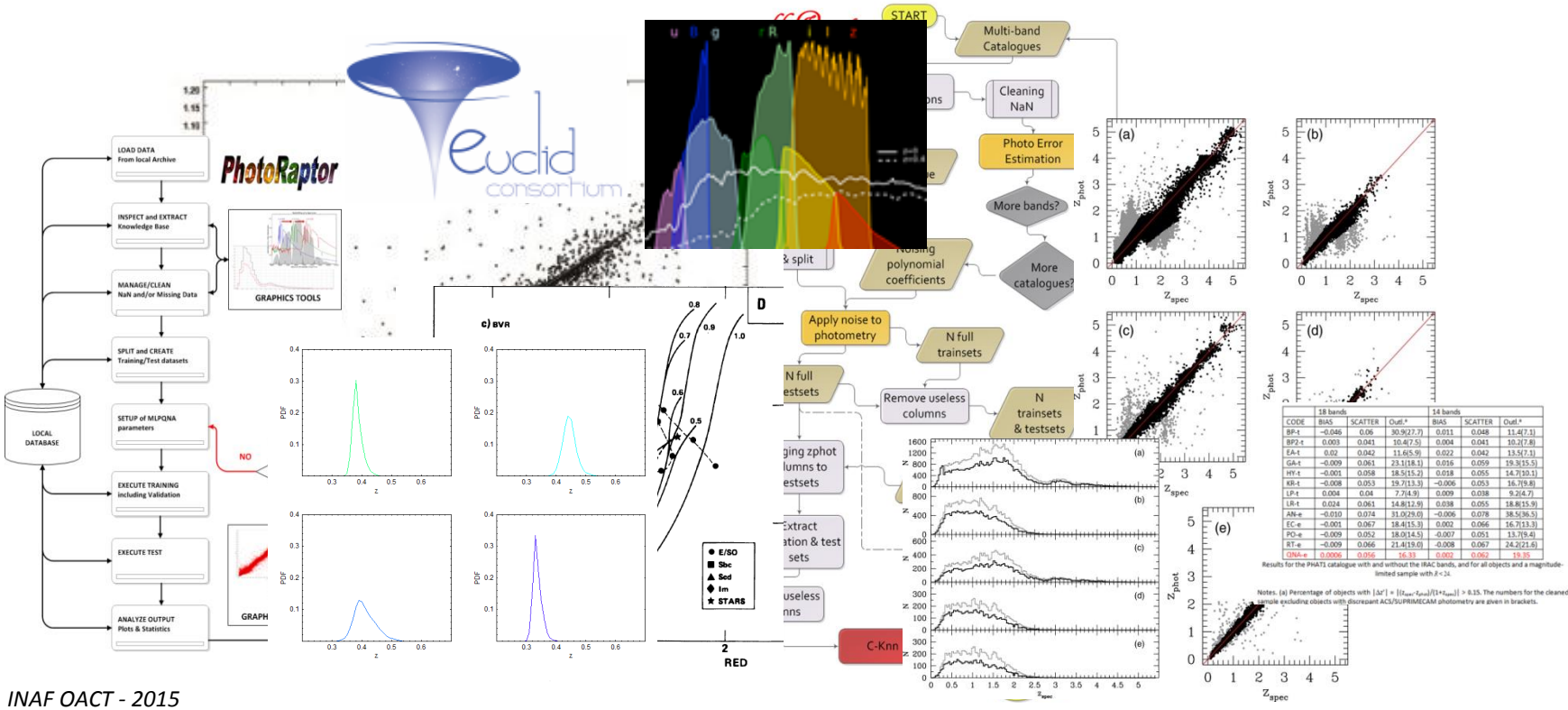
pdfRaptor

A software processing flow for photometric redshifts with machine learning methods

Valeria Amaro[1], Civita Vellucci[1], Stefano Cavuoti[2]
Massimo Brescia[2], Giuseppe Longo[1]

(1) Department of Physics – University Federico II, Via Cinthia 24, I-80126 Napoli, Italy
(2) INAF – Astronomical Observatory of Capodimonte, Via Moiariello 16, I-80131 Napoli, Italy



CODE	18 bands BIAS	18 bands SCATTER	18 bands Outl.*	14 bands BIAS	14 bands SCATTER	14 bands Outl.*
BP-t	−0.046	0.06	30.9(27.7)	0.011	0.048	11.4(7.1)
BP2-t	0.003	0.041	10.4(7.5)	0.004	0.041	10.2(7.8)
EA-t	0.02	0.042	11.6(5.9)	0.022	0.042	13.5(7.1)
GA-t	−0.009	0.061	23.1(18.1)	0.016	0.059	19.3(15.5)
HY-t	−0.001	0.058	18.5(15.2)	0.018	0.055	14.7(10.1)
KR-t	−0.008	0.053	19.7(13.3)	−0.006	0.053	16.7(9.8)
LP-t	0.004	0.04	7.7(4.9)	0.009	0.038	9.2(4.7)
LR-t	0.024	0.061	14.8(12.9)	0.038	0.055	18.8(15.9)
AN-e	−0.010	0.074	31.0(29.0)	−0.006	0.078	38.5(36.5)
EC-e	−0.001	0.067	18.4(15.3)	0.002	0.066	16.7(13.3)
PO-e	−0.009	0.052	18.0(14.5)	−0.007	0.051	13.7(9.4)
RT-e	−0.009	0.066	21.4(19.0)	−0.008	0.067	24.2(21.6)
QNA-e	0.0006	0.056	16.33	0.002	0.062	19.35

Results for the PHAT1 catalogue with and without the IRAC bands, and for all objects and a magnitude-limited sample with $\beta < 24$.

Notes. (a) Percentage of objects with $|\Delta z'| = |(z_{spec} - z_{phot})/(1+z_{spec})| > 0.15$. The numbers for the cleaned sample excluding objects with discrepant ACS/SUPRIMECAM photometry are given in brackets.

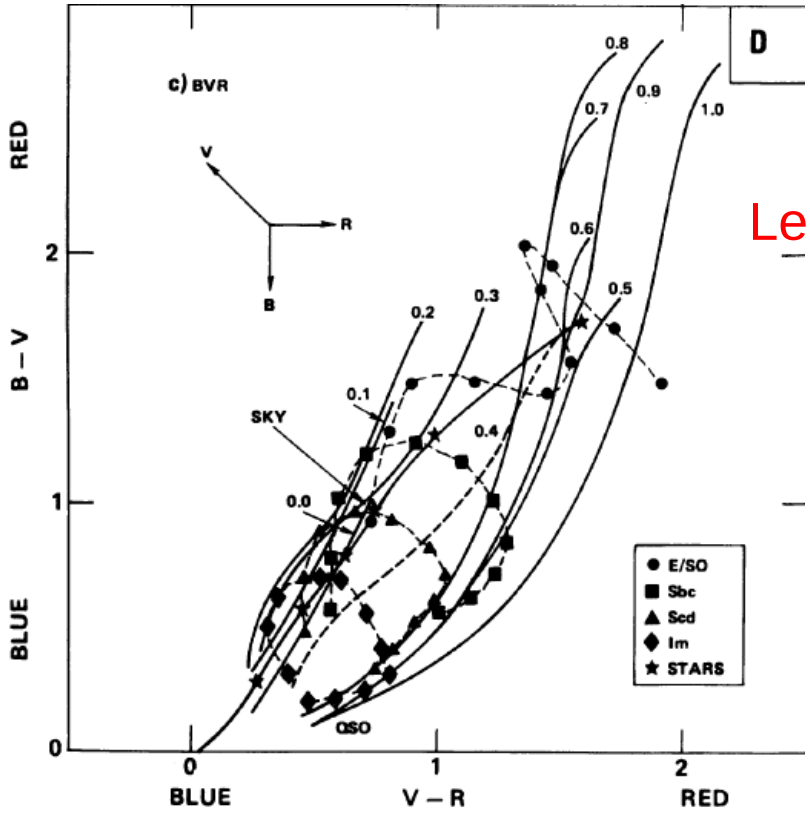
Why we need photo-z?

To measure the distance of objects

$$d = \frac{v}{H_0} \approx \frac{cz}{H_0}$$

To disentangle the degeneracies in the object classification

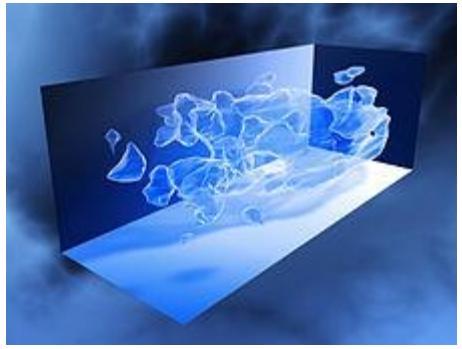
Cosmological parameters



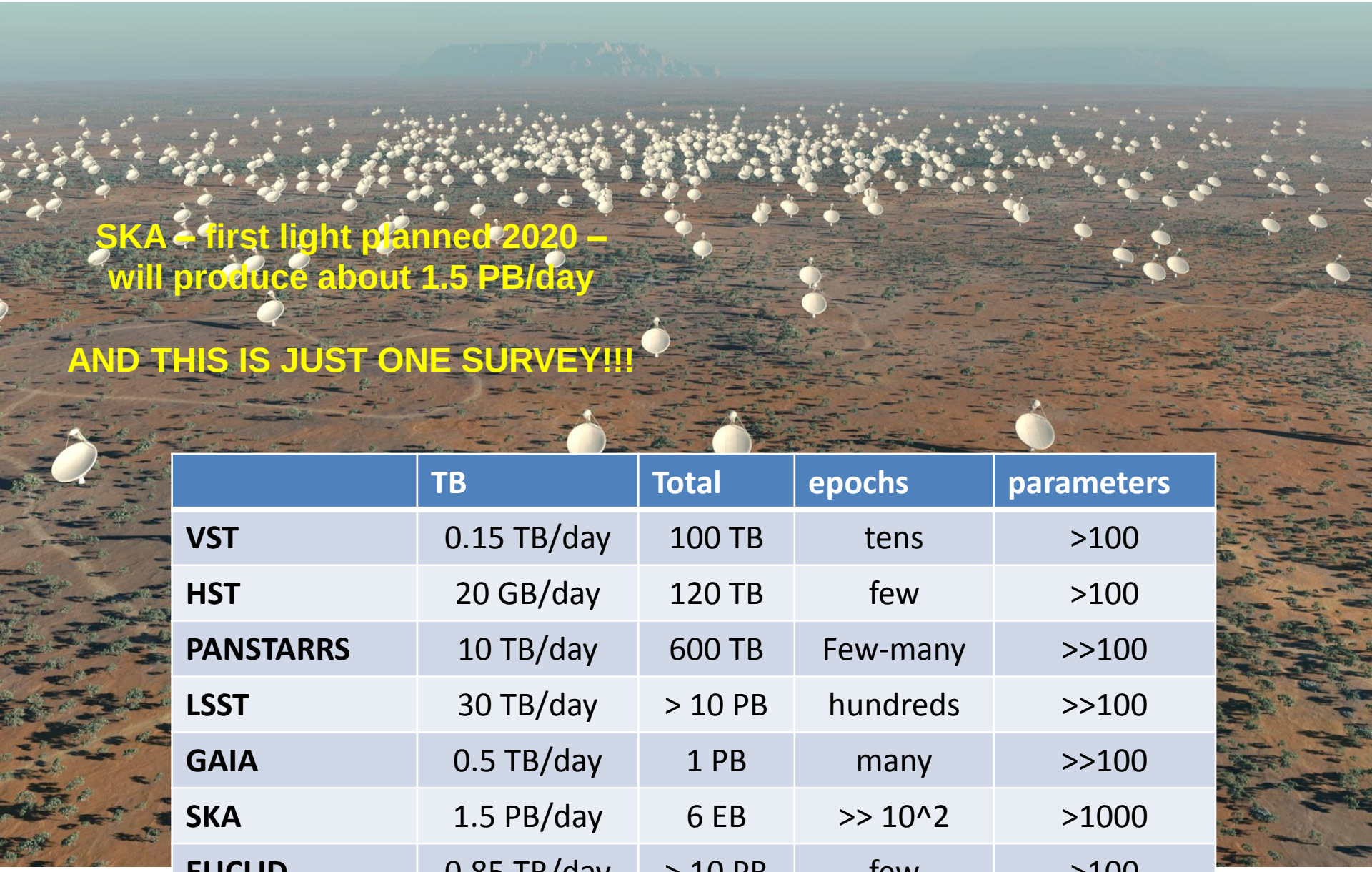
Lensing Effects

Dark Matter

Dark Energy



Data driven Astronomy



	TB	Total	epochs	parameters
VST	0.15 TB/day	100 TB	tens	>100
HST	20 GB/day	120 TB	few	>100
PANSTARRS	10 TB/day	600 TB	Few-many	>>100
LSST	30 TB/day	> 10 PB	hundreds	>>100
GAIA	0.5 TB/day	1 PB	many	>>100
SKA	1.5 PB/day	6 EB	>> 10^2	>1000
EUCLID	0.85 TB/day	> 10 PB	few	>100

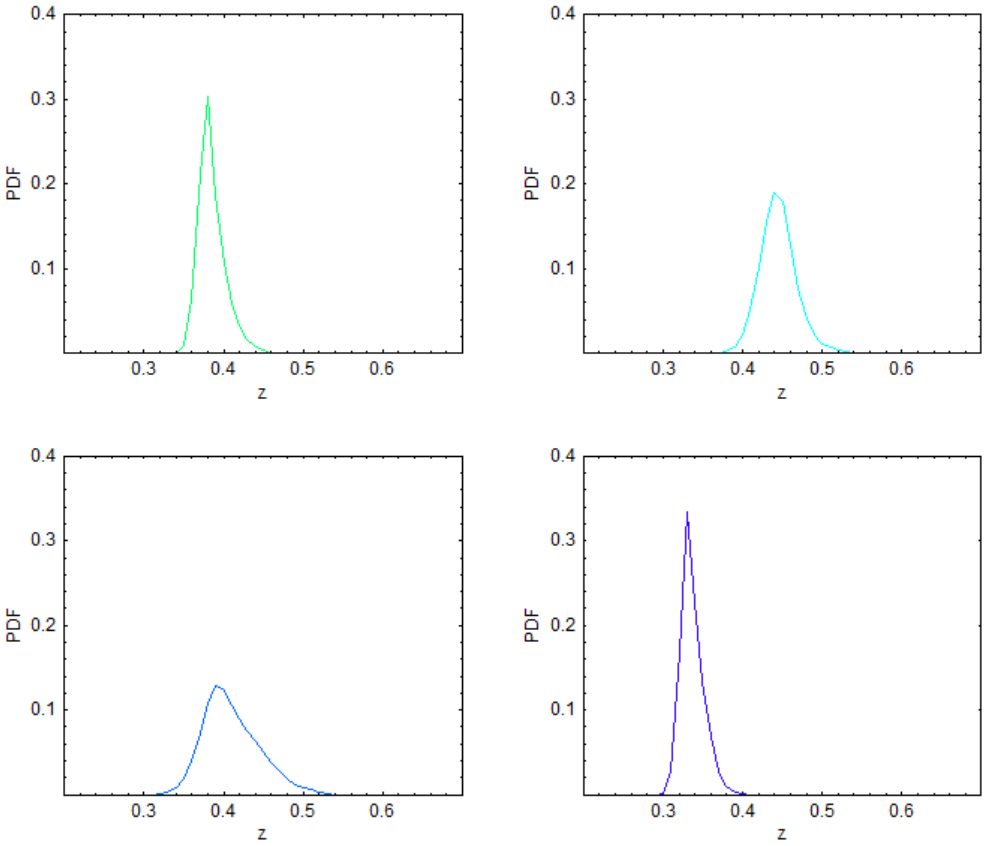
(SED fitting) Template Based

Use SED templates

Convolve with filter transmission curves

Fit object's fluxes (χ^2 minimization)

Output photometric redshift



They use SED models with priors (IMF, SFR, metallicity, age);
They suffer of mismatched templates

They provide a photo-z PDF by default

- *Hyper-z (Bolzonella et al. 2000, A&A 363, 476)*
- *BPZ (Benitez 2000, ApJ 536, 571)*
- *EAZY (Brammer et al. 2008, ApJ 676, 1503)*
- *Le PHARE (Arnouts et al. 2008, MNRAS 329, 355)*
- *LRT (Assef et al. 2008, ApJ 676, 286)*
- *ZEBRA (Feldmann et al. 2006, MNRAS 372, 565)*

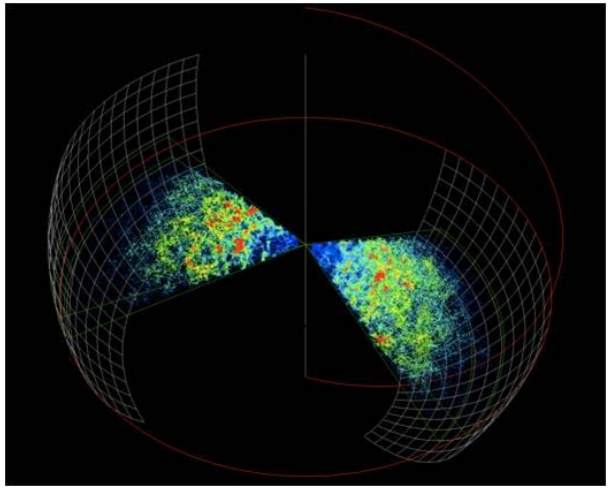
Empirical (data mining) Methods

Use a Knowledge Base of zspec

Learn hidden photometry-zspec correlation

Generalize learning on new photometric objects

Output photometric redshift



NO kind of priors and can find unknown relations;
They require huge Knowledge Base and mismatch outside training spec range

It is not immediate to provide a PDF for these methods
(*because they do not make any assumption on the physical features*)

- Nearest Neighbour (*Csabai & al. 2007, A N 328, 852*)
- Decision-tree (*BDT Gerdes et al 2010, ApJ 715, 823*)
- Direct fitting (*Barth 2002, AJ 124, 5*)
- Neural Networks (*ANNz Collister & Lahav 2004, PASP 116, 345*)
- Support Vector Machines (*Wadadekar 2005, PASP 117, 79*)
- Regression Trees & Random Forests (*Carliles et al 2010, ApJ 712, 511*)
- MLPQNA (*Brescia et al. 2013, ApJ, 772, 140*)

Photometric redshift PDF

One of the consequences of entering the era of precision cosmology is the widespread adoption of photometric redshift Probability Density Functions (PDFs).

Both current and future photometric surveys are expected to obtain images of billions of distinct galaxies. As a result, storing and analyzing all of these PDFs will be non-trivial and even more severe if a survey plans to compute and store multiple different PDFs.

We started our R&D process by a level-0 method (called ***base algorithm***), able to provide a rough PDF estimation of the photo-z for each single input object of the data sample used.
Then we are still under debugging a series of more complex methods based on a post-processing of photo-z production model.

The common element of such process is the machine learning model used to derive photo-z.
The model is MLPQNA (Multi Layer Perceptron trained by the Quasi Newton Algorithm), already successfully validated on several real cases.

Photo-z with MLPQNA

- **PHAT1 Contest** (*Cavuoti et al. 2012, A&A, 546, A13*)
- **GALEX+SDSS+UKIDSS+WISE QSOs** (*Brescia et al. 2013, ApJ, 772, 2, 140*)
- **CLASH-VLT** (*Biviano et al. 2013, A&A, 558, A1*)
- **EUCLID PHZ** (*Coupon et al. 2014, Challenge #1 internal report*)
- **SDSS DR9** (*Brescia et al. 2014, A&A, 568, A126*)
- **KiDS DR2** (*Cavuoti et al. 2015, MNRAS, accepted, in press*)
- **VST VOICE** (*Covone et al. 2015, in prep.*)
- **XMM** (*Vaccari et al. 2015, in prep.*)

MLPQNA Multi Layer Perceptron + Quasi Newton Algorithm

Pre-Processing
(on spectroscopic KB)
Feature Selection

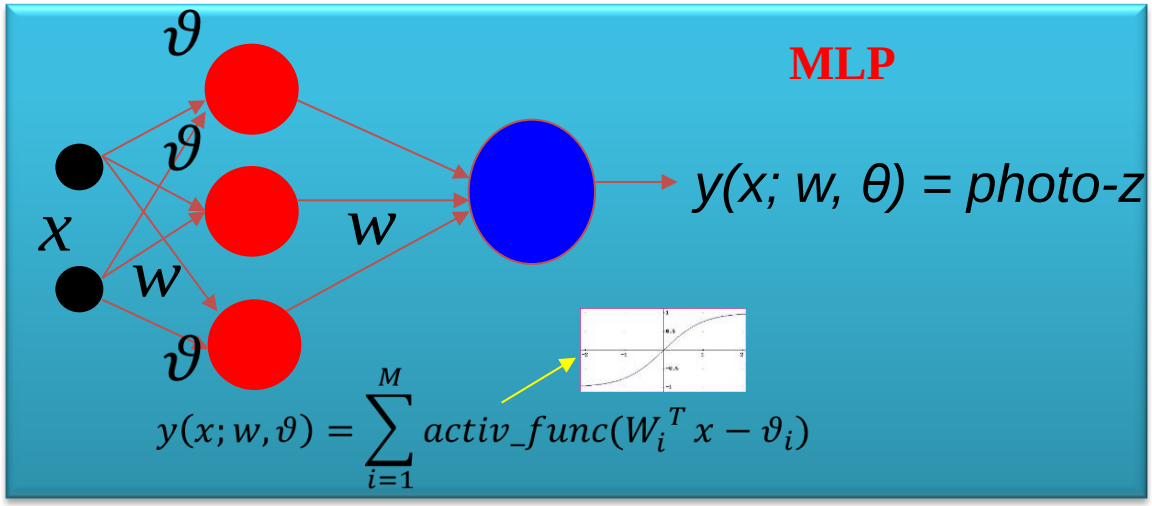
Split

Training Set | Test set

Processing (MLPQNA)

Post-processing

Statistical Analysis
and PDF output



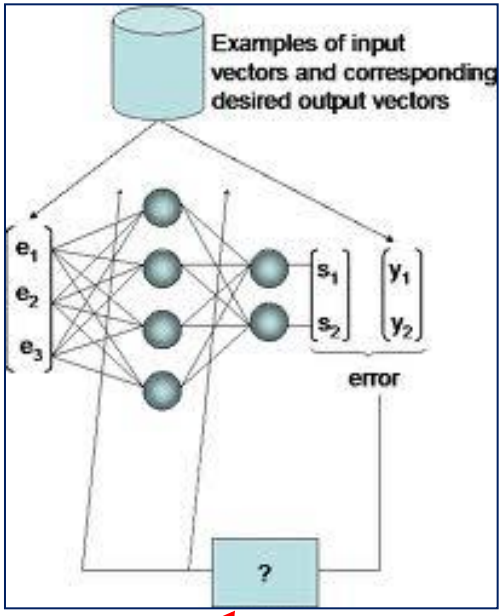
$$w^{k+1} = w^k + \alpha^k d^k$$

$$d^k \in R^N$$

DIRECTION OF SEARCH

$$\alpha^k \in R$$

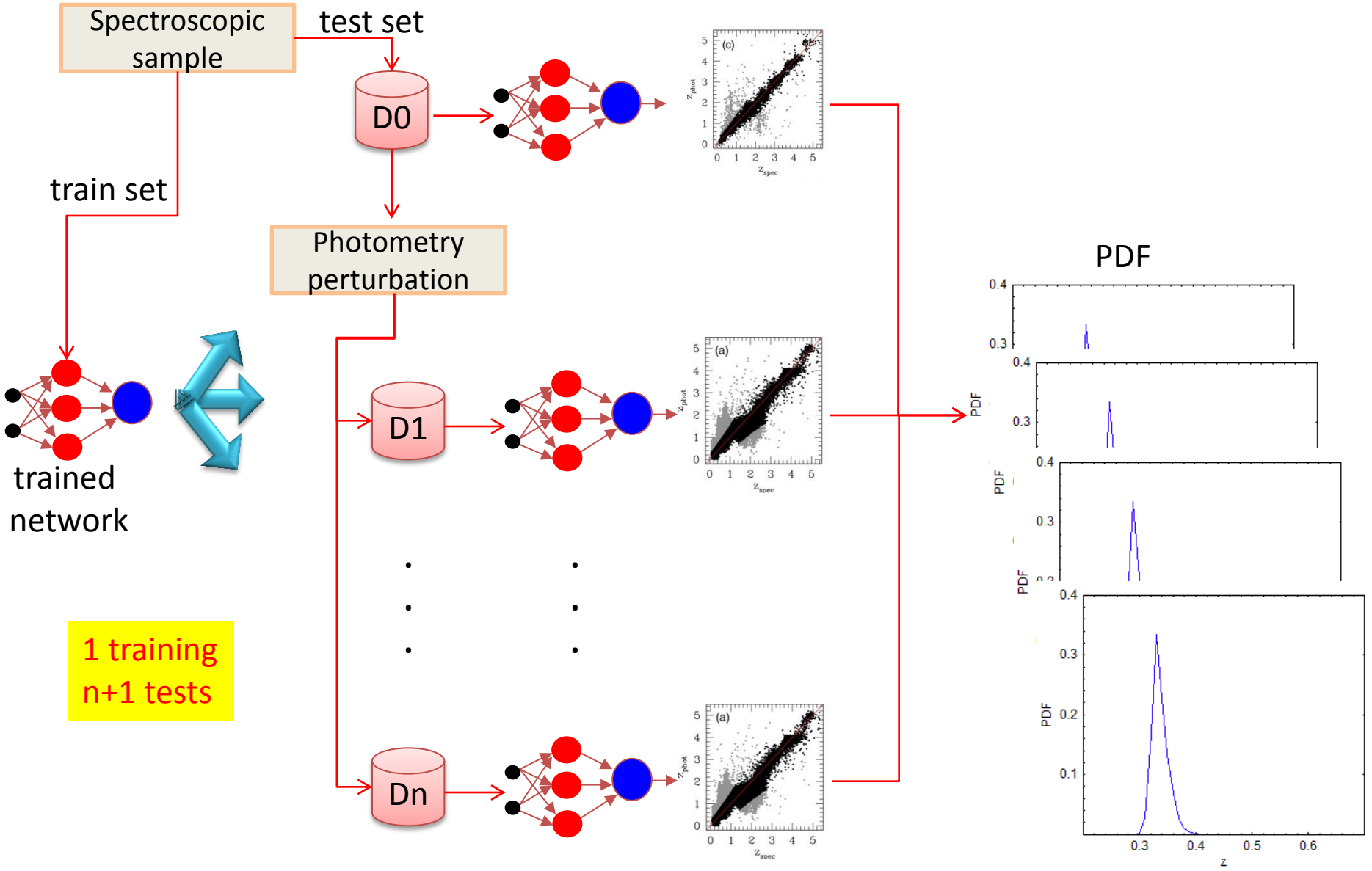
STEP



Hessian approx. (QNA)

$$\nabla^2 E(w^k) d^k = -\nabla E(w^k)$$

PDF base algorithm processing flow



pdfRaptor processing flow

Spectroscopic sample

D0

Photometry perturbation

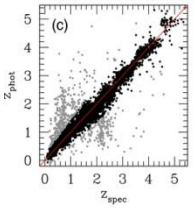
D1

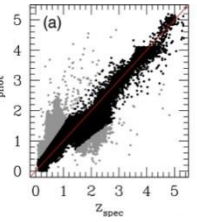
Dn

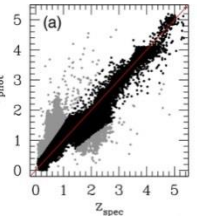
train + test sets

train + test sets

train + test sets



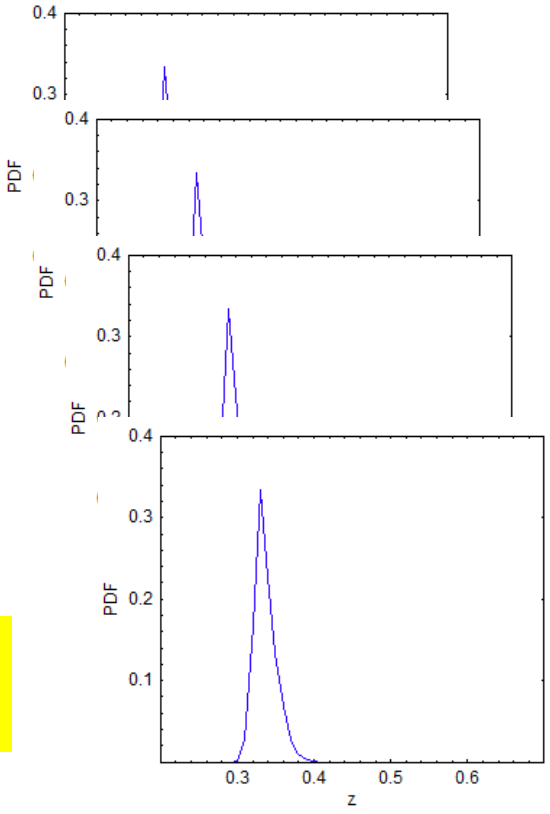




Post-processing model

n+1 training
n+1 tests

PDF



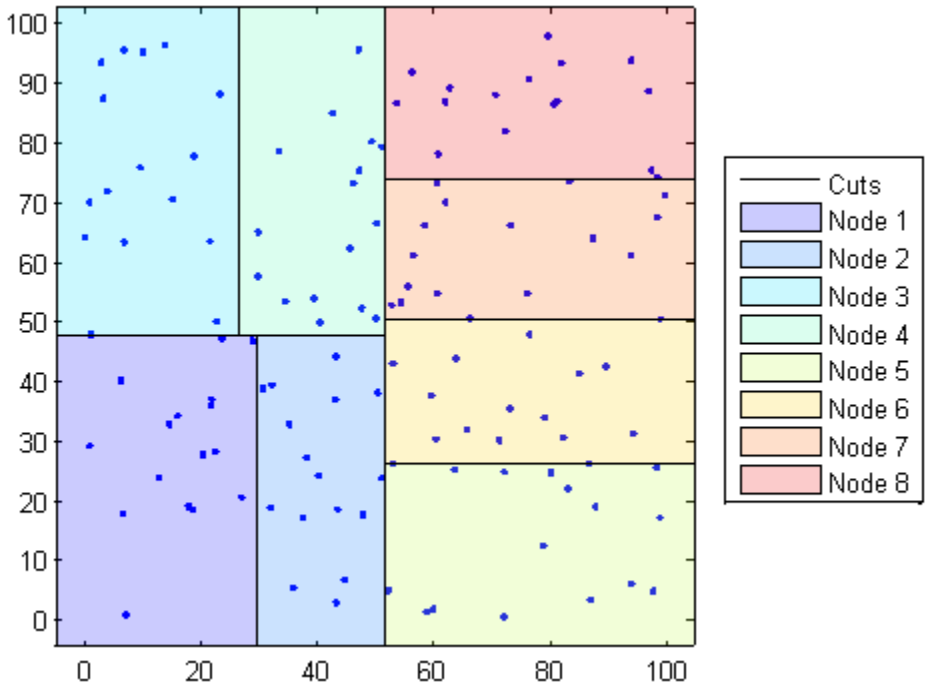
Post-processing for PDF (KD-TREE)

At a high level, a KD-TREE is a generalization of a binary search tree that stores points in *k*-dimensional space.

Post-processing Model
KD-TREE

The method uses the well-known KD-TREE algorithm to partition the photometric and spectroscopic Parameter Space on the base of, respectively, the photometric magnitudes and the zspec present within the used data.

The partitioning produces a series of bins and through the analysis of the associated standard deviations it could be possible to evaluate the trend of photometric error vs the spectroscopic one, giving the possibility to estimate the error distribution and to correlate both types of error trends.



Post-processing for PDF (K-NN)

At a high level, in a K-NN (K Nearest Neighbours) the input consists of the *k* closest training examples in the feature space. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its *k* nearest neighbors.

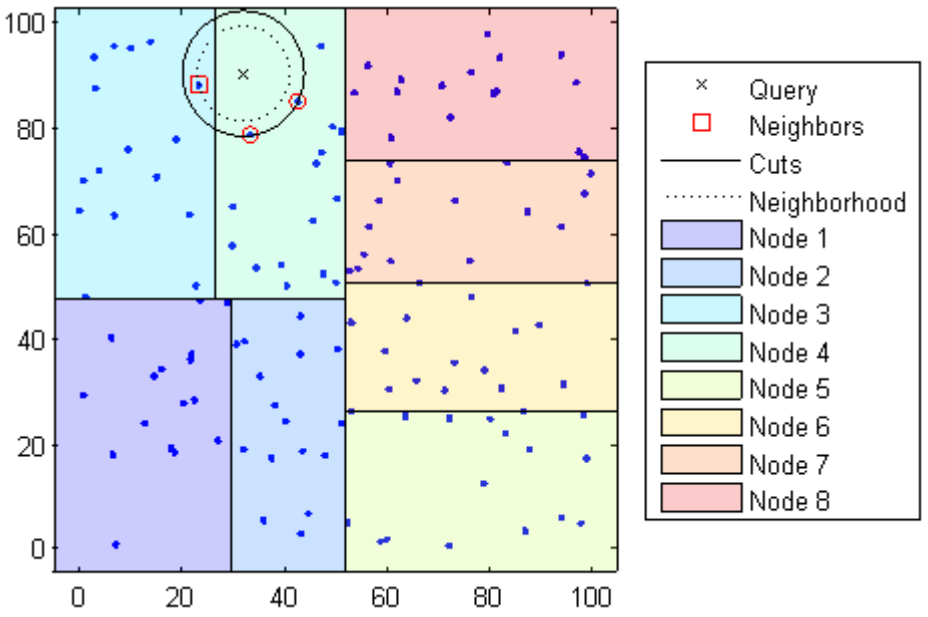
Post-processing Model
K-NN

The **K-NN** method is based on the extraction of arbitrary N objects within the test set closest to each single object selected within the *Evaluation set*.

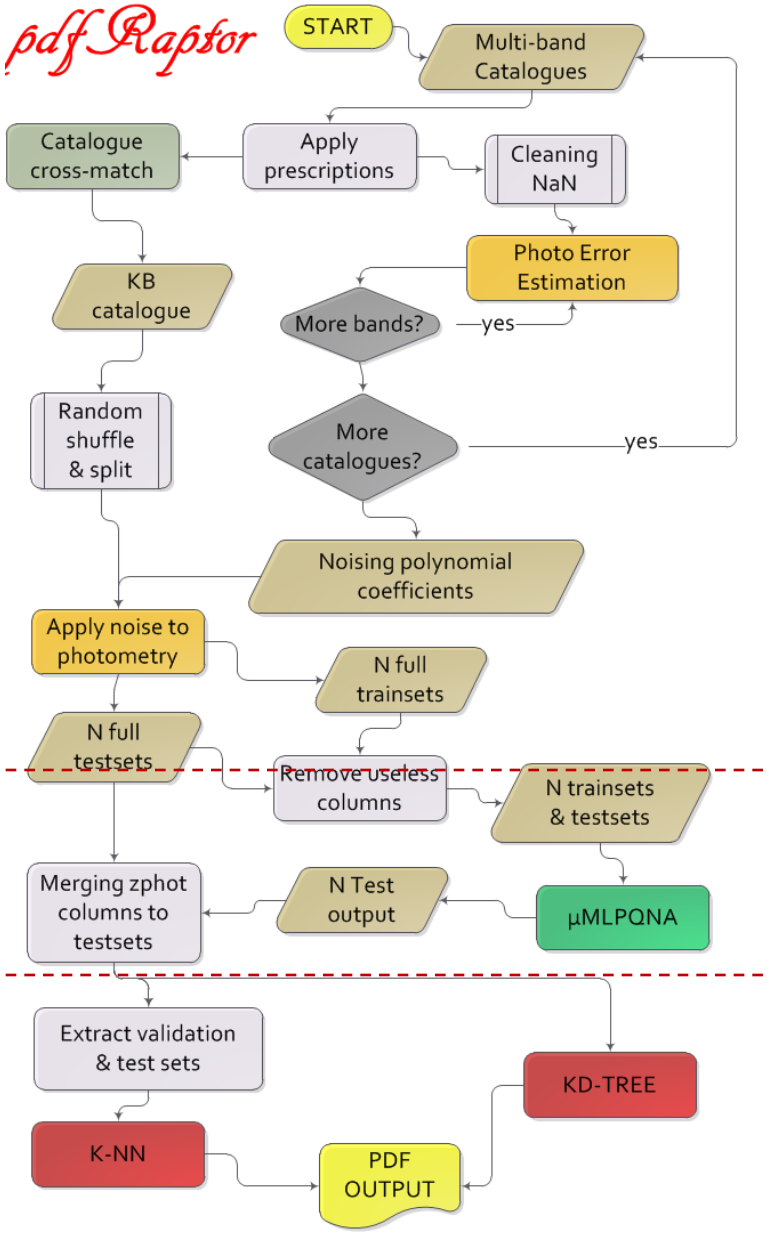
Here closest has to be intended in terms of euclidean distance among all photometric features of the objects.

The resulting distribution of the Δz is obtained by considering the N values for each object of the *Evaluation set*.

The associated error (bias ± σ) is the PDF estimation.



pdfRaptor pipeline architecture



Data Pre-processing: photometric evaluation and error estimation of the multi-band catalogue used as KB of the photo-z experiment.

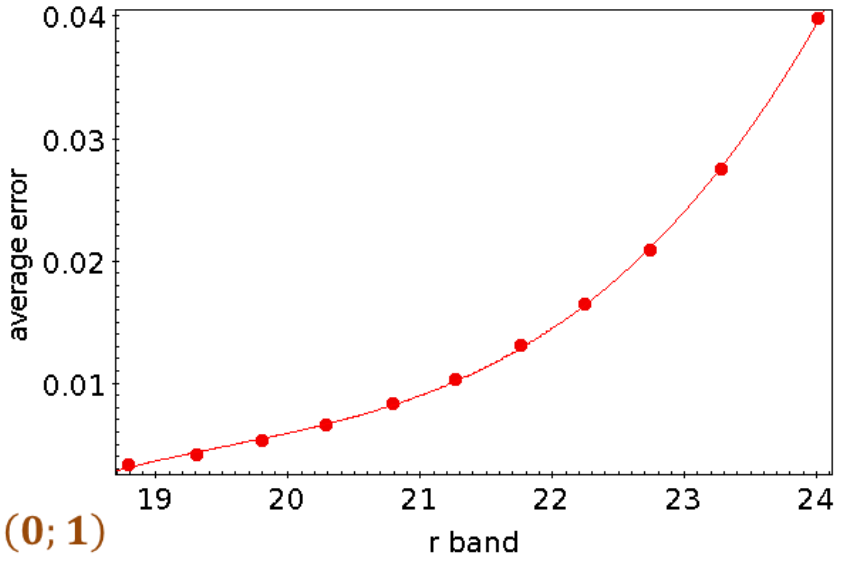
Photo-z calculation: training/test phase to be performed through the selected interpolative method (in this case µMLPQNA, which stands for multi-thread MLPQNA).

PDF calculation: methods designed and implemented to furnish a PDF evaluation for the photo-z produced.

Photometry perturbation

Given a dataset A, a normal distribution on A, and

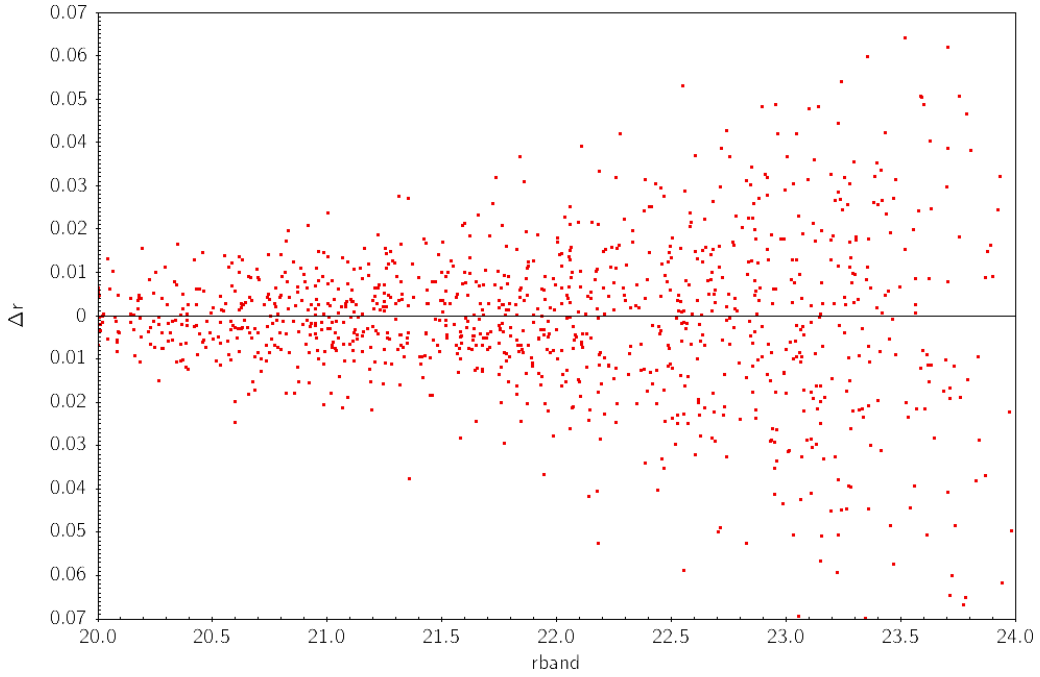
$N_{Samples}$	number of objects in a given dataset A
$N_{perturb}$	number of perturbations to be done
N_{mags},	number of affected magnitudes
p_b	polynomial used to perturb mag of band b
$alpha_b$	perturbation constant for the band b
$mag_b(o_i)$	mag value of the band b for the object o_i



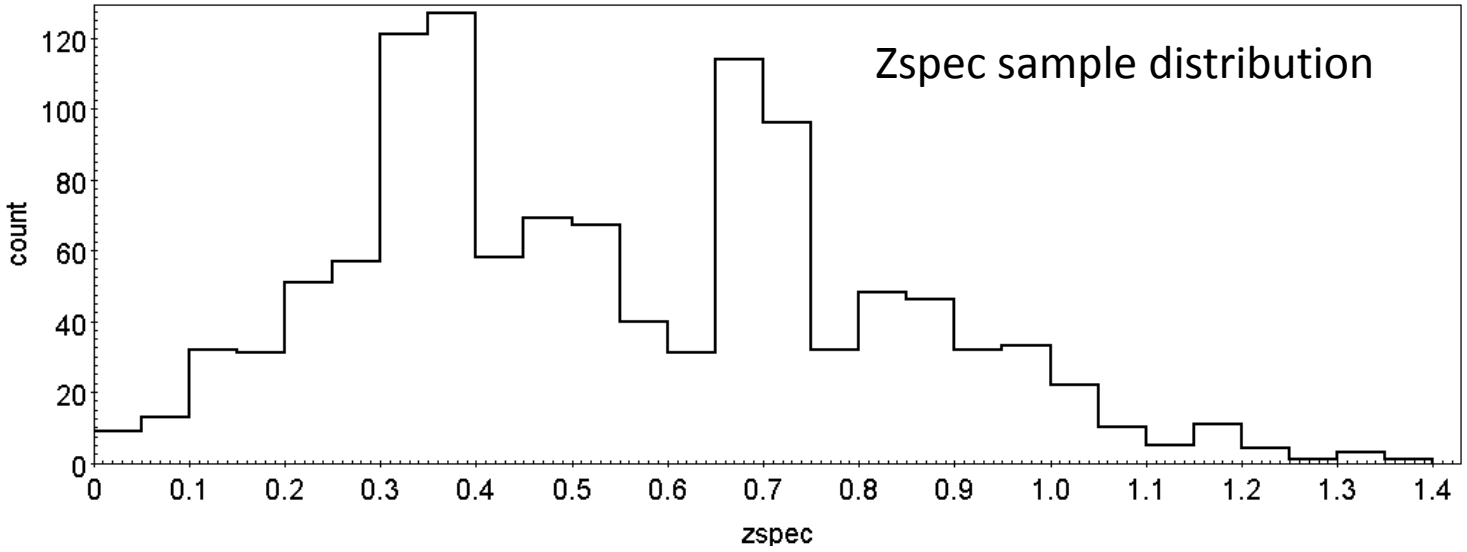
$$\boldsymbol{m_{ijperturbed}(o_i) = m_{ij} + alpha_b * p_b \circ \left(mag(o_i)\right) * N_A(0;1)}$$

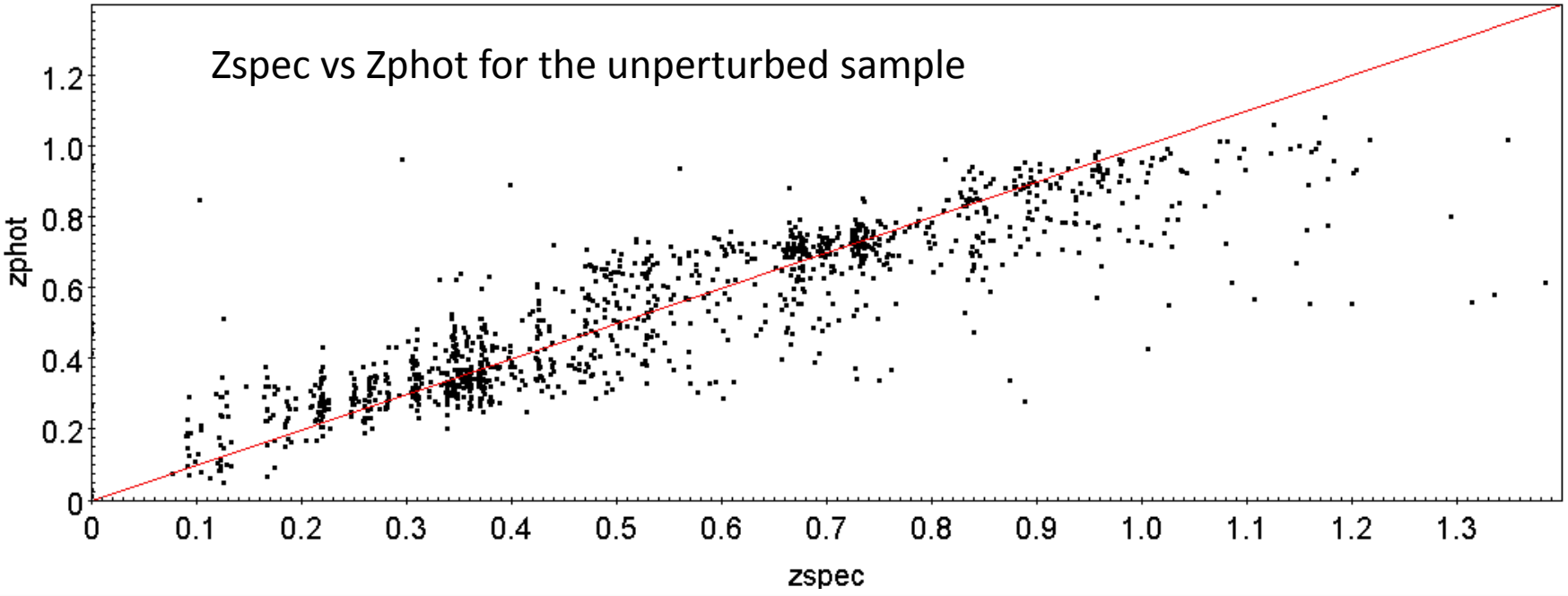
where the symbol "∘" stays for the scalar product,
$N_A(0;1)$is a normal distribution with the dimension of the dataset A to be perturbed, i.e. a distribution of a number $N_{Samples}$ of values in the interval (-1,1).

The variation of the percentage of noise is ensured by the randomly generated normal distribution at each step.

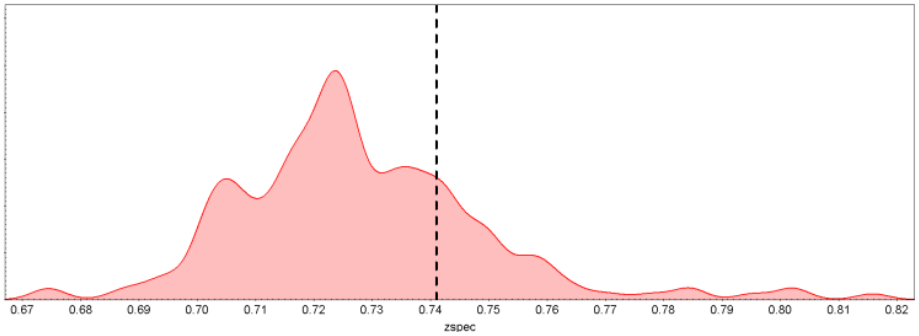


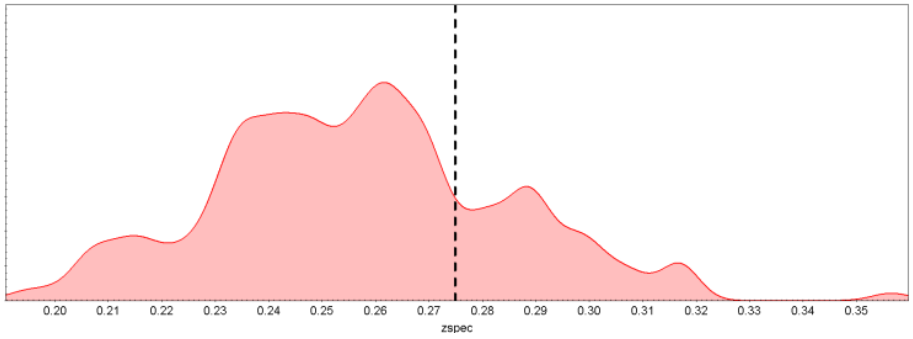
Knowledge base I/O





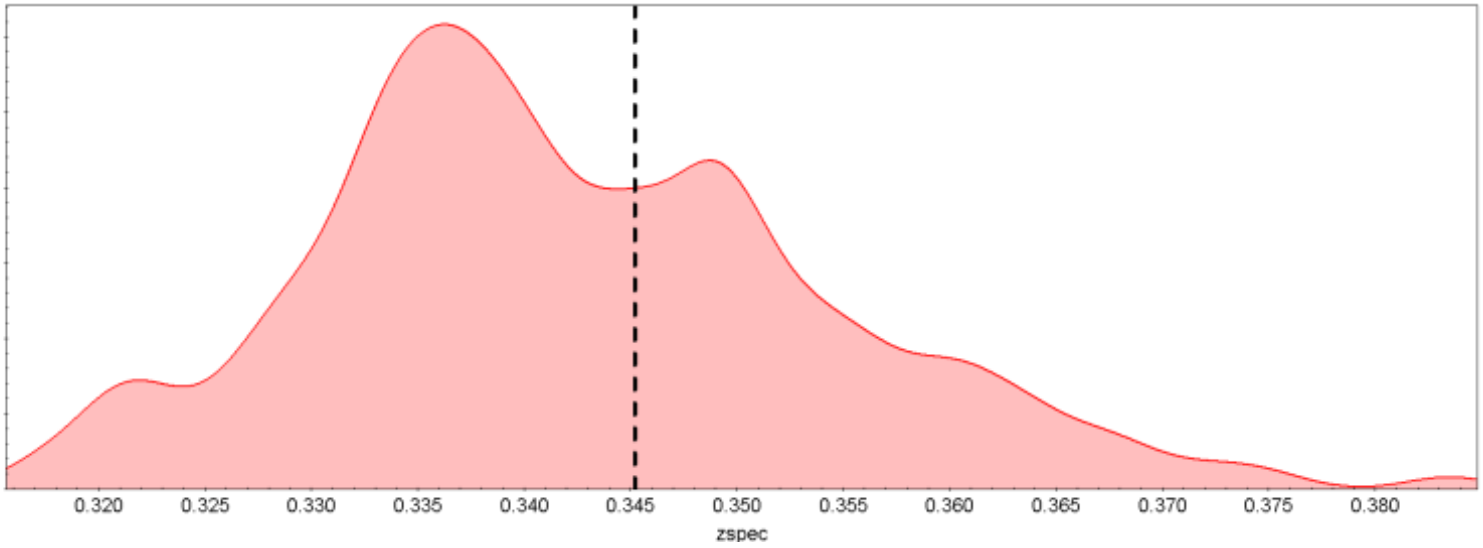
Base algorithm – PDF examples





zCosmos + Cosmos + Ukidss
Bands: R, I, Z, K
objects:2,887
(60%) 1,723 training set
(40%) 1,164 test set

Tests	Bias	σ
Unperturbed test set	0.0014	0.0087
Average on 200 perturbed test sets + unperturbed test set	0.0004	0.0085



Thanks!