ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Tales of photo-z, DBs and DES

Matías Carrasco Kind

NCSA/Department of Astronomy

University of Illinois at Urbana-Champaign

CEFCA Seminar October 23 $^{\text{th}}$, 2016

Motivation

- Photo-z Probability Density Functions needed
- Several methods / codes to compute photo-z
- Need for a meta-algorithm that combines multiple techniques
- PDFs are good **but** for large datasets, storage and I/O will be an issue
- Machine Learning and statistical tools

Motivation

- Photo-z Probability Density Functions needed
- Several methods/codes to compute photo-z
- Need for a meta-algorithm that combines multiple techniques
- PDFs are good **but** for large datasets, storage and I/O will be an issue
- Machine Learning and statistical tools

Motivation

- Photo-z Probability Density Functions needed
- Several methods/codes to compute photo-z
- Need for a meta-algorithm that combines multiple techniques
- PDFs are good **but** for large datasets, storage and I/O will be an issue
- Machine Learning and statistical tools

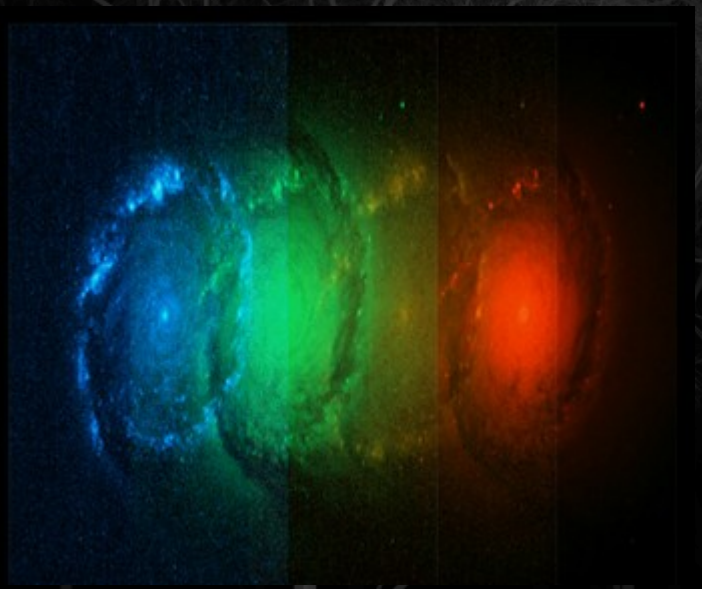
Motivation

- Photo-z Probability Density Functions needed
- 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 will be an issue
- Machine Learning and statistical tools

Motivation

- Photo-z Probability Density Functions needed
- 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 will be an issue
- Machine Learning and statistical tools

Photo-z PDF estimation (in 5 min.)



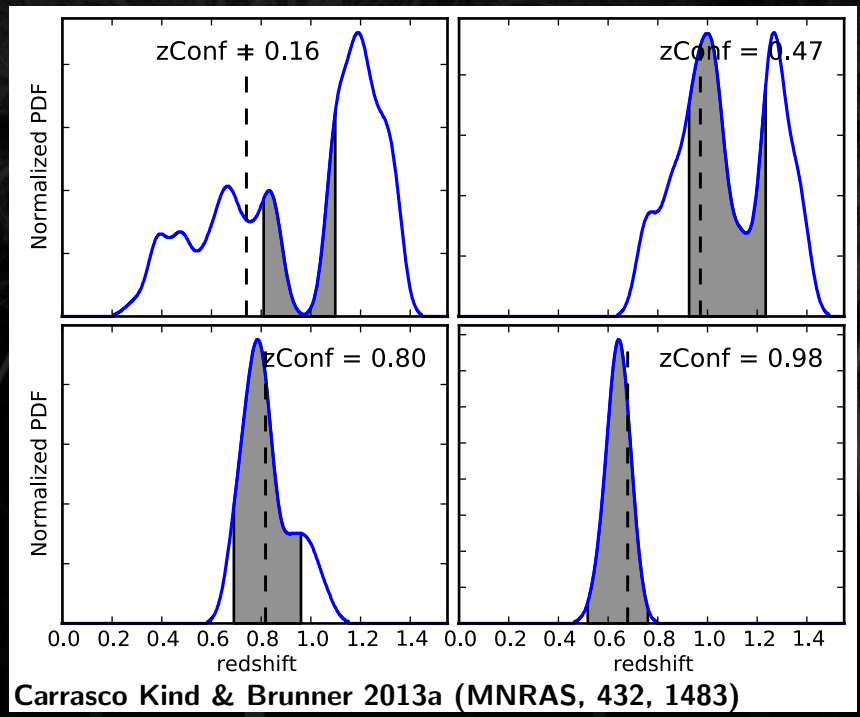
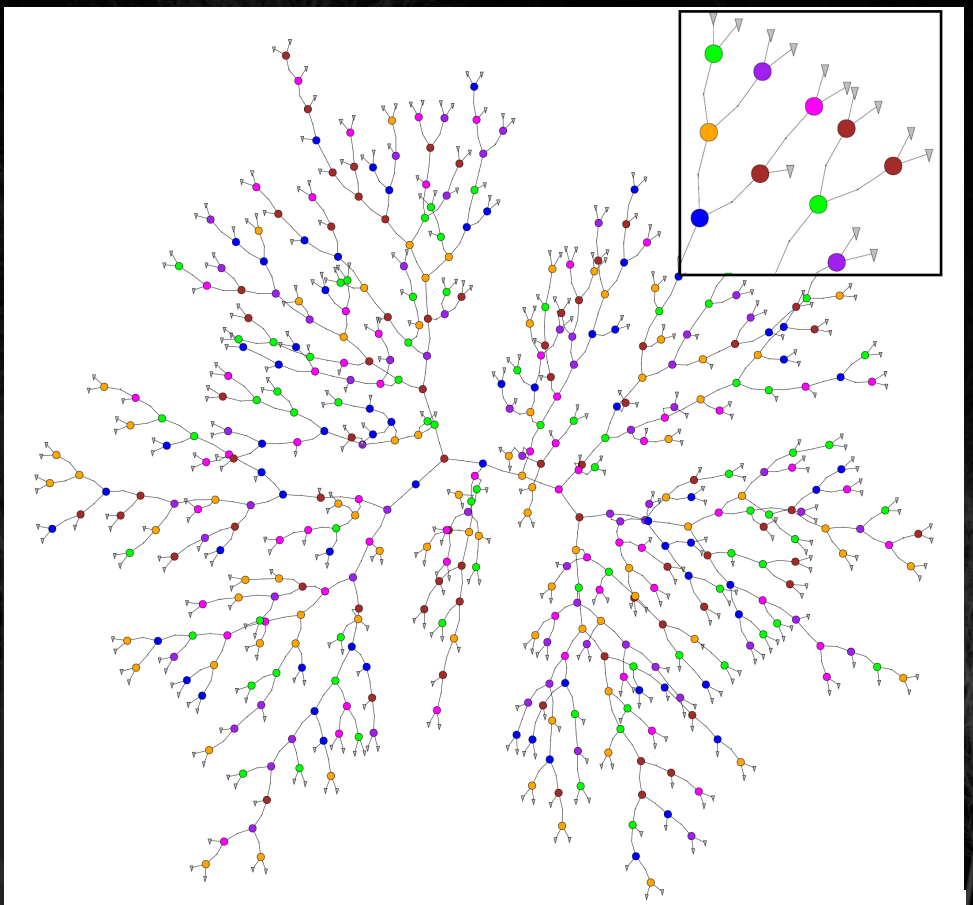


Photo-z PDF estimation: TPZ

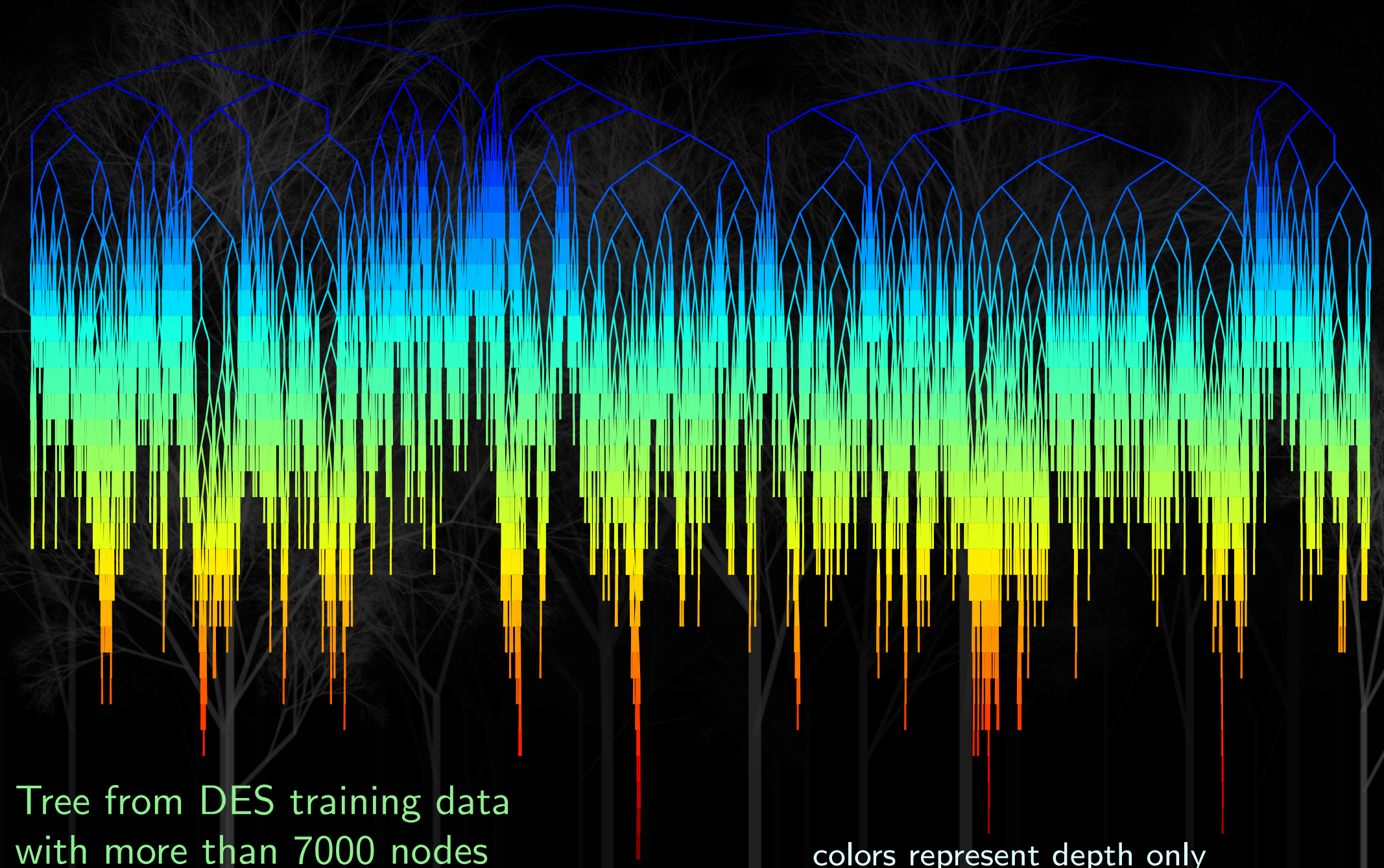
- TPZ (Trees for Photo-Z) is a supervised machine learning code
- Prediction trees and random forest
- Incorporate measurements errors and deals with missing values
- Ancillary information: expected errors, attribute ranking and others
- Application to the S/G



Carrasco Kind & Brunner 2013a (MNRAS, 432, 1483)

http://lcdm.astro.illinois.edu/code/mlz.html

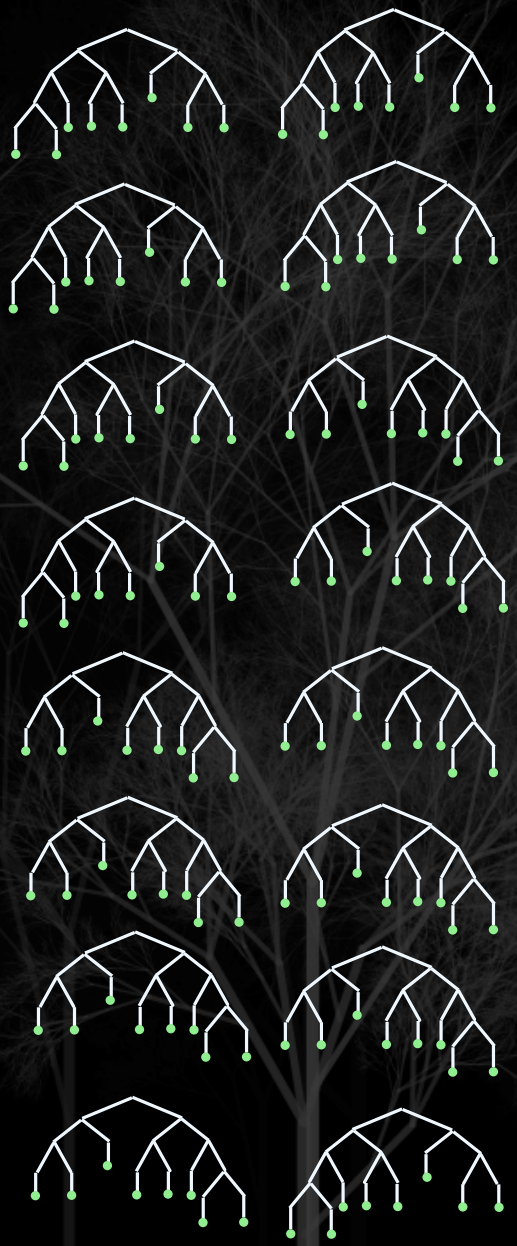
Photo-z PDF estimation: TPZ example



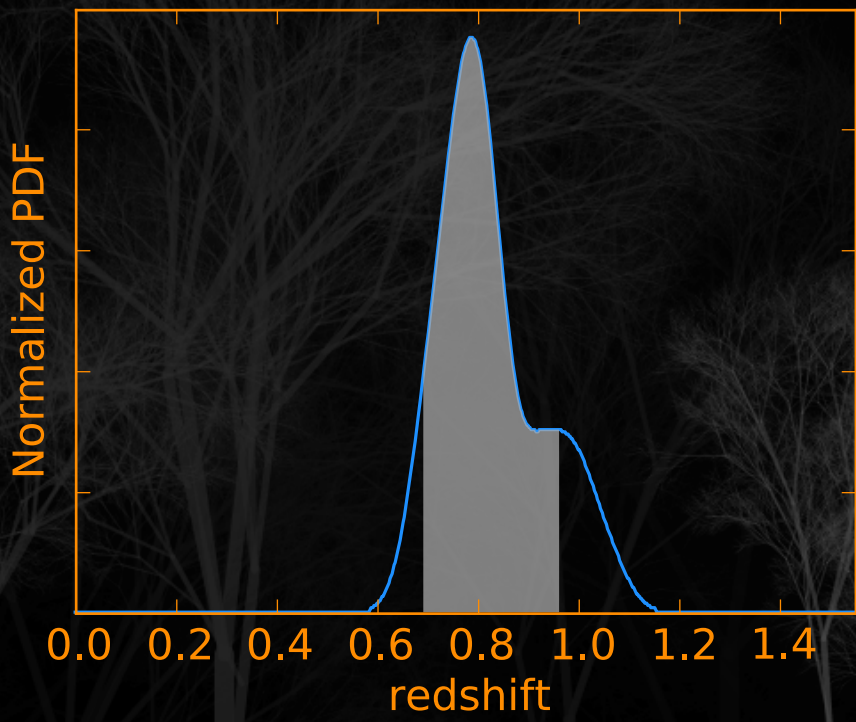
Tree from DES training data
with more than 7000 nodes

colors represent depth only

Photo-z PDF estimation: Random forest



Combine predictions from trees

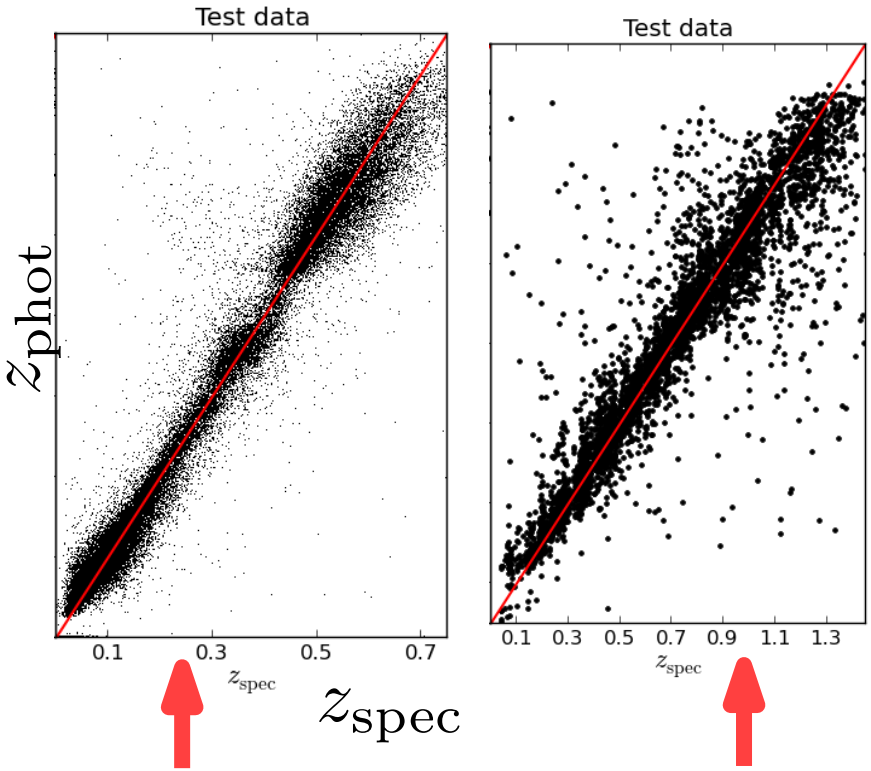


Trees are ideally uncorrelated and strong

Bootstrapping and error sampling

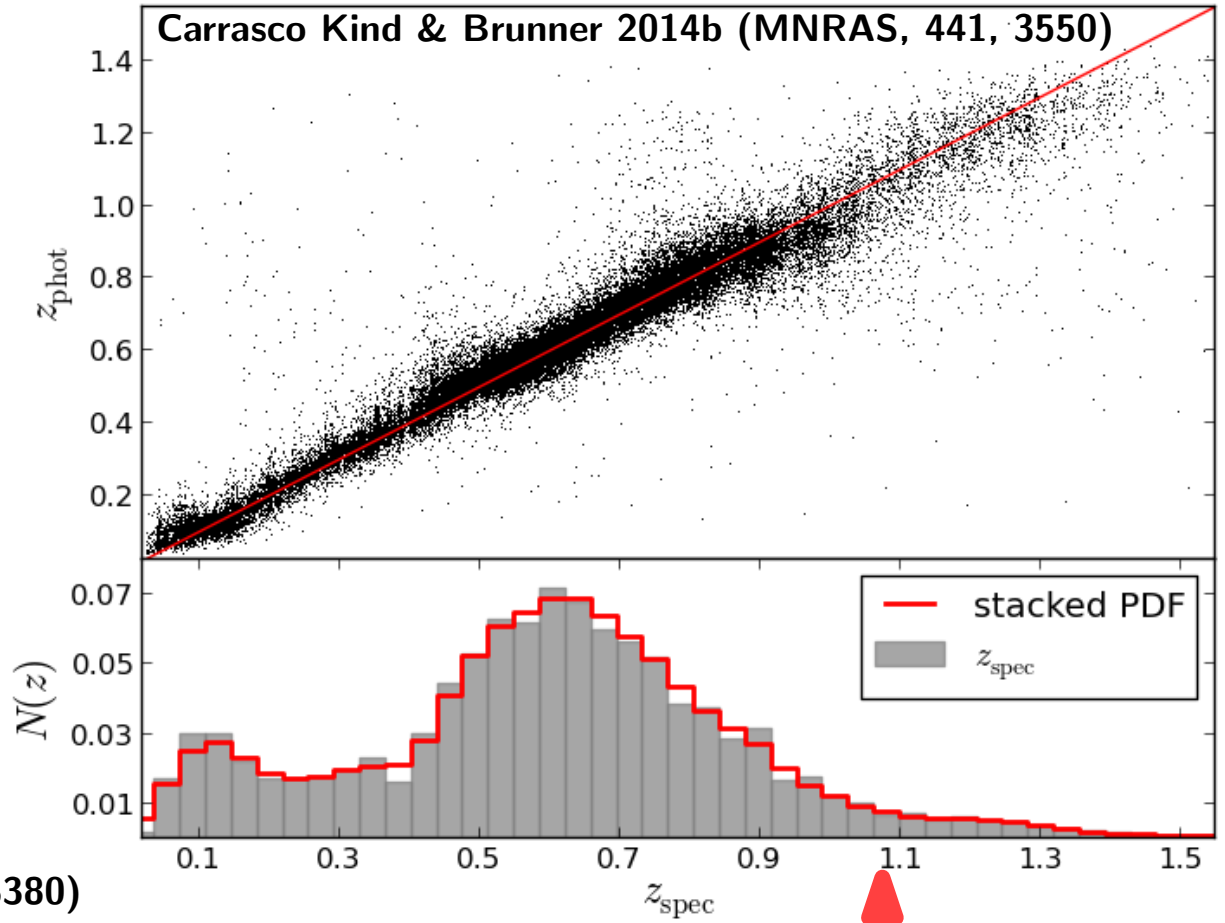
Random features at each node

Photo-z PDF estimation: TPZ applications



SDSS DR10 DEEP2

Carrasco Kind & Brunner 2014c (MNRAS, 442, 3380)

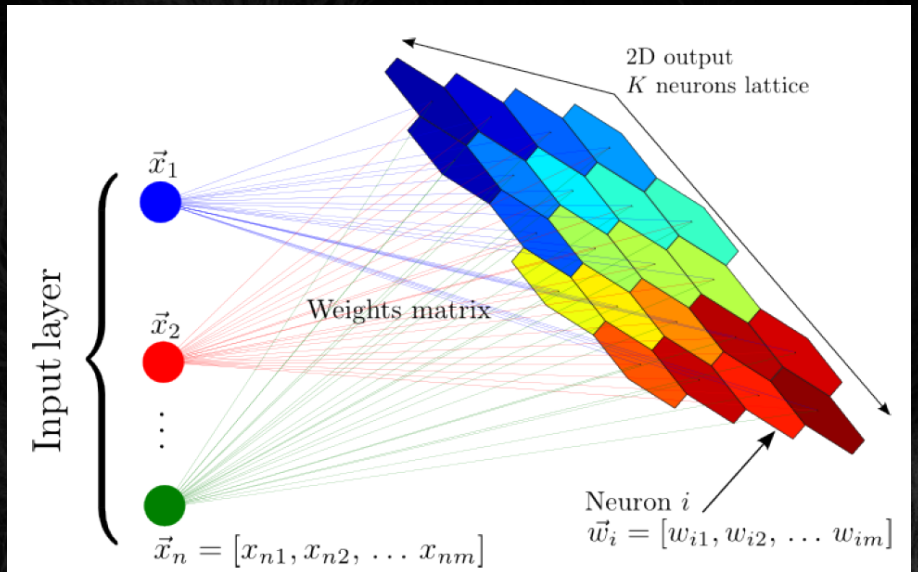


CFHTLenS

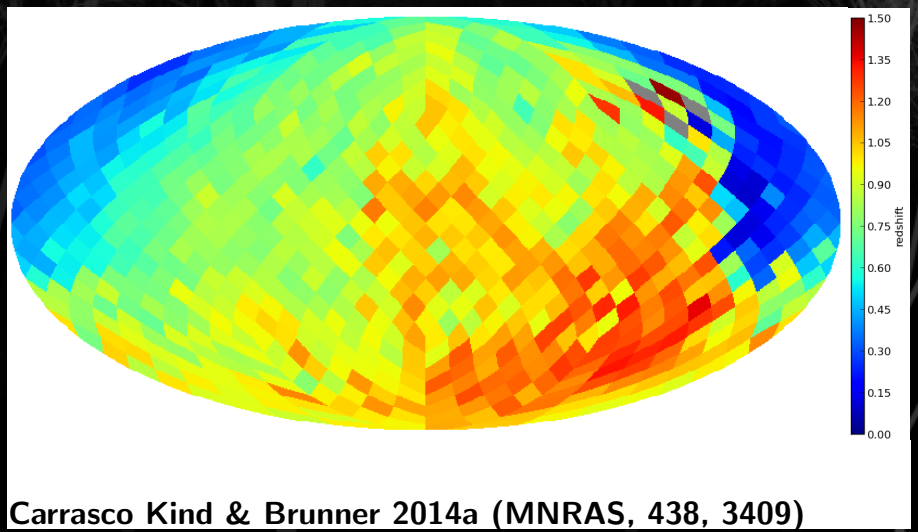
TPZ has been tested in several databases with remarkable results

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

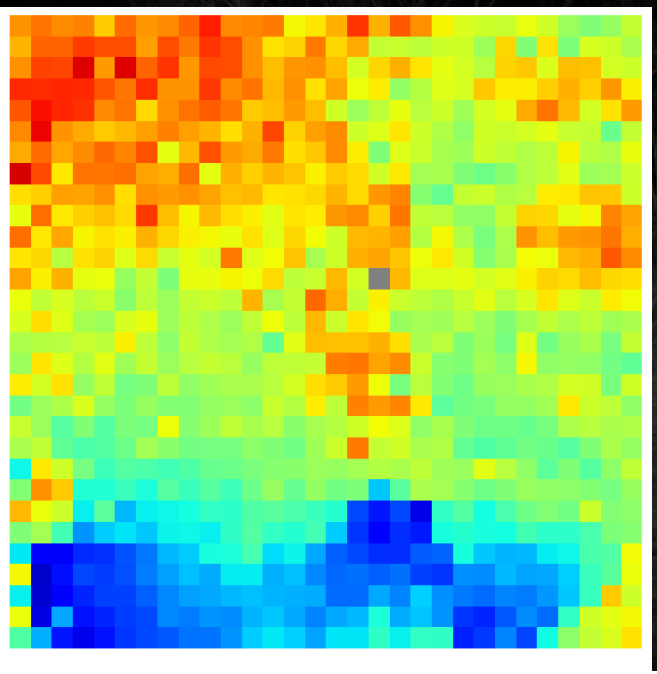
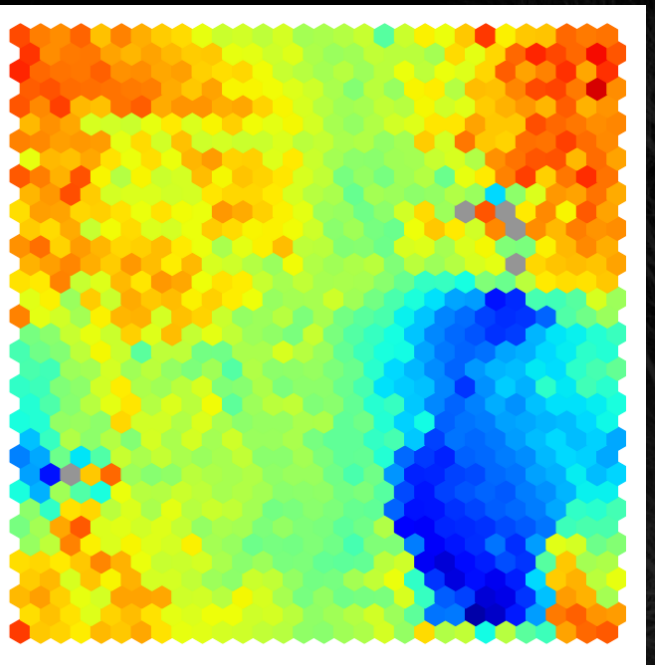
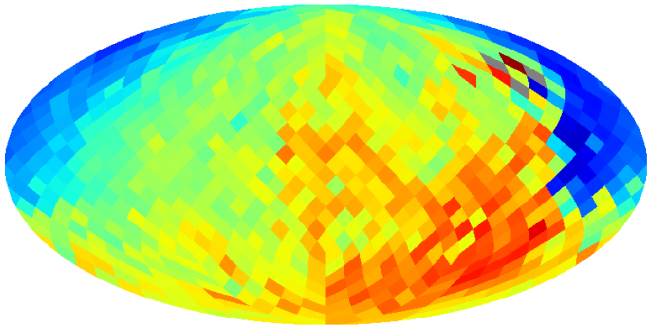


Carrasco Kind & Brunner 2014a (MNRAS, 438, 3409)



Carrasco Kind & Brunner 2014a (MNRAS, 438, 3409)

SOM topologies

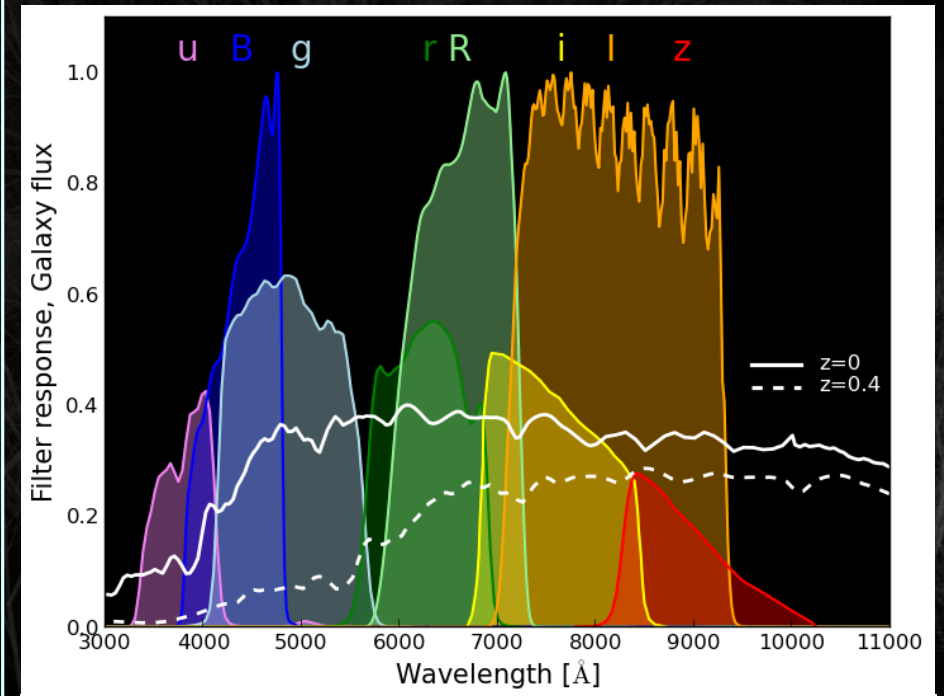




Carrasco Kind & Brunner 2014a (MNRAS, 438, 3409)

Different topologies can be used with or without periodic boundary conditions

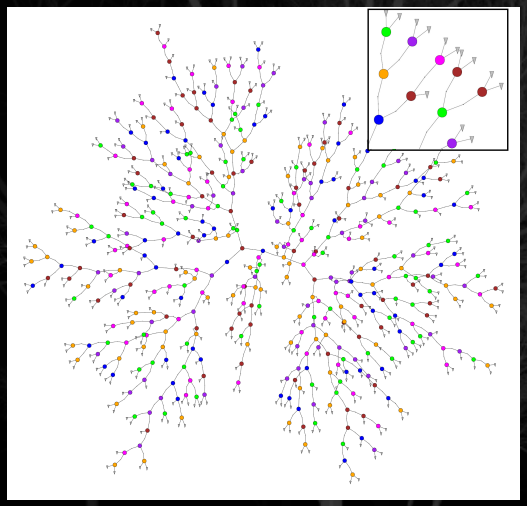
Photo-z PDF estimation: BPZ

- 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

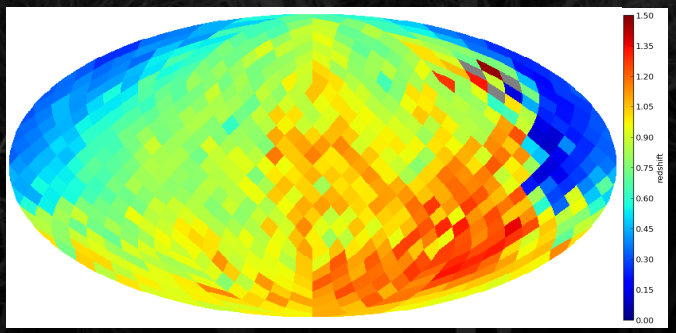


Carrasco Kind & Brunner 2014c (MNRAS, 442, 3380)

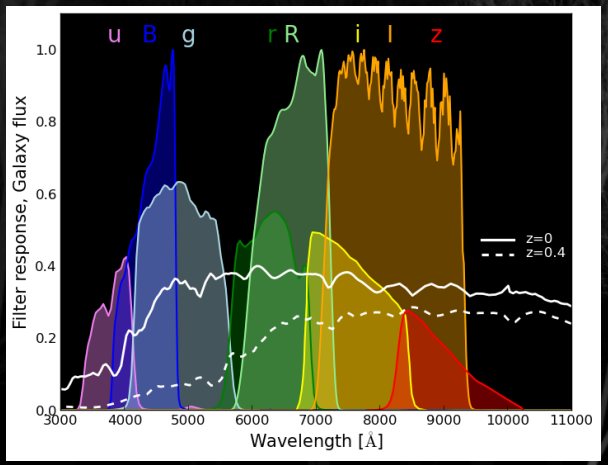
Photo-z PDF combination



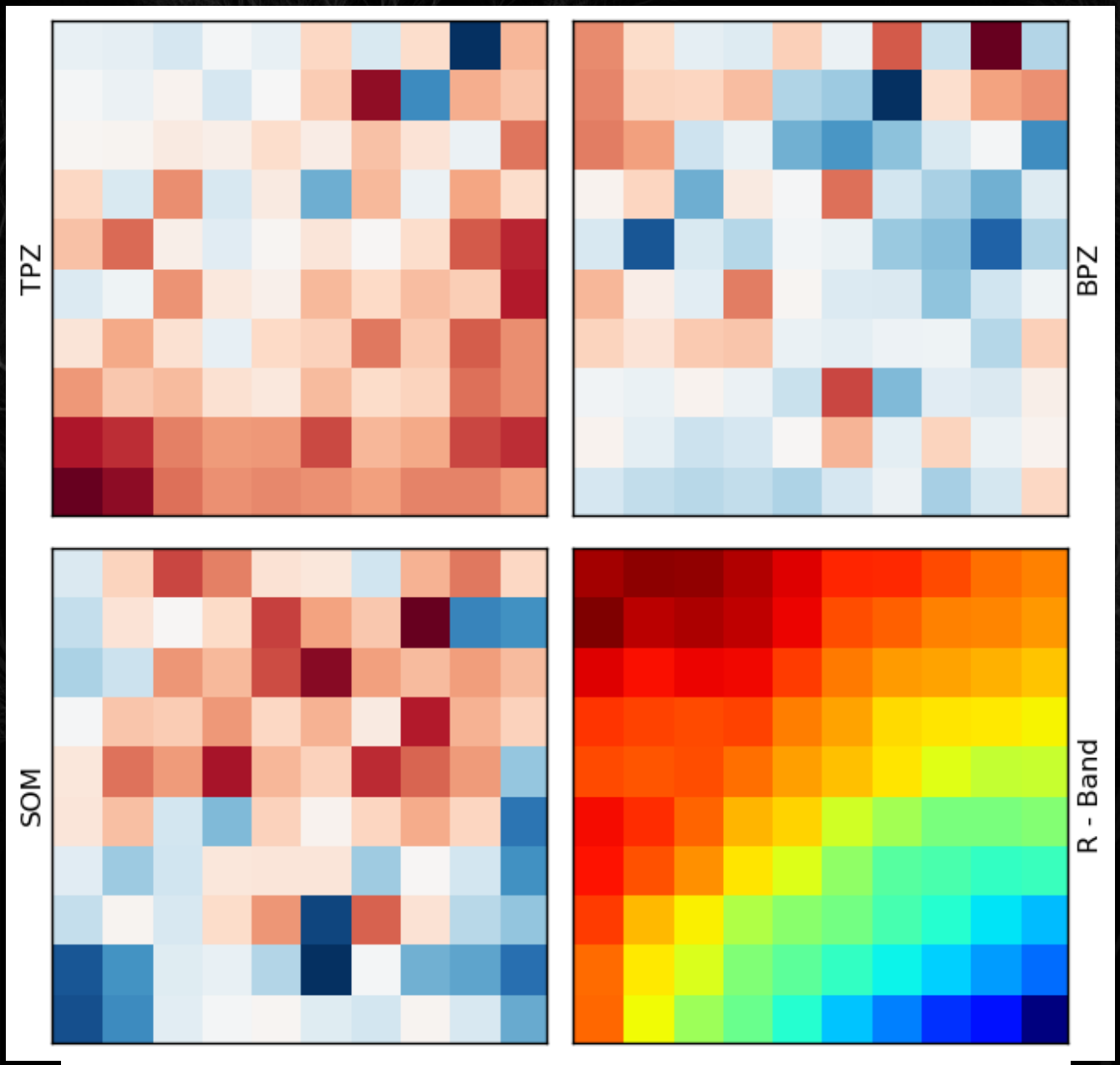
+



+

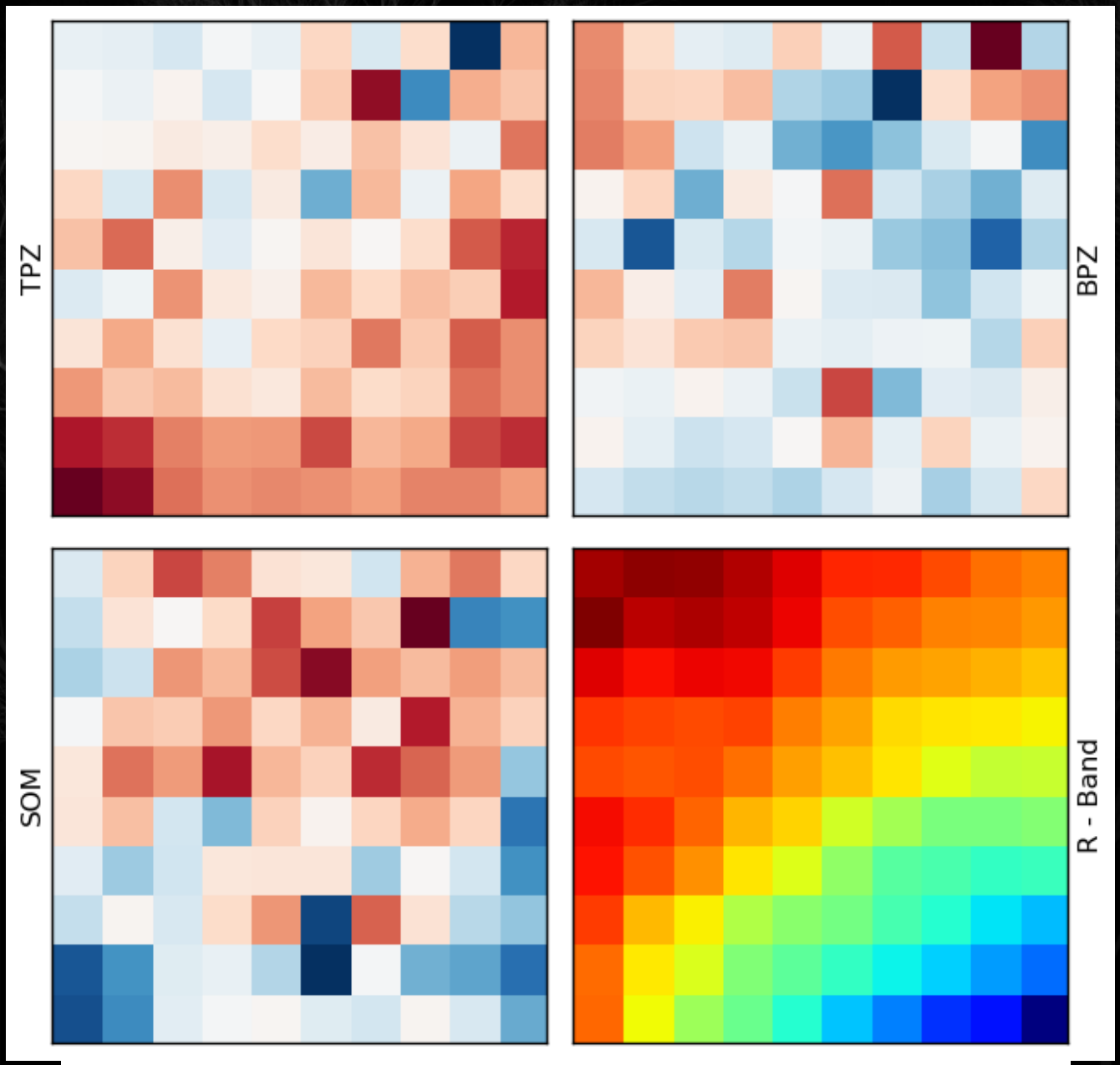


Bayesian framework



Carrasco Kind & Brunner 2014c (MNRAS, 442, 3380)

Bayesian framework



Carrasco Kind & Brunner 2014c (MNRAS, 442, 3380)

Our approach

Supervised method

+

Unsupervised method

+

Template fitting

+

Weighting scheme

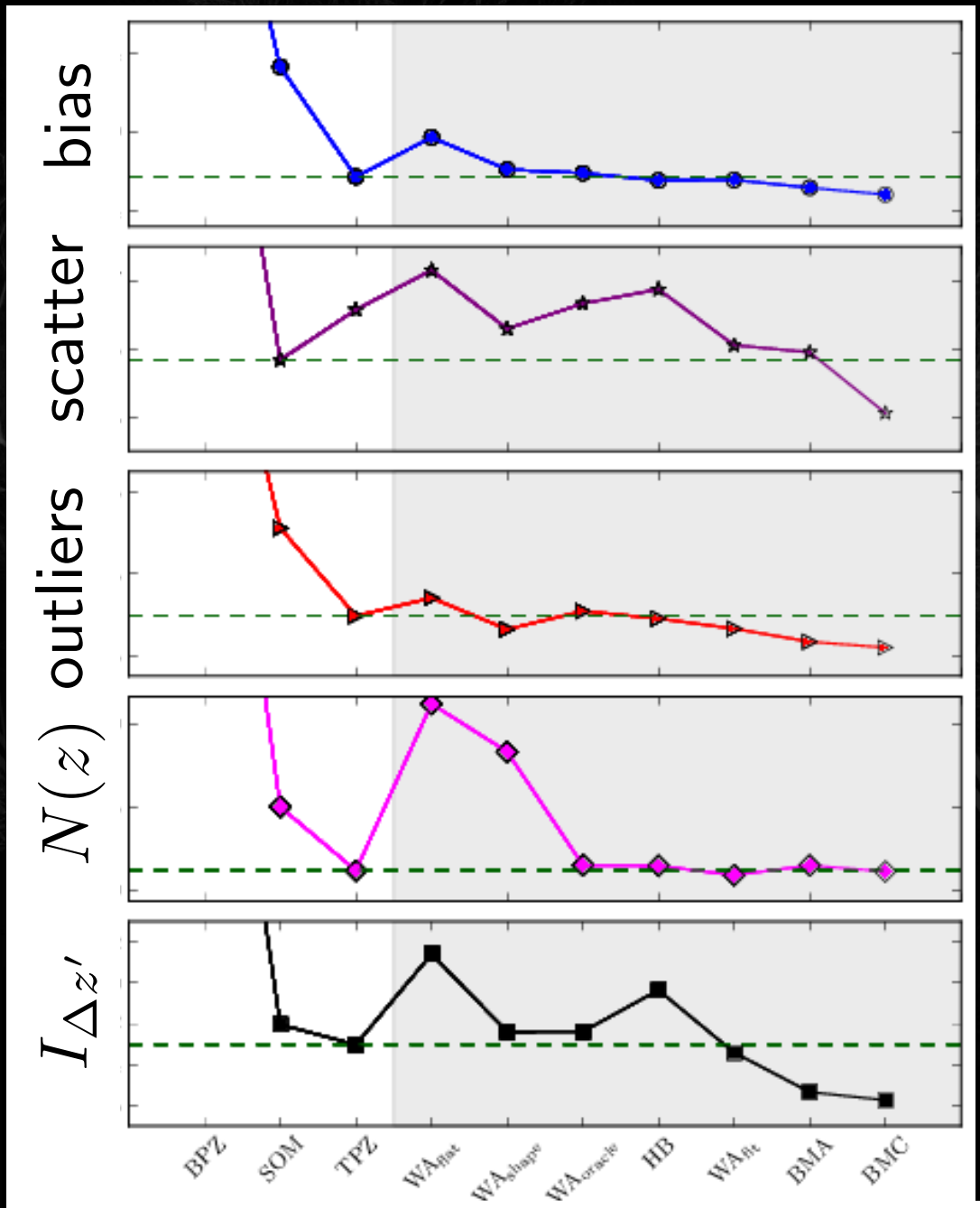
⇓

photo-z PDF

+

Outliers

Photo-z PDF combination: Results



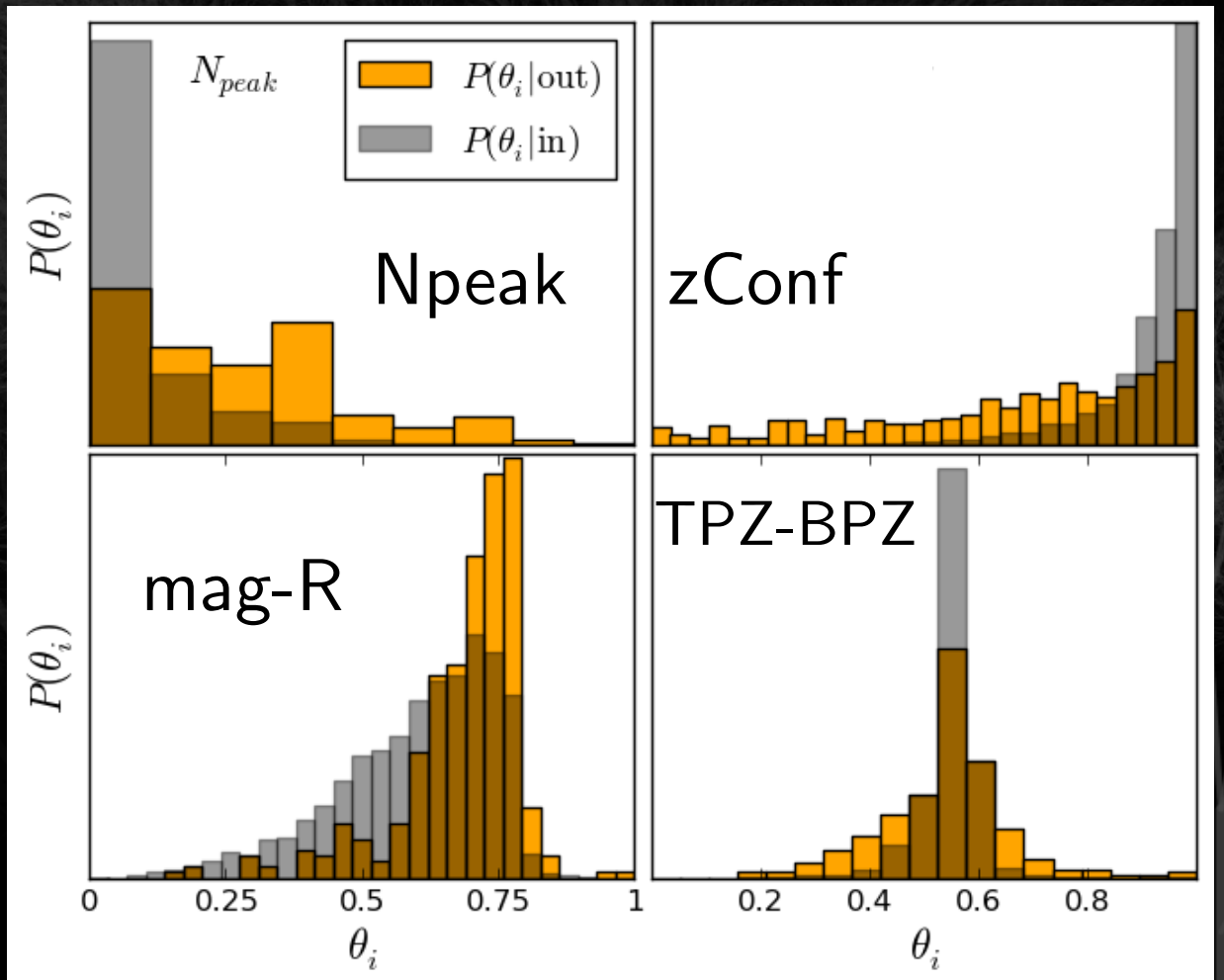
Carrasco Kind & Brunner 2014c (MNRAS, 442, 3380)

- Several combination methods
- Bayesian model averaging (BMA) and combination (BMC) are the best
- Same applies to S/G (Kim, Brunner & CK in prep.)

Photo-z PDF combination: Outliers

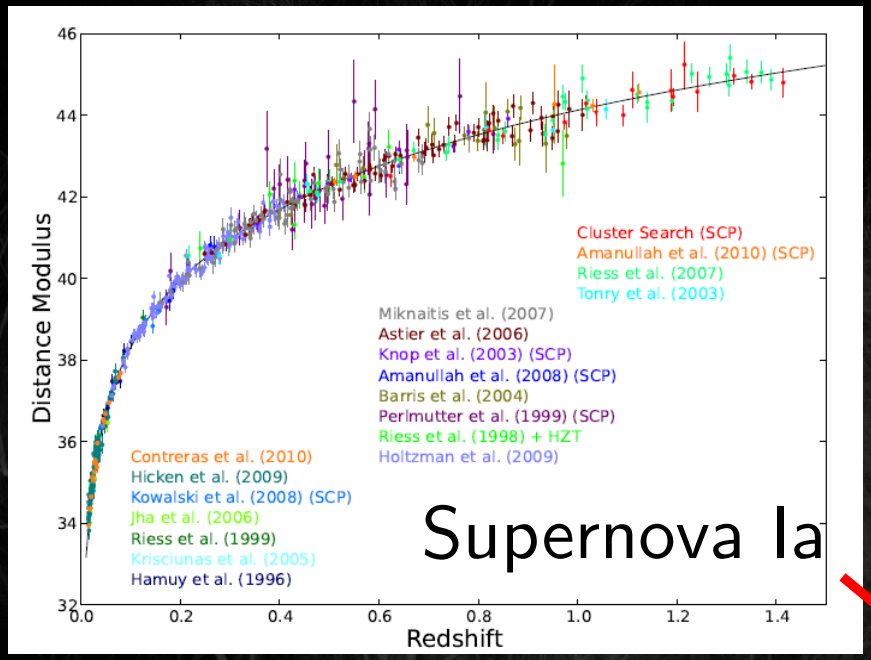
Naïve Bayes Classifier (same used for spam emails) to identify "spam" galaxies using information from multiple techniques

Each feature provides information about these two classes, and can be combined to make a stronger classifier

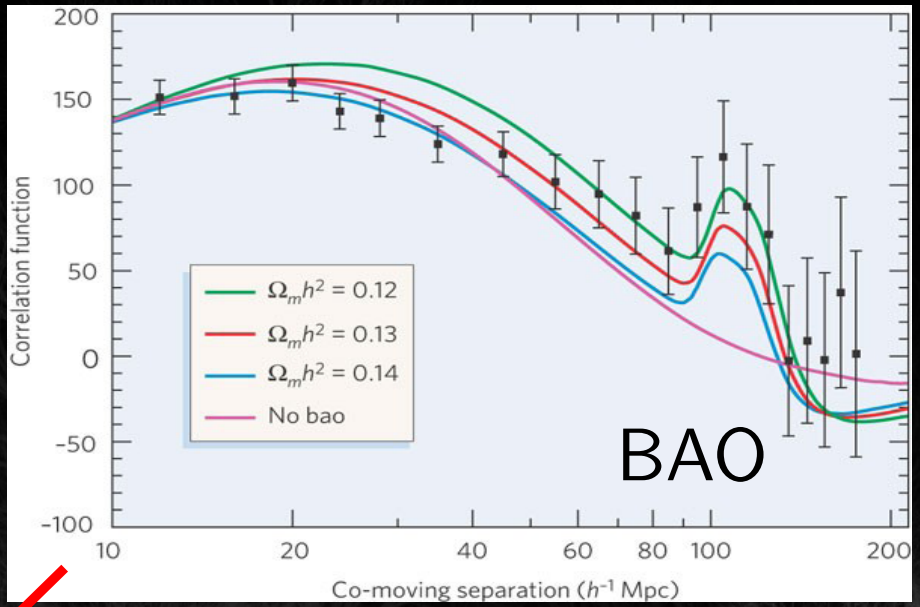


Carrasco Kind & Brunner 2014c (MNRAS, 442, 3380)

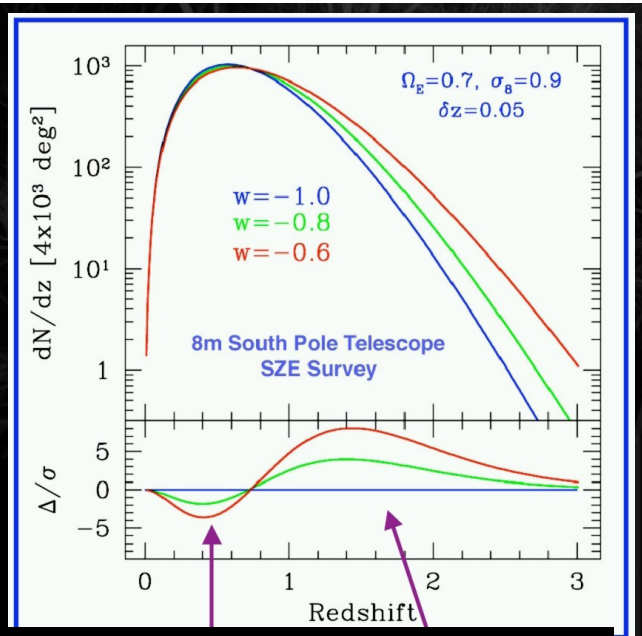
Observational Probes of Dark Energy



Supernova Ia



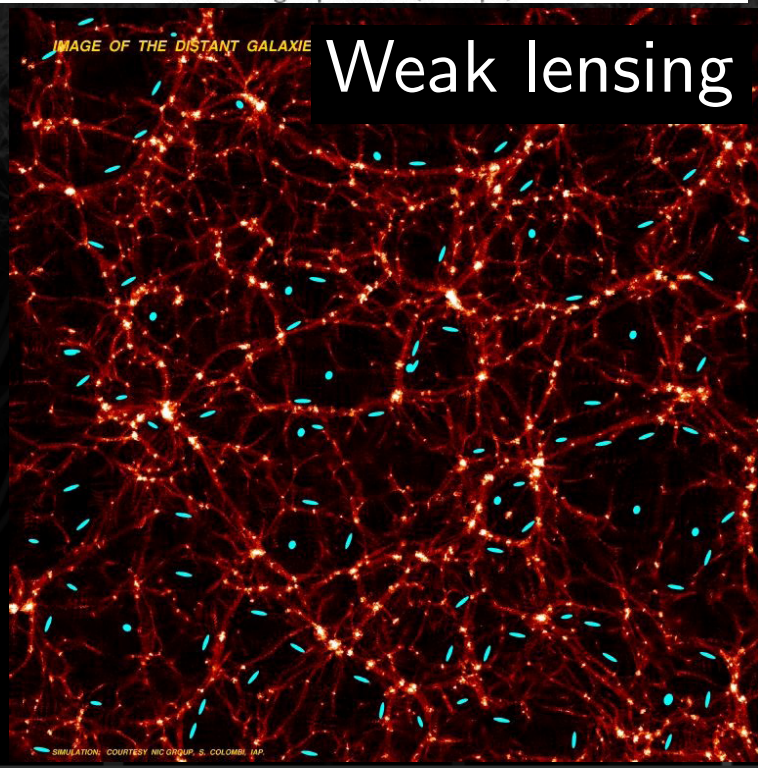
BAO



Cluster abundance

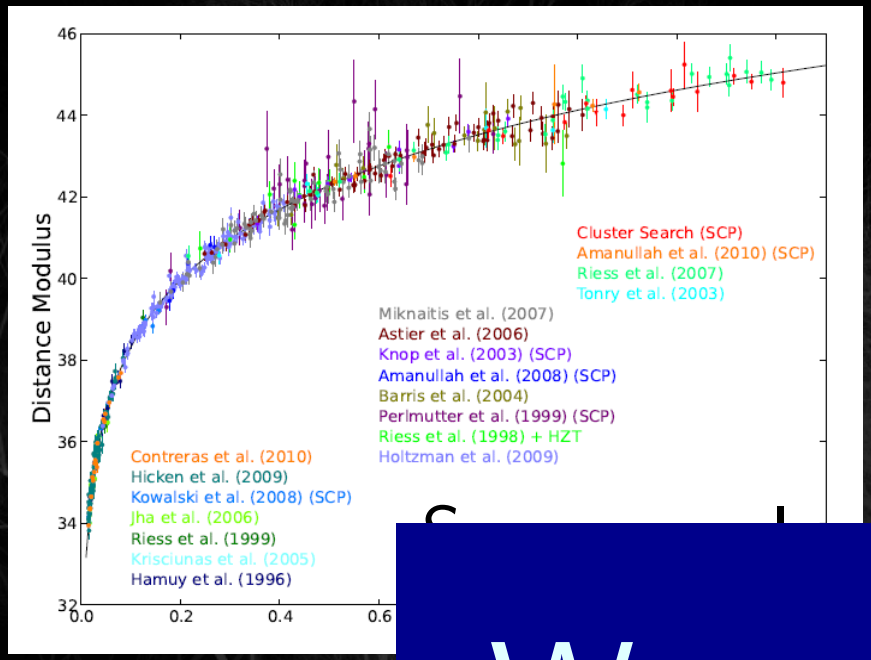
Geometry

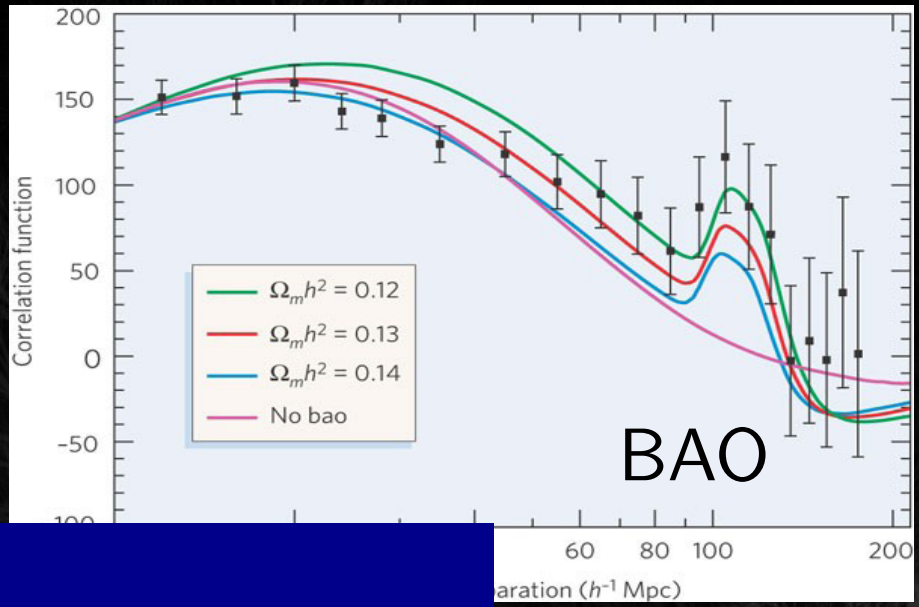
Growth of structure



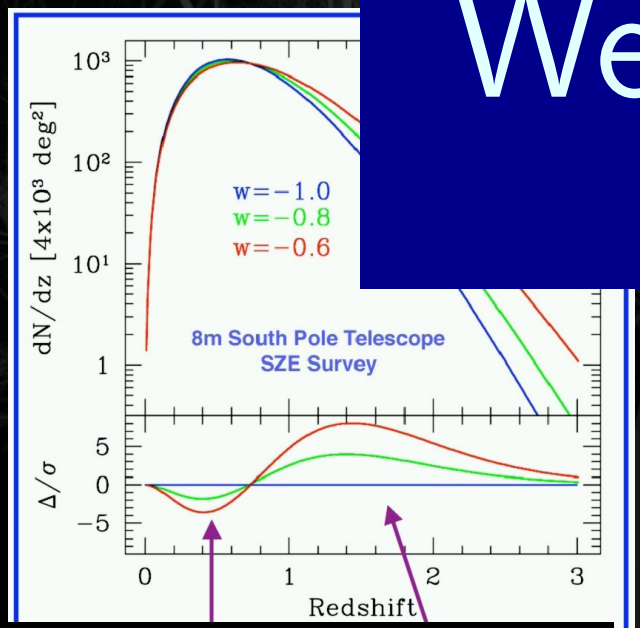
Weak lensing

Observational Probes of Dark Energy



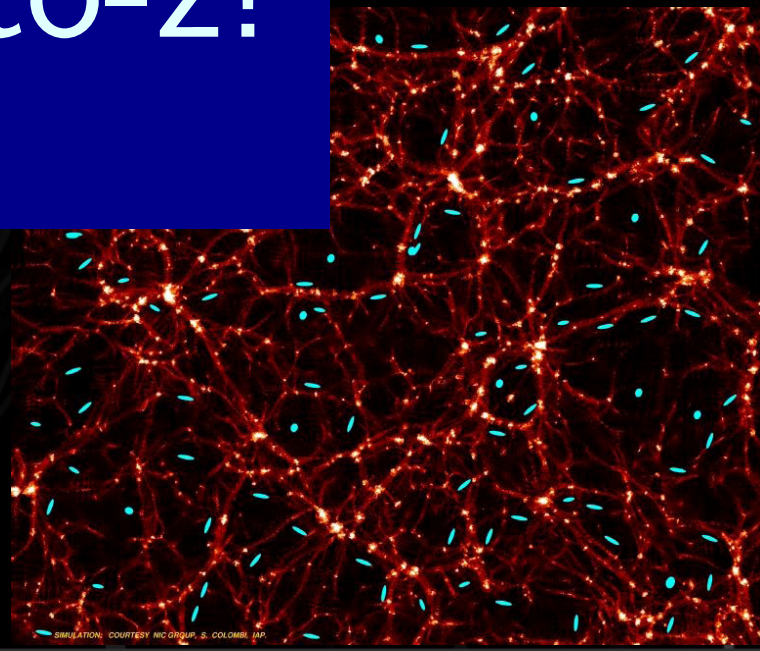


We need photo-z!



Cluster abundance

Growth of structure



Weak lensing

Photo-z for DES SV data

Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **445,** 1482–1506 (2014) doi:10.1093/mnras/stu1836

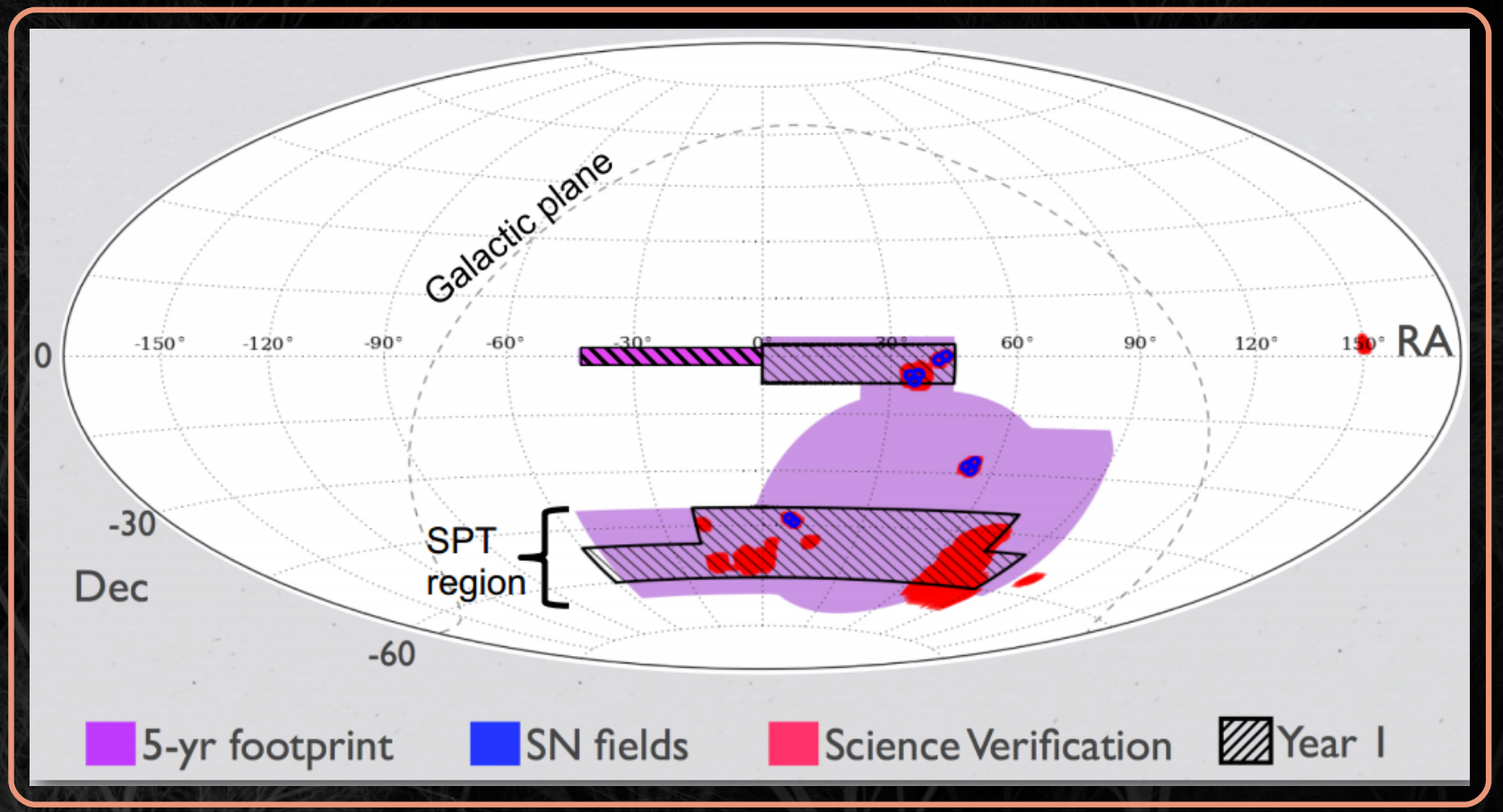
Photometric redshift analysis in the Dark Energy Survey Science Verification data

C. Sánchez,[1⋆] M. Carrasco Kind,[2,3] H. Lin,[4] R. Miquel,[1,5] F. B. Abdalla,[6] A. Amara,[7] M. Banerji,[6] C. Bonnett,[1] R. Brunner,[2] D. Capozzi,[8] A. Carnero,[9,10] F. J. Castander,[11] L. A. N. da Costa,[9,10] C. Cunha,[12] A. Fausti,[10] D. Gerdes,[13] N. Greisel,[14,15] J. Gschwend,[9,10] W. Hartley,[7,16] S. Jouvel,[6] O. Lahav,[6] M. Lima,[10,17] M. A. G. Maia,[9,10] P. Martí,[1] R. L. C. Ogando,[9,10] F. Ostrovski,[9,10] P. Pellegrini,[9] M. M. Rau,[14,15] I. Sadeh,[6] S. Seitz,[14,15] I. Sevilla-Noarbe,[18] A. Sypniewski,[13] J. de Vicente,[18] T. Abbot,[19] S. S. Allam,[4,20] D. Atlee,[21] G. Bernstein,[22] J. P. Bernstein,[23] E. Buckley-Geer,[4] D. Burke,[12,24] M. J. Childress,[25,26] T. Davis,[26,27] D. L. DePoy,[28,29] A. Dey,[21,30] S. Desai,[31,32] H. T. Diehl,[4] P. Doel,[6] J. Estrada,[4] A. Evrard,[13,33,34] E. Fernández,[1] D. Finley,[4] B. Flaugher,[4] J. Frieman,[4] E. Gaztanaga,[11] K. Glazebrook,[35] K. Honscheid,[36] A. Kim,[37] K. Kuehn,[38] N. Kuropatkin,[4] C. Lidman,[38] M. Makler,[39] J. L. Marshall,[28,29] R. C. Nichol,[8] A. Roodman,[12,24] E. Sánchez,[18] B. X. Santiago,[10,40] M. Sako,[22] R. Scalzo,[25] R. C. Smith,[19] M. E. C. Swanson,[3] G. Tarle,[13] D. Thomas,[8,41] D. L. Tucker,[4] S. A. Uddin,[26,35] F. Valdés,[21] A. Walker,[19] F. Yuan[25,26] and J. Zuntz[42]

Affiliations are listed at the end of the paper

Sanchez et al. (2015). First published paper using DES data!

Photo-z for DES SV data



- DES Science Verification data
- Photo-z code comparison and analysis
- Good benchmark results for future releases

Spectroscopic data

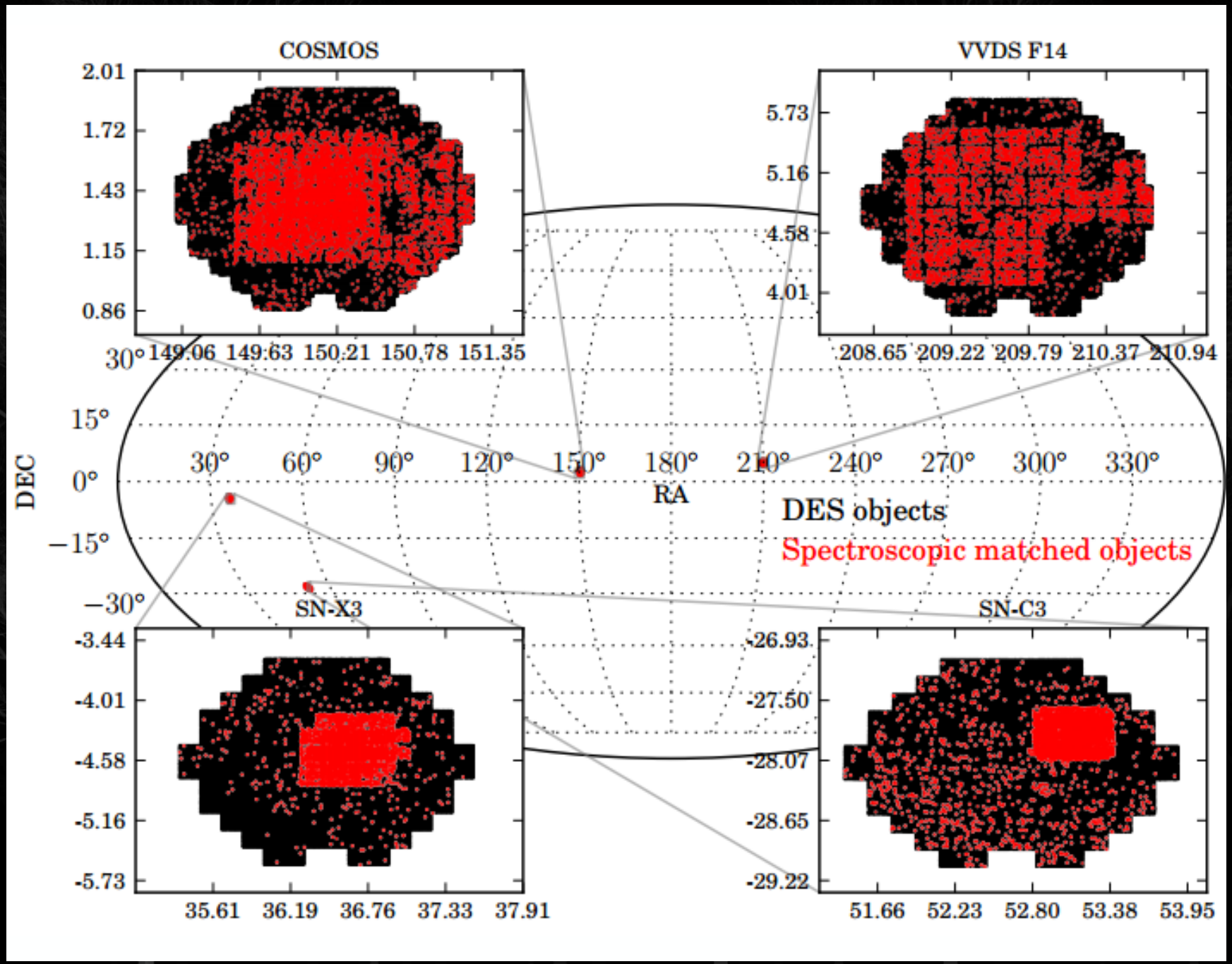
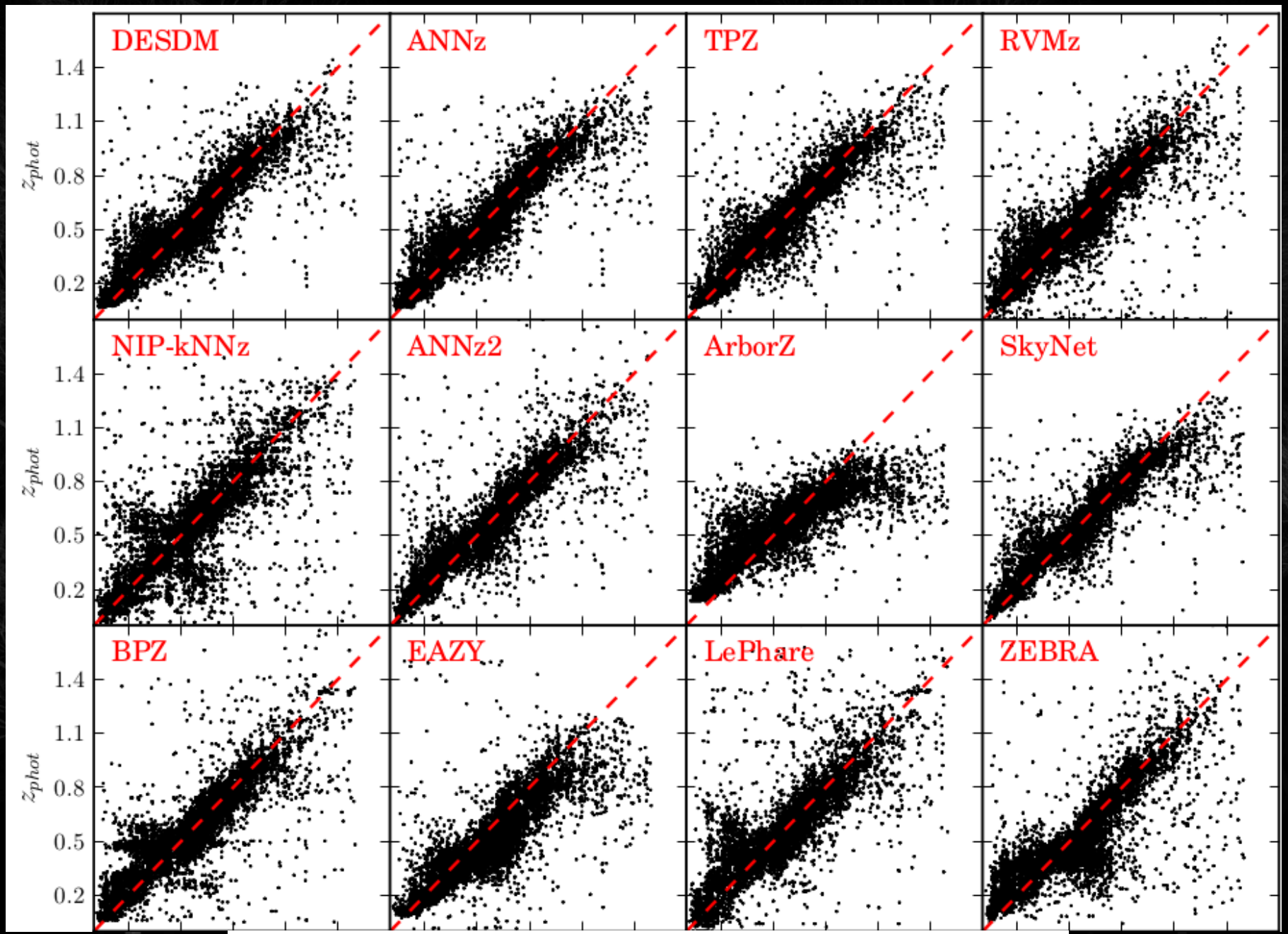


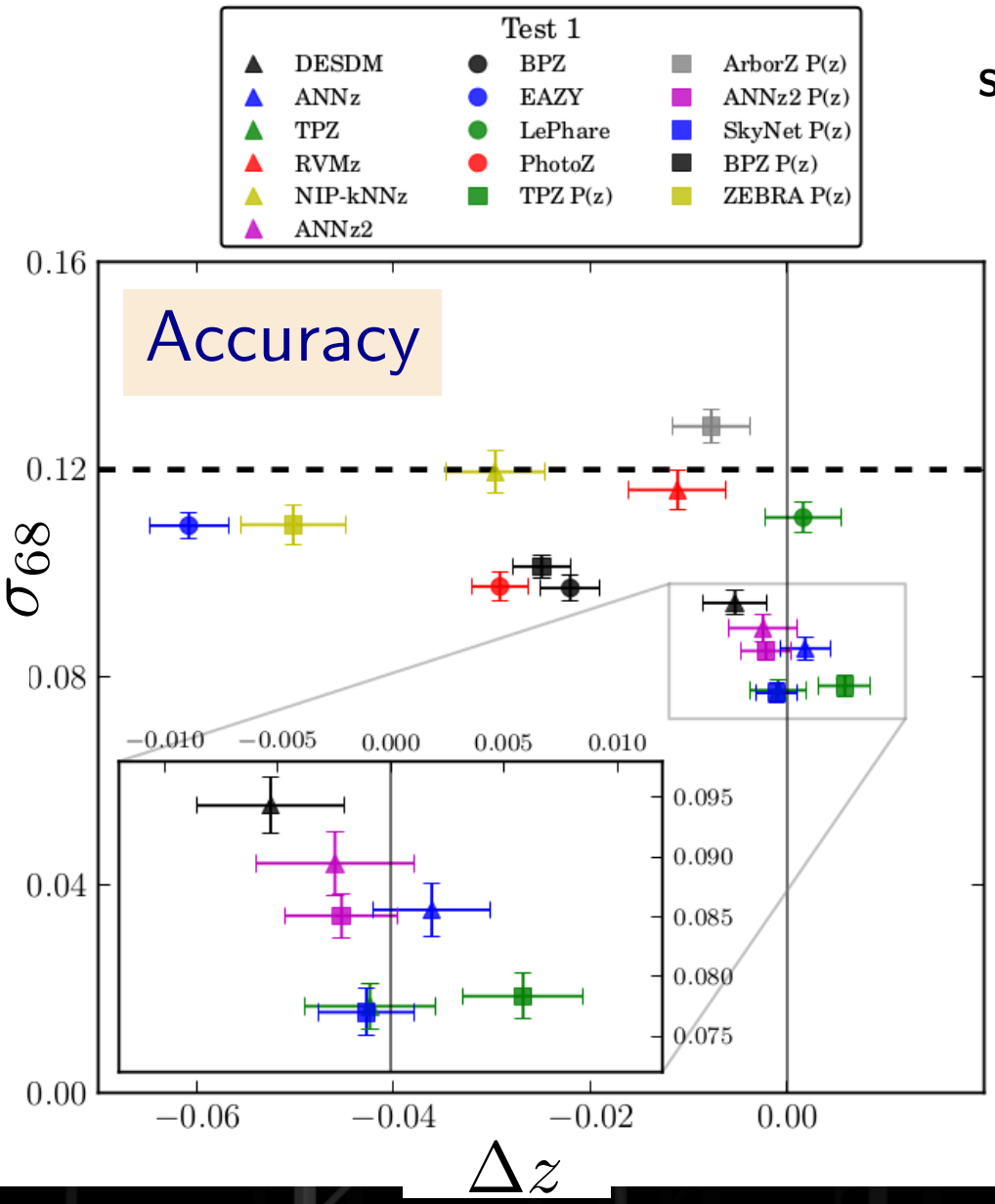
Photo-z PDF application: DES SV data

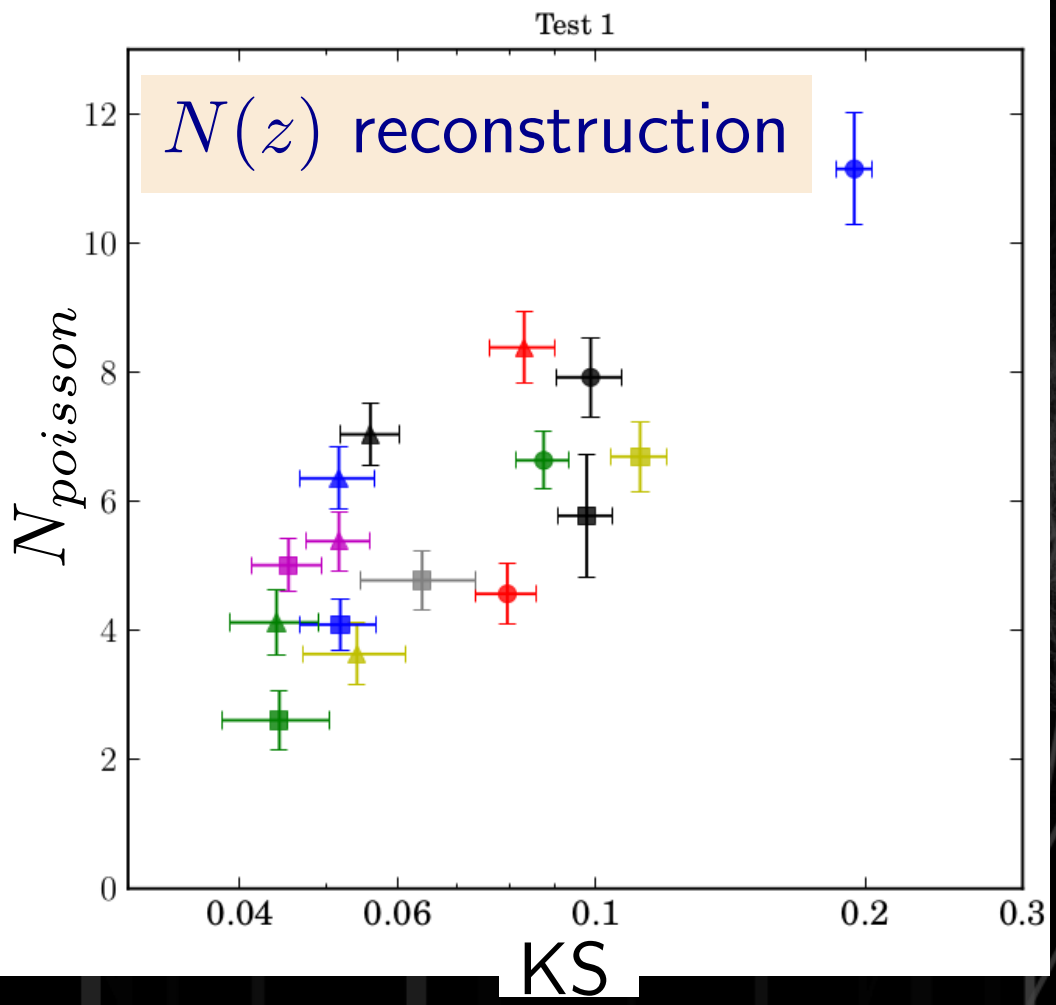


Sanchéz, Carrasco Kind, et al. 2014 (MNRAS, 445, 1482)

Photo-z for DES SV data

Sanchéz, Carrasco Kind, et al. 2014 (MNRAS, 445, 1482)

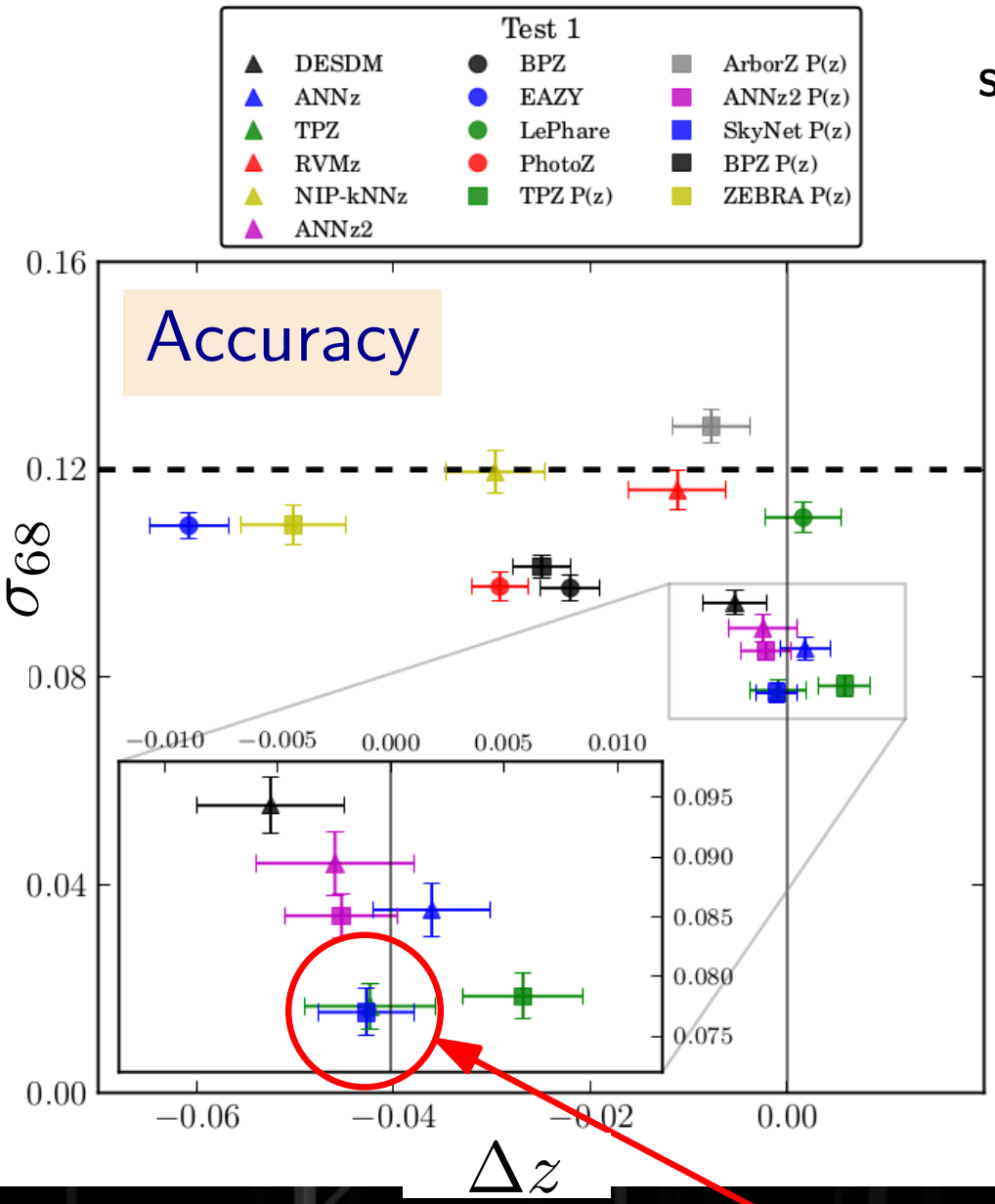


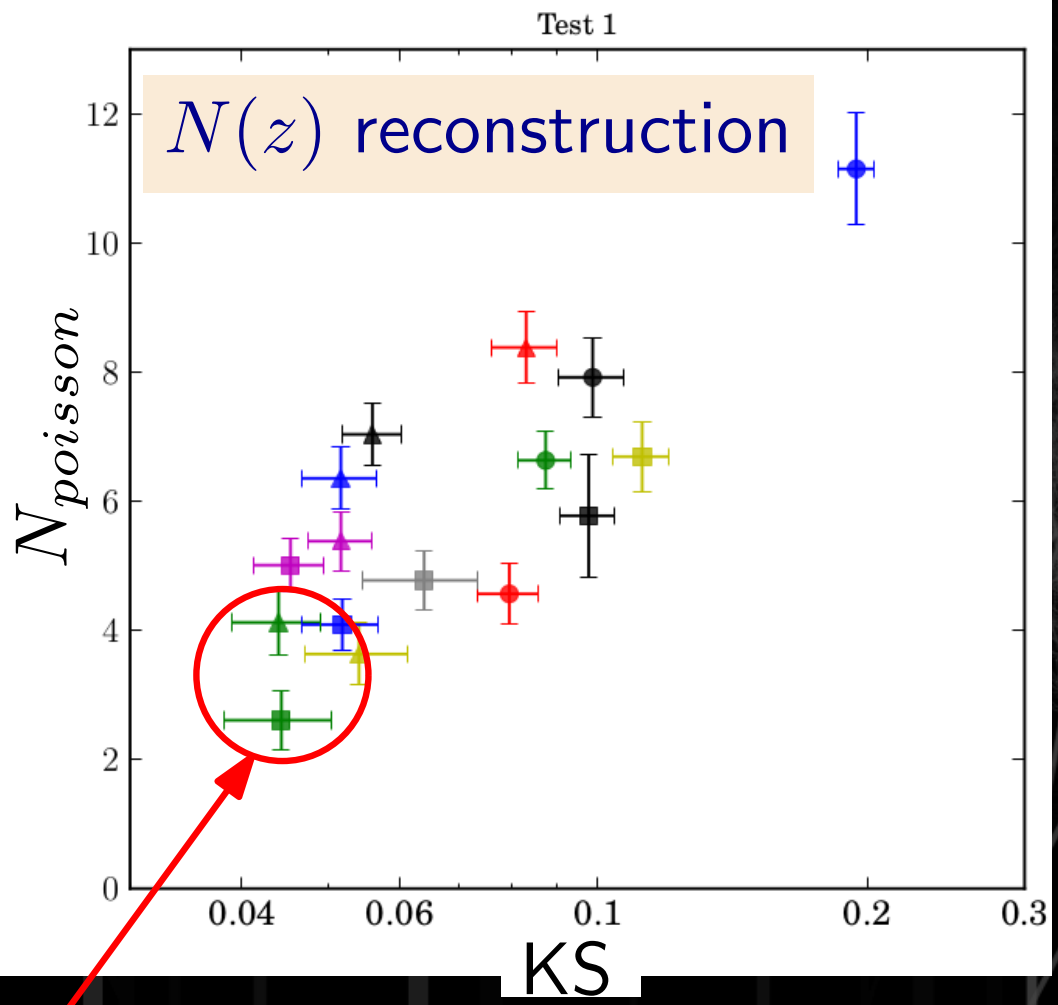


13 photo-z codes comparison

Photo-z for DES SV data

Sanchéz, Carrasco Kind, et al. 2014 (MNRAS, 445, 1482)





TPZ

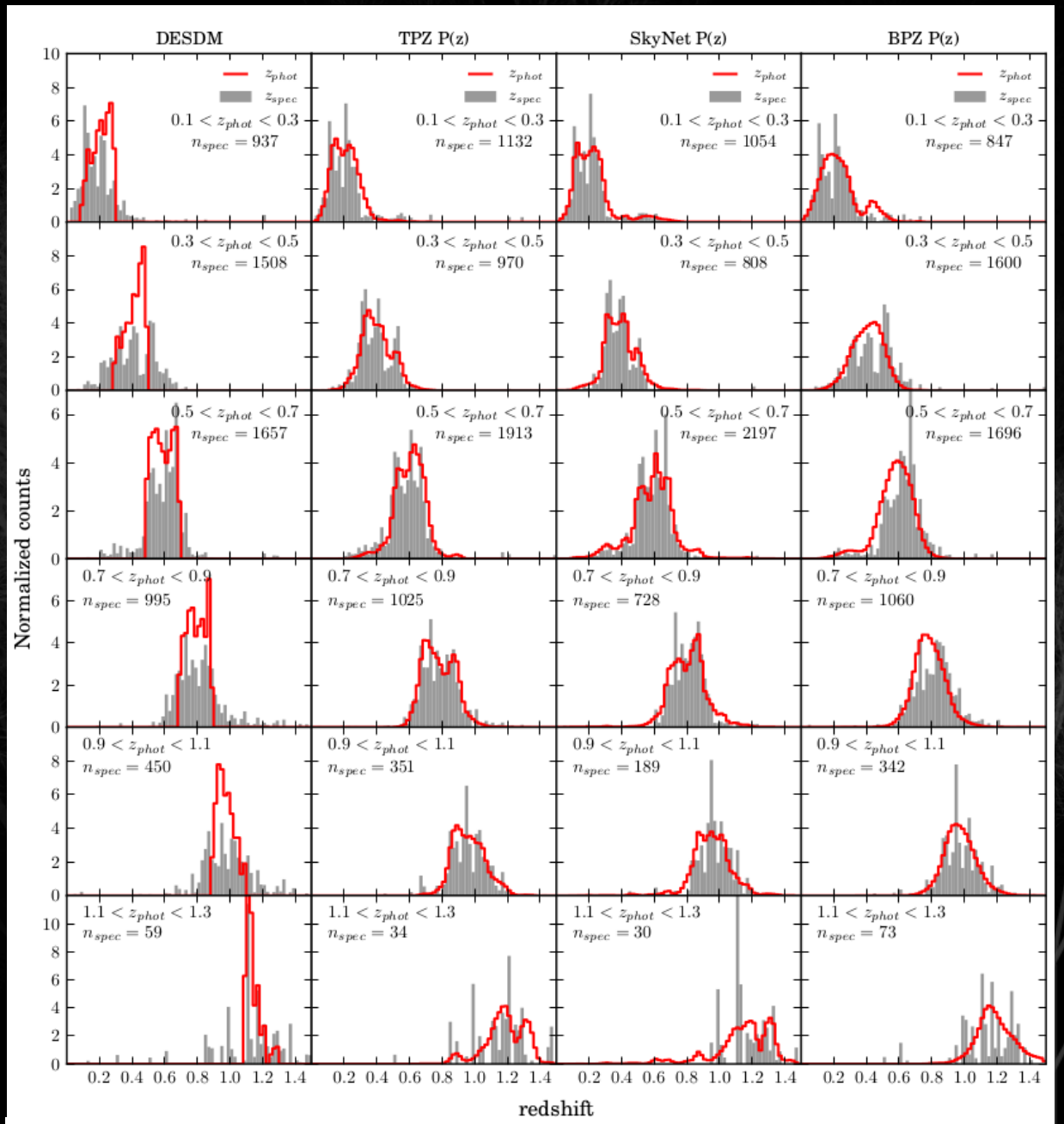
Photo-z for DES SV data

4 codes recommendation

Default, 2 training and 1 template

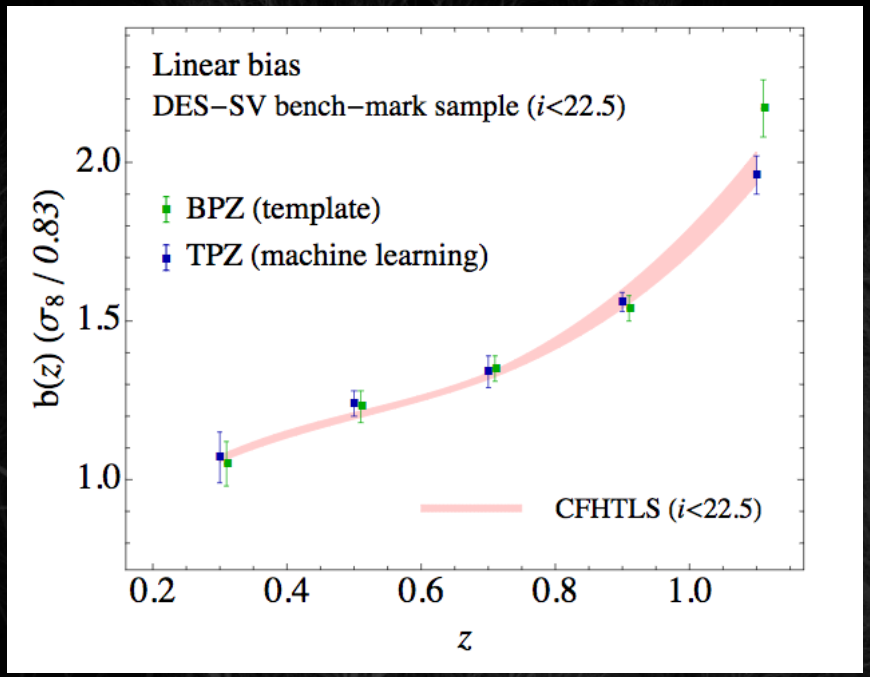
PDF methods are better for $N(z)$

Combination methods! (not used here)

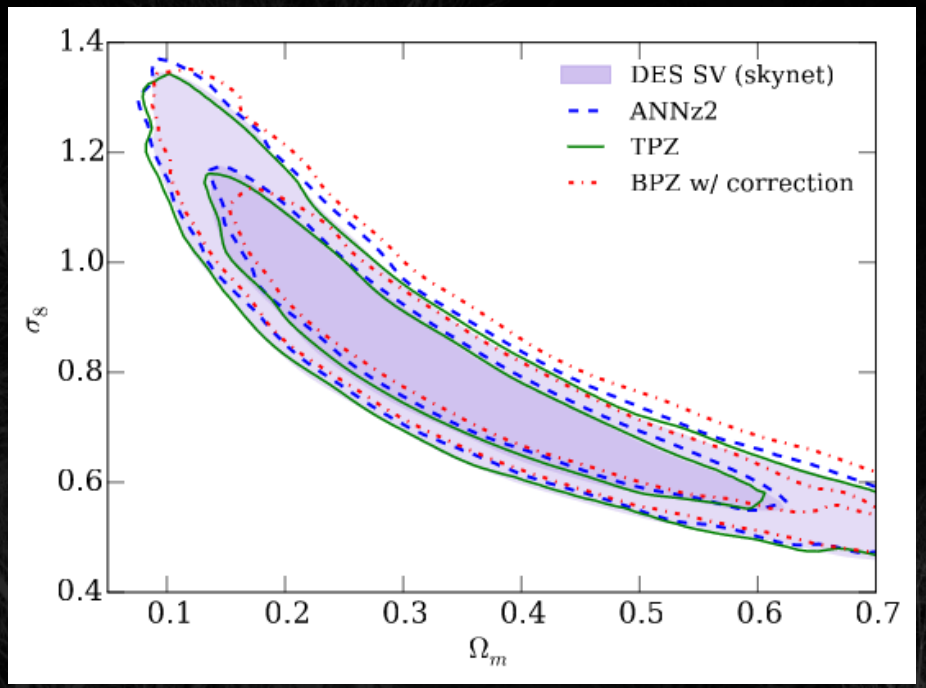


Sanchéz, Carrasco Kind, et al. 2014 (MNRAS, submitted)

Photo-z for DES SV data

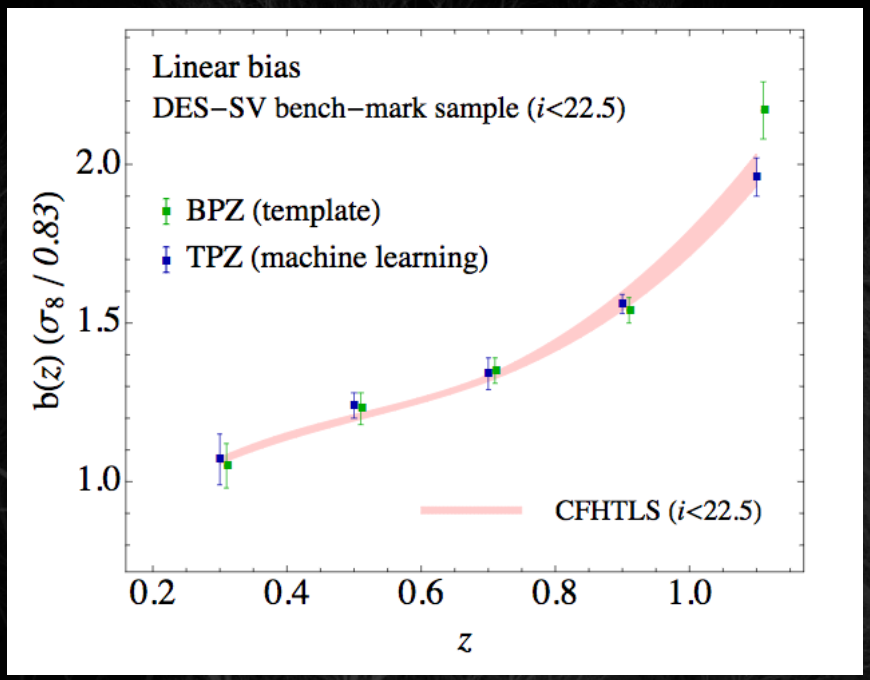


Galaxy clustering, photometric redshifts and diagnosis of systematics in the DES SV data (arXiv:1507.05360)



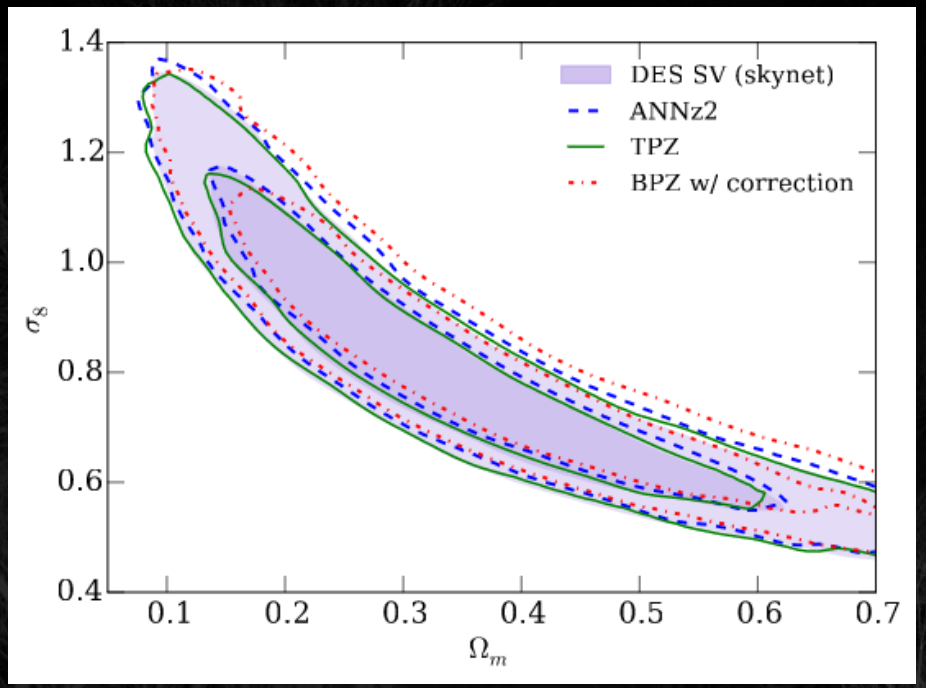
Cosmology from Cosmic Shear with DES SV Data (arXiv:1507.05552)

Photo-z for DES SV data



Galaxy clustering, photometric redshifts and diagnosis of systematics in the DES SV data (arXiv:1507.05360)

... and 14 other DES papers!



Cosmology from Cosmic Shear with DES SV Data (arXiv:1507.05552)

For the first time using Machine Learning photo-z in cosmology!

Technical Challenges

- Volume of data (SV: 25M, Y1: 130M, Y3: 250M)

- Training not only garbage in $\rightarrow$ garbage out BUT features in $\rightarrow$ features out, overfitting, manifolds, etc...

- Cosmic variance, biases, magnitude errors, uniformity of spec samples, cross matching, etc

- Software sustainability, distribution and license.

- PDF generation and storage. Multiple methods, multiple versions!

Photo-z PDF representation and storage

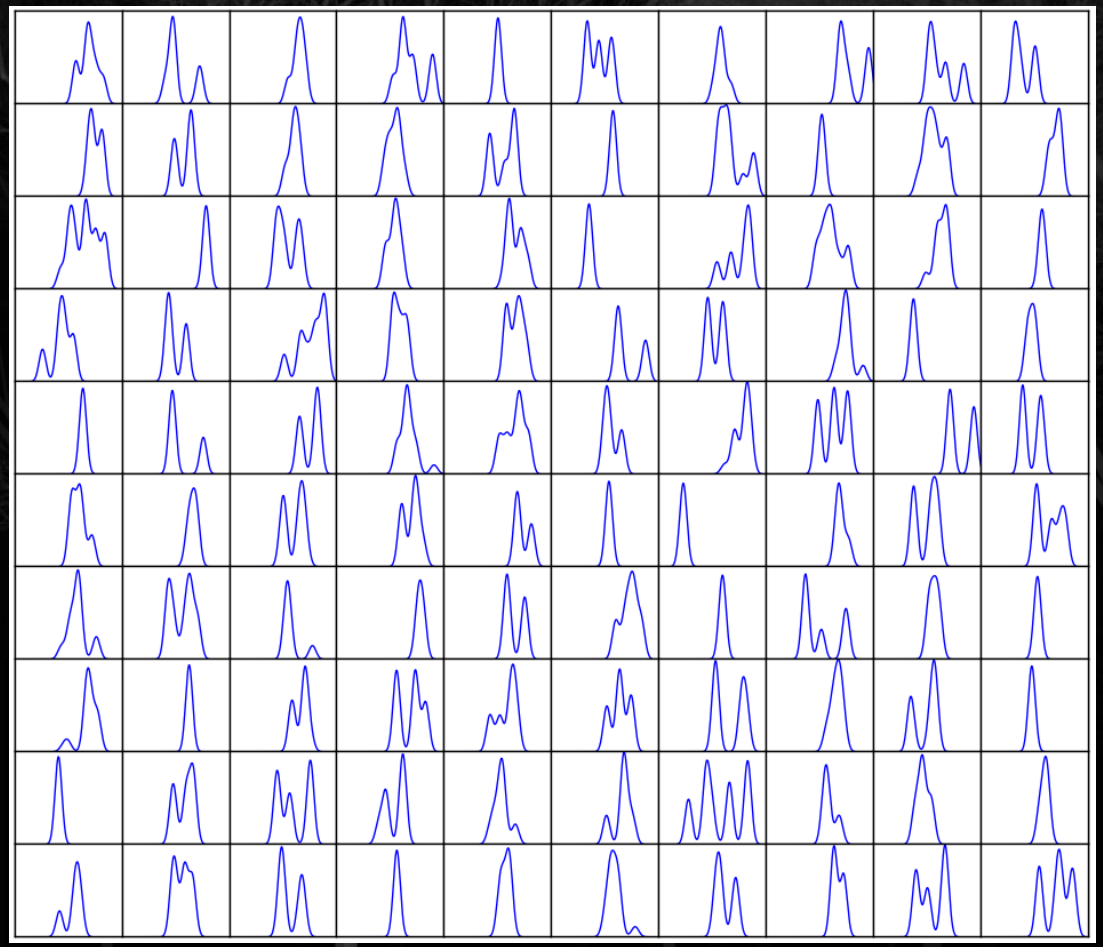


Photo-z PDF storage: Strategies

Single Gaussian fit

Multi-Gaussan fit

Monte Carlo sampling

Sparse representation techniques

Reduce number of points while increasing accuracy

Photo-z PDF storage: Strategies

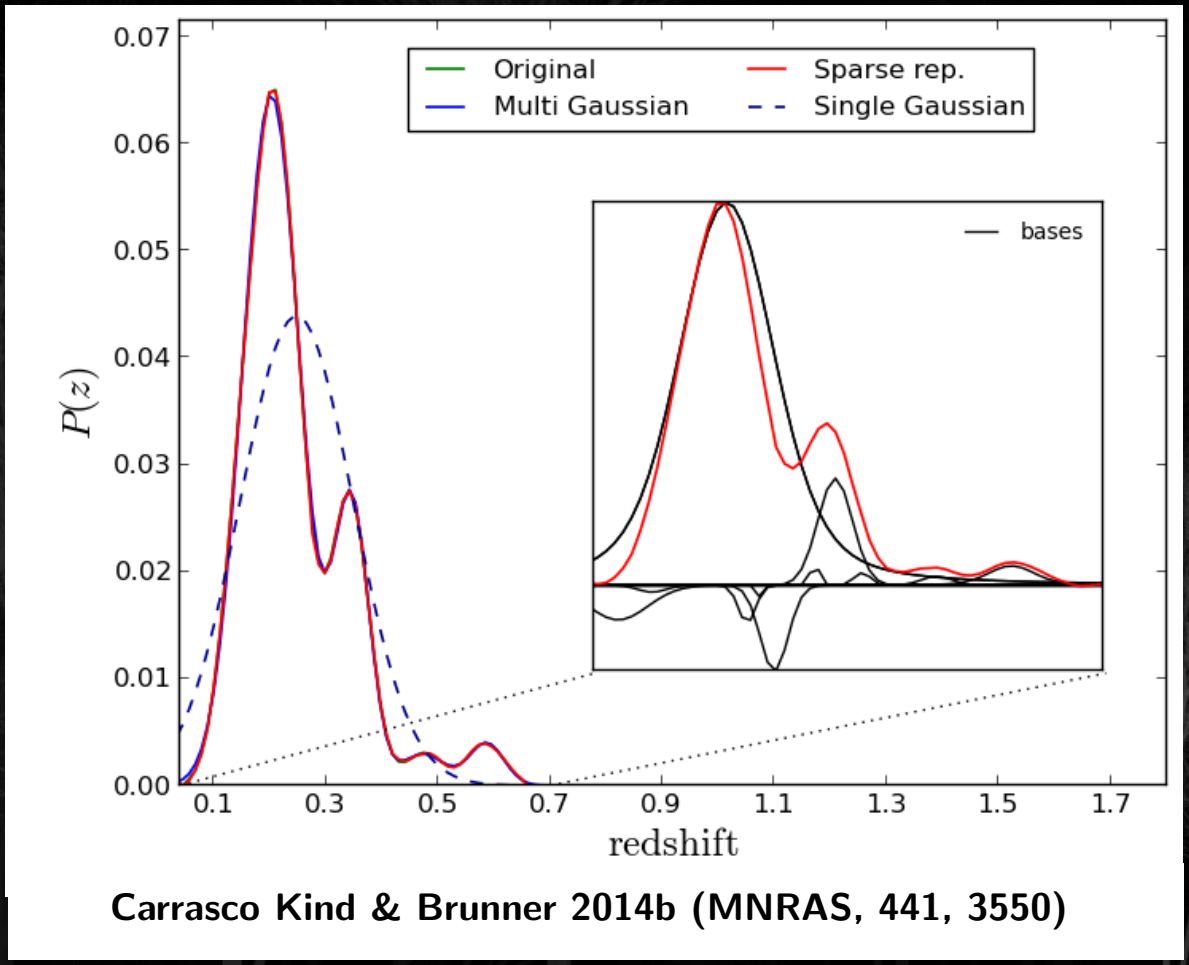
Single Gaussian fit

Multi-Gaussan fit

Monte Carlo sampling

Sparse representation techniques

Reduce number of points while increasing accuracy



Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)

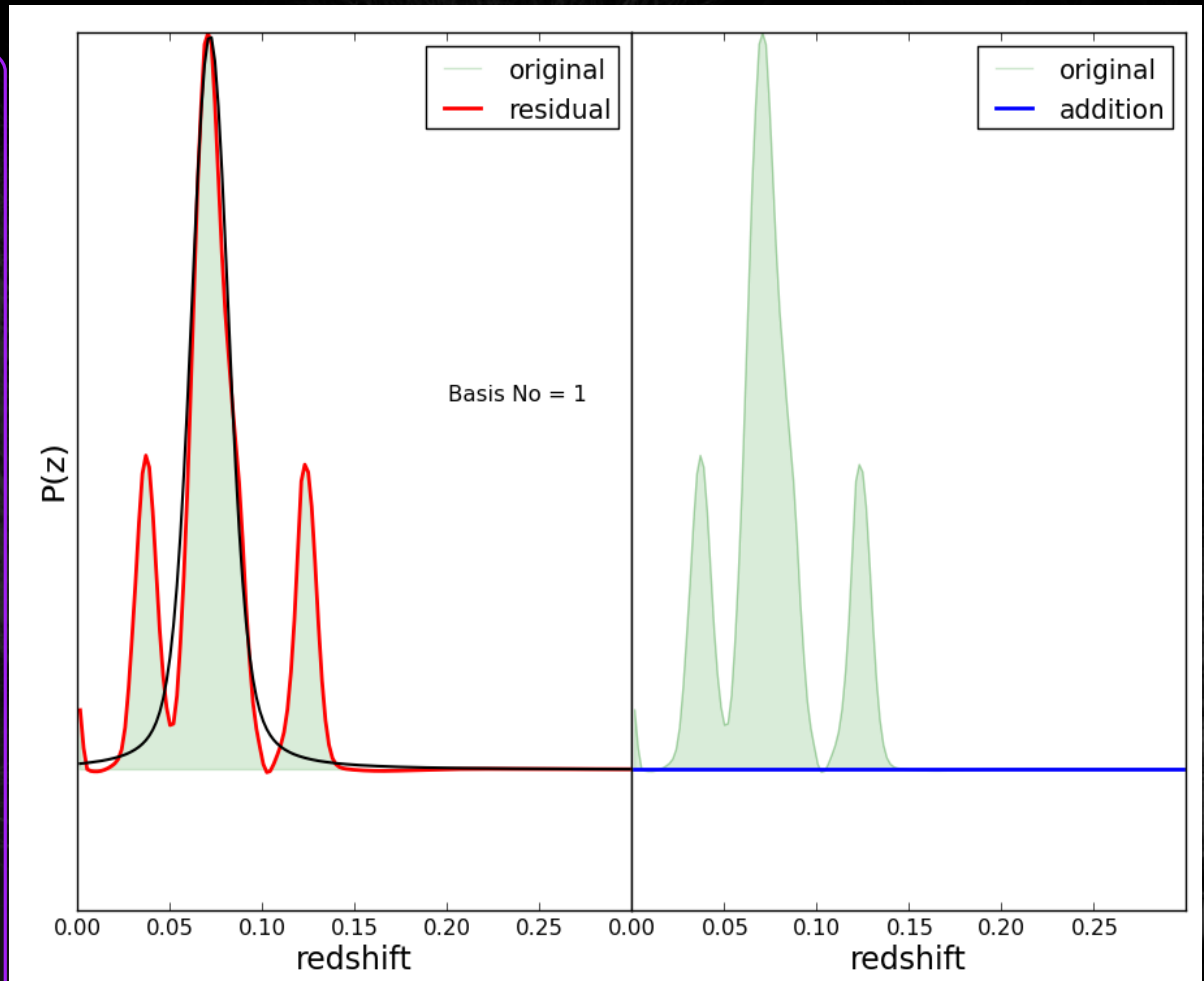
Photo-z PDF storage: Sparse representation

Use Gaussian and Voigt profiles as bases, need $N^2_{\rm original}$ bases

With only 10-20 bases achieve 99.9 % accuracy

Use 32-bits integer per basis, compression

Store Multiple PDFs



Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)

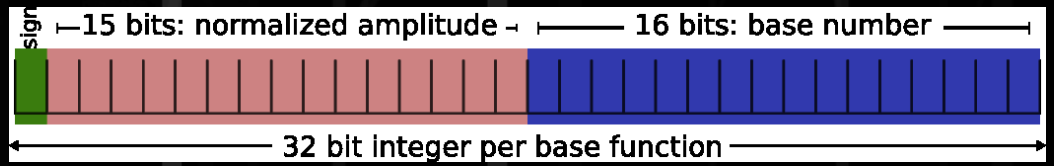


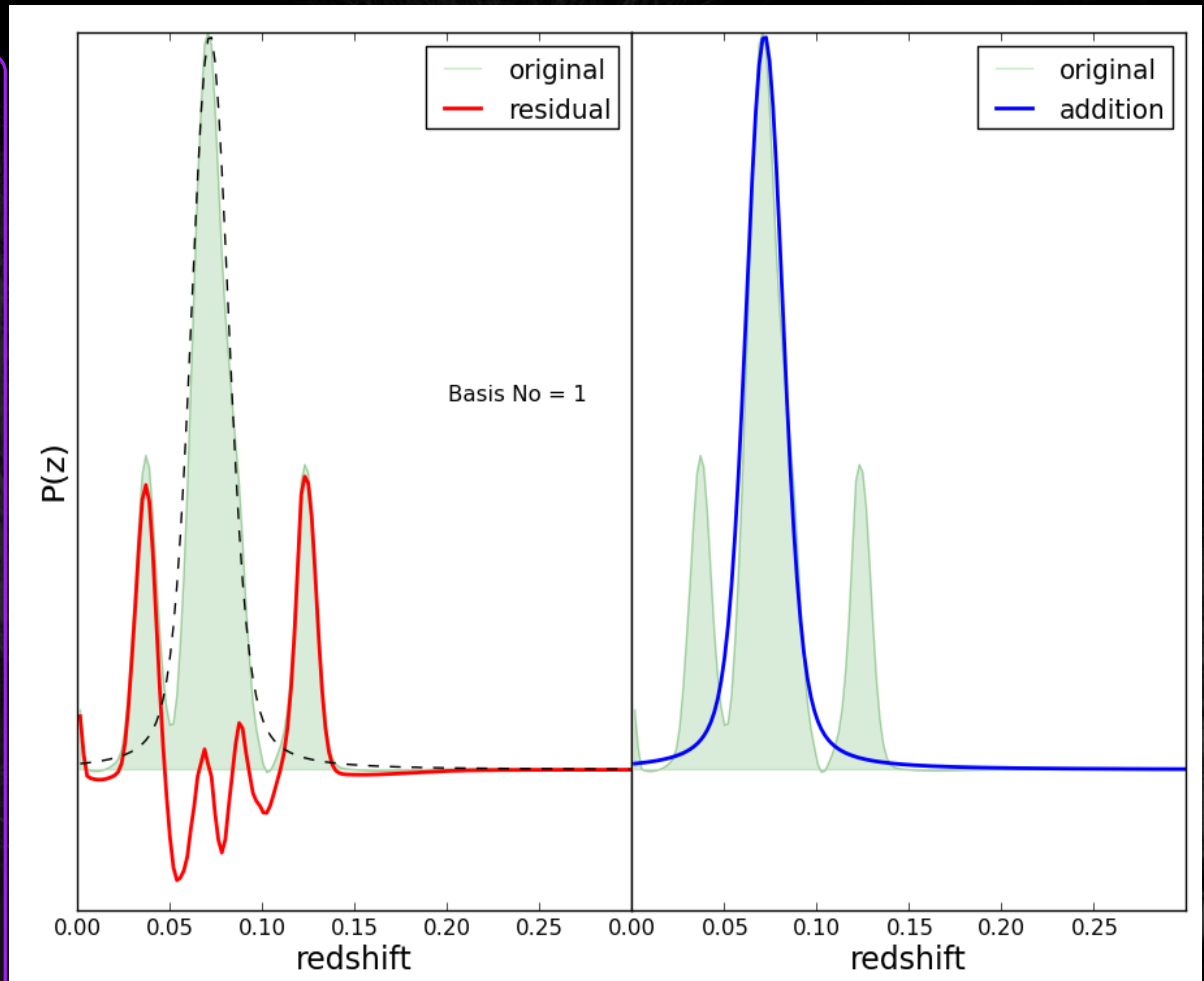
Photo-z PDF storage: Sparse representation

Use Gaussian and Voigt profiles as bases, need $N^2_{\rm original}$ bases

With only 10-20 bases achieve 99.9 % accuracy

Use 32-bits integer per basis, compression

Store Multiple PDFs



Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)

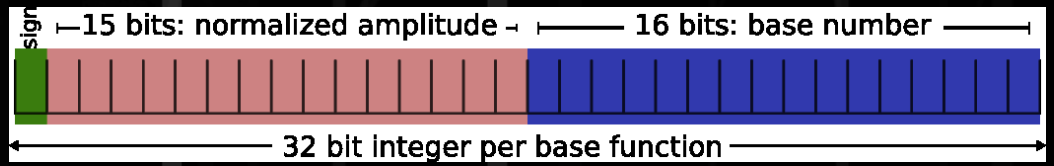


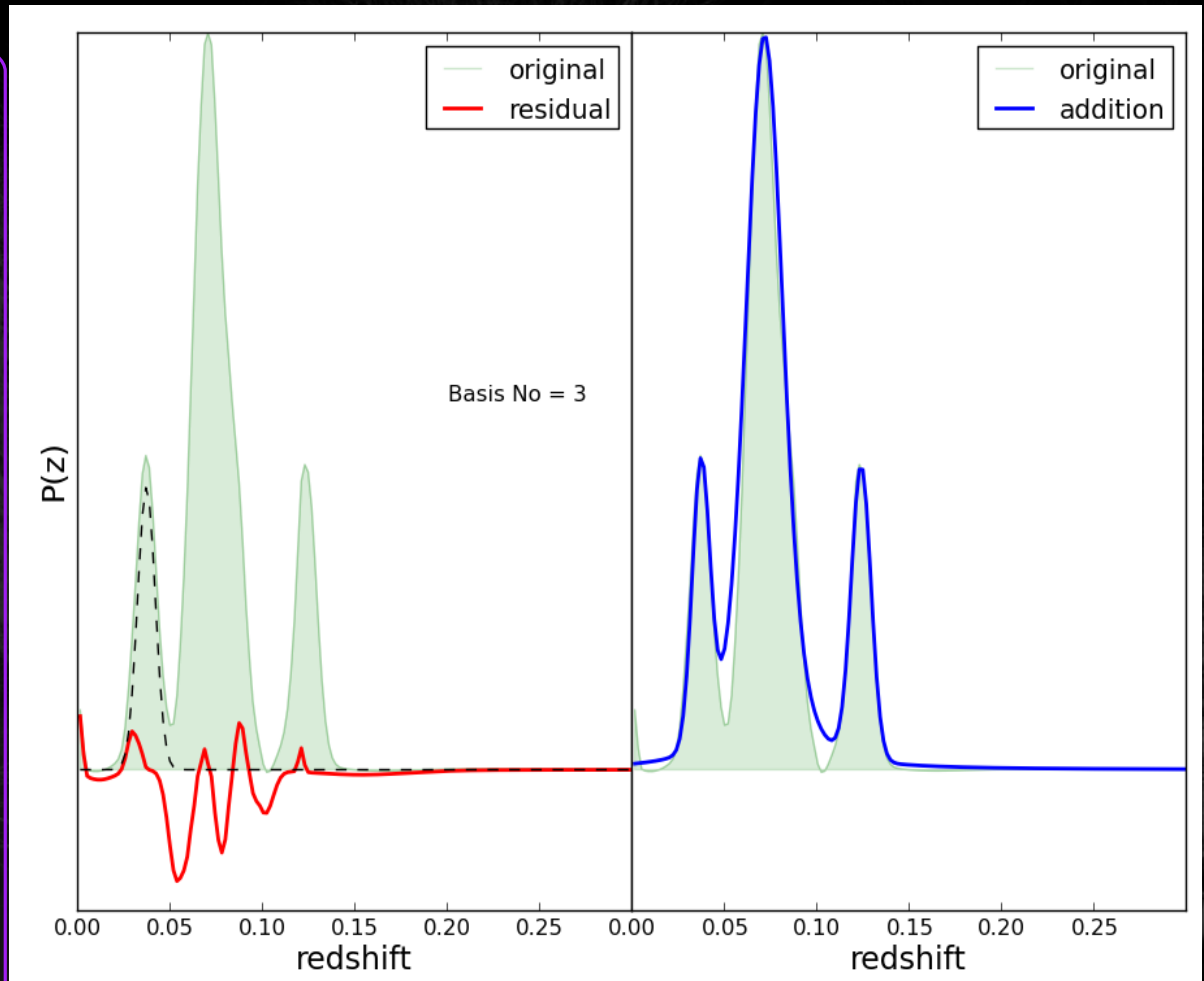
Photo-z PDF storage: Sparse representation

Use Gaussian and Voigt profiles as bases, need $N^2_{\rm original}$ bases

With only 10-20 bases achieve 99.9 % accuracy

Use 32-bits integer per basis, compression

Store Multiple PDFs



Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)

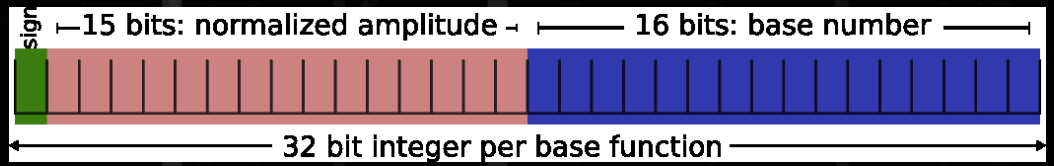


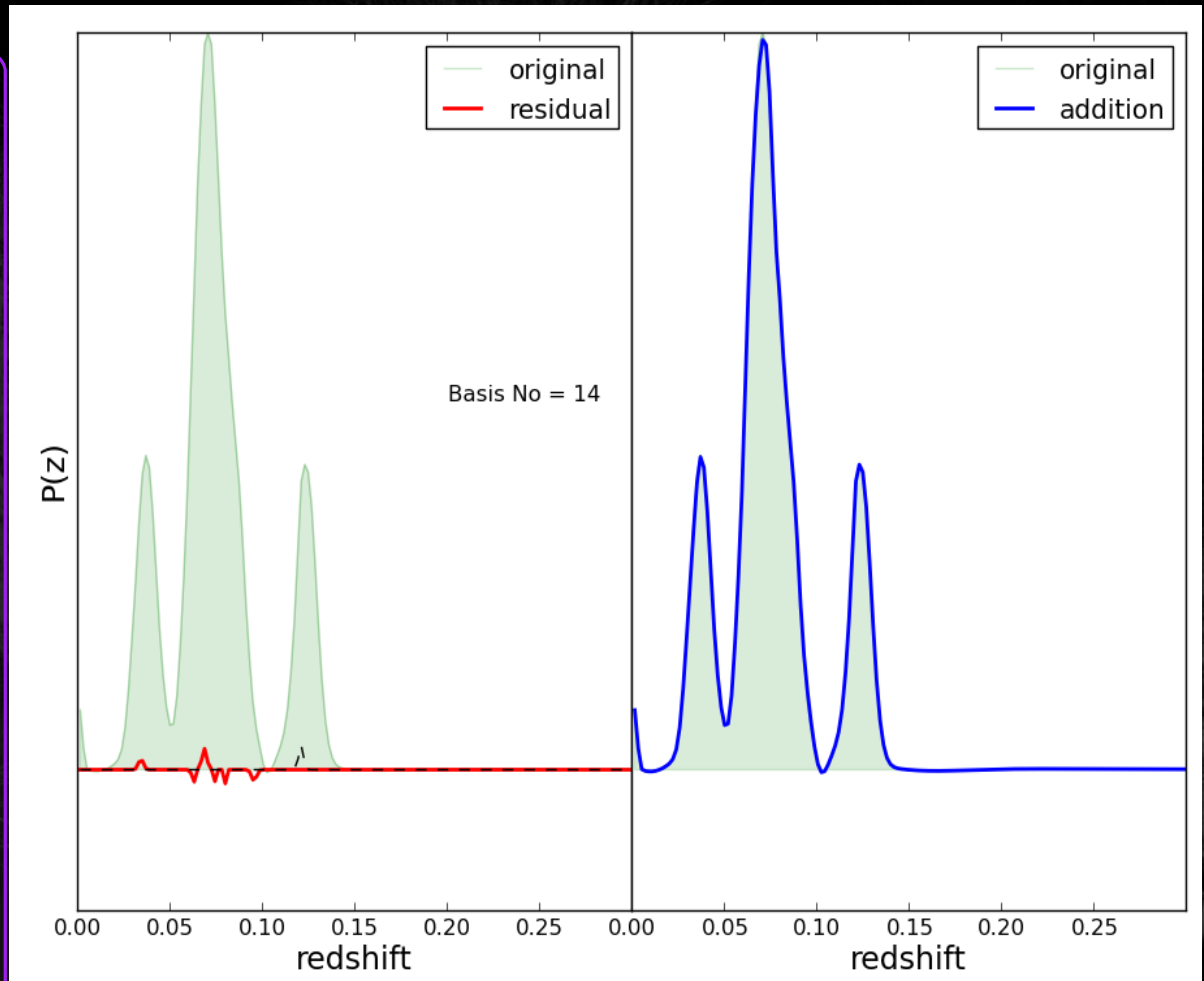
Photo-z PDF storage: Sparse representation

Use Gaussian and Voigt profiles as bases, need $N^2_{\rm original}$ bases

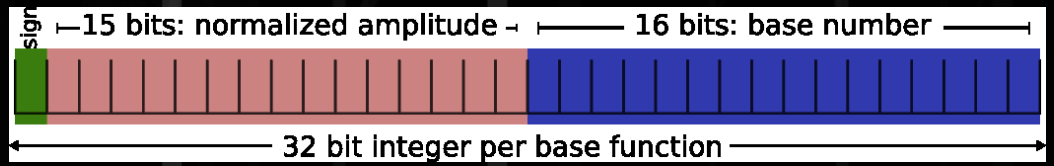
With only 10-20 bases achieve 99.9 % accuracy

Use 32-bits integer per basis, compression

Store Multiple PDFs



Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)



$N(z)$ and sparse representation

By definition:

$$N(z) = \sum_{k=1}^{N} \int_{z-\Delta z/2}^{z+\Delta z/2} P_k(z) dz$$

$N(z)$ and sparse representation

By definition:

$$N(z) = \sum_{k=1}^{N} \int_{z-\Delta z/2}^{z+\Delta z/2} P_k(z) dz$$

Using sparse representation, we represent each PDF pz_k as:

$\mathbf{pz}_k \approx \mathbf{D} \cdot \boldsymbol{\delta}_k$ $\mathbf{D}$ is the dictionary, $\boldsymbol{\delta}_k$ is the sparse vector, then

$$N(z) = \sum_{k=1}^{N} \boldsymbol{\delta}_k \cdot \int_{z-\Delta z/2}^{z+\Delta z/2} \mathbf{D} dz$$

Only bases are integrated

by precomputing:

$$\boldsymbol{\delta}_N = \sum_{k=1}^{N} \boldsymbol{\delta}_k \qquad \mathbf{I_D}(z) = \int_{z-\Delta z/2}^{z+\Delta z/2} \mathbf{d}_j dz \qquad j = 1, 2, \ldots, m$$

$N(z)$ and sparse representation

By definition:

$$N(z) = \sum_{k=1}^{N} \int_{z-\Delta z/2}^{z+\Delta z/2} P_k(z) dz$$

Using sparse representation, we represent each PDF pz_k as:

$\mathbf{pz}_k \approx \mathbf{D} \cdot \boldsymbol{\delta}_k$ $\mathbf{D}$ is the dictionary, $\boldsymbol{\delta}_k$ is the sparse vector, then

$$N(z) = \sum_{k=1}^{N} \boldsymbol{\delta}_k \cdot \int_{z-\Delta z/2}^{z+\Delta z/2} \mathbf{D} dz$$

Only bases are integrated

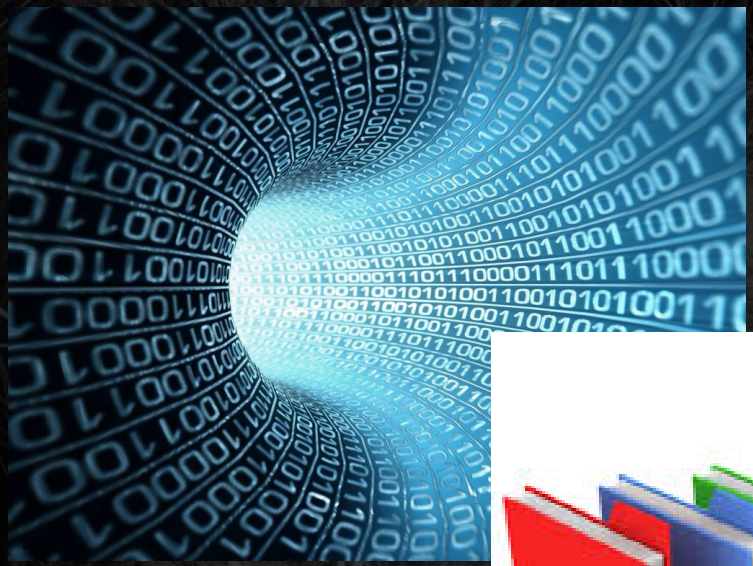
by precomputing:

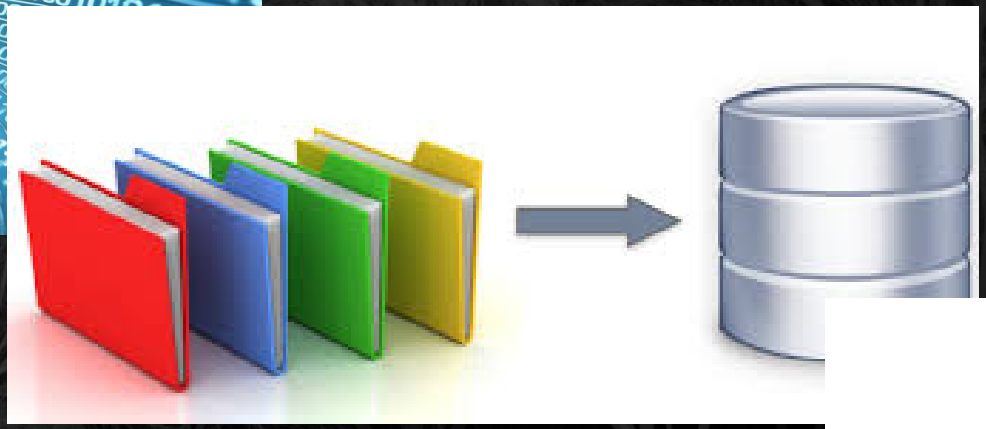
$$\boldsymbol{\delta}_N = \sum_{k=1}^{N} \boldsymbol{\delta}_k \qquad \mathbf{I_D}(z) = \int_{z-\Delta z/2}^{z+\Delta z/2} \mathbf{d}_j dz \qquad j = 1, 2, \ldots, m$$

$N(z)$ is reduce to a simple dot product

$$N(z) = \mathbf{I_D}(z) \cdot \boldsymbol{\delta}_N$$

How we can do it?



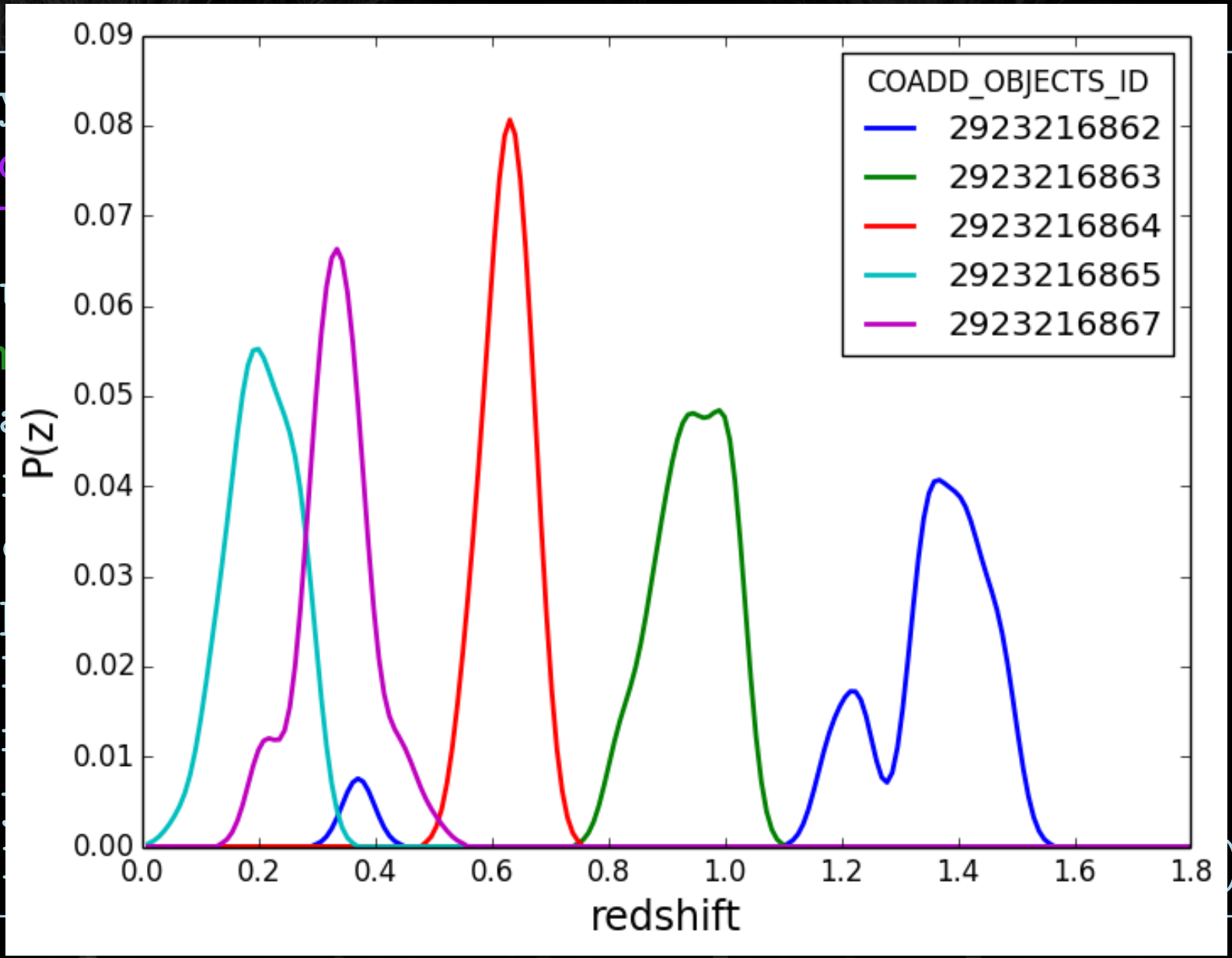


Getting some PDFs from DB

```
query="""
select COADD_OBJECTS_ID,TPZ from
PHOTOZ_PDF_SVA1_GOLD where rownum < 6"""
cc=cursor.execute(query)
#Handling and plot
df=ea.to_pandas(cc)
for i in xrange(5):
    cid=df.COADD_OBJECTS_ID.values[i]
    plt.plot(zbins,df.TPZ.values[i],
    lw=2,label=cid)
plt.xlabel('redshift',fontsize=17)
plt.ylabel('P(z)',fontsize=17)
plt.legend(loc=0, title='COADD_OBJECTS_ID')
```

Getting some PDFs from DB

```
query
sele
PHOT
cc=c
#Ha
df=e
for
plt.
plt.
plt.
```



Now using Sparse rep.

```
query="""
select COADD_OBJECTS_ID,PHZ.GET_PDF(TPZ) as
TPZ from PHOTOZ_SPARSE_SVA1_GOLD
where rownum < 6"""
cc=cursor.execute(query)
#Handling and plot
df=ea.to_pandas(cc)
for i in xrange(5):
    cid=df.COADD_OBJECTS_ID.values[i]
    plt.plot(zbins,df.TPZ.values[i],
    lw=2,label=cid)
plt.xlabel('redshift',fontsize=17)
plt.ylabel('P(z)',fontsize=17)
plt.legend(loc=0, title='COADD_OBJECTS_ID')
```

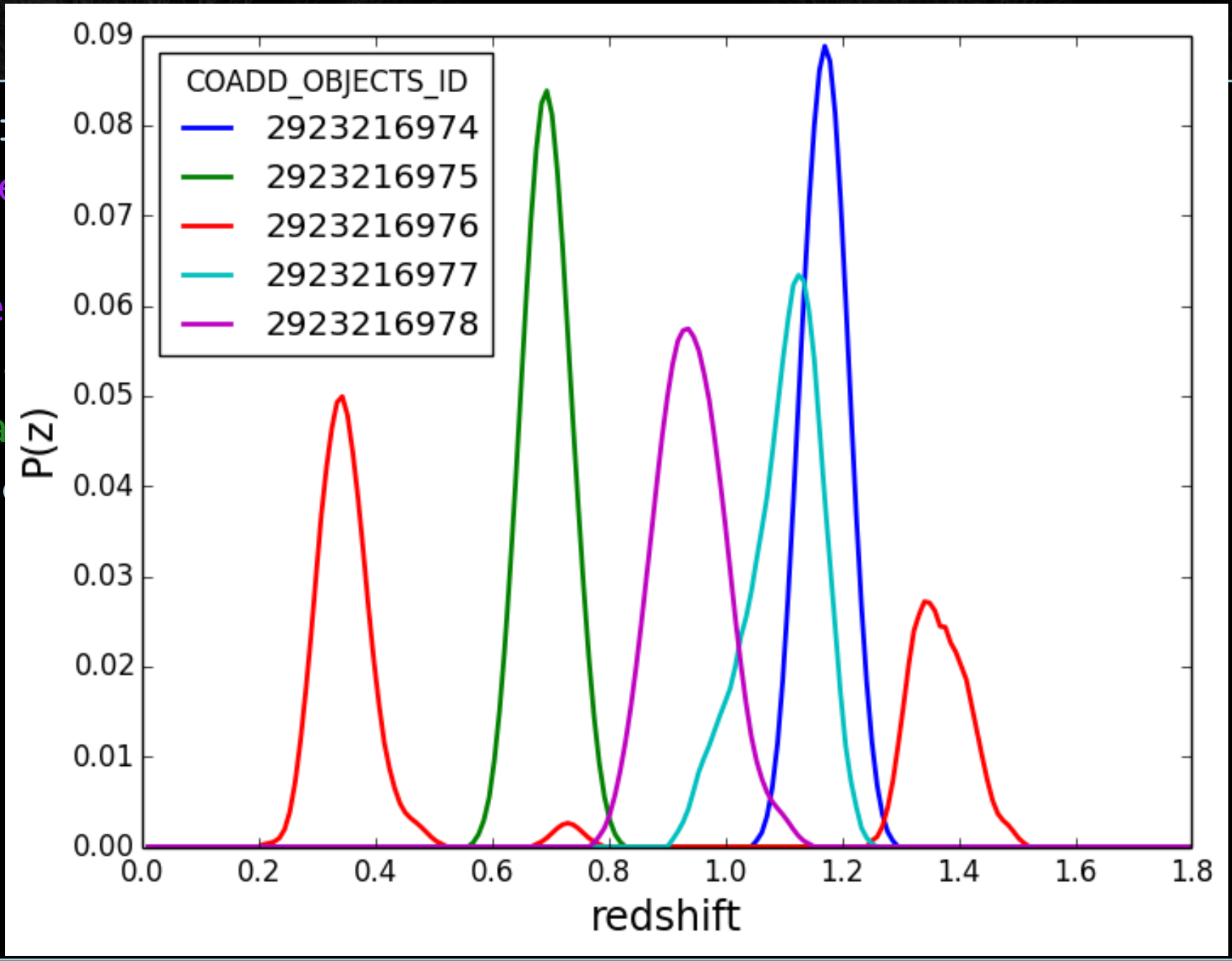
Now using Sparse rep.

```
query="""
select COADD_OBJECTS_ID,PHZ.GET_PDF(TPZ) as
TPZ from PHOTOZ_SPARSE_SVA1_GOLD
where rownum < 6"""
cc=cursor.execute(query)
#Handling and plot
df=ea.to_pandas(cc)
for i in xrange(5):
    cid=df.COADD_OBJECTS_ID.values[i]
    plt.plot(zbins,df.TPZ.values[i],
    lw=2,label=cid)
plt.xlabel('redshift',fontsize=17)
plt.ylabel('P(z)',fontsize=17)
plt.legend(loc=0, title='COADD_OBJECTS_ID')
```

Now using Sparse rep.

```
que
sel
TPZ
whe
cc=
#Ha
df=
for

plt
plt
plt
```

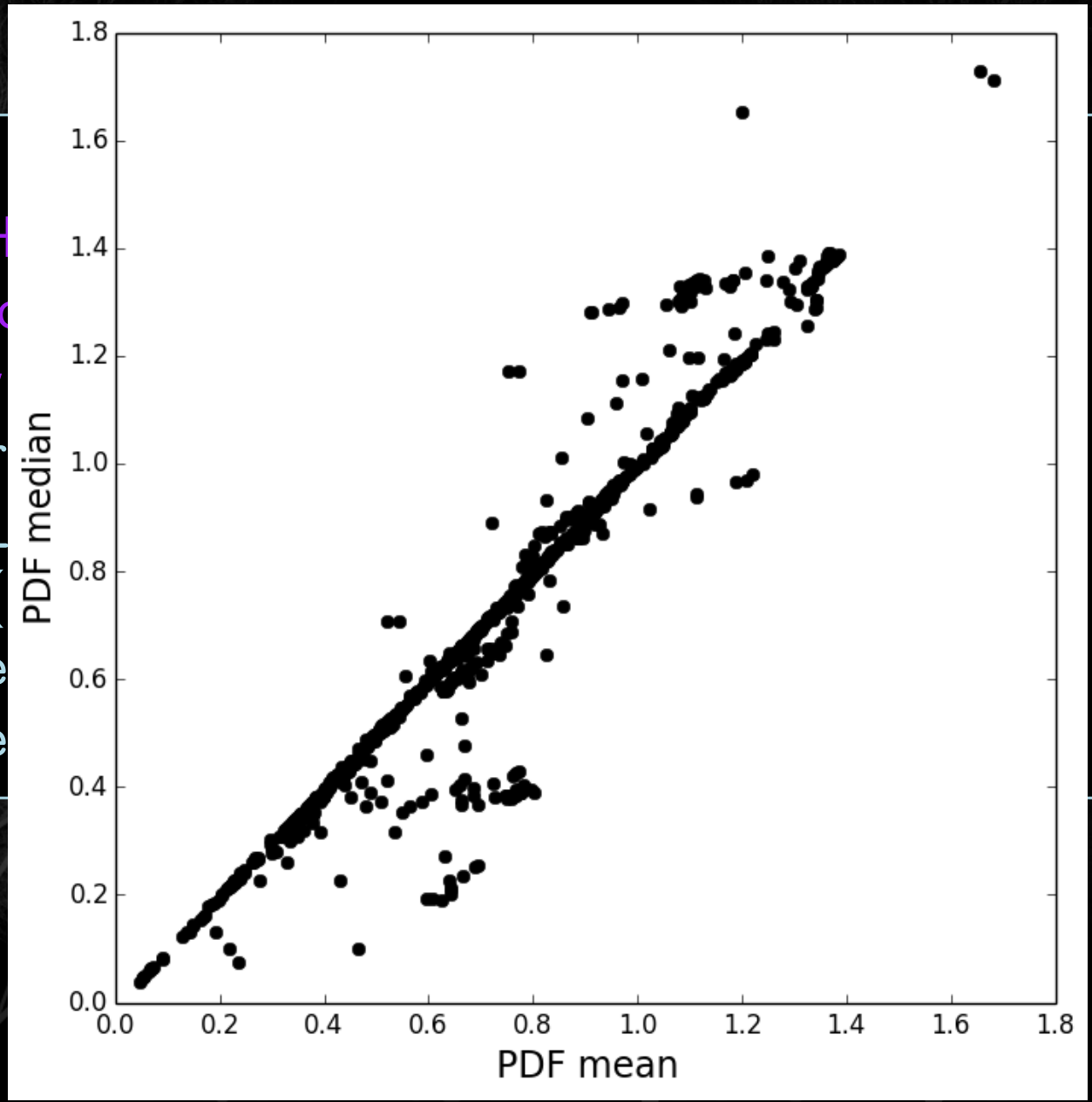


Getting metrics on the fly

```
query="""
Select PHZ.MEAN(tpz) mean, PHZ.MEDIAN(tpz)
median from PHOTOZ_PDF_SVA1_GOLD
where rownum < 1000"""
cc=cursor.execute(query)
df=ea.to_pandas(cc)
plt.plot(df.MEAN,df.MEDIAN,'ko')
plt.xlabel('PDF mean',fontsize=17)
plt.ylabel('PDF median',fontsize=17)
```

Getting metrics on the fly

```
query="""
Select PH                                    pz)
median fr
where row
cc=cursor
df=ea.to_
plt.plot(
plt.xlabe
plt.ylabe
```



Stacking PDFs in DB cluster!

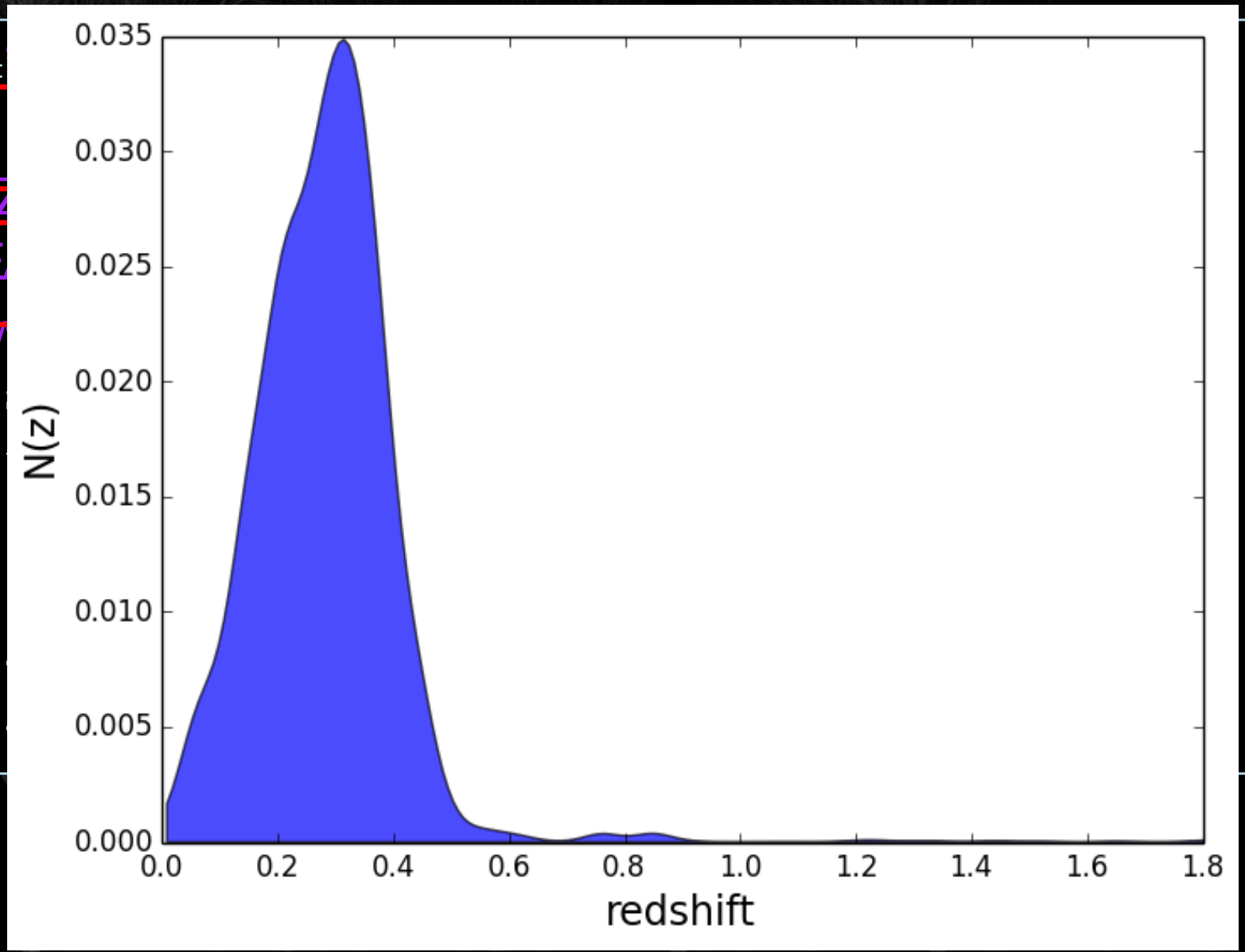
```
query="""
Select NZ(PHZ.TOTABLE(tpz)) as NZ from
PHOTOZ_PDF_SVA1_GOLD where
PHZ.MEAN(tpz) BETWEEN 0.1 and 0.4
and rownum < 100000"""
cc=cursor.execute(query)
df=ea.to_pandas(cc)
plt.fill_between(zbins,df.NZ.values[0],
     facecolor='blue',alpha=0.7)
plt.xlabel('redshift',fontsize=17)
plt.ylabel('N(z)',fontsize=17)
```

Stacking PDFs in DB cluster!

```
query="""
Select NZ(PHZ.TOTABLE(tpz)) as NZ from
PHOTOZ_PDF_SVA1_GOLD where
PHZ.MEAN(tpz) BETWEEN 0.1 and 0.4
and rownum < 100000"""
cc=cursor.execute(query)
df=ea.to_pandas(cc)
plt.fill_between(zbins,df.NZ.values[0],
     facecolor='blue',alpha=0.7)
plt.xlabel('redshift',fontsize=17)
plt.ylabel('N(z)',fontsize=17)
```

Stacking PDFs in DB cluster!

```
query=
Select
PHOTOZ
PHZ.ME
and rov
cc=cur
df=ea.
plt.fi
      f
plt.xl
plt.yl
```



Conclusions

✓ Compute photo-z PDF

Individual techniques (MLZ; arXiv:1303.7269, arXiv:1312.5753)

✓ Combine PDFs efficiently

Better than individual, outliers identification (arXiv:1403.0044)

✓ PDF Sparse Representation

99.9% accuracy in P(z) and N(z) with 15 points (arXiv:1404.6442)

✓ Uses of these tools!

Clustering, weak lensing, DES, DESDM, etc...

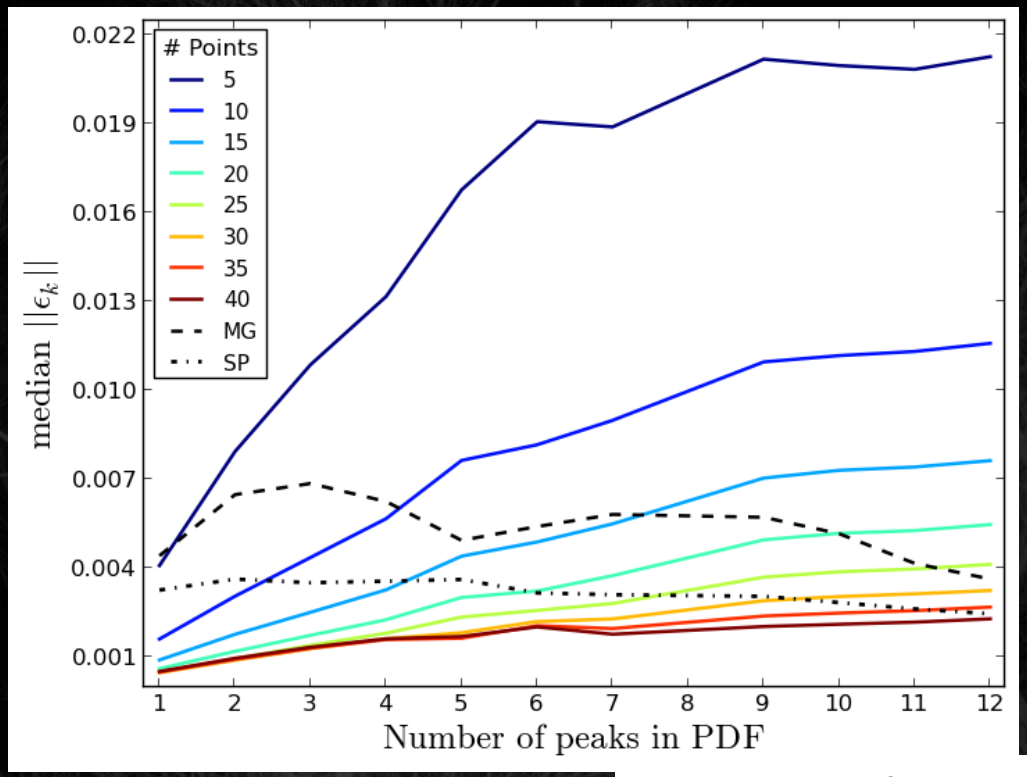
THANKS!

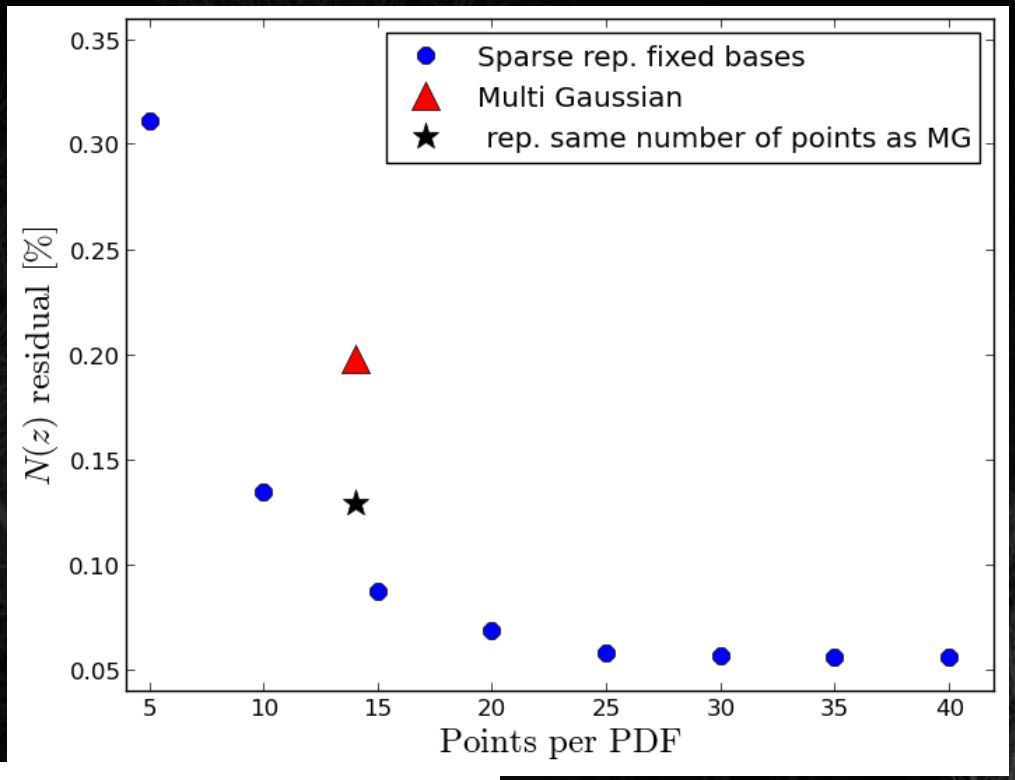
Questions?

Matias Carrasco Kind
NCSA/UIUC
mcarras2@ncsa.illinois.edu
http://matias-ck.com/
https://github.com/mgckind

EXTRA SLIDES

Photo-z PDF storage: Results



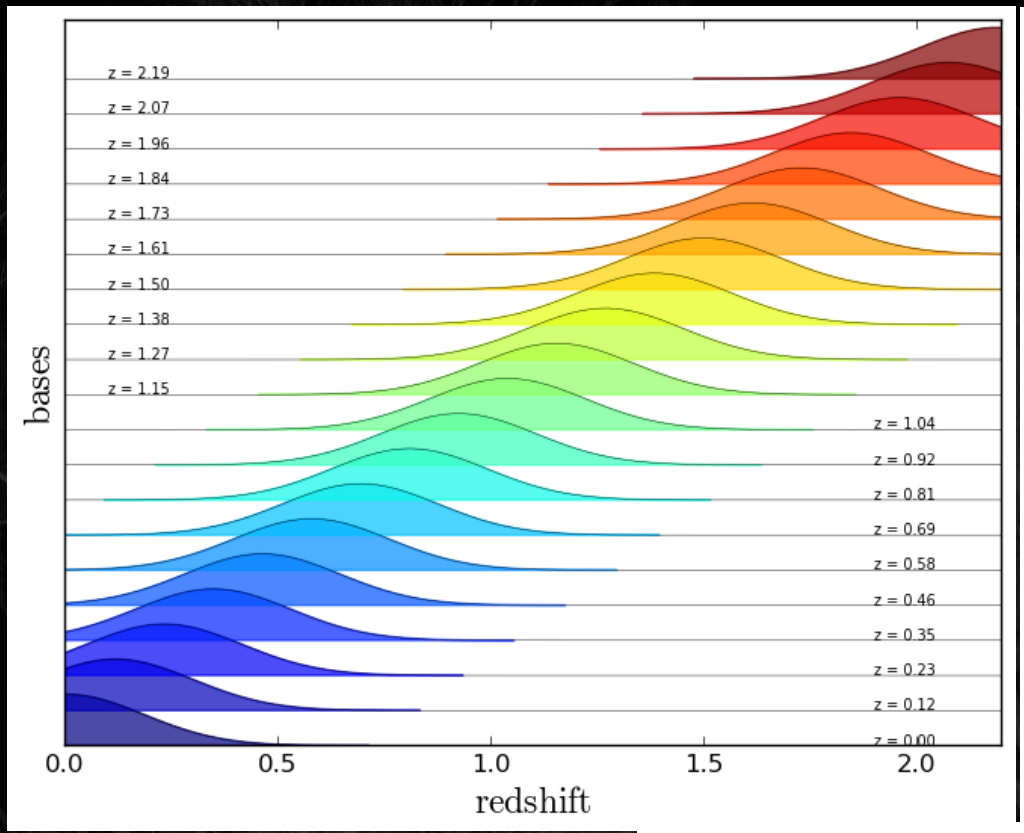


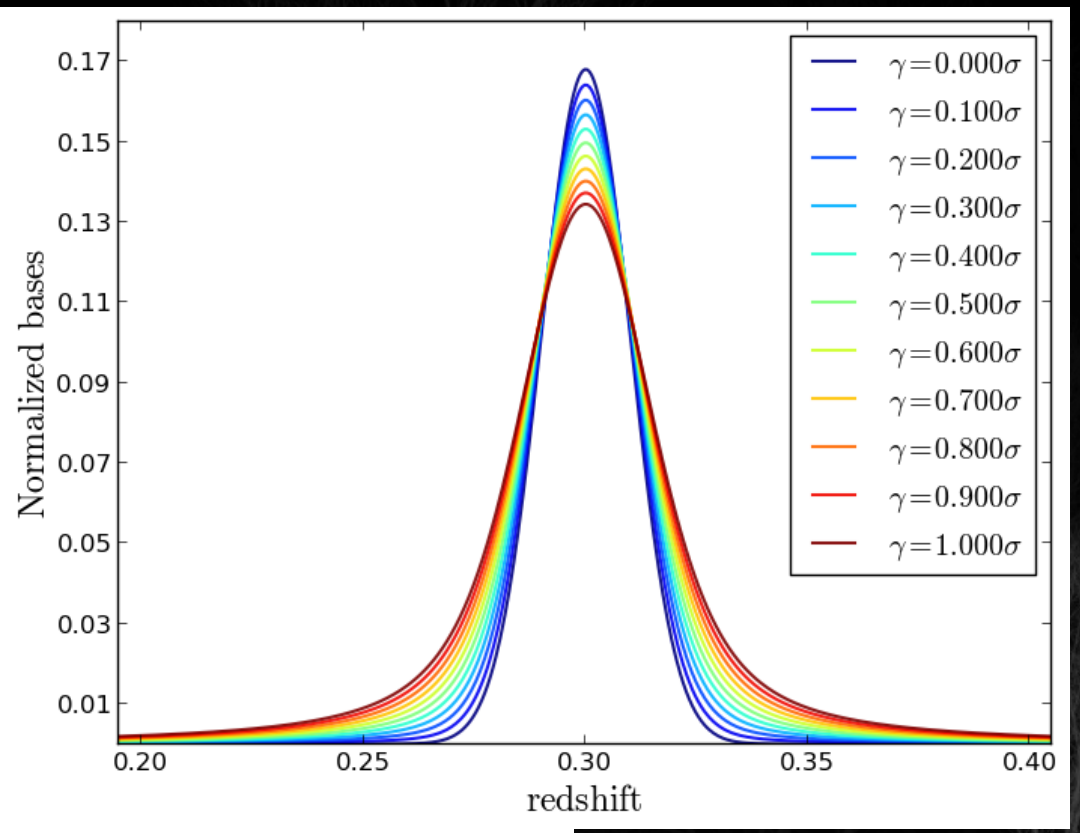
Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)

For PDFs with less than 4 peaks 5-10 points should be sufficient

Sparse representation gives more accurate and more compressed representation for N(z), 99.9% accuracy with 15 points (200 points originally)

Photo-z PDF storage: Dictionary





Carrasco Kind & Brunner 2014b (MNRAS, 441, 3550)

Combination of Gaussian and Voigt profiles

Covering the whole redshift space, at each location we have several bases

Photo-z PDF estimation: Error and validation

Out of Bag (cross-validation) data used to validate trees/maps

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

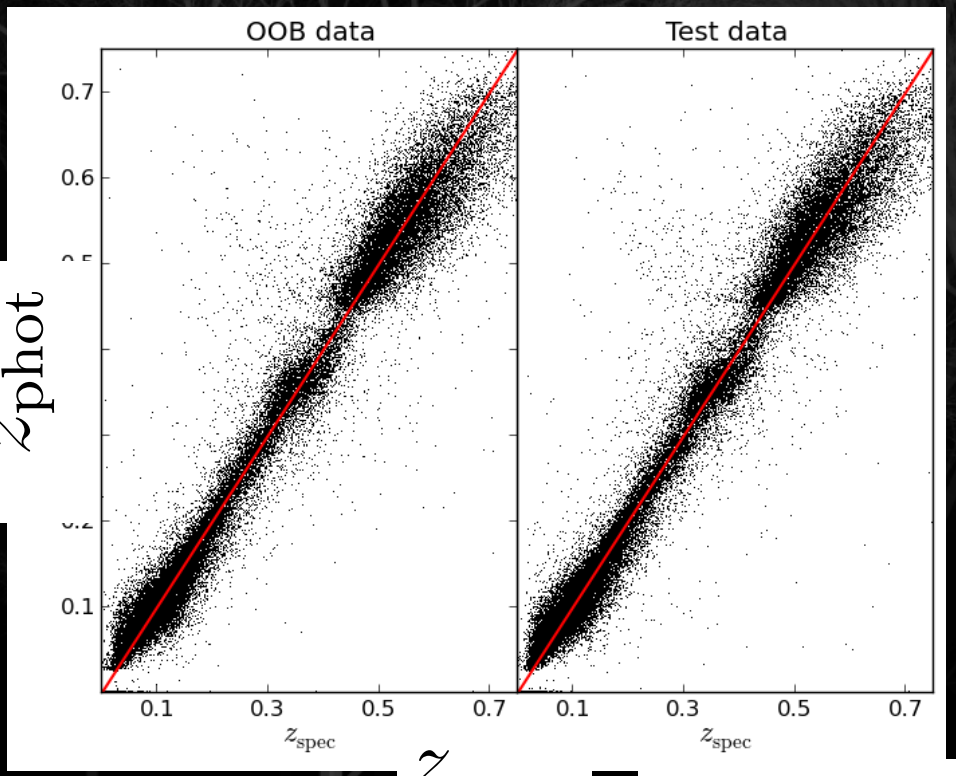
We can learn from the cross-validation data!

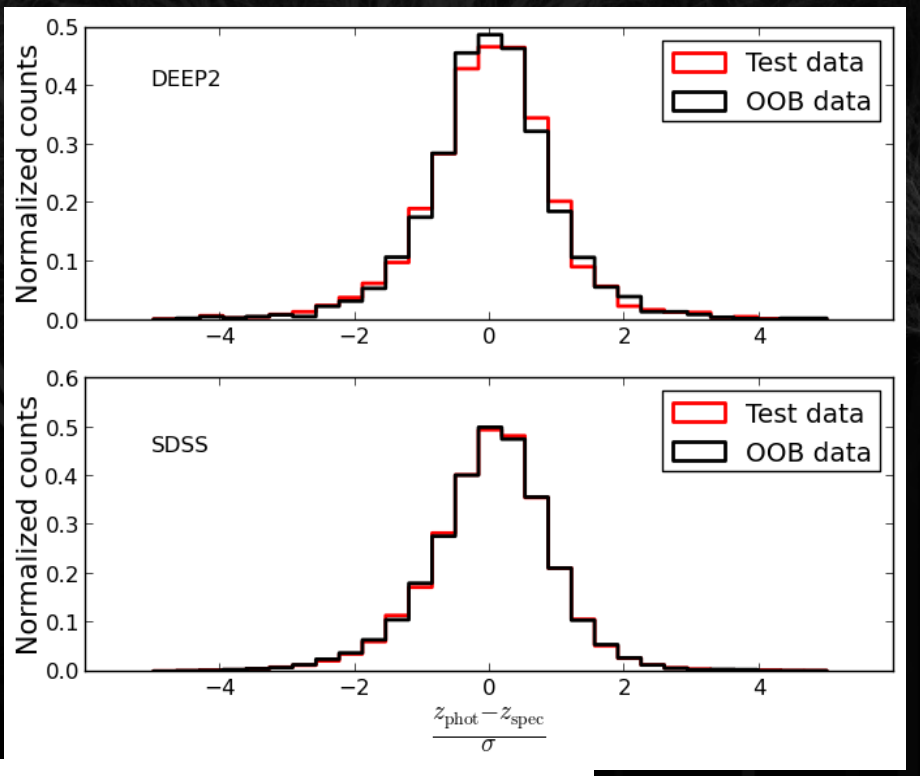
Photo-z PDF estimation: Error and validation

Out of Bag (cross-validation) data used to validate trees/maps

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

We can learn from the cross-validation data!





Carrasco Kind & Brunner 2014c (MNRAS, 442, 3380)

Photo-z PDF estimation: SOM 2D toy example

Suppose 2D data distributed in a given space

De-project the data in a 2D map

Each cell will contain objects with similar properties

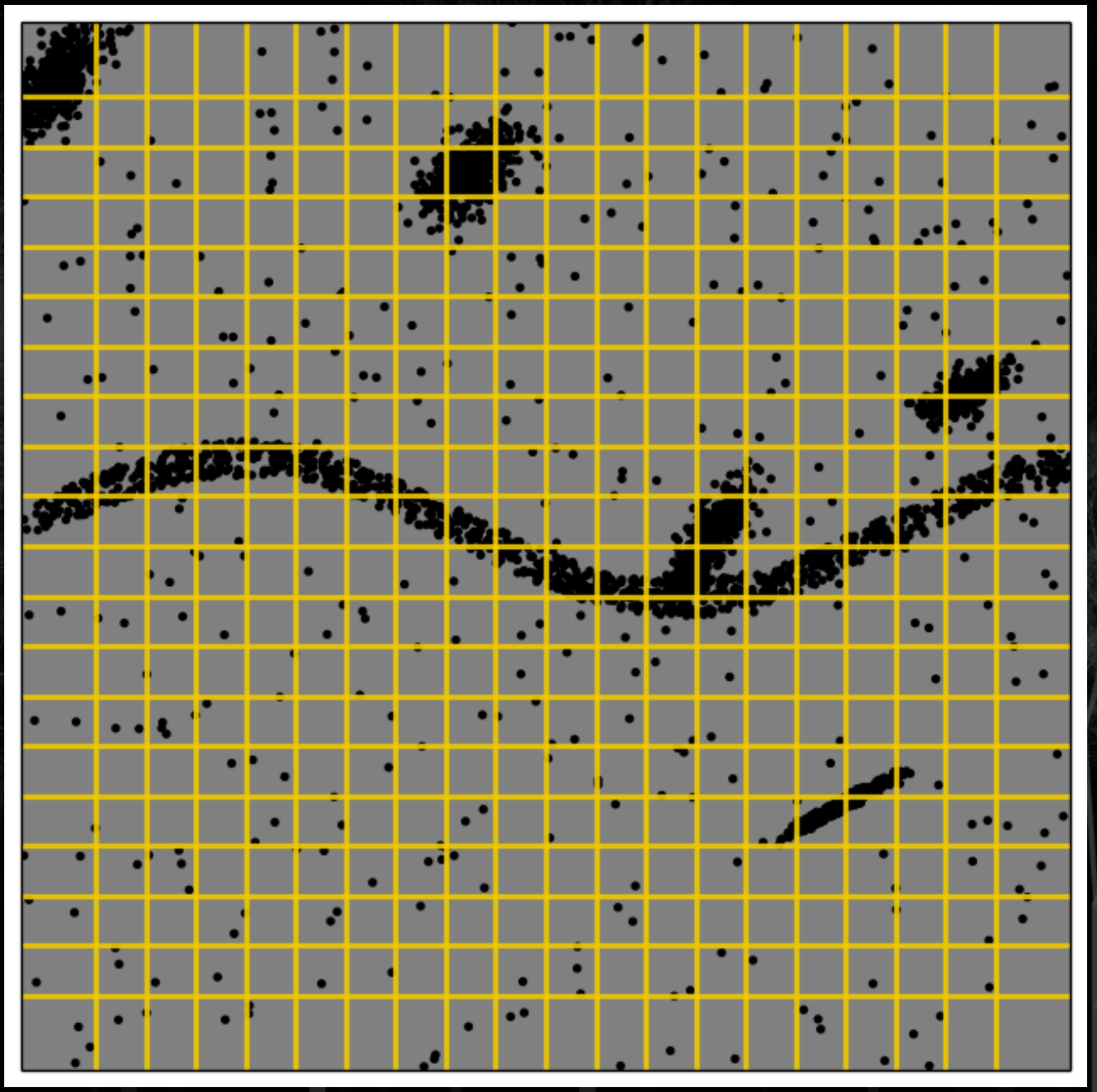


Photo-z PDF estimation: SOM 2D toy example

Suppose 2D data distributed in a given space

De-project the data in a 2D map

Each cell will contain objects with similar properties