This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 721463

SUNDIAL
H2020 Innovative Training Network

DATA SCIENCE UNINA

INAF
ISTITUTO NAZIONALE DI ASTROFISICA
NATIONAL INSTITUTE FOR ASTROPHYSICS

ARTIFICIAL INTELLIGENCE IN ASTRONOMY

I.e. Machine Learning successes and problems

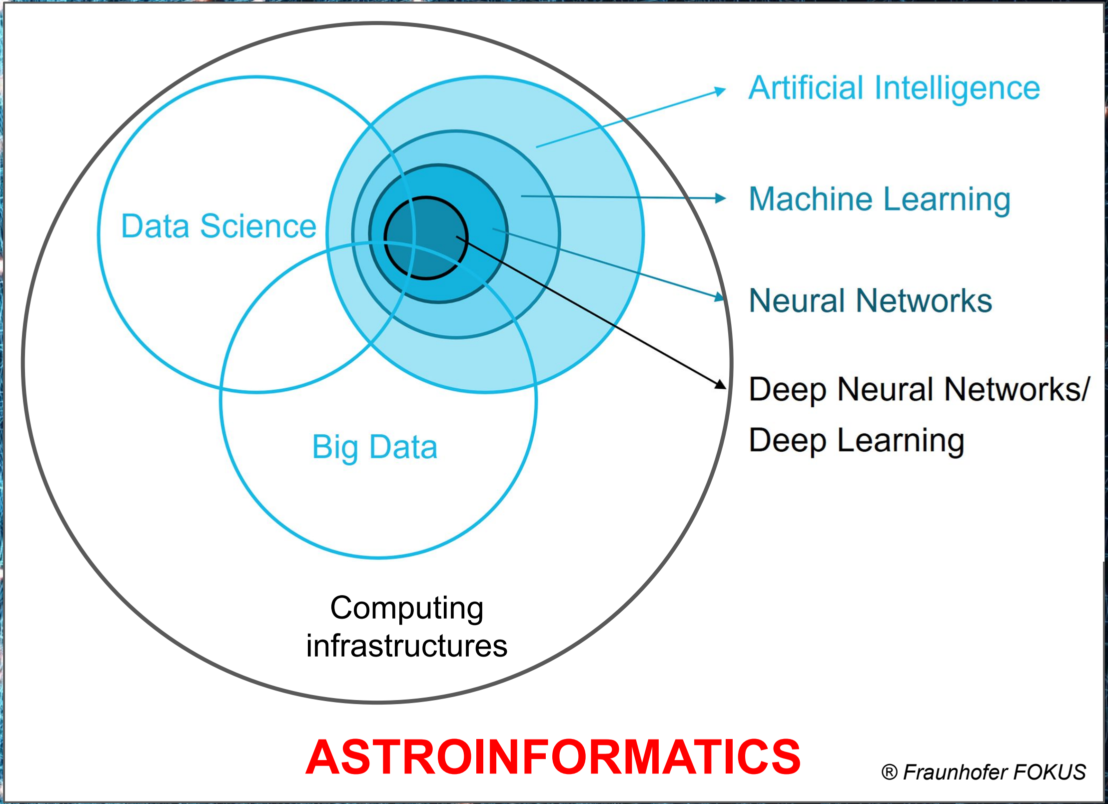
Giuseppe Longo
DATA SCIENCE INITIATIVE
University of Napoli Federico II - Italy
CINI Consortium
longo@na.infn.it

Special Thanks to the group:

Massimo Brescia
Stefano Cavuoti
Michele delli Veneri
Giuseppe Longo
Oleksandra Razim
Giuseppe Riccio
Olena Torbaniuk

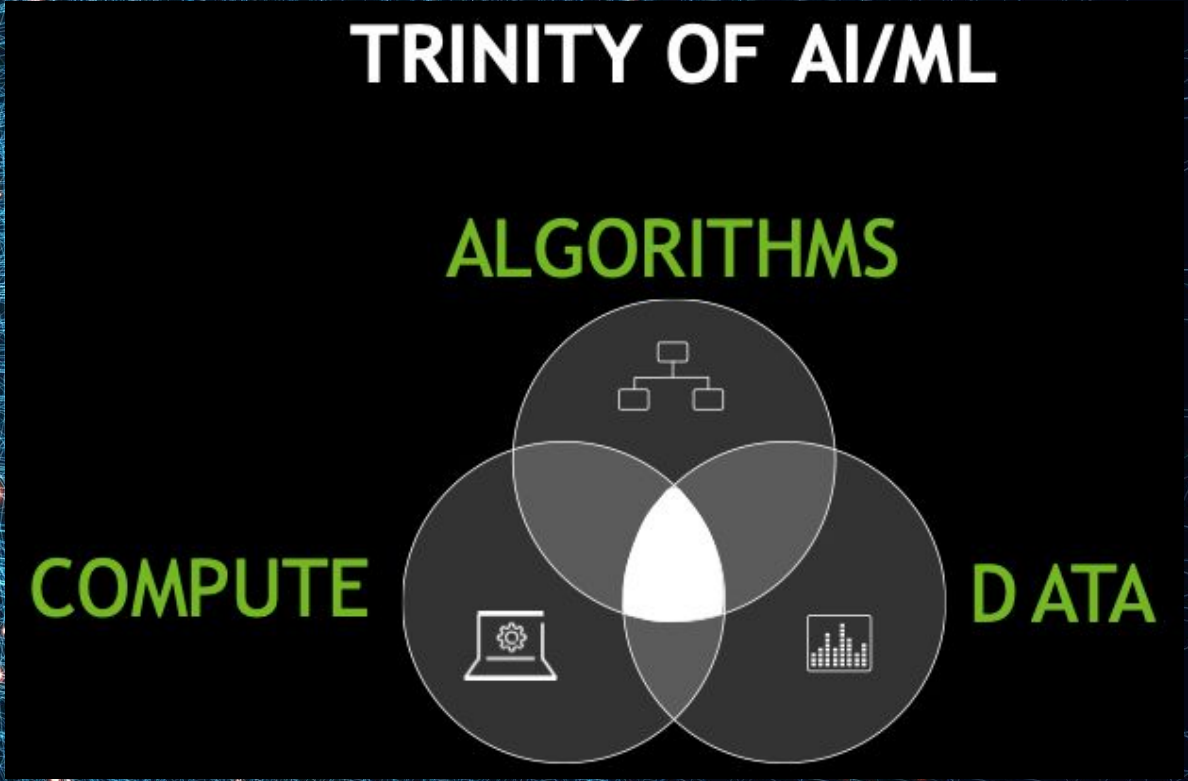
\+ & Many great students

Personal considerations



- **Artificial Intelligence is just a buzzword** (recently resurrected for marketing purposes)
- **Deep learning** is a subset of machine learning
- Machine learning, data mining, KDD, and statistical pattern recognition are different "nuances" of the same stuff

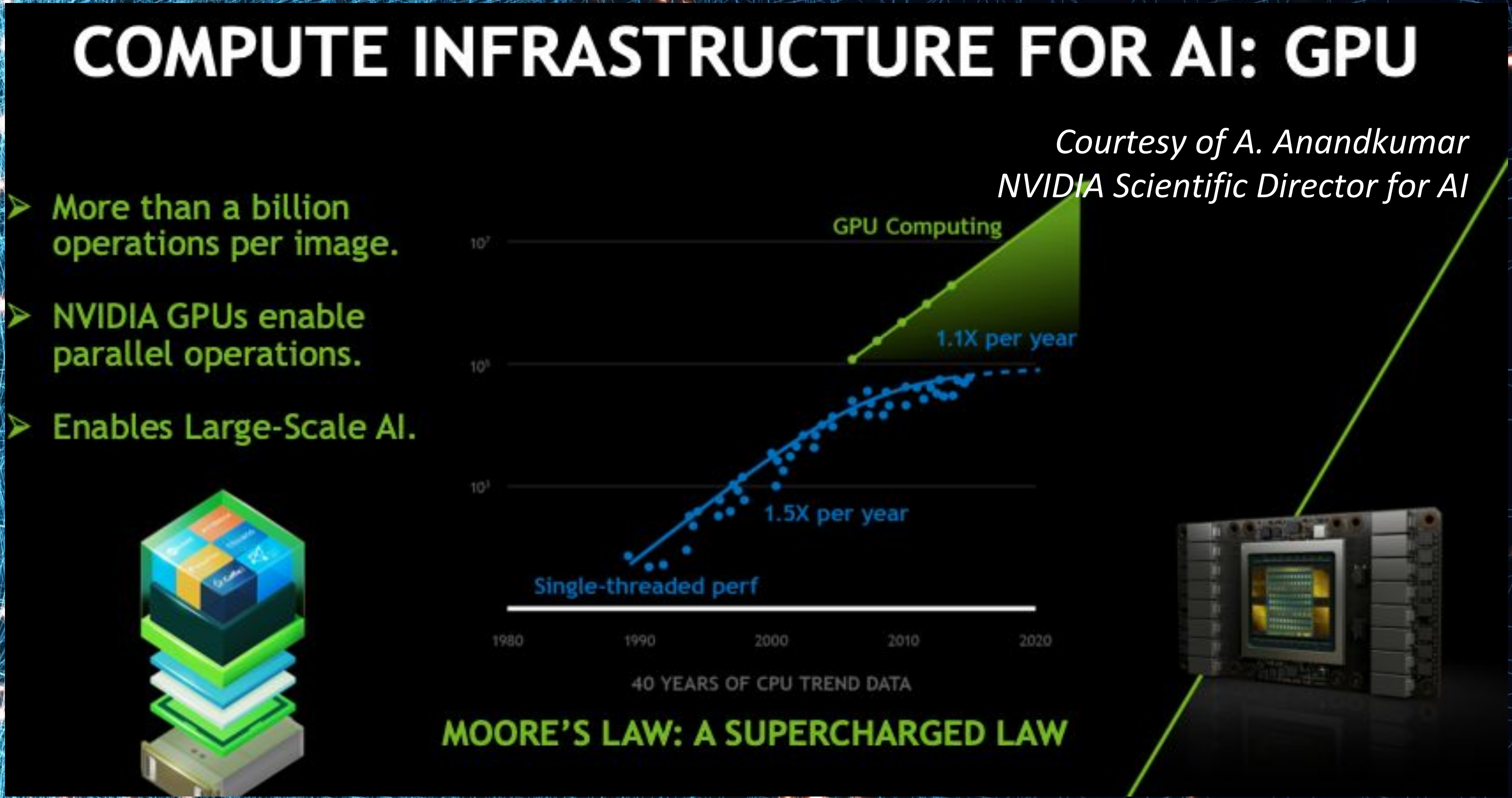
The trinity of AI/ML



DATA TYPES

Type of data	2003	2003-2009	2009-2019	superv.	unsuperv.	DL
Tabular data (vectors)	Yes	Yes	Yes	Y	Y	Y
Time Series	Yes	No	Yes	Y	Y	Y
Astrometry			Yes	Y	Y	?
Images (1 band)	Yes	yes	Yes	Y	Y	Y
Multiband			Yes	Y	Y	Y
Spectra	Yes	yes	Yes	Y	Y	Y
Data Cubes			Yes	Y	Y	Y
Simulations	no	no	yes	y	y	y

1. Computing



For "money rich" communities coping with "data rich" problems, computing is **NOT YET** an unsolvable problem (SKA 2 ???)

The astroinformatics field is exploding (2003 vs 2019)

2003 - Special issue of the International Journal of Neural Networks on "Neural Network Analysis of Complex Scientific Data", Eds. Tagliaferri R., Longo G., D'Argenio B.
2010 - N.M. Ball and R.J. Brunner, 2010, arXiv:0804.3413

2019 - Focus Issue on Machine Learning in Astronomy, Publications of The Astronomical Society of the Pacific, Eds. Longo G., Merenyi E. & Tino P.
2019 - Papers presented at "Astroinformatics 2019", Pasadena July
2019 - review (in press)

WARNING: does not cover "Bayesian" and similar approaches.

TASKS AND SCIENCE CASES - I

Task	2003	2003-2009	2009-2019	superv.	Unsuperv.	DL	Notes
S/G separation	yes	Yes	yes	Y	y	?	ANN, CNN
Galaxy properties Morphology Properties SFR Evolution	yes	yes	yes	Y	y	y	ANN, SVM, PPS; CNN,
Spectral classification	yes	yes	yes	Y	y	y	ANN, SVM, RF
Image segmentation	yes		yes	y	y	y	ANN, GAN
Noise removal	yes		yes	Y	y	no	SVM, ANN
Photometric redshifts (galaxies)	yes	Yes	yes	Y	y	y	SVM, ANN, RF, CNN, KNN, + other
Variable objects	yes	Yes	yes	y	y	y	SVM, DT, ANN, RF, CNN
Stellar evolution models	yes		yes	y	n	n	ANN
Outlier detection		Yes	yes	Y	y	y	ANN, RF, CNN
Search for AGN		Yes	yes	Y		y	SVM, ANN, CNN

Task	2003	2003-2009	2009-2019	superv.	Unsuperv.	DL	Notes
Solar activity		yes	yes	Y	n	n	
Galactic studies Interstellar Medium Open clusters Stellar associations			yes	y	y	Y	GAME, ANN, GNG, DBSCAN,
Planetary studies Surface morph		yes	yes	Y	Y	n	SVM, ANN, ADABOOST, CNN
Asteroids			yes	Y		Y	CNN
Exoplanets			yes	y	y	y	DBSCAN, ANN
Gravitational lensing			yes	y		y	GAN, CNN
Dark matter			yes	Y		Y	GAN
Magnetic fields			yes	Y			ANN
Instrumentation Monitoring & control			yes	Y	Y	Y	SVM, ANN, expert systems
Data reduction and data logs			yes	Y	Y		ANN

Algorithms: open problems

- **How to evaluate performances**
  statistical indicators are not always unambiguous

- **How to evaluate effects of errors (we need PDFs)**

- Not all features are significant for the task, hence the **need to reduce dimensionality** (most relevant, all-relevant, Data Driven Approach?)

- **Proper coverage of OPS:**
  how to control biases in the training set

- **Missing data are still a problem**

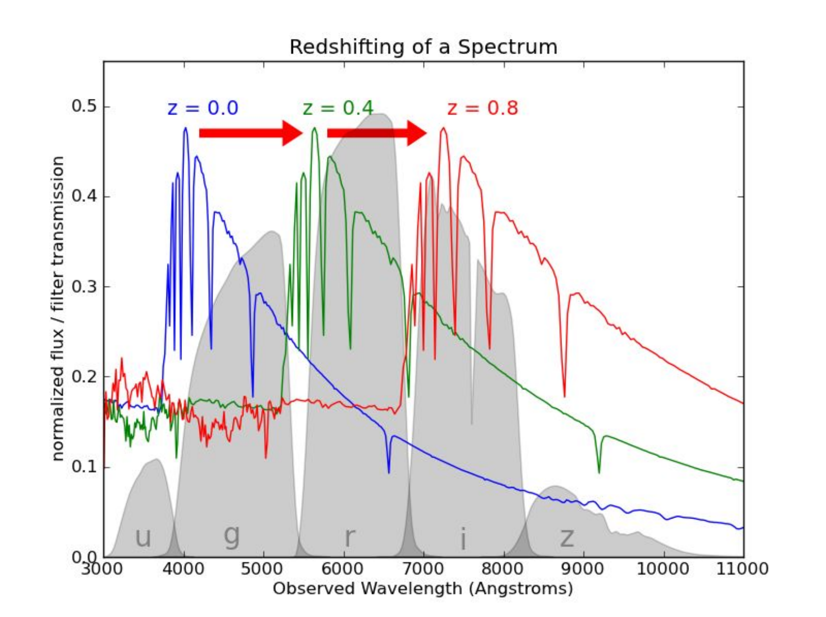


Photo-z as a template case of supervised ML

- More than 220 papers in the last 10 years
- Different surveys (almost all), many wavelengths
- Different coverages of OPS
- Wide range of science applications

Summarising the work by many:

Massimo Brescia, Stefano Cavuoti
& Valeria Amaro, Alex Razim, Giuseppe Riccio, Michele delli Veneri and others.

In theory, ML photo-z methods are simple.....

Use a set of "accurate" templates to infer the hidden function *f* which maps the vector space **X** onto the scalar z

$$f : \mathbf{X} \rightarrow z \text{ where: } \quad \mathbf{X} \equiv x_1, x_2, \ldots x_n$$

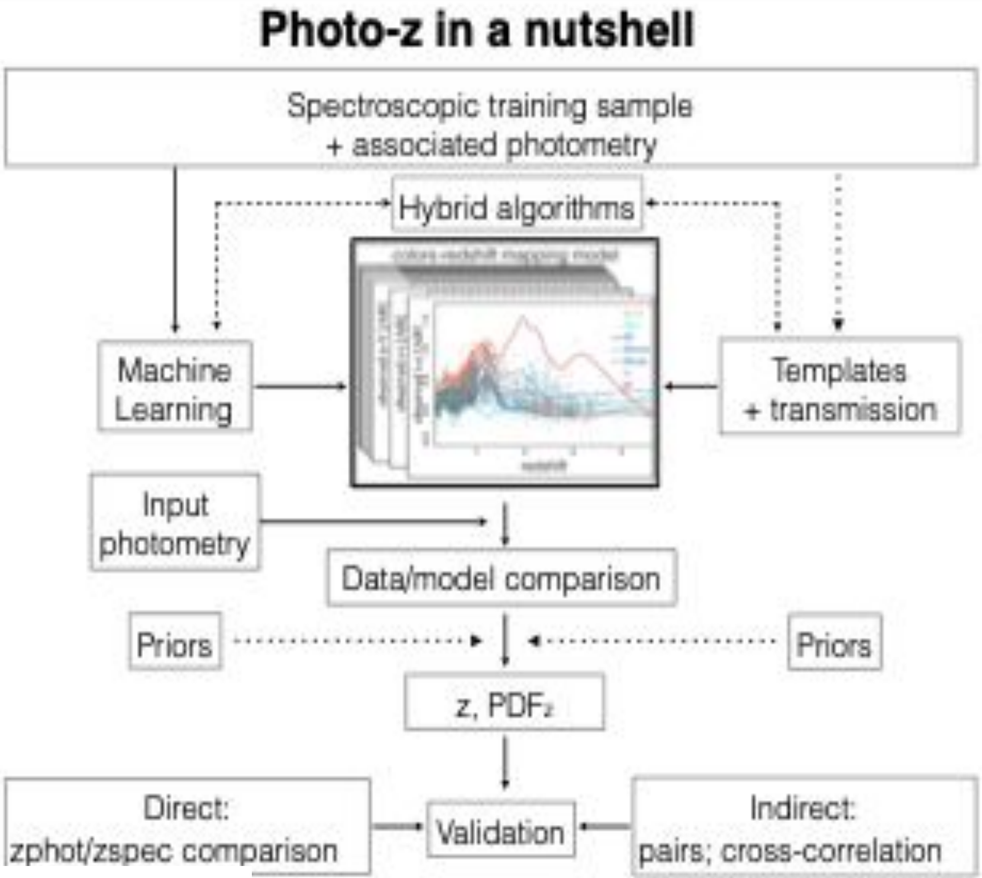
Is the vector space defined by the input features and z is the target function

Empirical methods use a subset of the objects (TRAINING SET) for which the spectroscopic redshifts (the target) are known, to infer the mapping function ***f***

Performances are then evaluated on a second disjoint dataset (TEST SET) for which the target is known and which has not been used during the training (BLIND TEST)

Usually accurate, no assumptions on underlying physics, almost independent on zero points, photometric calibrations, etc.

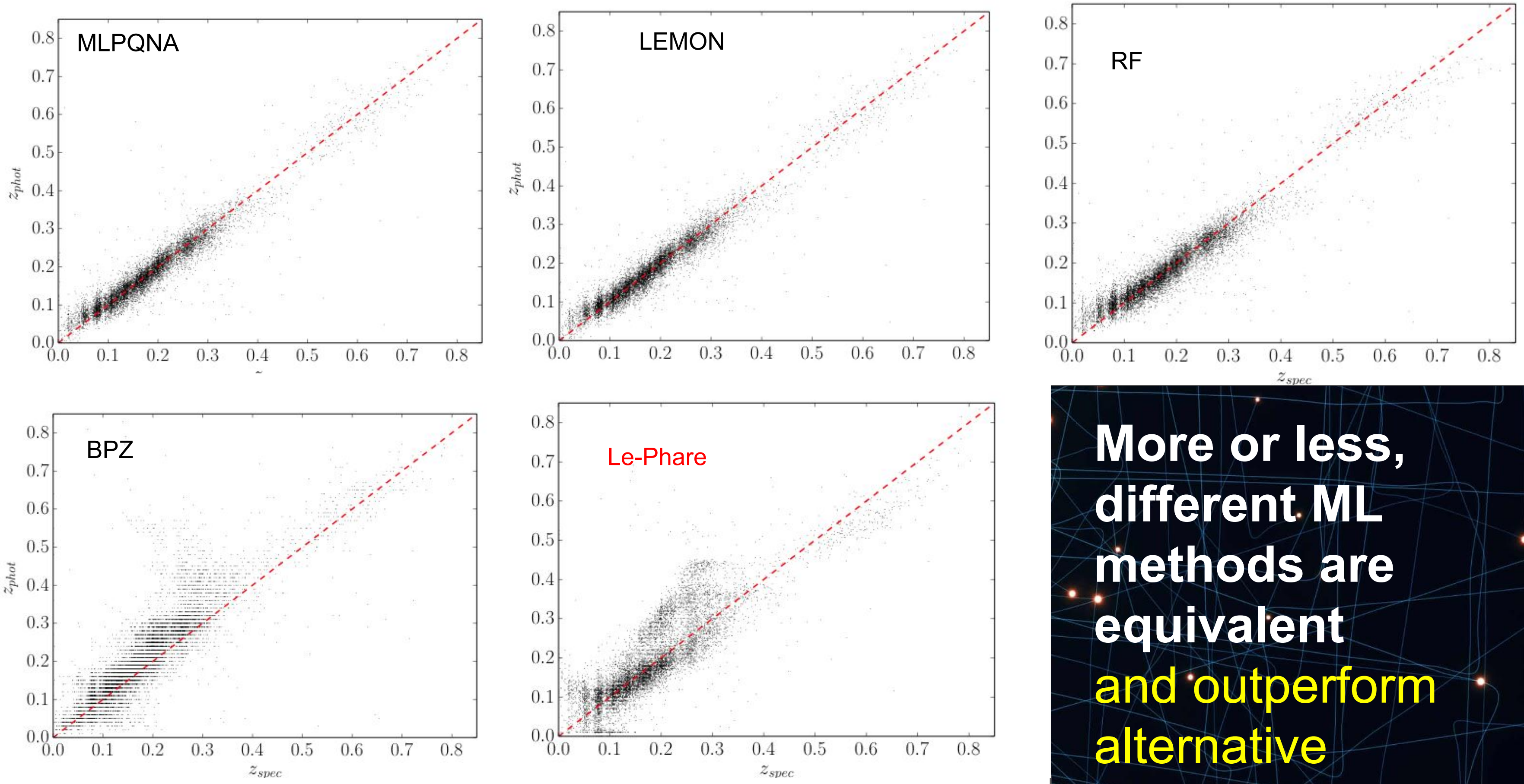
They are limited to the portion of the parameter space covered by the training set. Many problems in dealing with errors



Salvato, Hoyle 2018

DATA RICH REGIME (large training set)

All methods have been applied: decision trees, random forest, SVM, SOM, MLP in different nuances, genetic algorithms, deep learning, etc...



E.g. Cavuoti, et al., MNRAS, 2016 on KiDS data

More or less, different ML methods are equivalent and outperform alternative approaches

DATA RICH REGIME

ALL METHODS PERFORM WELL, BUT....

- **FEATURE SELECTION**

  Modern digital surveys produce huge amounts of measured parameters (e.g. SDSS ca. 550, KiDS more than 400, etc.)

  Merging more surveys makes the number of parameters explode.

  Number of examples is and will be forcefully limited

  different strategies to cope with it but no clear cut, unique solution....

FEATURE SELECTION

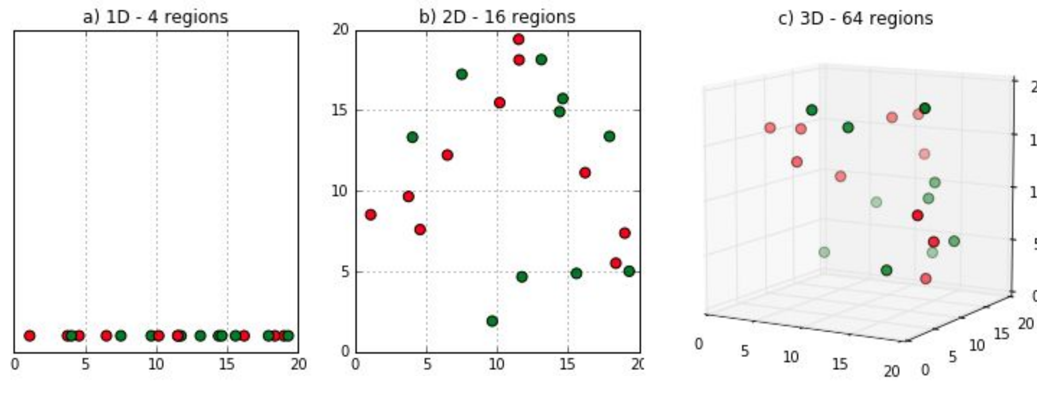
Finding optimal number and combination of parameters for a given task

Increasing the number of parameters means that the density of training points (examples) decreases
This leads to a loss in interpolation capabilities

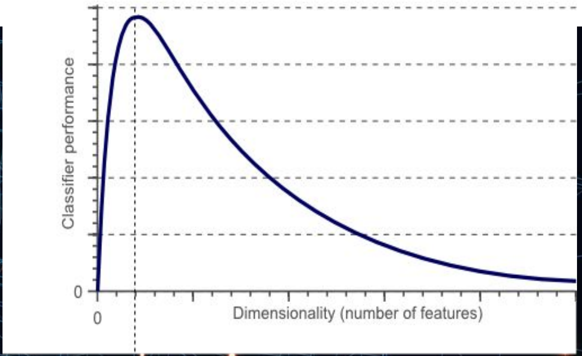
At the same time the <u>volume of an inscribing hypersphere</u> of dimension d and with radius 0.5 can be calculated as:

$$V(d) = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} 0.5^d.$$

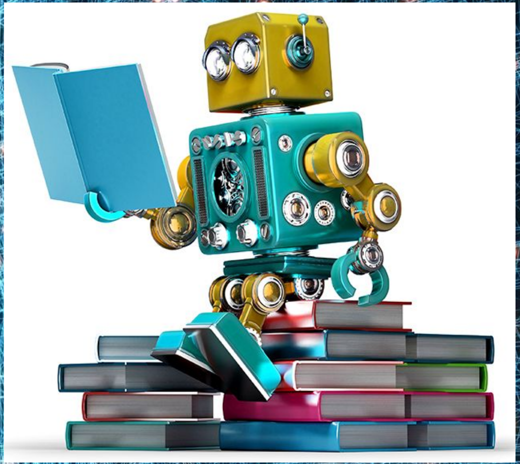
Figure shows how the volume of this hypersphere changes when the dimensionality increases:



The performance changes when the dimensionality increases, we have a peak and then a decrease, this leads to the importance of a "feature selection"



Optimal number of features



Feature selection

- Preselection based on common sense or on the **opinion of the experts**
- **Empirical** (try all) → **Most relevant**
  - **Forward selection**
  - **Data driven approach**
- **All relevant**

Brescia et al 2013, ApJ, 772, 140

Survey	Bands	Name of feature	Synthetic description
GALEX	nuv, fuv	mag, mag_iso	Near and Far UV total and isophotal mags
		mag_Aper_1 mag_Aper_2 mag_Aper_3	phot. through 3, 4.5 and 7.5 arcsec apertures
		mag_auto and kron_radius	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	PSF fitting magnitude in Y, J, H, K bands
		AperMag3, AperMag4, AperMag6	aperture photometry through 2, 2.8 & 5.7″ circular aperture in each band
		HallMag, PetroMag	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

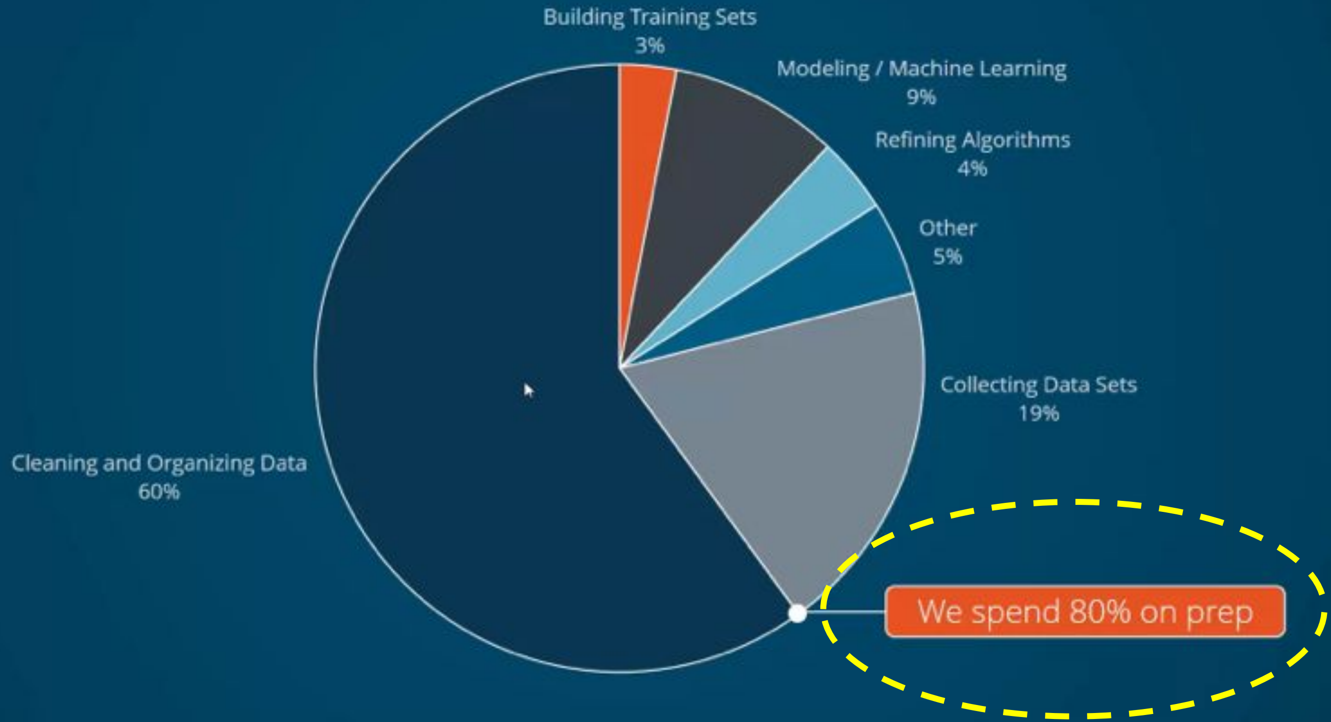
Traditional (empirical) approach:
First selection of features based on expertise
Trial and error on different combinations

Hundreds of experiments
Very demanding in terms of time

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

What data scientists spend the most time doing



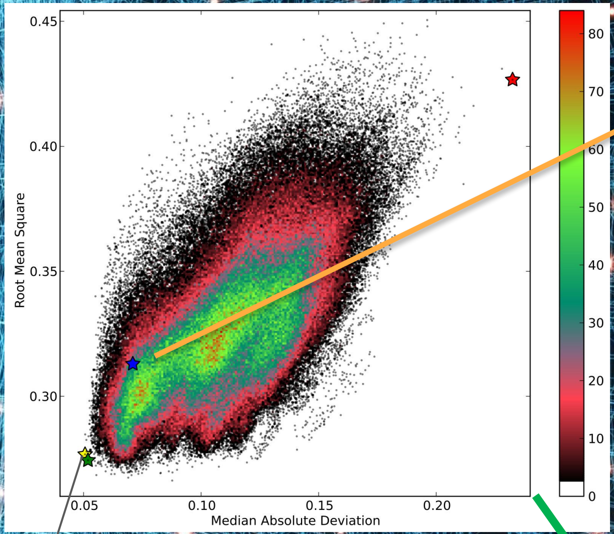
https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

A brute force approach (from K. Polsterer, Heidelberg, 2015)

QSOs from SDSS

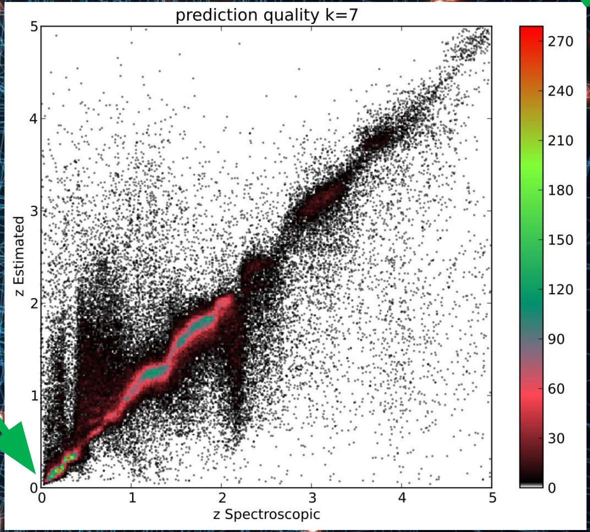
$$\frac{n!}{(n-r)!r!} = 341{,}055 \ \text{combinations}$$

- 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 2011
Traditional feature selection

Best results are comparable to Brescia et al. 2014



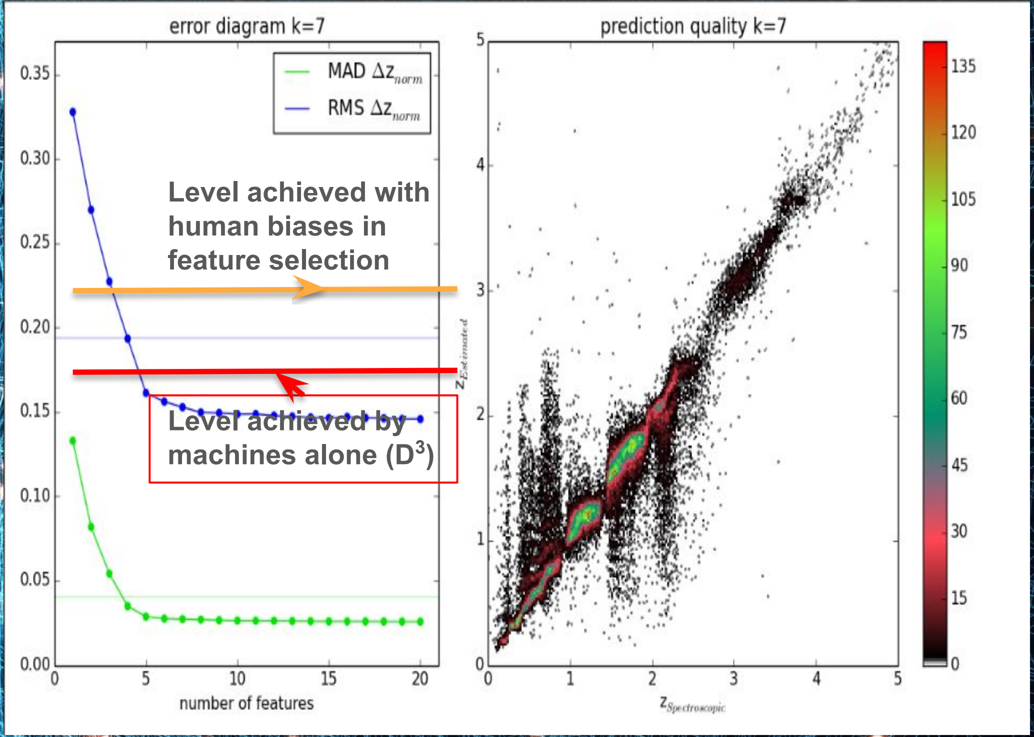
Best combination

$u_{model} - g_{model}$
$g_{psf} - r_{model}$
$z_{psf} - r_{model}$
$i_{psf} - z_{model}$

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 (therefore just add the most significant feature strategy)



You hit a plateau at 10 features.

Accuracy twice better

These 10 features do not make sense to an astronomer

(afterwards ... there might be some explanation)

$u_{psf} - g_{petr}$

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

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

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

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

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

$i_{psf} - i_{petr}$

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

$g_{mod} - g_{petr}$

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

Return of the features, D'Isanto, Cavuoti et al. 2018

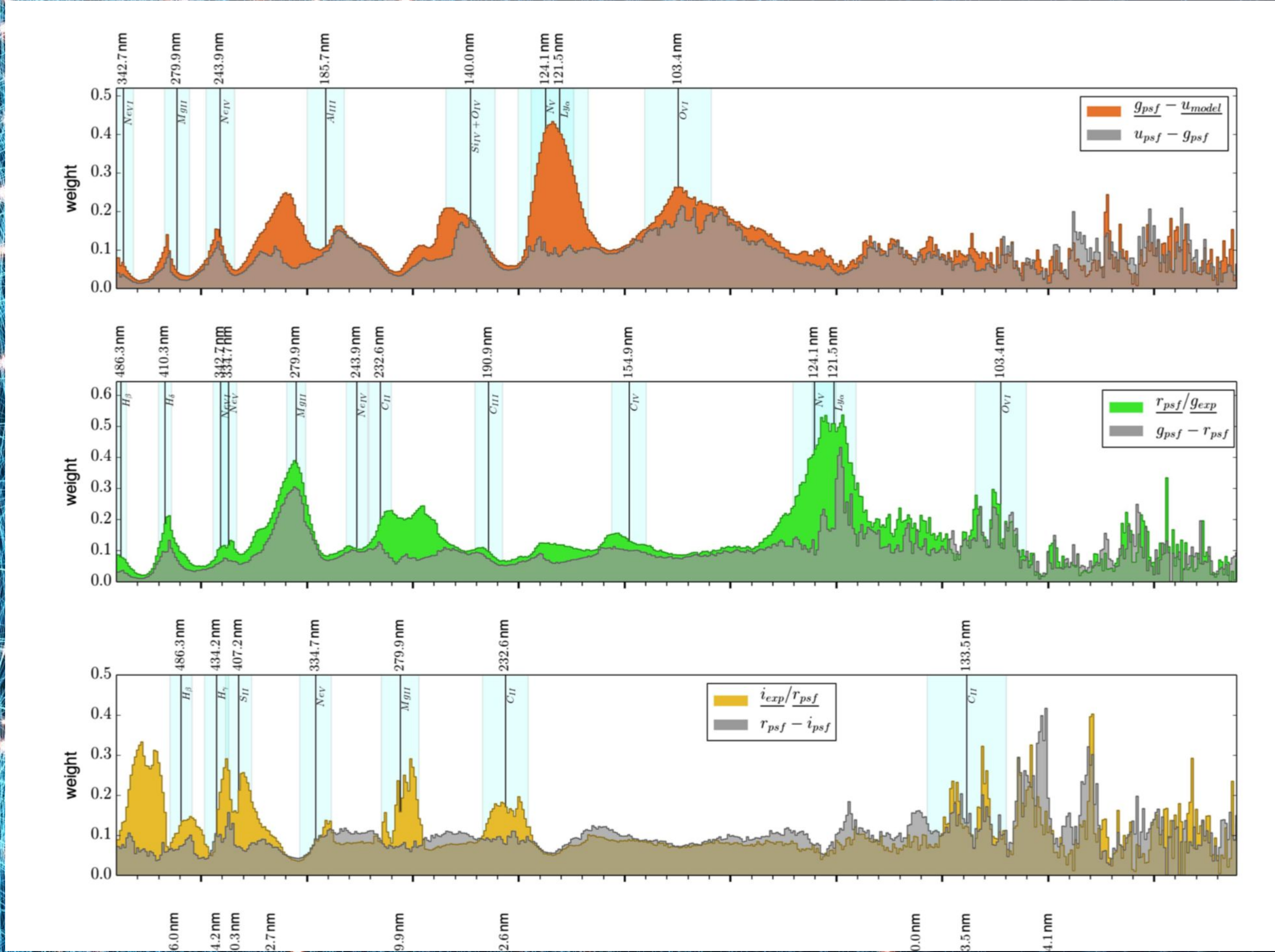
- Same data set... 4250 features
- Method: KNN in GPU Implementation
- Greedy forward selection strategy

– $r_{psf} - r_{petro}$
– $i_{psf} - i_{dev}$
– $\underline{i_{psf}} / \underline{z_{model}}$
– $\underline{r_{psf}} / \underline{i_{exp}}$.

Table 3. Summary of the scores obtained with the RF and DCMDN models in the three experiments.

Exp	Set	# Features	Mean	RMSE	NMAD
DR7a	$\mathrm{Classic}_{10}$	10	−0.024	0.163	0.051
	Best_4	4	−0.023	0.163	0.080
	Best_{10}	10	−0.014	0.124	0.044
	DCMDN	65 536	−0.020	0.145	0.043
DR7b	$\mathrm{Classic}_{10}$	10	−0.030	0.180	0.059
	Best_4	4	−0.027	0.183	0.087
	Best_{10}	10	−0.019	0.145	0.050
	DCMDN	65 536	−0.024	0.171	0.032
DR7+9	$\mathrm{Classic}_{10}$	10	−0.033	0.207	0.073
	Best_4	4	−0.032	0.206	0.100
	Best_{10}	10	−0.023	0.174	0.060
	DCMDN	65 536	−0.027	0.184	0.037

Notes. The DCMDN automatically extracted 65 536 features for each experiment. The resulting scores are also given.



An example of why these features are relevant.

Feature importance of some features in the Best10 set composed by magnitudes from neighbouring bands.

The results are compared to the classic features using PSF magnitudes of the same bands.

Based on the characteristics of the *ugriz* filters, the wavelengths indicating the start, centre, and end of the overlapping regions are used to overplot the positions of particular quasar emission lines using Eq. (2).

In optically selected samples and in presence of large knowledge base, the photo-z problem is saturated by ca. 10 features whose nature strongly depends on the data (no transfer from one data set to the other)

Computationally intensive (extremely), and difficult (if not plain impossible) for large panchromatic heterogeneous surveys

The Features which carry most of the information are not those usually selected by the astronomer but....

... astronomers prefer to understand the selected features (and if possible to associate them to physical properties)...

Feature selection - All relevant

Brescia 2018

PHiLAB (Parameter Handling investigation LABoratory)

Aims at finding tall the features with carry useful information for a given problem

Based on two concepts: «**shadow features**» and **Naïve-LASSO regularization** and exploiting Random Forest model as importance computing engine.

SHADOW FEATURES represent the noisy versions of the real ones and their calculated importance can be used to estimate the relevance of the real features.

A shadow feature for each real one is introduced by randomly shuffling its values among the N samples of the given dataset.

Kursa & Rudnicki 2010, Journal of Statistical Software, 36, 11

LASSO penalizes regression coefficients with an L_1-norm penalty, shrinking many of them to zero. Features with non-zero regression coefficients are "selected".

Regularization in Machine Learning is a process of introducing additional information to solve learning overfitting or to perform Feature Selection in a sparse Parameter Space. The regularization is a penalty term added to any loss function L.

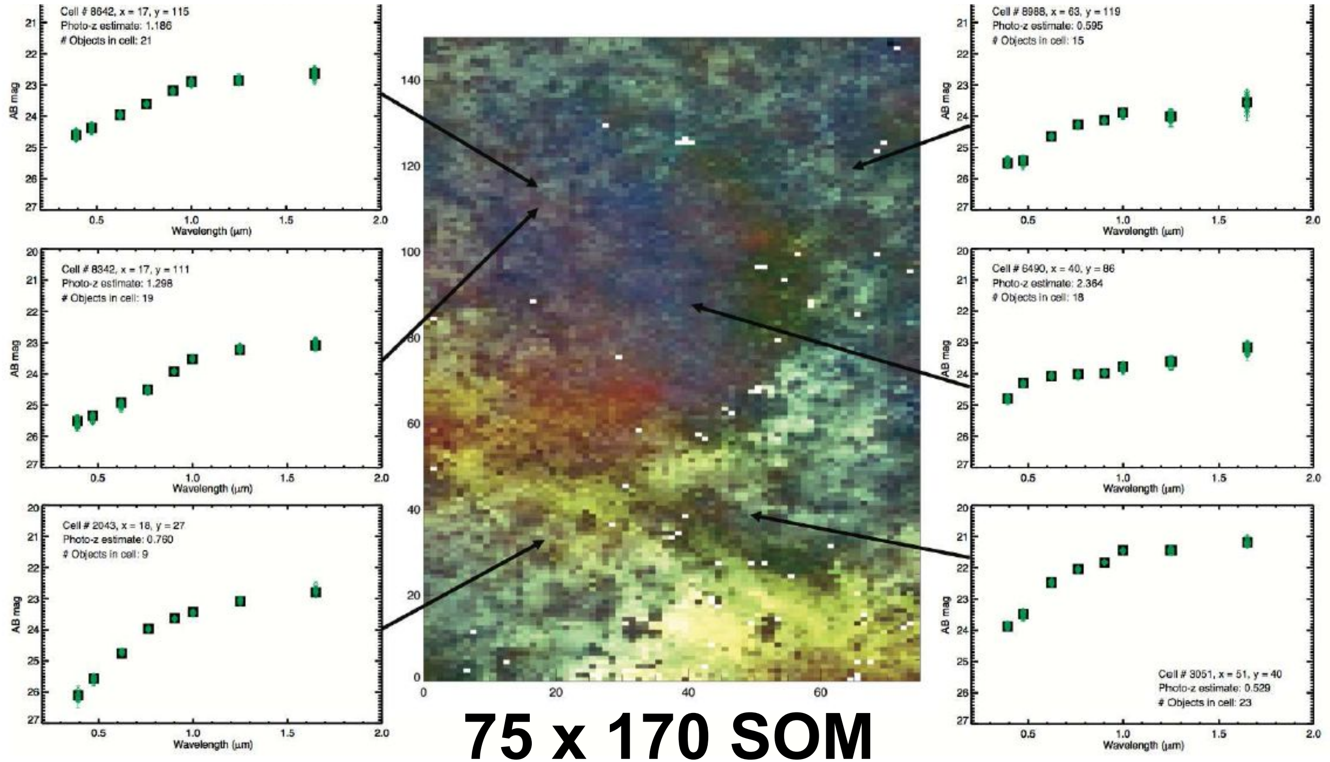
$$min_f \sum_{i=1}^{n} L(f(x)) + \boldsymbol{\lambda} \mathbf{L}_{\mathbf{1}-norm}(\mathbf{w})$$

Hara & Maehara 2016, Proceedings of NIPS 2016, Barcelona, Spain

DATA RICH REGIME

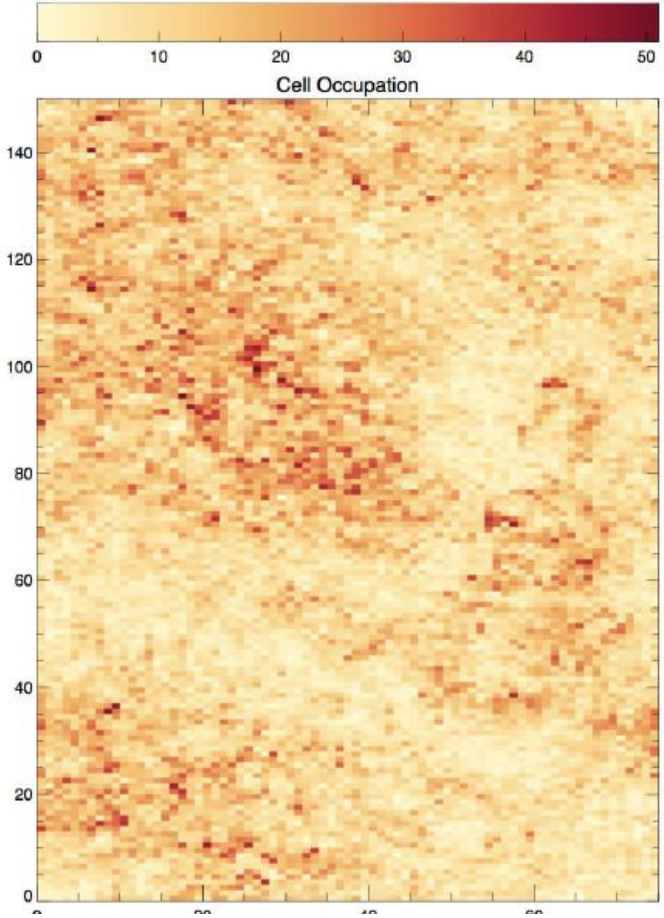
- **Coverage of OPS (Biases in training set)**
  - The OPS is not uniformously covered by the Training set
  - **Do training and test set cover the same OPS?**

Masters et al., 2015, APJ

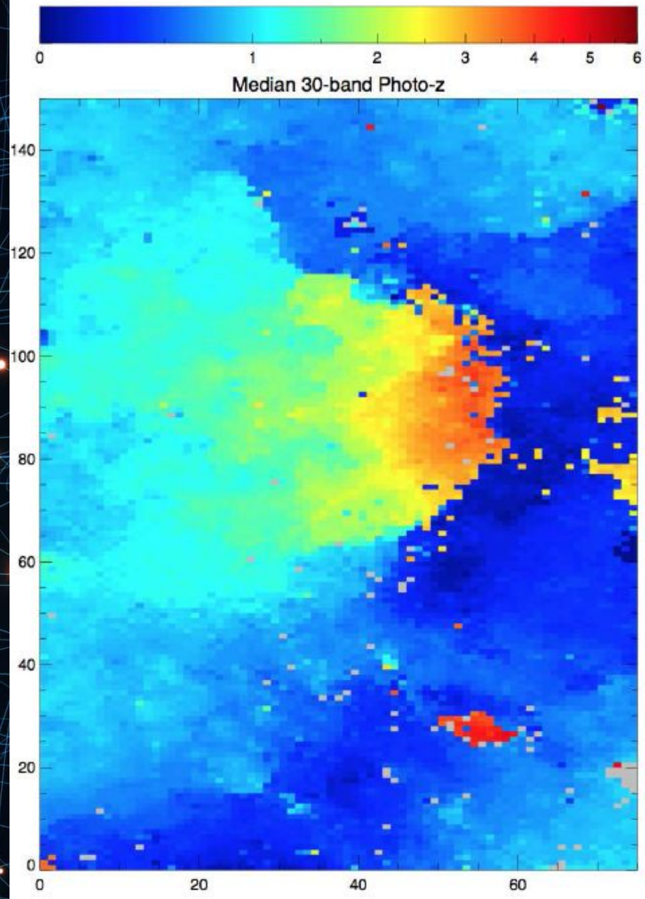


75 x 170 SOM

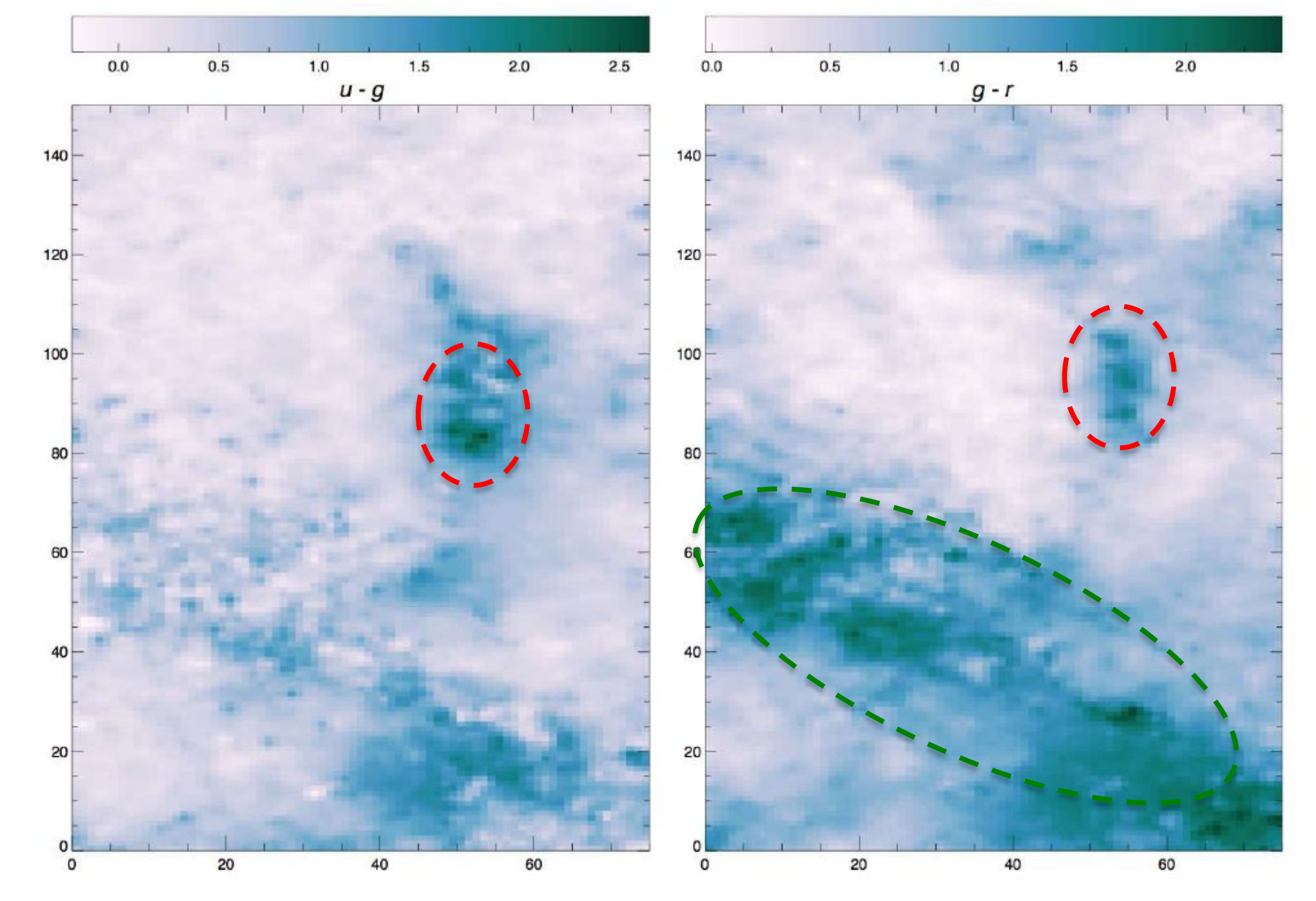
COSMOS data (EUCLIDISED) and converted to "pseudo-Euclid" photometric system: u,g,r,i,z,Y,J,H; Spectroscopic data from COSMOS master catalogue



Density of galaxies in the color space (OPS)



Projection of redshift in the OPS



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

Passive and dusty
galaxies at low redshift

DATA POOR REGIME

Most astronomical literature deals with

- Optically selected samples
- Large spectroscopic knowledge bases
  - More or less uniform coverage of OPS
- Negligible fraction of missing data

Future panchromatic surveys will deal with

- Non optically selected samples (radio, X ray, etc.)
- Reduced spectroscopic knowledge bases
  - Non uniform and incomplete coverage of parameter space (very sparse)
  - Spectroscopic KB extracted from different regions of the sky (e.g. pencil beam surveys, etc.)
- Huge fraction of missing data

A Comparison of Photometric Redshift Techniques for Large Radio Surveys

Norris, Salvato, Longo, Brescia et al., 2019, ArXiv:1902.05188

The survey **EMU - Evolutionary Map of the Universe**, to be performed with ASKAP will observe Ca. 70 million galaxies

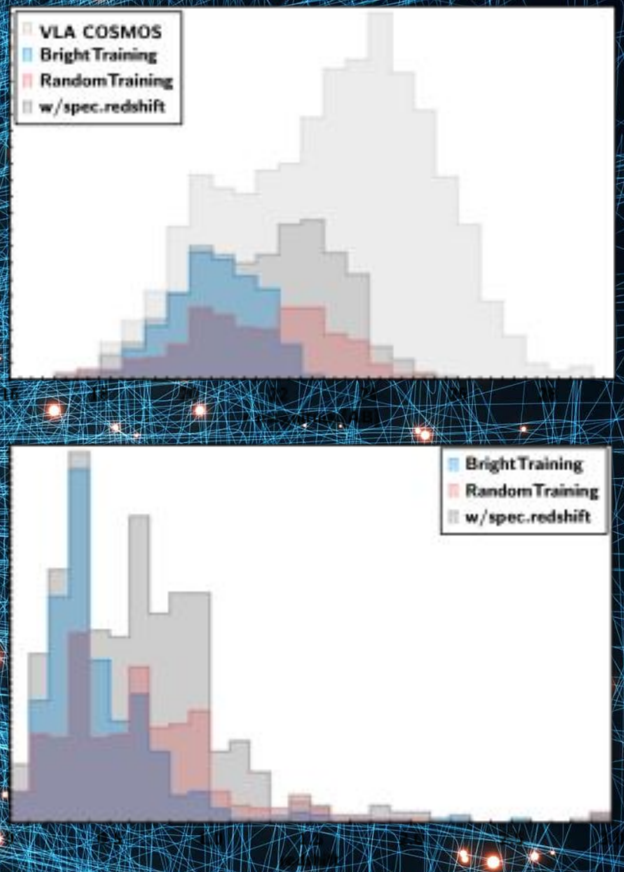
Radio selected samples are dominated (ca 50%) by **starburst** and and **high-z radio loud AGN** (Norris, 2011, 2013). These objects are usually faint and underrepresented in optically selected samples.

The median redshift sample of EMU will be ca z=1.2, while most optically selected samples have median redshift at z=0.5/0.7

Test DATA: VLA-COSMOS 1.4 GHz sample

2242 sources with optical counterparts (Sargent et al. 2010).
757 soTest DATA: VLA-COSMOS 1.4 GHz sample
form the “spectroscopic KB”. (91 (XMM) + 158 (Chandra) X-ray sources).
45 features (photometric measurements)

Small training sets
Poor coverage of OPS
Strongly biased
Incomplete data



16 sets of experiments:
(combinations of...)

1. **Luminosity biases (B or R)**
   Training on shallower sample
   Bright (50%) or **R**andom

2. **Depth (deep or Shallow)**
   Deep: train on deepest data available
   Shallow:: train on data at the same depth of EMU

3. **Radio fluxes** (**Y** or **N**)
   Inclusion of the radio fluxes in the OPS

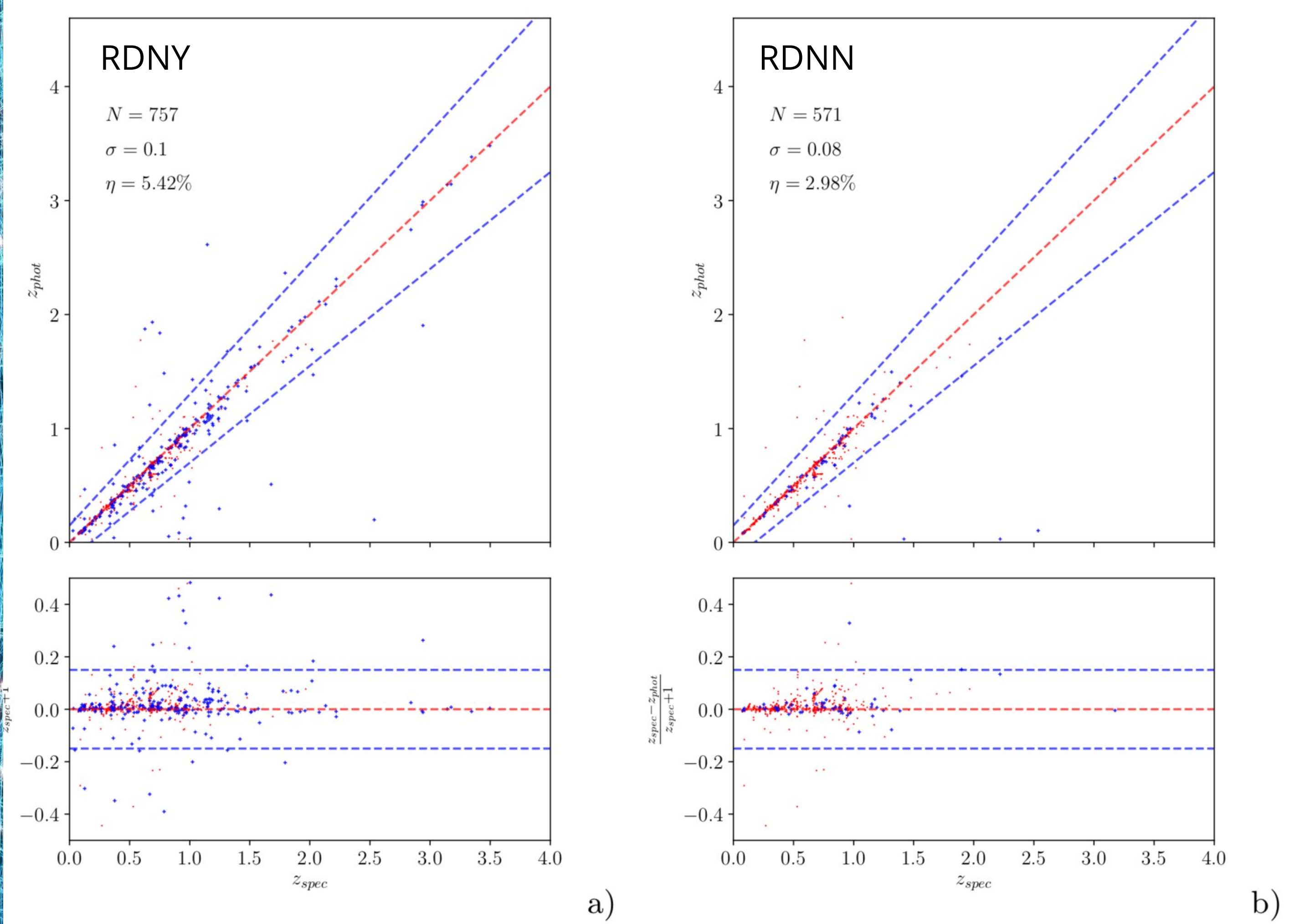
4. **X-ray AGN** (**Y** or **N**)
   Included (not) in the training set

Experiment		A1	B1	C1	D1	E1	F1	G1	H1	A2	B2	C2	D2	E2	F2	G2	H2
Code		BDNY	BDYY	BDNN	BDYN	BSNY	BSYY	BSNN	BSYN	RDNY	RDYY	RDNN	RDYN	RSNY	RSYY	RSNN	RSYN
Training set size		391	391	302	302	391	391	302	302	343	343	278	278	343	343	278	278
Max test set size		366	366	457	457	366	366	457	457	416	416	481	481	416	416	481	481
kNN	N=	366	366	293	293	366	366	293	438	414	414	322	322	414	414	322	322
	NMAD=	0.15	0.15	0.13	0.14	0.1	0.48	0.1	err	0.05	0.05	0.05	0.04	0.23	0.24	0.22	0.22
	η=	56	58	58	59	31	95	28	95	18	18	11	11	49	52	49	52
	β=	44	42	27	26	69	5	46	5	82	82	60	60	51	48	34	32
RF-JHU	N=	366	366	438	438	366	366		438	414	414	467	467	414	414	467	467
	NMAD=	0.11	0.12	0.12	0.12	43	0.45		err	0.07	0.07	0.07	0.07	0.09	0.09	0.1	0.1
	η=	28	27	28	30	95	95		95	15	15	16	16	20	19	21	19
	β=	72	73	69	67	5	5		5	85	85	82	82	80	81	77	79
RF-NA	N=	366	366	293	293	366	366	293	293	414	414	322	322	414	414	322	322
	NMAD=	0.13	0.12	0.16	0.17	0.11	0.09	0.12	0.12	0.07	0.07	0.06	0.06	0.13	0.13	0.11	0.1
	η=	33	25	86	83	28	22	35	33	14	15	8	7	36	36	28	25
	β=	67	75	9	11	72	78	42	43	86	85	62	62	64	64	48	50
MLPQNA	N=	366	366	293	293	366	366	293	293	414	414	322	322	414	414	322	322
	NMAD=	0.2	0.25	0.15	0.14	0.13	0.12	0.08	0.09	0.06	0.06	0.05	0.05	0.12	0.14	0.11	0.12
	η=	80	88	36	31	40	40	22	27	17	19	14	13	36	38	27	32
	β=	20	12	41	44	60	60	50	47	83	81	58	58	64	62	49	46
Le Phare	N=	757		571		509		549		757		571		509		549	
	NMAD=	0.02		0.01		0.08		0.08		0.02		0.01		0.08		0.08	
	η=	5		3		22		23		5		3		22		23	
	β=	95		73		52		56		95		73		52		56	

Table 3. Results of the 16 experiments. Line 2 of the header gives the code as described in §3: Bias (Bright/Random), IR Depth (Deep/Shallow), Radio (Y/N), X-ray (Y/N). Column 1: method name; column 2: metric: N=number of redshifts estimated, σ=standard deviation of estimated-true, η=percentage of outliers, β= overall success rate, expressed as a percentage, as defined in the text.

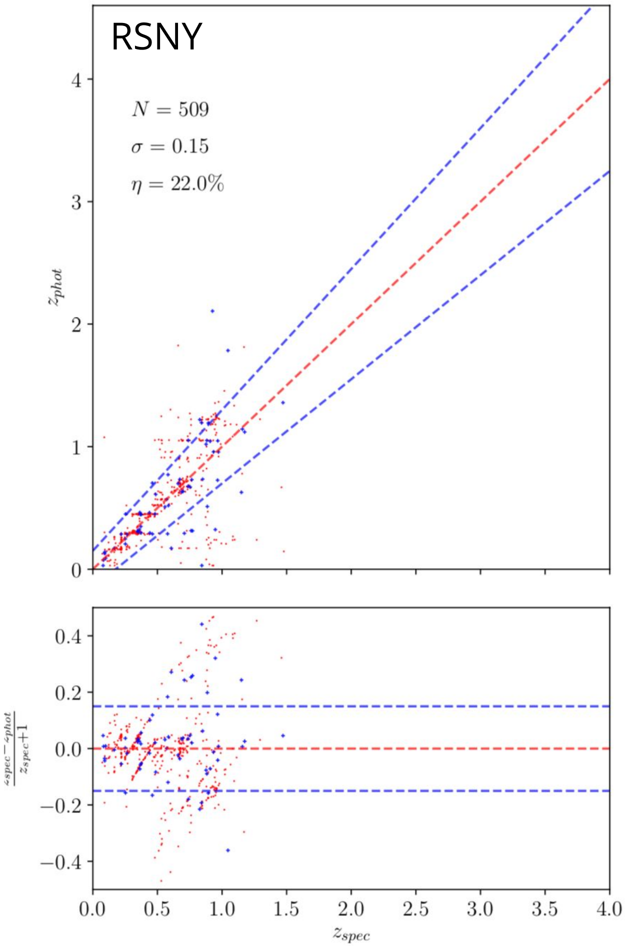
Random Forest (2 implementations), MLPQNA, LE-Phare (SED), BPZ (hybrid), K-NN

Le Phare: SED fitting

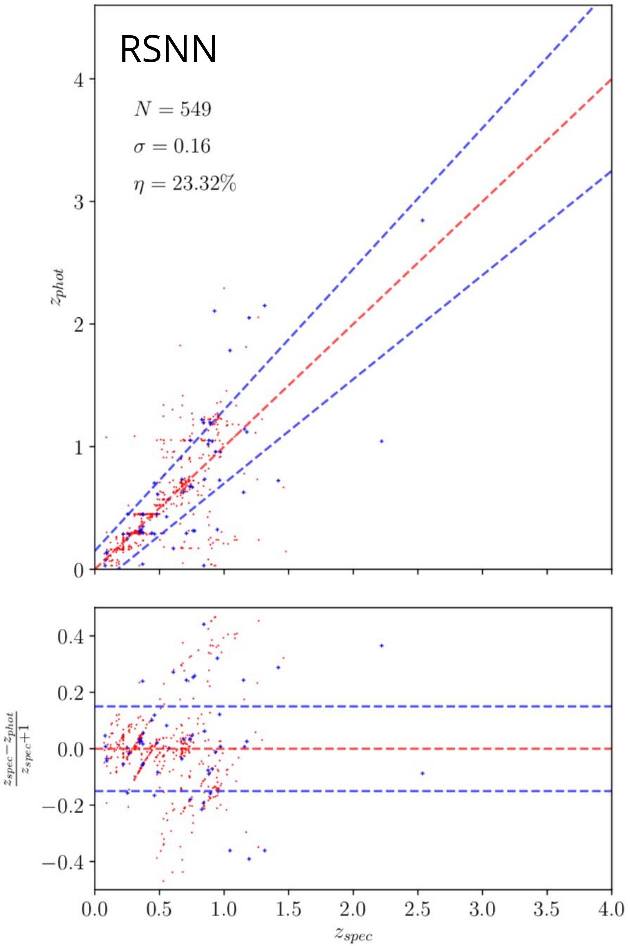


Blu: AGN
Red: non-AGN

Makes use of full COSMOS wavelength coverage



c)

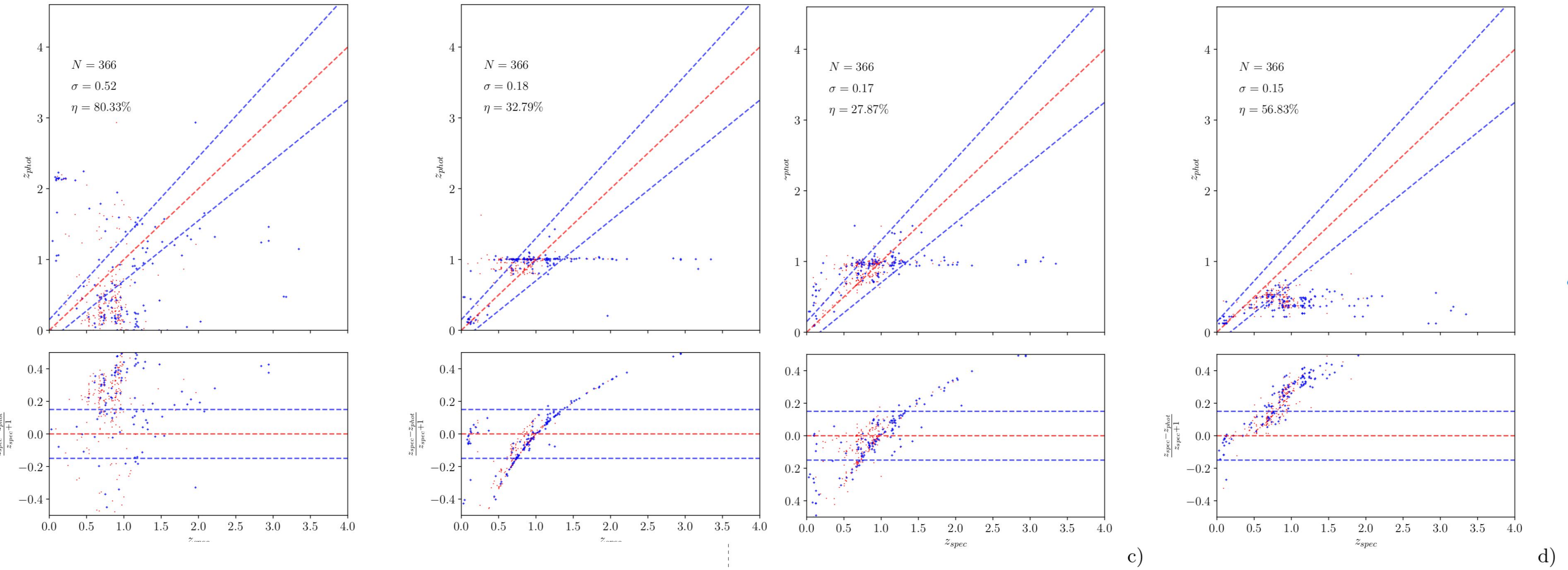


d)

Le Phare

Blu: AGN
Red: non-AGN

Blu: AGN
Red: non-AGN



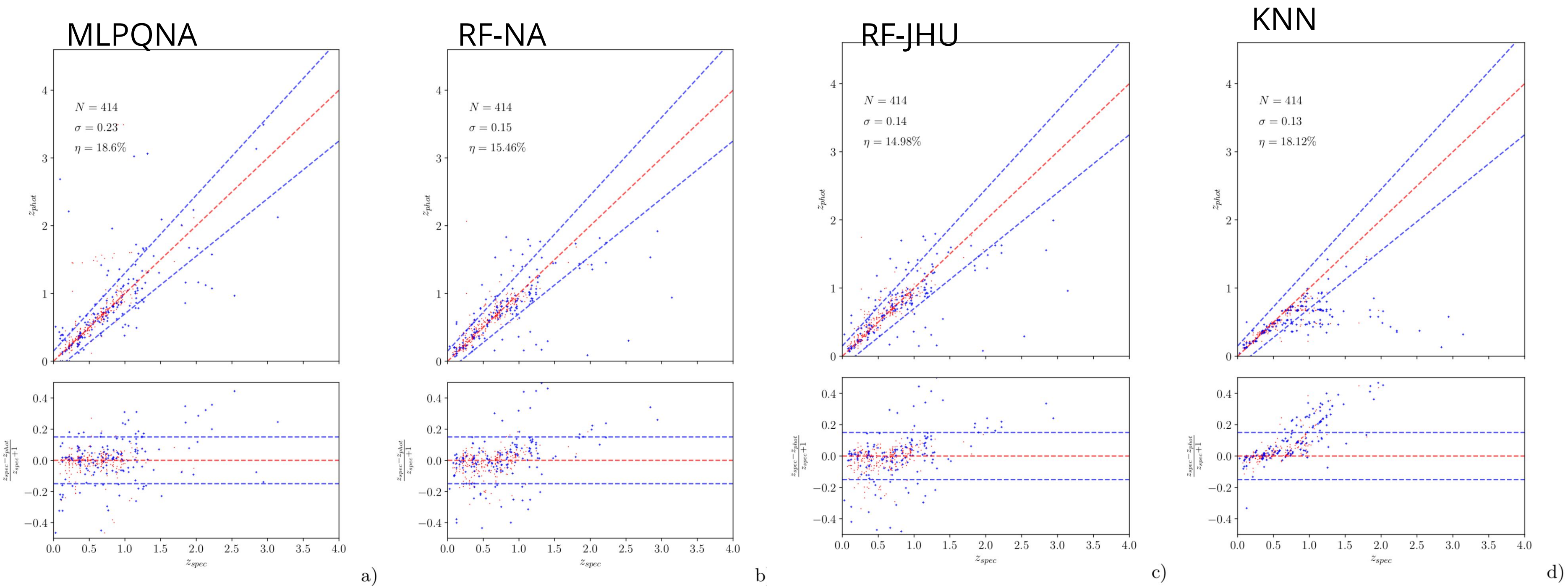
MLPQNA

RF-NA

RF-JHU

KNN

Exp. A1/BDNY:
most realistic for radio surveys (trained on bright 50%)



Exp. B2/RDYY

(random training, deep sample, radio fluxes used, conf. AGN in the training)

Data overabundance vs annotated data scarcity

Common to many (most) domains

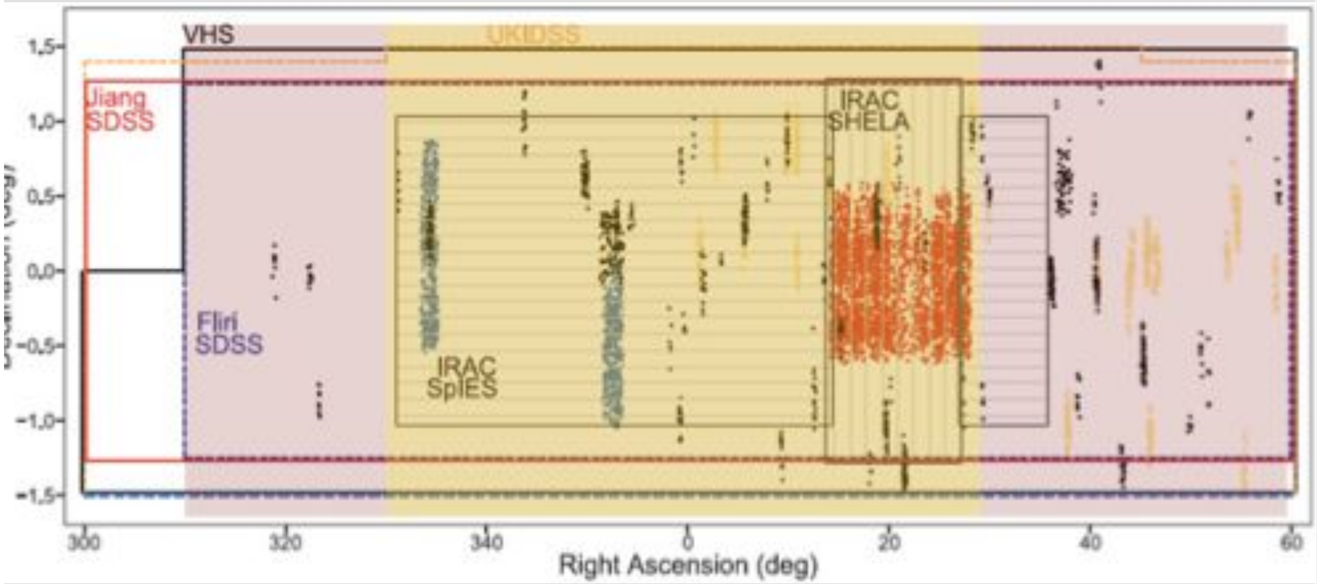
...different strategies to cope with it but no clear cut, unique solution....

Crowdsourcing
Semi-supervised learning
Generative adversarial networks
Active Learning
Domain adaptation/transfer learning
SImulations
Domain knowledge and structure

Photometric Redshifts for X-ray selected Active Galactic Nuclei in the eROSITA era

M. Brescia[1★], M. Salvato[2†], S. Cavuoti[1,3,4‡], T. T. Ananna[5,6], G. Riccio[1], S. M. LaMassa[7], C. M. Urry[6] and G. Longo[3,4]

Sample composed by ca. 7.000 sources in Stripe 82 with X ray counterpart (La Massa et al. 2017)



igure 1. Map of the original multi-wavelength coverage of Stripe 82X area discussed in A17. The total area extends for ~ 2.5° in eclination and 120° in Right Ascension. The dots represent X-ray sources, respectively, from XMM-Newton AO13 (red), AO10 (blue), chival XMM-Newton sources (yellow) and Chandra sources (black). While standard photo-z are generated for the entire area (in red), e selection of the best features discussed in the first part of the paper is obtained considering only the sources in the yellow area.

Filter	BAND DEPTH								
	NOMINAL	BEST	SDSS	SDSS & VHS	SDSS & IRAC	SDSS & WISE	SDSS VHS & IRAC	SDSS VHS & WISE	SDSS VHS IRAC & WISE
FUV	21.99	—	—	—	—	—	—	—	—
NUV	21.99	—	—	—	—	—	—	—	—
u	31.22	28.54	28.54	28.54	28.54	28.54	28.54	28.54	28.54
g	28.77	24.20	24.39	24.20	24.39	24.39	24.20	24.20	24.20
r	27.13	23.25	23.43	23.25	23.43	23.43	23.25	23.25	23.25
i	27.21	22.35	23.49	22.64	23.49	22.45	22.64	22.35	22.35
z	30.46	22.42	23.35	22.46	22.99	22.42	22.46	22.42	22.08
J	24.74	21.64	—	24.64	—	—	21.64	21.64	21.51
H	24.15	22.87	—	22.87	—	—	21.61	22.87	21.61
K	22.60	21.63	—	21.63	—	—	21.63	21.63	21.63
Juk	23.44	—	—	—	—	—	—	—	—
Huk	22.69	—	—	—	—	—	—	—	—
Kuk	22.41	—	—	—	—	—	—	—	—
CH1_SPIES	24.27	$20.82^{\dagger}$	—	—	$21.64^{\dagger}$	—	$21.06^{\dagger}$	—	$20.49^{\dagger}$
CH1_SHELA	22.80		—	—		—		—	
CH2_SPIES	22.88	20.49†	—	—	21.41†	—	$21.07^{\dagger}$	—	$20.22^{\dagger}$
CH2_SHELA	23.88		—	—		—		—	
W1	21.16	20.71	—	—	—	20.71	—	20.71	20.61
W2	20.74	20.59	—	—	—	20.63	—	20.63	20.59
W3	18.20	18.04	—	—	—	18.11	—	18.11	18.04
W4	16.15	16.06	—	—	—	16.13	—	16.13	15.94
N. of sources	5990	2290	4855	3218	2293	3291	1620	2696	1380
N. of sources w/ z_{spec}	2933	1686	2793	2218	1596	2160	1279	1935	1121
N. of sources w/ $F_X > 10^{-14}$	2351	1249	2025	1649	1051	1619	888	1445	793
N. of sources w/ $F_X > 10^{-14}$ and z_{spec}	1550	1025	1483	1309	857	1256	758	1174	683

Table 1. Summary table for depth, amount of sources and redshift coverage. The first column refers to the nominal depth of the entire sample of reliable counterparts in Stripe 82X, as presented in A17. The following columns refer to the magnitudes reached in the various experiments, i.e., the faintest magnitude reported in the Stripe 82X catalogue for the various sub-samples for which the photo-z have been computed. The values in the column ***BEST*** represent the faintest magnitudes of the sub-sample of sources in the yellow area of Fig. 1, used for the features analysis performed with ΦLAB, (Sec. 3.1). The bands marked with a — symbol have been discarded from that specific experiment.

- FUV and NUV magnitudes and corresponding errors from GALEX all-sky survey (Martin et al. 2005);
- u, g, r, i, z SDSS AUTO magnitudes and corresponding errors from Fliri & Trujillo (2016);
- J, H, K from VISTA (Irwin et al. 2004). As shown in A17 additional data in J_{UK},H_{UK},K_{UK} data from UKIDSS (Lawrence et al. 2007) are available for the same area but were not used in this paper;
- 3.6 and 4.5 μm magnitudes and corresponding errors from IRAC. Here two complementary surveys are used: SPIES (Timlin et al. 2016) and SHELA (Papovich et al. 2016). Given the similarity of the two surveys, we do not differentiate sources belonging to one or another;
- W1, W2, W3, W4 magnitudes and corresponding errors from AllWISE (Wright et al. 2010).

FS with PhiLab

feature	importance	feature	importance
R-Z	14.51%	J-K	0.40%
G-I	12.44%	U-CH1	0.35%
CH1-CH2	7.50%	H-CH1	0.34%
U-G	6.00%	R-CH2	0.33%
Z-W1	5.84%	U-I	0.33%
Z-CH1	4.24%	R-W2	0.33%
G-R	4.03%	K-CH1	0.33%
K	3.14%	R-W1	0.31%
G-Z	3.03%	U	0.30%
I-W1	2.00%	U-J	0.30%
I-CH2	1.94%	G-W2	0.27%
H	1.81%	G-CH2	0.24%
R-I	1.67%	I-J	0.23%
I-CH1	1.51%	CH1-W2	0.23%
J	1.45%	G	0.22%
H-K	1.34%	J-CH2	0.21%
R	1.21%	G-CH1	0.21%
I	1.21%	G-K	0.20%
W1	1.18%	J-W1	0.20%
I-Z	1.08%	H-W2	0.20%
Z	0.99%	K-CH2	0.17%
H-W1	0.97%	K-W2	0.16%
K-W1	0.83%	U-W1	0.16%
Z-W2	0.83%	Z-J	0.16%
CH2-W1	0.77%	U-K	0.15%
Z-CH2	0.68%	R-CH1	0.14%
U-R	0.68%	H-CH2	0.13%
U-Z	0.62%	CH1	0.13%
G-W1	0.56%	Z-H	0.13%
J-CH1	0.54%	U-H	0.12%
Z-K	0.52%	J-W2	0.09%
I-W2	0.50%	I-K	0.08%
J-H	0.46%	R-K	0.07%
W1-W2	0.45%	—	—

Table 2. Results of the feature analysis (percentages of estimated feature importance) performed with ΦLAB in the case of the parameter space composed by considering all magnitudes and colours available.

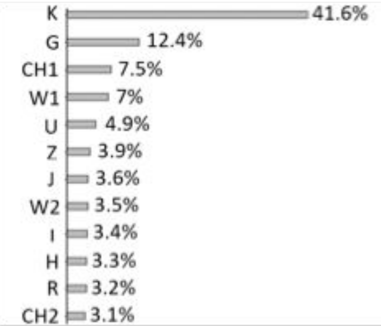


Figure 3. Results of the feature analysis performed with ΦLA
The importance of each feature is estimated for the case in whi
only magnitudes are considered for the sample *BEST magopt*.

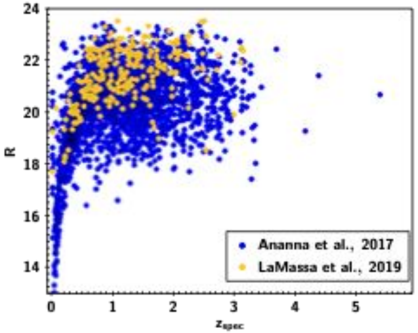


Figure 2. Redshift and magnitude distribution for the sources with spectroscopic redshift. The blue sources were presented in A17 and have been used in this work as training and blind test samples. The 258 yellow sources are on average fainter and were recently presented in LaMassa et al. (2019). They are used as additional blind test sample.

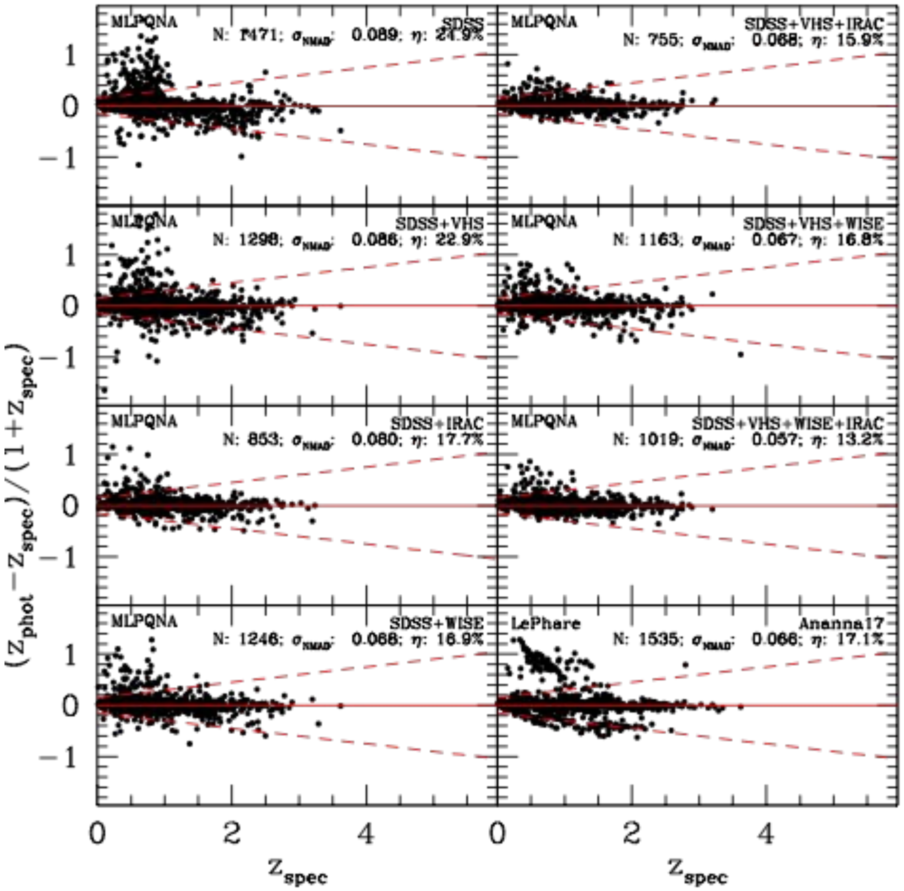
Spectroscopic KB

Due to different depths …. need to handle missing data

(i) SDSS, VHS, WISE & IRAC (*sdssVWI*)
(ii) SDSS, VHS & WISE (*sdssVW*);
(iii) SDSS, VHS & IRAC (*sdssVI*);
(iv) SDSS & WISE (*sdssW*);
(v) SDSS & IRAC (*sdssI*);
(vi) SDSS & VHS (*sdssV*);
(vii) SDSS.

	Number of sources	$\|bias\|$	σ	σ_{68}	σ_{NMAD}	η
A17	258	0.0066	0.292	0.129	0.089	27.07
sdss	227	0.0037	0.367	0.158	0.129	33.48
sdssV	135	0.0357	0.322	0.211	0.149	41.48
sdssW	144	0.0073	0.288	0.173	0.137	34.03
sdssI	110	0.0119	0.202	0.184	0.163	40.91
sdssVW	111	0.0459	0.272	0.167	0.143	33.33
sdssVI	58	0.0343	0.255	0.161	0.116	32.76
sdssVWI	25	0.0298	0.151	0.152	0.104	32.00
MLPQNA$_{\mathrm{merged}}$	229	0.0182	0.270	0.192	0.154	38.43

f all statistical results for the new sample of 258 spectroscopic redshifts presented in LaMassa et al. 2019.



igure 5. Comparison between spectroscopic redshift and photo-z for the sources cut at the eROSITA flux and divided on the basis o vailable photometric points. For comparison, the result from A17 is reported in the lower right panel of the figure. By comparing the ccuracy and the fraction of outliers in every panel with the corresponding row in Table 8, we see that computing photo-z using only DSS for bright X-ray sources is not recommended.

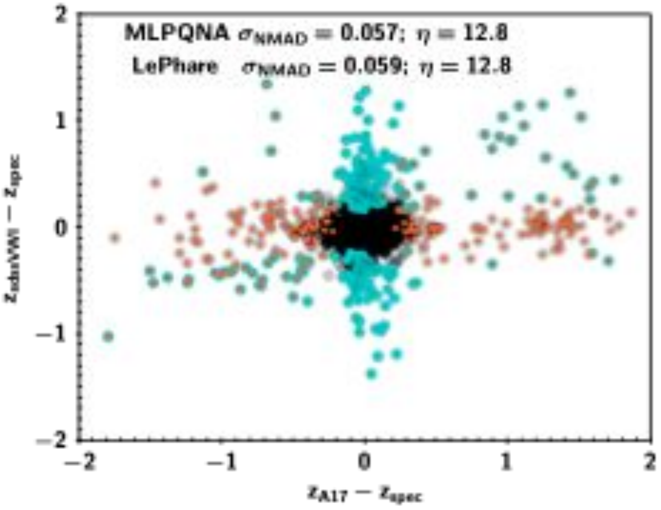


Figure 6. Difference between spectroscopic redshift and photo-z computed via MLPQNA and LePhare for the sub-sample of 1679 sources with SDSS, VHS, WISE and IRAC photometry. Sources that are outliers for MLPQNA (LePhare) are plot in cyan (orange). For this sub-sample the accuracy and fraction of outliers are very similar for the two methods. However, the majority of the outliers are such only for one of the two algorithms.

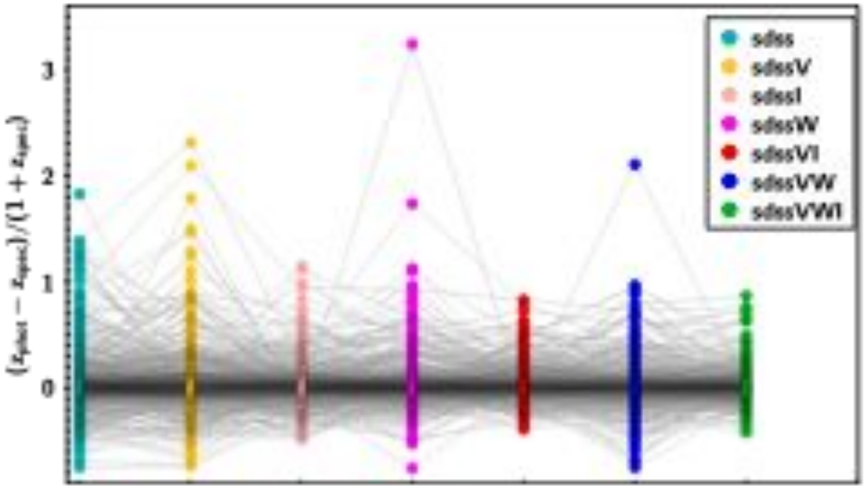
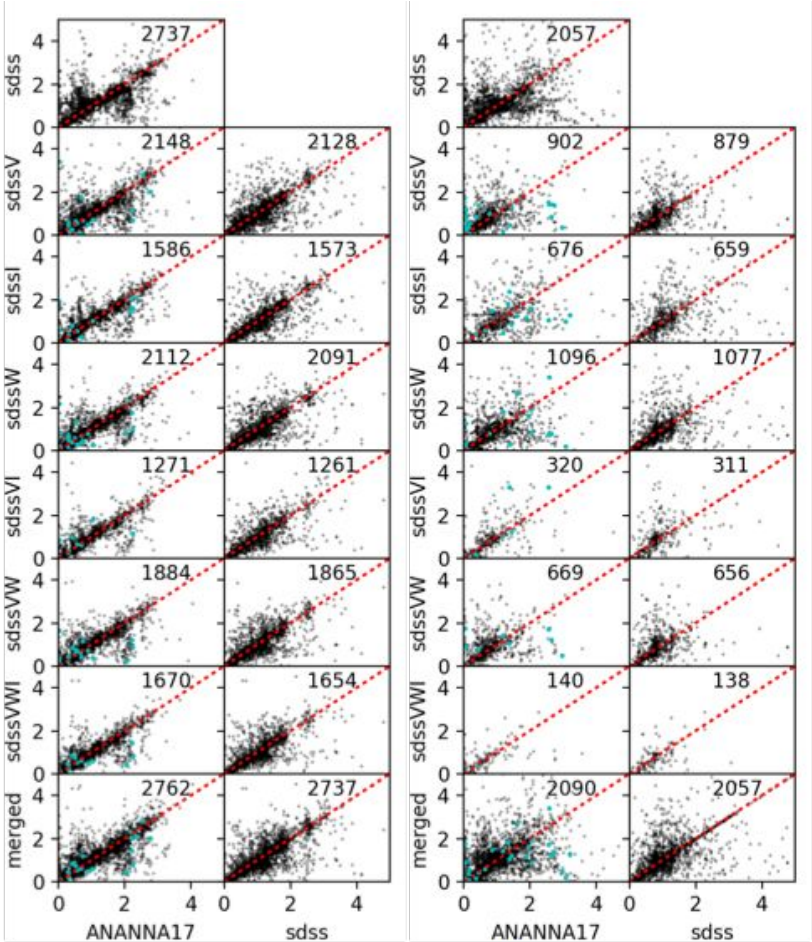


Figure 7. One-to-one comparison of accuracy for photo-z computed via MLPQNA with different combinations of photometry. For this plot only sources present in all the subsamples have ben used.



Comparison between photo-z computed via SED fitting (A17) and MLPQNA for the sample for which spectroscopic information is, respectively, available (left panel) and not available (right panel).

The cyan points indicate the sources for which the redshift could be computed only after considering supplementary photometry in addition to SDSS.

Some conclusions on upervised methods

- **If large annotated, reliable data sets are available, all methods are substantially equivalent (DL, RF, MLPQNA, K-NN, etc.)**
  - Need for extensive feature selection (different approaches substantially equivalent)
  - Differences are in the range of a few % which are usually negligible when errors are properly taken into account

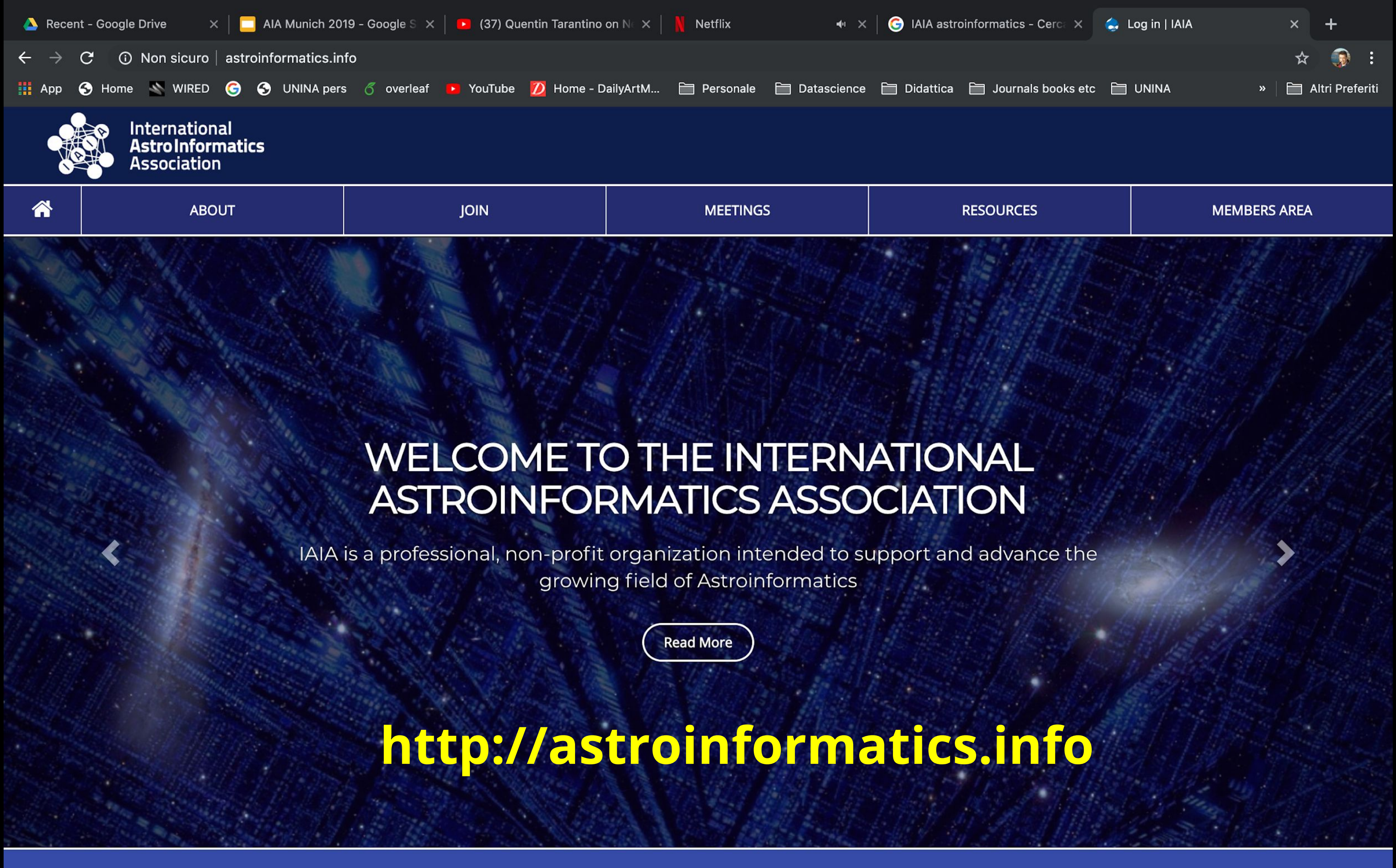
- **If data are heterogeneous (depth, coverage, etc.) or biased… methods matter**
  - DL substantially useless, RF or KNN outperformed by normal MLP's (better at generalising ?)
  - Handling biases and understanding results becomes the crucial part.
  - Lots of work remains to be done to be able to apply these methods to future surveys

- **The scientific exploitation of future large survey projects requires better "annotated data"**

... Globally, the shortfall for data scientists is projected to be between five million and 10 million. For SA to have "healthy participation" in SKA, the country will need 200 data scientists when the project is live ...

Peter Quinn, 2019

Thanks for the attention



http://astroinformatics.info