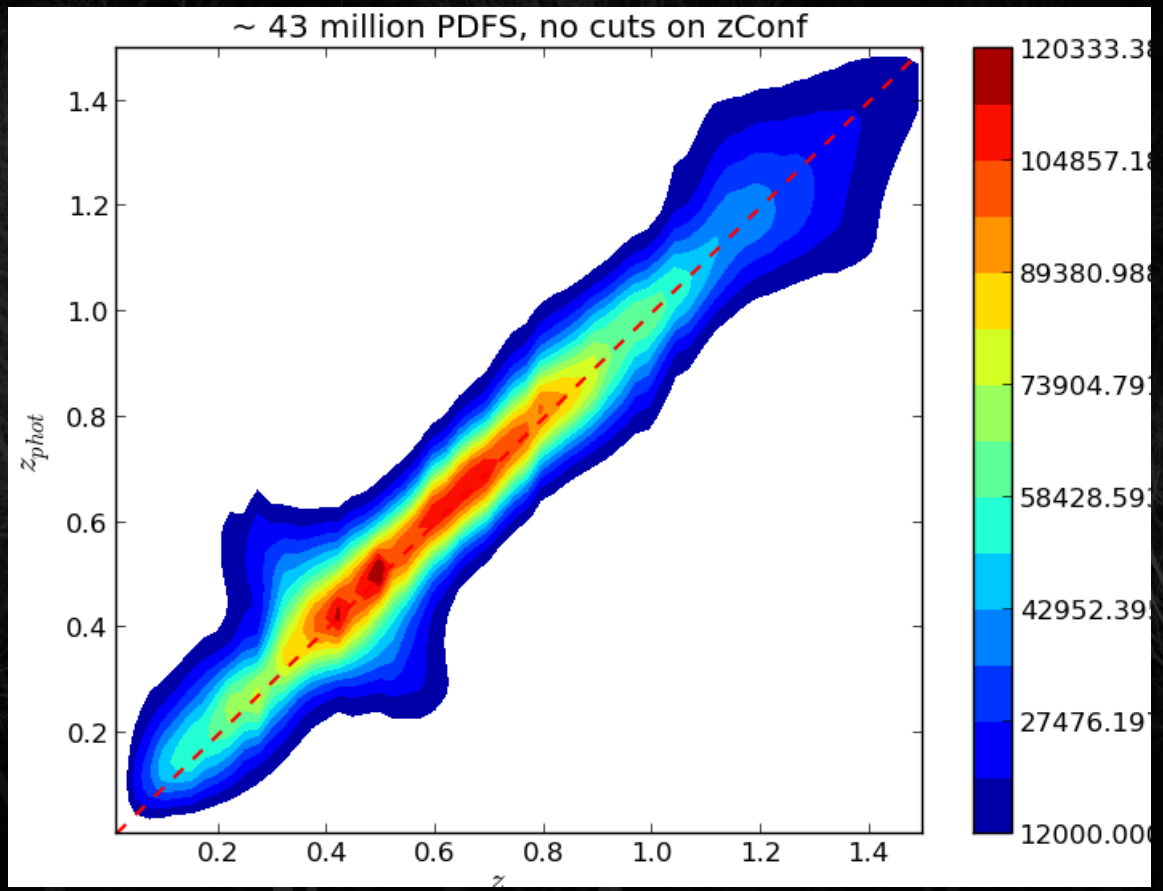
100,00 for training

5 magnitudes only

$\sim$ 0.17 sec per PDF

Store 43 million PDFs for analysis

No outlier removal



Photometric redshift PDFs using TPZ

Metrics

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

$< \Delta z >= 0.0088$

$< |\Delta z| >= 0.089$

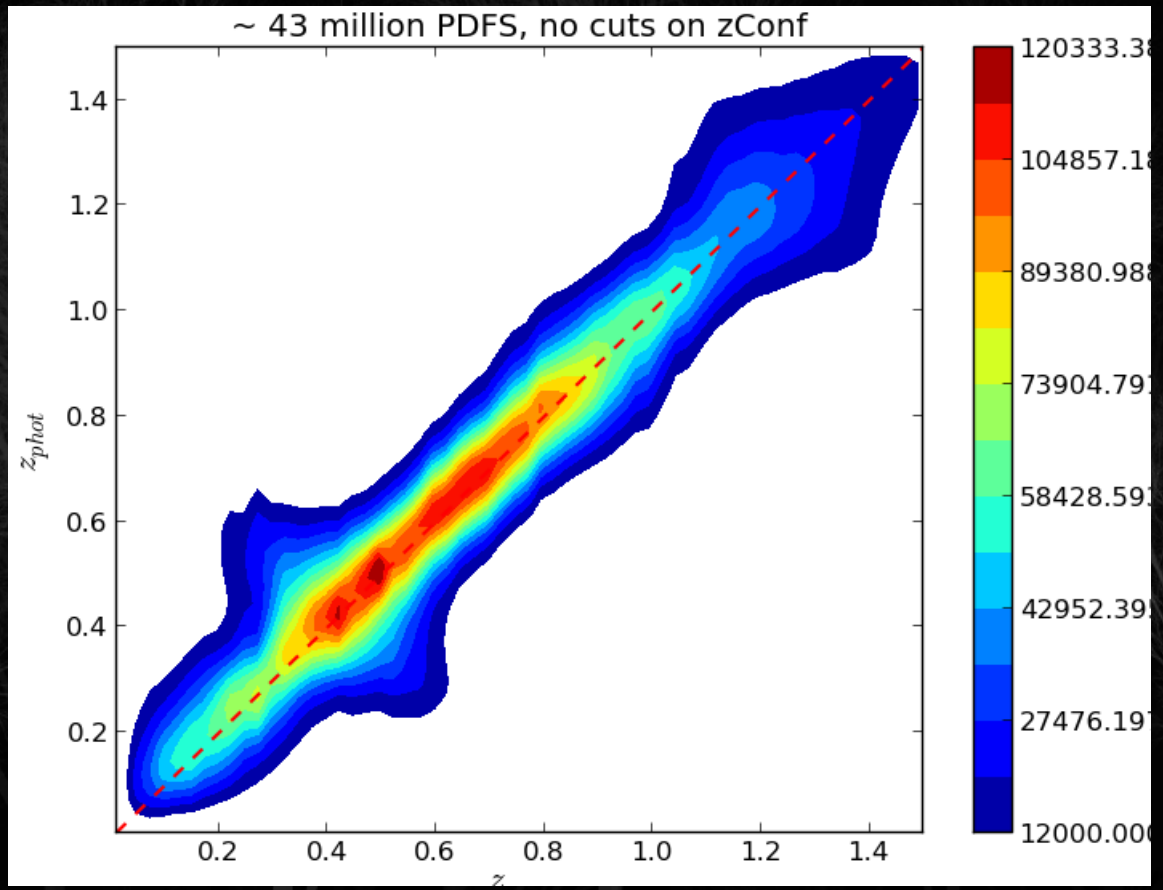
$\sigma_{\Delta z} = 0.1421$

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

$\sigma_{68} = 0.0885$

$frac > 2\sigma = 0.0531$

$frac > 3\sigma = 0.0207$



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