Data-rich astronomy: mining synoptic sky surveys

Stefano Cavuoti
Department of Physics – University Federico II – Napoli

Supervisors:

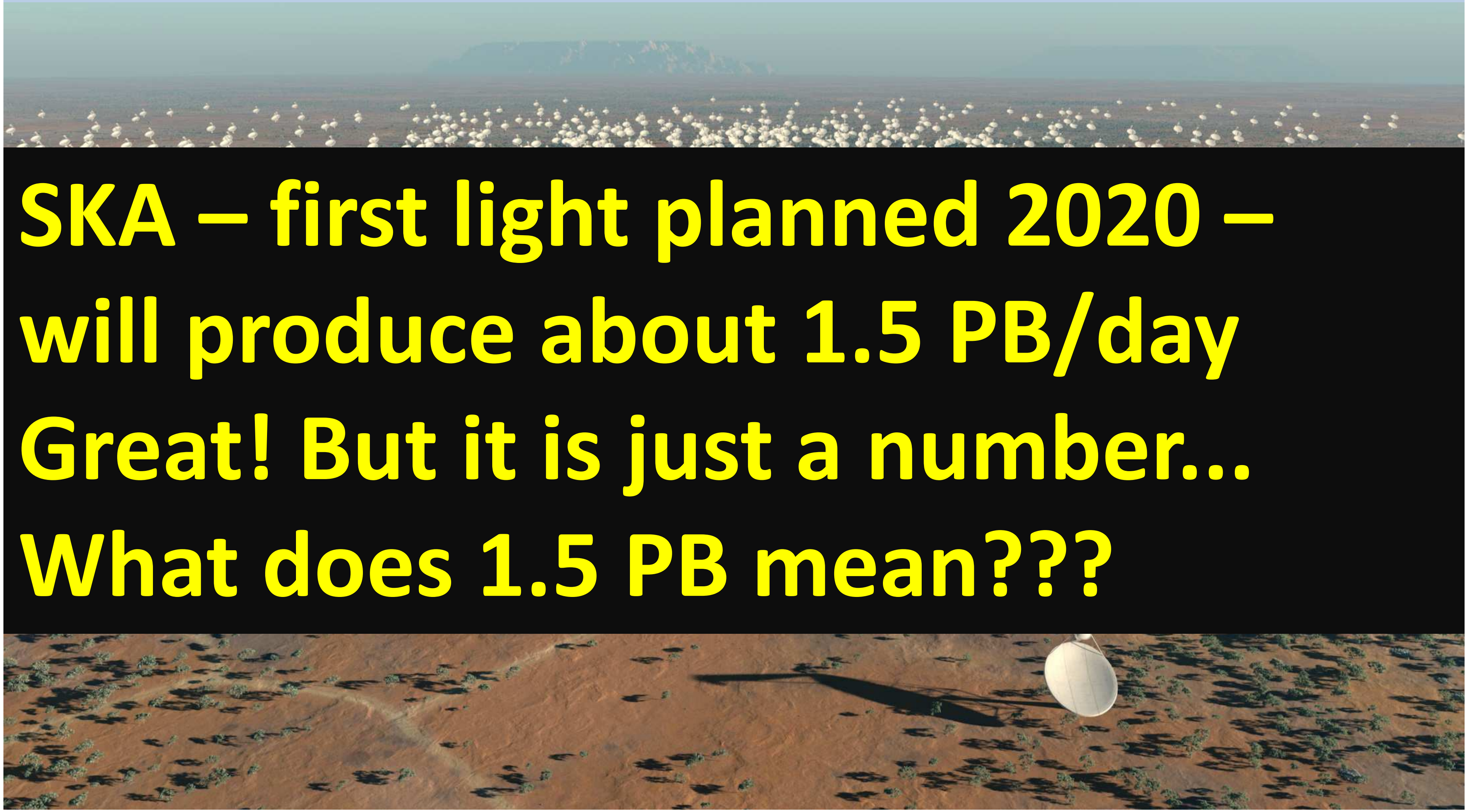
Giuseppe Longo
Department of Physics – University Federico II – Napoli

Massimo Brescia
INAF – Capodimonte Astronomical Observatory – Napoli

PhD Defence – 28 May 2013

The first part of the title of my thesis: Data-Rich Astronomy

We all know that astronomy has became a data rich science, but do we grasp the depth of the problem?



SKA – first light planned 2020 – will produce about 1.5 PB/day

Great! But it is just a number...

What does 1.5 PB mean???

SKA WILL ALSO FILL ABOUT 1.000.000.000 AMAZON KINDLE PER DAY

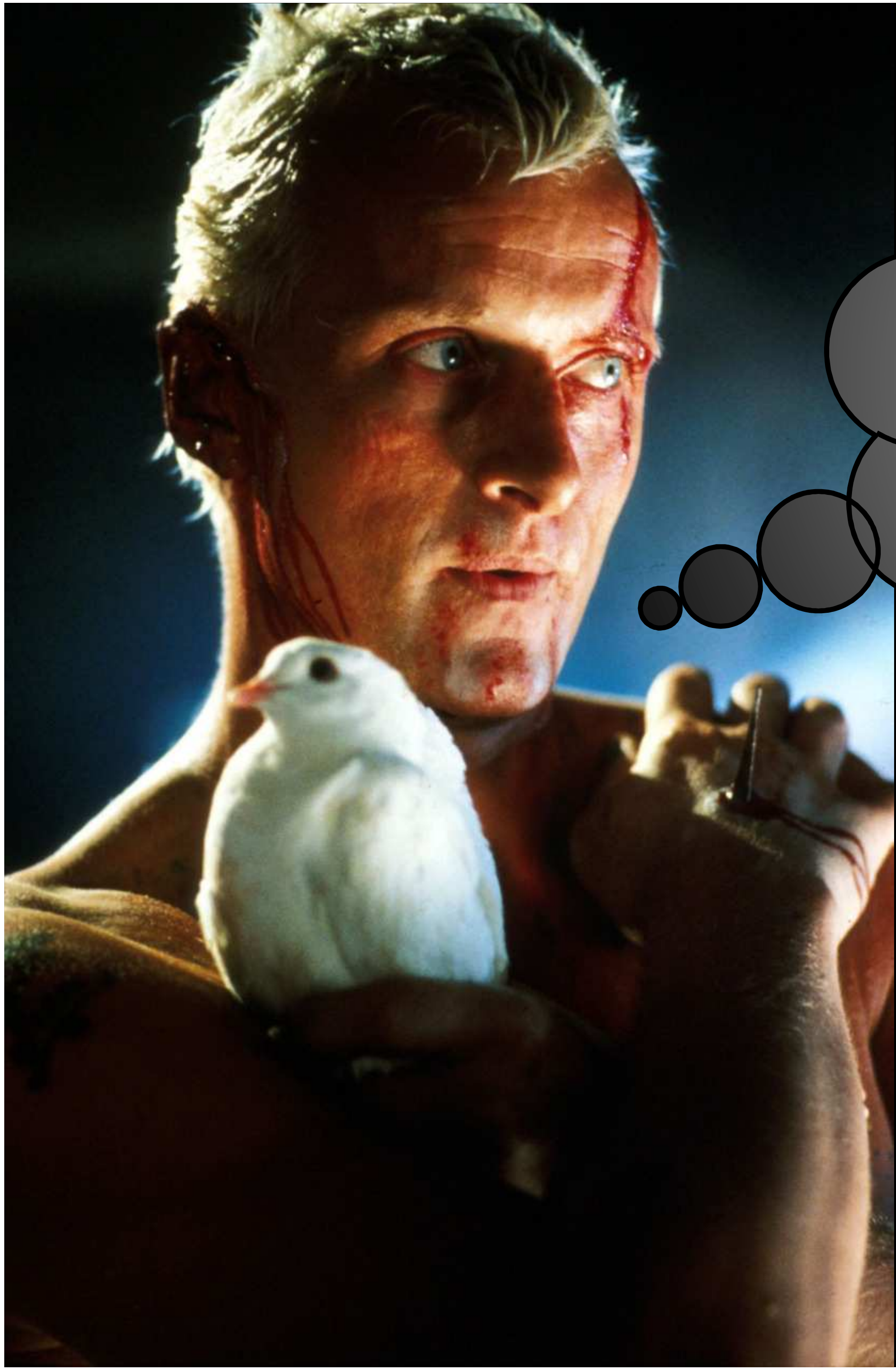
The largest library in the world is the **Library of Congress, Washington, D.C., USA with ONLY 30.000.000 books…**

US Census Bureau (December 2010) estimates for 2020 is 7.7 billion of person…

So to SEE each day the amount of SKA data, each person in the world should read about 100.000 books per day…

ARE YOU READY FOR THIS???

AND THIS IS JUST ONE SURVEY!!!



I've seen things you people wouldn't believe. Attack ships on fire off the shoulder of Orion. I've watched c-beams glitter in the dark near the Tannhäuser Gate. **All those ... *moments* will be lost in time, like tears...in rain.**

Time to die…

ROY EFFECT:

(Blade Runner)

MOST DATA WILL **NEVER** BE SEEN BY HUMANS!!!

ASTRO-INFORMATICS

										A	S	T	R	O	N	O	M	Y		
							P	H	Y	**S**	I	C	S							
								D	A	**T**	A	M	I	N	I	N	G			
										R	E	S	E	A	R	C	H			
							A	L	G	**O**	R	I	T	H	M	S				
		S	C	A	L	A	B	I	L	**I**	T	Y								
I	N	F	O	R	M	A	T	I	O	**N**	T	E	C	H	N	O	L	O	G	Y
A	D	V	A	N	C	E	D	S	O	**F**	T	W	A	R	E					
					O	N	T	O	L	**O**	G	Y								
									G	**R**	I	D								
										M	O	D	E	L	I	N	G			
P	A	R	A	L	L	E	L	I	Z	**A**	T	I	O	N						
										T	E	C	H	N	O	L	O	G	Y	
						M	A	C	H	**I**	N	E	L	E	A	R	N	I	N	G
										C	L	O	U	D						
										S	T	A	T	I	S	T	I	C	S	

SEMANTIC TUNING:

X-informatics are the application of information technology to discipline X, with emphasis on persistent data stores. Astro-Informatics, Bio-Informatics, Chem-Informatics, Meteo-Informatics and so on.

BEYOND THE SEMANTIC:

These fields share the same traits: they all aim at acquiring new viewpoints and models by applying informatics-based approaches to existing fields such as biology. They also share the same methodology: the generation of huge amount of data with the help of advanced sensor and observation technologies, and the fast search and knowledge discovery from large-scale databases.

Astroinformatics: a new era for Astronomy?

You take the **Blue Pill**,
The story ends. You wake up in your bed and believe whatever you want to believe.
You take the **Red Pill**,
You stay in Wonderland and I show You how deep the rabbit hole goes

I'm only offering You the **TRUTH**... Nothing more.

My Thesis Work

I tried to use the Astroinformatics paradigm and tools to tackle several problems…

…well sometimes I also needed to create that tools…

Algorithmic Aspects:

- **GAME**
- **MLPQNA**
- **SVM**

Technological Aspects

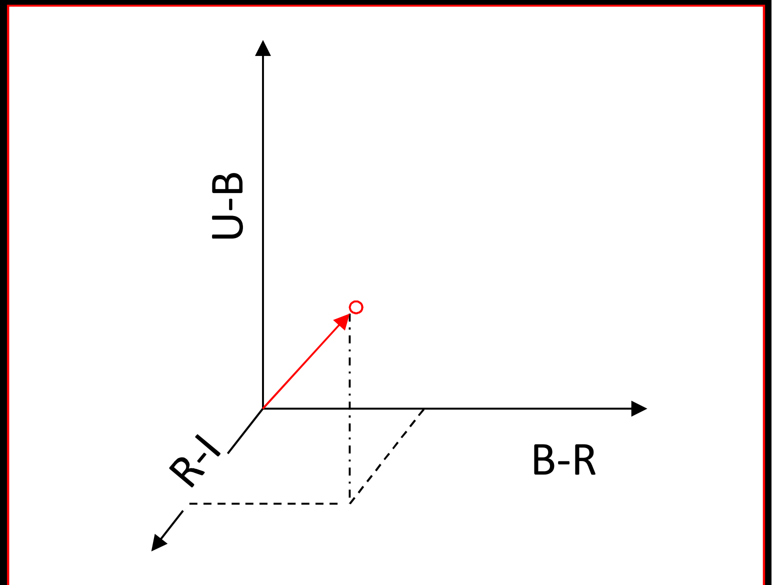
- **DAMEWARE**
- **STraDiWa**

Scientific Aspects:

- **AGN classification**
- **comparison of catalogue extracting methods**
- **EUCLID Mission**
- **globular cluster classification**
- **photometric redshifts**
- **transients detection and modelization**

This talk is focused on the yellow items

PHOTOMETRIC REDSHIFTS AS AN INVERSE PROBLEM



Why do we need (photometric or spectroscopic) redshifts?

- Obviously to measure the distance of objects
- To disentangle the degeneracies in the object classification
- Cosmological parameters
- Lensing Effects
- Dark Energy
- Dark Matter

OK! But why are Photometric Redshifts crucial?

SDSS DR9 Facts	
Sky coverage	14,555 square degrees
Catalog objects	932,891,133
Galaxy spectra	1,457,002
Quasar spectra	228,468
Star spectra	668,054

932,891,133 PHOTOMETRIC OBJECTS
2,353,524 SPETTROSCOPIC OBJECTS
~ 400 times more objects!!!

A short History: (see e.g. Yee 1998 for a review)

- **Baum (1962)**
  Colors of early type galaxies measured from 9 bands with a photometer were turned into a low resolution SED to determine distances of galaxy clusters relative to other clusters of galaxies.

- **Koo (1985)**
  Colors (from photographic plate material) were compared to colors expected for synthetic Bruzual-Charlot SEDs. Redshifts were estimated from iso-z lines in colorcolor diagrams.

- **Loh & Spillar (1986)**
  used $\chi 2$-minimization for redshift estimates

- **Pello and others**
  developed a method of "permitted" redshifts; the intersection of the permitted redshift intervalls for all galaxy colors measured defines "the" redshift of a galaxy.

- **Photometric redshifts have become very popular since the middle of the 1990s**
  - well calibrated, deep multi-waveband data (HDF, other deep fields, SDSS)
  - representative spectroscopic data sets available to test method (Keck, VLT, SDSS...)
  - better cost efficiency if only approximate redshift is needed

Photometric Redshifts: Methods

Template based:

color-space tessellation, $\chi 2$-minimization, maximum likelihood, Bayesian ...

uses physical information: SED's (sizes, compactness...),
... and therefore biased
extrapolates reasonably ok into unknown territory

Learning based:

Nearest Neighbour, Kd-tree, Direct fitting, Neural Networks, Support Vector Machines, Kernel Regression, Regression Trees & Random Forests...

ignores physical information: and therefore unbiased,
can uncover unknown dependencies
requires large training set, bad in extrapolation

Photometric redshifts: the Data Mining approach

Photometric redshifts are treated as a regression problem (i.e. function approximation) , hence a DM problem:

$$\mathbf{X} \equiv \{x_1, x_2, x_3, \ldots x_N\} \quad \textbf{input vectors}$$

$$\mathbf{Y} \equiv \{x_1, x_2, x_3, \ldots x_M\} \quad \textbf{target vectors} \quad M << N$$

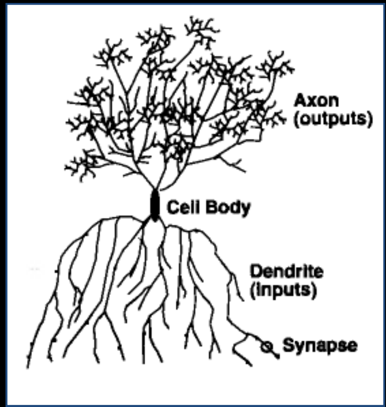
$$\textbf{find} \quad \hat{f}: \quad \hat{\mathbf{Y}} = \hat{f}(\mathbf{X}) \quad \textbf{is a good approximation of } \mathbf{Y}$$

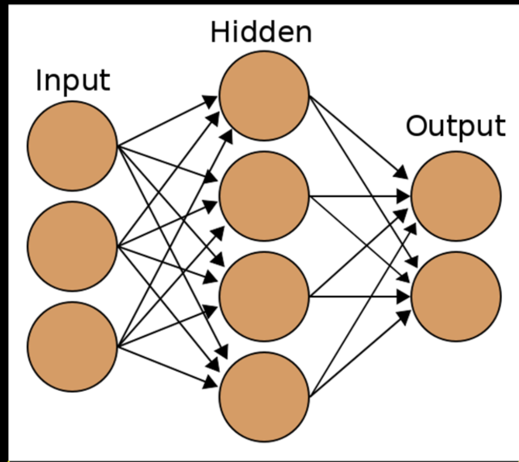
KB = Knowledge Base

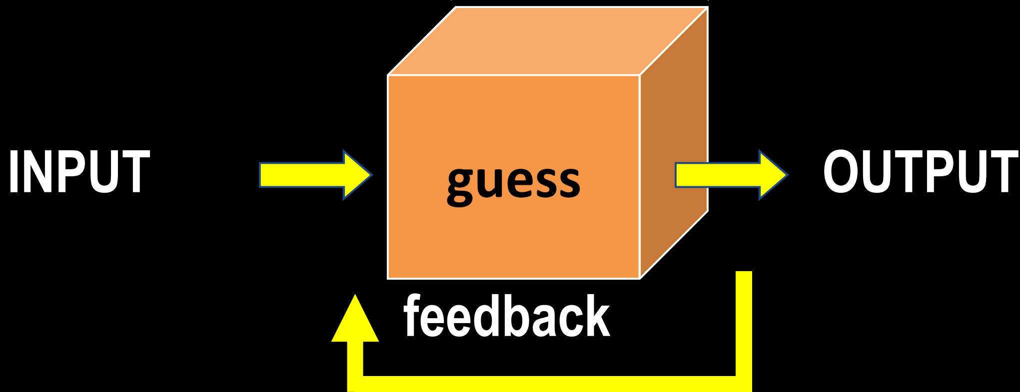
KB(from VO)
(set of templates) → Mapping function → Knowledge (phot-z's)

Our Photometric Redshift Method - MLP

A Multi Layer Perceptron is a mathematical operator that mimics the brain behavior:





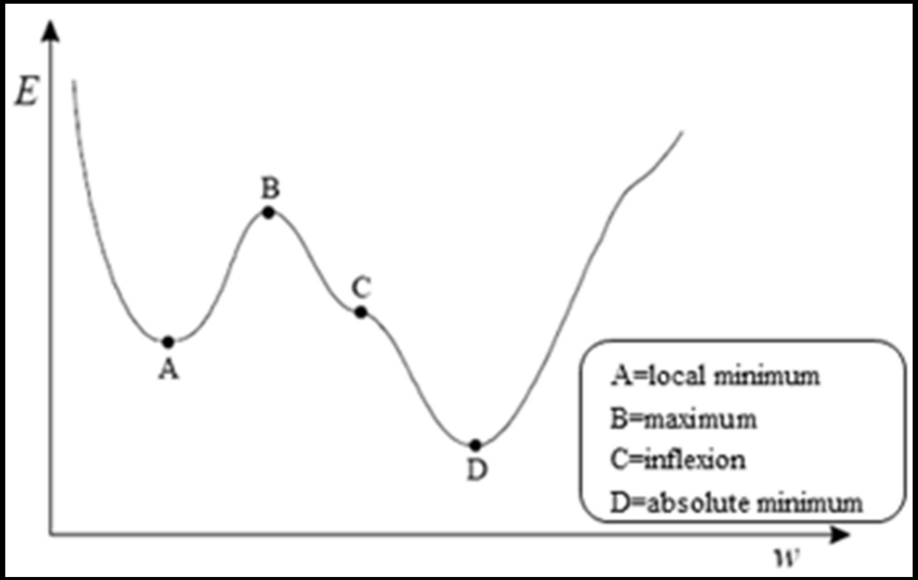


Neurons are connected by «activation functions» we have different kind of MLP changing the way with they found the best solution

Our Photometric Redshift Method - MLPQNA

MLP may be trained in several ways, we implement and tested some of them (Back Propagation, Genetic Algorithm and Quasi Newton Algorithm).

QNA are based on Newton's method to find the stationary point of a function, where the gradient is 0. Newton's method assumes that the function can be locally approximated as a quadratic in the region around the optimum, and use the first and second derivatives (gradient and Hessian) to find the stationary point.



We used MLPQNA with great results both in regression and classification cases, the redshift estimation that follows are the regression use cases.

Our Photometric Redshift Environment - DAME Program

DAME Program is a joint effort between University Federico II, Caltech and INAF-OACN, aimed at implementing (as web 2.0 apps and services) a scientific gateway for data exploration on top of a virtualized distributed computing environment.

Multi-purpose data mining with machine learning Web App REsource

Extensions
- DAME-KNIME
- ML Model plugin

Specialized web apps for:
- text mining (VOGCLUSTERS)
- Transient classification (STraDiWA)
- EUCLID Mission Data Quality

Web Services:
- SDSS mirror
- WFXT Time Calculator
- GAME (GPU+CUDA ML model)

http://dame.dsf.unina.it/
Science and management Documents
Science cases
Newsletters

http://www.youtube.com/user/DAMEmedia
DAMEWARE Web Application media channel

My specific contributions to the development of DAME

- I developed the first prototype of DAME,
- I was and still am the person in charge for the DMM (Data Mining Models) package design and implementation.
- I supervised, checked and tested each released plugin, finally,
- I was involved in the ideation, implementation and test of several models, such as GAME (also in the CUDA version), MLPQNA and SVM
- I was involved in the implementation and test of the new plugin procedure
- And finally in the definition of the future development of the suite.

Photometric Redshifts

- Implementation of a new method (MLPQNA)
- Test of the method in the PHAT1 contest
- Understanding that we could violate the Haykin theorem

Hence…

- Galaxies:
  - SDSS
- Quasars (Feature Selections, and Outliers Understanding in progress)
  - SDSS
  - SDSS + GALEX
  - SDSS + UKIDSS
  - SDSS + GALEX + UKIDSS
  - SDSS + GALEX + UKIDSS + WISE

Why Quasars?
Quasar Candidates Cosmology

PHoto-z Accuracy Testing – PHAT1 CONTEST

The PHAT consists of a **competition** engaged by involving all relevant players (Hildebrandt et al 2010) with the "*aim to evaluate different (theoretical/empirical) methods to extract photo-z from an ensemble of ground-based and space observation catalogues in several bands, composed to perform photometric redshift prediction evaluation tests of several models, both theoretical and empirical, based on the training/statistics of given spectroscopic redshifts*". The imaging dataset is obtained in the **GOODS-North** (Great Observatories Origins Deep Survey Northern field.
The total features of object (**1984**) **patterns** are indeed based on **18 bands**.

In this contest, in fact, **only 515 objects** were made available with the corresponding spectroscopic redshift, while for the remaining 1469 objects the related spectroscopic redshift has been hidden from all participants.

A&A 523, A31 (2010)
DOI: 10.1051/0004-6361/201014885

Astronomy & Astrophysics

PHAT: PHoto-z Accuracy Testing★

H. Hildebrandt[1], S. Arnouts[2], P. Capak[3], L. A. Moustakas[4], C. Wolf[5], F. B. Abdalla[6], R. J. Assef[7], M. Banerji[8], N. Benítez[9], G. B. Brammer[10], T. Budavári[11], S. Carliles[12], D. Coe[4], T. Dahlen[13], R. Feldmann[14], D. Gerdes[15], B. Gillis[16], O. Ilbert[17], R. Kotulla[18,19], O. Lahav[6], I. H. Li[20], J.-M. Miralles[21], N. Purger[22], S. Schmidt[23], and J. Singal[24]

Statistical Indicators

$$\Delta z = (zspec - zphot)$$

$$\text{bias} = \frac{\sum_{i=1}^{N} \Delta z_i}{N}$$

$$\text{MAD} = Median(|\,\Delta z - Median(\Delta z)\,|)$$

$$\text{standard deviation } \sigma = \sqrt{\frac{\sum_{i=1}^{N}\left[\Delta z_i - \left(\frac{\sum_{i=1}^{N}\Delta z_i}{N}\right)\right]^2}{N}}$$

$$\Delta z' = (zspec - zphot)/(1 + zspec)$$

$$\text{bias}_{norm} = \frac{\sum_{i=1}^{N} \Delta z'_i}{N}$$

$$\text{MAD}_{norm} = Median(|\,\Delta z' - Median(\Delta z')\,|)$$

$$\sigma_{norm} = \sqrt{\frac{\sum_{i=1}^{N}\left[\Delta z'_i - \left(\frac{\sum_{i=1}^{N}\Delta z'_i}{N}\right)\right]^2}{N}}$$

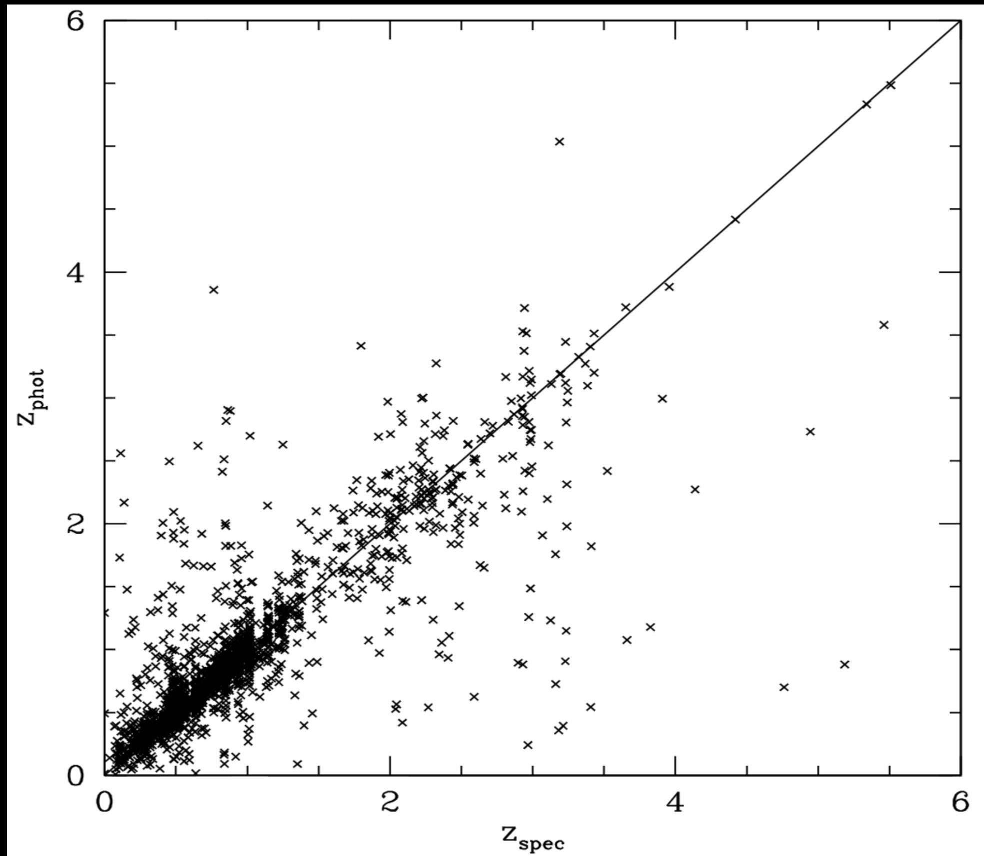
Astronomy & Astrophysics manuscript no. aa19755-12
September 5, 2012

Photometric redshifts with the quasi Newton algorithm (MLPQNA). Results in the PHAT1 contest

S. Cavuoti[1,2], M. Brescia[2,1], G. Longo[1,2,3], and A. Mercurio[2]

Filter	Instrument	$m_{\mathrm{lim.;AB}}$
U	MOSAIC@KPNO-4 m	27.1[a]
B	SUPRIMECAM@Subaru	26.9[a]
V	SUPRIMECAM@Subaru	26.8[a]
R	SUPRIMECAM@Subaru	26.6[a]
I	SUPRIMECAM@Subaru	25.6[a]
Z	SUPRIMECAM@Subaru	25.4[a]
F435W	ACS@HST	27.8[b]
F606W	ACS@HST	27.8[b]
F775W	ACS@HST	27.1[b]
F850LP	ACS@HST	26.6[b]
J	ULBCAM@UH-2.2 m	24.1[c]
H	ULBCAM@UH-2.2 m	23.1[c]
HK	QUIRC@UH-2.2 m	22.1[c]
K	WIRC@Hale-5 m	22.5[d]
3.6 μm	IRAC@Spitzer	25.8[e]
4.5 μm	IRAC@Spitzer	25.8[e]
5.8 μm	IRAC@Spitzer	23.0[e]
8.0 μm	IRAC@Spitzer	23.0[e]

18 bands (near UV → mid IR)



Best among all empirical methods
bias ~ 0,0006
σ_{norm}= 0.05
$|\Delta z|>1\sigma$ = 16.33%

PHAT1 CONTEST - RESULTS

A	18-band; $\|\Delta z\| \leq 0.15$			14-band; $\|\Delta z\| \leq 0.15$			18-band; $R < 24$; $\|\Delta z\| \leq 0.15$			14-band; $R < 24$; $\|\Delta z\| \leq 0.15$		
Code	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %
QNA	0.0006	0.056	16.3	0.0028	0.063	19.3	0.0002	0.053	11.7	0.0016	0.060	13.7
AN-e	–0.010	0.074	31.0	–0.006	0.078	38.5	–0.013	0.071	24.4	–0.007	0.076	32.8
EC-e	–0.001	0.067	18.4	0.002	0.066	16.7	–0.006	0.064	14.5	–0.003	0.064	13.5
PO-e	–0.009	0.052	18.0	–0.007	0.051	13.7	–0.009	0.047	10.7	–0.008	0.046	7.1
RT-e	–0.009	0.066	21.4	–0.008	0.067	24.2	–0.012	0.063	16.4	–0.012	0.064	18.4
B	18-band; $\|\Delta z\| \leq 0.5$			14-band; $\|\Delta z\| \leq 0.5$			18-band; $R < 24$; $\|\Delta z\| \leq 0.5$			14-band; $R < 24$; $\|\Delta z\| \leq 0.5$		
Code	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %
QNA	–0.0028	0.114	3.8	–0.0046	0.125	3.8	–0.0039	0.101	1.7	–0.0039	0.101	1.7
AN-e	–0.036	0.151	3.1	–0.035	0.173	4.2	–0.047	0.130	1.4	–0.047	0.130	1.4
EC-e	–0.007	0.120	3.6	–0.003	0.114	3.6	–0.015	0.106	1.9	–0.015	0.106	1.9
PO-e	–0.013	0.124	3.1	0.001	0.107	2.3	–0.020	0.098	1.2	–0.020	0.098	1.2
RT-e	–0.031	0.126	3.2	–0.028	0.137	3.6	–0.034	0.111	1.4	–0.034	0.111	1.4
C	18-band; $z_{sp} \leq 1.5$, $\|\Delta z\| \leq 0.15$			14-band; $z_{sp} \leq 1.5$, $\|\Delta z\| \leq 0.15$			18-band; $z_{sp} > 1.5$, $\|\Delta z\| \leq 0.15$			14-band; $z_{sp} > 1.5$, $\|\Delta z\| \leq 0.15$		
Code	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %
QNA	–0.0004	0.053	14.6	0.0001	0.061	16.6	0.0074	0.072	26.3	0.0222	0.070	35.0
AN-e	–0.017	0.070	27.6	–0.010	0.076	33.6	0.051	0.078	50.7	0.045	0.077	66.4
EC-e	–0.003	0.065	16.1	–0.000	0.064	14.5	0.015	0.077	32.3	0.015	0.077	29.5
PO-e	–0.012	0.049	12.6	–0.011	0.047	9.4	0.019	0.075	48.3	0.026	0.074	37.7
RT-e	–0.016	0.062	19.6	–0.014	0.064	21.1	0.040	0.072	31.8	0.039	0.071	41.9

WARNING: Still limited by the Haykin Theorem!!!

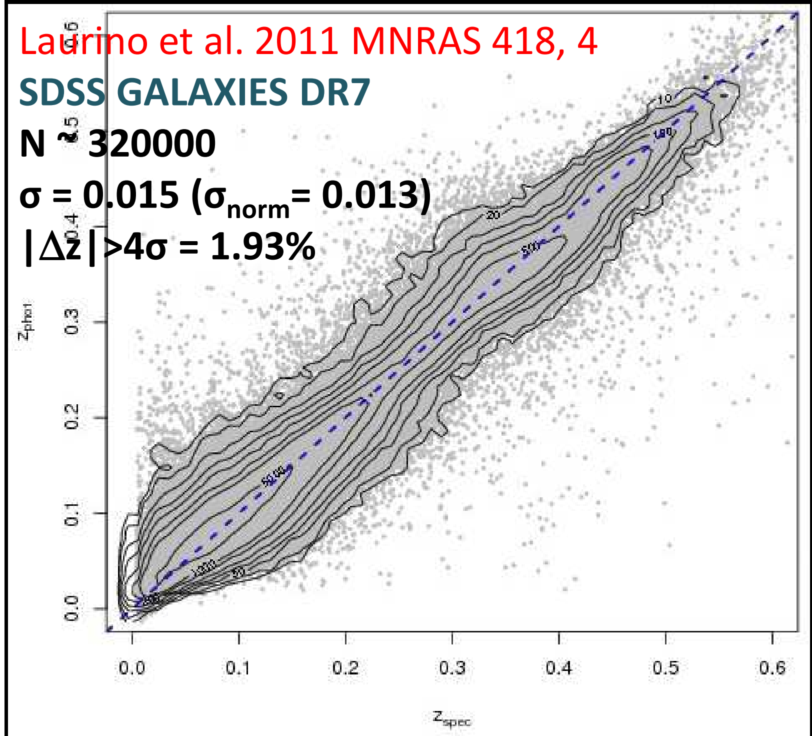
Haykin Theorem

The so called Haykin Theorem stated that:
“A second hidden layer is almost useless”

Bengio & LeCun (2007) have proved that complex problems, in which the mapping function is highly non linear and the local density of data in the parameter space is very variable, are better matched by deep networks with more than one hidden computational layer.

With our experiments, we proved that (both with galaxies and Quasars) the photo-z mapping function, with such dataset, is so complex that requires the second hidden layer despite the Haykin theorem!!!

SDSS DR9 Galaxies



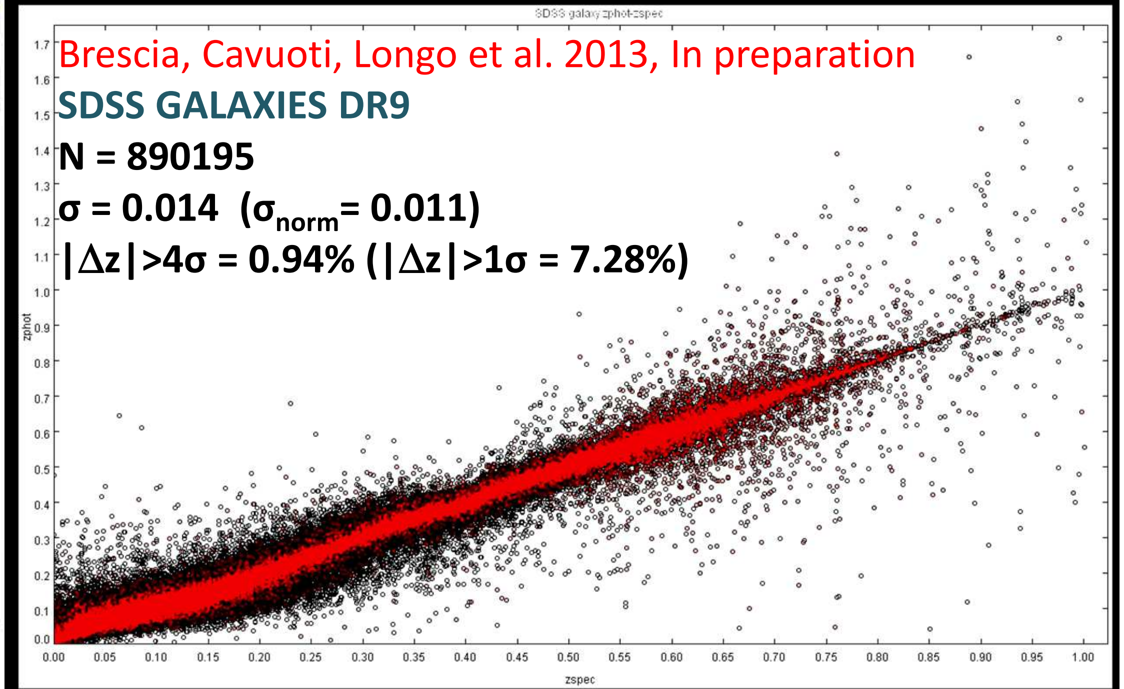
Laurino et al. 2011 MNRAS 418, 4

SDSS GALAXIES DR7

N ~ 320000

σ = 0.015 (σ_{norm}= 0.013)

$|\Delta z|>4\sigma$ = 1.93%



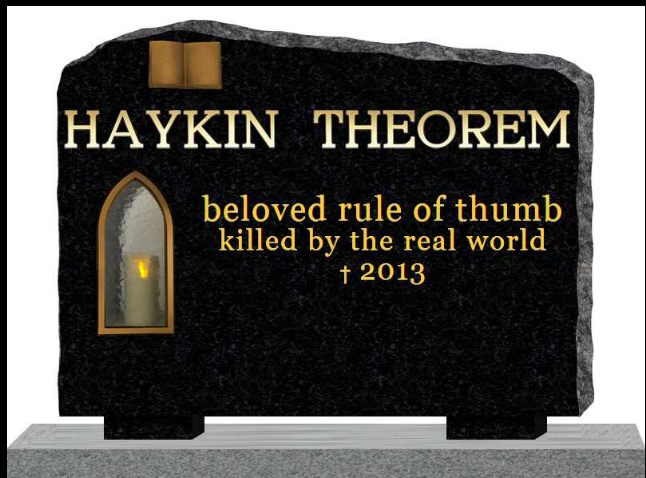
Brescia, Cavuoti, Longo et al. 2013, In preparation

SDSS GALAXIES DR9

N = 890195

σ = 0.014 (σ_{norm}= 0.011)

$|\Delta z|>4\sigma$ = 0.94% ($|\Delta z|>1\sigma$ = 7.28%)



QSO Redshifts

For the Quasars SDSS bands are not enough…

Thanks to the federation of database and using the VO tools we retrieve the data from four surveys: SDSS, GALEX, UKIDSS and WISE obtaining:

• SDSS,	~100k objects	z limit ~ 5
• SDSS+GALEX	~45k objects	z limit ~ 3.5
• SDSS+UKIDSS	~30k objects	z limit ~ 5
• SDSS+UKIDSS+GALEX	~15k objects	z limit ~ 2.8
• SDSS+UKIDSS+GALEX+WISE	~14k objects	z limit ~ 2.8

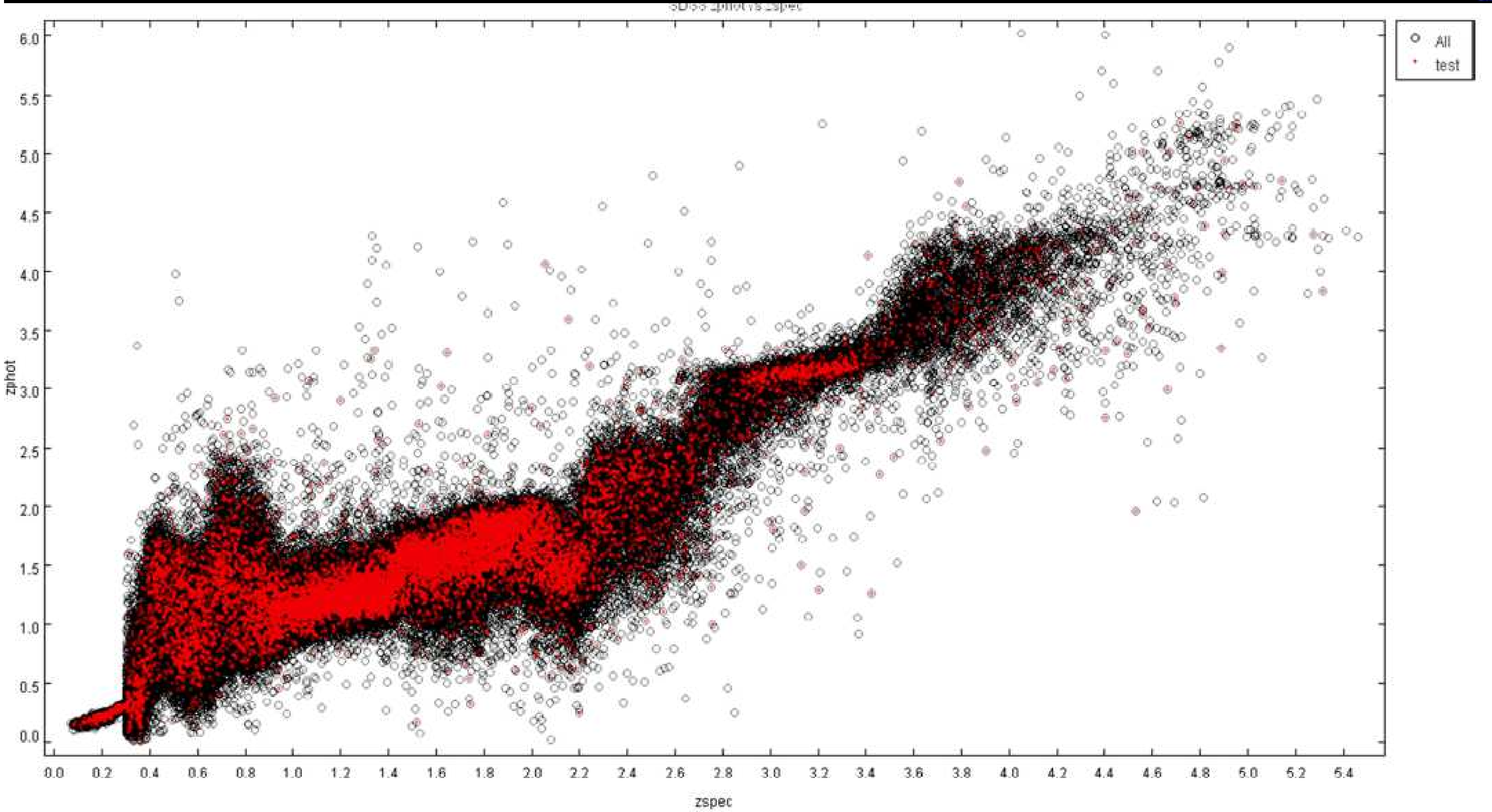
Once I had the data three new questions arise…

- Which magnitudes are the best for this work?
- It's better to use magnitudes straightforward or colors? Or a combination of both (colors + reference mag)?
- Adding bands reduces number of templates. Which factor is dominant?

And after many (ca. 100) experiments we choose:

- Color + reference mag
- 2 hidden layers
- SDSS psf mag
- GALEX mag iso
- UKIDSS hall mag
- WISE mag iso

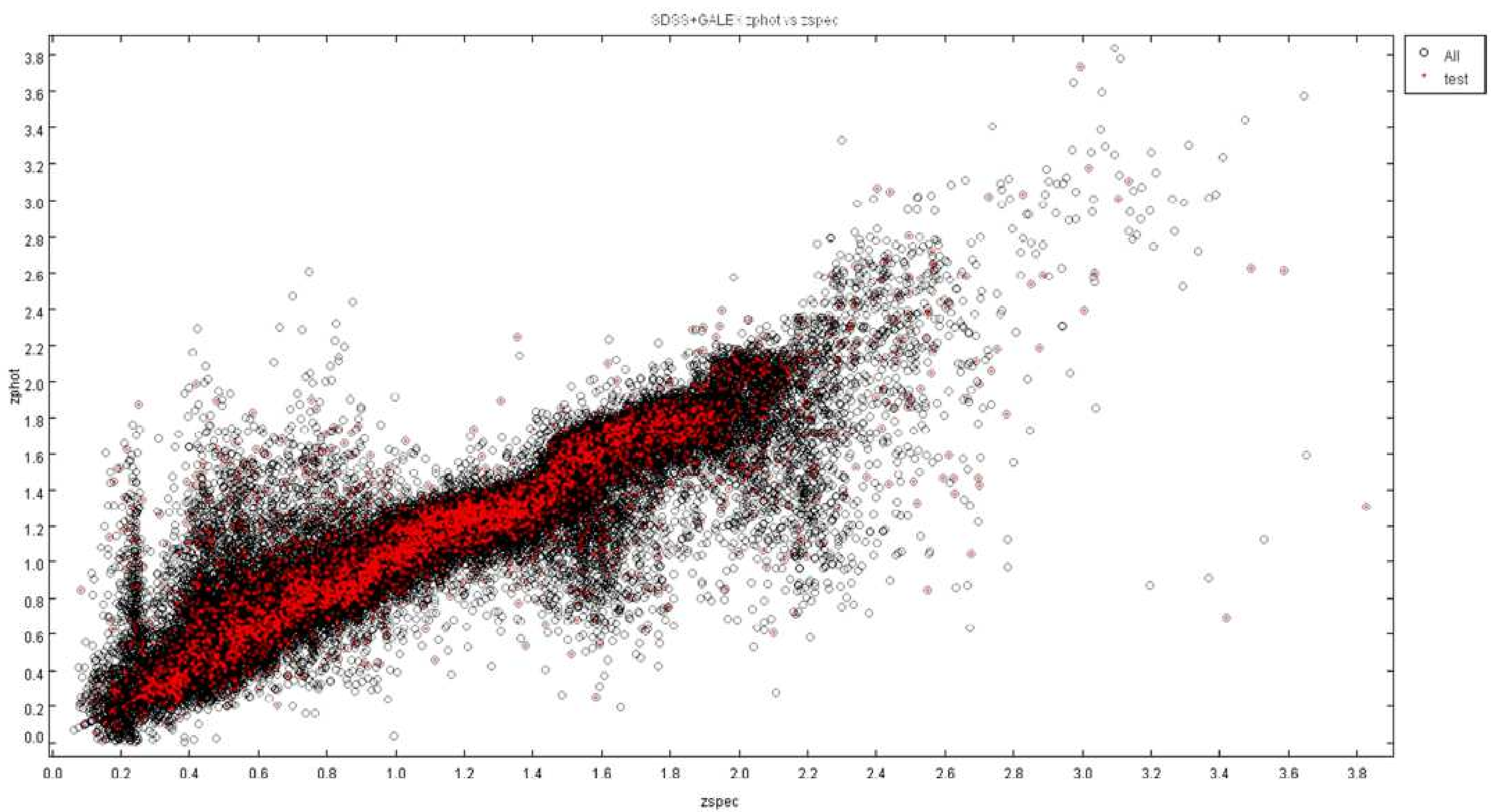
QSO Redshift – SDSS



Ref.	|bias|	sigma	MAD	RMS	|biasnorm|	snorm	MADnorm	RMSnorm
MLPQNA	0.007	0.25	0.102	0.26	0.032	0.15	0.039	0.17
Bovy 2012		0.46						
Laurino 2011	0.210	0.28	0.110	0.35	0.095	0.16	0.041	0.19
Ball 2010		0.35			0.095	0.18		
Richards 2009		0.52			0.115	0.28		

105759
objects

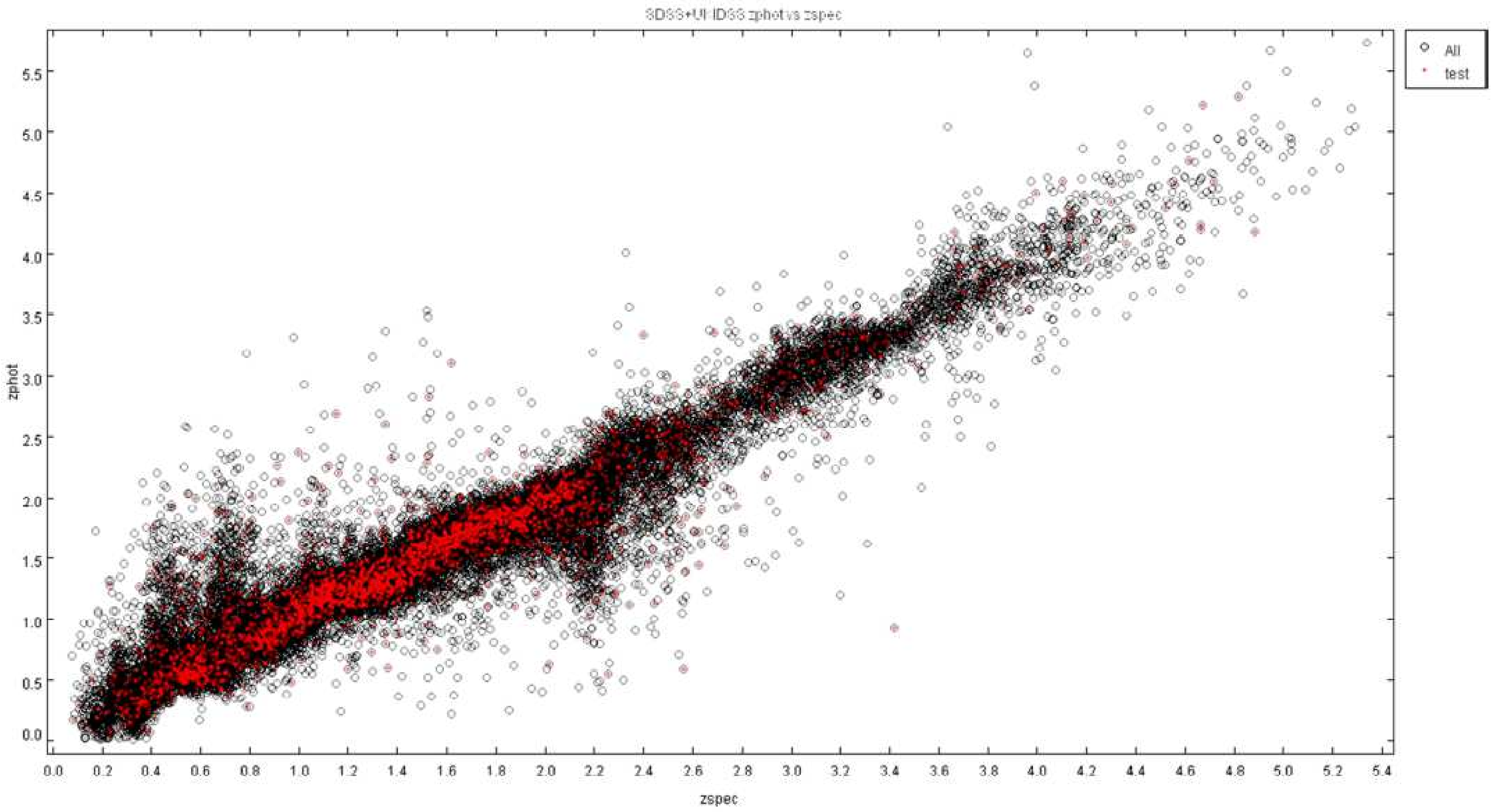
QSO Redshift – SDSS + GALEX



Ref.	|bias|	sigma	MAD	RMS	|biasnorm|	snorm	MADnorm	RMSnorm
MLPQNA	0.003	0.21	0.060	0.22	0.012	0.11	0.029	0.12
Bovy 2012		0.26						
Laurino 2011	0.13	0.21	0.061	0.25	0.058	0.29	0.029	0.11
Ball 2010		0.23			0.06	0.12		
Richards 2009		0.37			0.071	0.18		

44688 objects

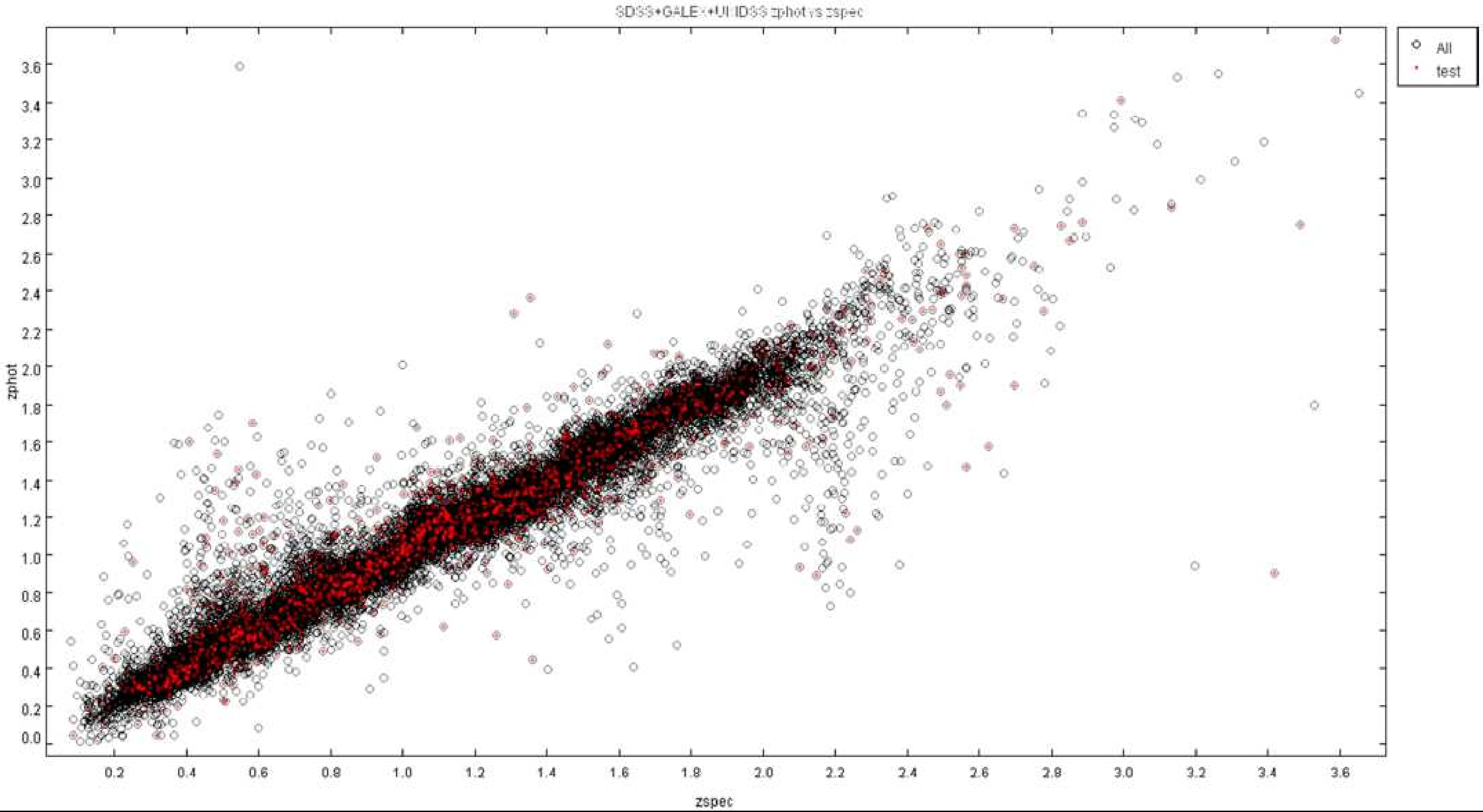
QSO Redshift – SDSS + UKIDSS



Ref.	|bias|	sigma	MAD	RMS	|biasnorm|	snorm	MADnorm	RMSnorm
MLPQNA	0.003	0.21	0.084	0.21	0.010	0.11	0.040	0.11
Bovy 2012		0.28						

31094
objects

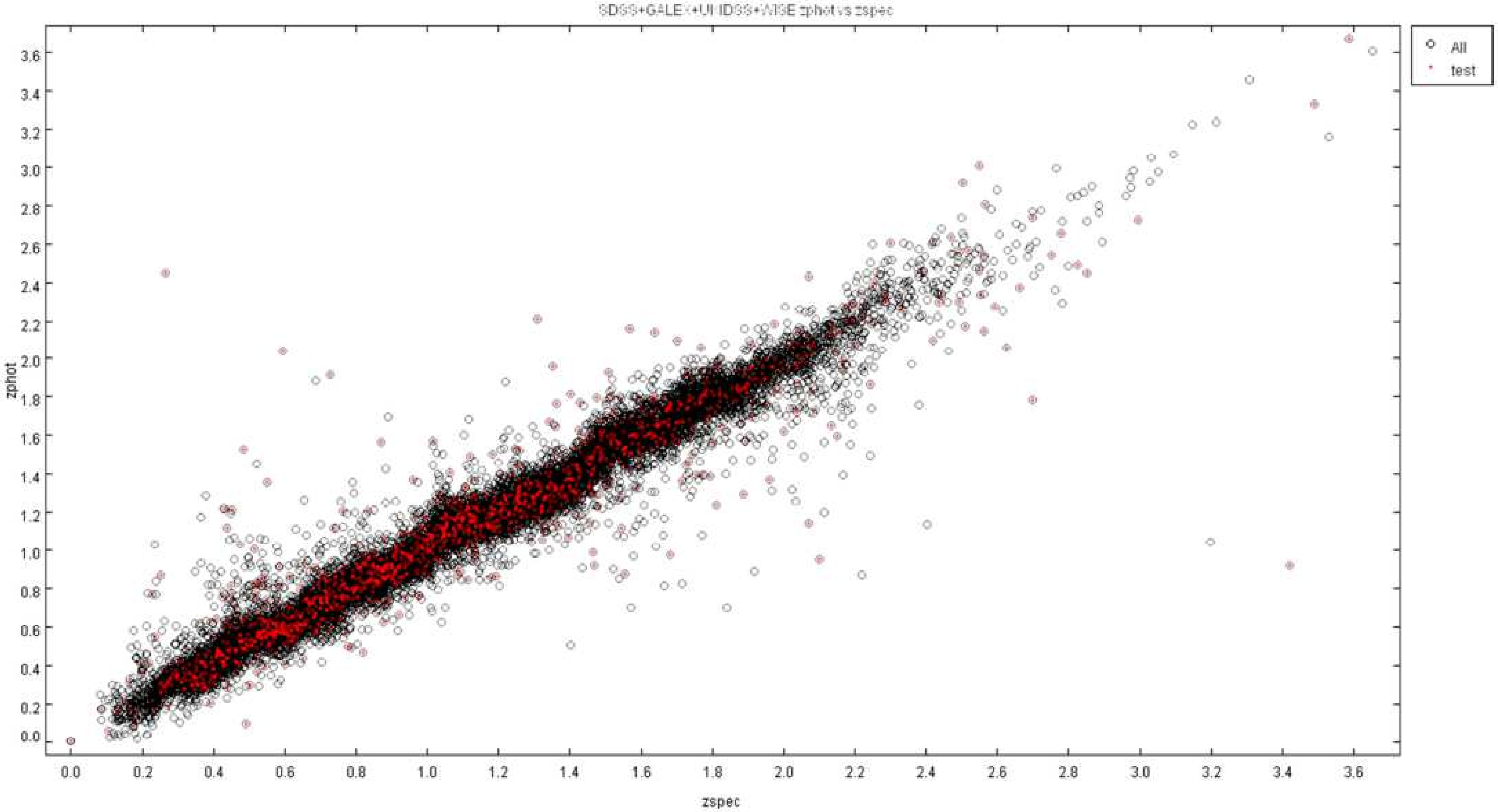
QSO Redshift – SDSS + UKIDSS + GALEX



Ref.	|bias|	sigma	MAD	RMS	|biasnorm|	snorm	MADnorm	RMSnorm
MLPQNA	0.005	0.15	0.072	0.15	0.006	0.075	0.036	0.075
Bovy 2012		0.21						

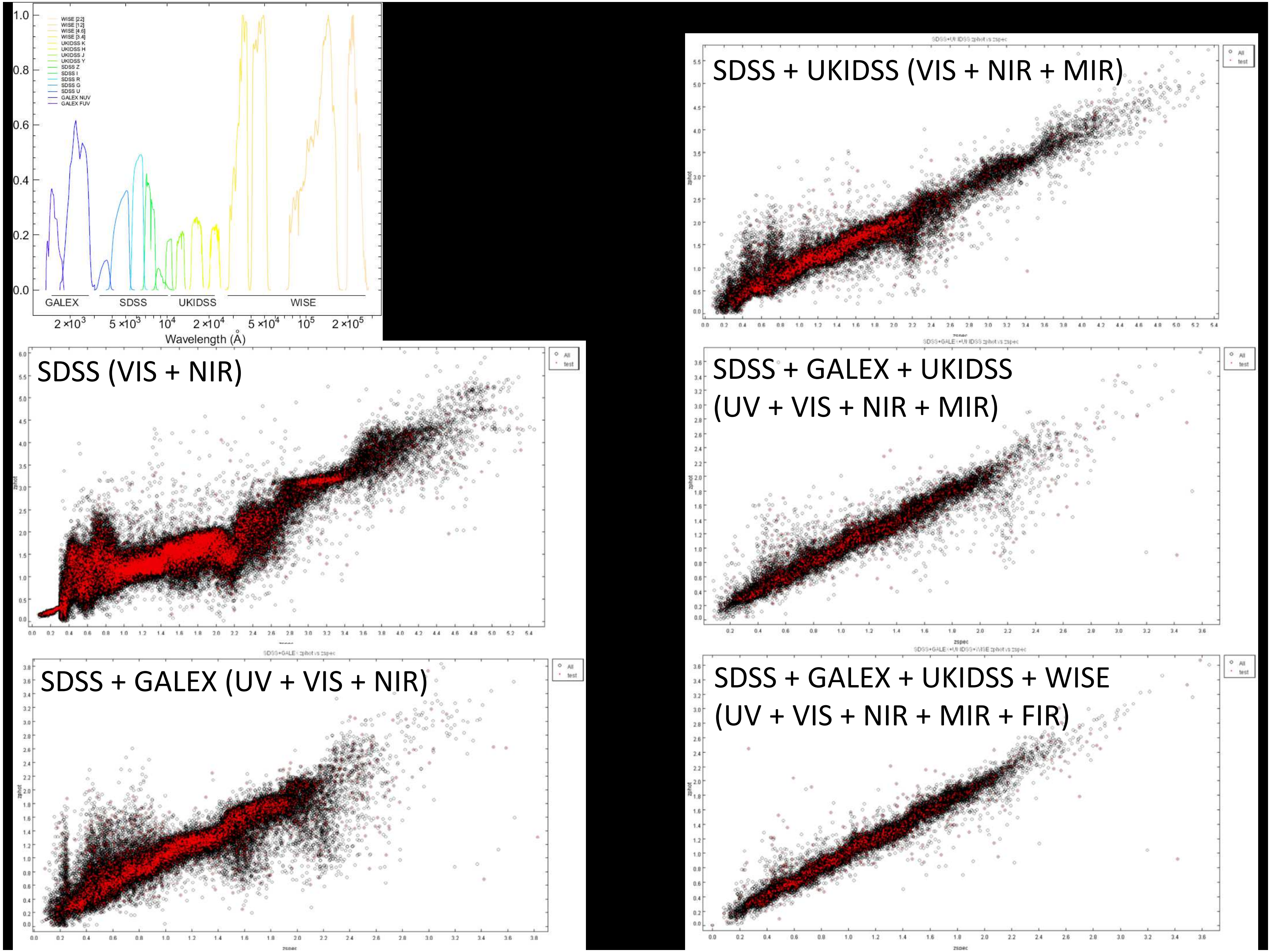
14588
objects

QSO Redshift – SDSS + UKIDSS + GALEX + WISE



Ref.	|bias|	sigma	MAD	RMS	|biasnorm|	snorm	MADnorm	RMSnorm
MLPQNA	0.003	0.15	0.063	0.15	0.005	0.15	0.063	0.15

14291
objects


GALEX
SDSS
UKIDSS
WISE
Wavelength (Å)
SDSS + UKIDSS (VIS + NIR + MIR)
SDSS (VIS + NIR)
SDSS + GALEX + UKIDSS
(UV + VIS + NIR + MIR)
SDSS + GALEX (UV + VIS + NIR)
SDSS + GALEX + UKIDSS + WISE
(UV + VIS + NIR + MIR + FIR)
zphot
zspec
All
test

QSO Redshift overal comparison

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

Exp	$BIAS(\Delta z)$	$\sigma(\Delta z)$	$MAD(\Delta z)$	$RMS(\Delta z)$
SDSS				
MLPQNA	0.007	0.25	0.102	0.26
Bovy et al.	-	0.46	-	-
Laurino et al.	0.210	0.28	0.110	0.35
Ball et al.	-	0.35	-	-
Richards et al.	-	0.52	-	-
SDSS + GALEX				
MLPQNA	0.003	0.21	0.060	0.22
Bovy et al.	-	0.26	-	-
Laurino et al.	0.13	0.21	0.061	0.25
Ball et al.	-	0.23	-	-
Richards et al.	-	0.37	-	-
SDSS + UKIDSS				
MLPQNA	0.001	0.25	0.066	0.26
Bovy et al.	-	0.28	-	-
SDSS + GALEX + UKIDSS				
MLPQNA	0.0009	0.18	0.043	0.19
Bovy et al.	-	0.21	-	-
SDSS + GALEX + UKIDSS + WISE				
MLPQNA	0.002	0.15	0.040	0.15

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

QSO Redshift conclusions

What we learned:

- Additional Bands are more important than additional points in the training sets;
- Wing Degeneracies fade out with wavelength coverage;
- Photometric redshift are complex enough to require the violation of the "Haykin theorem".

Brescia, M.; Cavuoti, S.; D'Abrusco, R.; Longo, G.; Mercurio, A.; 2013, Photo-z prediction on WISE-GALEX-UKIDSS-SDSS Quasar Catalogue, based on the MLPQNA model, **Apj, accepted (in press)**


CLASSIFICATION PROBLEMS:
AGN
Globular Clusters in external galaxies
Variable Sky (started)

AGN CLASSIFICATION

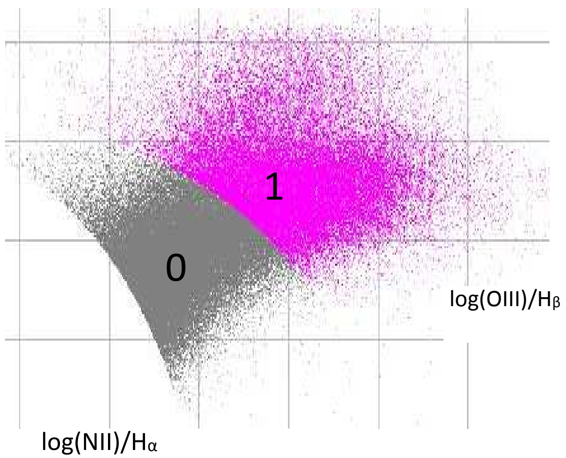
Photometric parameters used for training of the NNs and SVMs:

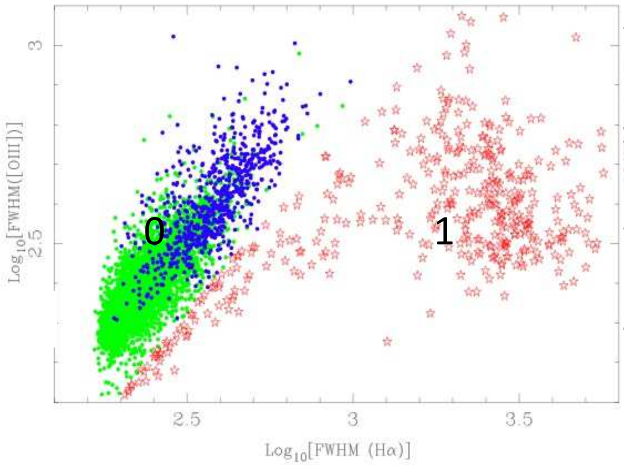
petroR50_u, petroR50_g, petroR50_r, petroR50_i, petroR50_z
concentration_index_r
fibermag_r
$(u-g)_{dered}$, $(g-r)_{dered}$, $(r-i)_{dered}$, $(i-z)_{dered}$
dered_r
photo_z_corr

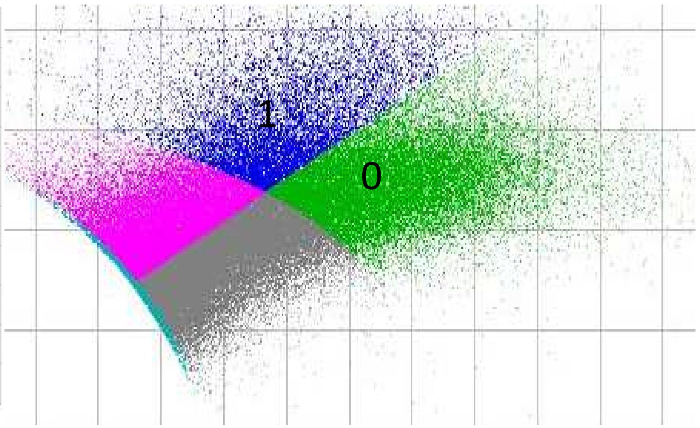
1° Experiment: AGN -> 1, Mixed-> 0

2° Experiment: Type 1 -> 1, Type 2 -> 0

3° Experiment: Seyfert -> 1, LINERs ->0







Cavuoti, S.; Brescia, M.; D'Abrusco, R.; Longo, G.; Photometric AGN Classification in the SDSS with Machine Learning Methods **to be Submitted to MNRAS**

AGN CLASSIFICATION RESULTS

Sample	Parameters	BoK	Algorithm	e_{tot}	C(MLP)
Experiment (1) AGN detection	SDSS photometric parameters + photo redshift	**BPT plot +Kewley's line**	SVM / MLP	~74% / ~76%	AGN~55% / NotAGN ~87%
Experiment (2) Type 1 vs. Type 2	SDSS photometric parameters + photo redshift	**Catalogue of Sorrentino et al.+Kewley's line**	SVM / MLP	e_{typ1}~82% e_{typ2}~86% / e_{typ2}~99% e_{typ1}~98%	Type1 ~99% / Type2 ~100%
Experiment (3) Seyfert Vs. LINERs	SDSS photometric parameters + photo redshift	**BPT plot+Heckman's+Kewley's lines**	SVM / MLP	Sey~78% / LIN~80%	Sey~53% / LIN~92%

- Checking the trained NN with a dataset of sure not AGN just 12.6% are false positive
- False positive surely not AGN (according BoK) are 0.89%

Globular Cluster Recognition

NGC1399 Dataset

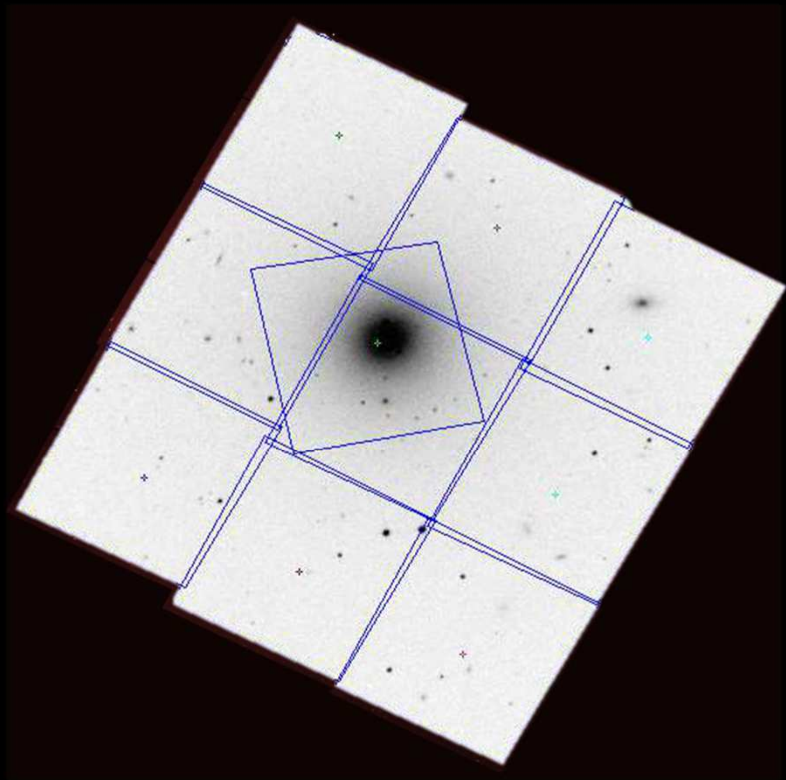
NGC1399 (~20 Mpc) is an ideal target because allows to probe a large fraction of the galaxy and still resolve GC sizes.

9 HST V-band (f606w) observations, drizzled to super-Nyquist sampling the ACS PSF (2.9 pc/pix).

Chandra ACIS-I + ACIS-S

ACS *g-z* colors for central region

Ground-based *C-R* photometry for part of the sources over the whole field



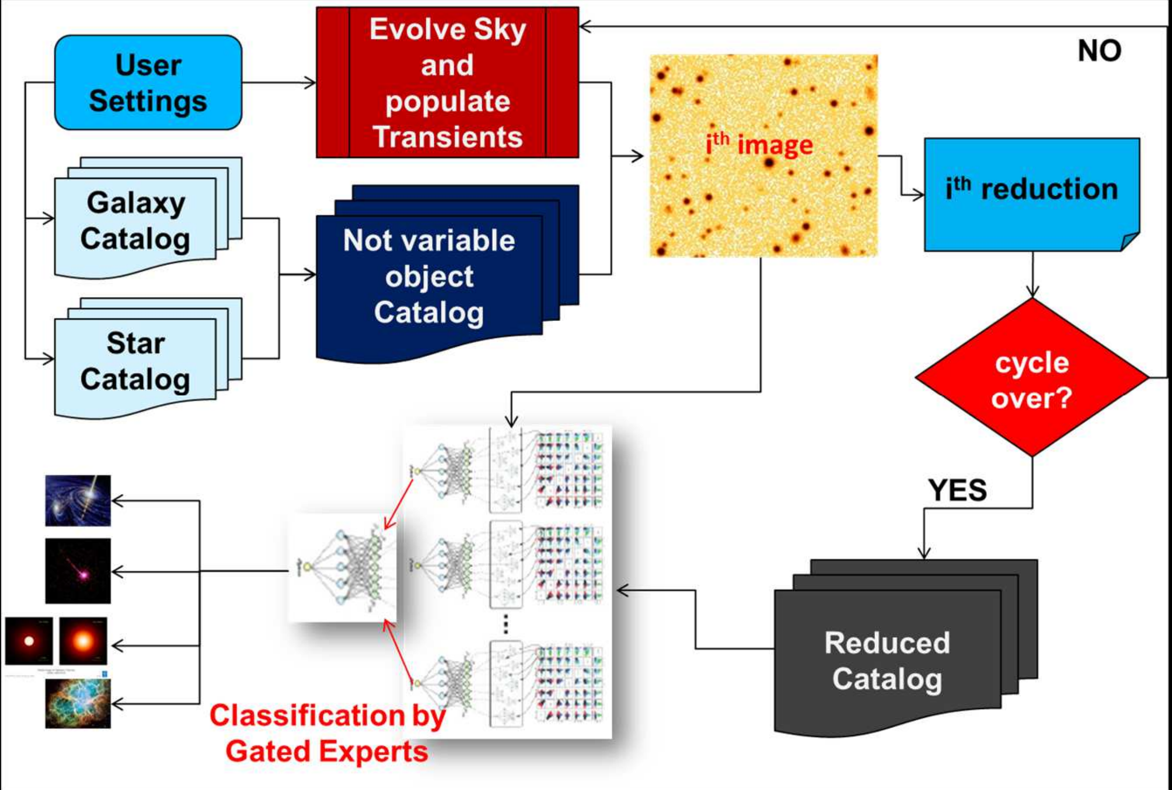
Brescia, M.; Cavuoti, S.; Paolillo, M.; Longo, G.; Puzia, T.; 2012, The detection of Globular Clusters in galaxies as a data mining problem, **MNRAS, Volume 421, Issue 2, pp. 1155-1165, available at arXiv:1110.2144v1**

Quality and pruning results

Type of experiment	Missing features	Figure of merit	MLPQNA	GAME	SVM	MLPBP	MLPGA
Complete patterns	–						
		class.accuracy	**98.3**	82.1	90.5	59.9	66.2
		completeness	**97.8**	73.3	89.1	54.1	61.4
		contamination	**1.8**	18.7	7.7	42.2	35.1
No par. 11	11						
		class.accuracy	**98.0**	81.9	90.5	59.0	62.4
		completeness	**97.6**	79.3	88.9	56.1	62.2
		contamination	**1.6**	19.6	7.9	43.1	38.8
Only optical	8, 9, 10, 11						
		class.accuracy	93.9	86.4	90.9	70.3	76.2
		completeness	91.4	78.9	88.7	54.0	65.1
		contamination	5.9	13.9	8.0	33.2	24.6
Mixed	5, 8, 9, 10, 11						
		class.accuracy	94.7	86.7	89.1	68.6	71.5
		completeness	92.3	81.5	88.6	52.8	63.8
		contamination	5.0	16.6	8.1	37.6	30.1

- **isophotal magnitude** (feature 1);
- **3 aperture magnitudes** (features 2–4) obtained through **circular apertures of radii 2**, **6** and **20 arcsec**, respectively;
- **Kron radius**, **ellipticity** and the **FWHM** of the image (features 5–7);
- **4 structural parameters** (features 8–11) which are, respectively, the **central surface brightness**, the **core radius**, the **effective radius** and the **tidal radius**;

STraDiWA



Prototipation of a web tool (**STraDiWA**, *Sky Transient Discovery Web Application*) for detection and classification of transients from simulated images.

The pipeline includes an automatic system for the extraction of the catalogues from syntetic images.

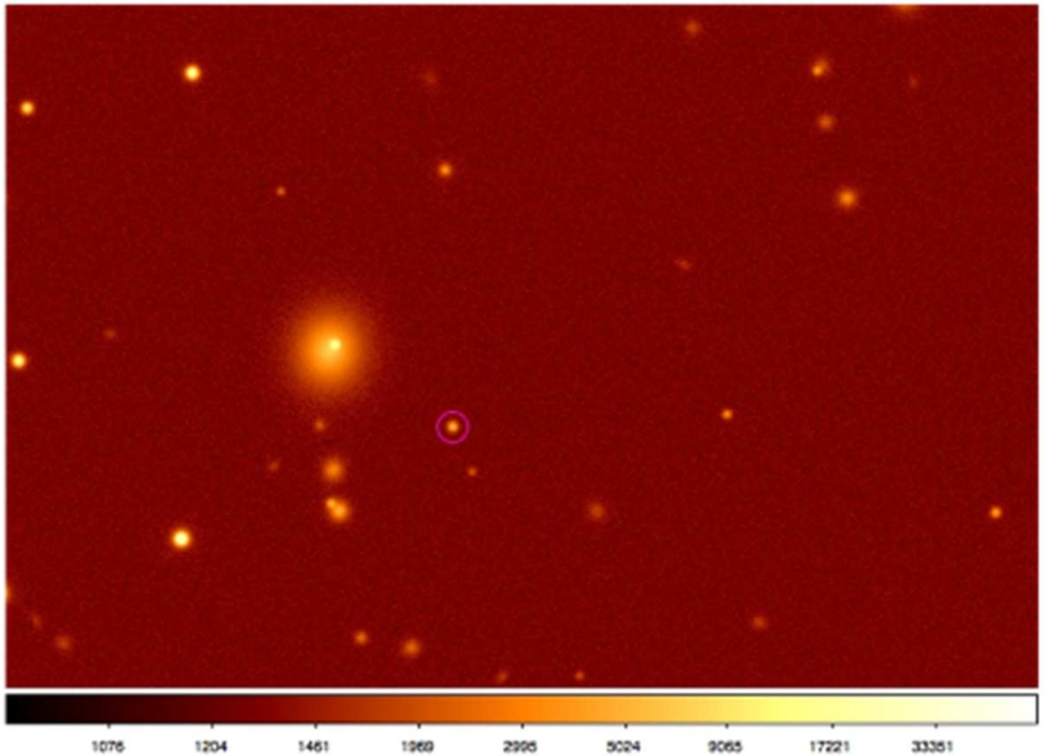
Modeling of transients, Cepheids and Supernovae Ia

Annunziatella, M.; Mercurio, A.; Brescia, M.; **Cavuoti, S.**; Longo, G, "*Inside catalogs: a comparison of source extraction software*", **2013**, **PASP**

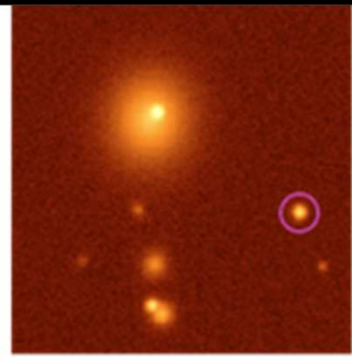
Cepheid example – VST Image simulated

A classical Cepheid is modeled:

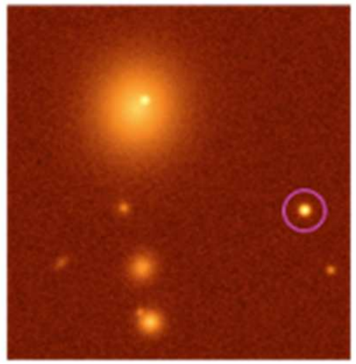
- Assuming a sinusoidal law.
- Imposing a PL luminosity relation.
- We used the coefficients for the mean PL relation calibrated in Bono et al. 2010 and references therein.



Stamp of the image of a Classical Cepheid of Period of about 20 days at t = 0 days.



Cepheid at t = 13 d.



Cepheid at t = 35 d.

SN-1a example – VST Image simulated

A classical SN-Ia is modeled:

- using an analytical function, used in Contardo, Leibundgut, and Vacca 2000 for the fit of a sample of type Ia Supenovae.

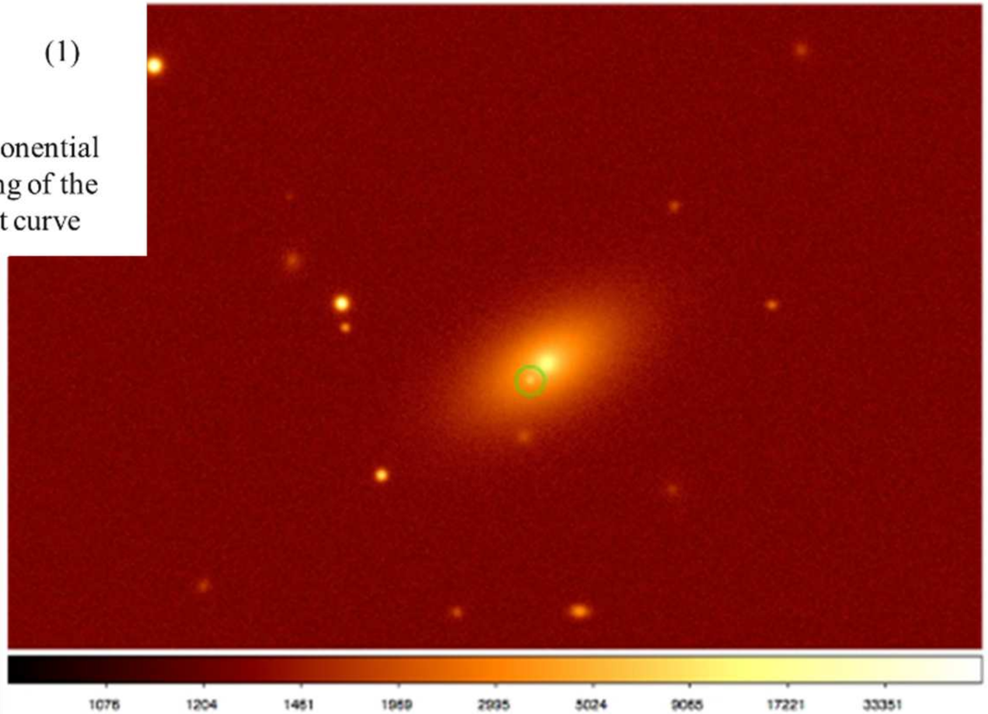
Linear decay

Second Maximum phase

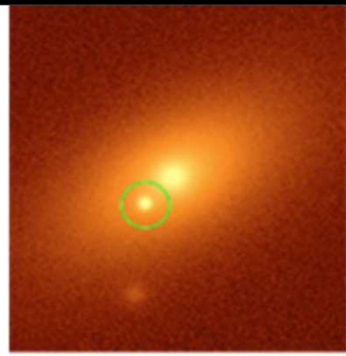
$$m(t) = \frac{f_0 + \gamma(t - t_0) + g_0 e^{\frac{(t-t_0)^2}{2\sigma_0^2}} + g_1 e^{\frac{(t-t_1)^2}{2\sigma_1^2}}}{(1 - e^{\frac{\tau - t}{\theta}})} \quad (1)$$

First Maximum

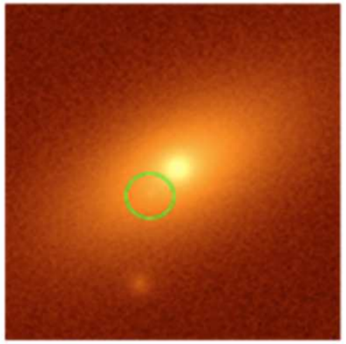
Exponential rising of the light curve



Stamp of the image of a type Ia at t = 0 d. and B band



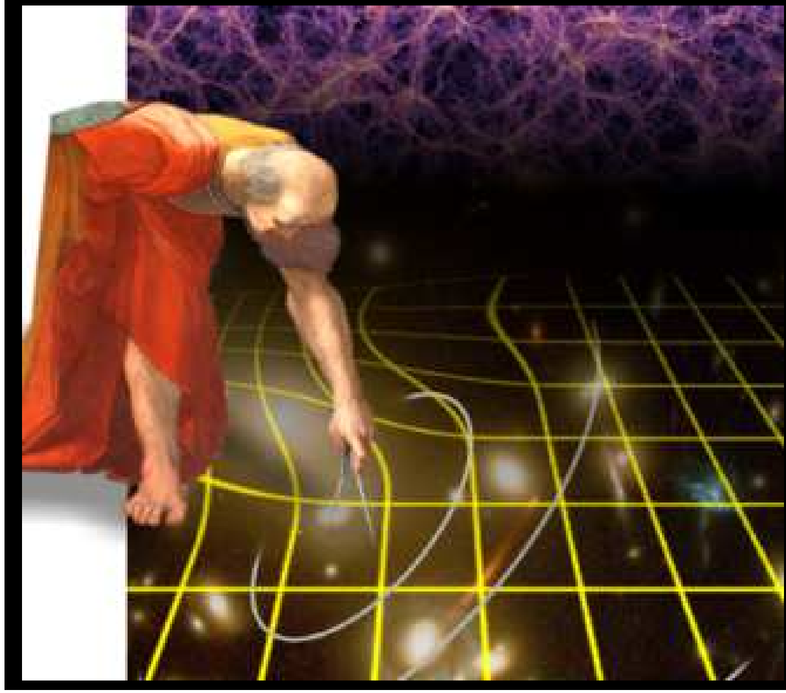
Stamp of a type Ia at t = 13 d.

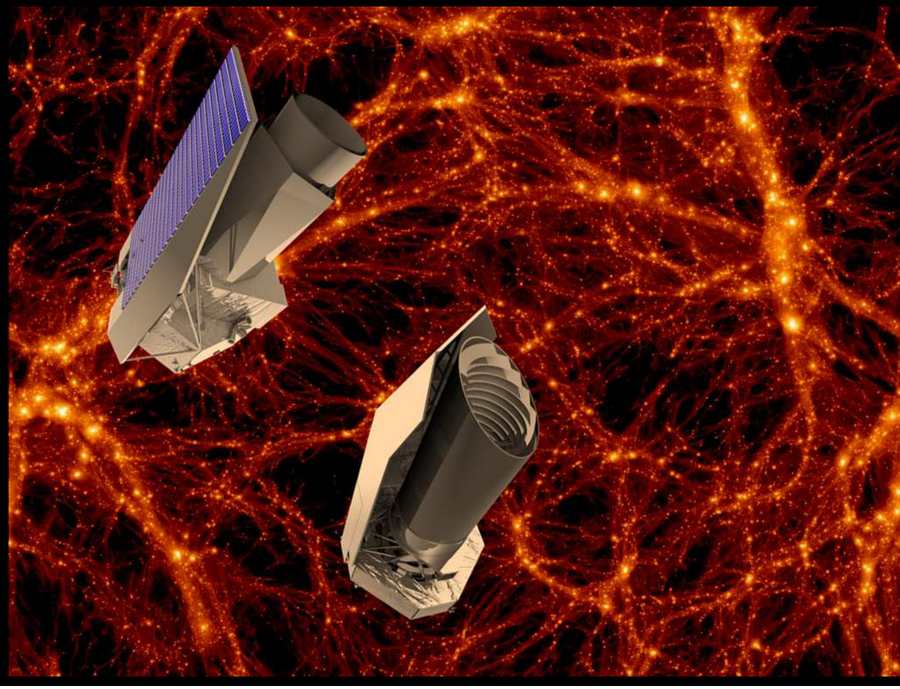


Stamp of a type Ia at t = 63 d.

For Completeness…EUCLID

Euclid is an ESA mission medium class mission selected for launch in 2019 in the Cosmic Vision 2015-2025 programme to map the geometry of the dark Universe. The mission will investigate the distance-redshift relationship and the evolution of cosmic structures by measuring shapes and redshifts of galaxies and clusters of galaxies out to redshifts ~2, or equivalently to a look-back time of 10 billion years. In this way, Euclid will cover the entire period over which dark energy played a significant role in accelerating the expansion.



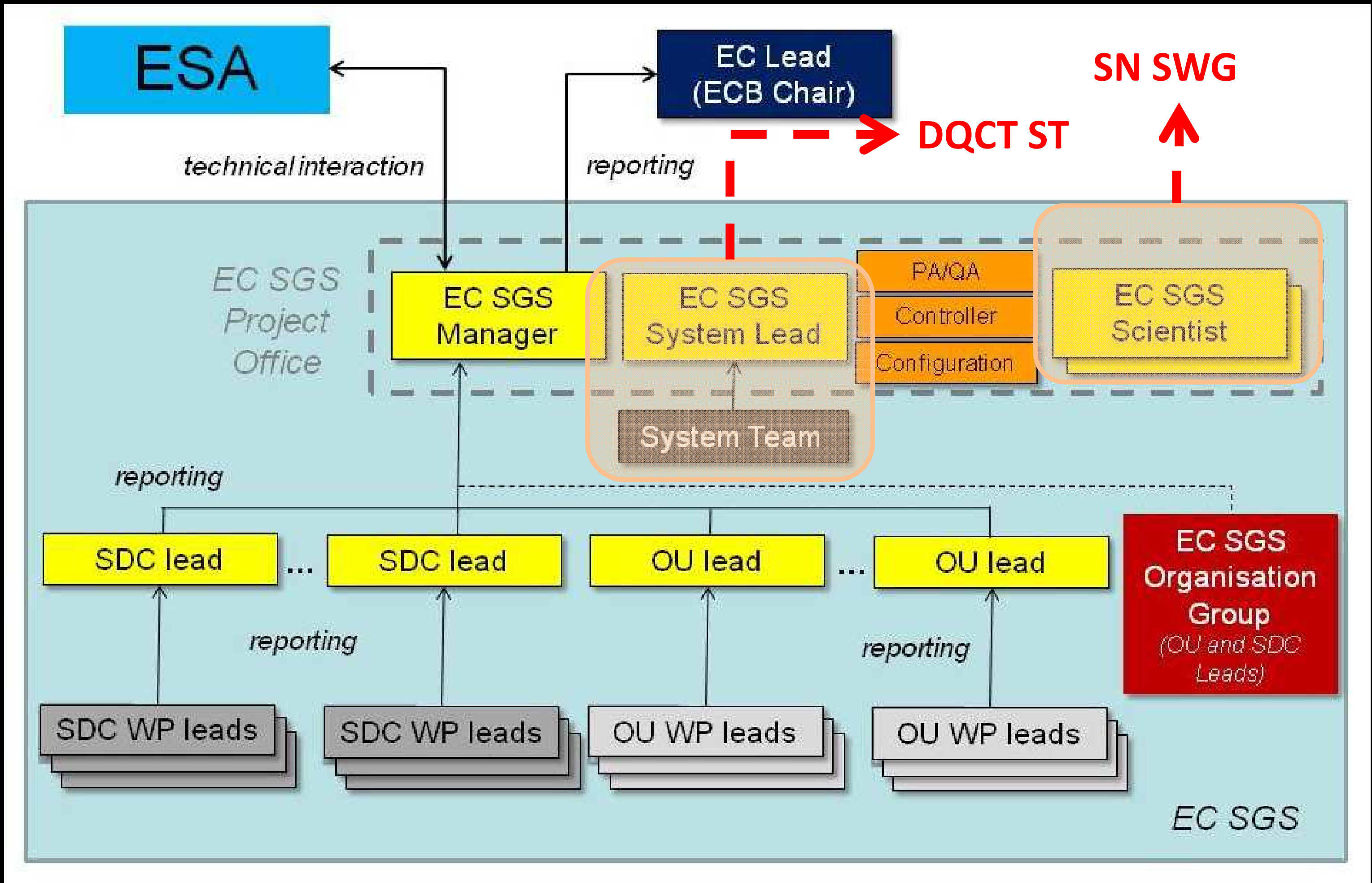


For Completeness...EUCLID

In the Euclid project, I'm involved, since Jan 2012 in two tasks:

- Science Team (Italy, Norway and Finland) for the design and development of Data Quality Common Tools
- Science Working group for the Legacy Science requirements definitions dedicated to transient objects detection and classification.

Recently, after the stunning results that we obtained, we joined also the photometric redshift team!!!



Publications I - Refeered Papers

Technological

1. Brescia, M.; Cavuoti, S.; Garofalo, M.; Guglielmo, M.; Longo, G.; Nocella, A.; Riccardi, S.; Vellucci, C.; Djorgovski, G.S.; Donalek, C.; Mahabal, A. Data Mining in Astronomy with DAME. **to be Submitted to PASP**

Algorithmic

2. Cavuoti, S.; Garofalo, M.; Brescia , M.; Paolillo, M.; Pescape', A.; Longo, G.; Ventre, G.; GPUs for astrophysical data mining. A test on the search for candidate globular clusters in external galaxies, **New Astronomy**

Scientific

3. Cavuoti, S.; Brescia, M.; D'Abrusco, R.; Longo, G.; Photometric AGN Classification in the SDSS with Machine Learning Methods **to be Submitted to MNRAS**
4. Brescia, M.; Cavuoti, S.; D'Abrusco, R.; Longo, G.; Mercurio, A.; 2013, Photo-z prediction on WISE-GALEX-UKIDSS-SDSS Quasar Catalogue, based on the MLPQNA model, **Submitted to Apj (in press)**
5. Annunziatella, M.; Mercurio, A.; Brescia, M.; Cavuoti, S.; Longo, G.; 2012, Inside catalogs: a comparison of source extraction software, **PASP, Vol. 125, Nr. 923, pp. 68-82**
6. Cavuoti, S.; Brescia, M.; Longo, G.; Mercurio, A.; 2012, Photometric Redshifts with Quasi Newton Algorithm (MLPQNA). Results in the PHAT1 Contest, **A&A, Vol. 546, A13, pp. 1-8**
7. Brescia, M.; Cavuoti, S.; Paolillo, M.; Longo, G.; Puzia, T.; 2012, The detection of Globular Clusters in galaxies as a data mining problem, **MNRAS, Volume 421, Issue 2, pp. 1155-1165, available at arXiv:1110.2144v1**
8. Brescia, M.; Cavuoti, S.; Longo, G., Photometric Redshifts for all galaxies in the SDSS DR9 with the MLPQNA method", in preparation, **to be submitted to A&A**

Publications II - Proceedings

1. **Cavuoti, S.**; Brescia, M.; Longo, G., 2012, Data mining and Knowledge Discovery Resources for Astronomy in the Web 2.0 Age, Proceedings of SPIE Astronomical Telescopes and Instrumentation 2012, Software and Cyberinfrastructure for Astronomy II, Ed.(s): N. M. Radziwill and G. Chiozzi, Volume 8451, RAI Amsterdam, Netherlands, July 1-4 **refeered proceeding**
2. **Cavuoti, S.**; Garofalo, M.; Brescia, M.; Pescape', A.; Longo, G.; Ventre, G., Genetic Algorithm Modeling with GPU Parallel Computing Technology" in "Neural Nets and Surroundings, Smart Innovation, Systems and Technologies", Vol. 19, p. 11, Springer **refeered proceeding**
3. Brescia, M., Cavuoti, S., Djorgovski, G.S., ,Donalek, C., Longo, G.,,Paolillo, M., "Extracting knowledge from massive astronomical data sets", 2012, in "Astrostatistics and Data Mining", Springer Series in Astrostatistics, Volume 2, Springer Media New York, ISBN 978-1-4614-3322-4 **volume contribute**
4. Brescia M., Cavuoti S., D'Abrusco R., Laurino O., Longo G. "DAME: A distributed data mining and exploration framework within the Virtual Observatory",, 2011, in "Remote Instrumentation for eScience and Related Aspects", F. Davoli et al. (eds.), Springer Science+Business Media, LLC 2011, ISBN 978-1-4614-0508- **volume contribute**
5. Brescia M., **Cavuoti, S.**, Djorgovski, G.S., ,Donalek, C., Longo, G., Paolillo, M., 2011, Extracting knowledge from massive astronomical data sets, arXiv:1109.2840, to appear in Astrostatistics and data mining in large astronomical databases, L.M. Barrosaro et al. eds, Springer Series on Astrostatistics, 15 pages **invited review**.
6. **Cavuoti, S.**; Brescia, M.; Longo, G.; Garofalo, M.; Nocella, A.; 2012,DAME: A Web Oriented Infrastructure for Scientific Data Mining and Exploration,Science - Image in Action. Edited by Bertrand Zavidovique (Universite' Paris-Sud XI, France) and Giosue' Lo Bosco (University of Palermo, Italy) . Published by World Scientific Publishing Co. Pte. Ltd., 2012. ISBN 9789814383295, pp. 241-247
7. Djorgovski, S. G.; Longo, G., Brescia, M., Donalek, C., **Cavuoti, S.**, Paolillo, M., D'Abrusco, R., Laurino, O., Mahabal, A., Graham, M., DAta Mining and Exploration (DAME): New Tools for Knowledge Discovery in Astronomy. American Astronomical Society, AAS Meeting #219, #145.12, Tucson, USA, January 08-12
8. Brescia, M.; **Cavuoti, S.**; D'Abrusco, R.; Laurino, O.; Longo, G.; 2010, DAME: A Distributed Data Mining & Exploration Framework within the Virtual Observatory, INGRID 2010 Workshop on Instrumenting the GRID, Poznan, Poland, in Remote Instrumentation for eScience and Related Aspects, F. Davoli et al. (eds.), Springer Science+Business Media, LLC 2011, DOI 10.1007/978-1-4614-0508-5 17
9. Brescia, M.; Longo, G.; Castellani, M.; **Cavuoti, S.**; D'Abrusco, R.; Laurino, O., 2012, DAME: A DistributedWeb Based Framework for Knowledge Discovery in Databases, 54th SAIT Conference, Astronomical Observatory of Capodimonte, Napoli, Italy, May 6, Mem. S.A.It. Suppl. Vol. 19, 324

Publications III – Technical Reports

1. Brescia, M.; Annunziatella, M.; Cavuoti, S.; Longo, G.; Mercurio, A.; STraDiWA Project Sky Transient Discovery Web Application SOFTWARE Documentation DAME-DOC-NA-0003-Rel1.0
2. Cavuoti, S.; Riccardi, S.; Guglielmo M.; DAMEWARE Installation and Deployment Developer Manual DAME-MAN-NA-0019-Rel1.0
3. Fiore, M.; Cavuoti, S.; Data Mining Plugin User/Administration Manual VONEURAL-MAN-NA-0005-Rel1.6
4. Fiore, M.; Cavuoti, S.; Data Mining PluginWizard User Manual VONEURALMAN-NA-0004-Rel1.3
5. Cavuoti, S.; Mercurio, A.; Annunziatella, M.; Brescia, M.; Variable Sky Objects Simulation and Detection Workflow Simulation Package Procedure DAME-PRO-NA-0010Rel2.0
6. Brescia, M.; Cavuoti, S.; Garofalo, M.; Nocella, A.; Riccardi S.; DAME Web Application REsource Design Summary DAMEWARE-SDD-NA-0018-Rel1.0
7. Cavuoti, S.; Di Guido, A.; Data Mining Suite 2.0 Software Design Description IEEE 1016 Component Data Mining Model VONEURAL-SDD-NA-0008-Rel2.0
8. Brescia, M.; Annunziatella, M.; Cavuoti, S.; Longo, G.; Mercurio, A.; STraDiWA Sky Transient Discovery Web Application Description of the Workflow SOFTWARE Specifications DAME-SPE-NA-0011-Rel1.0
9. Di Guido, A.; Fiore, M.; Cavuoti, S.; Brescia M.; DMPlugin Description Report Beta release ofWeb Application Data Mining Model Technical Report DAME-TRE-NA-0016-Rel1.0
10. Brescia, M.; Cavuoti, S.; DAMEWARE Web Application REsource Internal Test Report DAME-TRE-NA-0019Rel1.0
11. Brescia, M.; Cavuoti, S.; Photo-z prediction on PHAT1 Catalogue, based on MLPQNA regression model DAMEWARE-VER-NA-0008-Rel1.0

Conclusions, in the middle of the white Rabbit Hole…

Well, in conclusion… we have not yet (we'll never do) concluded,
in reality: we just started…

We obtained a lot of great results about redshifts and about the other issues, but this is not the core of this talk.

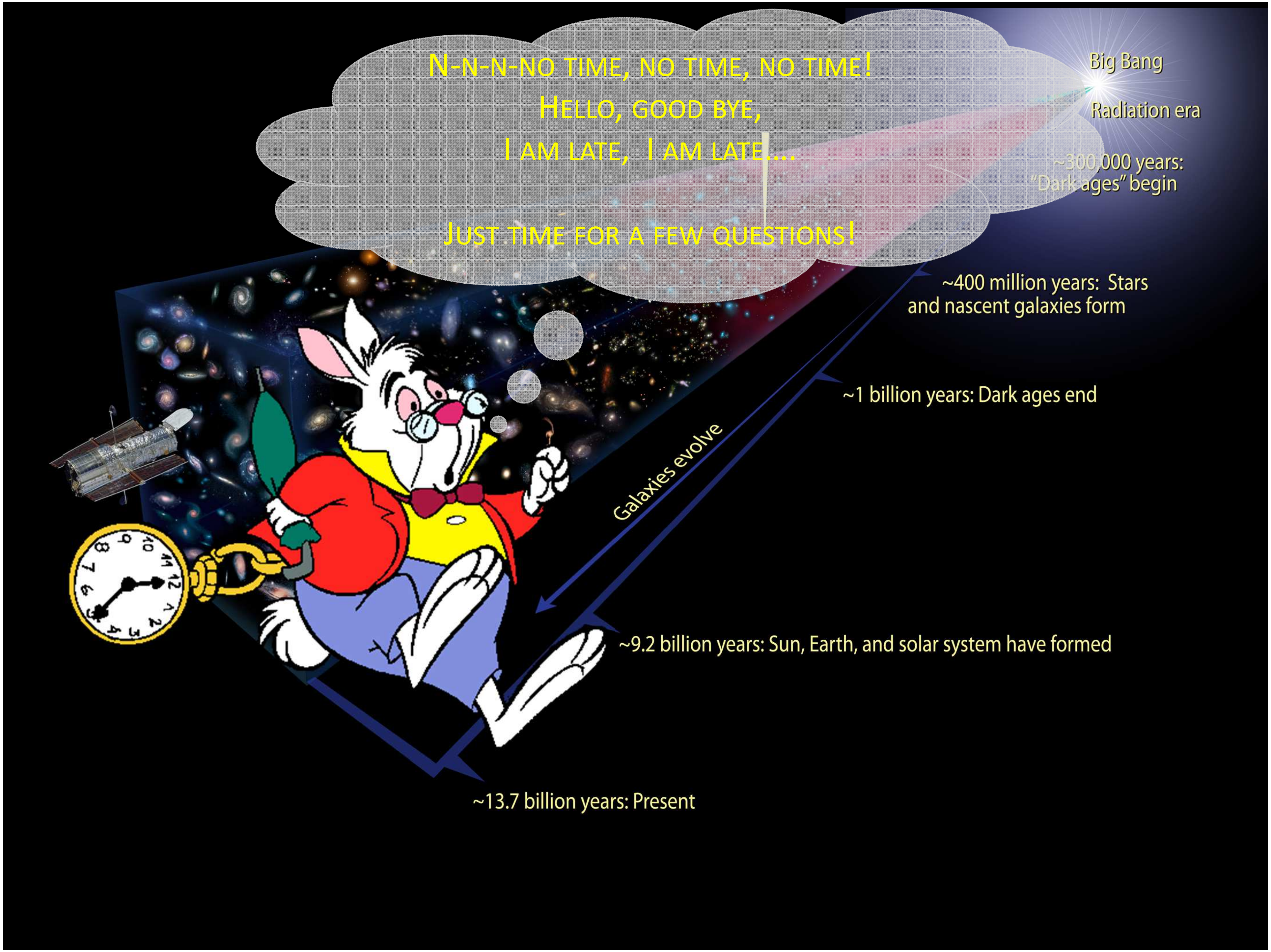
THE CORE IS:

For the Red Pill consumers: YES

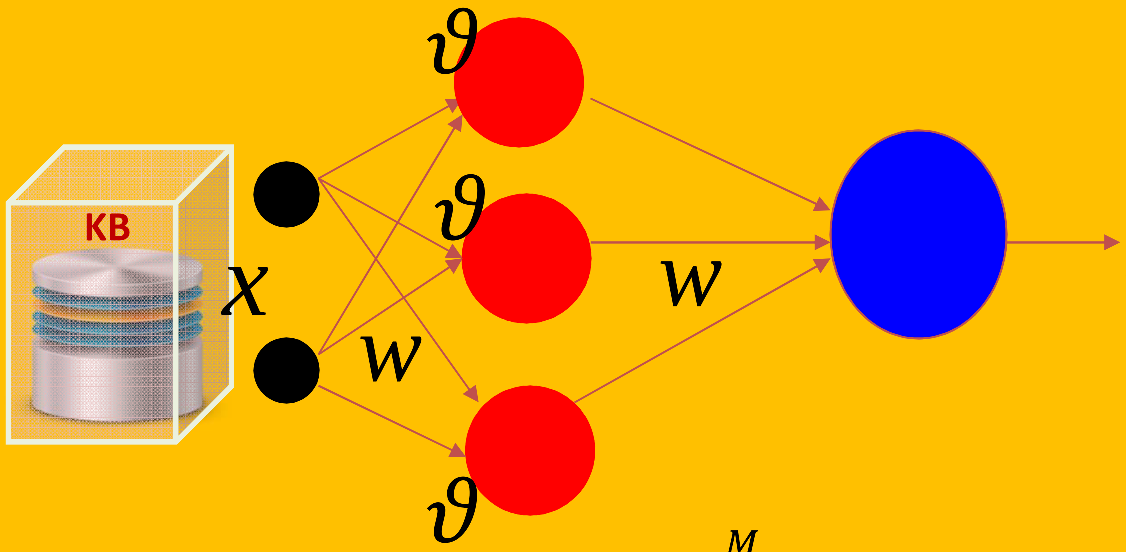
Astroinformatics is opening a new wide and encouraging door, and a new era of observational Astronomy has started.

For the Blue Pill consumers:

Don't worry, tomorrow you forget everything, you just have a little déjà vu…

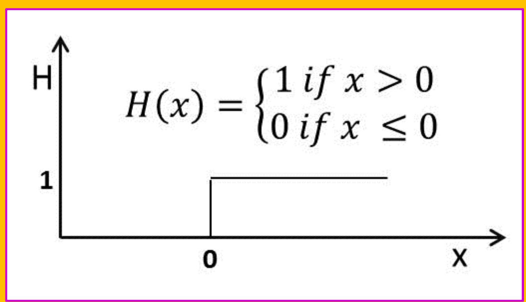

N-n-n-no time, no time, no time!
Hello, good bye,
I am late, I am late....
Just time for a few questions!
Big Bang
Radiation era
~300,000 years:
"Dark ages" begin
~400 million years: Stars
and nascent galaxies form
~1 billion years: Dark ages end
Galaxies evolve
~9.2 billion years: Sun, Earth, and solar system have formed
~13.7 billion years: Present

Multi Layer Perceptron

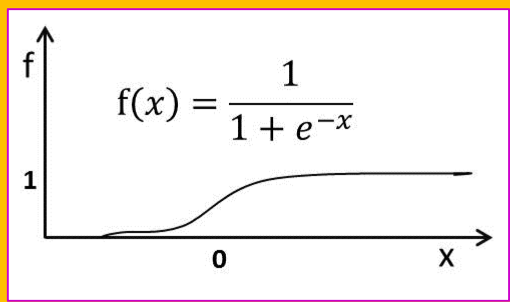


$$y(x;w,\vartheta) = GC/noGC$$

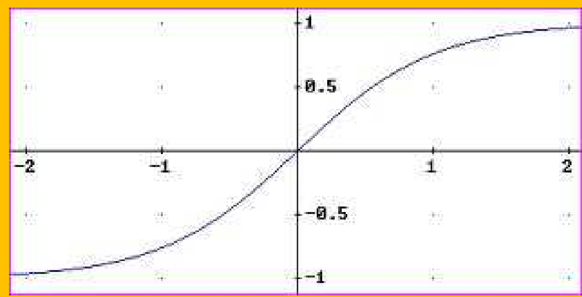
$$y(x;w,\vartheta) = \sum_{i=1}^{M} activ\_func(W_i^T x - \vartheta_i)$$



Heaviside



Sigmoidal

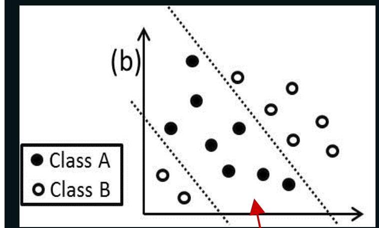


Hyperbolic tangent

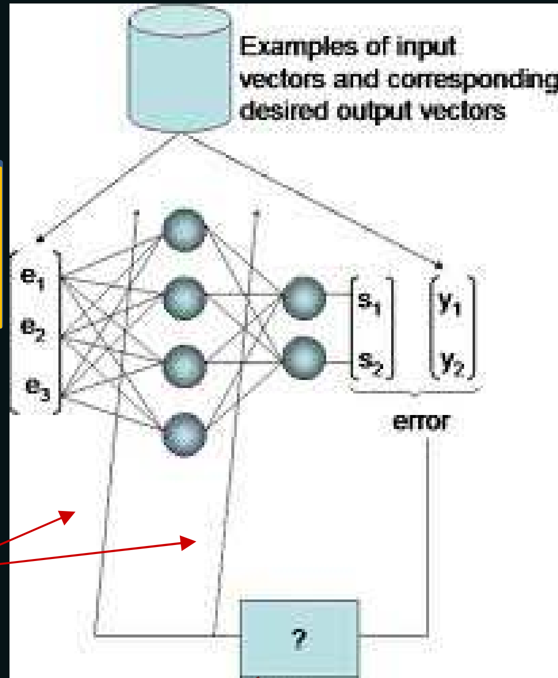
MLP learning phase

$$\min_{w} E(w) = \frac{1}{2P}\sum_{p=1}^{P} E_p(w) = \frac{1}{2P}\sum_{p=1}^{P}(y(x^p;w)-d^p)^2$$

E_p is a measure of the **error related to the *p*-th pattern**



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



$d^k \in R^N$ **DIRECTION OF SEARCH**

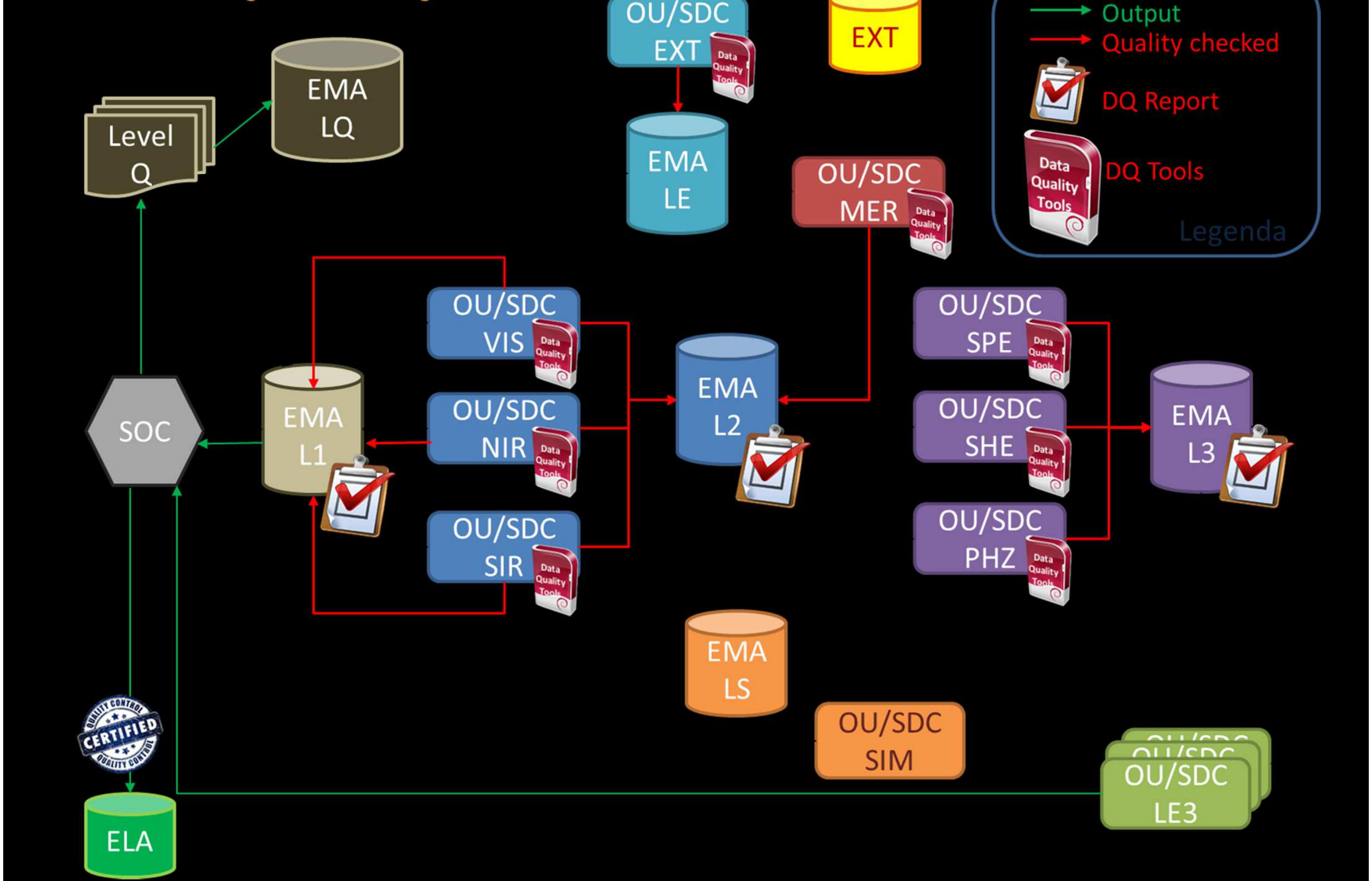
$\alpha^k \in R$ **STEP**

$$d^k = -\nabla E(w^k)$$ **Descent gradient (BP)**

$$d^k = genetic\ operators$$ **Genetic Algorithms (GA)**

$$\nabla^2 E(w^k) d^k = -\nabla E(w^k)$$ **Hessian approx. (QNA)**

EUCLID (DQCT)



"*Data Quality in Euclid*", Euclid Consortium Scientific Ground Segment document code EUCL-OAC-SGS-TN-00085 (ESA EUCLID Official Archive)

MDS with: $N > 10^9$, $D \gg 100$, $K > 10$

N = no. of data vectors,
D = no. of data dimensions
K = no. of clusters chosen,
K_{max} = max no. of clusters tried
I = no. of iterations, M = no. of Monte Carlo trials/partitions

K-means: $K \times N \times I \times \mathbf{D}$
Expectation Maximization: $K \times N \times I \times \mathbf{D^2}$
Monte Carlo Cross-Validation: $M \times K_{max}^2 \times N \times I \times \mathbf{D^2}$
Correlations ~ $N \log N$ or N^2, ~ D^k ($k \geq 1$)
Likelihood, Bayesian ~ N^m ($m \geq 3$), ~ D^k ($k \geq 1$)
SVM > ~ $(NxD)^3$

→ **Lots of computing power**

QSO Redshift conclusions

GALEX	SDSS	UKIDSS	WISE	*bias* (Δz)	σ (Δz)
		Service Experiments			
X	X	X	X	0.0033	0.174
X[1,2]	X	X[6]	X	-0.0001	0.152
X[3]	X	X[6]	X	-0.0016	0.165
X[1]	X	X[6]	X	0.0054	0.151
X[2]	X	X[6]	X	-0.0026	0.151
X[4,5]	X	X[6]	X	-0.0008	0.152
X[1,2,3]	X	X[6]	X	0.0041	0.163
X[2,3]	X	X[6]	X	-0.0033	0.155
		X[6,7]		-0.0091	0.299
		X[7]		0.0111	0.465
		X[6]		-0.0081	0.294
		Four Survey Experiment			
X[2]	X	X[6]	X	-0.0026	0.151
		Three Survey Experiment			
X[2]	X	X[6]		-0.0046	0.152
X[2]	X		X	0.0025	0.162
	X	X[6]	X	-0.0032	0.179
X[2]		X[6]	X	0.0110	0.203
		Two Survey Experiment			
		X[6]	X	0.0045	0.236
X[2]			X	0.0175	0.288
	X	X[6]		-0.0027	0.210
	X		X	-0.0039	0.197
X[2]	X			-0.0055	0.240
X[2]		X[6]		0.0133	0.238
		One Survey Experiment			
			X	0.0165	0.297
	X			-0.0162	0.338
X[1,2]				0.0550	0.419
		X[6]		-0.0081	0.294

[1] *mag*
[2] *mag_iso*
[3] *mag_Aper 1, 2 and 3*
[4] *mag_auto*
[5] *kron_radius*
[6] *HallMag*
[7] *PetroMag*

What we learned:

- Additional Bands are more important than additional points in the training sets;
- Wing Degeneracies fade out with the band coverage;
- Photometric redshift are complex enough to require the violation of the Haykin theorem.