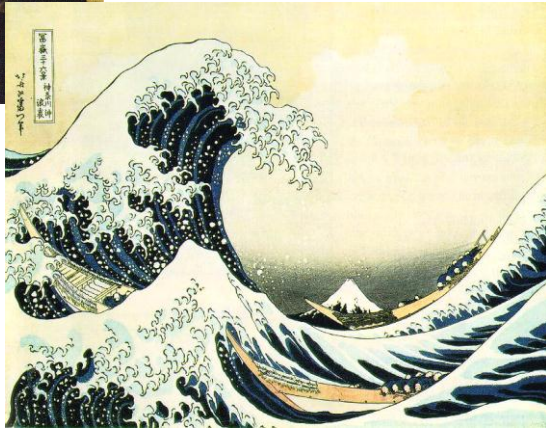
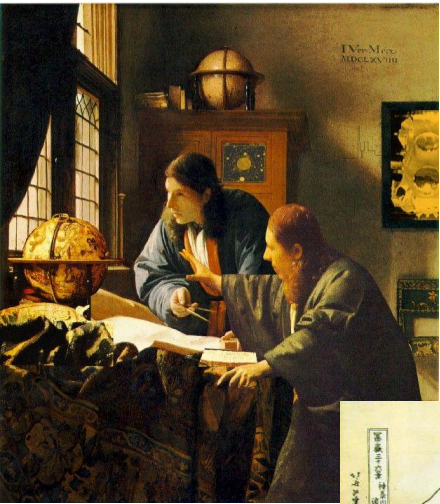


# Astronomical data mining (machine learning)



We would all testify to the growing gap between the generation of data and our *understanding* of it

...

*Ian H. Witten & E. Frank, Data Mining, 2001*

**G. Longo**

University Federico II in Napoli

**In collaboration with**

M. Brescia (INAF)

R. D'Abrusco (CfA – USA)

G.S. Djorgovski (Caltech – USA)

O. Laurino – (CfA – USA) ... **and the DAME team**



# An overview of the topics:

1. Generalities

2. A few words about the DAME (Data Mining and Exploration) application

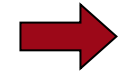
3. An application of DAME to photometric redshifts of galaxies and QSOs (a complex DM workflow)

4. Some concluding remarks

*Frate Luca Pacioli, founding father of Algebra  
Capodimonte Museum, Napoli (Italy)*



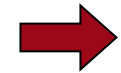
# As a result of large surveys and **VO efforts** we have entered an era where



***Most data will never be seen by humans!***

The need for data storage, network, database-related technologies, standards, etc.

Information complexity is also greatly increasing



***Most knowledge hidden behind data complexity is potentially lost***

Most (if not all) empirical relationships known so far depend on 3 parameters .... (e.g. fundamental plane of E galaxies and bulges).

Simple universe or rather human bias?



***Most data (and data constructs) cannot be comprehended by humans directly!***

The need for data mining, KDD, data understanding technologies, hyperdimensional visualization, AI/Machine-assisted discovery



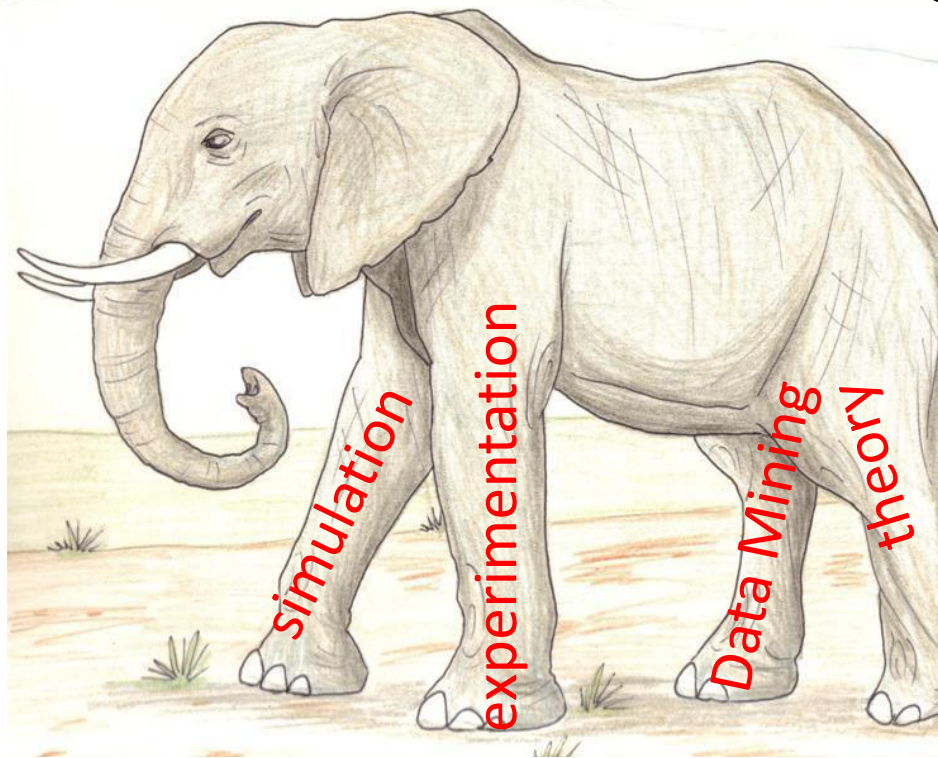
## This trend is common to many fields

Data Mining, computer science, etc. have become the “fourth leg of science” (besides theory, experimentation and simulations)

### The FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE



- Synergy between different worlds is required
- Sociological and academic issues to be solved (formation, infrastructures, and so on)



# So far restricted choice of problems due to lack of suitable KB's

Longo et al. 2003	Ball & Brunner 2009	BoK
S/G separation	S/G separation	Y
Morphological classification of galaxies <i>(shapes, spectra)</i>	Morphological classification of galaxies <i>(shapes, spectra)</i>	Y
Spectral classification of stars	Spectral classification of stars	Y
Image segmentation	Image segmentation	
Noise removal <i>(grav. waves, pixel lensing, images)</i>	-----	
Photometric redshifts <i>(galaxies)</i>	Photometric redshifts <i>(galaxies, QSO's)</i>	Y
Search for AGN	Search for AGN and QSO	Y
Variable objects	<b>Time domain</b>	
Partition of photometric parameter space for specific group of objects	Partition of photometric parameter space for specific group of objects	Y
Planetary studies (asteroids)	Planetary studies (asteroids)	Y
Solar activity	Solar activity	Y
<b>Interstellar magnetic fields</b>	----	
<b>Stellar evolution models</b>	----	

## **Very incomplete list of activity going on since 2010**

- **IJCAI-09 Workshop on Machine Learning and AI Applications in Astrophysics and Cosmology**, Pasadena 2009
- **Astroinformatics: AAS n. 215**, Washington 2009
- **First IG-KDD meeting, IVOA Interop**, Victoria
- **Astroinformatics 2010: Caltech**, Pasadena 2010
- **ANITA Workshop**, Perth 2011
- **Second IG-KDD meeting IVOA Interop**, Napoli may 2011
- **GREAT School**, La Palma 2011
- **Astroinformatics 2011**, Sorrento September 2011
- Many other dedicated sessions at larger meetings (IEEE, SIAM, etc.)

Firefox - Astroinformatics 2011 x The IJCAI-09 Workshop on Machine L... x +

http://dame.dsf.unina.it/astroinformatics2011.html ☆ - Machine learning pasadena workshop

Il sito è SICURO - Invia notifica

# AstroInformatics2011

Sorrento, September 25 - 29, 2011

## Main Menu

- Home
- SOC and LOC
- Acknowledgments
- Location
- Program
- Registration
- Participants
- Accommodation
- Social Events
- Napoli info
- Weather

## Welcome to AstroInformatics 2011 Conference

### Sorrento, September 25-29 (Sun-Thu), 2011

Astronomy is rapidly becoming exponentially data-rich, and the tasks of data management, data exploration, and knowledge discovery become central to the research enterprise, bringing along many technical and methodological challenges. Information and Communication Technology could also provide the stage where to interact, publish, preserve and disseminate knowledge.

The newly emerging field of AstroInformatics is a bridge between scientific challenges posed by the exponential growth of data volumes and complexity in astronomy, and applied statistics, computer science, and engineering. Our goal is to engage a broader community of astronomers and computing and informatics professionals in developing and applying new tools and techniques for the data-rich astronomy in the 21st century. A key component of this is training of a new generation of computationally empowered students and scientists.

The main focus of this year's conference is on practical tools for knowledge discovery and exploration in large and complex data sets. We will also have topical workshops on Practical AstroSemantics, Computational Education for Scientists, and the WorldWide Telescope

There will be a modest number of invited review talks to serve as a basis for the discussion, and a lot of discussion, some of it led by panels. Contributed papers will be accommodated as posters.

Reception: evening of September 25



skype  
apimaia  
è in linea



11:36  
02/06/2011

<http://dame.dsf.unina.it/astroinformatics2011.html>

Maximum number of attendees: ca. 50. Register soon





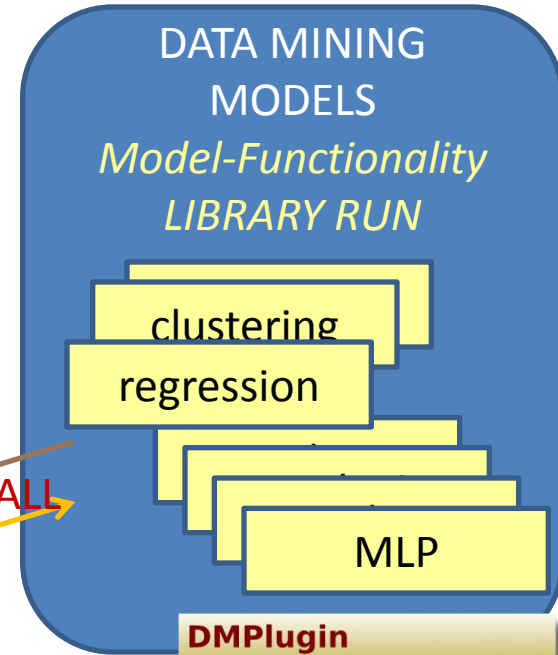
# The DAME architecture



user



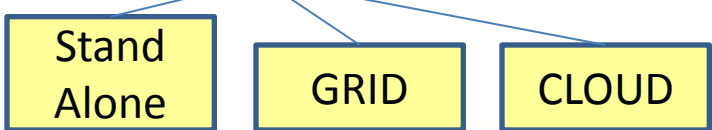
Client-server AJAX  
(Asynchronous JAVa-  
Xml) based;  
interactive web app  
based on Javascript  
(GWT-EXT);



HW env virtualization;  
Storage + Execution LIB  
Data format conversion



Restful, Stateless Web Service  
experiment data, working  
flow trigger and supervision  
Servlets based on XML  
protocol




# DAME front-end



Browser tabs: G., F..., P., B..., N., h..., C., W., V..., S W., S..., V..., I..., I..., W., D., M.

Address bar: http://dame.scope.unina.it:8080/FrontEnd/

DAME Application



**Workspace Manager**

New Workspace

Workspace [Upload] [Experiment] [Rename] [Delete]

No items to show.

**Files Manager**

Download	File	Last Access	Delete
No items to show.			

**Experiment Manager**

Experiment	Status	Last Access	Delete
No items to show.			

Help SECTION

# DAME plugin wizard



DMPlugin Application Wizard

File Help

### Plugin Informations

Name: Example  
Documentation: http://www.someurl.edu/url  
Version: 1.0  
Domains: clustering

### Owner Informations

Owner Name: John Smith  
Owner Mail: john@someurl.edu

### Running Modes Informations

Train	<input checked="" type="checkbox"/>	Documentation: http://www.someurl.edu/#train	Running Time: 0
Test	<input type="checkbox"/>	Documentation:	Running Time: 0
Run	<input type="checkbox"/>	Documentation:	Running Time: 0
Full	<input type="checkbox"/>	Documentation:	Running Time: 0

### Components

- Train
  - Fields
    - someField
  - Input Files
    - inputFile
  - Output Files
    - output

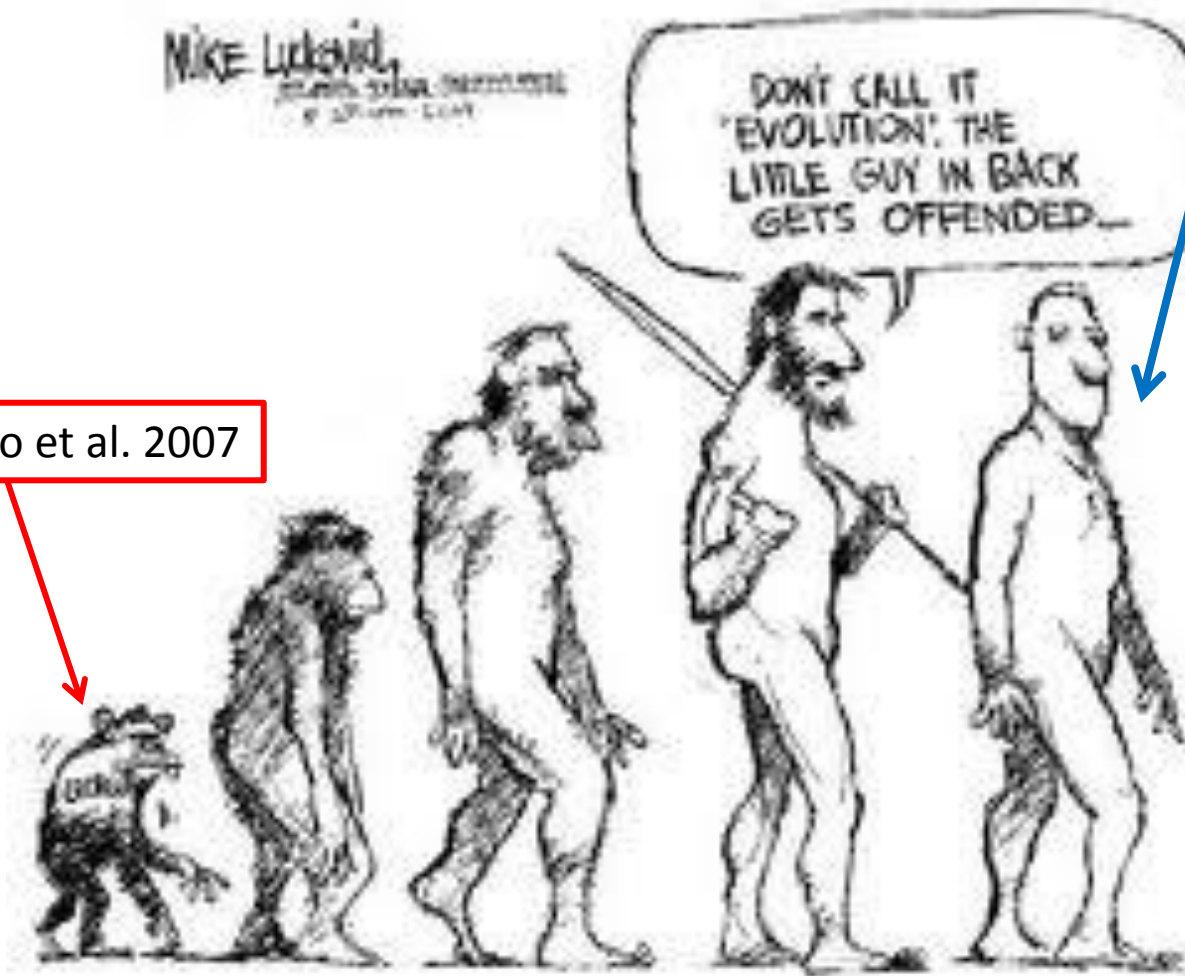
Name: output  
Description: Output File  
Format: votable  
 is Partial  
Save

Add Delete Edit

# PHOTOMETRIC REDSHIFTS: the evolution of a DM method

Laurino, D'Abrusco et al. 2011

D'Abrusco et al. 2007



# GOAL

to produce a DM workflow capable to automatically derive phot-  
z's for all extragalactic objects (galaxies and quasar), including  
QSO candidate selection

## Data used :

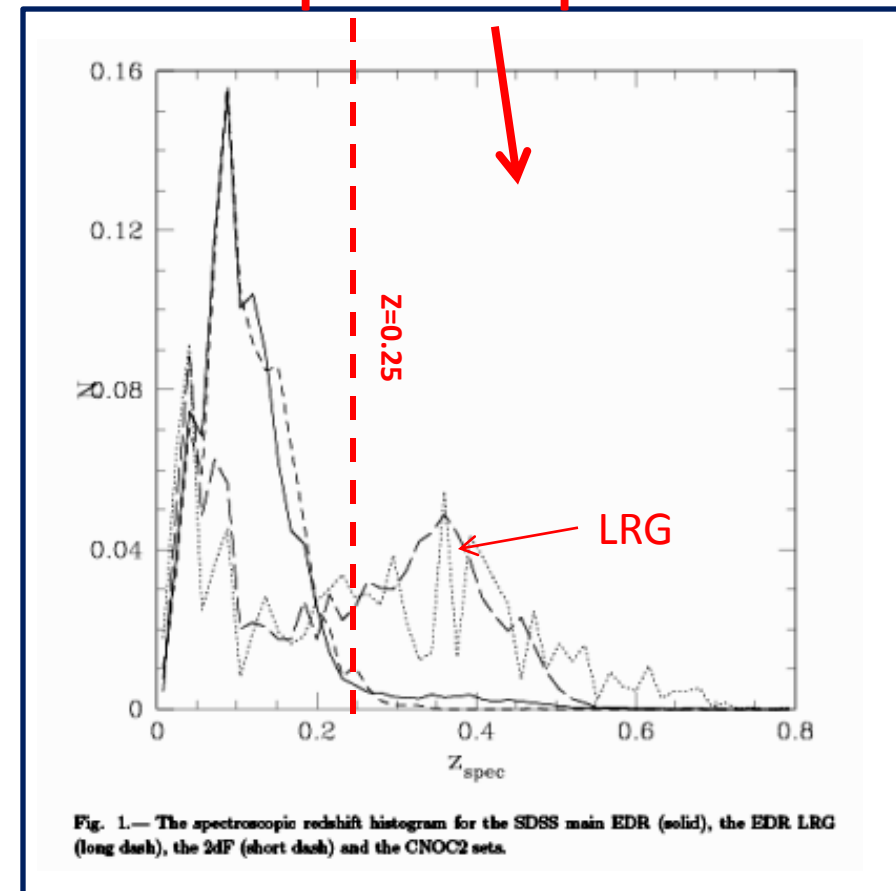
**SDSS:**  $10^8$  galaxies in 5 optical bands;  
**BoK:** spectroscopic redshifts for  
 $10^6$  galaxies  
**BoK:** incomplete and **biased**.

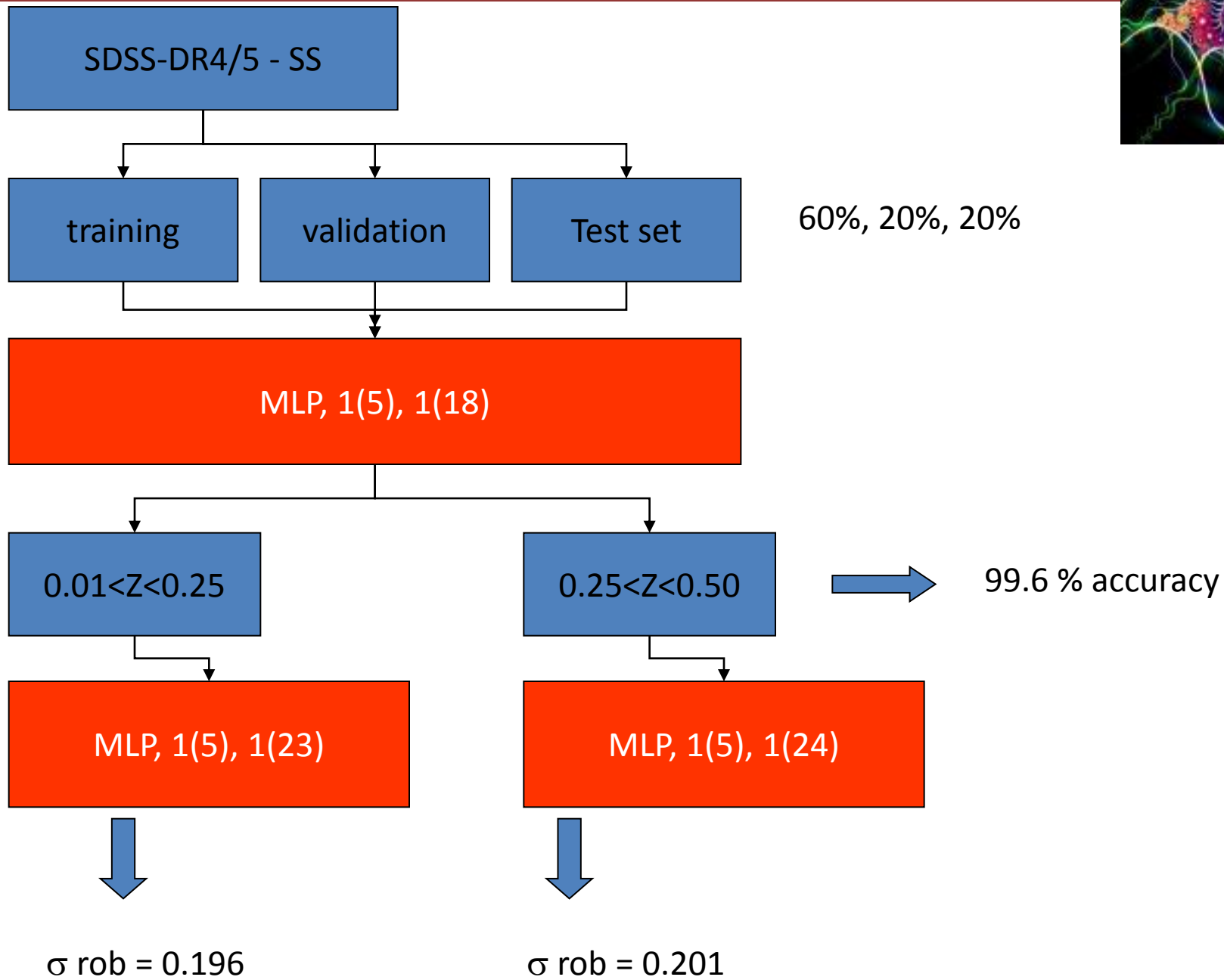
UKIDSS

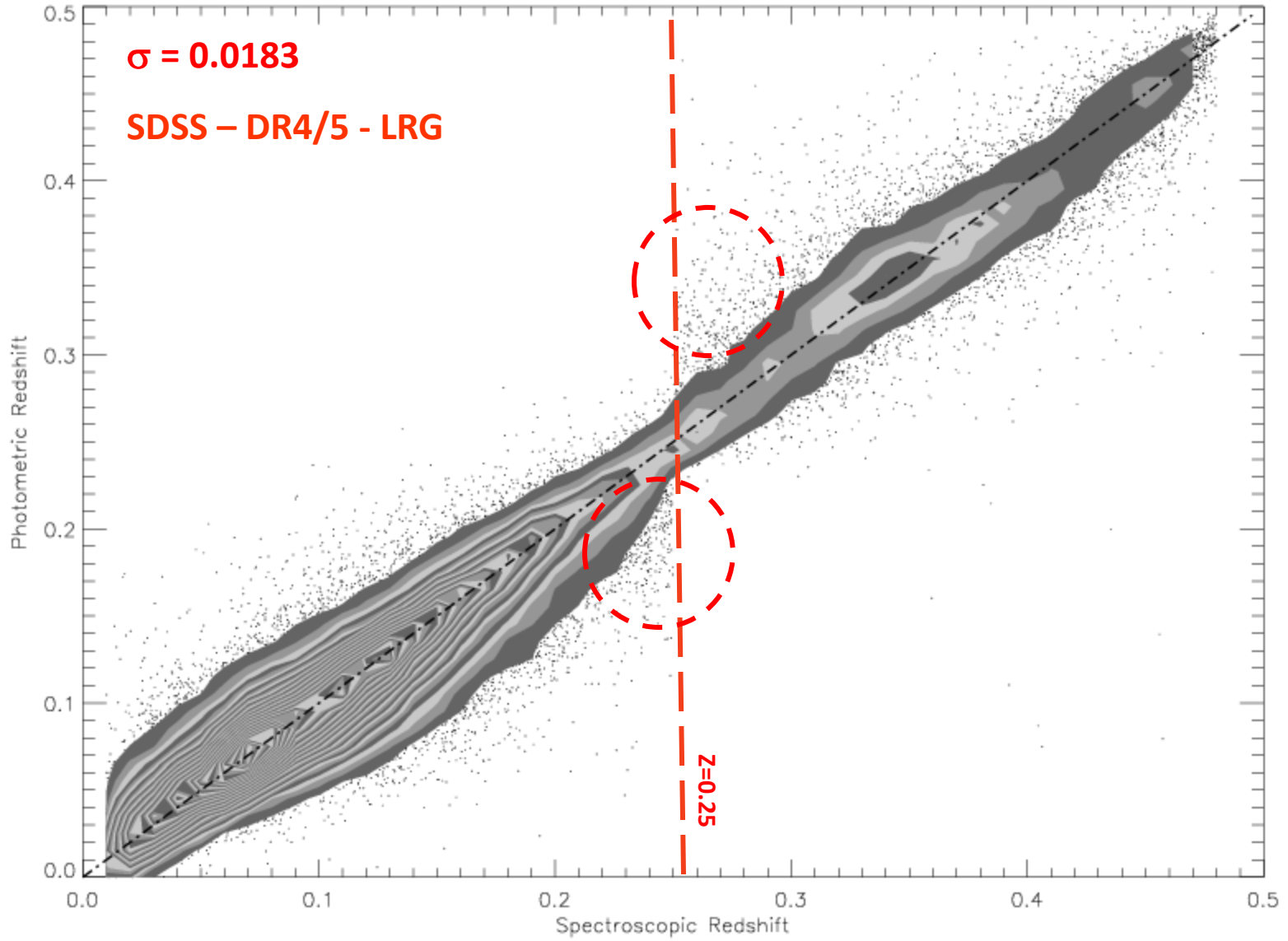
Galex



→ Spectroscopic KB SDSS

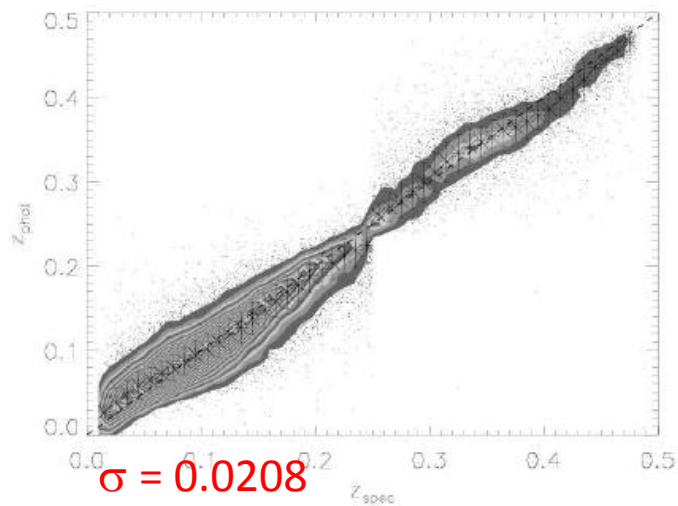






# Uneven coverage of parameter space:

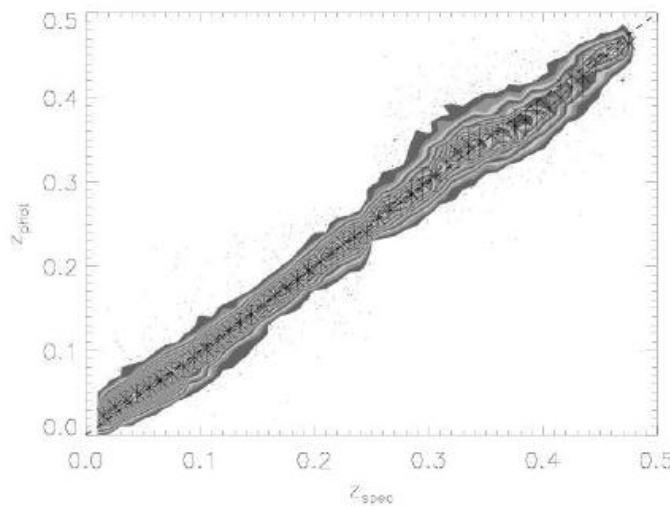
General galaxy sample



$$\sigma = 0.0208$$

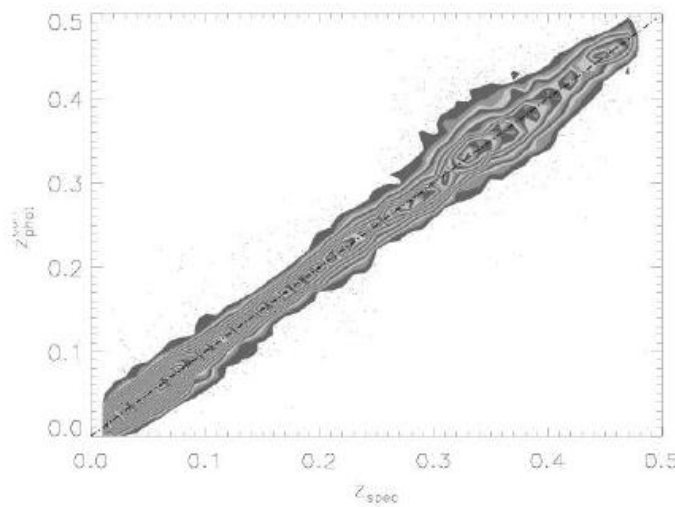
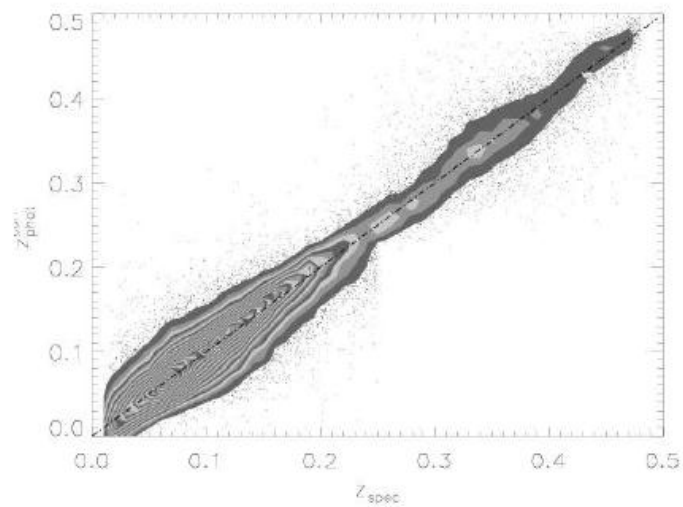
$$\Delta z = -0.0029$$

LRG sample



$$\sigma = 0.0178$$

$$\Delta z = -0.0011$$



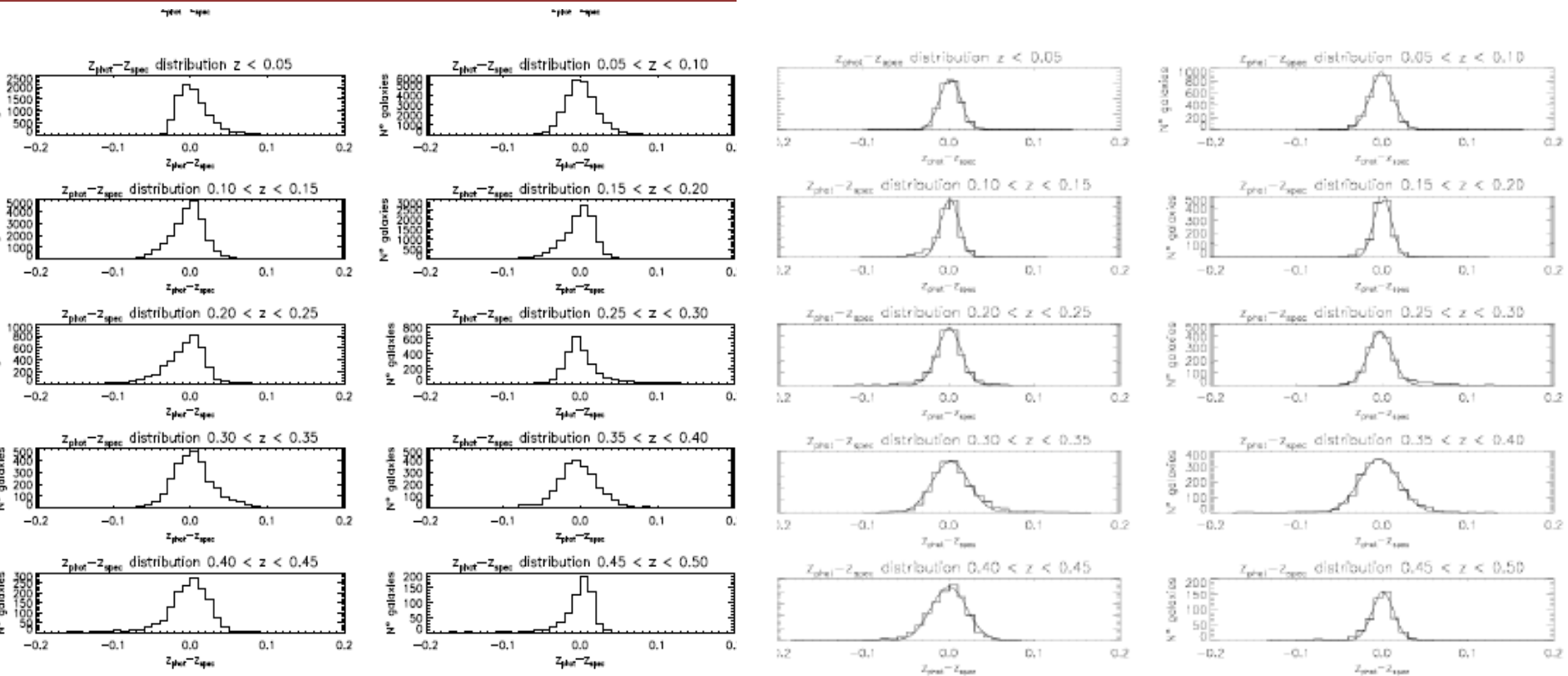
Non LRG only

$$\sigma = 0.0363$$

$$\Delta z = -0.0030$$



# Errors can be easily evaluated



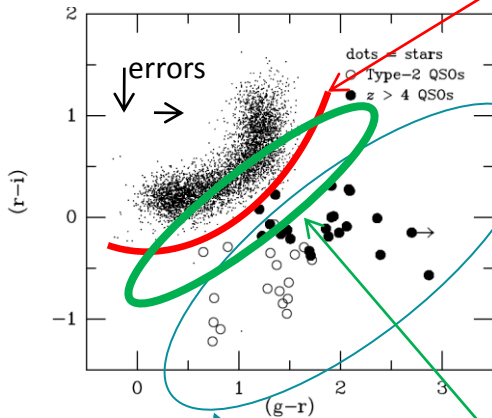
General galaxy sample

LRG sample

**And are, on average, well behaved but incompletely defined....**

# STEP 2: Photometric selection of candidate QSO's (as a clustering problem)

Traditional way to look for candidate QSO in 3 band survey



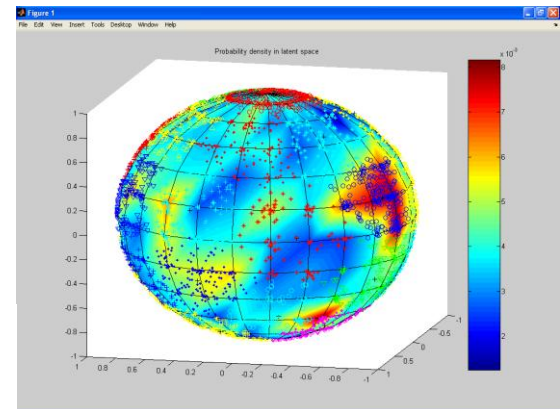
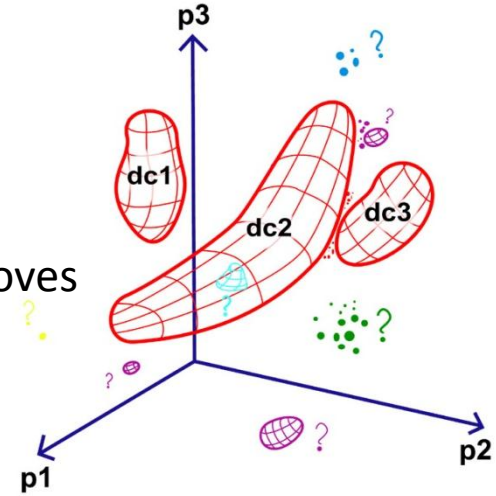
Cutoff line

Candidate QSOs for spectroscopic follow-up's

Ambiguity zone

Adding one feature improves separation...

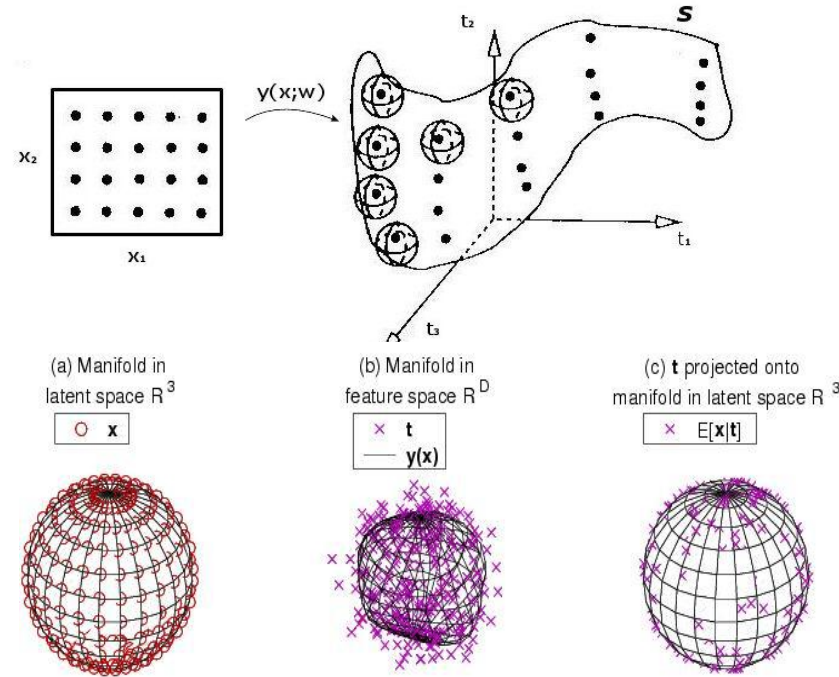
A Generic Machine-Assisted Discovery Problem: Data Mapping and a Search for Outliers



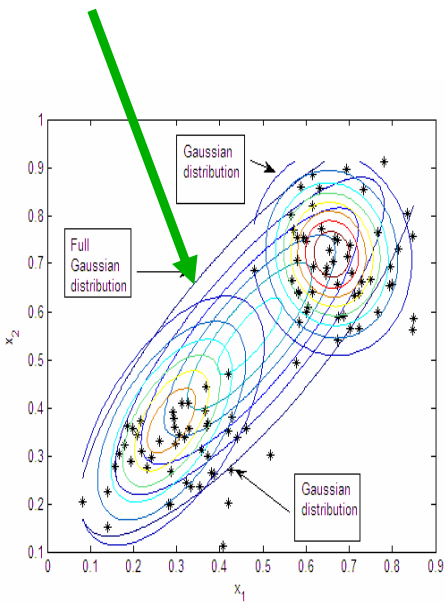
PPS projection of a 21-D parameter space showing as blue dots the candidate quasars. Notice better disentanglement

## Step 1: Unsupervised clustering with PPS - Probabilistic Principal Surfaces

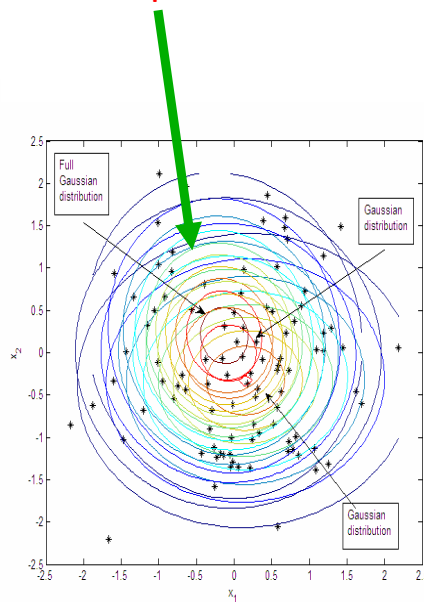
PPS determines a large number of distinct groups of objects: nearby clusters in the color space are mapped onto the surface of a sphere.



Not replaced!



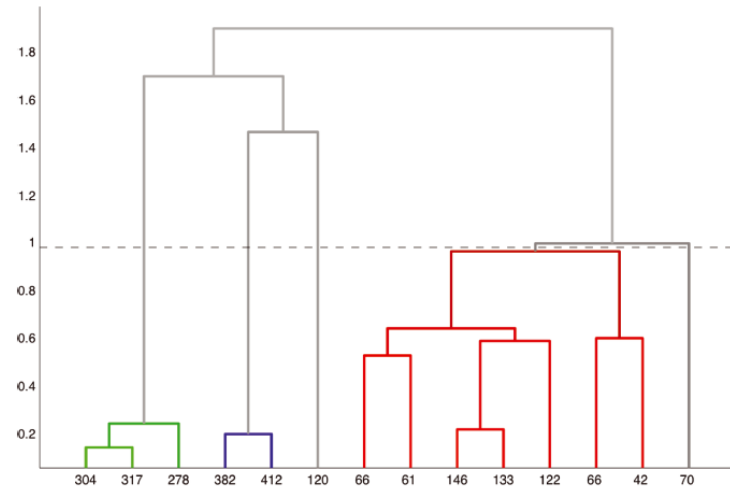
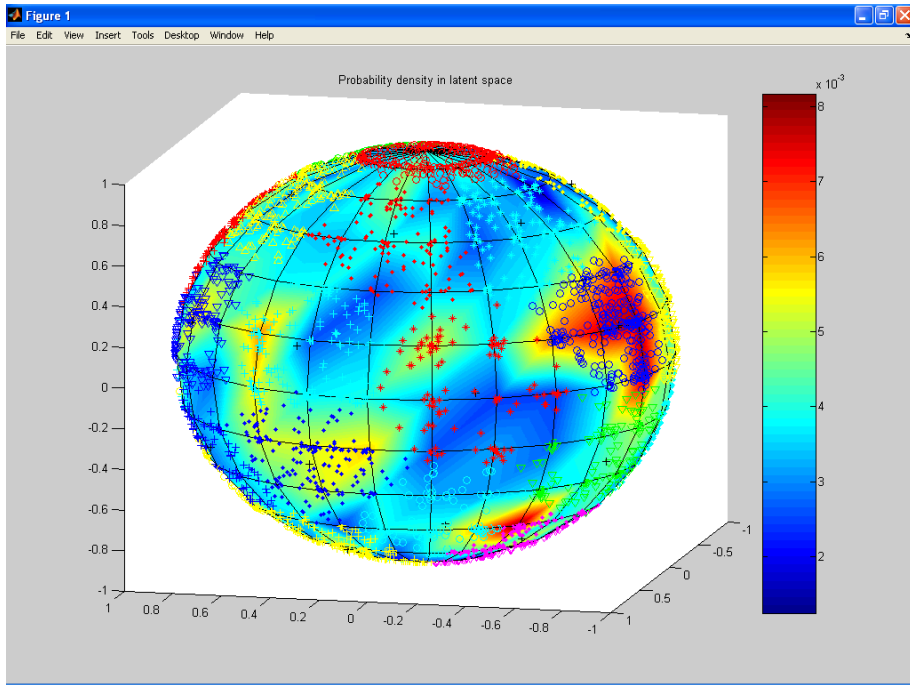
Replaced!



## Step 2: Cluster agglomeration

NEC aggregates clusters from PPS to a (a-priori unknown) number of final clusters.

- Plateau analysis:** final number of clusters  $N(D)$  is calculated over a large interval of  $D$ , and critical value(s)  $D_{th}$  are those for which a plateau is visible.
- Dendrogram analysis:** the stability threshold(s)  $D_{th}$  can be determined observing the number of branches at different levels of the graph.



**Figure 7.** An example of dendrogram (see text for details) used as a representation of an agglomerative clustering process performed by the NEC algorithm. On the x and y axes are, respectively, reported the numeric labels of the initial clusters and the values of the dissimilarity threshold. The dashed line intersects the lines of the tree-like graph associated with the clusters produced by the clustering process when the value of the dissimilarity threshold has been fixed to  $T_{cr} = 1$ .

To determine the critical dissimilarity  $D_{th}$  threshold we rely not only on a stability requirement.

A cluster is successful if fraction of confirmed QSO is higher than assumed fractionary value (**Th**)

$D_{th}$  is required to maximize **NSR**  $NSR = \frac{\text{Number of successful clusters}}{\text{Number of total clusters}}$

The process is recursive: feeding merged unsuccessful clusters in the clustering pipeline until no other successful clusters are found.

The overall efficiency of the process  $e_{tot}$  is the sum of weighed efficiencies  $e_i$  for each generation:

$$e_{tot} = \sum_{i=1}^n e_i$$

# Data and experiments

## Data samples:

1. **Optical**: sample derived from SDSS database table “Target” queried for QSO candidates, containing  $\sim 1.11 \cdot 10^5$  records and  $\sim 5.8 \cdot 10^4$  confirmed QSO (‘specClass == 3 OR specClass == 4’).
2. **Optical + NIR**: sample derived from positional matching (‘best’) between SDSS-DR3 database view “Star” queried for all objects with spectroscopic follow-up available and detection in all 5 bands (u,g,r,i,z) with high reliability for redshift estimation and line-fitting classification (‘specClass’) and high S/N photometry, and UKIDSS-DR1 star-like (‘mergedClass == -1’) objects fully detected in each of the four Survey bands (Y,J,H,K) and clean photometry. **This sample is formed by 2192 objects.**

## Experiments:

### Optical (1)

candidate QSO

4 colours

### Optical+NIR (2)

star-like objects

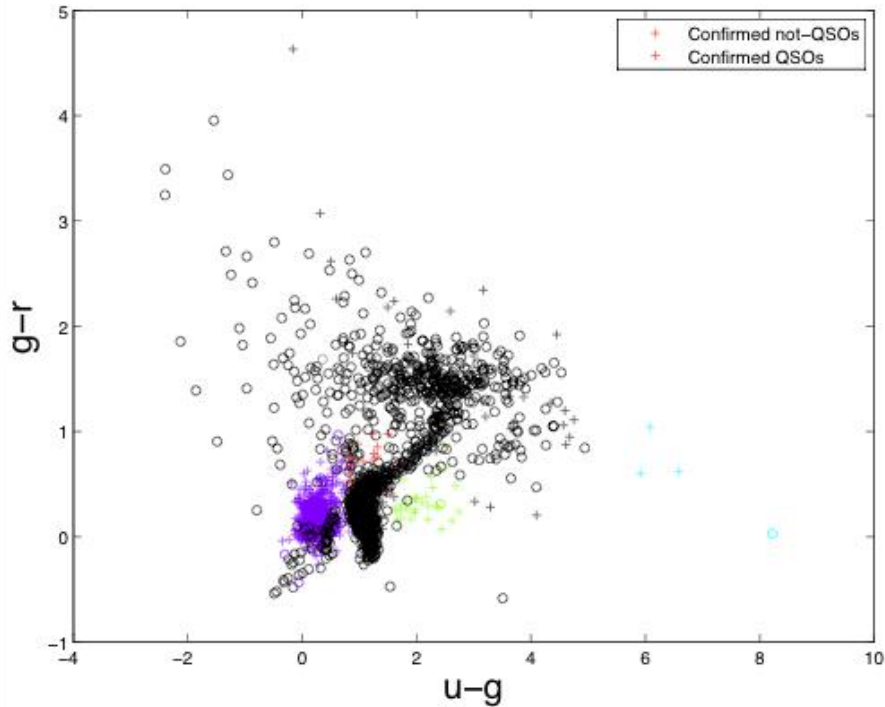
4 + 3 colours

### Optical (3)

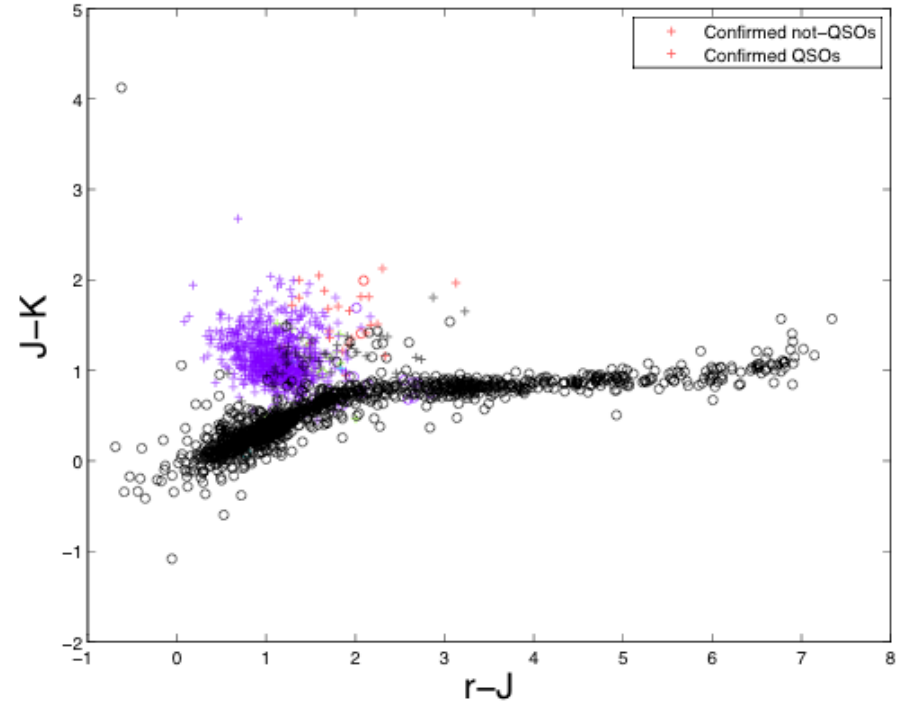
star-like objects

4 colours

**u - g vs g - r**

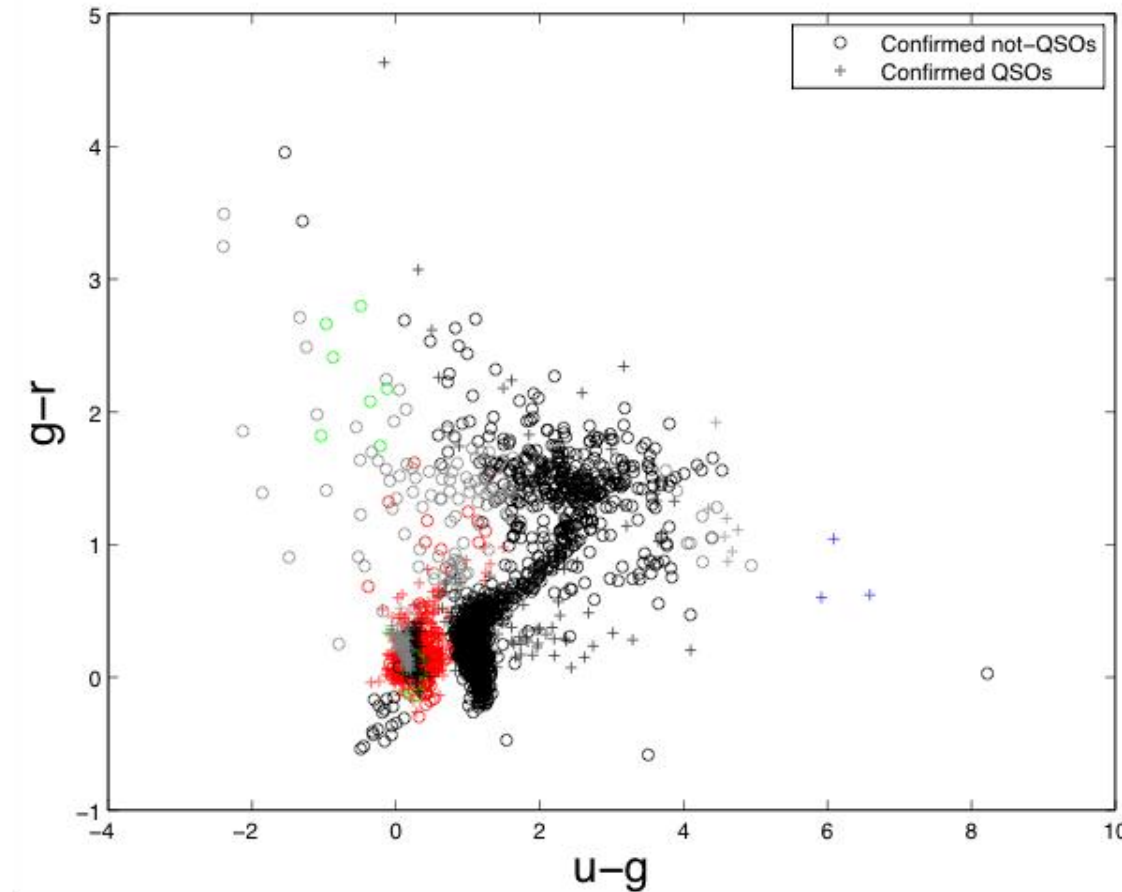


**r - J vs J - K**



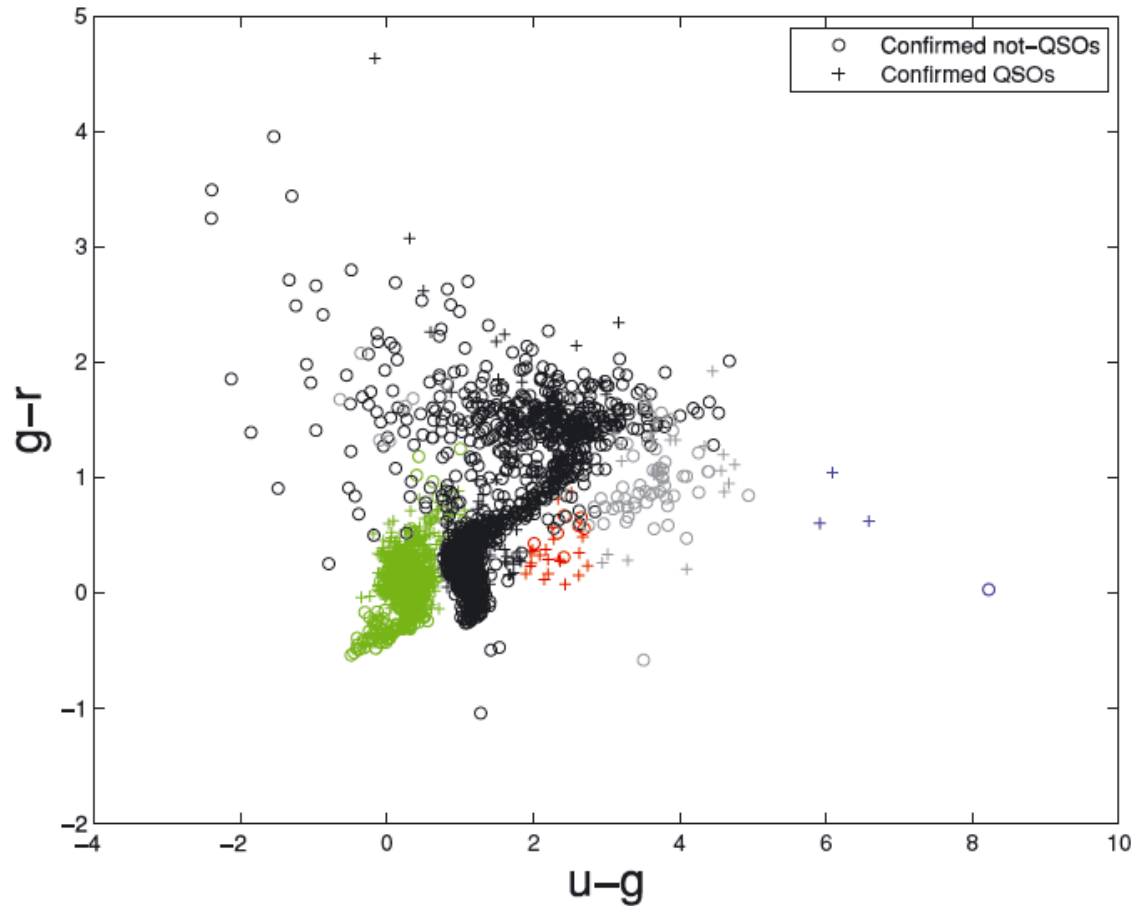
**Only a fraction (43%) of these objects have been selected as candidate QSO's by the first SDSS targeting algorithm in first instance:** the remaining sources have been included in the spectroscopic program because they have been selected in other spectroscopic programmes (mainly stars).

**u - g vs g - r**

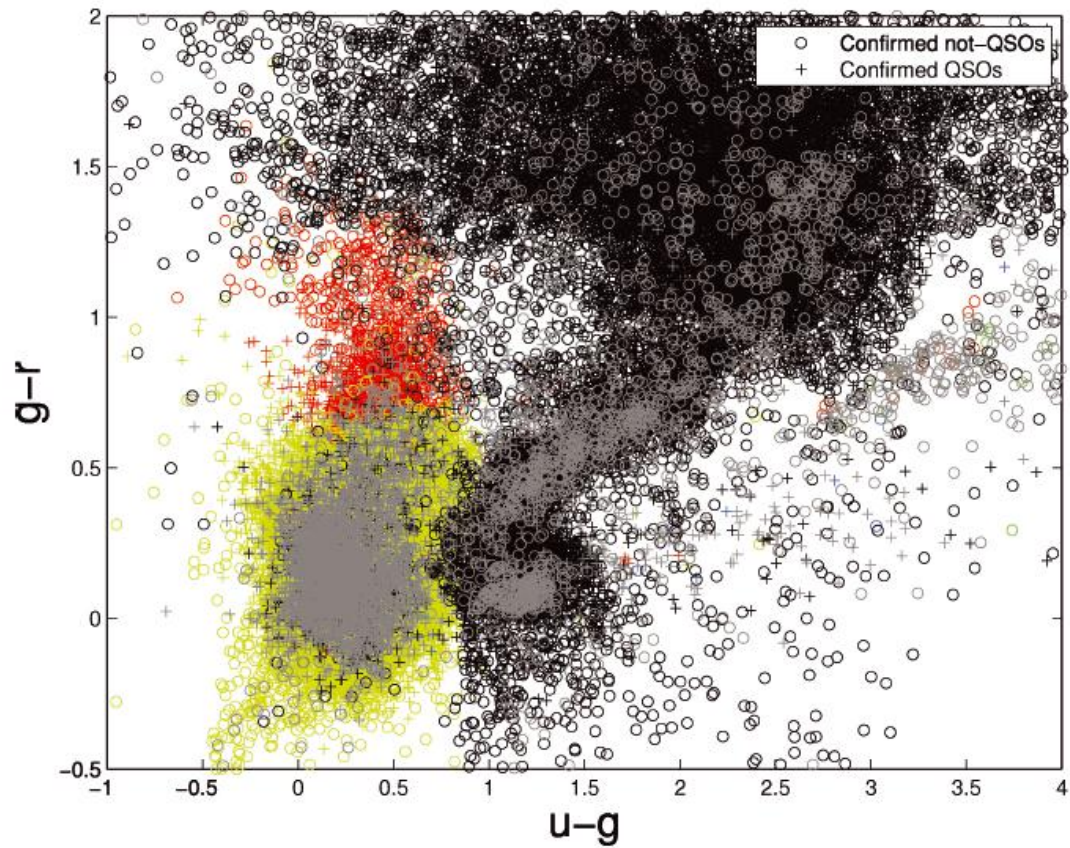


In this experiment the clustering has been performed on the same sample of the previous experiment, using only optical colours.





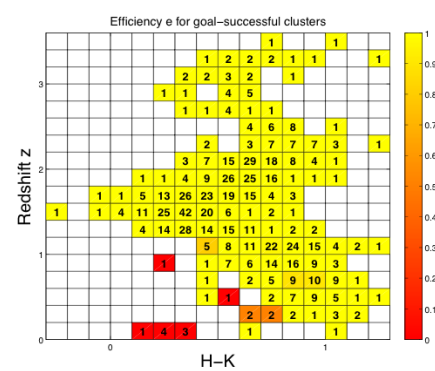
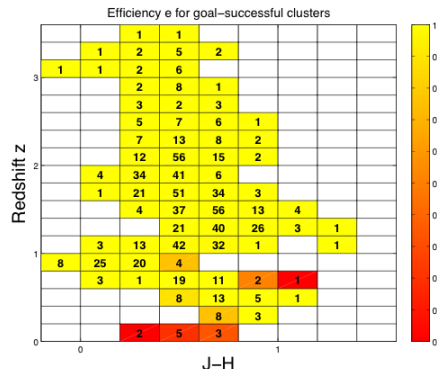
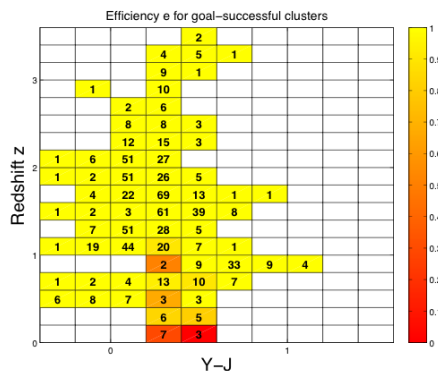
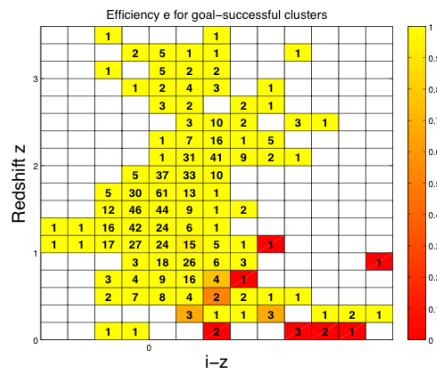
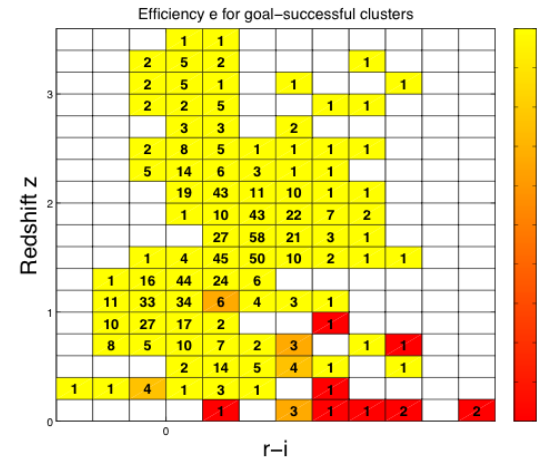
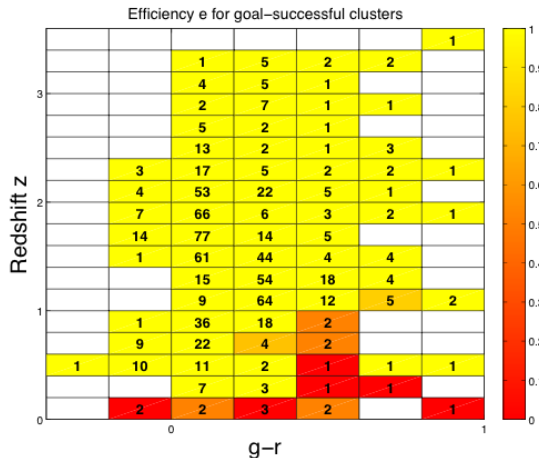
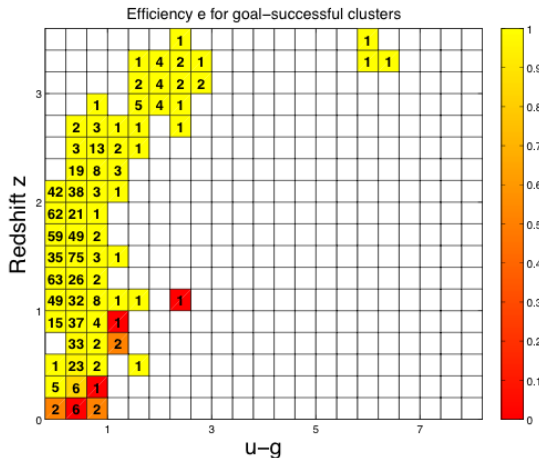
**Figure 12.** Distribution of S-A sample points in the  $(u - g)$  versus  $(g - r)$  plane after the labelling phase of the first experiment. Black and grey symbols are associated, respectively, to members of the ‘notgoal-successful’ and ‘unsuccessful’ clusters, while each single ‘goal-successful’ cluster is drawn using a different colour. All confirmed QSOs and not QSOs, regardless of their membership, are represented by crosses and circles, respectively. Three final ‘goal-successful’ clusters, green, red and blue, respectively, are selected.



**Figure 22.** Distribution of S-S sample points in the  $(u - g)$  versus  $(g - r)$  plane after the labelling phase of the fourth experiment. Black and grey symbols are associated, respectively, to members of the ‘notgoal-successful’ and ‘unsuccessful’ clusters, while each single ‘goal-successful’ cluster is drawn using a different colour. All confirmed QSOs and not QSOs, regardless of their membership, are represented by crosses and circles, respectively.

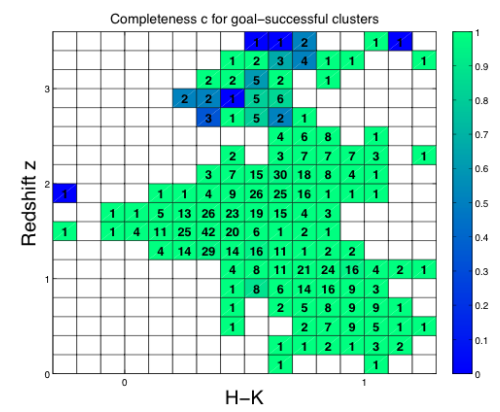
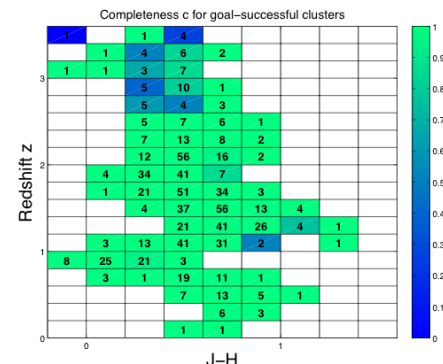
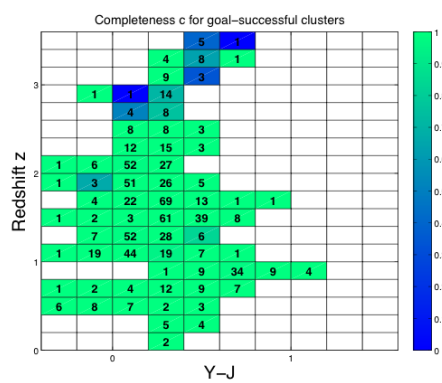
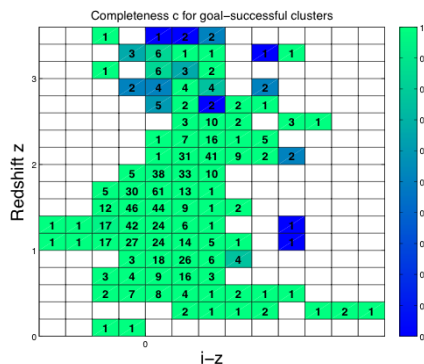
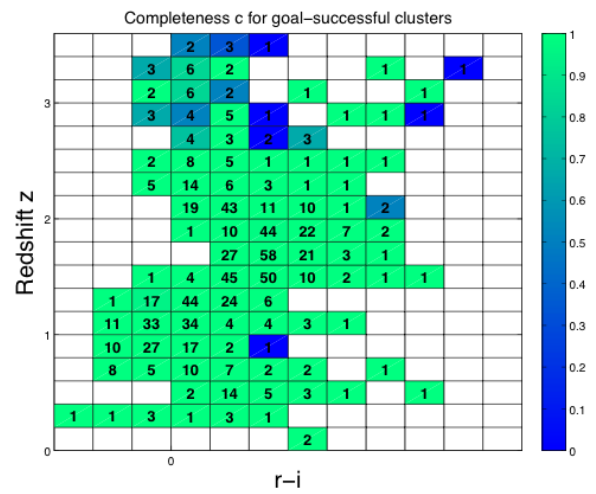
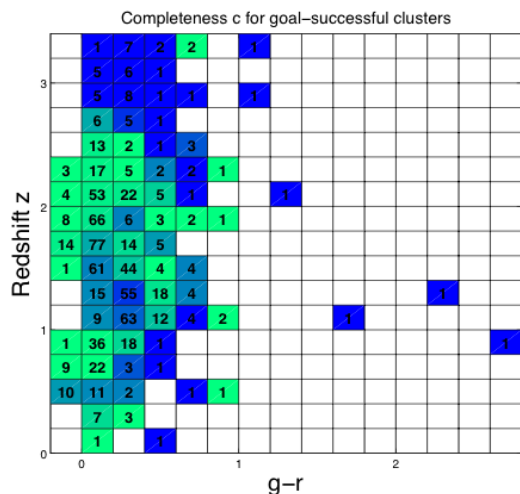
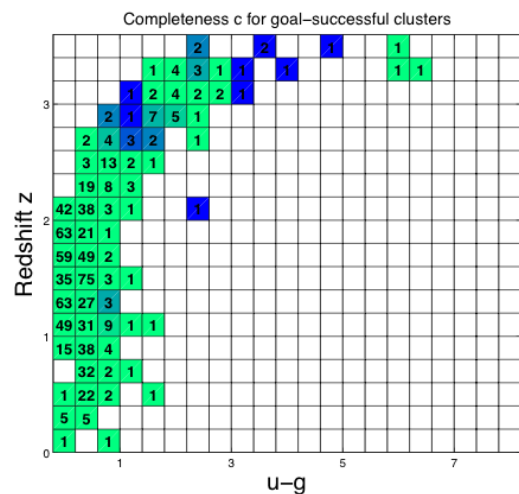
Experiment	Colours	$e_{\text{tot}}$ (per cent)	$c_{\text{tot}}$ (per cent)
1	Natural	81.5	89.3
1	$(u - r, g - i, r - z, i - u)$	81.7	89.5
1	$(u - i, g - z, r - g, i - r)$	80.8	89.3
1	$(u - z, g - u, r - z, i - r)$	82.0	89.0
1	$(u - g, g - z, r - i, i - z)$	81.4	89.7
2	Natural	92.3	91.4
2	$(u - r, g - i, r - z, i - u, Y - H, J - K, H - Y)$	92.3	91.5
2	$(u - i, g - z, r - g, i - r, Y - K, J - Y, H - J)$	92.7	91.8
2	$(u - z, g - u, r - z, i - r, Y - H, J - K, H - Y)$	91.9	90.9
2	$(u - Y, g - H, r - J, i - K, z - u, Y - g, H - r)$	91.0	91.0
2	$(u - H, g - J, r - K, i - u, z - g, Y - z, H - i)$	90.9	91.2
2	$(u - J, g - K, r - u, i - g, z - r, Y - i, H - z)$	92.2	91.5
2	$(u - z, g - K, H - J, z - Y, r - u, z - i, i - H)$	92.6	91.4
3	Natural	97.3	94.3
3	$(u - r, g - i, r - z, i - u)$	97.1	94.8
3	$(u - i, g - z, r - g, i - r)$	97.0	93.9
3	$(u - z, g - u, r - g, i - r)$	97.3	94.0
3	$(u - g, g - z, r - i, i - z)$	96.9	94.9
4	Natural	94.5	93.0
4	$(u - r, g - i, r - z, i - u)$	95.2	93.9
4	$(u - i, g - z, r - g, i - r)$	95.0	94.0
4	$(u - z, g - u, r - g, i - r)$	95.4	94.4
4	$(u - g, g - z, r - i, i - z)$	95.7	94.6

The catalogue of  $1.83 \times 10^6$  candidate quasars is publicly available at the URL: [http://voneural.na.infn.it/catalogues\\_qsos.html](http://voneural.na.infn.it/catalogues_qsos.html)



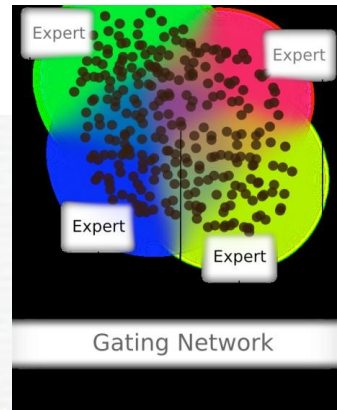
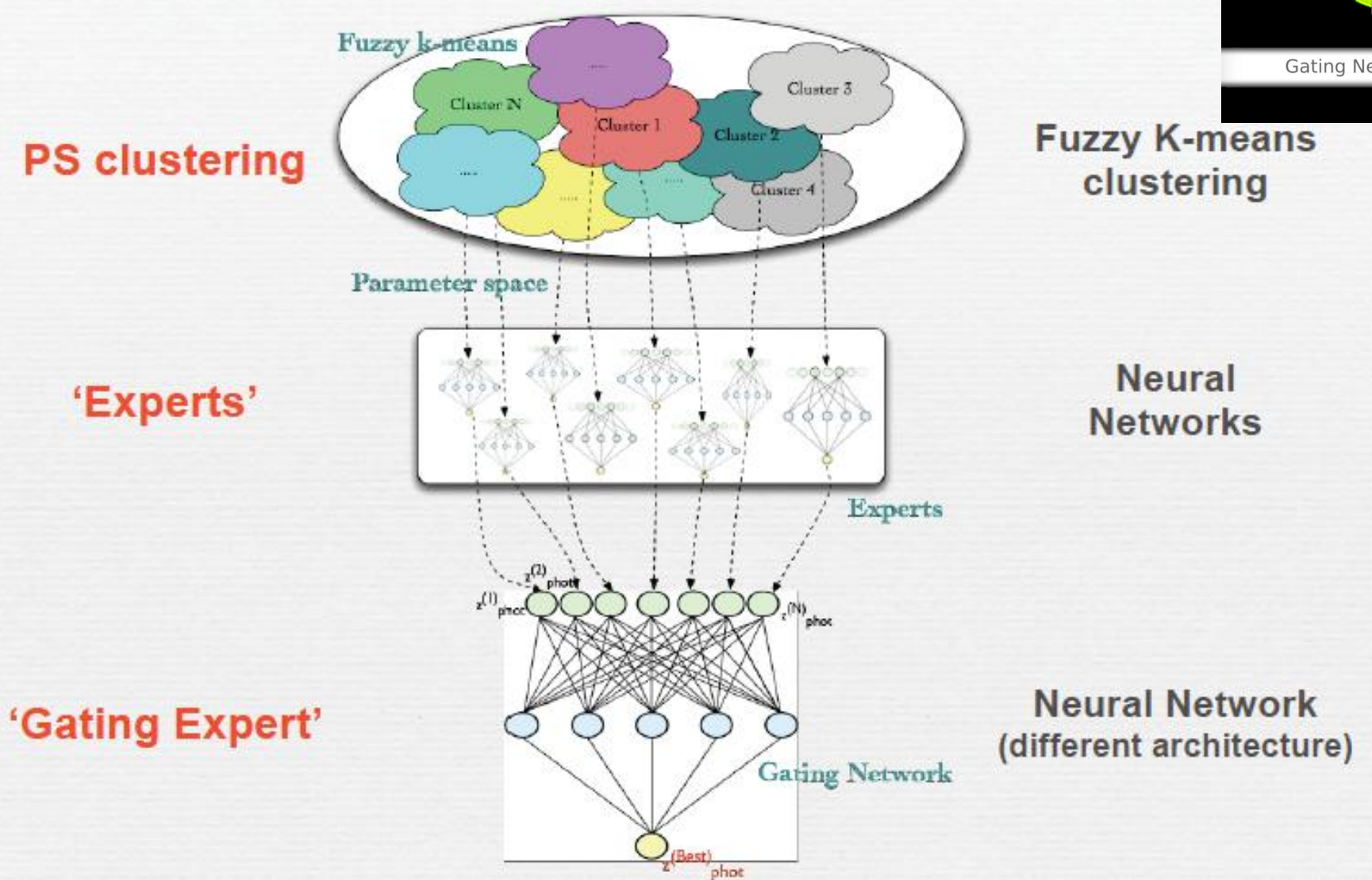
**Experiment 2:**  
local values of  $e$

# Experiment 2: local values of $c$

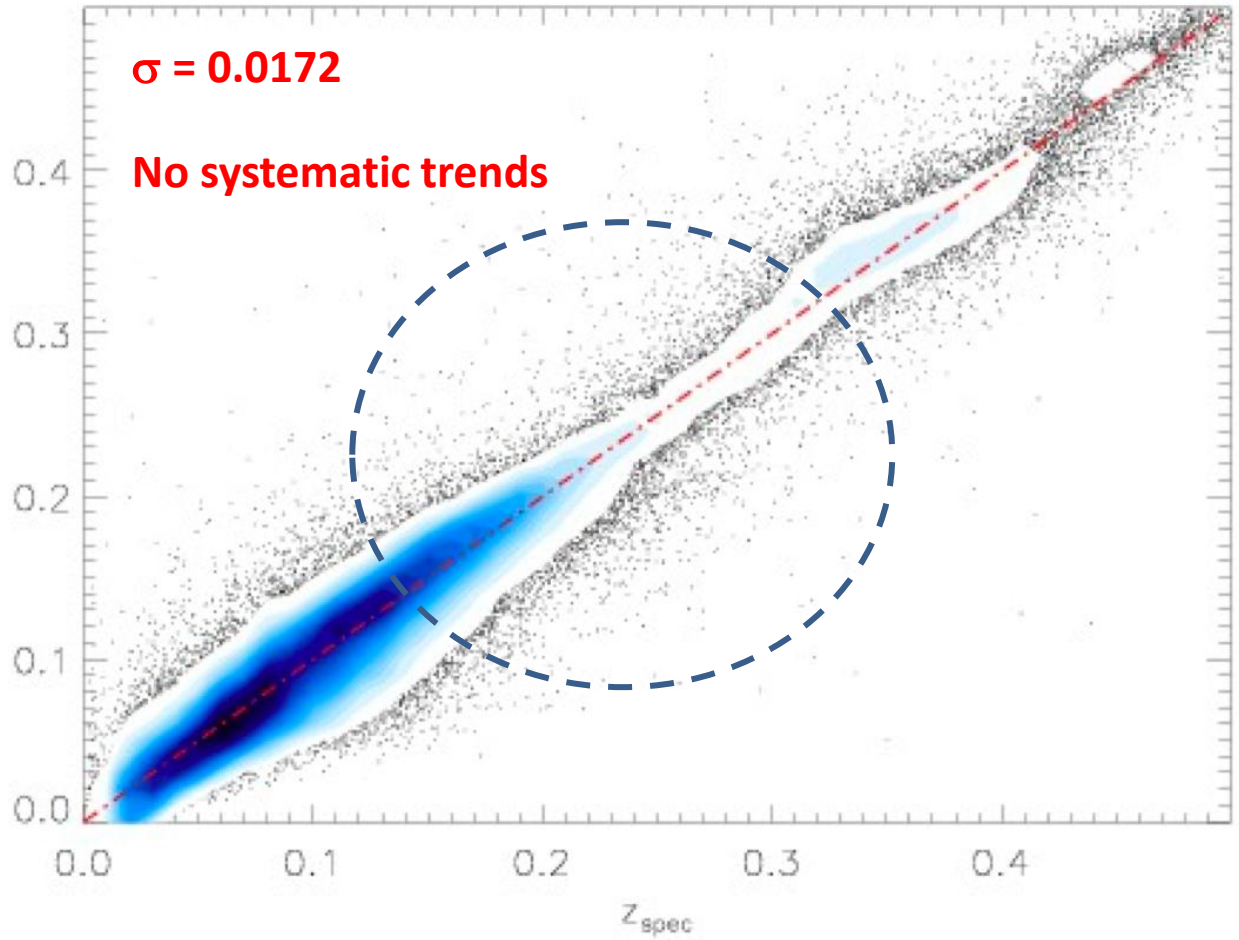
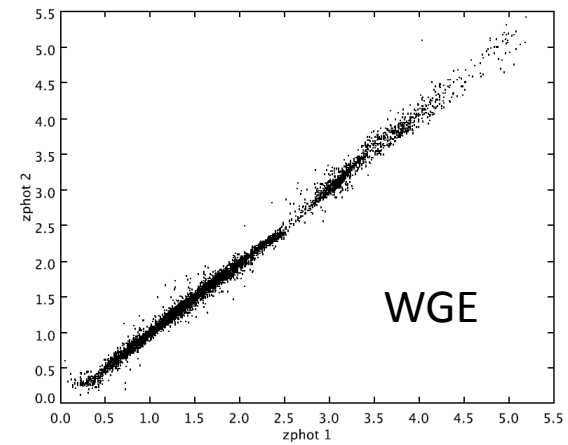
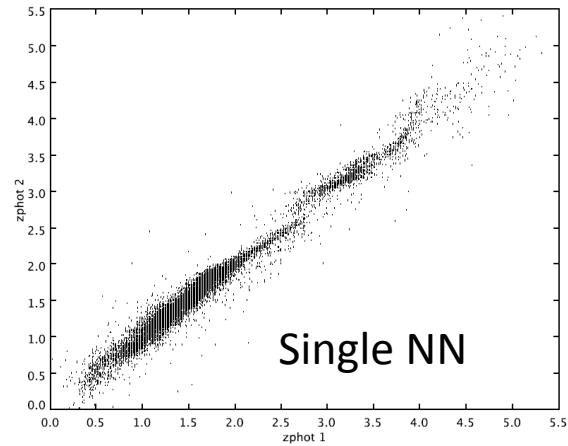


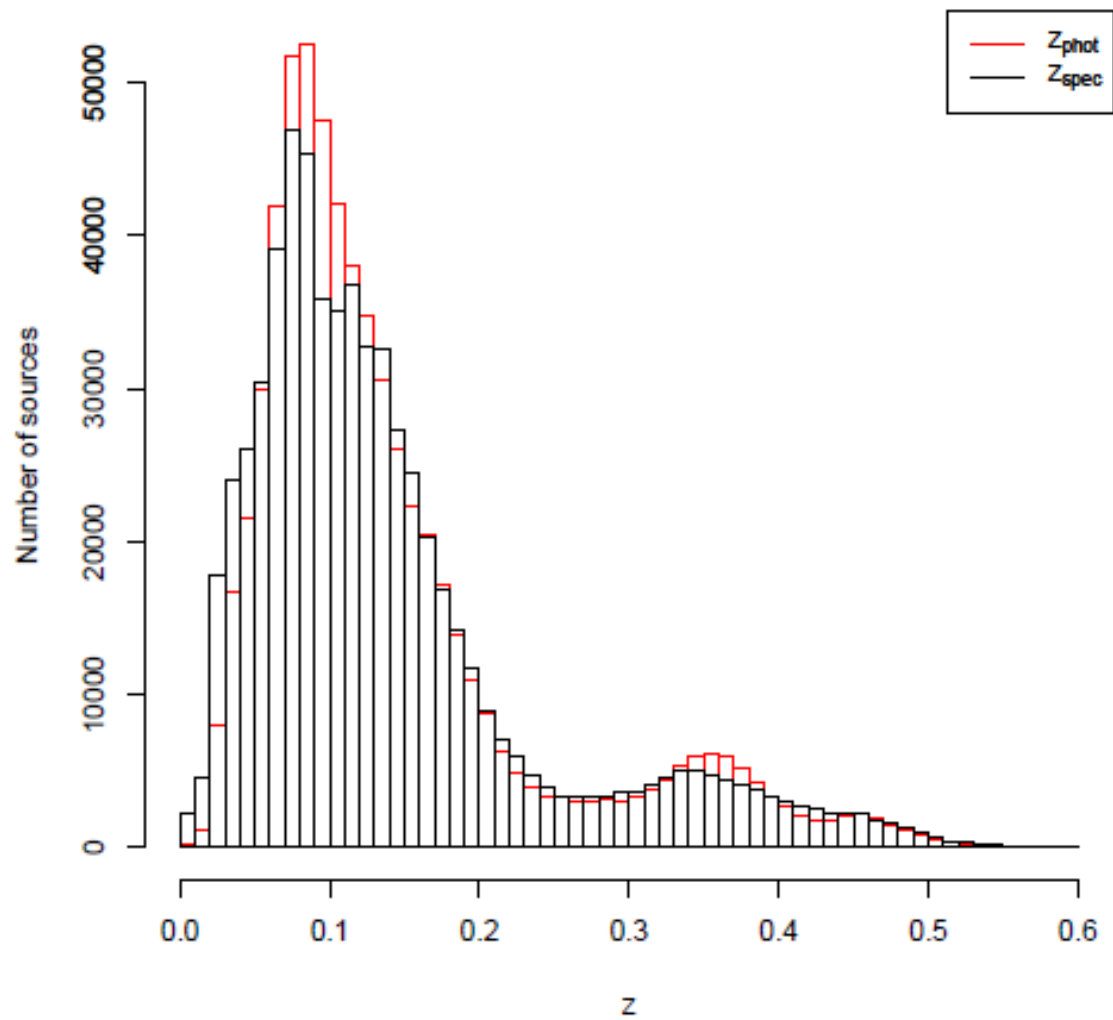
# BUT ... LET'S GO BACK TO PHOT-Z

## Photometric redshifts: the method



# Galaxies





**Figure 4.** Histograms of the distribution of spectroscopic and photometric redshifts for the sample of SDSS galaxies with optical photometry, belonging to the KB used to train the WGE in the first experiment.



# QSO candidates experiment

SDSS only

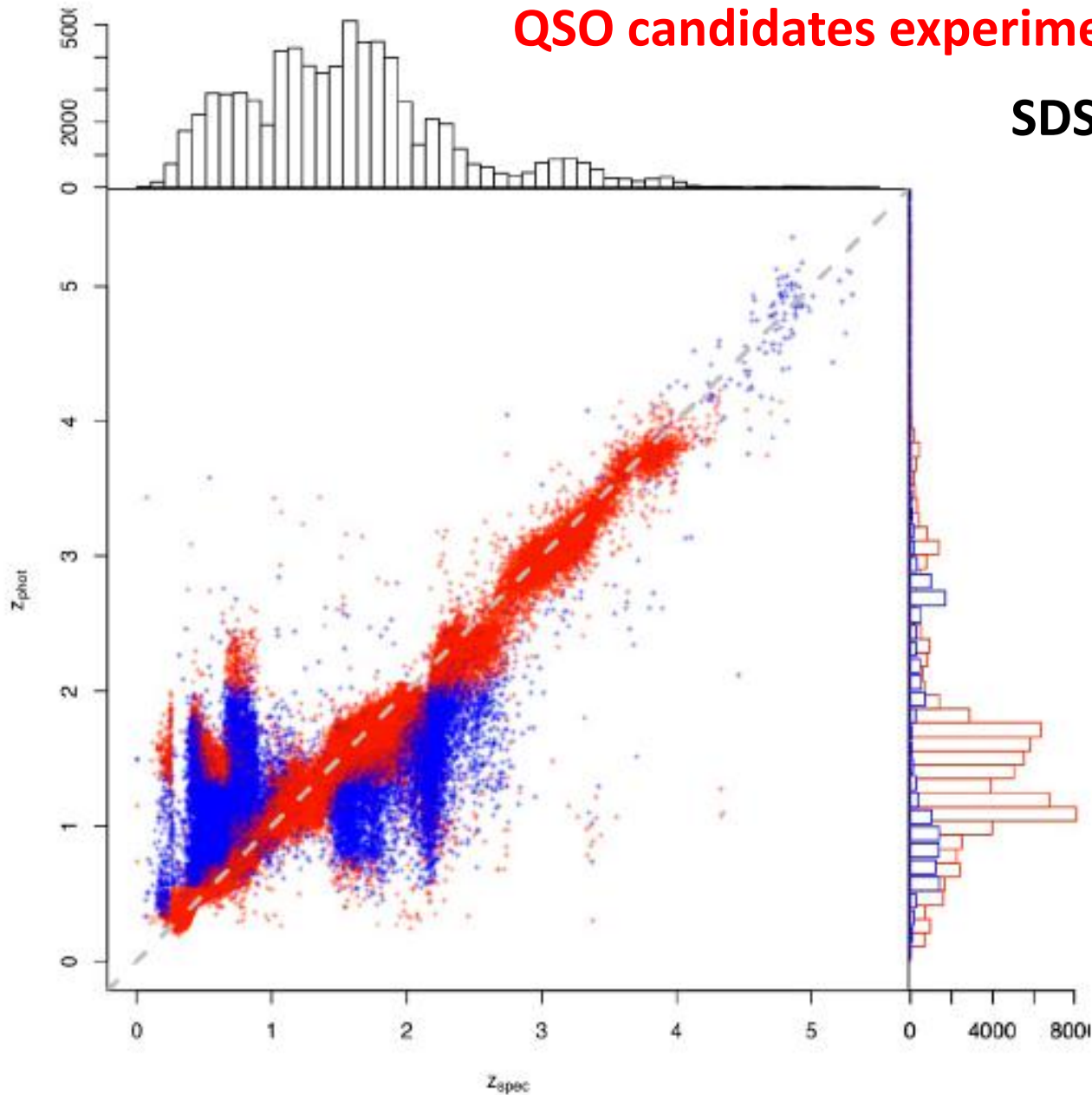
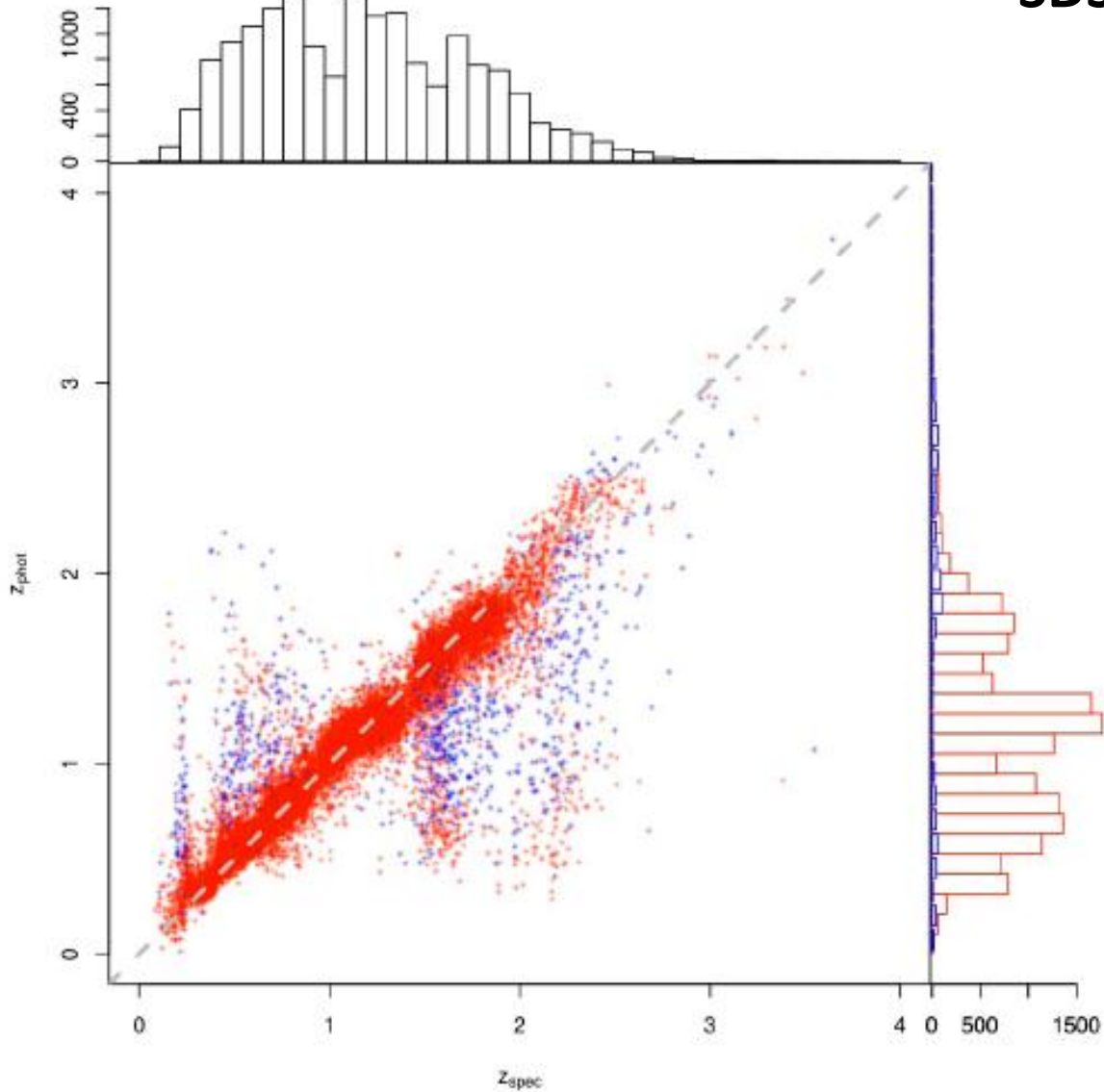
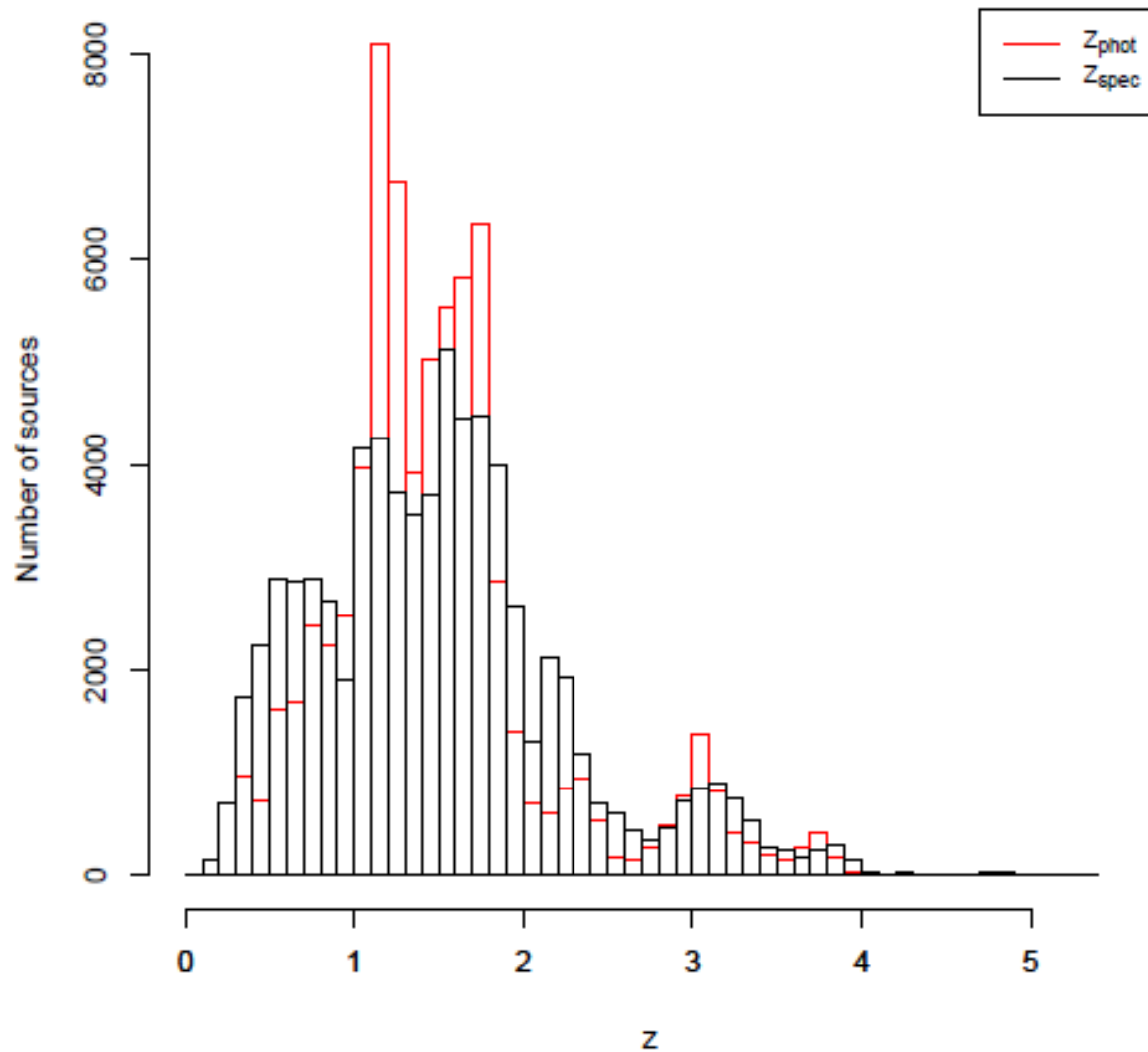


Figure 16. Scatterplot of the spectroscopic vs photometric redshifts for the KB of the second experiment (quasars with optical photometry), with marginal histograms for reliable and unreliable photometric redshift estimates according to the quality flag  $q$ .

## SDSS + Galex



**Figure 18.** Scatterplot of the spectroscopic vs photometric redshifts for the KB of the third experiment (quasars with optical and ultraviolet photometry), with marginal histograms for reliable and unreliable photometric redshift estimates according to the quality flag  $q$ .



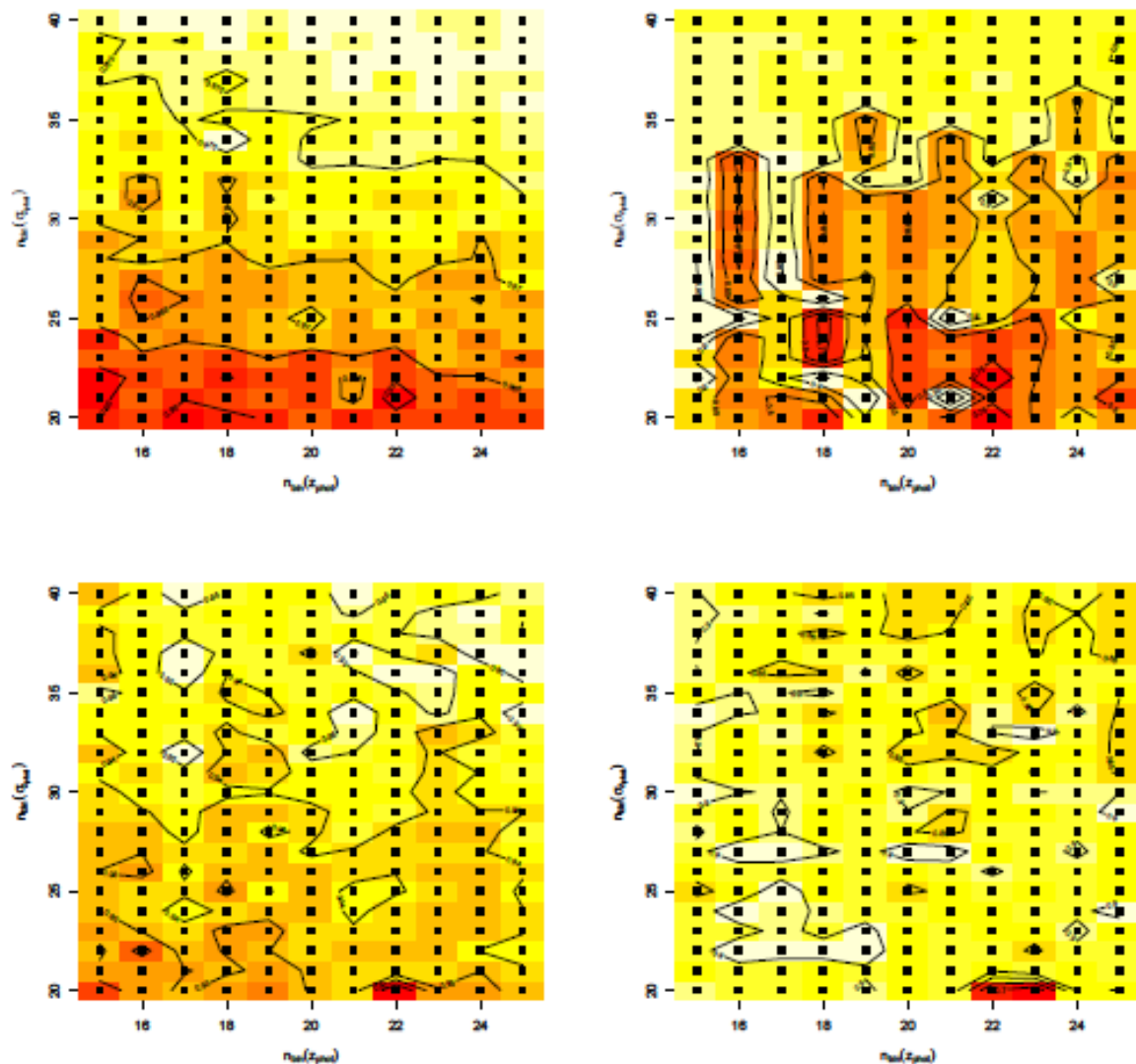
**Figure 6.** Histograms of the distribution of spectroscopic and photometric redshifts for the sample of SDSS quasars with optical photometry, belonging to the KB used to train the WGE in the second experiment.

## Errors taken into account:

- **Input noise**: error propagation on the input parameter (Ball et al. 2008)
- **Model variance**: different models make differing predictions (Collister & Lahav 2004)
- **Model bias**: different models are affected by different biases.
- **Target noise**: in some regions of the parameter space, data may represent poorly the relation between featured and targets (*Laurino 2009*).

**Table 4.** Statistical diagnostics of photometric redshifts reconstruction for all the experiments discussed in this paper and for relevant papers in the literature. The first column (Exp. 1) contains the diagnostics for the experiment for the determination of the photometric redshifts of the optical galaxies from the SDSS catalog described in paragraph 6.1, while the columns (Exp. 2) and (Exp. 3) describe the diagnostics for the experiments concerning the determination of the photometric redshifts for quasars with optical and optical+ultraviolet photometry respectively (more details in paragraphs 6.2 and 6.3. Some of the same statistical diagnostics are shown for some significant papers from the literature, respectively (D’Abrusco et al. 2007) for optical galaxies in column (1), (Ball et al. 2008) for both optical and optical+ultraviolet quasars in the columns indicated as (2) and (Richards et al. 2009) for Other results are discussed in more details in the section 9.

Diagnostic	Exp. 1	(1)	Exp. 2	(2)	(3)	Exp. 3	(2)	(3)
$\langle \Delta z \rangle$	0.015	0.021	0.21	-	-	0.13	-	-
RMS( $\Delta z$ )	0.021	0.074	0.35	-	-	0.25	-	-
$\sigma^2(\Delta z)$	$2.9 \cdot 10^{-4}$	$5.0 \cdot 10^{-4}$	0.08	0.123	0.27	0.044	0.054	0.136
MAD( $\Delta z$ )	0.011	0.012	0.11	-	-	0.061	-	-
%( $\Delta z < 0.01, < 0.1$ )	43.4	41.1	50.7	54.9	63.9	68.1	70.8	74.9
%( $\Delta z < 0.02, < 0.2$ )	72.4	68.4	72.3	73.3	80.2	86.5	85.8	86.9
%( $\Delta z < 0.03, < 0.3$ )	86.9	83.4	80.5	80.7	85.7	91.4	90.8	91.0
$\sigma^2(\Delta z < 0.01, < 0.1)$	$8.2 \cdot 10^{-6}$	$8.2 \cdot 10^{-6}$	$7.9 \cdot 10^{-4}$	-	-	$7.6 \cdot 10^{-4}$	-	-
$\sigma^2(\Delta z < 0.02, < 0.2)$	$3.0 \cdot 10^{-5}$	$3.1 \cdot 10^{-5}$	0.003	-	-	0.023	-	-
$\sigma^2(\Delta z < 0.03, < 0.3)$	$6.1 \cdot 10^{-5}$	$6.3 \cdot 10^{-5}$	0.005	-	-	0.039	-	-
$\langle \Delta z_{\text{norm}} \rangle$	0.014	0.017	0.095	0.095	0.115	0.058	0.06	0.071
RMS( $\Delta z_{\text{norm}}$ )	0.019	0.037	0.19	-	-	0.11	-	-
$\sigma^2(\Delta z_{\text{norm}})$	$1.8 \cdot 10^{-4}$	$1.1 \cdot 10^{-3}$	0.025	0.034	0.079	0.086	0.014	0.031
MAD( $\Delta z_{\text{norm}}$ )	0.009	0.011	0.041	-	-	0.029	-	-
%( $\Delta z_{\text{norm}} < 0.01, < 0.1$ )	48.3	45.6	77.3	-	-	87.4	-	-
%( $\Delta z_{\text{norm}} < 0.02, < 0.2$ )	77.2	73.5	87.3	-	-	94.0	-	-
%( $\Delta z_{\text{norm}} < 0.03, < 0.3$ )	90.1	87.0	91.8	-	-	96.4	-	-
$\sigma^2(\Delta z_{\text{norm}} < 0.01, < 0.1)$	$8.3 \cdot 10^{-6}$	$8.2 \cdot 10^{-6}$	$6.2 \cdot 10^{-4}$	-	-	$5.6 \cdot 10^{-4}$	-	-
$\sigma^2(\Delta z_{\text{norm}} < 0.02, < 0.2)$	$3 \cdot 10^{-5}$	$3.0 \cdot 10^{-5}$	0.002	-	-	0.001	-	-
$\sigma^2(\Delta z_{\text{norm}} < 0.03, < 0.3)$	$5.8 \cdot 10^{-5}$	$6.0 \cdot 10^{-5}$	0.004	-	-	0.002	-	-



**Figure 15.** Plots of the efficiency (left column) and of the completeness (right column) of the process of selection of the catastrophic outliers as functions of the two parameters  $n_{\text{bin}}(z_{\text{phot}})$  and  $n_{\text{bin}}(\sigma_{z_{\text{phot}}})$  involved in the procedure for the determination of the quality flag  $q$ . The upper plots are associated to the experiment for the evaluation of the photometric redshifts for the optical SDSS quasars, while the lower plots are associated to the third experiment for the estimation of the photometric redshifts of the SDSS quasars with optical and ultraviolet photometry.

**Table 5.** Accuracy of the reconstruction of the photometric redshifts for the three experiments described in this paper as a function of the number of sources composing the KBs. Robust estimates of the robust standard deviation of the  $\Delta z$  variable, obtained with the MAD algorithm are provided together with the percentages of sources with  $\Delta z < 0.3$  and  $\Delta z < 0.03$  for the experiments involving the quasars and the galaxy respectively.

Cardinality KB	Exp. 1	$\sigma_{rob}$		%( $\Delta z < 0.03, < 0.3$ )		
		Exp. 2	Exp. 3	Exp. 1	Exp. 2	Exp. 3
$5 \cdot 10^2$	0.035	0.392	0.201	68.3	60.3	79.2
$10^3$	0.027	0.245	0.167	71.1	70.1	85.6
$5 \cdot 10^3$	0.019	0.181	0.102	82.9	74.2	91.6
$10^4$	0.018	0.165	0.100	83.2	78.4	90.4
$5 \cdot 10^4$	0.017	0.143	-	86.3	81.6	-
$10^5$	0.018	-	-	87.6	-	-
$5 \cdot 10^5$	0.018	-	-	88.9	-	-
Whole KB	0.017	0.143	0.089	90.1	79.4	91.3

# Pro's of the WGE method

Method	Dataset	Variance	$\frac{\sigma^2}{1+z}$	$\mu\left(\frac{\Delta z}{1+z}\right)$	$\% \Delta_{0.1}$	$\% \Delta_{0.2}$	$\% \Delta_{0.3}$
<i>k</i> NN	S	<b>0.123</b>	<b>0.034</b>	0.095	54.9	73.3	80.7
<i>k</i> NNPDF	S	–	–	–	53.8	72.4	79.8
CZR	S	0.265	0.079	0.115	<b>63.9</b>	<b>80.2</b>	<b>85.7</b>
WGE	S	0.142	0.059	0.032	48.8	70.3	78.9
WGE+err	S	<b>0.133</b>	0.056	<b>0.025</b>	48.7	71.4	80.4
<i>k</i> NN	SG	<b>0.054</b>	<b>0.014</b>	0.060	70.8	85.8	90.8
<i>k</i> NNPDF	SG	–	–	–	71.8	86.4	90.8
CZR	SG	0.136	0.031	0.071	<b>74.9</b>	<b>86.9</b>	91.0
WGE	SG	0.058	0.030	0.022	67.9	85.2	91.1
WGE+err	SG	0.057	0.029	<b>0.012</b>	69.3	86.2	<b>91.3</b>

- **WGE provides errors and flag outliers:** it is trained to recognize distinct regimes in both  $z_{\text{phot}}$  and  $\sigma_{z_{\text{phot}}}$ ;
- **Scalability:** WGE is able to crunch very large datasets with limited computational resources;

- **Fast training:** WGE readily improves to the data rate of very large throughputs;
- **WGE is versatile:** fits well with different sources and with general regression and classification problems;
- **WGE is general:** can combine different methods (not based on data mining).



# Conclusions I

- Machine learning is a powerful tool which even now may outperform traditional approaches, even more so with the very large datasets of the future. **But:**
  - The implementation of successful methods requires a strict interaction between mathematicians, computer scientists and **domain experts**, and it may be a lengthy procedure
  - Requires (large to very large) computing infrastructures (*horses not chickens...*)