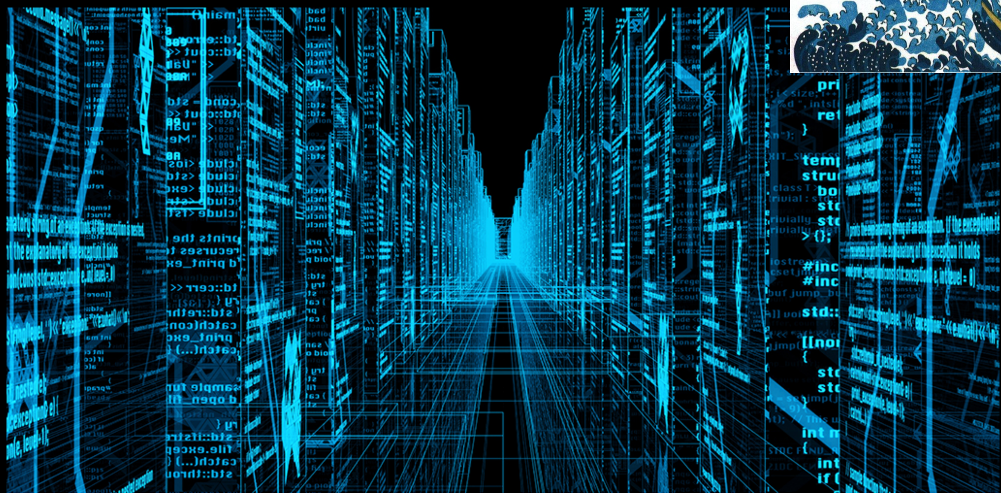
The ART of getting science from big data with machine learning: the case of photometric redshifts

G. Longo, M. Brescia, S, Cavuoti,
V. Amaro, C. Vellucci, and others

1. Department of Physics, University Federico II in Naples (I)
2. INAF – Astronomical Observatory of Capodimonte (I)



Università degli Studi Federico II

Istituto Nazionale di Astrofisica - INAF

CENTER FOR DATA-DRIVEN DISCOVERY

California Institute of Technology

cost

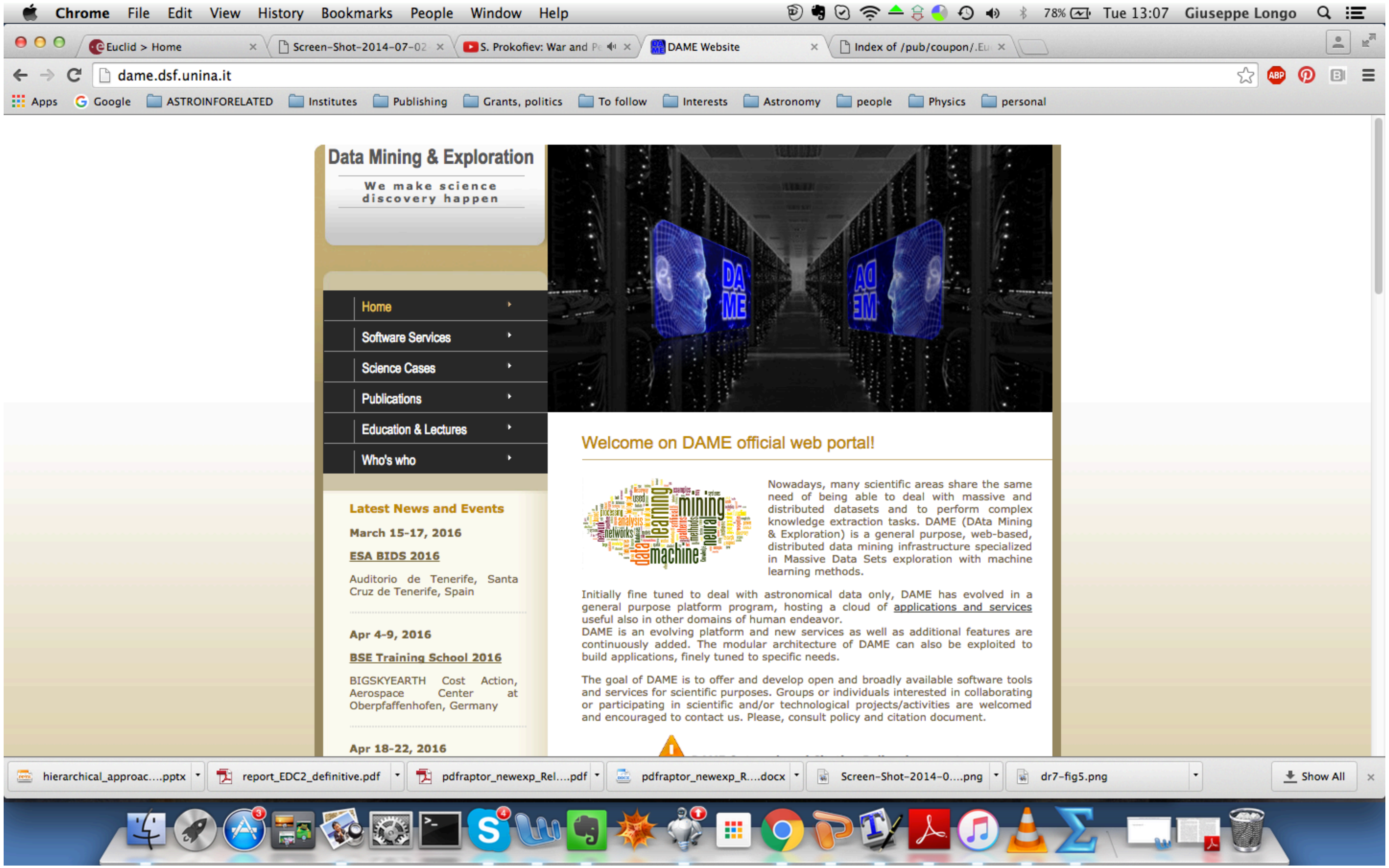
TD-1403

COST Action "Big-Sky Earth"

SUNDIAL
H2020 Innovative Training Network

DAME

All software, papers, discussion, demos, etc. are available here: http://dame.dsf.unina.it/



Distilled problems as derived from our experience **on many data sets…**

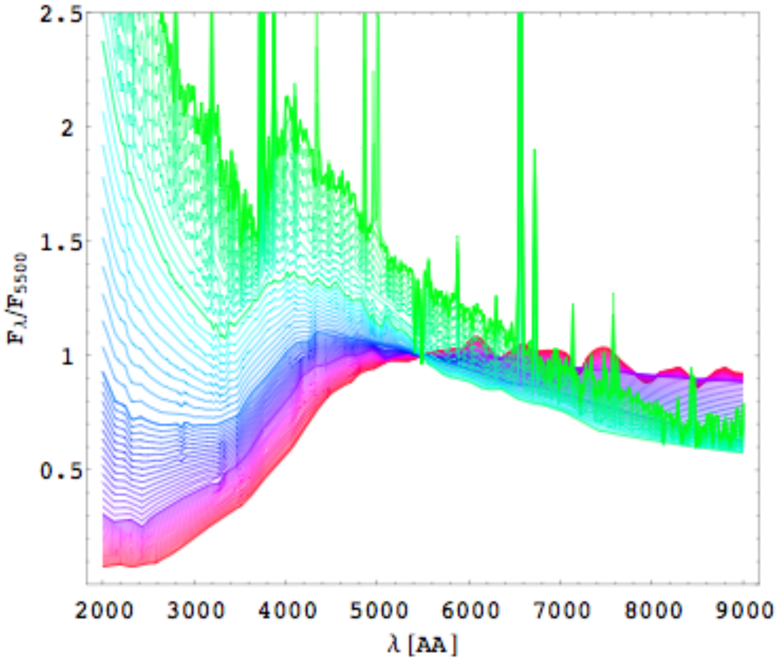
…on many problems….

1. Coverage of the Observed Parameter Space by the knowledge base (biases, outliers, peculiar and rare objects, etc… nature of the sample, etc.)
2. Choice of the method
3. Feature selection
4. Missing data
5. Error estimation

Sloan Digital Sky Survey (SDSS)
Kilo Degree Survey (KiDS)
SUBARU/HSC-COSMOS (also EMU)
Euclid (DC1 & 2)
VST-VOICE
CRTS
LSST

Star/Galaxy xlassification
Classificcation of galaxies (emission lines, AGN, starburst, etc.)
Metallicty
Star formation rates
Young stellar objects (via Lactea Project, cf.Molinari's talk)
Photometric redshifts
….
Non astronomical data sets
(biomedical and geophysics)

Two approaches SED (Spectral Energy Distribution) fitting



Library of M template spectra (M<100)

Convolve with filter bandpasses for a specific survey

Stretch templates for redshift (z) assuming constant step Dz in an interval range z_{min} , z_{max}

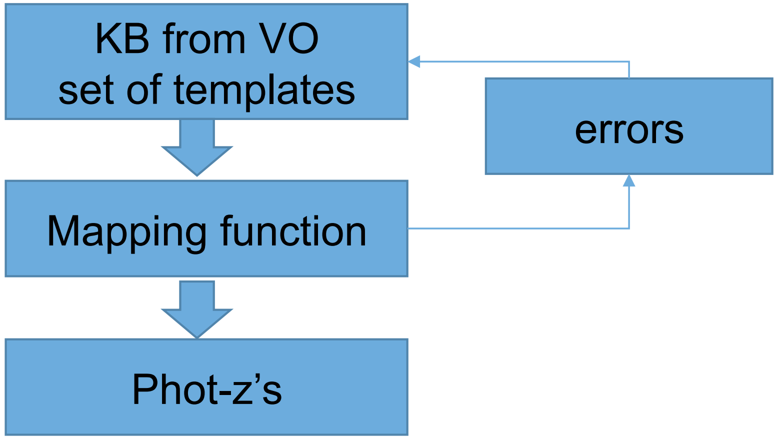
$$SED(T_i, z_{\min} + n\Delta z) \quad i \in \{1, M\},\ \mathrm{n} \in \left\{1, \frac{z_{\max} - z_{\min}}{\Delta z}\right\}$$

Find best fitting i,j using any optimization method

Templates: either synthetic or observed

Arbitrary choice of templates, lots of assumptions on physics
Strong dependence on zero points, photometric calibrations, etc.

But they go very deep, well beyond the spectroscopic limit



Machine learning methods

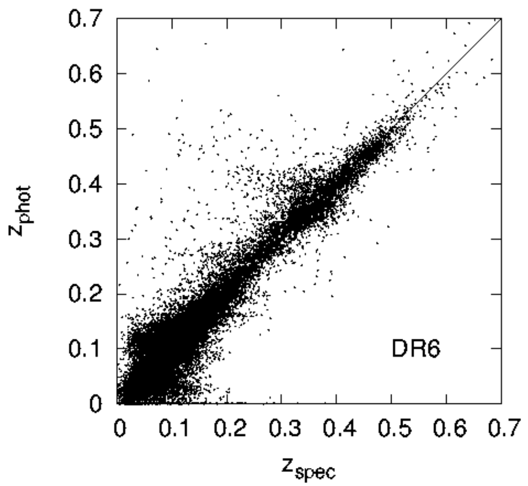
Library of true template spectra (large samples) from real objects (training set)

Use examples to find the mapping function

More accurate than SED fitting

But:

Need for the training set to properly cover the OPS
Need to select proper set of features
Need to properly handle missing data

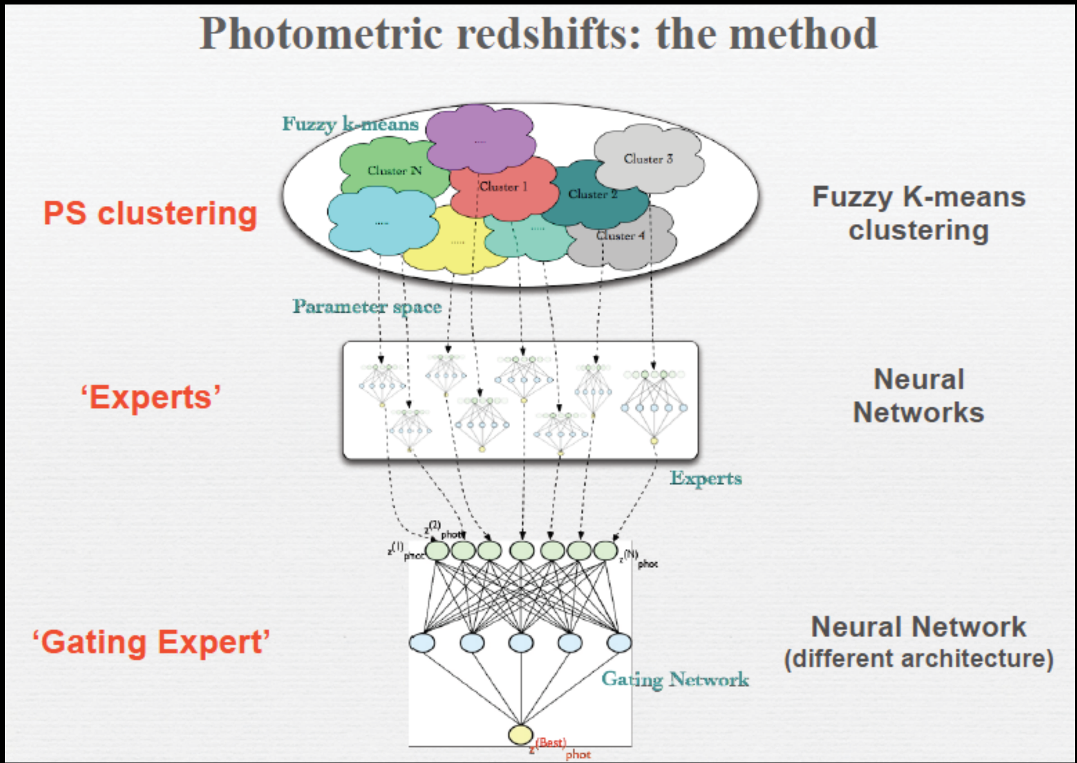


Models are almost irrelevant
SVM (various flavours), (MLP’s - many implementations); Decision Trees, RF (various flavours), kNN, etc…

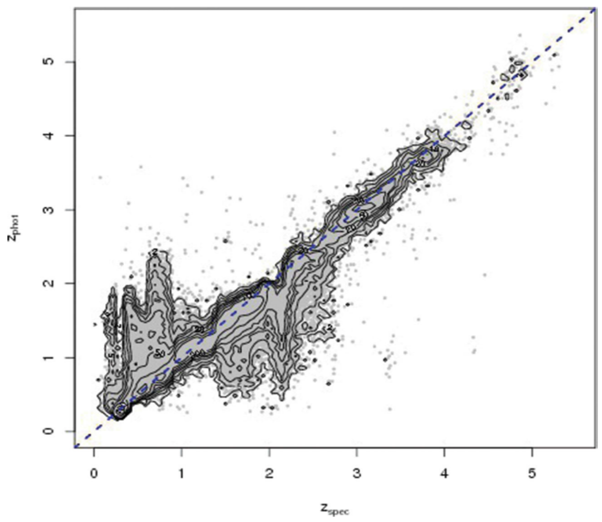
Photo-z for Quasars: **Astroinformatics of galaxies and quasars: a new general method for photometric redshifts estimation**, O. Laurino, R. D'Abrusco, G. Longo, and G. Riccio, MNRAS, 2011, 418, 2165 (arXiv/1107.3160);

WGE: Weak Gated Expert

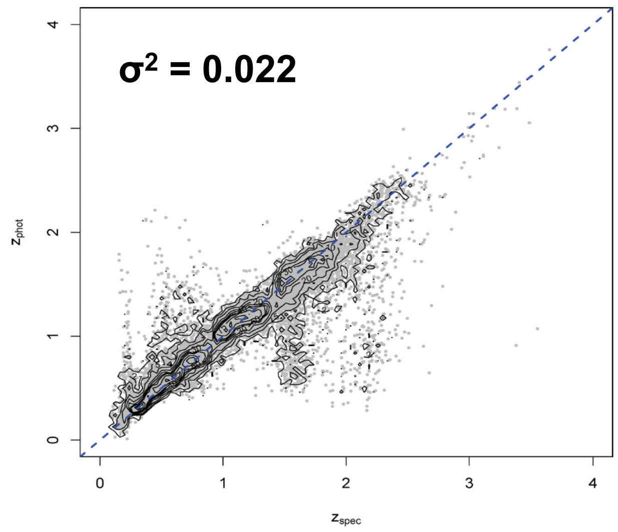
Data from the unresolved objects SDSS catalogue

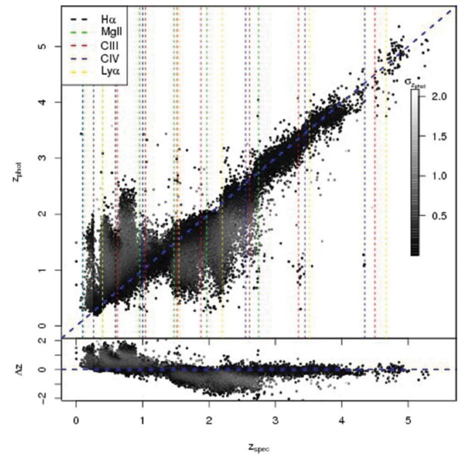


Optical bands only



Optical + UV bands





›per panel, it is shown the scatter plot of the spectroscopic versus photometric redshifts evaluated with the WGE method for the members periment for the quasars extracted from the SDSS catalogue with optical photometry, while in the lower panel the scatter plot of the ft z_{spec} versus Δz variable is shown for the same sources. All points are colour coded according to the value of the errors $\sigma_{z_{phot}}$ as evaluated ertical dashed lines represent the redshift at which the most luminous emission lines characterizing quasars spectra shift off the SDSS ue to redshift. Most of the features of the plot are associated to one or more of these lines.

1.st lesson: Additional Info are always needed to understand systematics

Ex. Position of emission lines relative to filter bands

Second method on same OBJECTS: MLPQNA

Survey	Bands	Name of feature	Synthetic description
GALEX	nuv, fuv	mag, mag_iso mag_Aper_1 mag_Aper_2 mag_Aper_3 mag_auto and kron_radius	Near and Far UV total and isophotal mags phot. through 3, 4.5 and 7.5 arcsec apertures magnitudes and Kron radius in units of A or B
SDSS	u, g, r, i, z	psfMag	PSF fitting magnitude in the u g, r, i, z bands.
UKIDSS	Y, J, H, K	PsfMag AperMag3, AperMag4, AperMag6 HallMag, PetroMag	PSF fitting magnitude in Y, J, H, K bands aperture photometry through 2, 2.8 & 5.7″ circular aperture in each band Calibrated magnitude within circular aperture r_hall and Petrosian magnitude in Y, J, H, K bands
WISE	W1, W2, W3, W4	W1mpro, W2mpro, W3mpro, W4mpro	W1: 3.4 μm and 6.1″ angular resolution; W2: 4.6 μm and 6.4″ angular resolution; W3: 12 μm and 6.5″ angular resolution; W4: 22 μm and 12″ angular resolution. Magnitudes measured with profile-fitting photometry at the 95% level. Brightness upper limit if the flux measurement has SNR< 2
SDSS	-	z_{spec}	Spectroscopic redshift

Parameter space more complex and need for Feature selection

[Photometric redshifts for quasars in multiband surveys](), M. Brescia, S. Cavuoti, R. D'Abrusco, A. Mercurio, G. Longo, 2013, ApJ, 772, 140 (astro-ph: 1305.5641)

Table 6. Catastrophic outliers evaluation and comparison between the residual $\sigma_{clean}(\Delta z_{norm})$ and $NMAD(\Delta z_{norm})$. The reported number of objects, for each cross-matched catalog, is referred to the test sets only. Catastrophic outliers are defined as objects where $|\Delta z_{norm}| > 2\sigma\,(\Delta z_{norm})$. The standard deviation $\sigma_{clean}(\Delta z_{norm})$ is calculated after having removed the catastrophic outliers, i.e. on the data sample for which $|\Delta z_{norm}| \leq 2\sigma\,(\Delta z_{norm})$

Exp	n. obj.	$\sigma\,(\Delta z_{norm})$	% catas. outliers	$\sigma_{clean}\,(\Delta z_{norm})$	$NMAD\,(\Delta z_{norm})$
SDSS	41431	0.15	6.53	0.062	0.058
SDSS + GALEX	17876	0.11	4.57	0.045	0.043
SDSS+UKIDSS	12438	0.11	3.82	0.041	0.040
SDSS+GALEX+UKIDSS	5836	0.087	3.05	0.040	0.032
SDSS+GALEX+UKIDSS+WISE	5716	0.069	2.88	0.035	0.029

2-nd lesson:
Adding more parameters may improve performances … but….

Table 4. Comparison among the performances of the different references. MLPQNA is related to our experiments, based on a four-layers network, trained on the mixed (colors + reference magnitudes) datasets. In some cases the comparison references are not reported, due to the missing statistics. Column 1: reference; columns 2-6, respectively: bias, standard deviation, MAD, RMS and NMAD calculated on $\Delta z_{norm} = (z_{spec} - z_{phot}) / (1 + z_{spec})$ related to the test sets. For the definition of the parameters and for discussion see text.

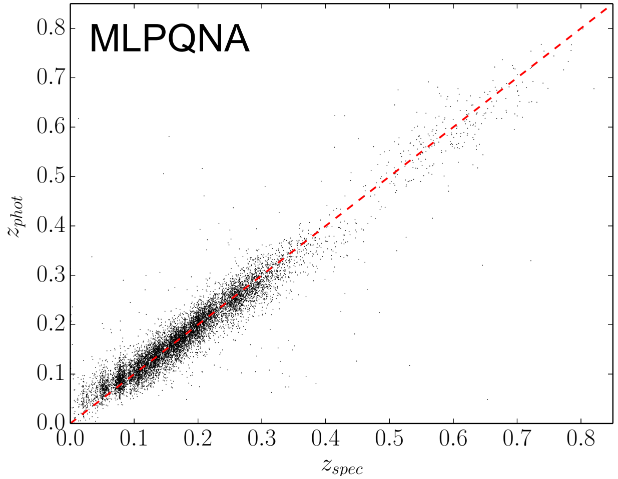
Exp	$BIAS(\Delta z_{norm})$	$\sigma(\Delta z_{norm})$	$MAD(\Delta z_{norm})$	$RMS(\Delta z_{norm})$	$NMAD(\Delta z_{norm})$
			SDSS		
MLPQNA	0.032	0.15	0.039	0.17	0.058
Laurino et al.	0.095	0.16	0.041	0.19	-
Ball et al.	0.095	0.18	-	-	-
Richards et al.	0.115	0.28	-	-	-
			SDSS + GALEX		
MLPQNA	0.012	0.11	0.029	0.11	0.043
Laurino et al.	0.058	0.29	0.029	0.11	-
Ball et al.	0.06	0.12	-	-	-
Richards et al.	0.071	0.18	-	-	-
			SDSS + UKIDSS		
MLPQNA	0.008	0.11	0.027	0.11	0.040
			SDSS + GALEX + UKIDSS		
MLPQNA	0.005	0.087	0.022	0.088	0.032
			SDSS + GALEX + UKIDSS + WISE		
MLPQNA	0.004	0.069	0.020	0.069	0.029

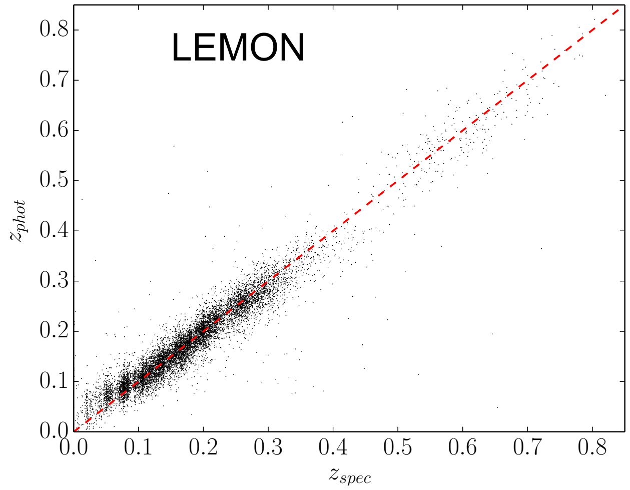
Table 5. Comparison in terms of outliers percentages among the different references. In some cases the comparison references are not reported, due to the missing statistics. Column 1: reference; Column 2-3 are fractions of outliers at different σ based on $\Delta z = (z_{spec} - z_{phot})$; Column 4-5 are the fractions of outliers at different σ based on $\Delta z_{norm} = (z_{spec} - z_{phot}) / (1 + z_{spec})$. The column 4 reports our catastrophic outliers, defined as $|\Delta z_{norm}| > 2\sigma(\Delta z_{norm})$.

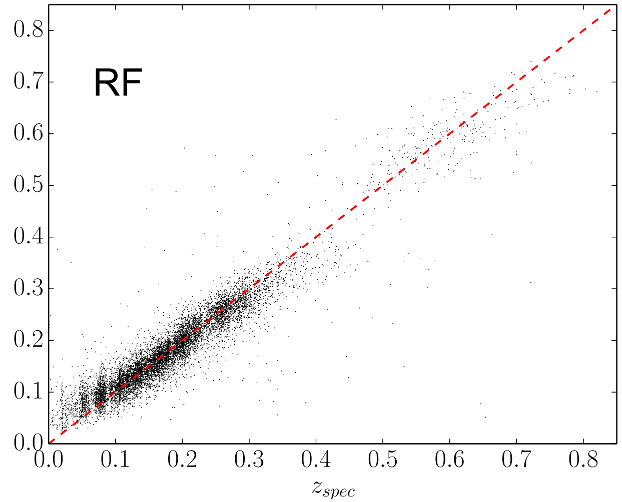
Exp	Outliers ($\|\Delta z\|$)		Outliers ($\|\Delta z_{norm}\|$)	
	$> 2\sigma(\Delta z)$	$> 4\sigma(\Delta z)$	$> 2\sigma(\Delta z_{norm})$	$> 4\sigma(\Delta z_{norm})$
	SDSS			
MLPQNA	7.68	0.38	6.53	1.24
Bovy et al.		0.51		
	SDSS + GALEX			
MLPQNA	4.88	1.61	4.57	1.37
Bovy et al.		1.86		
	SDSS + UKIDSS			
MLPQNA	4.00	1.73	3.82	1.38
Bovy et al.		1.92		
	SDSS + GALEX + UKIDSS			
MLPQNA	2.86	1.47	3.05	0.23
Bovy et al.		1.13		
	SDSS + GALEX + UKIDSS + WISE			
MLPQNA	2.57	0.87	2.88	0.91

Different Machine Learning methods of different complexity (MLPQNA is simpler than WGE) lead to similar results with a slight edge for MLPQNA

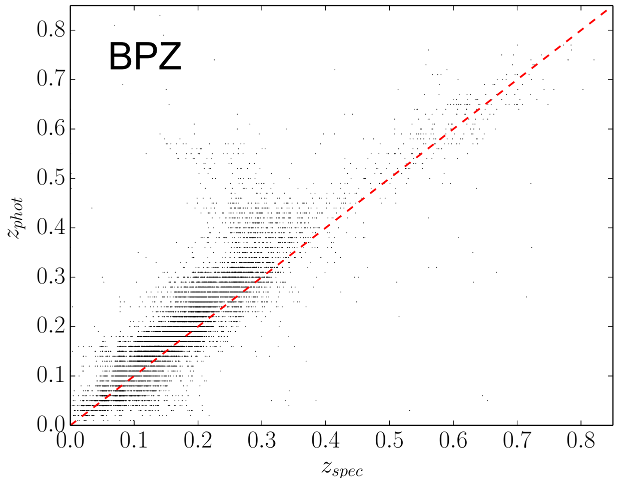
A few selected resuls from a large variety of methods applied to the same data set and problem

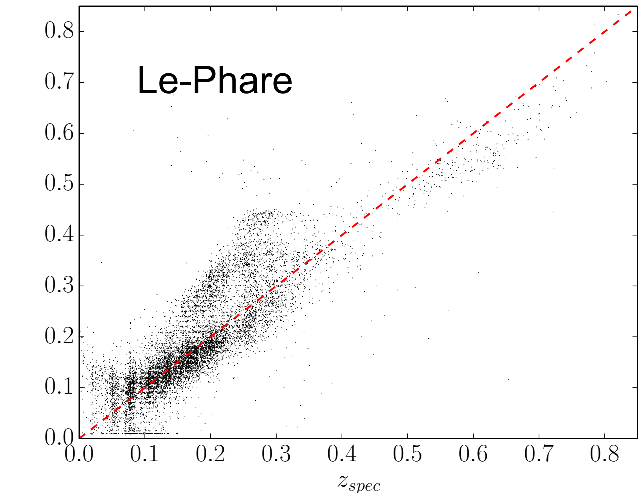






Cavuoti, Tortora, Brescia, Longo et al., MNRAS, 2016





More or less, different ML methods are equivalent

(no need to look for the latest fashionable method … just to produce one paper more….)

Room for improvement is elsewhere

Feature selection

Coverage of the observed parameter space

- Missing or uneven spectroscopic coverage
- Peculiar objects (different populations result from different selection criteria)
- How to go beyond the spectroscopic limit

Missing Data

- Need to handle differently “non detections” and “non observed”

(Cavuoti et al. in preparation)

Evaluation of errors

- Probability distribution function
- Proper choice of statistical indicators

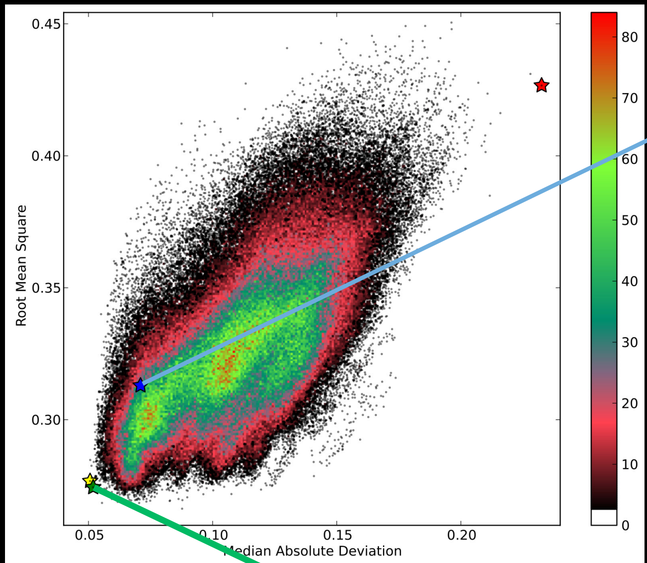
Photometric redshifts for QSO's … a data driven approach

(from K. Polsterer, Heidelberg, 2015)

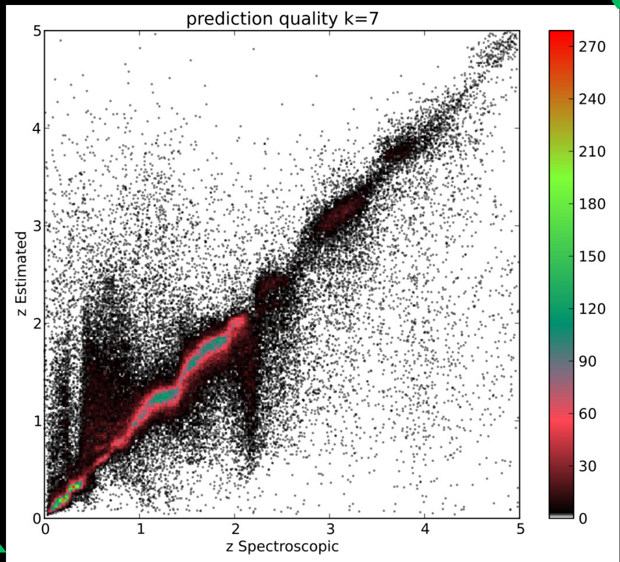
$$\frac{n!}{(n-r)!r!} = 341{,}055 \ \text{ combinations}$$

One does not know a-priori which features are the most relevant

Use all 55 significant photometric features to select the most significant 4



Laurino et al.
Traditional feature selection



Best combination

$u_{model} - g_{model}$

$g_{psf} - r_{model}$

$z_{psf} - r_{model}$

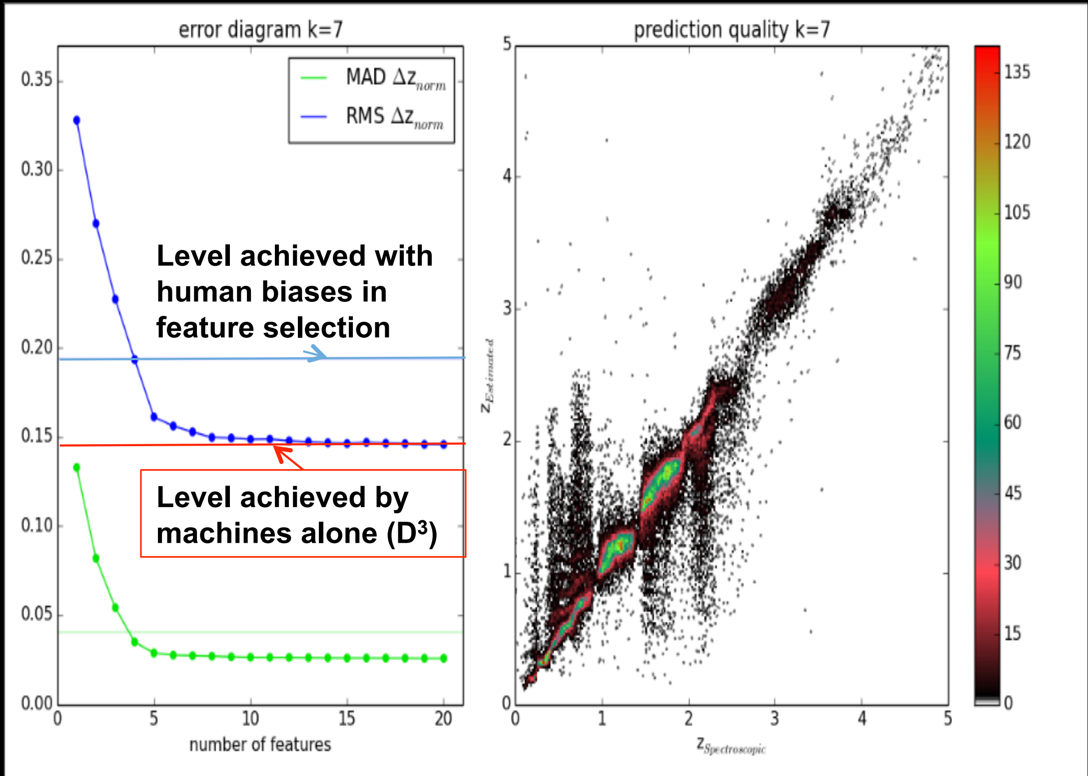
$i_{psf} - z_{model}$

Results comparable to Brescia et al. 2014

Photometric redshifts for SDSS QSO (From K. Polsterer)

PSF, Petrosian, Total magnitudes + extinction + errors ….. 585 features…. Let us find the best combination of 10, 11, 12 etc… using FEATURE ADDITION

For just 10 features ….. 1,197,308,441,345,108,200,000 combinations



You hit a plateau at 10 features.

Accuracy twice better

These 10 features do not make sense to an astronomer

$$u_{psf} - g_{petr}$$
$$dered(z_{pdf}) - dered(i_{petr})$$
$$dered(g_{psf}) - dered(r_{mod})$$
$$dered(r_{psf}) - dered(z_{mod})$$
$$\sqrt{\sigma^2_{g_{petr}} - \sigma^2_{r_{model}}}$$
$$dered(r_{mod}) - dered(i_{mod})$$
$$i_{psf} - i_{petr}$$
$$dered(z_{psf}) - dered(r_{petr})$$
$$g_{mod} - g_{petr}$$
$$\sqrt{\sigma^2_{g_{petr}} - \sigma^2_{r_{petr}}}$$

$u_{psf} - g_{petr}$

$dered(z_{pdf}) - dered(i_{petr})$

$dered(g_{psf}) - dered(r_{\mathrm{mod}})$

$dered(r_{psf}) - dered(z_{\mathrm{mod}})$

$\sqrt{\sigma^2_{g_{petr}} - \sigma^2_{r_{\mathrm{mod}el}}}$

$dered(r_{\mathrm{mod}}) - dered(i_{\mathrm{mod}})$

$i_{psf} - i_{petr}$

$dered(z_{psf}) - dered(r_{petr})$

$g_{\mathrm{mod}} - g_{petr}$

$\sqrt{\sigma^2_{g_{petr}} - \sigma^2_{r_{petr}}}$

Afterwards … astronomers may find explanations ….
(Capak, private comm.)

Filter leaks, etc…

Lesson to be learned

Features which carry most of the information are not those usually selected by the astronomer on the basis of his/her personal experience….

Let the data speak for themselves ?

Feature Selection

Behind the concept of Feature Selection, there is the property of **feature importance and relevance** in the context of a parameter space used to approach any prediction/classification task with machine learning methodology.

The **importance** of a feature is the relevance of its informative contribution to the solution of a learning problem.

An effective FS should avoid the time-consuming exhaustive exploration of the parameter space and should take into account what is known about its features, i.e. their variability in the given knowledge base domain, not forgetting to take care of the curse of dimensionality problem.

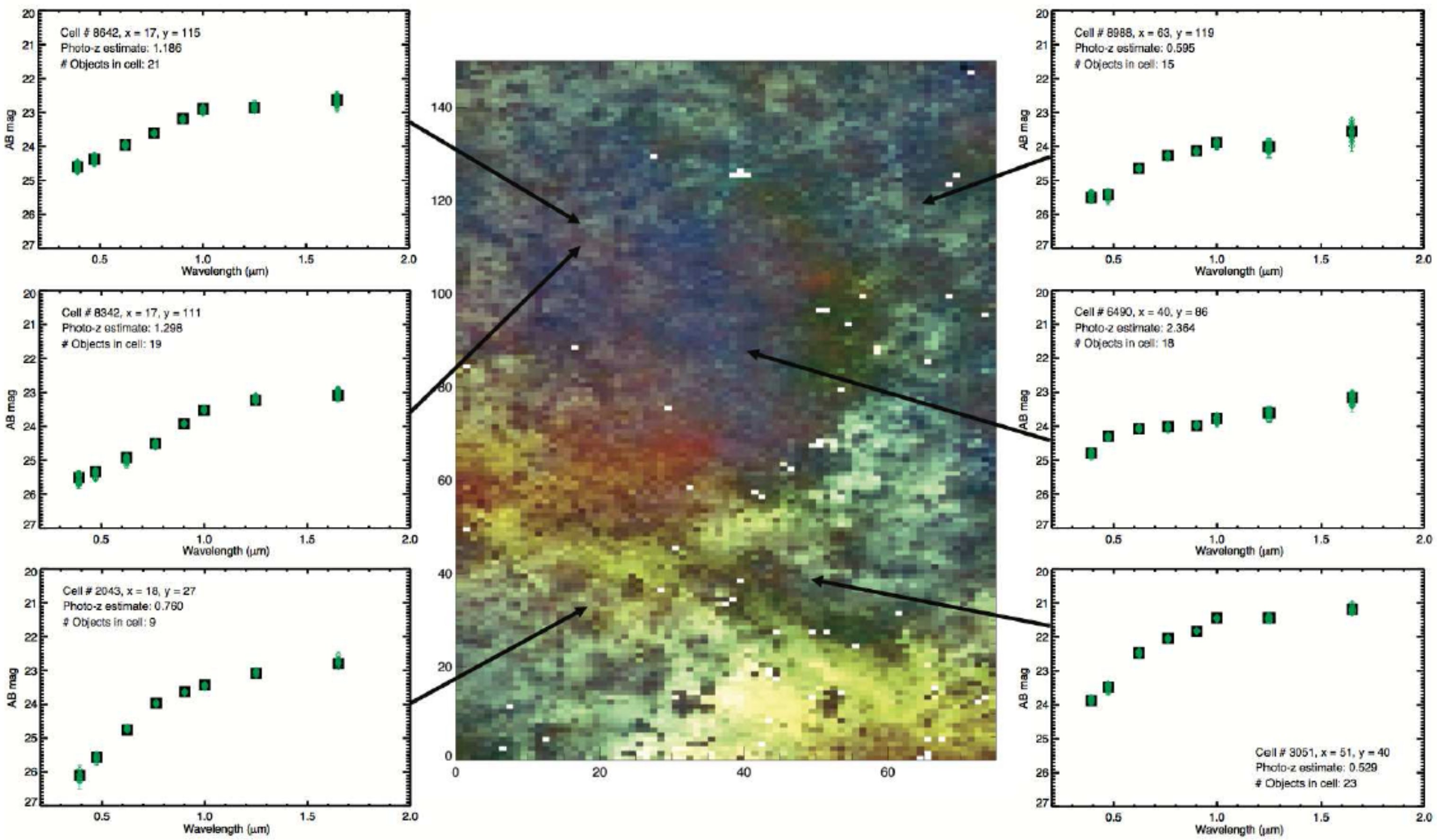
We have designed a FS method (*Brescia et al., in prep.*), based on a combination of Random Forest, Logistic Regression and L_x-norm regularization, able to overcome known statistical limitations of importance obtained by Random Forest, and by exploiting the virtuous regression control mechanism induced by the regularization concept, as already positively experimented in the learning rule of our MLPQNA neural network method (*Brescia et al. 2013, ApJ 772, 2, 140*).

We started to validate such method in some astrophysical contexts, resulting highly promising, for example, in the star forming evolutionary classification problem (**see talk of S. Molinari, presented yesterday**) and currently under test in the COSMOS galaxy photo-z and multi-survey (from UV to NIR) quasar photo-z prediction use cases.

Training set coverage of OPS

Masters et al., 2015, Astrop. Journal,

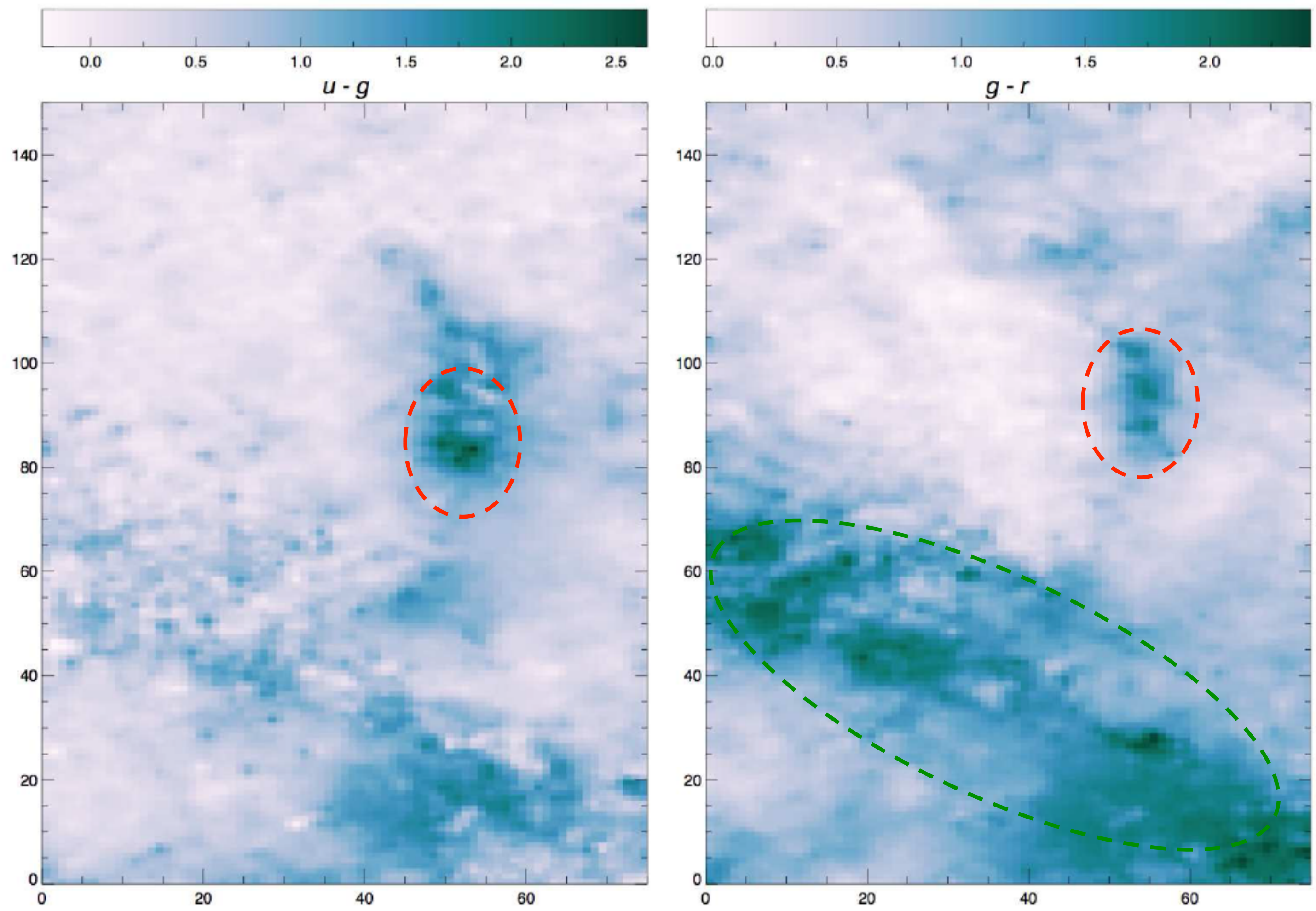
Exploring the parameter space using SOM



75 x 170 SOM

COSMOS data (EUCLIDISED)

OPS: u,g,r,i,z,Y,J,H



Ly –alpha break
u-g at 2.5<z<3.0
g-r at 3<z<4

Passive and dusty galaxies at low redshift

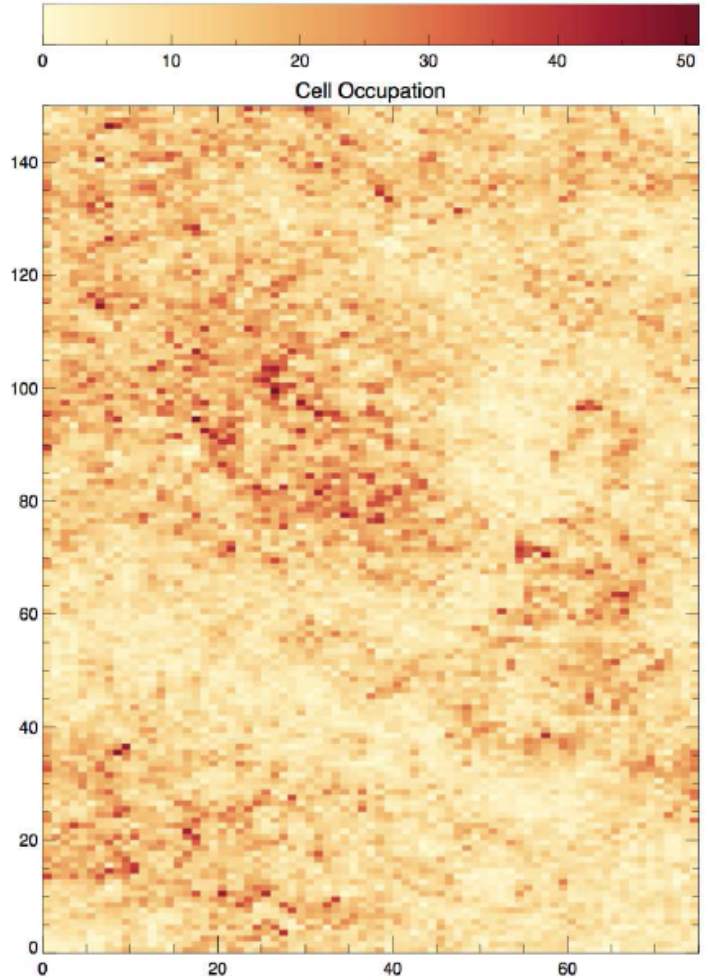
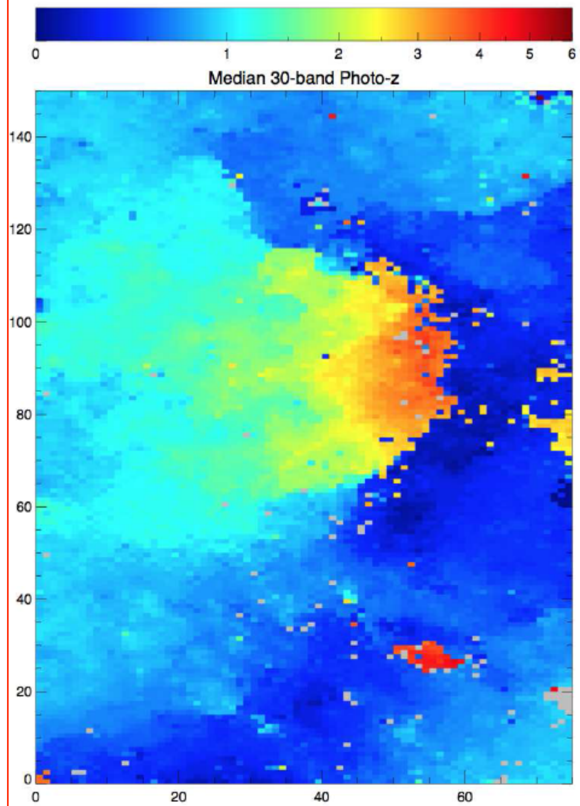


FIG. 3.— The SOM colored by the number of galaxies in the overall sample associating with each color cell. The coloration is effectively our estimate of $\rho(\vec{C})$, or the density of galaxies as a function of position in color space.

How the training set populates the “Euclid” parameter space

Poor coverage of many areas.

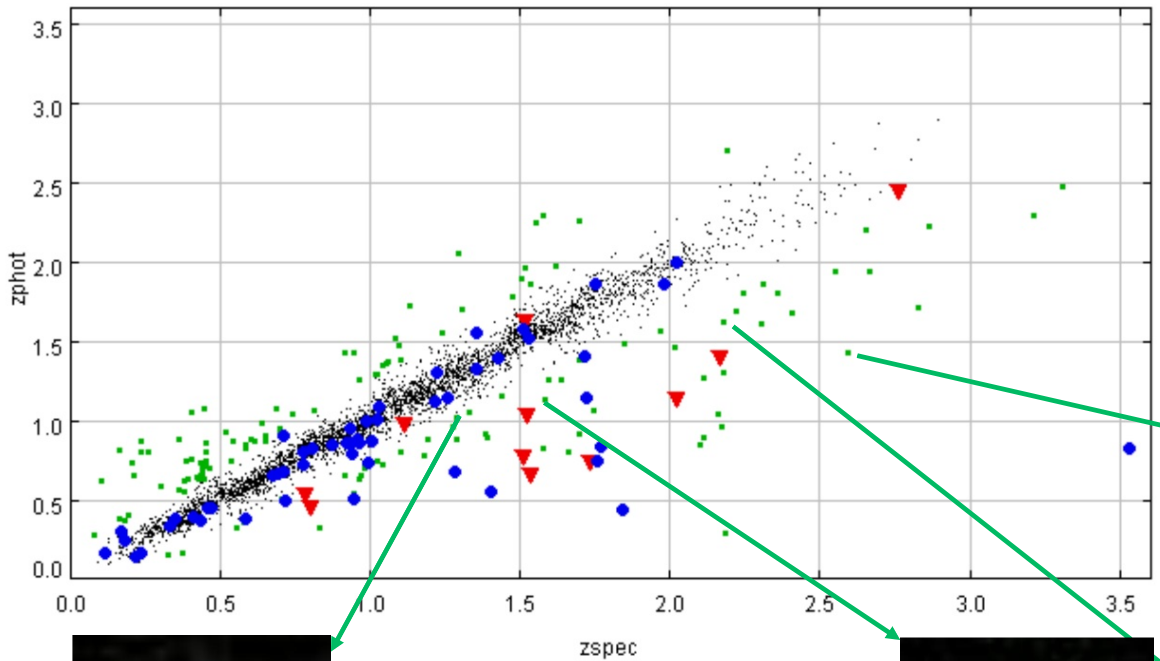


NO data….
… NO Results

Distribution of redhifts projected on the SOM

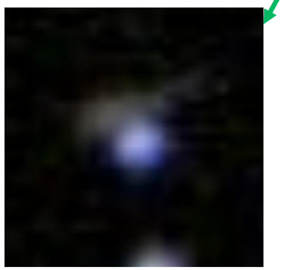
Catastrophic outliers as peculiar objects ?

(Petrillo Laurea Thesis 2013, University of Naples)

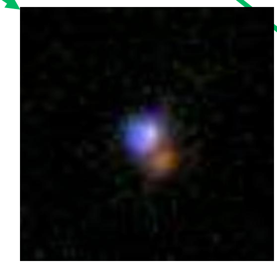


- Blu dots: blazars
- Green dots: unknown CO's
- Red triangles: gravitationally lensed quasars

Gravitational lens candidates



Peculiar objects



How about standard quality flags?

SDSS provides a complete set of quality flags extrapolated from astronomers expertise

PSF_FLUX_INTERP	8%	21%
INTERP_CENTER	10%	29%
DEBLEND_NOPEAK	0%	3%
science_primary=0	11%	24%
nuv_flags	11%	18%
fuv_artifact	18%	24%

Inspection of flags for CO's shows that these flag are practically useless to discriminate CO's

SOME IMPORTANT FLAGS ARE MISSING IN DBs…. For instance:

SO FAR NO CHECK FOR DEPENDENCE ON VARIABILITY (AGN)

Most studies on SDSS which is almost simultaeous in all optical bands

Crosscorrelation with other catalogues to check for variability (e.g. CRTS)

Very bad problem (poorly explored)

Future surveys will produce non optically selected samples (largely dominated by AGN)

Result on EMU like sample extracted from COSMOS (Salvato M. et al. 2017, in preparation)

Sample dominated by radio loud and X ray detected AGN

16 experiments with a variety of ML and SED fitting methods

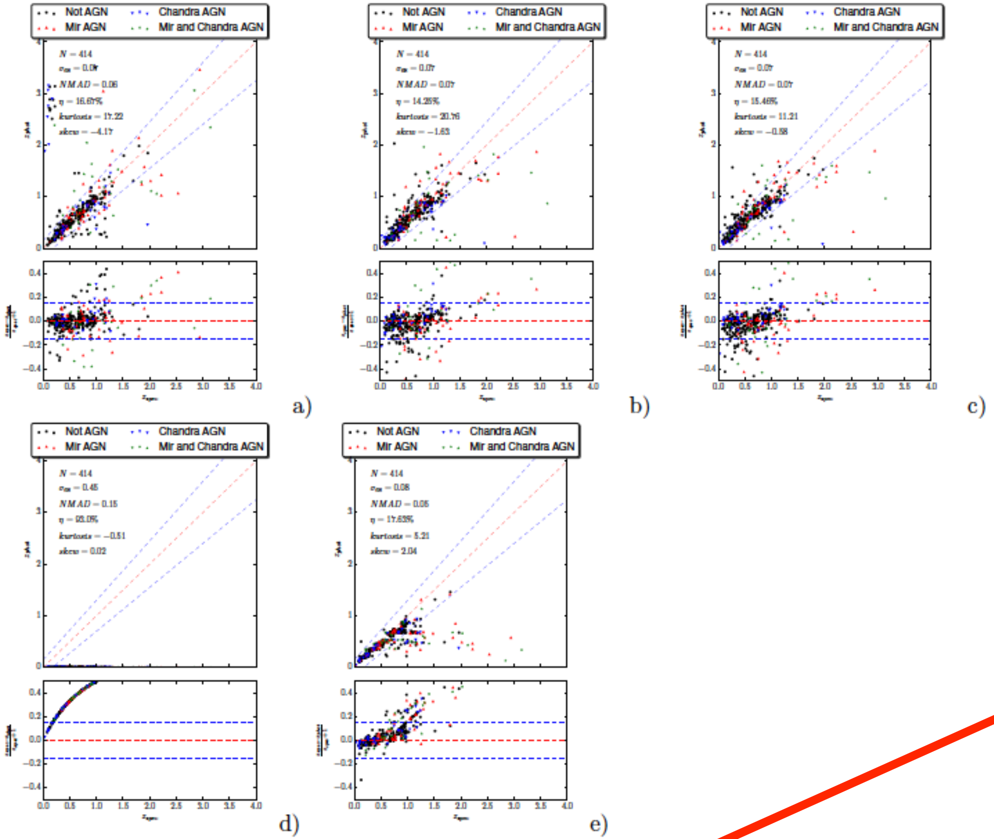
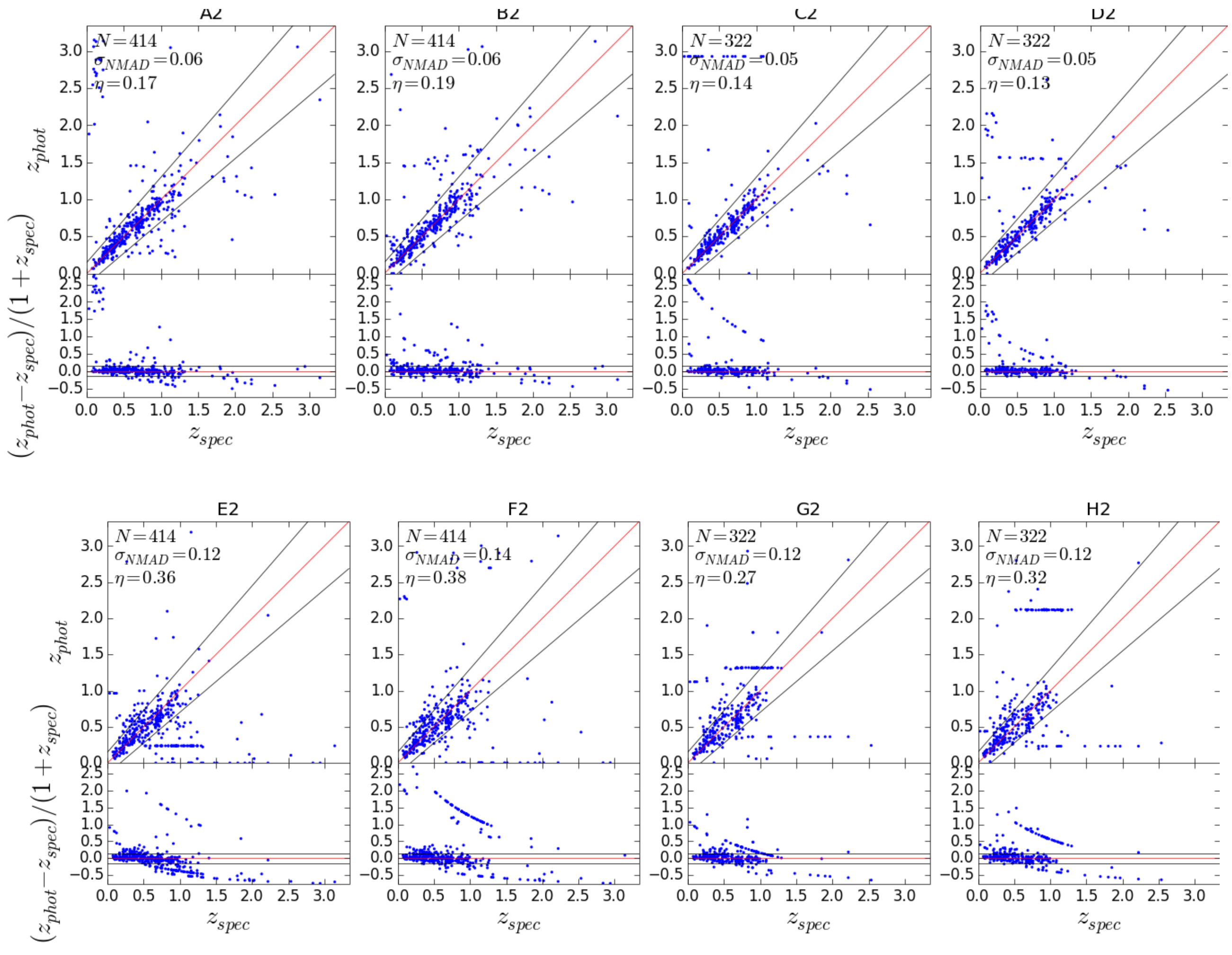


Fig. 13.— Summary of the results obtained in the experiment (RDNY) with the various methods. Performances are estimated on the blind set. Panel a) MLPQNA. Upper plot: scatter plot of spectrscopic redhifts for objects in th etest set against the photometric redshift estimate. Lower plot: normalised residuals against redshifts. Panel b) same as for panel a but for RF-NA. Panel c): Carliles. Panel e: Zinn.

Id.	CODE	KB bias	depth	Radio	X-AGN
A1	BDNY	BR	DEEP	N	Y
B1	BDYY	BR	DEEP	Y	Y
C1	BDNN	BR	DEEP	N	N
D1	BDYN	BR	DEEP	Y	N
E1	BSNY	BR	SHAL	N	Y
F1	BSYY	BR	SHAL	Y	Y
G1	BSNN	BR	SHAL	N	N
H1	BSYN	BR	SHAL	Y	N
A2	RDNY	RND	DEEP	N	Y
B2	RDYY	RND	DEEP	Y	Y
C2	RDNN	RND	DEEP	N	N
D2	RDYN	RND	DEEP	Y	N
E2	RSNY	RND	SHAL	N	Y
F2	RSYY	RND	SHAL	Y	Y
G2	RSNN	RND	SHAL	N	N
H2	RSYN	RND	SHAL	Y	N

Table 2: Summary of the experiments. Column 1: running id; column 2: identification code; column 3: Bright (BR) or Random (RND) training set; column 4: shallowness of ancillary data; column 5: radio fluxes used (Y) or not used (N) in training; column 6: bright X ray detected AGN included (Y) or not included (N) in the training set. **General consideration: while working on the**

MLPQNA

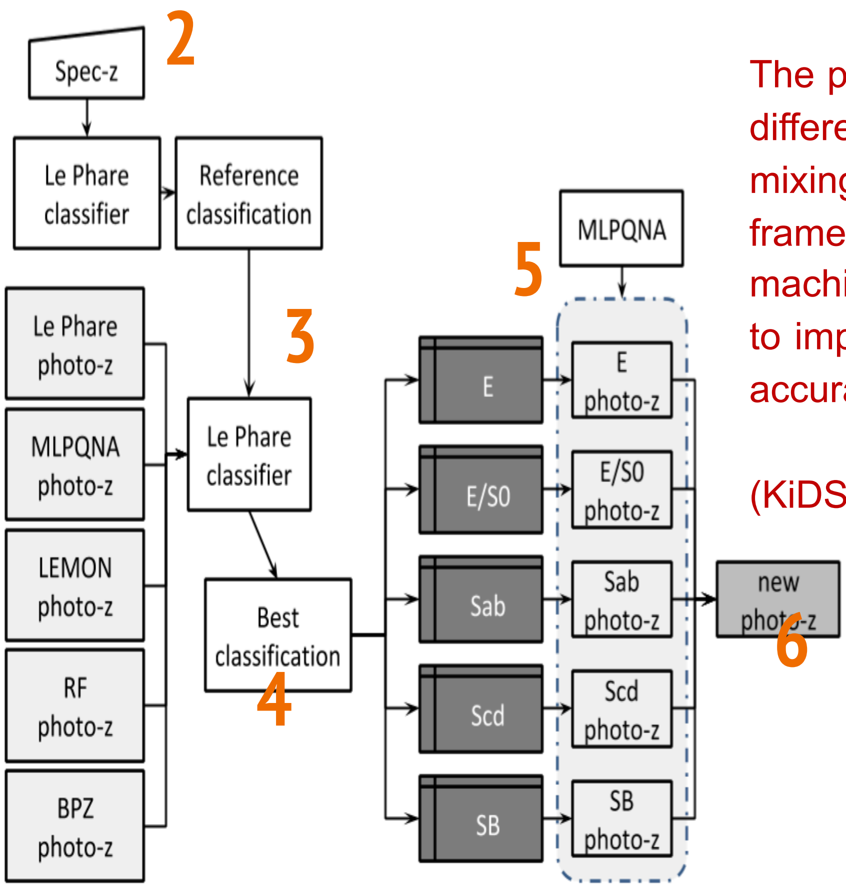


Id.	CODE	KB bias	depth	Radio	X-AGN
A1	BDNY	BR	DEEP	N	Y
B1	BDYY	BR	DEEP	Y	Y
C1	BDNN	BR	DEEP	N	N
D1	BDYN	BR	DEEP	Y	N
E1	BSNY	BR	SHAL	N	Y
F1	BSYY	BR	SHAL	Y	Y
G1	BSNN	BR	SHAL	N	N
H1	BSYN	BR	SHAL	Y	N
A2	RDNY	RND	DEEP	N	Y
B2	RDYY	RND	DEEP	Y	Y
C2	RDNN	RND	DEEP	N	N
D2	RDYN	RND	DEEP	Y	N
E2	RSNY	RND	SHAL	N	Y
F2	RSYY	RND	SHAL	Y	Y
G2	RSNN	RND	SHAL	N	N
H2	RSYN	RND	SHAL	Y	N

Concept Idea – virtuous cooperation between SED fitting and ML

Cavuoti et al. 2017, MNRAS 466, 2

1. Derive traditional photo-z's with all methods;
2. Use Le Phare bounded with spec-z's to obtain a reference classification;
3. Use Le Phare bounded with photo-z's to perform a series of classifications;
4. Identify the best classification using as ground truth the reference classification (step 2);
5. Perform a photo-z regression by training MLPQNA on separated subsets specific for each class;
6. Recombine the output.



The proposed workflow, involving different methodologies by mixing in a single collaborative framework SED fitting and machine learning models, is able to improve the photo-z prediction accuracy by ~10%.

(KiDS-DR2 data)

How to take into account photometric, initialization errors, and model dependent errors to produce a pseudo-PDF

SED fitting produces pseudo-PDFs using the fits to the different templates

ML methods need a different approach

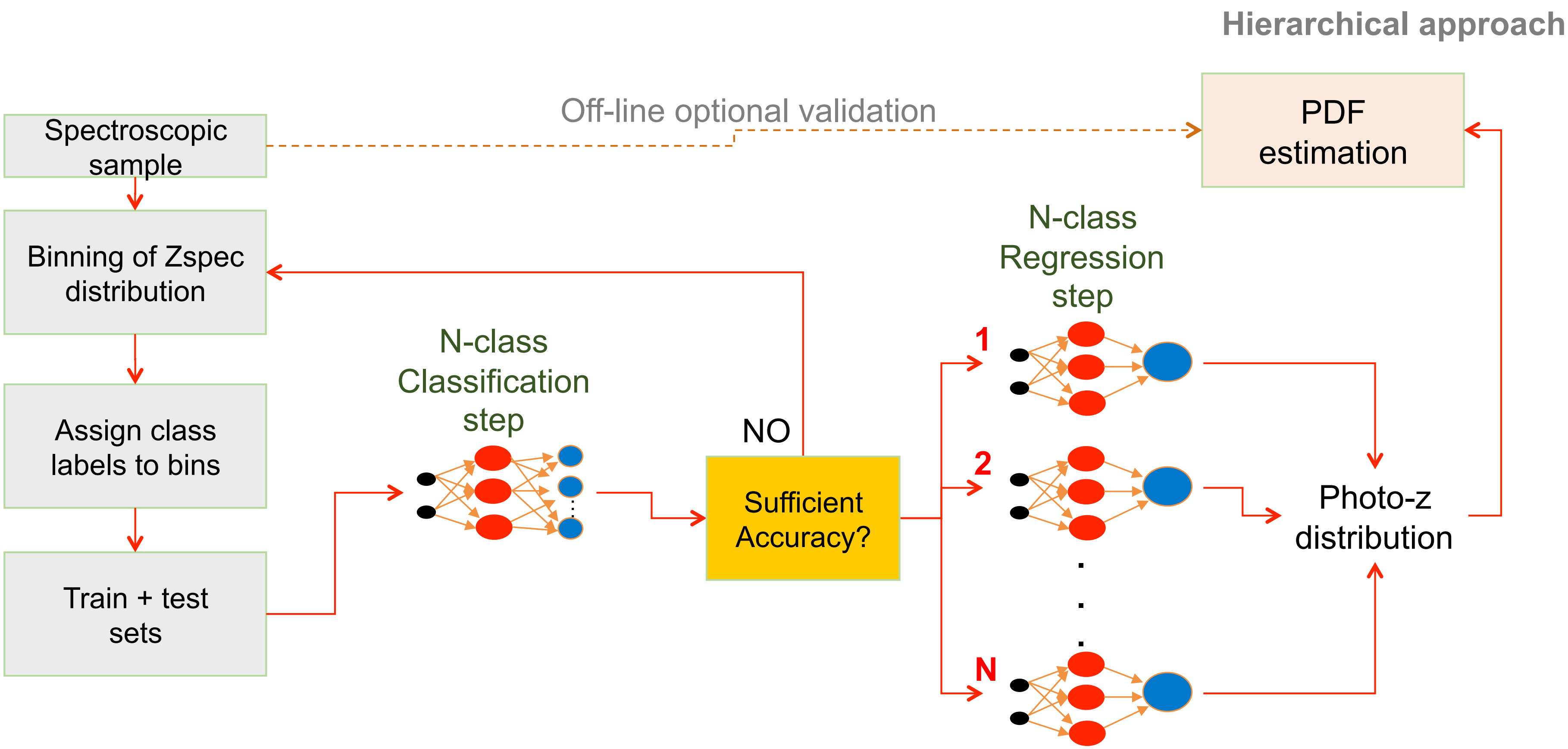
- Internal errors (initialization of weights)
- Photometric errors
- Errors in the KB (misclassified objects, poor coverage of OPS, peculiar objects, etc)

→

METAPHOR

Brescia, Cavuoti, Amaro, Vellucci & Longo 2016, 2017 (in prep)

PDF base algorithm processing flow



An application to KiDS (Kilo Degree Survey- KiDS)

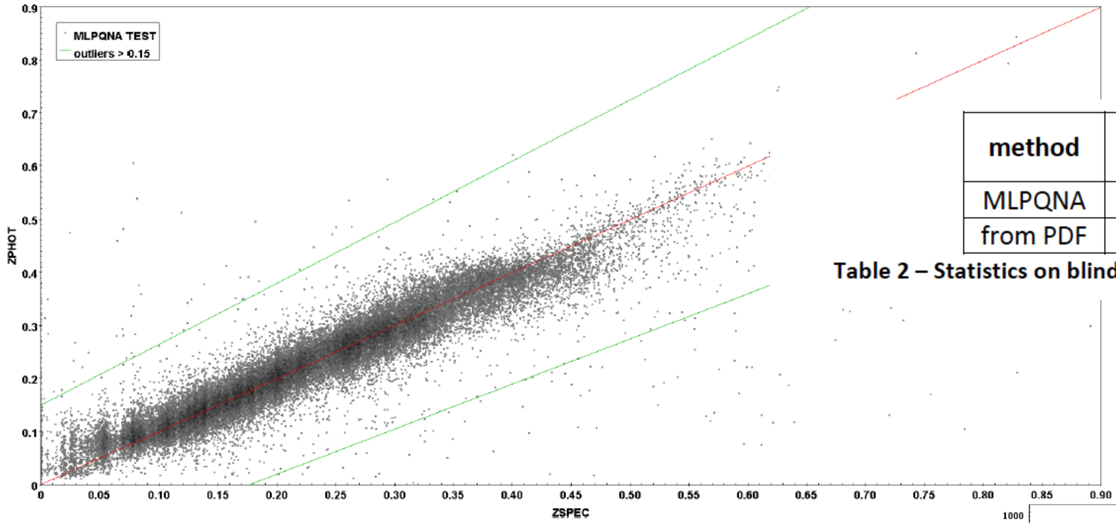


Figure 3 – MLPQNA based photo-z VS zspec plot for the blind test set. Green lines are referred to the outl (@0.15).

method	bias $\vert\Delta z\vert$	bias $\vert\Delta z/(1+z)\vert$	σ_{68} $\Delta z/(1+z)$	NMAD $\Delta z/(1+z)$	outliers % $\vert\Delta z/(1+z)\vert>0.15$
MLPQNA	0.00010	0.00088	0.021	0.021	0.40
from PDF	0.0086	0.0063	0.022	0.021	0.39

Table 2 – Statistics on blind test set for the photo-z estimates with MLPQNA, before and after PDF calculation.

- $f_{0.05}$ is the integral of the stacked PDF within the interval (in units of z) [-0.05, +0.05][2] = 92.8%
- $f_{0.15}$ is the integral of the stacked PDF within the interval (in units of z) [-0.15, +0.15][3] = 99.6%
- $\langle\Delta z\rangle$ is the bias of the stacked PDF weighted on its frequency[4] = -0.0014

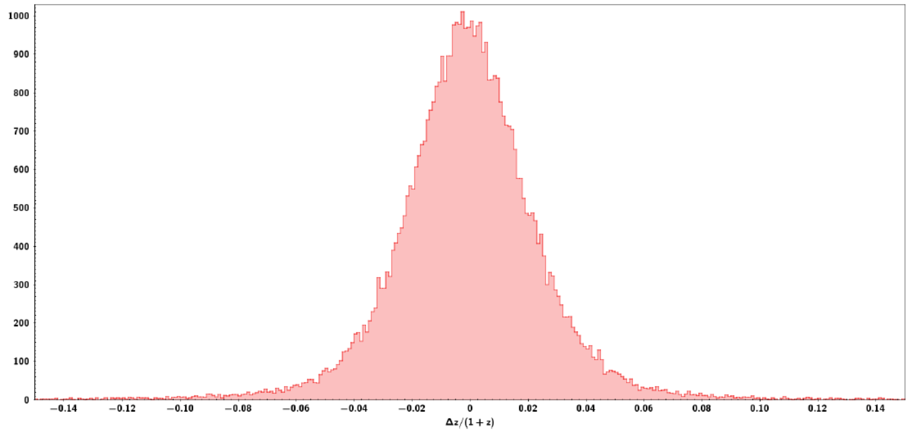


Figure 4 – Distributions of Δz/(1+z) for the photo-z obtained by MLPQNA.

1. **Machine Learning is an ART based on hard work and a deep understanding of each step involved in the process**
   (i.e. IT CANNOT BE IMPROVISED just because there are user friendly packages available)…. The simpler is the method the more difficult is to obtain robust and stable results…

- Need to take into account a priori information
- Need to have a deep understanding of the data themselves (selection effects introduced by previous classification steps)
- Combination of various methods can help

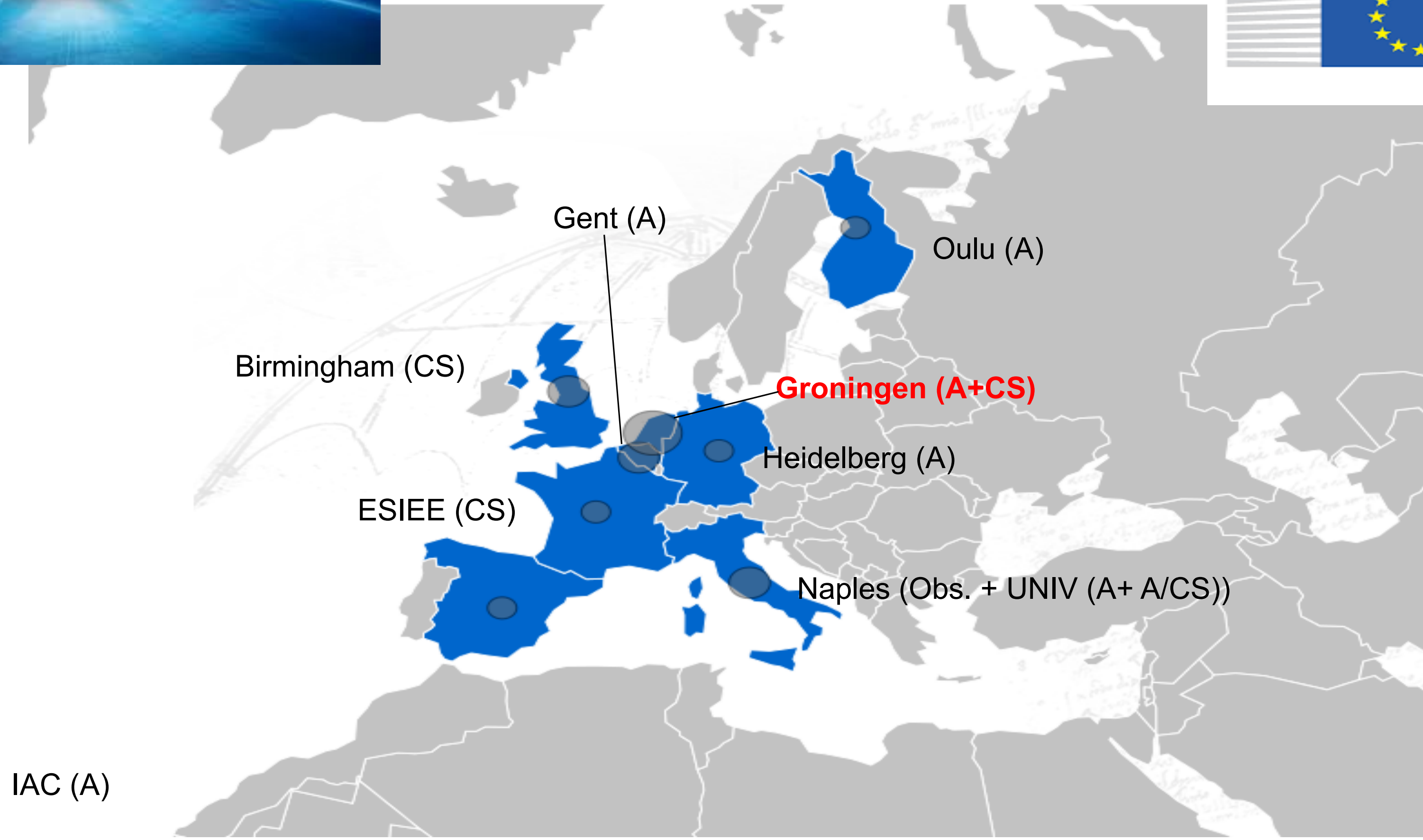
2. To optimise the use of ML in future surveys we need:

- to redefine the way we measure the observable parameters (very probable) and assign quality flags (definitely true)
- to optimise the coverage of the parameter space via specific spectroscopic campaigns (true)
- large computing power for feature selection phase (true) and smarter algorithms for FS

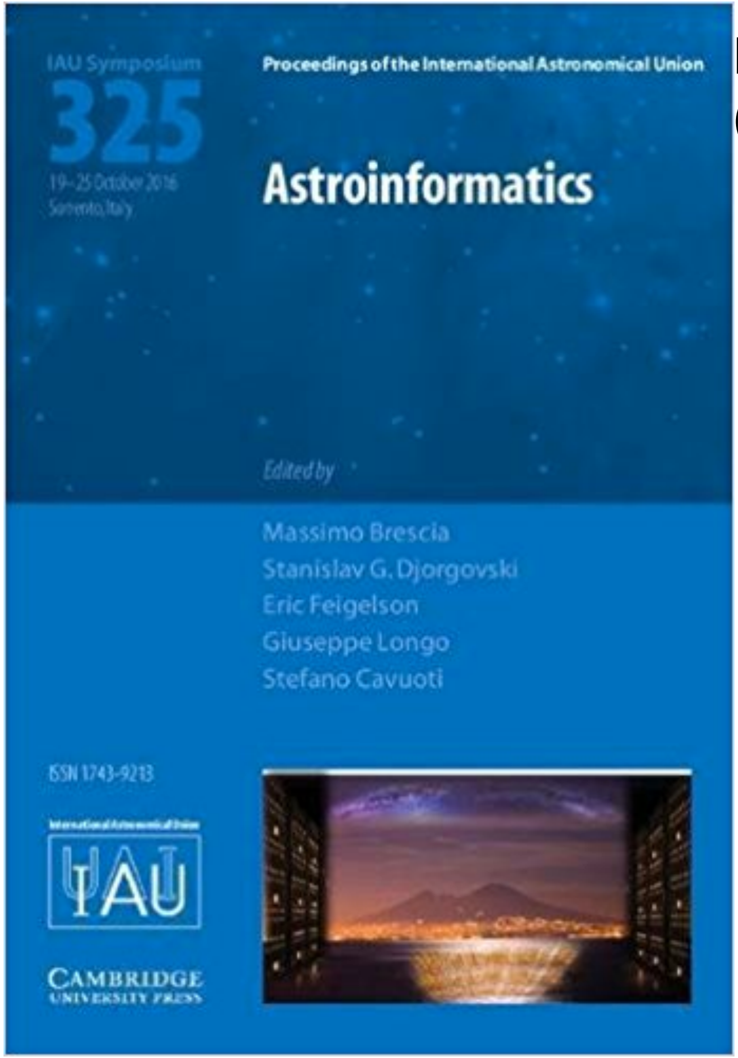
3. Suggestions to end users.

- Watch out for statistical indicators.... Often they do not mean much
- Check for biases in the imput catalogues

SUNDIAL Partners



First Astroinformatics ITN
Funded by EU



Proceedings of the IAU Symposium n. 325 (Astroinformatics 2016) - just published

Astroinformatics 2017

7-10 November

South Africa

Thank you