# METAPHOR

## Machine learning Tool for Accurate PHOtometric Redshifts

**V. Amaro**[1], S. Cavuoti[2]
M. Brescia[2], C. Vellucci[1], G. Longo[1]

1 - University of Napoli Federico II, Naples
2 - INAF- Astronomical Observatory of Capodimonte, Naples

# Summary

*Cavuoti et al., 2016, Submitted to MNRAS*

1. The METAPHOR structure and workflow

1. Testing METAPHOR on SDSS-DR9 data
    a. General applicability (tested on 3 interpolative methods - MLPQNA, KNN and RF)
    b. Comparison with one SED template fitting (LePhare method) just as benchmark

1. Deriving PDF's and evaluation of the performance

1. Preliminary testing on ESO KiDS public DR2

# Photo-z PDFs for Machine Learning are still an open issue

A reliable PDF should be able to:

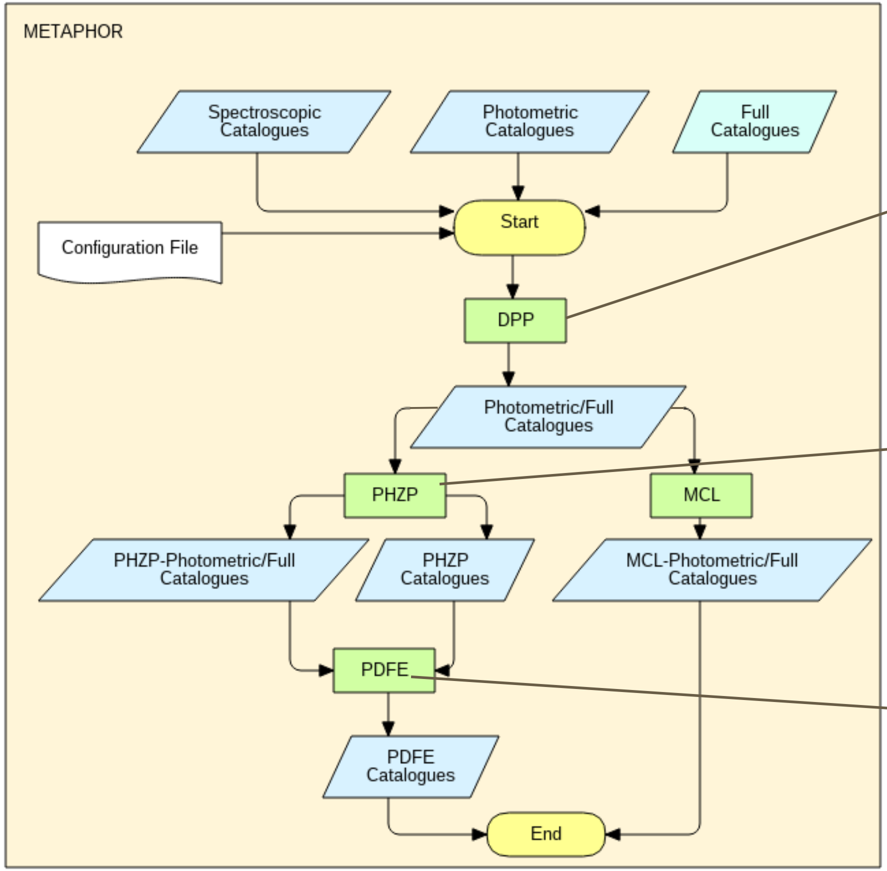
1) evaluate photometric error distributions;
2) assess the correlation between spectroscopic and photometric Errors;
3) disentangle photometric uncertainties from those intrinsic to the method itself.

Many PDF methods for ML developed over the past years, mostly based on:

- Supervised methods (ANN, RF, MLP, used both as regressors and classifiers)
- Unsupervised methods (SOMs, random atlas)

*Rau et al. 2015, MNRAS, 452*
*Carrasco & Brunner 2013, MNRAS, 432*
*Carrasco & Brunner 2013, MNRAS, 438*
*Carrasco & Brunner 2013, MNRAS, 442*
*Bonnet 2013, MNRAS, 449*
*Sadeh et al. 2015, arXiv:1507.00490*
*Speagle et al. 2015, arXiv:1510.08073*

# METAPHOR workflow

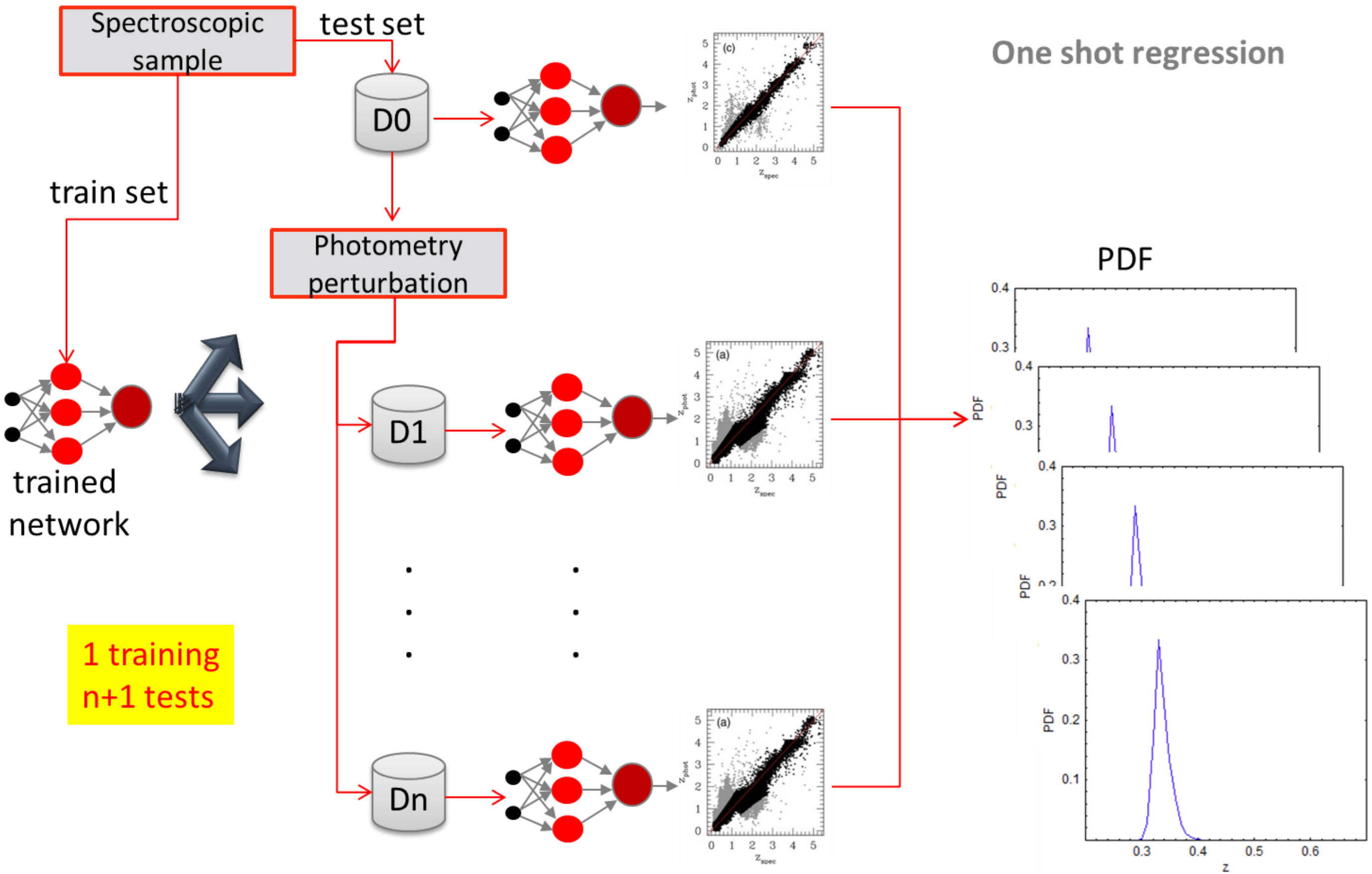



**Data pre-processing:**
KB preparation
photometry perturbation

**Photo-z Prediction:**
train/test phase by means of any arbitrary ML model (in our case the multi-threading MLPQNA)

**Probability Density Function Estimation**

# METAPHOR



Spectroscopic sample
test set
train set
D0
Photometry perturbation
D1
Dn
trained network
1 training
n+1 tests
One shot regression
PDF

# PDF estimation algorithm

After one training on the not perturbed training set and having produced N perturbations of the blind test set:

Submit to the network, N+1 test sets (N perturbed + original one) thus obtaining N+1 estimates of photo-z's;

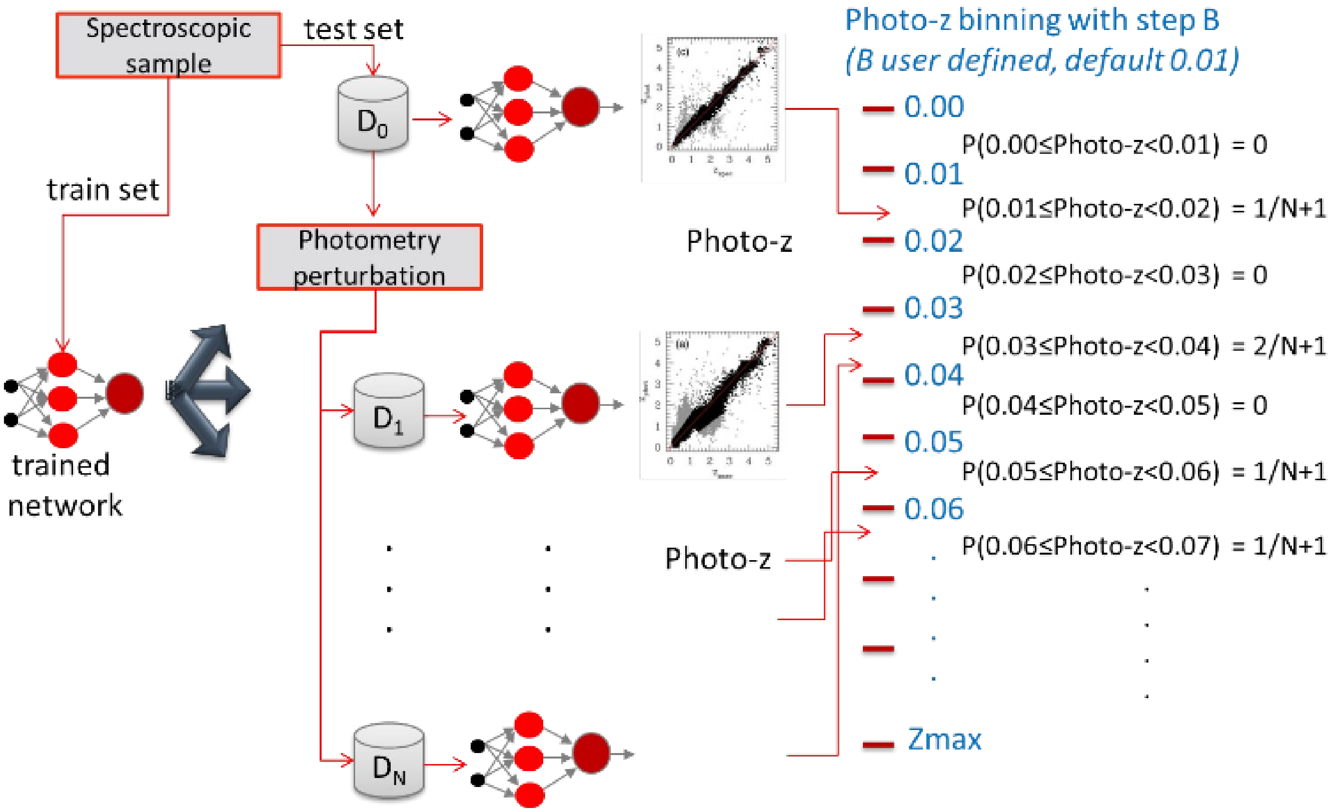
Binning in photo-z (according to the chosen precision). Let us call "B" the bin;

Calculate the number of photo-z's for each bin: said it C, then calculate the relative probability as P:

$$C_{B,i} \in [Z_i, Z_{i+B}[ \qquad\qquad P(Z_i \leq \text{Photo-z} \leq Z_{i+B}) = C_{B,i}/(N+1)$$

Calculate statistics for the resulting PDFs (the set of all probabilities obtained at the previous step)

# PDF estimation algorithm scheme



# The photometry perturbation procedure

The **perturbation procedure** consists of two steps:

**Photometry error estimation**: for which the basic idea is that a binning of photometric bands in which a polynomial fitting of the mean error values should be able to reproduce the intrinsic trend of the inner distribution

**Issue: right choice of the bin amplitude in order to minimize the risk of information losses (aliasing, masking), somehow overcome by:**

**Photometry perturbation**:

Variable Gaussian noise added to photometry, weighted by the polynomial fit;

Parametrization of the method through the use of a different multiplicative constant for each band in order to ensure *flexibility* ( different choice of bands and catalogues, different quality of photometry).

# Photometry error estimation

Polynomial fitting **steps**:

1. Binning the chosen band;
2. Extract statistical errors (μ, σ) from each bin;
3. Perform polynomial fitting with specified order;
4. Compare the fit to σ distributions to verify that for each bin the fitting error tolerance is within 1σ: generation of a boolean flag (True in the case that all the bins fulfill this condition; False otherwise);
5. If the quality flag is set to False, then increase the polynomial degree and go to step 4

# The first METAPHOR test data (SDSS DR9)

| Band | lower limit | upper limit |
|---|---|---|
| u | 17.0 | 26.8 |
| g | 16.0 | 24.9 |
| r | 15.4 | 22.9 |
| i | 15.0 | 23.3 |
| z | 14.5 | 23.0 |

We used a sample of the SDSS-DR9 spectroscopic catalogue, prepared as follows:

Objects classified as galaxies with the specClass flag "*galaxy*";

*psfMag* type magnitudes and relative errors;

Removed missing detections in any of the five SDSS photometric bands (NaN entries);

Selected objects with PhotoFlags ≠ 0 ( thus removing objects that could not be real, or with suspicious deblending or with photometry affected by cosmic rays or bleed trails);

The final KB consisted of ~50,000 training and ~100,000 test set objects

Used features: 5 mags: u, g, r, i, z and 4 derived colours

**MLPQNA has been successfully tested as a classifier/regressor in a variety of scientific cases**

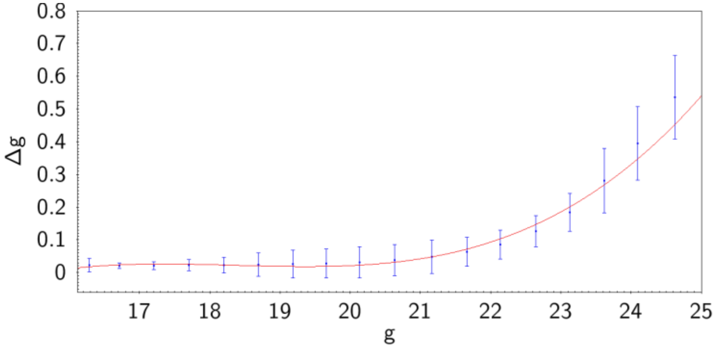
*Brescia et al. (2012, MNRAS, 421; 2013, ApJ, 772; 2014, PASP, 126)*

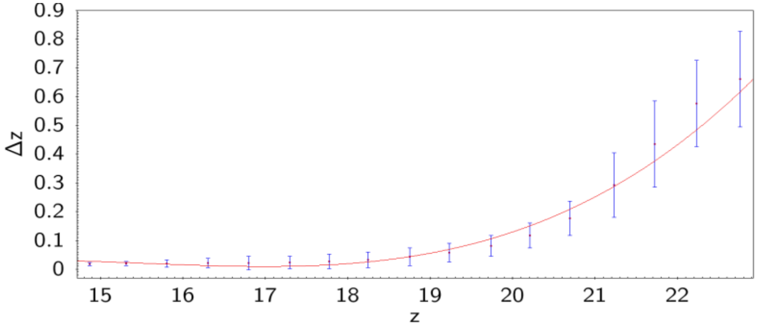
*Cavuoti et al. (2012, A&A, 546; 2014, MNRAS, 437; 2014, IAU Symposium, Vol. 306; 2015, MNRAS, 452; 2015, Exp. Astronomy, Springer, Vol.39; 2016, A Cooperative approach among methods for photometric redshifts estimation: an application to KiDS data.* Submitted to MNRAS);

**In the case of SDSS-DR9 data we already produced a photo-z catalogue for ~143 million galaxies**

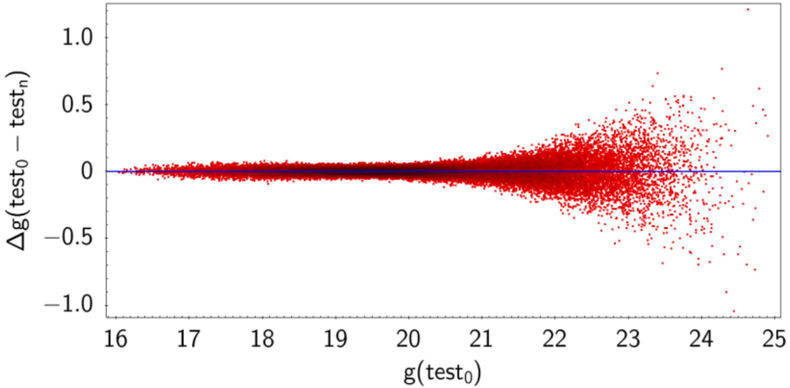
*Brescia et al., 2014, A&A, 568 + VizieR On-line Data Catalog:J/A+A/568/A126*

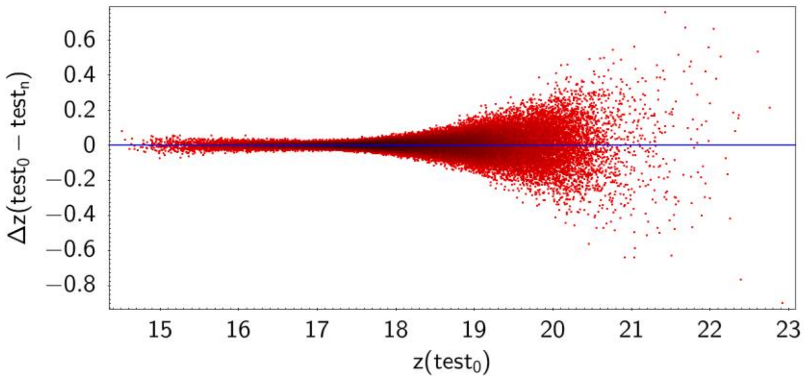
# Photometry perturbation examples (for SDSS DR9)




$$m_{ij}^{*}=m_{ij}+a_i f_i(m_{ij})*\mathrm{gaussRandom}_{[-1,1]}$$



# Results for photo-z's and stacked PDF

**Statistics for the residuals $\Delta z=(z_{spec}-z_{phot})/(1+z_{spec})$:**

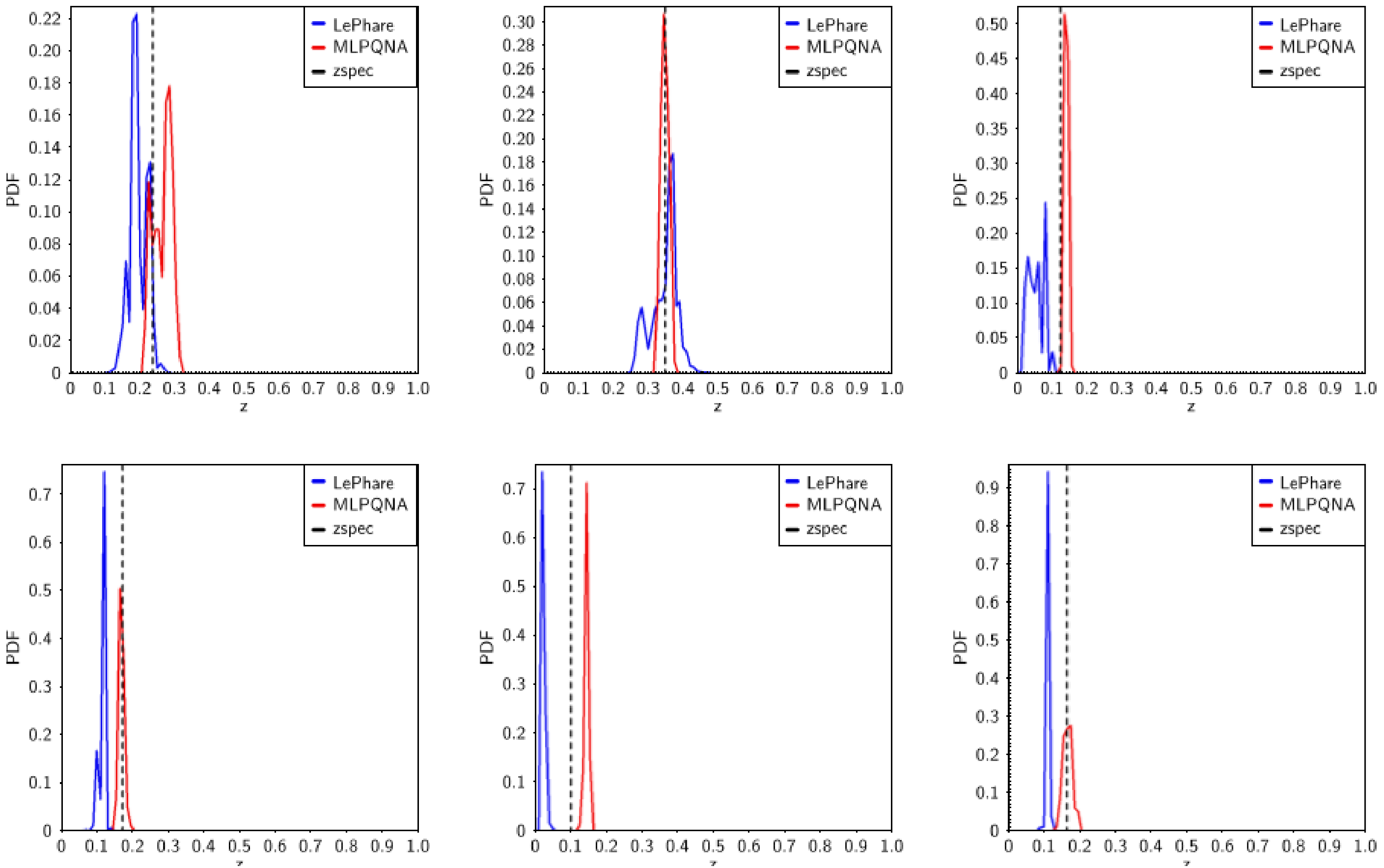
**Comparing MLPQNA to KNN, RF and *Le Phare***

| Estimator | MLPQNA | KNN | RF | *Le Phare* |
|---|---|---|---|---|
| *bias* | 0.0063 | 0.0029 | 0.0035 | 0.0009 |
| $\sigma$ | 0.024 | 0.026 | 0.025 | 0.060 |
| $\sigma_{68}$ | 0.020 | 0.020 | 0.019 | 0.035 |
| *NMAD* | 0.017 | 0.018 | 0.018 | 0.030 |
| *skewness* | 0.075 | 0.330 | 0.015 | −18.076 |
| *outliers* > 0.15 | 0.12% | 0.15% | 0.15% | 0.69% |

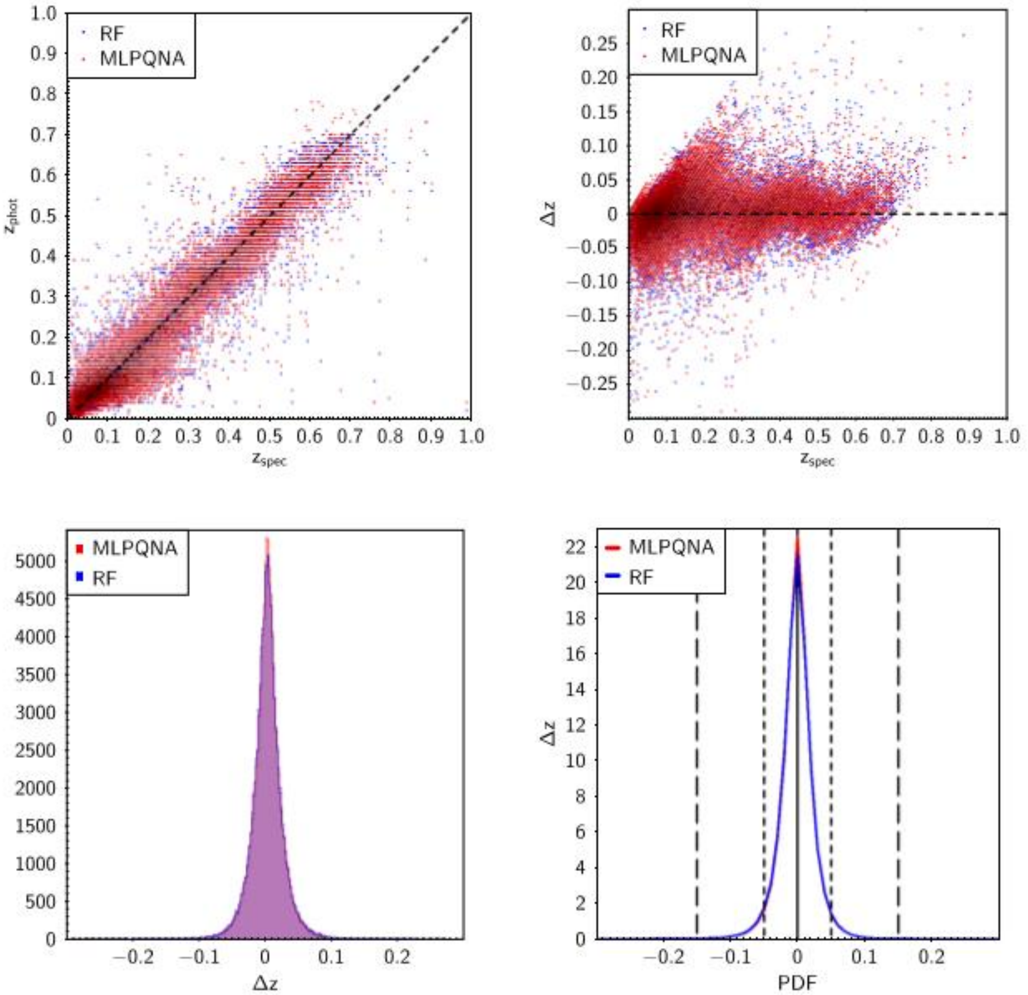
| Estimator | MLPQNA | KNN | RF | *Le Phare* |
|---|---|---|---|---|
| $f_{0.05}$ | 92.5% | 92.0% | 92.1% | 71.2% |
| $f_{0.15}$ | 99.7% | 99.8% | 99.7% | 99.1% |
| $\langle \Delta z \rangle$ | −0.0014 | −0.0018 | −0.0016 | 0.0131 |

- Statistics of interpolative methods are comparable;
- Le Phare skewness is ~200 times more asymmetric;
- On the $f_{0.05}$ for stacked PDFs, interpolative methods reach best performance.

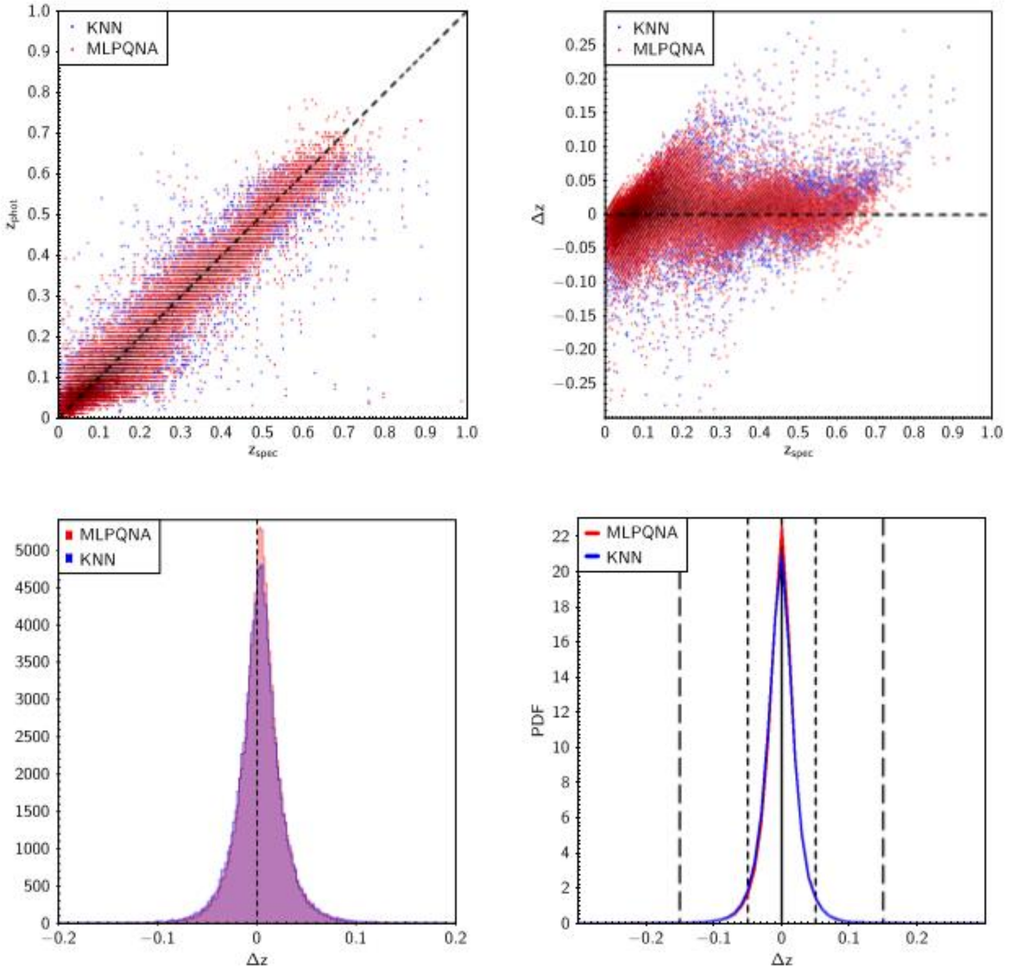
# Individual PDFs: some examples



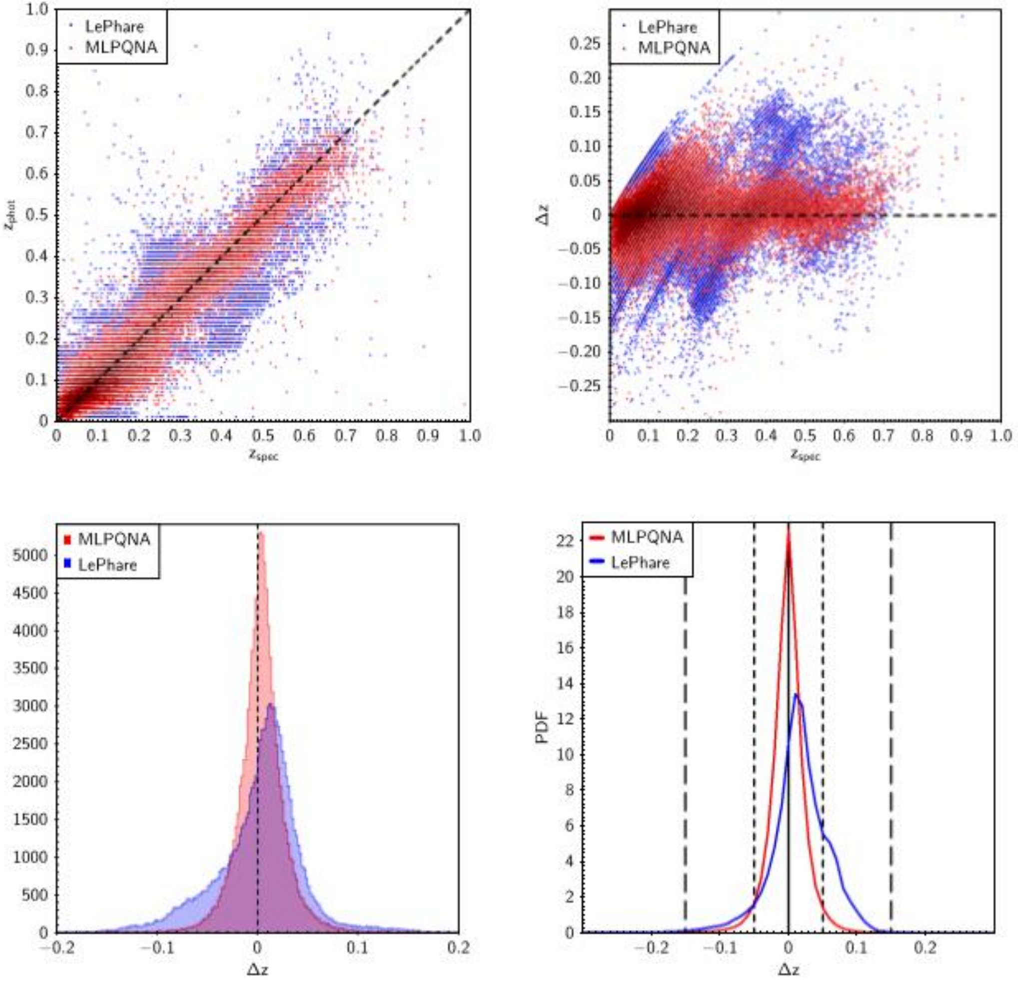
# MLPQNA vs RF



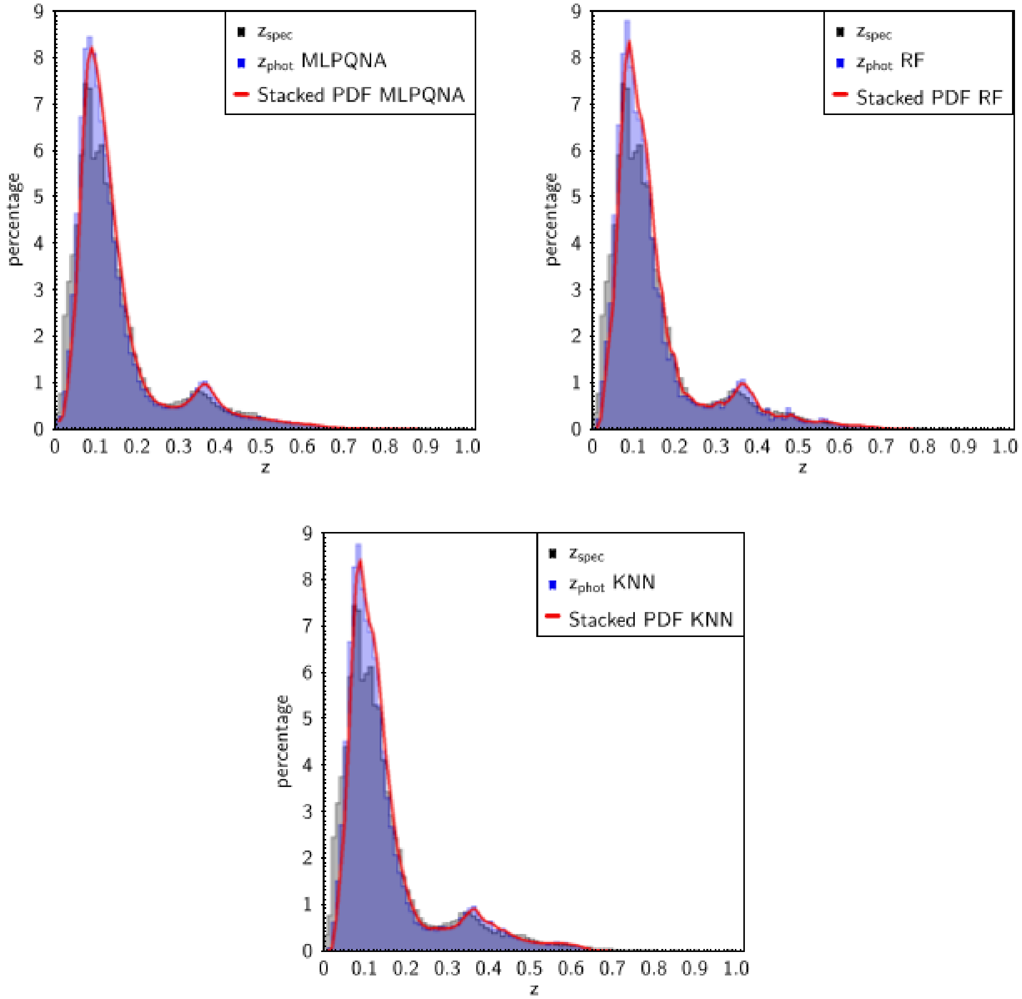
# MLPQNA vs KNN



# MLPQNA vs Le Phare



# Stacked PDFs vs $z_{spec}$ distributions



# METAPHOR in Euclid OU-PHZ Data Challenge #2

METAPHOR and MLPQNA are used within the Organization Unit for photo-z's in the ESA Euclid Mission. Several internal contests are under development, whose goal is to select the most suitable methods to derive official photo-z PDFs for top and legacy science.

2 catalogues used (based on zCosmos and simulations) by splitting one field on RA:

- A calibration catalogue, without RA and DEC, but with spectroscopy (training set) and photometry;
- A verification catalogue, without zspec's but with RA, DEC and photometry.

Data pre-processing on the training set:

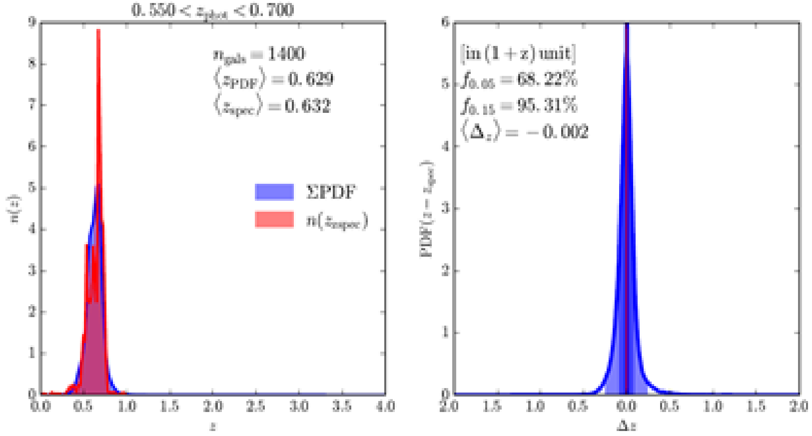
- Reliable zspec's based on provided quality flags (Salvato+2016 in prep.);
- Corrected magnitudes (depths within 5σ; Mag err < 1);
- Application of some photometric prescriptions;
- 8 optical/NIR photo-bands available: g, r, i, z, VIS, Y, J, H;
- Features used: 17 = 8 bands + 9 colours, i.e. g-r, r-I, i-z, z-Y, Y-J, J-H, VIS-Y, VIS-J, VIS-H;

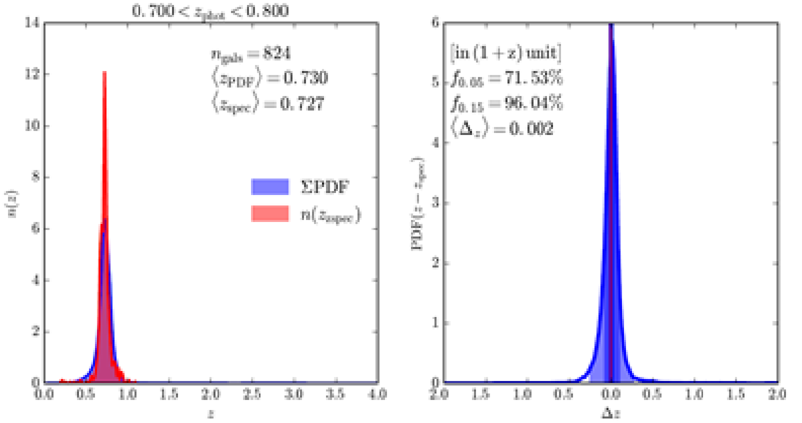
**The final KB consisted of 8,234 training and 3,535 blind test set objects (random split 70% / 30%)**

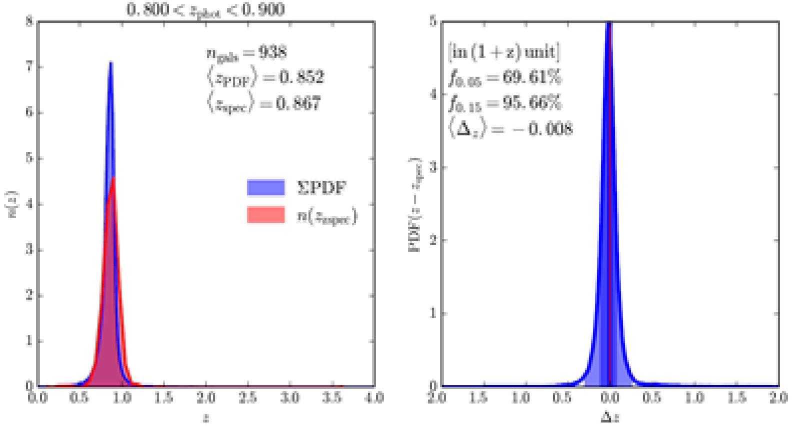
**Euclid requirements: f0.05 ≥ 68%, f0.15 ≥ 90%, <Δz> = 0.002**

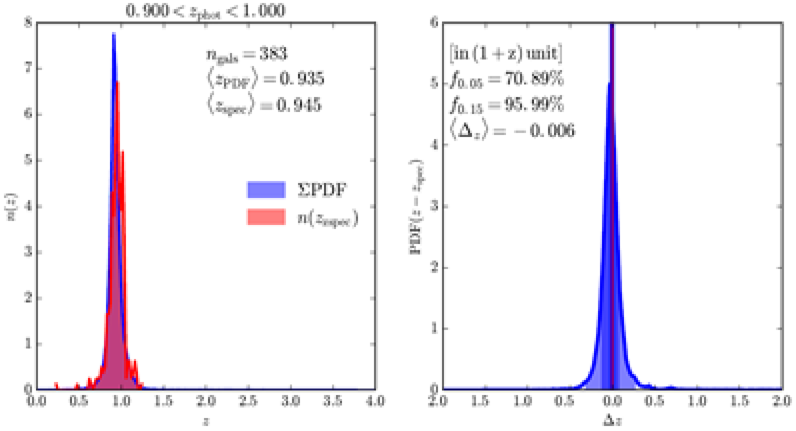
# EDC2 results with METAPHOR

*Courtesy J. Coupon and EUCLID OU-PHZ Team*

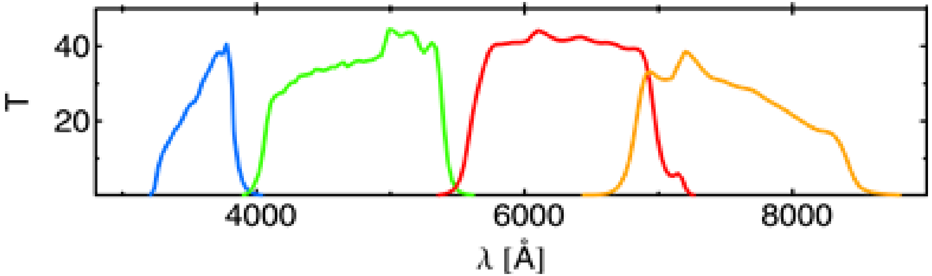


# METAPHOR in the ESO KiDS public data (DR2)




**KiDS DR2** (*de Jong et al. 2015*)**, *griz* photometry with SDSS and GAMA spectra as KB**

excluded objects with low photometric quality (i.e. with flux error > 1 magnitude);

removed objects having at least one missing band

selected objects with zero IMA-FLAGS in the **g**, **r** and **i** bands (i.e. sources flagged as located in proximity of saturated pixels, star haloes, image border or reflections, or within noisy areas). The **u** band was not considered in such selection since its masked regions are less extended than in the other three KiDS bands.

**The final KB consisted of 15,180 training and 10,067 blind test objects**

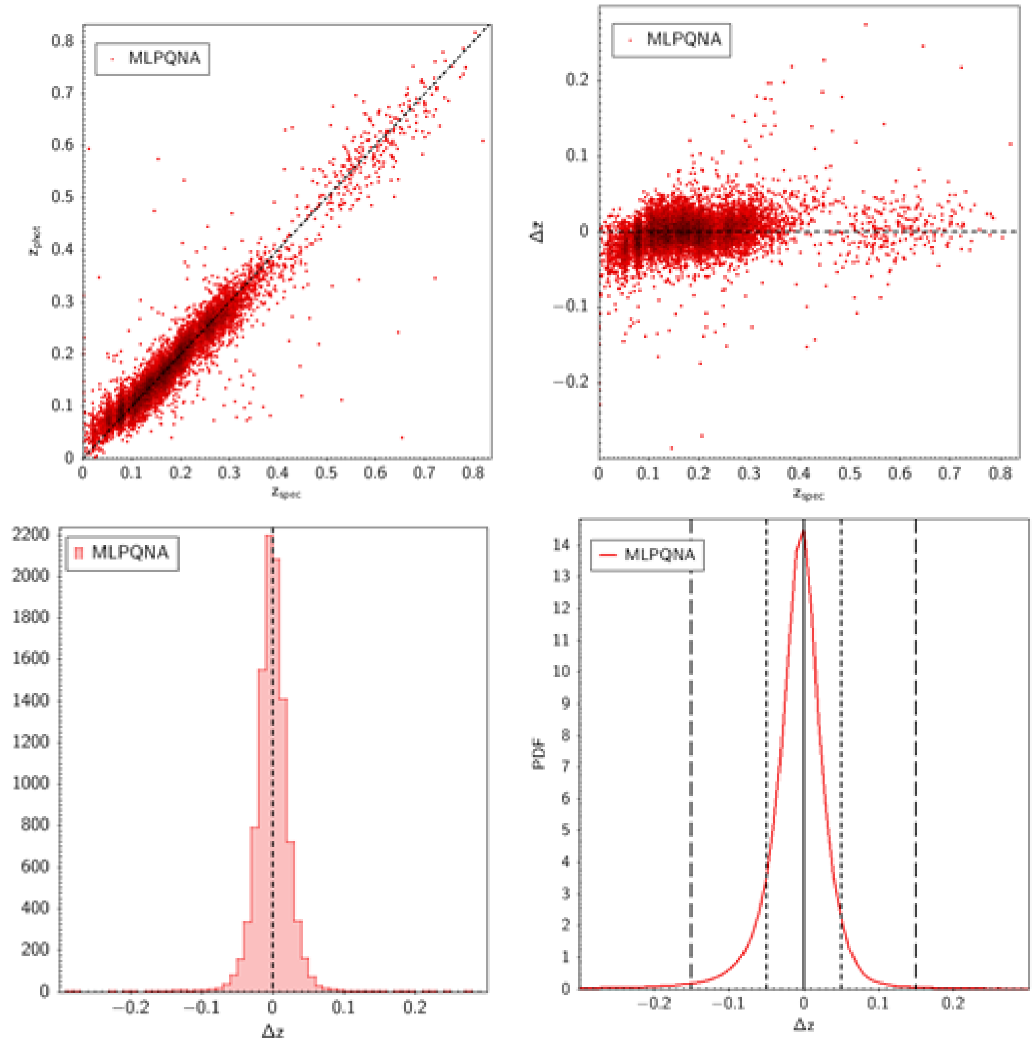
**In the case of KiDS-DR2 data we already produced a photo-z catalogue for ~1 million galaxies**

*Cavuoti, S., Brescia, M., Tortora, C., et al., 2015, MNRAS, 452*

# KiDS results with METAPHOR

| Estimator | MLPQNA |
|---|---|
| mean | −0.0007 |
| sigma | 0.026 |
| sigma68 | 0.018 |
| NMAD | 0.018 |
| outliers>0.15 | 0.31% |

| Estimator | MLPQNA |
|---|---|
| $f_{0.05}$ | 81.3% |
| $f_{0.15}$ | 98.4% |
| $\langle \Delta z \rangle$ | −0.0084 |



# Conclusions

METAPHOR can be applied with any empirical photo-z estimation model. It is able to take into account photometric errors due to both measurements and method itself;

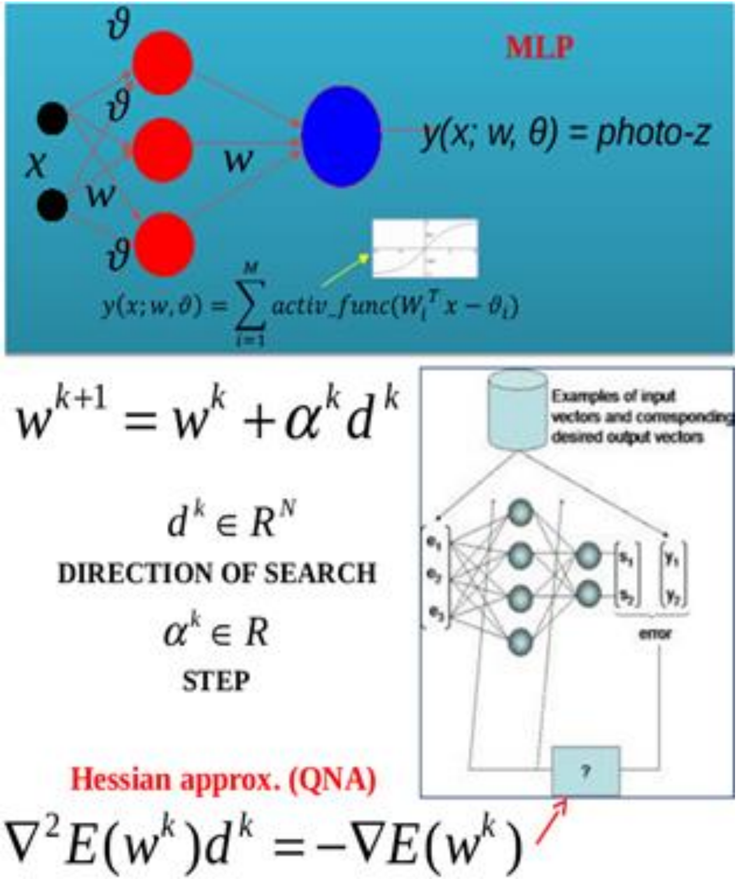
It is one of the candidate tools for the production of photo-z PDFs within the OU-PHZ of the ESA Euclid Mission, where not yet public data challenges are still in progress. Results of last challenge internally circulated 3 days ago, confirming promising performance of METAPHOR;

Highlights:

Empirical methods perform very well in $f_{0.05}$ and in PDF symmetry;

The stacked distributions of ML PDFs are almost indistinguishable from the distribution of spectroscopic redshifts;

# MLPQNA - Multi Layer Perceptron + Quasi Newton



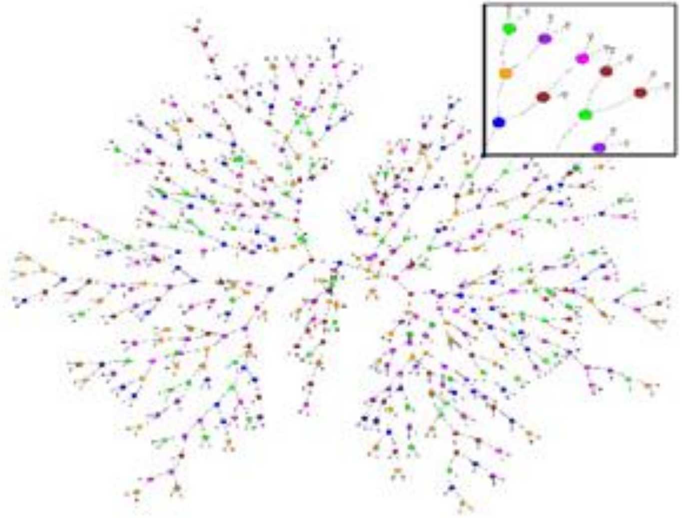

- Multi Layer Perceptron with feed-forward trained by Quasi Newton Learning rule
- It allows to find the stationary points of a function by approximating the Hessian matrix of the training error through a cyclic gradient calculation

*Brescia et al., 2013, ApJ, 772, 140; 2014, PASP, 126, 942*

# Random forest

Supervised method able to learn by creating a *random forest* (bootstrap, replica with replacement of objs of the train set) of *decision trees* (classifiers or regressors);

The split in branches of the original node, encompassing all the training set, proceeds recursively along the feature that maximize the information about the classes quared errors (regressors).



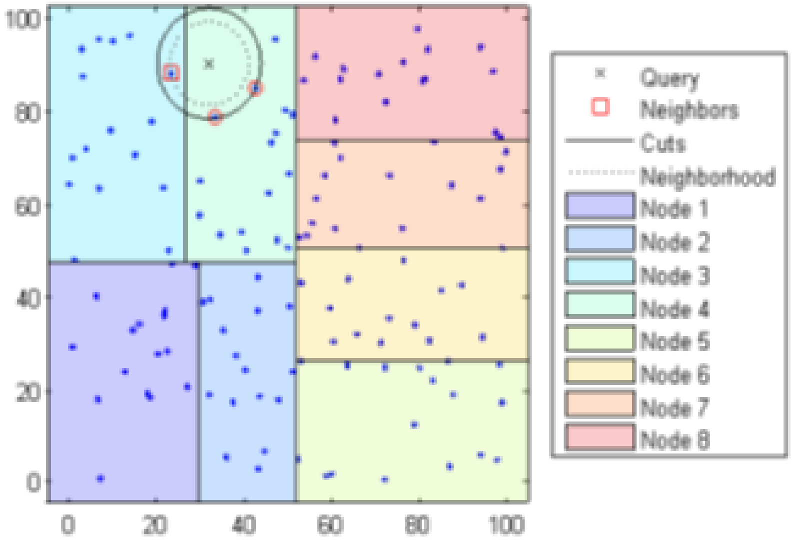
- The splitting proceeds until a terminal leaf node is created, matching an a priori defined stopping criterion;
- The objects in the terminal nodes are thus characterized by having same data properties;
- The photo-z estimation is the mean of all the bootstrapped replica of the objs in the terminal leaves.

*Breiman, 2001*

# K-Nearest Neighbours

- Given the N neighbors in the training set, for each obj of the test set the photo-z estimation is obtained by the mean of the N neighbors targets

- The neighborhood is the Euclidean distance among the features of the parameter space




*Cover & Hart (1967)*

# Le Phare SED Template Fitting

Observed magnitudes are matched with those observed by a set of SED models;
SED templates are red-shifted in step of Δz ( e.g. 0.01) and convolved with the filter transmission curves;
At the end the photo-z's are found, by minimizing the chi-squared:

$$\chi^2(z,T,A) = \sum_{i=1}^{N_f} \left( \frac{F^f_{\mathrm{obs}} - A \times F^f_{\mathrm{pred}}(z,T)}{\sigma^f_{obs}} \right)^2$$

varying the three free parameters z (redshift), T (spectral type), A (normalization factor).

*Arnouts et al. 1999*