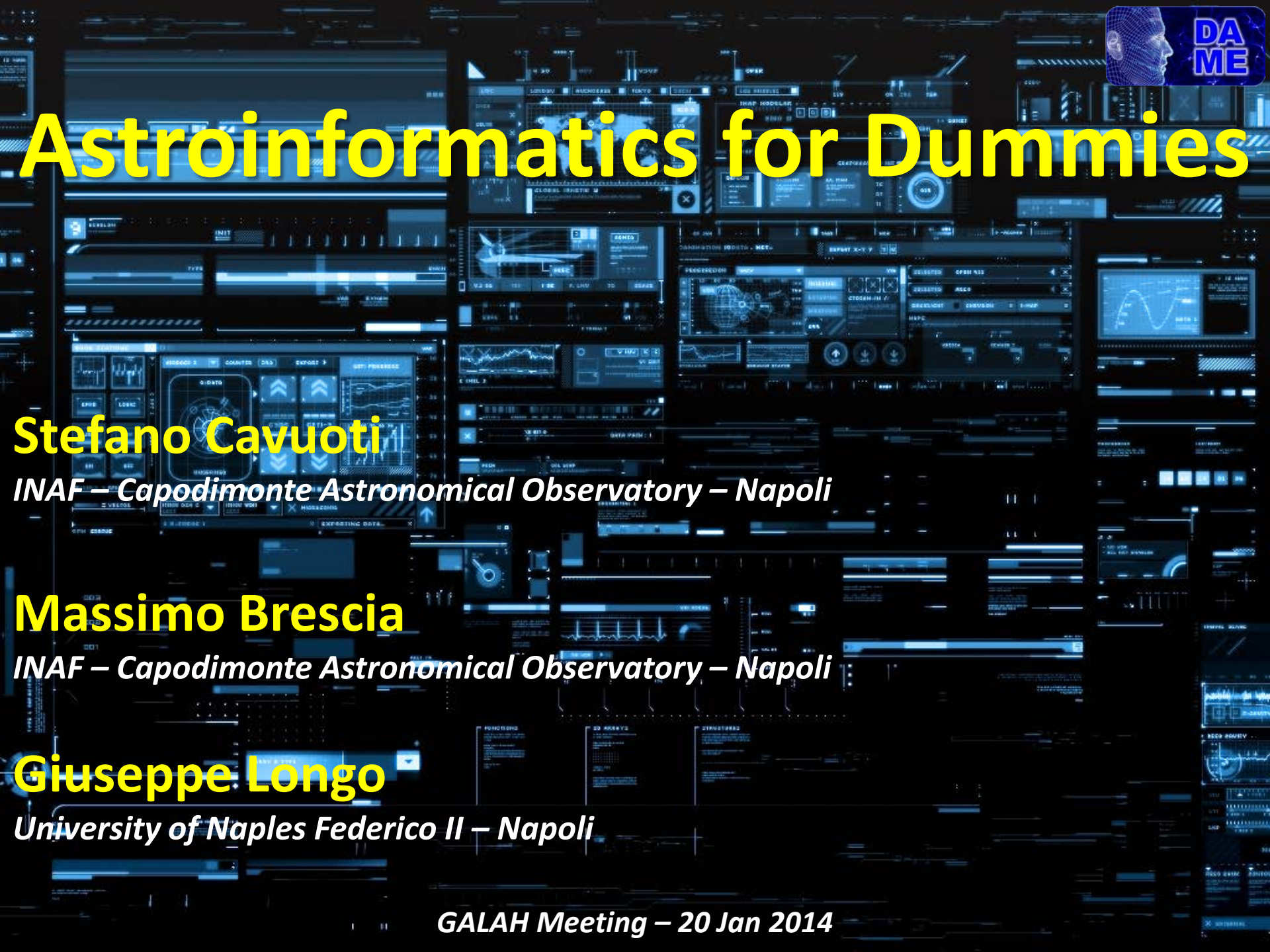


Astroinformatics for Dummies

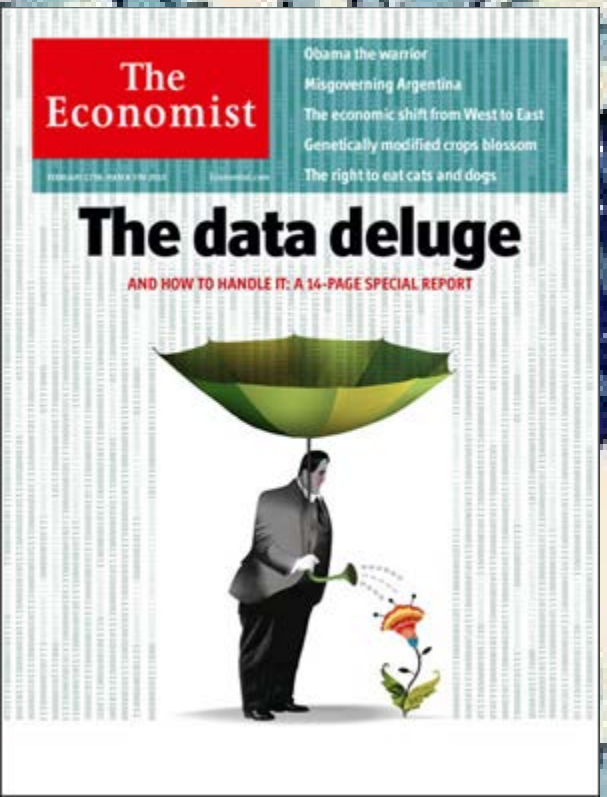
Stefano Cavuoti
INAF – Capodimonte Astronomical Observatory – Napoli

Massimo Brescia
INAF – Capodimonte Astronomical Observatory – Napoli

Giuseppe Longo
University of Naples Federico II – Napoli

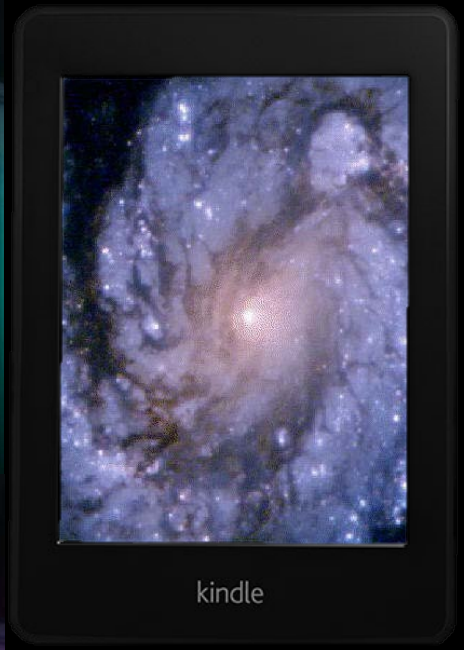
GALAH Meeting – 20 Jan 2014

The problem:
Data-Rich Astronomy



We all know that astrophysics 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 10.000 books per day…

ARE YOU READY FOR THIS???

AND THIS IS JUST ONE SURVEY!!!

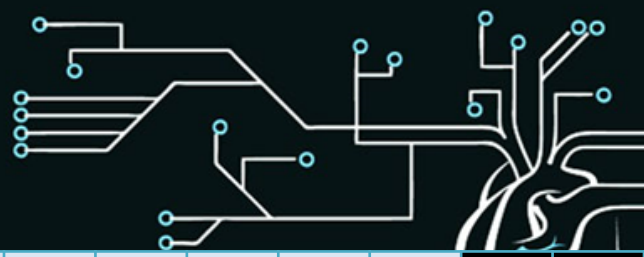
I've seen things you people wouldn't believe. Attack ships on fire off the shoulder of Orion. I've watched c-beams glitter in the dark near the Tannhäuser Gate. **All those ... *moments* will be lost in time, like tears...in rain.**

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 is the application of information technology to X disciplines, with emphasis on persistent data stores.
Astro-, Bio-, Chem-, 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.

Data Mining

One of the crucial part of Astroinformatics is **Data Mining**

Data Mining is the process concerned with automatically uncovering patterns, associations, anomalies, and statistically significant structures in large and/or complex data sets

Therefore it includes all those disciplines which can be used to uncover useful information in the data

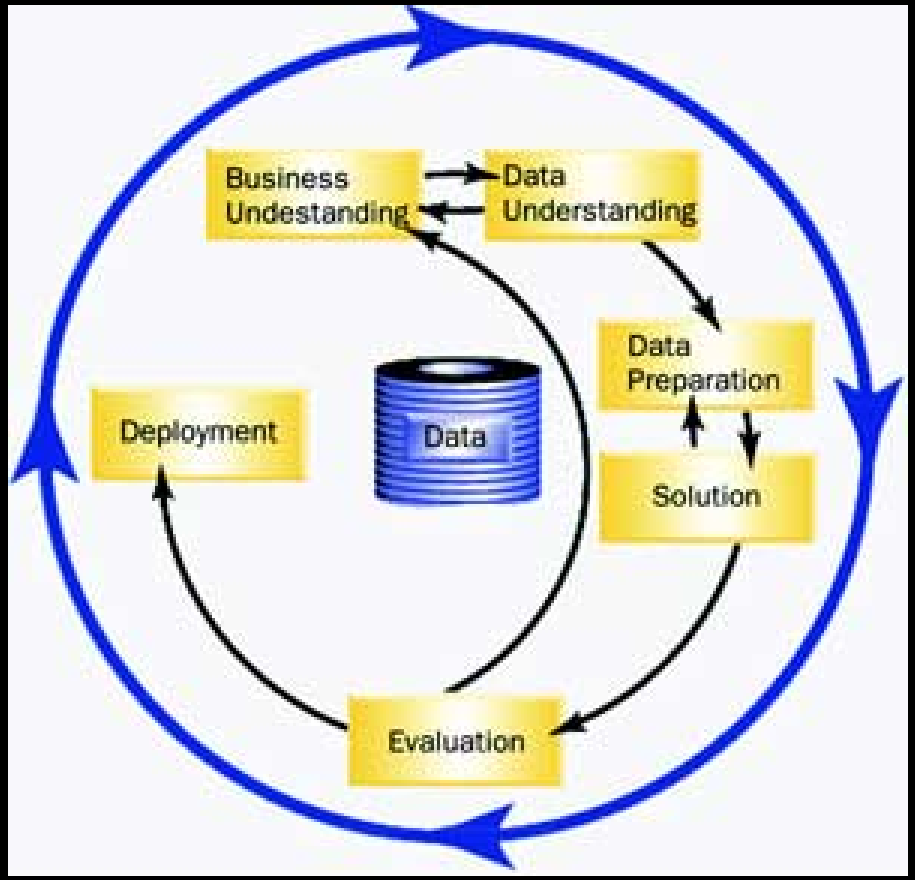
What is new is the confluence of the most mature offshoots of many disciplines with technological advances

As such, its contents are «user defined» and more than a new discipline is an ensemble of different methodologies originated in different fields

In particular in this talk we will focus on **Machine Learning**

Data Mining 11-step virtuous cycle

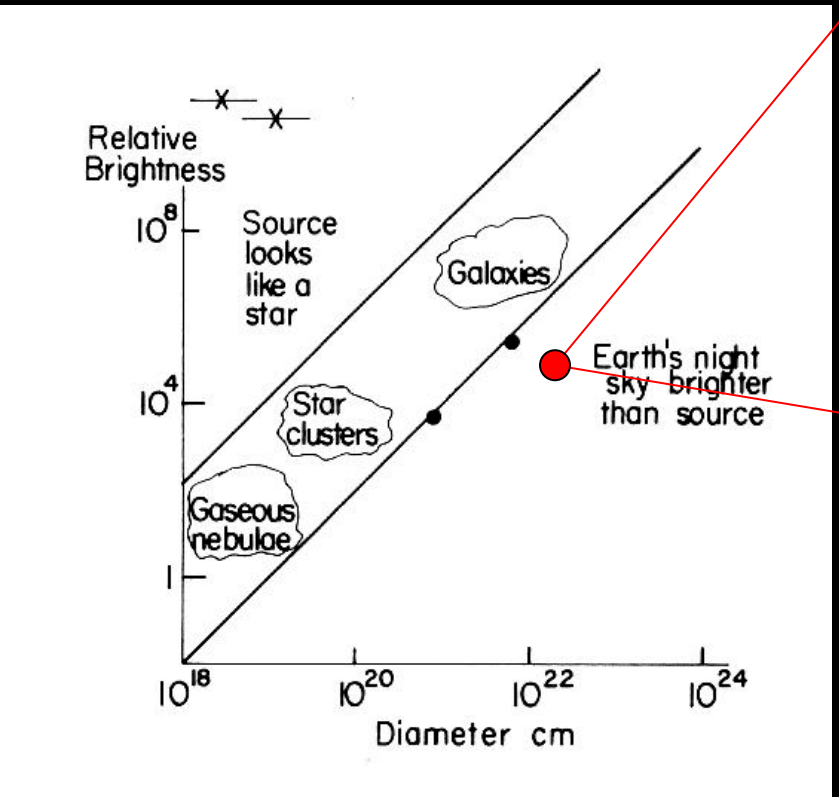
1. Translate any opportunity (problem) into DM opportunity (problem)
2. Select appropriate data
3. Get to know the data
4. Create a model set
5. Fix problems with the data
6. Transform data to bring information
7. Build/Evolve models
8. Assess models
9. Deploy models
10. Assess results
11. Go to 2

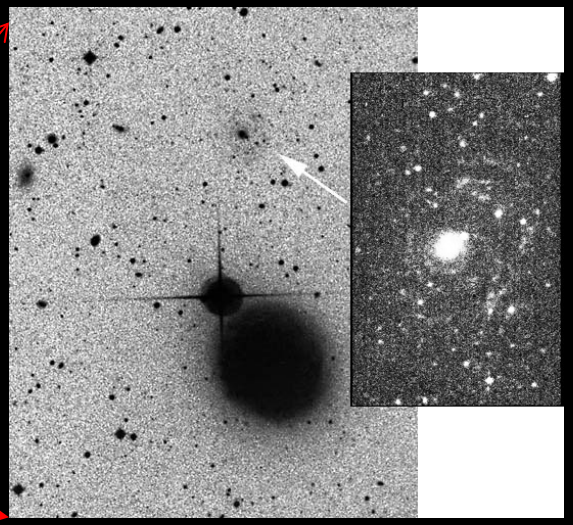


Parameter Space in Astrophysics

The astronomical parameter space is of high dimensionality, still sparsely covered and poorly sampled:

every time you improve either coverage or sampling you make new discoveries





Malin 1

a new type of low surface brightness galaxies (Malin, 1991)

Machine Learning

Machine learning: Field of study that gives computers the ability to learn without being explicitly programmed.

Arthur Samuel (1959).

February 24, 1956, Arthur Samuel's Checkers program, which was developed for play on the IBM 701, was demonstrated to the public on television.

In 1962, self-proclaimed checkers master Robert Nealey played the game on an IBM 7094 computer…

…the computer won.

May 11, 1997 – Deep Blue defeats Kasparov

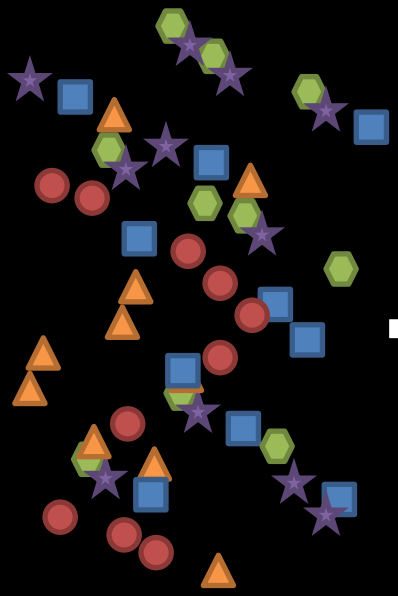
Machine Learning: Supervised

A Supervised Method tries to reproduce a bias, extending a preexisting knowledge on new patterns…

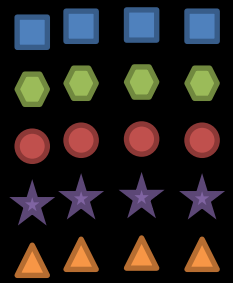
- Good for interpolation of data, bad for extrapolations;
- They need extensive bases of knowledge (i.e. uniformly sampling the parameter space) which are difficult to obtain;
- Errors are easy to evaluate;
- Relatively easy to use;
- They reproduce all biases and preconceived ideas present in the KB.

Supervised Methods are subdivided into: Classification and Regression algorithms

Machine Learning: Supervised



Knowledge Base

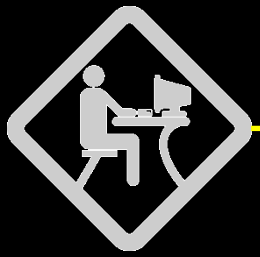


Train Set

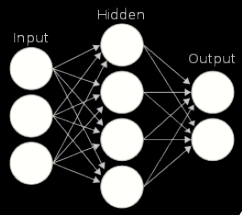
Train

Blind Test Set

Test



Analysis of results



Trained
Network

The World

New Knowledge

Machine Learning: Unsupervised

Unsupervised Methods (UM) are applied without any a priori knowledge…
They cluster the data relying on their statistical properties. The understanding only takes place through labeling (very limited Knowledge Base or KB).

"a blind man in a dark room - looking for a black cat - which isn't there"

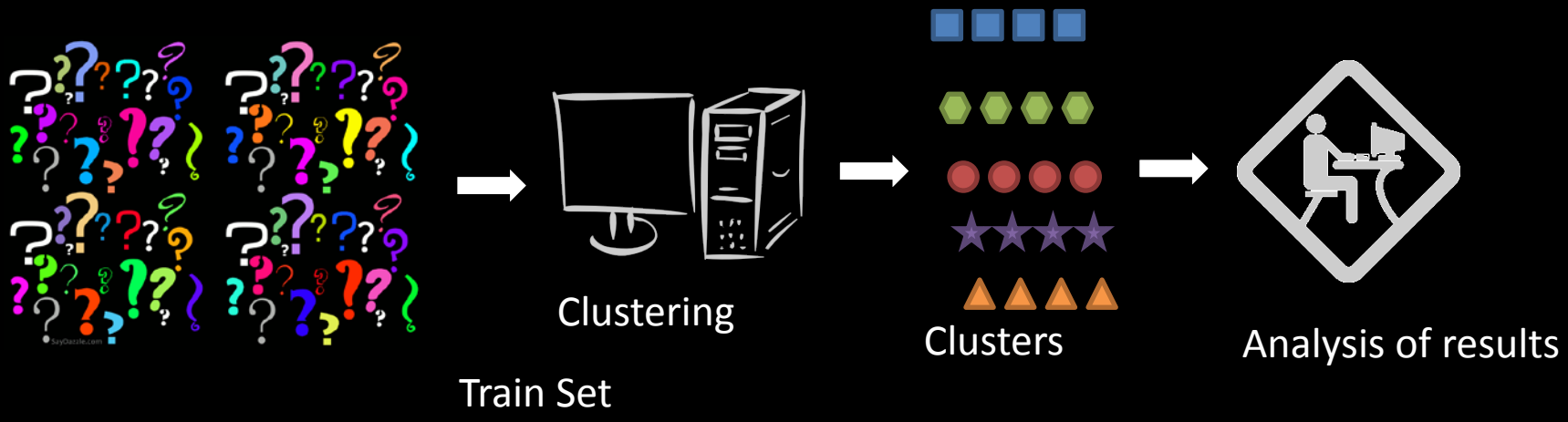
Charles Bowen

to be honest it is full of cats, the problem is to find the cat that interests us…

UM:

- need little or none a-priori knowledge;
- do not reproduce biases present in the KB;
- require more complex error evaluation (through complex statistics);
- are computationally intensive;
- are not user friendly *(… more an art than a science; i.e. lot of experience required)*

Machine Learning: Unsupervised



Machine Learning: Unsupervised

Train Set

Clustering

Clusters

Analysis of results

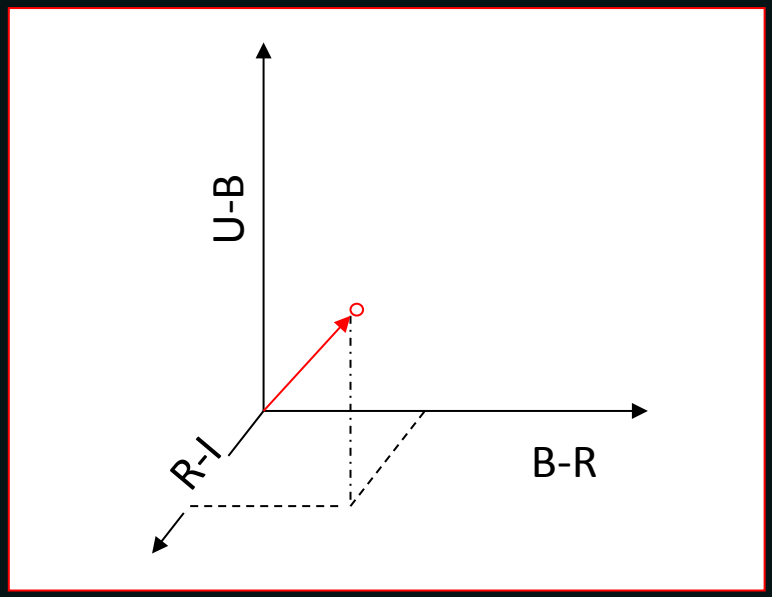
Analysis of results

Analysis of results

New Knowledge


REGRESSION PROBLEMS:
Photometric Redshift
Galaxy and Quasars

PHOTOMETRIC REDSHIFTS AS AN INVERSE PROBLEM



Why do we need (photometric or spectroscopic) redshifts?

- 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!!!

Photometric Redshifts: Methods

Template based:

color-space tessellation, χ2-minimization, maximum likelihood, Bayesian...

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

Learning based:

Nearest Neighbor, 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:

$$\mathbf{X} \equiv \{x_1, x_2, x_3, \ldots x_N\} \quad \textbf{input vectors}$$
$$\mathbf{Y} \equiv \{x_1, x_2, x_3, \ldots x_M\} \quad \textbf{target vectors} \quad M << N$$
$$\textbf{find} \quad \hat{f}: \ \hat{\mathbf{Y}} = \hat{f}(\mathbf{X}) \quad \textbf{is a good approximation of } \mathbf{Y}$$

KB = Knowledge Base

KB(from VO)
(set of templates)

→

Mapping function

→

Knowledge (photo-z)

Statistical evaluation

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

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

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

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

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

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

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

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

Average statistical indicators such as bias and standard deviation, however, provide only part of the information which allows to correctly evaluate the performances of a method and, for instance, they provide only very little evidence of the systematic trends which are observed as a sudden increase in the residuals spread over specific regions of the redshift space

Statistical evaluation

$$\mathrm{bias(x)} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\sigma(x) = \sqrt{\frac{\sum_{i=1}^{N}\left[x_i - \left(\frac{\sum_{i=1}^{N} x_i}{N}\right)\right]^2}{N}}$$

$$\mathrm{MAD(x)} = Median\left(|x|\right)$$

$$\mathrm{NMAD(x)} = 1.48 \times Median\left(|x|\right)$$

Where x can be:

$$\Delta z = (z_{spec} - z_{phot})$$

or

$$\Delta z_{norm} = (z_{spec} - z_{phot})/(1 + z_{spec})$$

Average statistical indicators such as bias and standard deviation, provide part of the information which allows to correctly evaluate the performances of a method and, for instance, they provide only very little evidence of the systematic trends which are observed as a sudden increase in the residuals spread over specific regions of the redshift space.

The rest of important information could be retrieved from the analysis of outliers (in particular the catastrophic ones).

Catastrophic outliers

$$|\Delta z_{norm}| > 2\sigma\left(\Delta z_{norm}\right)$$

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 **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 to 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]

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

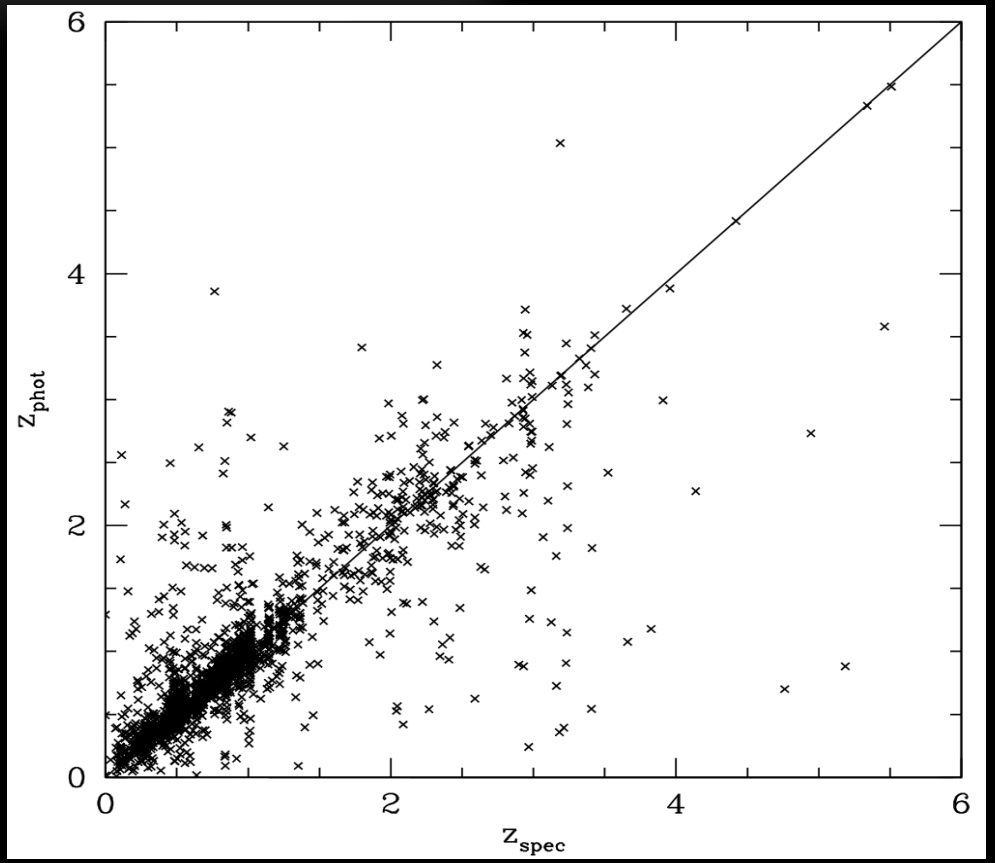
PHAT1 CONTEST

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

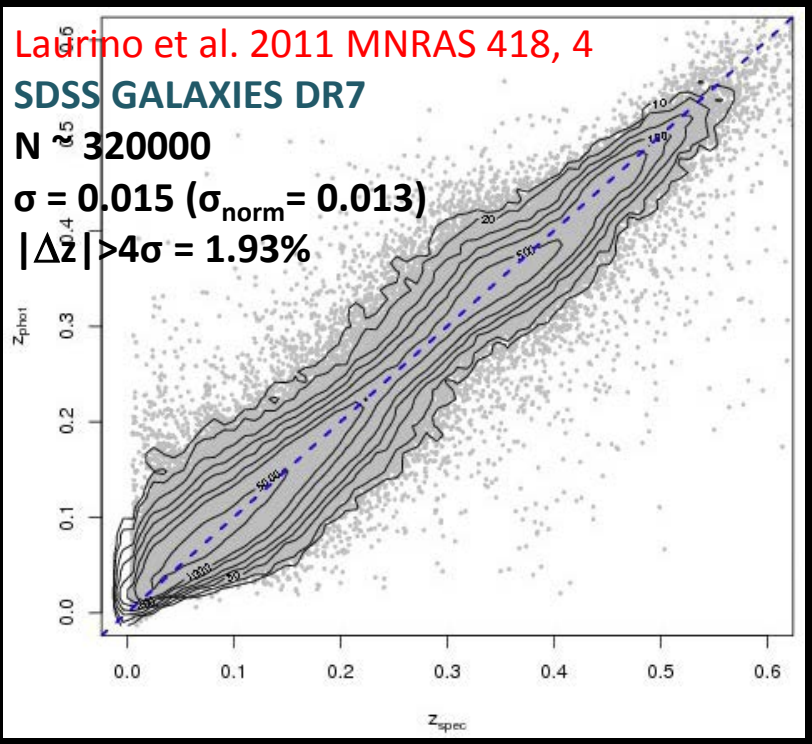
σ_{norm}= 0.05

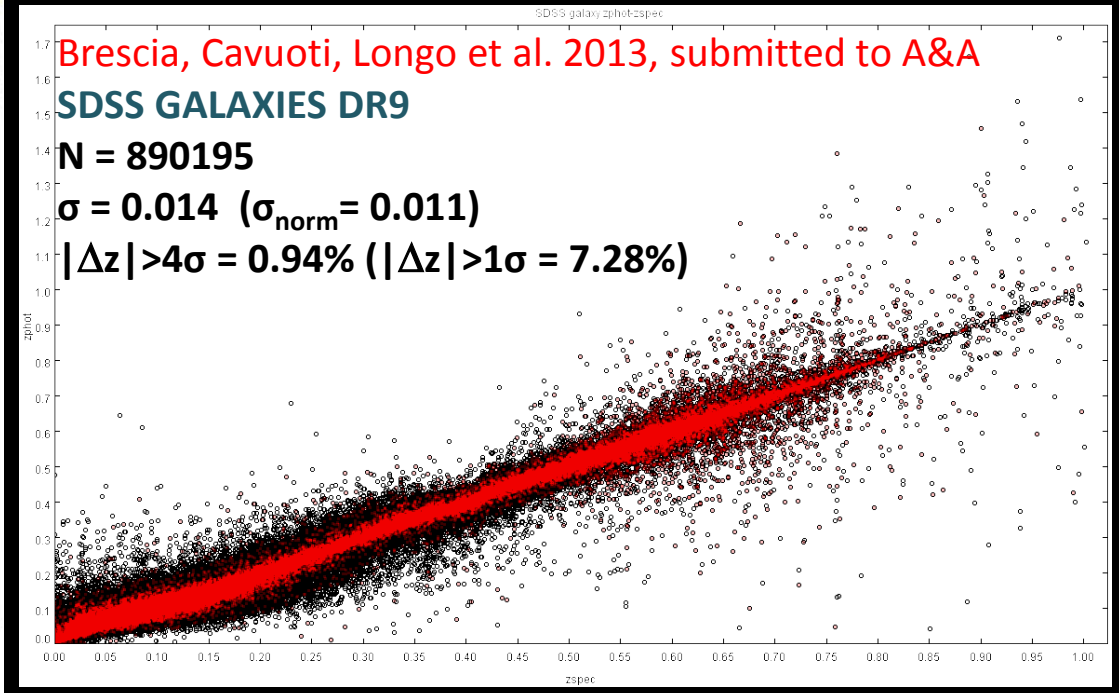
$|\Delta z| > 0.15$ = 16.33%

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

Photo-z for SDSS DR9 Galaxies





An application of a machine learning based method to the estimation of photometric redshifts for the galaxies in the SDSS Data Release 9 (SDSS-DR9). Photometric redshifts for more than 129 million galaxies were produced and made available at the URL: http://dame.dsf.unina.it/catalog/DR9PHOTOZ/

The obtained redshifts have a normalized standard deviation σ_{norm} = 0.011, which decreases to 0.009 after the rejection of catastrophic outliers. This result is better or comparable with what was already available in the literature but present a smaller number of catastrophic outliers

Photo-z for QSO

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

Having the data, three new questions arise…

- Which magnitudes are the best for this work?
- Is it better to use magnitudes 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 ISO mag
- UKIDSS HALL mag
- WISE ISO mag

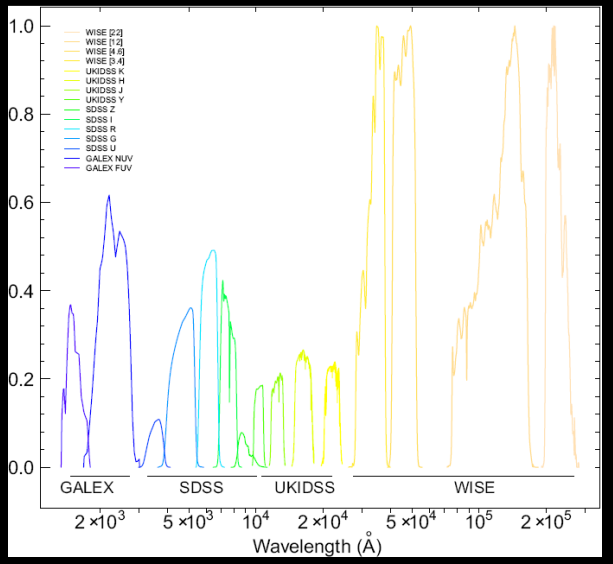
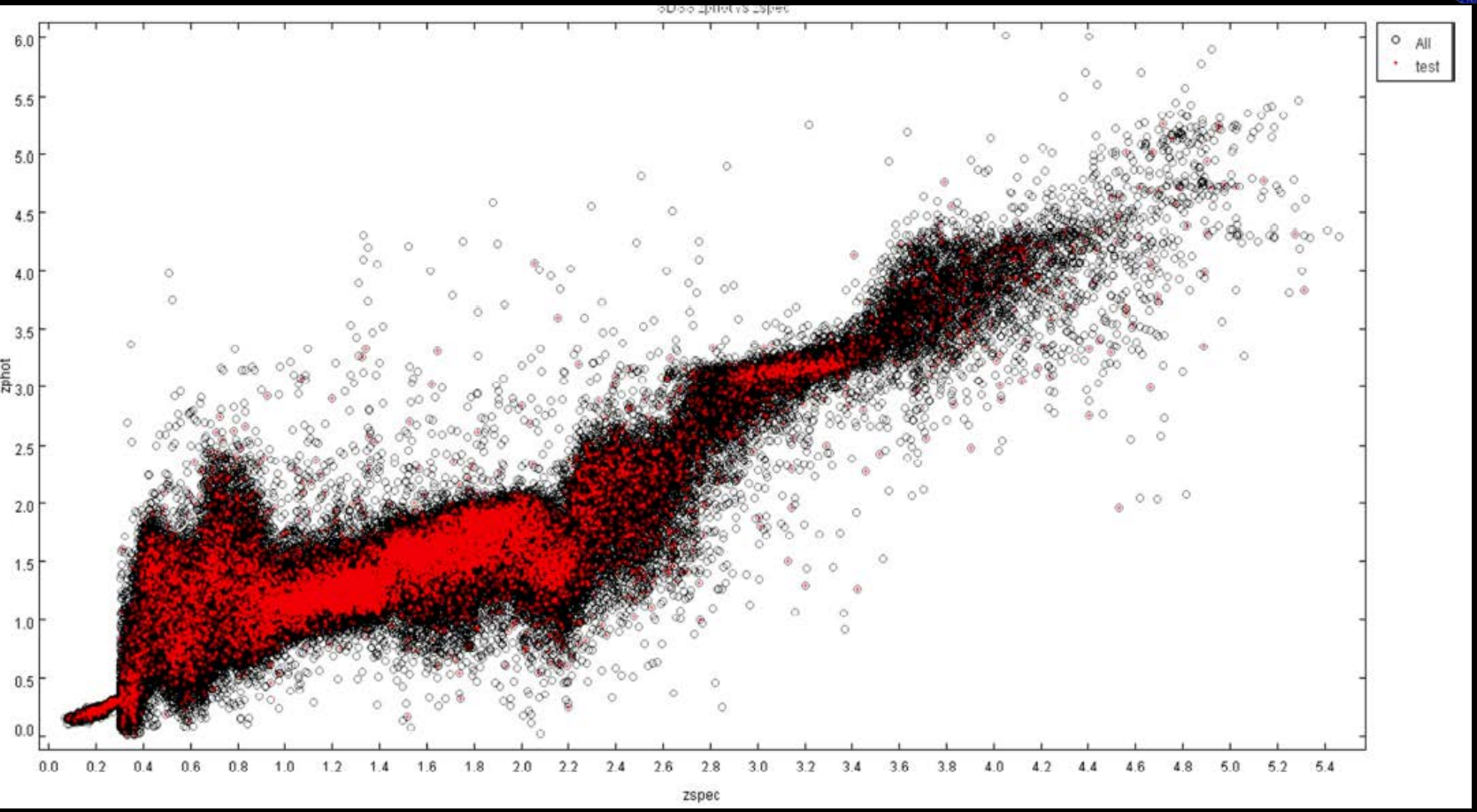


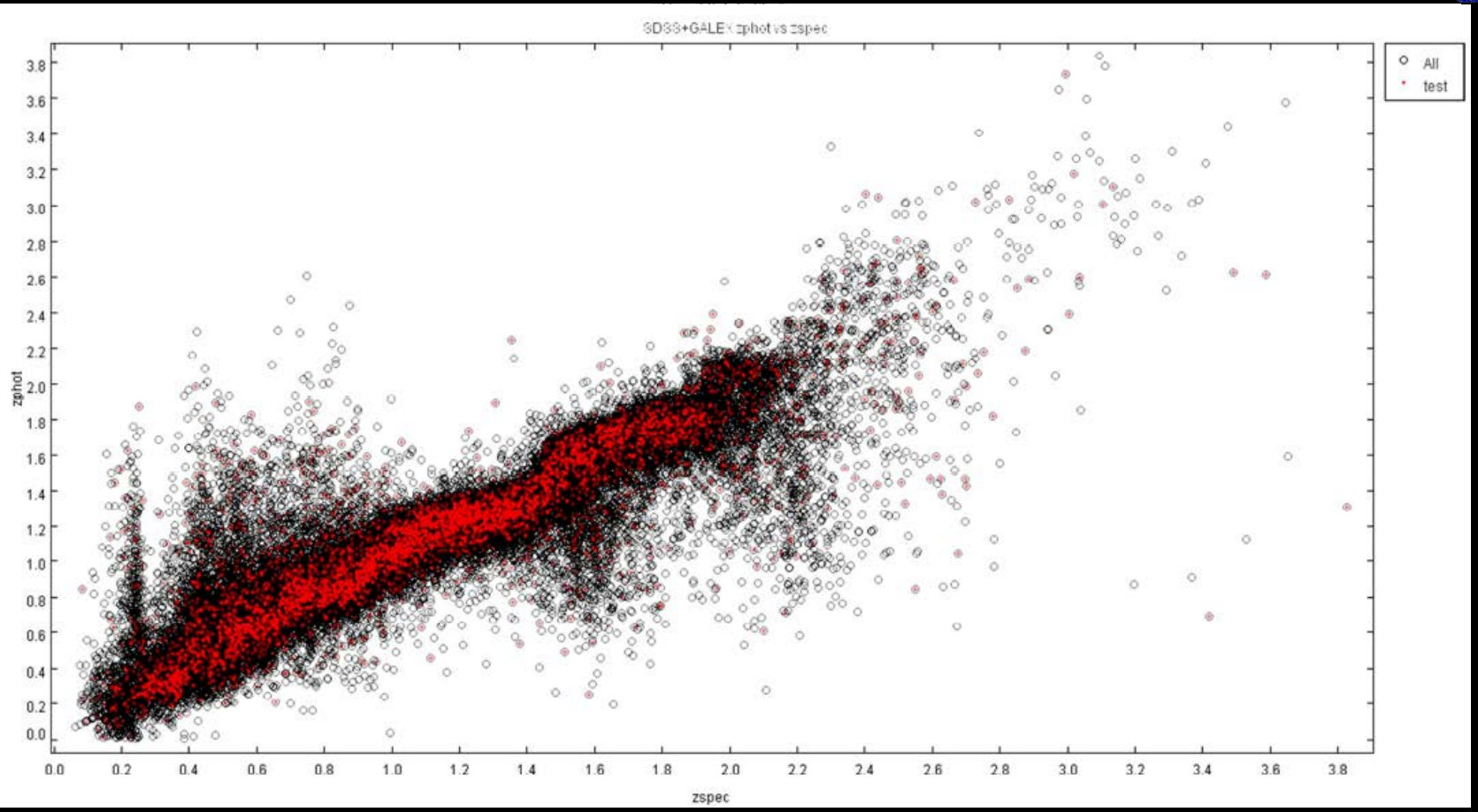
Photo-z for QSO: SDSS



Ref.	|bias|	sigma	MAD	RMS	$\|bias_{norm}\|$	s_{norm}	MAD_{norm}	RMS_{norm}
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

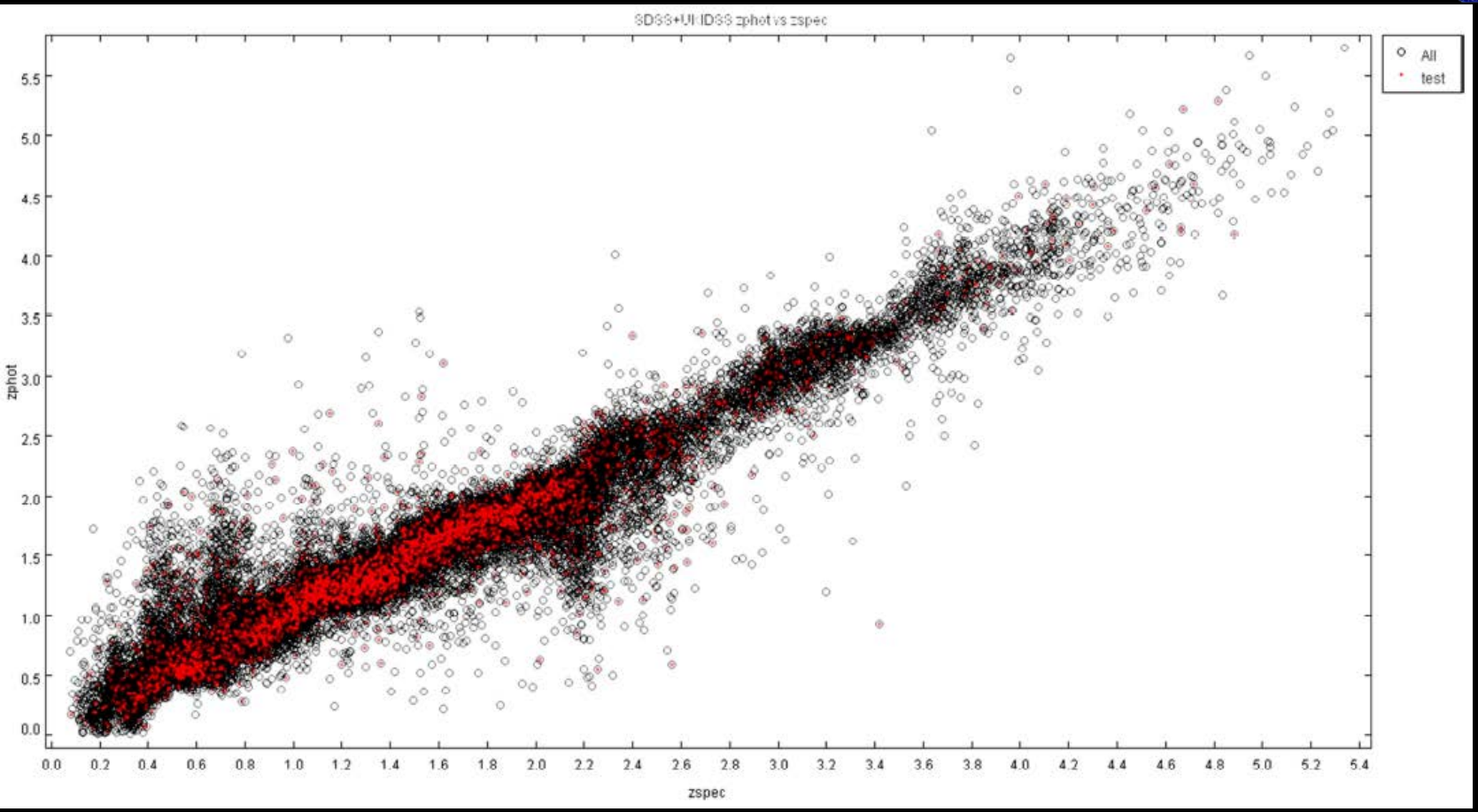
Photo-z for QSO: SDSS + GALEX



Ref.	|bias|	sigma	MAD	RMS	$\|bias_{norm}\|$	s_{norm}	MAD_{norm}	RMS_{norm}
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

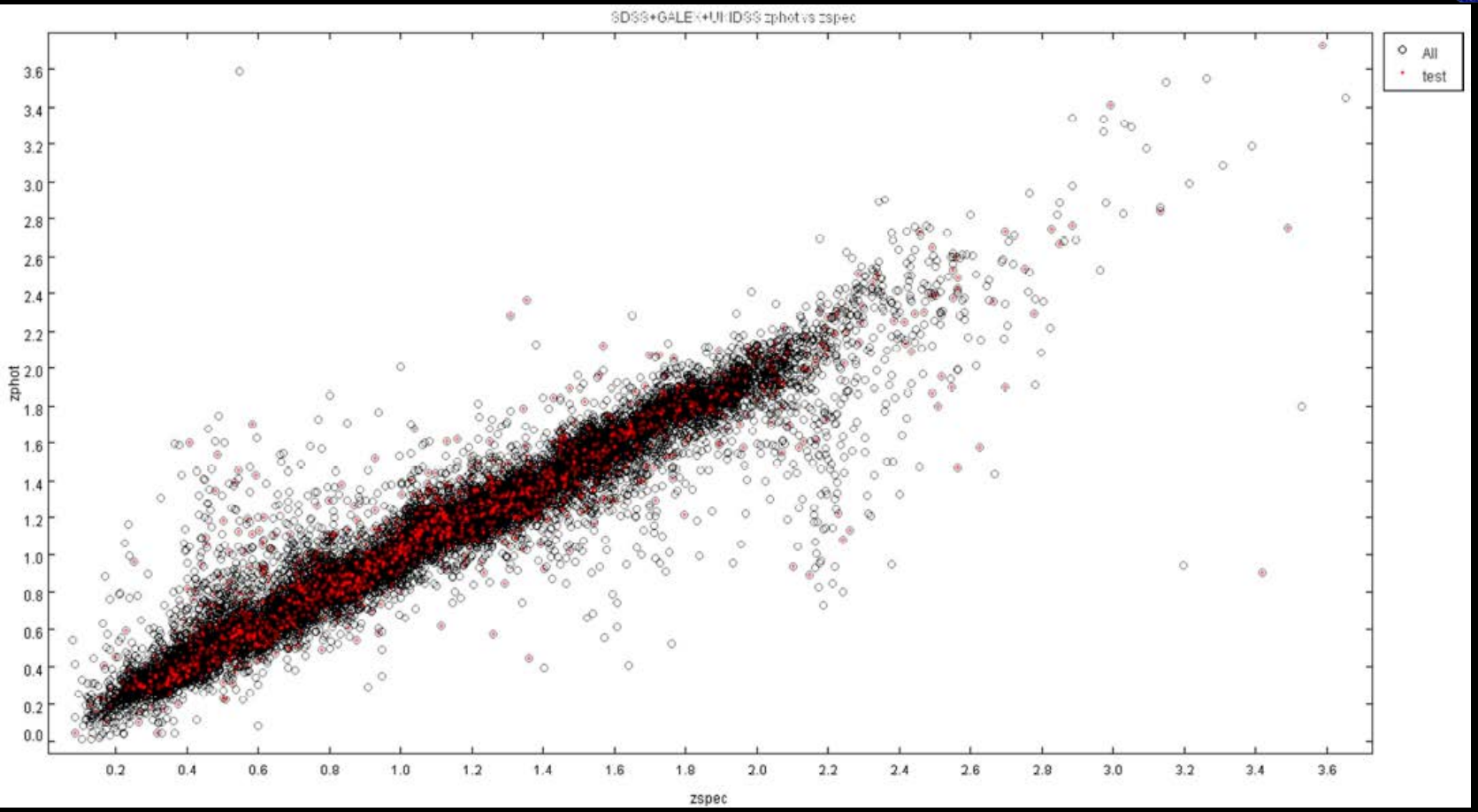
Photo-z for QSO: SDSS + UKIDSS



Ref.	|bias|	sigma	MAD	RMS	$\|bias_{norm}\|$	s_{norm}	MAD_{norm}	RMS_{norm}
MLPQNA	0.003	0.21	0.084	0.21	0.010	0.11	0.040	0.11
Bovy 2012		0.28						

31094 objects

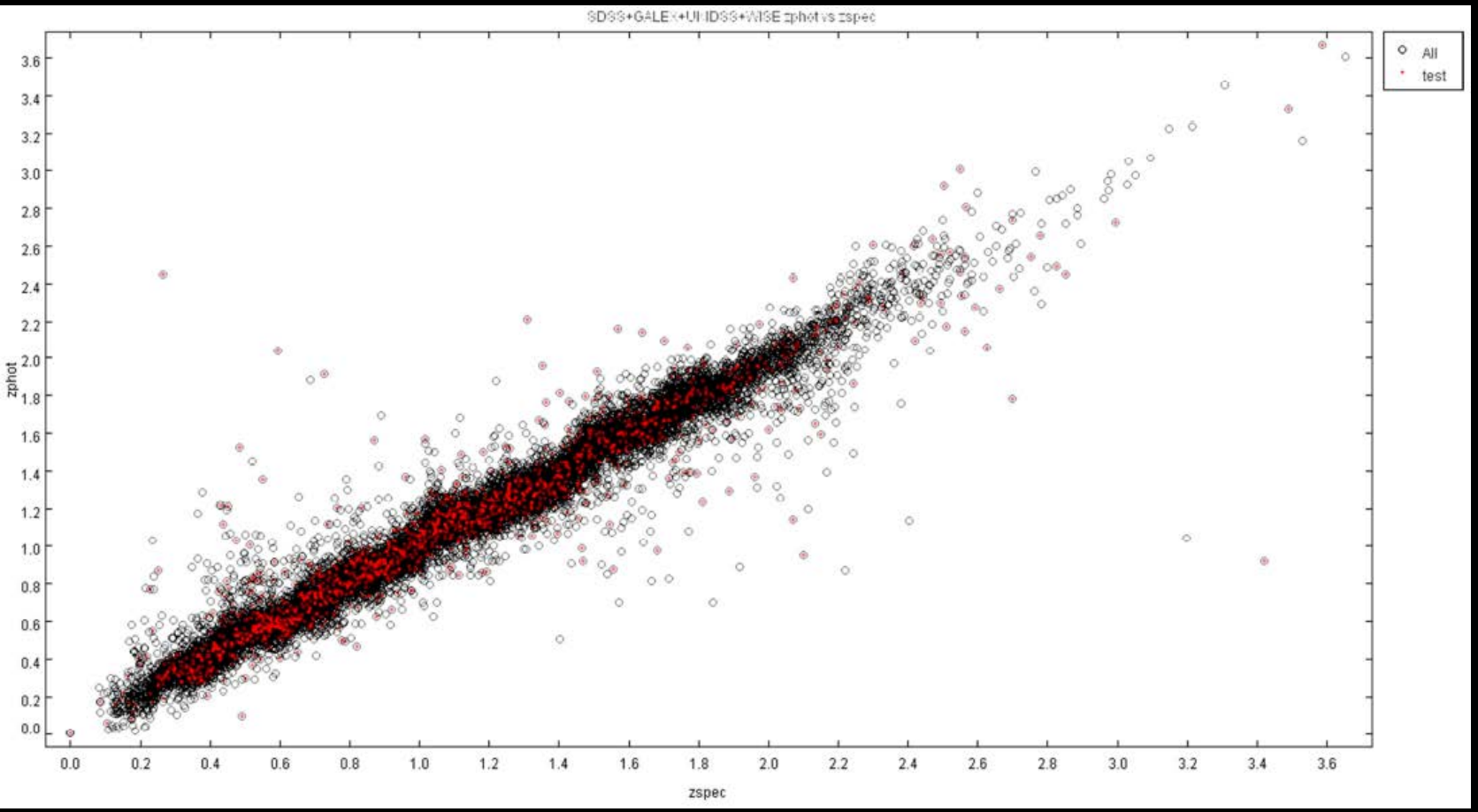
Photo-z for QSO: SDSS + UKIDSS + GALEX



Ref.	|bias|	sigma	MAD	RMS	$\|bias_{norm}\|$	s_{norm}	MAD_{norm}	RMS_{norm}
MLPQNA	0.005	0.15	0.072	0.15	0.006	0.075	0.036	0.075
Bovy 2012		0.21						

14588 objects

Photo-z for QSO: 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

Photo-z for QSO: overall 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	Outliers ($\lvert\Delta z\rvert$)		Outliers ($\lvert\Delta z_{norm}\rvert$)	
	$> 2\sigma(\Delta z)$	$> 4\sigma(\Delta z)$	$> 2\sigma(\Delta z_{norm})$	$> 4\sigma(\Delta z_{norm})$
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

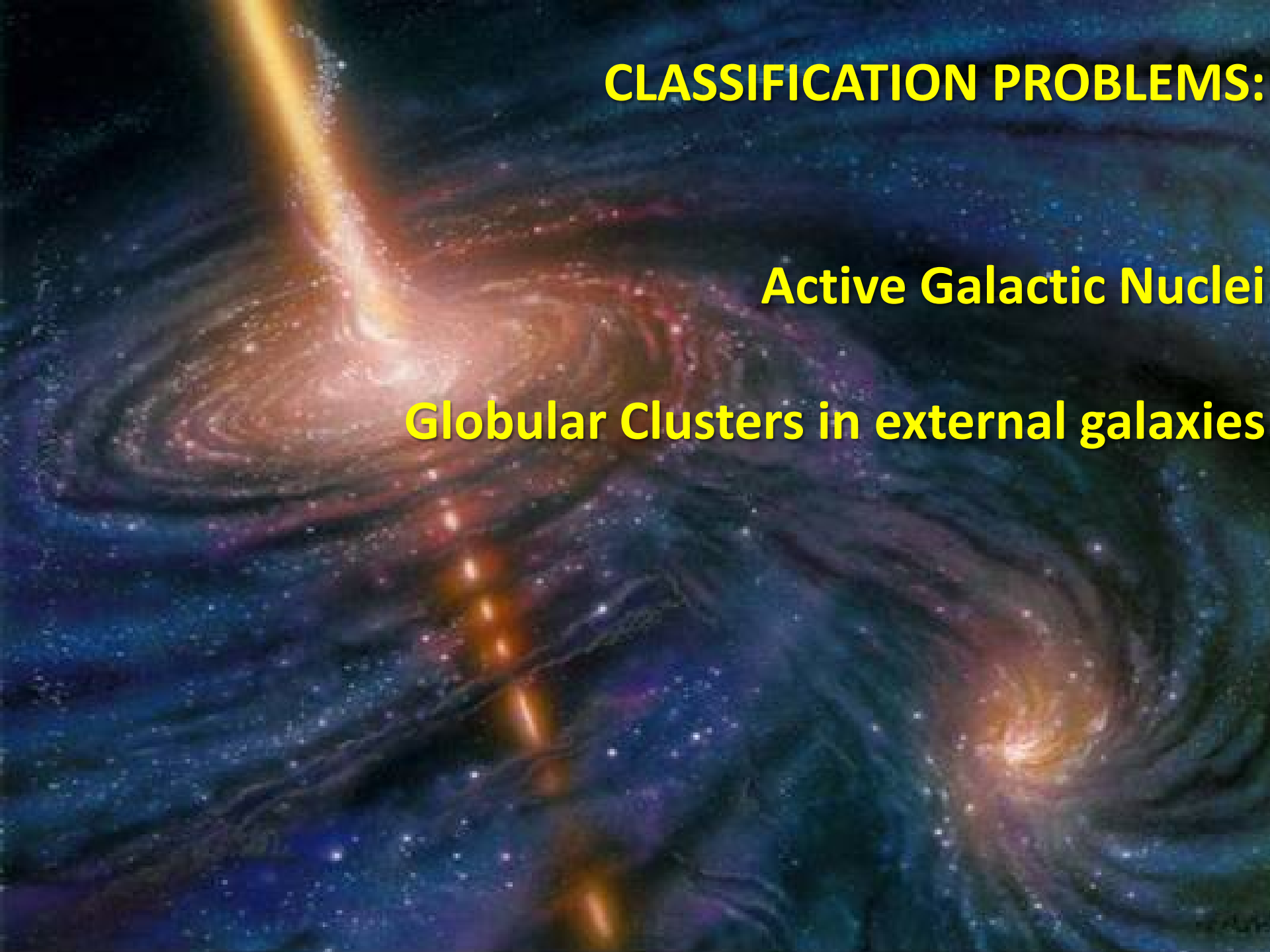
Exp	$BIAS(\Delta z)$	$\sigma(\Delta z)$	$MAD(\Delta z)$	$RMS(\Delta z)$
SDSS				
MLPQNA	0.007	0.25	0.102	0.26
Bovy et al.	-	0.46	-	-
Laurino et al.	0.210	0.28	0.110	0.35
Ball et al.	-	0.35	-	-
Richards et al.	-	0.52	-	-
SDSS + GALEX				
MLPQNA	0.003	0.21	0.060	0.22
Bovy et al.	-	0.26	-	-
Laurino et al.	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

Photo-z for QSO: conclusions

What we learned:

- Additional bands are more important than additional points in the training sets;
- Wing degeneracies fade out with wavelength coverage;
- Photometric redshifts 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, 772, 2, 140***

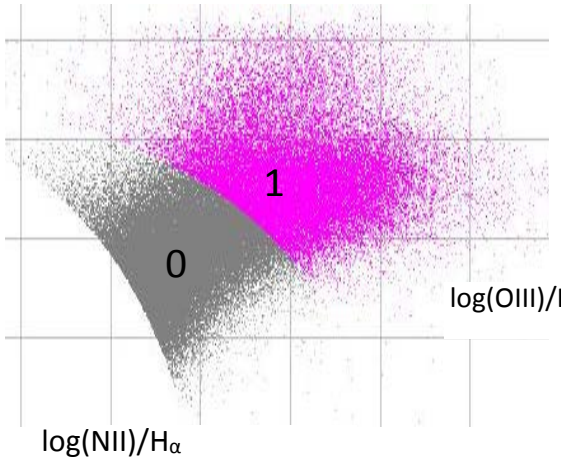

CLASSIFICATION PROBLEMS:
Active Galactic Nuclei
Globular Clusters in external galaxies

AGN CLASSIFICATION

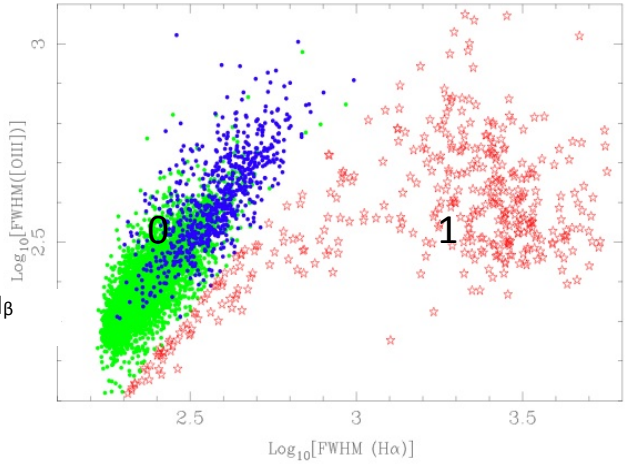
Photometric parameters used for training of the NNs and SVMs:

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

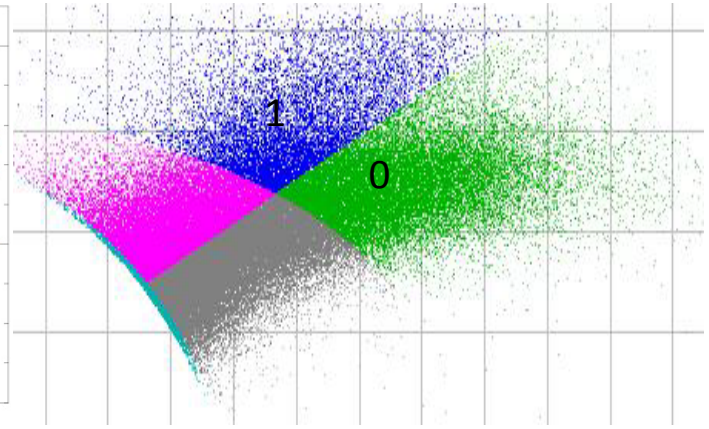
1° Experiment:
AGN -> 1, Mixed-> 0



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



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



Cavuoti, S.; Brescia, M.; D'Abrusco, R.; Longo, G.; Paolillo, M.; 2014, Photometric classification of emission line galaxies with Machine Learning methods, ***MNRAS, 437, 1, 968-975***

AGN CLASSIFICATION RESULTS

Sample	Parameters	KB	Algorithm	e_{tot}
Experiment (1) AGN detection	SDSS photometric parameters + photo redshift	**BPT plot +Kewley's line**	*SVM* *MLP*	~74% ~76%
Experiment (2) Type 1 vs. Type 2	SDSS photometric parameters + photo redshift	**Catalogue of Sorrentino et al.+ Kewley's line**	*SVM* *MLP*	~82% e~72%
Experiment (3) Seyfert Vs. LINERs	SDSS photometric parameters + photo redshift	**BPT plot+Heckman's+Kewley's lines**	*SVM* *MLP*	~78% ~74%

- Checking the trained NN with a dataset of sure not AGN, just 12.6% are false positive
- False positive surely not AGN (according KB) are 0.89%
- **ONLINE CATALOG AVAILABLE AT** http://vizier.cfa.harvard.edu/viz-bin/VizieR?-source=J/MNRAS/437/968

Globular Cluster Recognition

NGC1399 Dataset

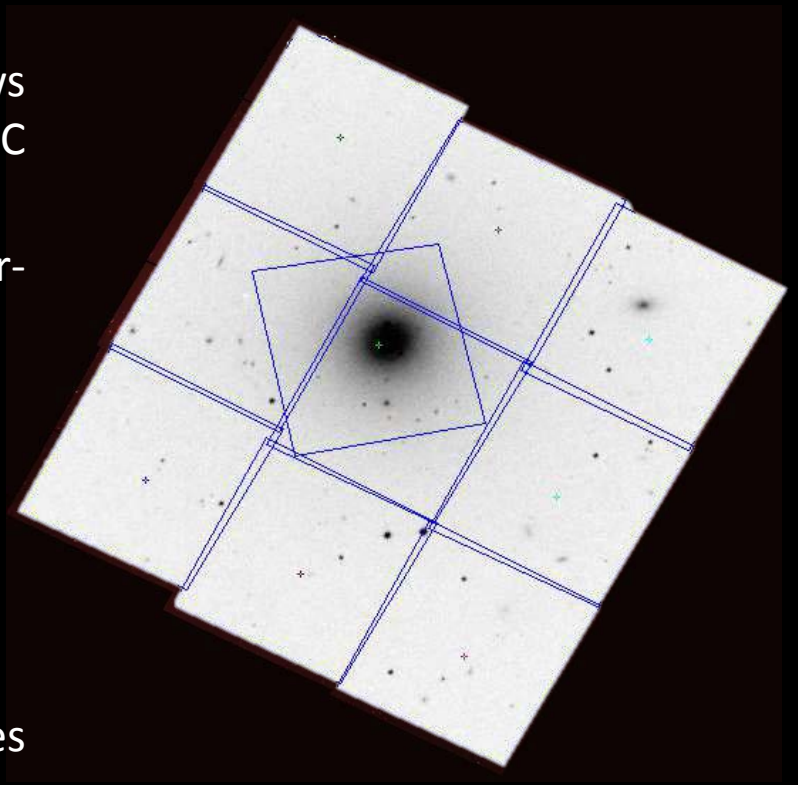
NGC1399 (~20 Mpc) is an ideal target because it allows to probe a large fraction of the galaxy and still resolve GC sizes.

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, 421, 2, 1155-1165***

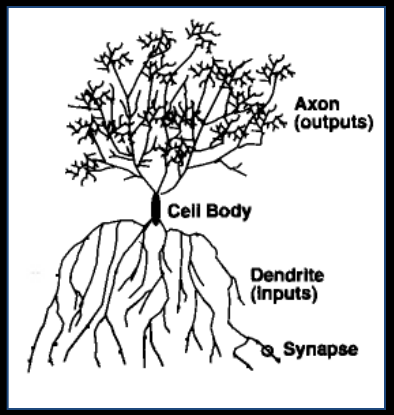
Quality and pruning results

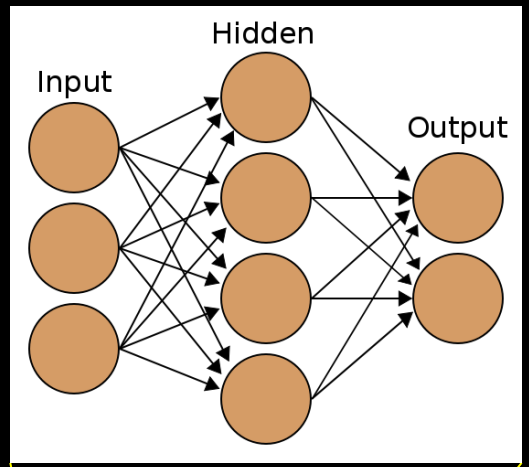
Type of experiment	Missing features	Figure of merit	MLPQNA	GAME	SVM	MLPBP	MLPGA
Complete patterns	–						
		class.accuracy	**98.3**	82.1	90.5	59.9	66.2
		completeness	**97.8**	73.3	89.1	54.1	61.4
		contamination	**1.8**	18.7	7.7	42.2	35.1
No par. 11	11						
		class.accuracy	**98.0**	81.9	90.5	59.0	62.4
		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**;

Multi Layer Perceptron

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

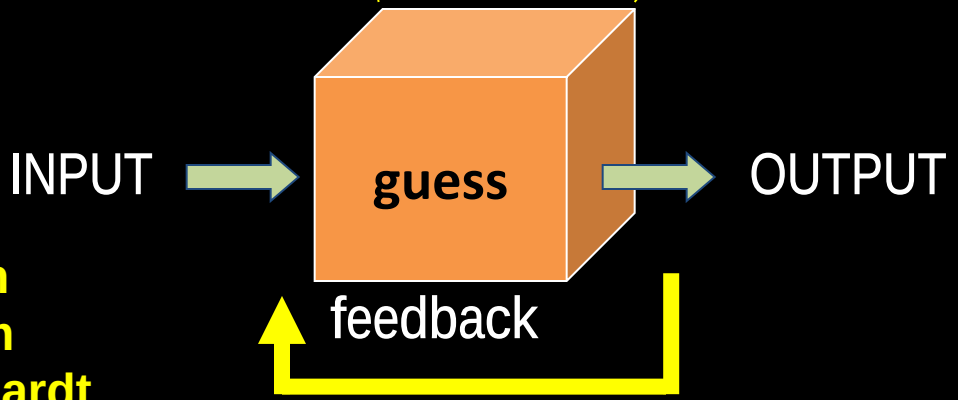




Neurons are connected by «activation functions» we have different kind of MLP changing the way with they found the best solution

Training rules:

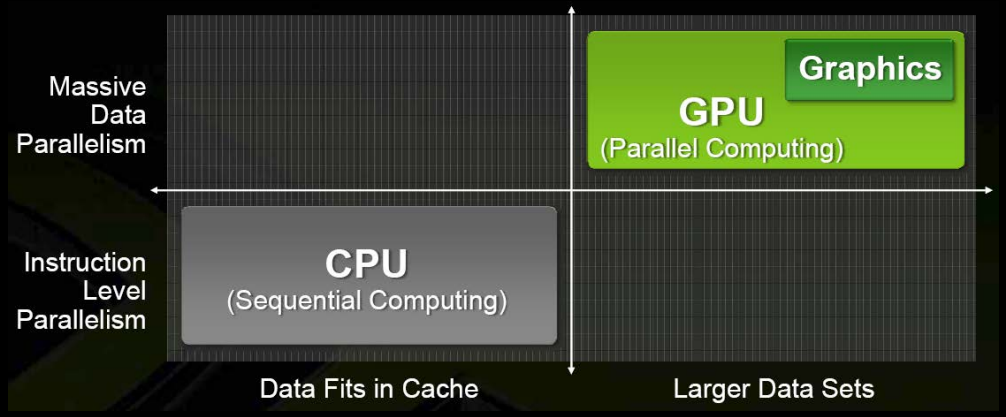
- **Quasi Newton**
- **Back Propagation**
- **Genetic Algorithm**
- **Levenberg Marquardt**

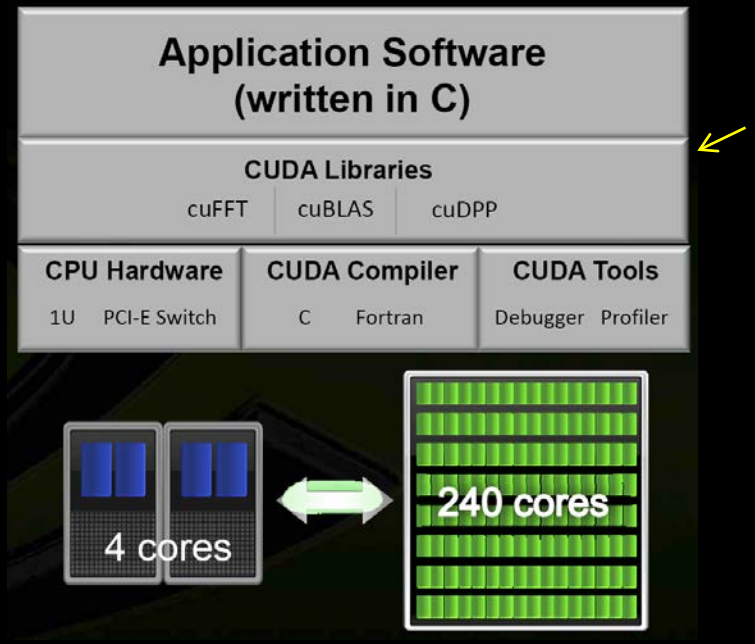


… GPU technology?

The Graphical Processing Unit is specialized in highly parallel computation (exactly what graphics rendering is about). So, more transistors can be devoted to data processing rather than data caching and flow control.

«GPU have evolved to the point where many real world apps are easily implemented on them and run significantly faster than on multi-core systems. Future computing architectures will be hybrid systems with parallel-core GPUs working in tandem with multi-core CPUs» Jack Dongarra, Director of the Innovative Computing Laboratory The University of Tennessee





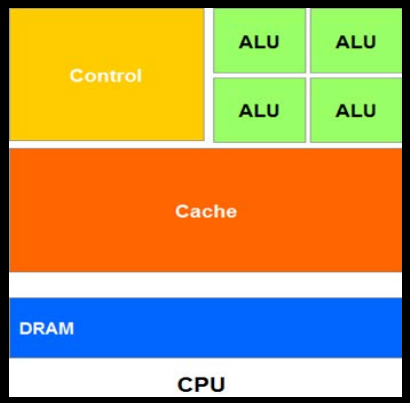
DAME - FMLPGA
Fast Multi Layer Perceptron Genetic Algorithm

FMLPGA is a Soft Computing model developed in order to solve supervised regression or classification problems, scalable for Massive Data Sets (MDS).

It is embarrassingly parallel.

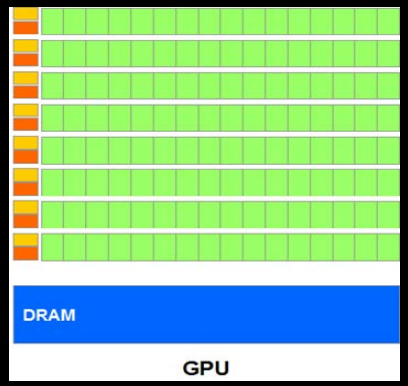
GPU vs CPU

Multi-core CPU

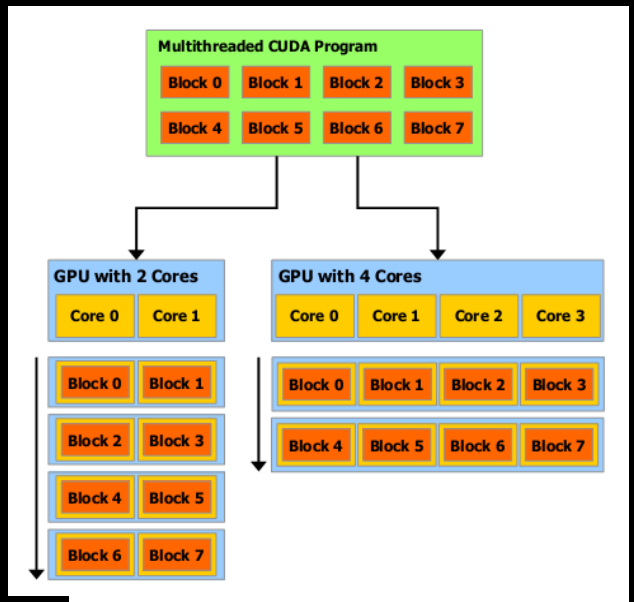


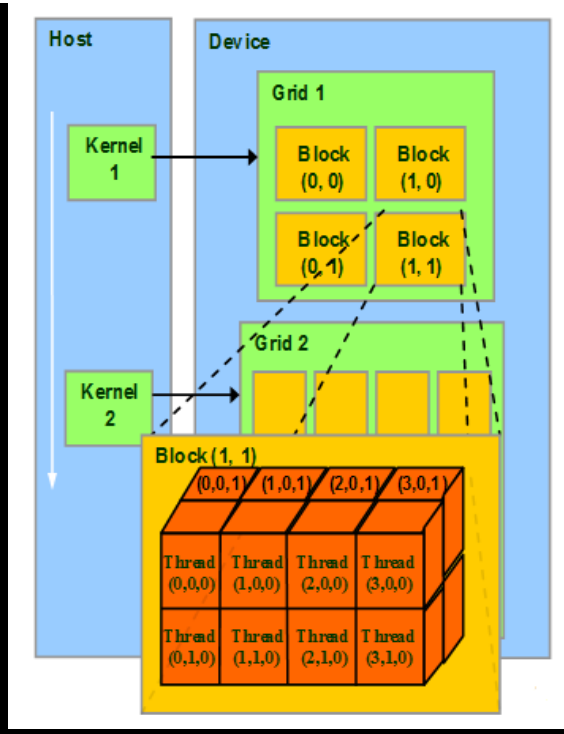
- Composed by few cores, designed to maximize the sequential code efficiency;
- Large cache memory to reduce latency time to access data and/or complex instruction execution;
- Sophisticate control logic to handle instruction flow (pipelining and multi-threading).

Many-core GPU



- Composed by many cores (hundreds), designed to execute parallel code;
- Memory structures with negligible access time to perform contemporary simple instructions;
- Simple control logic (the only bottleneck could be the communication with the CPU host);

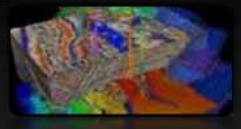




CUDA cost-benefit ratio

"The PGI compiler is now showing us just how powerful it is. On the software we are writing, it's many times faster on the NVIDIA card. We are very pleased and excited about the future uses. It's like owning a personal supercomputer."

Dr. Kerry Black University of Melbourne



Large Oil Company

3x in 7 days

Solving billions of equations iteratively for oil production at world's largest petroleum reservoirs

UNIVERSITY of HOUSTON
CULLEN COLLEGE of ENGINEERING

Univ. of Houston

Prof. M.A. Kayali

20x in 2 days

Studying magnetic systems for innovations in magnetic storage media and memory, field sensors, and biomagnetism

THE UNIVERSITY OF MELBOURNE

Uni. Of Melbourne

Prof. Kerry Black

65x in 2 days

Better understand complex reasons by lifecycles of snapper fish in Port Phillip Bay

УГАТУ

Ufa State Aviation

Prof. Arthur Yuldashev

7x in 4 Weeks

Generating stochastic geological models of oilfield reservoirs with borehole data

GAMESS-UK

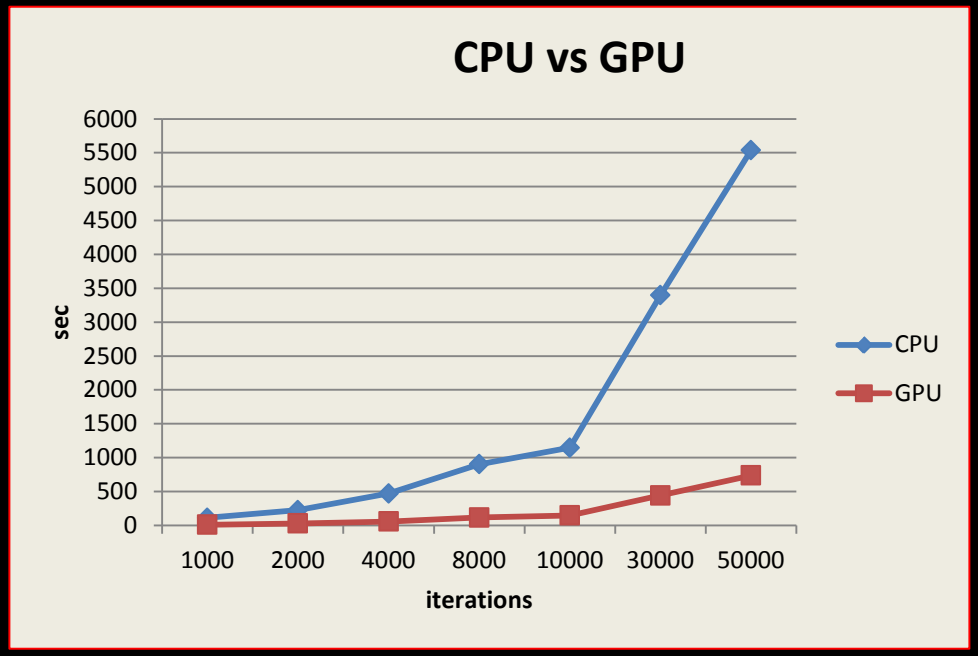
Dr. Wilkinson, Prof. Naidoo

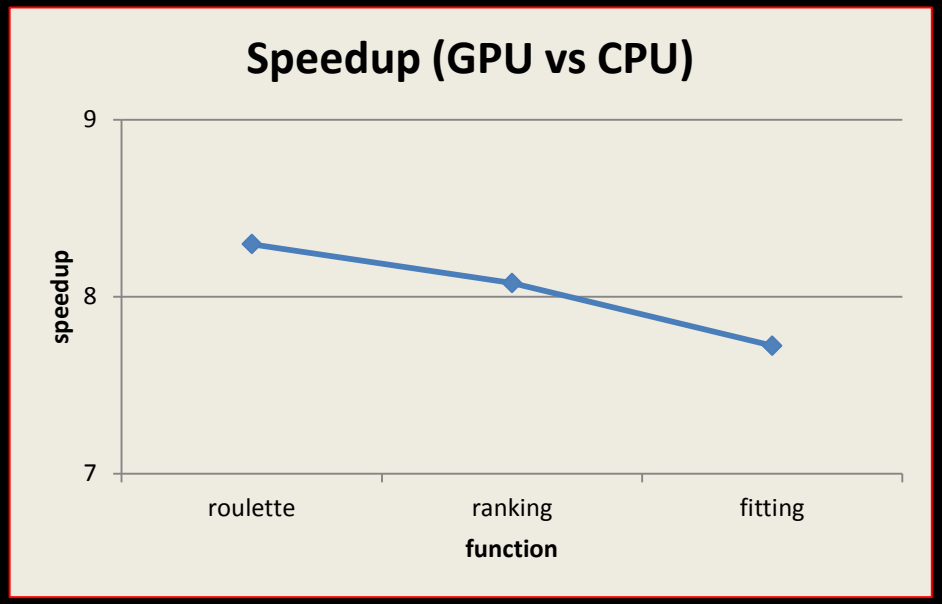
10x

Used for various fields such as investigating biofuel production and molecular sensors.

CUDA – Our Experience

With our first test (FMLPGA, available on DAMEWARE) we obtain a speedup of **8x** during a bachelor thesis work (**1 month**)





Our Tools 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 services:
- text mining (VOGCLUSTERS)
- Transient classification (STraDiWA)
- EUCLID Mission Data Quality

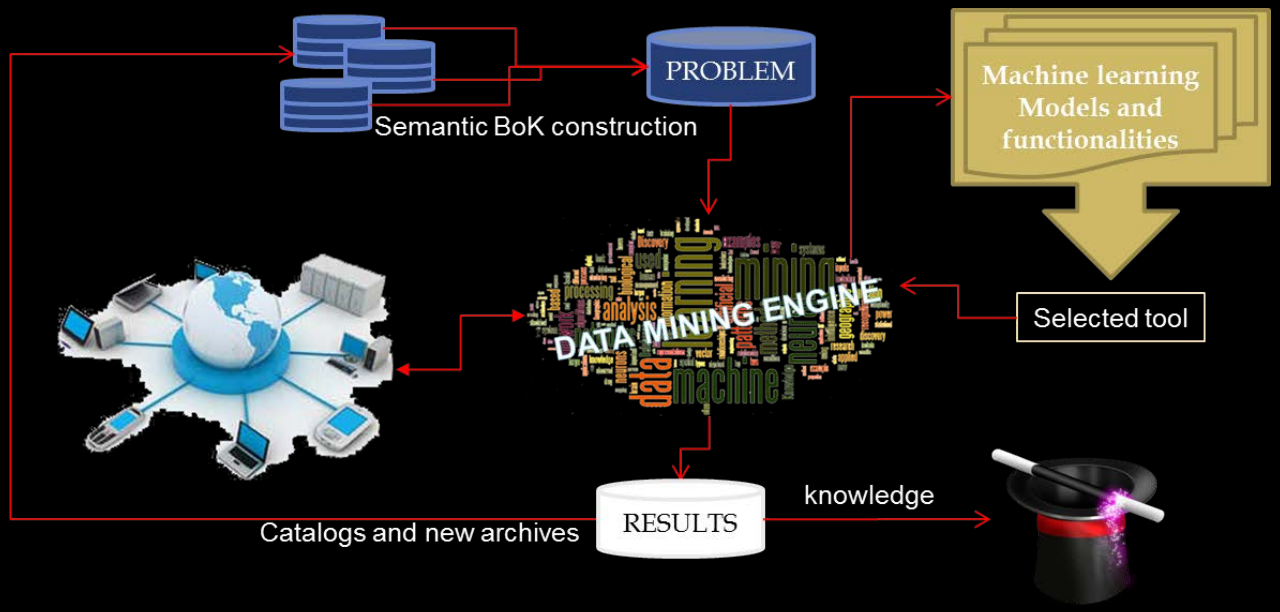
Other Services:
- CLASH-VLT Data Archive
- PhotoRaptor
- GPU-based models

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

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

DAMEWARE

Inspired by human brain features: high-parallel data flow, generalization, robustness, self-organization, pruning, associative memory, incremental learning, genetic evolution.



Multi Layer Perceptron
trained by:
- Back Propagation
- Quasi Newton
- Genetic Algorithm
- Levenberg Marquardt

Support Vector Machines

Genetic Algorithms

Evolving SOM

Self Organizing Maps

K-Means

Multi-layer Clustering

Principal Probabilistic Surface

Bayesian Networks

Random Decision Forest

Discrete Wavelet Transform

next ...

Classification

Regression

Clustering

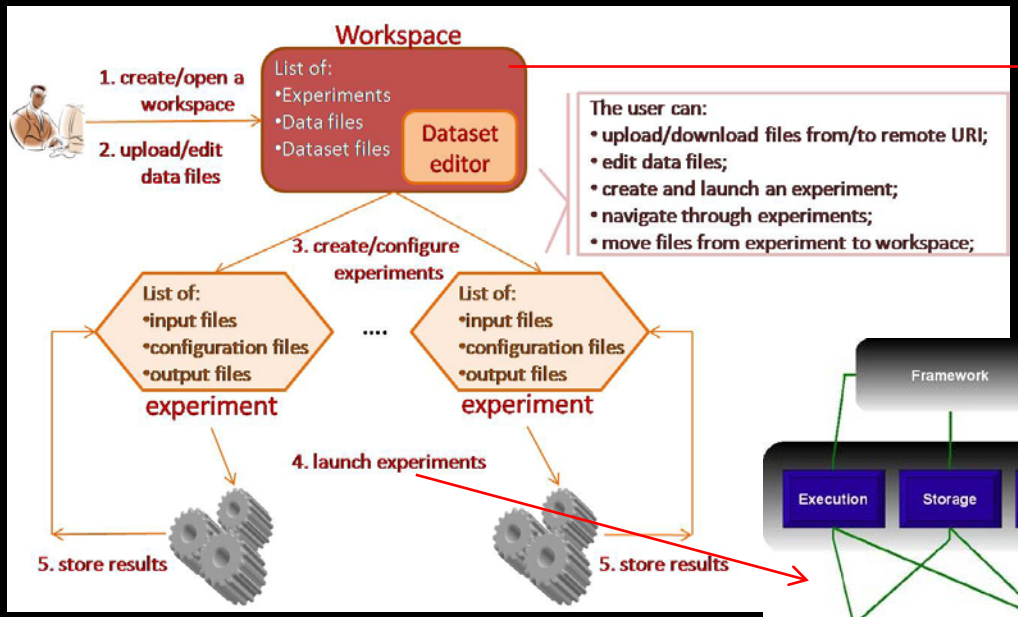
Feature Extraction

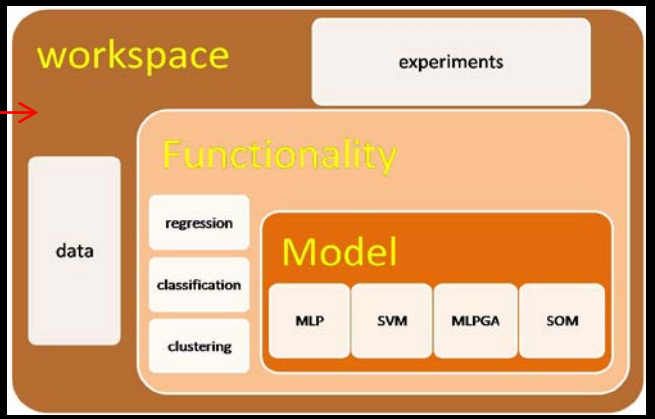
DAMEWARE

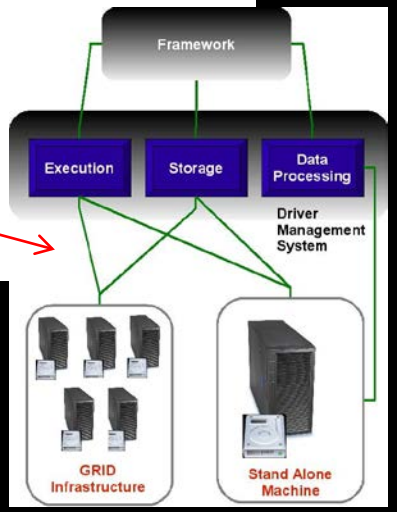
Based on the X-Informatics paradigm, it is multi-disciplinary platform (until now X = Astro)

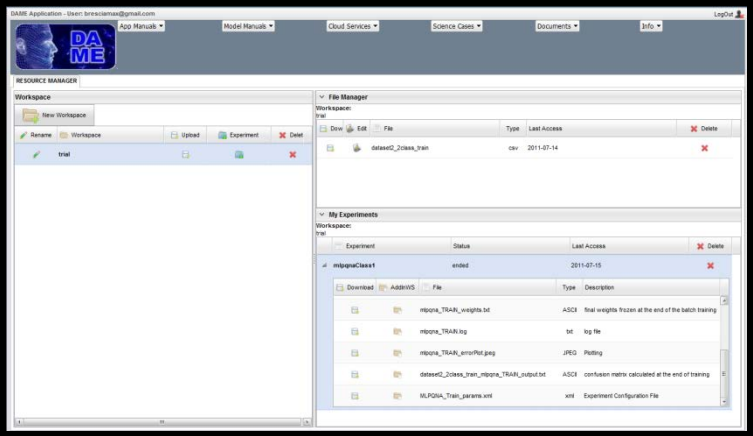
End users can remotely exploit high computing and storage power to process massive datasets (in principle they can do data mining on their smartphone…)

User can automatically plug-in his own algorithm and launch experiments through the Suite via a simple web browser







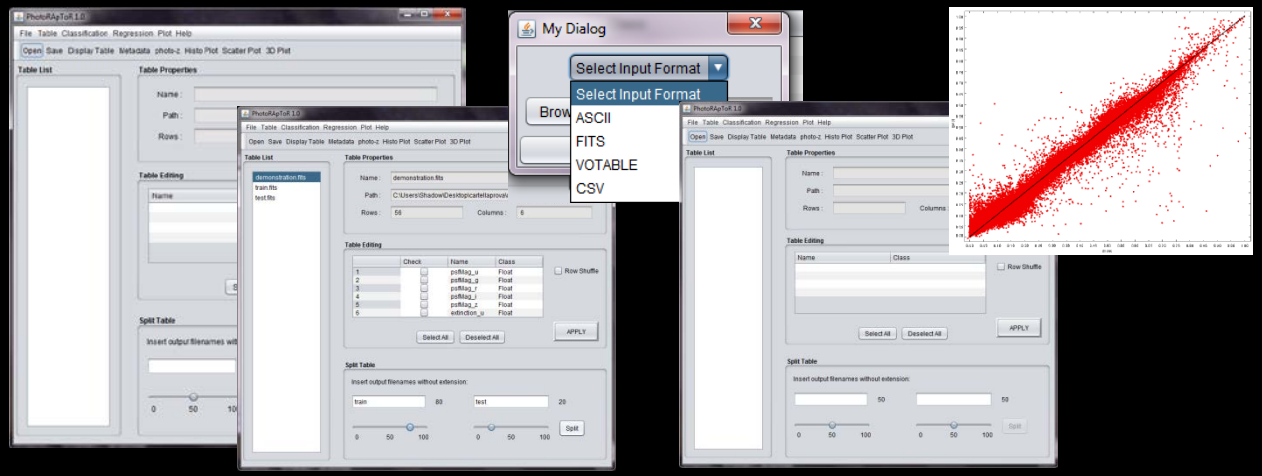


PhotoRaptor

(Photometric Research Application To Redshifts)

Java Desktop Application (multi-platform, Win7/8, Linux, Mac)
Dataset manipulation (plotting, editing, split, metadata selection, ordering, shuffling…)
MLPQNA-based photo-z estimation
General-purpose classification/regression problem solving
Post-processing (visualizing, statistical analysis)

Originally developed for Redshift, it became a multipurpose Desktop Application
First official release planned at the end of February 2014



DA ME

PhotoRaptor

Pre-Processing
(on spectroscopic KB)
Feature Selection

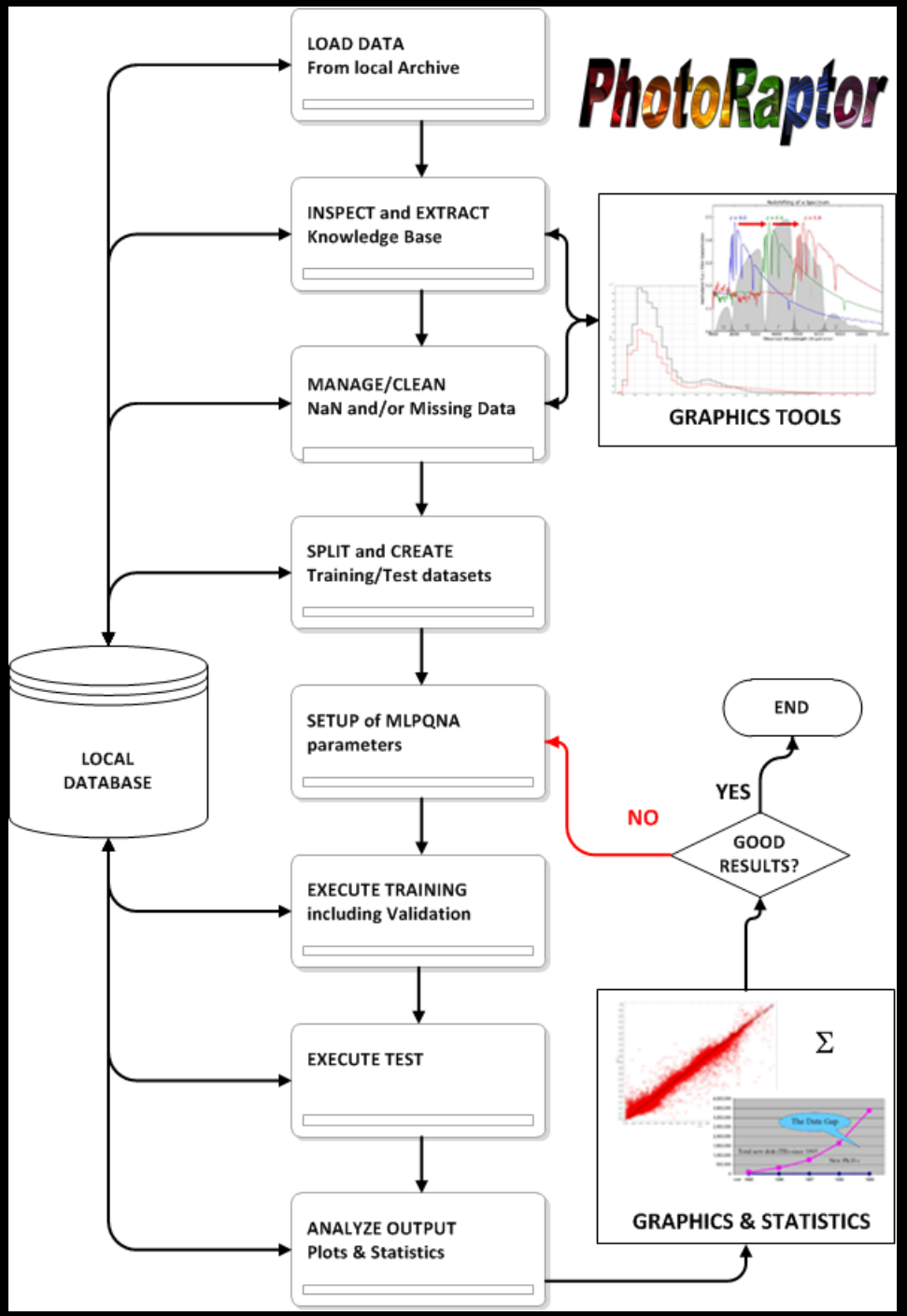
Split

Training Set Test dataset

Processing (MLPQNA)

Post-processing

Statistical Analysis
and Generalization



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'll just have a 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

References

- ✓ Cavuoti, S.; Garofalo, M.; Brescia , M.; Paolillo, M.; Pescape', A.; Longo, G.; Ventre, G.; 2014 **New Astronomy 26, 12-22**
- ✓ Cavuoti, S.; Brescia, M.; D'Abrusco, R.; Longo, G.; Paolillo, M.; 2014, **MNRAS 437, 1, 968-975**
- ✓ Brescia, M.; Cavuoti, S.; D'Abrusco, R.; Longo, G.; Mercurio, A.; 2013, **ApJ 772, 2, 140**
- ✓ Annunziatella, M.; Mercurio, A.; Brescia, M.; Cavuoti, S.; Longo, G.; 2012, **PASP 125, 923, 68-82**
- ✓ Cavuoti, S.; Brescia, M.; Longo, G.; Mercurio, A.; 2012, **A&A 546, A13, 1-8**
- ✓ Brescia, M.; Cavuoti, S.; Paolillo, M.; Longo, G.; Puzia, T.; 2012, **MNRAS 421, 2, 1155-1165**
- ❖ Brescia, M.; Cavuoti, S.; Longo, G., V. De Stefano, 2014, Photometric Redshifts for all galaxies in the SDSS DR9 with the MLPQNA method", **submitted to A&A**
- ❖ Brescia, M.; Longo, G.; Cavuoti, S.; Djorgovski, G.S.; Donalek, C.; Mahabal, A.A.; Garofalo, M.; Nocella, A.; Guglielmo, M.; Albano, G.; Esposito, F.; Manna, F.; Di Guido, A.; D'Abrusco, R.; Fiore, M.; 2014, *Mining massive astronomical data sets. The DAMEWARE framework*, **Submitted to Computing in Astronomy, Special Issue of Computer Journal, IEEE, ISSN: 0018-9162**
- ❑ 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* In prep. **to be Submitted to PASP**
- ❑ De Stefano, V.; Cavuoti, S.; Brescia, M.; Longo, G.; 2014, *Photometric redshift estimation with PhotoRApToR*, **in prep. To be submitted to Astronomy & Computing**