Astronomical data Mining:

An application to the photometric redshifts of galaxies and QSOs

Giuseppe Longo

University Federico II in Napoli, longo@na.infn.it

In coll. with

M. Brescia, R. D'Abrusco, O.Laurino & the ***DAME*** team

IPAC-Pasadena, August 5 2009

The company which is making the journey…

University Federico II

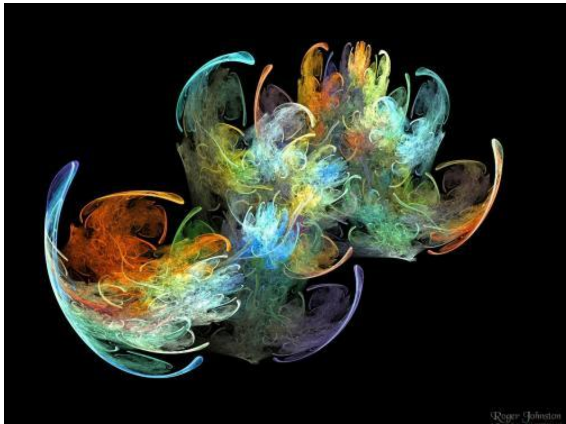
- Massimo Brescia (project manager)
- Stefano Cavuoti
- Raffaele D'Abrusco
- Giancarlo D'Angelo (GRID)
- Natalia V. Deniskina
- Michelangelo Fiore (student)
- Mauro Garofalo
- Omar Laurino (project engineer)
- Giuseppe Longo (Principal Investigator)
- Francesco Manna (student)
- Alfonso Nocella
- Civita Vellucci (student)

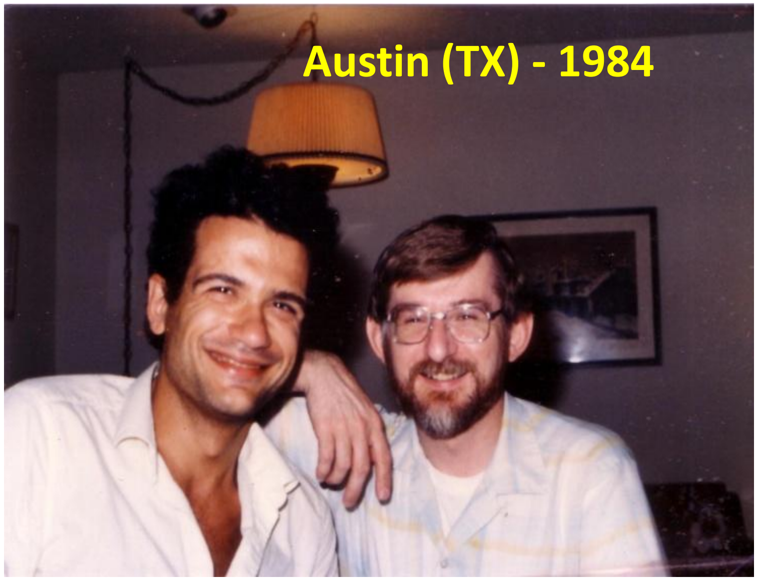
Caltech

- G.S. Djorgovski
- C. Donalek
- A. Mahabal

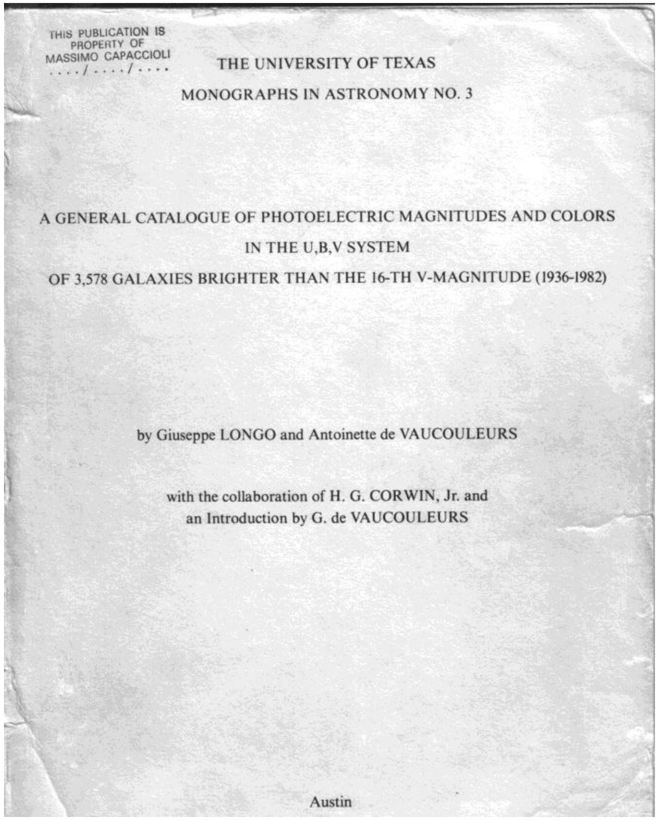
Summary of the talk

- Data Mining and astronomy
  - Why DAME and what is DAME
    - Photometric redshifts and galaxy phot-z's in DAME
      - A DM "pipeline" for QSO's (candidate selection and phot-z's)
        - Some general considerations on the future





Most of us have done it for their whole life


THIS PUBLICATION IS PROPERTY OF MASSIMO CAPACCIOLI

THE UNIVERSITY OF TEXAS

MONOGRAPHS IN ASTRONOMY NO. 3

A GENERAL CATALOGUE OF PHOTOELECTRIC MAGNITUDES AND COLORS IN THE U,B,V SYSTEM OF 3,578 GALAXIES BRIGHTER THAN THE 16-TH V-MAGNITUDE (1936-1982)

by Giuseppe LONGO and Antoinette de VAUCOULEURS

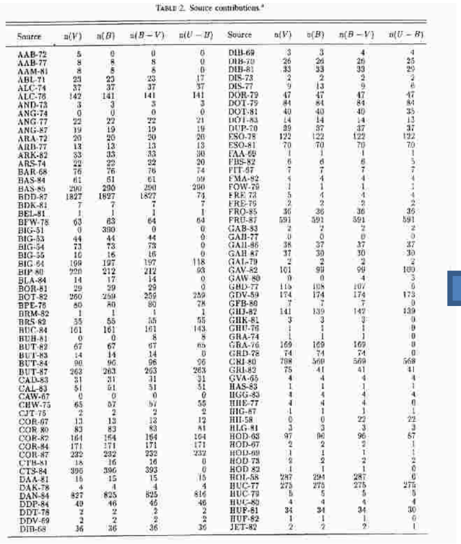
with the collaboration of H. G. CORWIN, Jr. and an Introduction by G. de VAUCOULEURS

Austin

Compilation of photoelectric multiaperture photometry

Through standard luminosity profile curves to derive "Extrapolation corrections"

…. in order to derive Total Magnitudes of galaxies



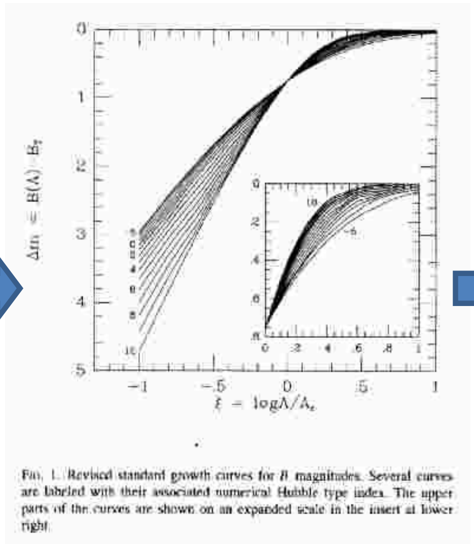
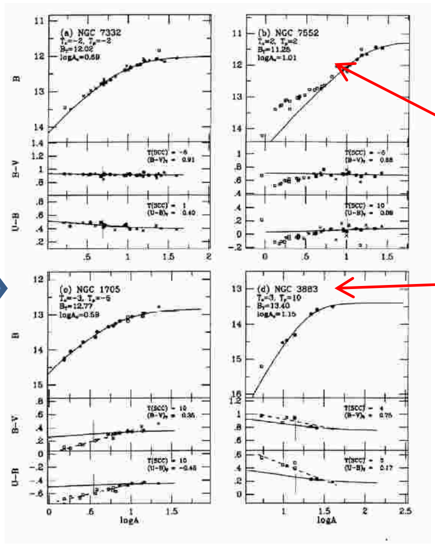


Fig. 1. Revised standard growth curves for B magnitudes. Several curves are labeled with their associated numerical Hubble type index. The upper parts of the curves are shown on an expanded scale in the insert at lower right.



Knowledge

Data

Base of Knowledge (BoK)

Model

Data Mining is not only new astronomy.

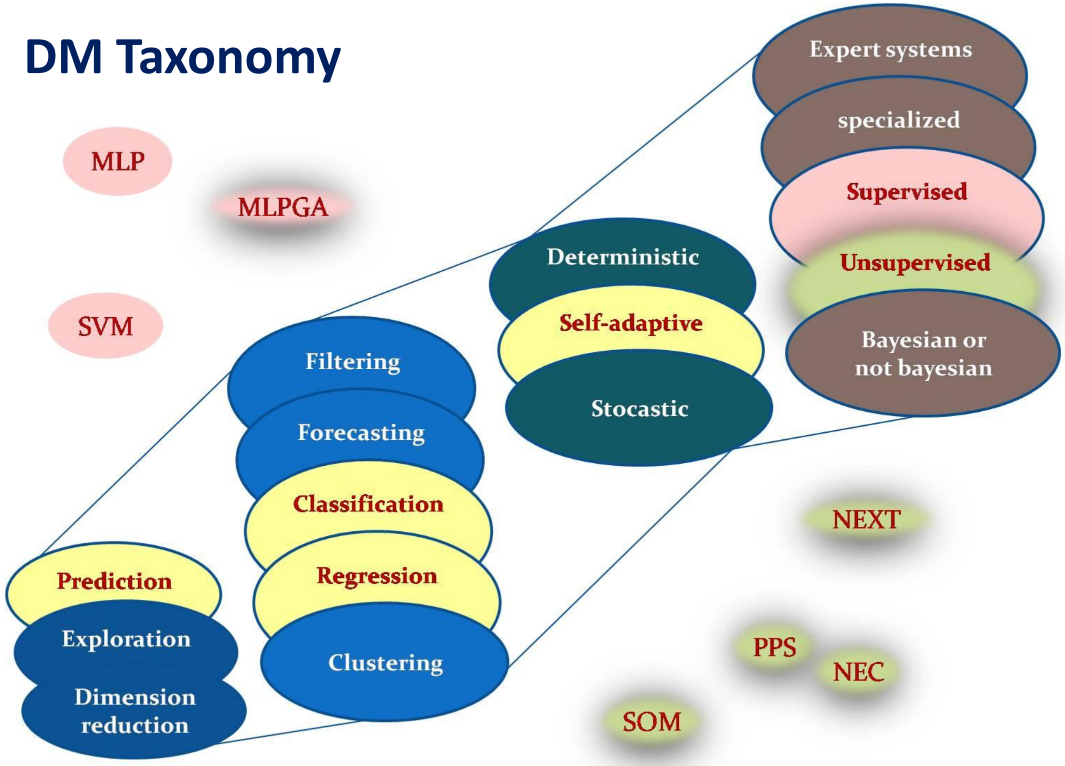
In many cases (***but NOT ALL***) it is just the name we give to rather usual stuff when it needs to be performed fast and on billions of records of COMPLEX data

Human brain is not sufficient

Machine learning methods

Data Mining is the activity of extracting **USEFUL** information from **COMPLEX** data using Statistical Pattern Recognition and Machine Learning methods.

DM Taxonomy



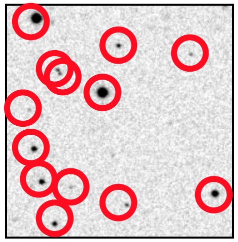
1. To catalogue the known (classification)
2. Characterize the unknown (clustering)
3. Find functional dependencies (regression)
4. Find exceptions (outliers)

Supervised Methods

Patterns are learnt from extensive set of templates (Base of Knowledge = BoK)

Unsupervised Methods

Patterns are discovered using the data themselves



p={isophotal, petrosian, aperture magnitudes concentration indexes, shape parameters, etc.}

$$p^1 = \{RA^1, \delta^1, t, \{\lambda_1, \Delta\lambda_1, f_1^{1,1}, \Delta f_1^{1,1}, \ldots, f_1^{1,m}, \Delta f_1^{1,m}\}, \ldots, \{\lambda_n, \Delta\lambda_n, f_n^{1,1}, \Delta f_n^{1,1}, \ldots, f_n^{1,m}, \Delta f_n^{1,m}\}\}$$

$$p^2 = \{RA^2, \delta^2, t, \{\lambda_1, \Delta\lambda_1, f_1^{2,1}, \Delta f_1^{2,1}, \ldots, f_1^{2,m}, \Delta f_1^{2,m}\}, \ldots, \{\lambda_n, \Delta\lambda_n, f_n^{2,1}, \Delta f_n^{2,1}, \ldots, f_n^{2,m}, \Delta f_n^{2,m}\}\}$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$p^N = \{RA^N, \delta^N, t, \{\lambda_1, \Delta\lambda_1, f_1^{N,1}, \Delta f_1^{N,1}, \ldots, f_1^{N,m}, \Delta f_1^{N,m}\}, \ldots\}$$

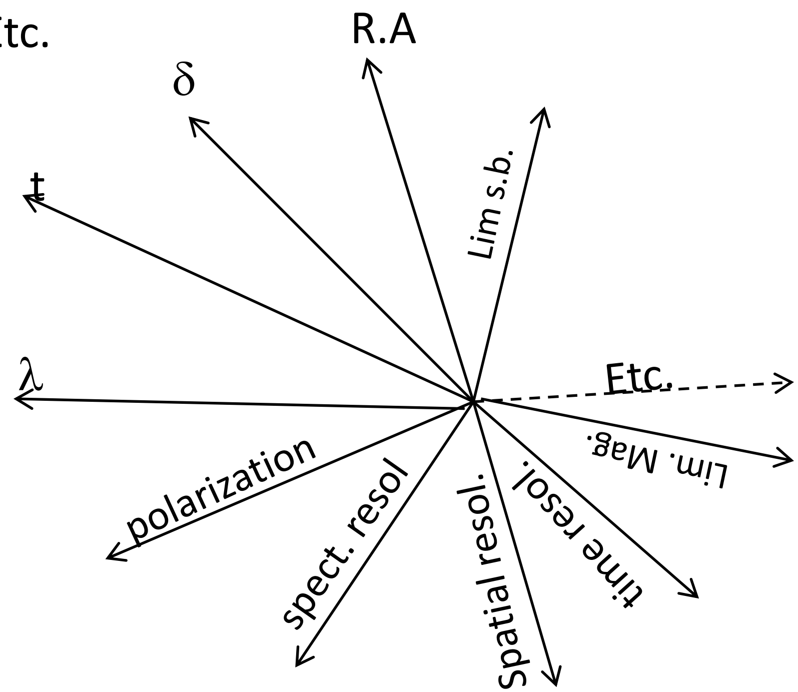
$$D = 3 + m \times n$$

The scientific exploitation of a multi band, multiepoch (K epochs) universe implies to search for patterns, trends, etc. among N points in a DxK dimensional parameter space:

N >10^9, D>>100, K>10

Any observed (simulated) datum *p* defines a point (region) in a subset of R^N. Es:

- RA and dec
- time
- λ
- experimental setup (spatial and spectral resolution, limiting mag, limiting surface brightness, etc.) parameters
- fluxes
- polarization
- Etc.



$$p \in \Re^N \qquad N >> 100$$

The parameter space concept is crucial to:

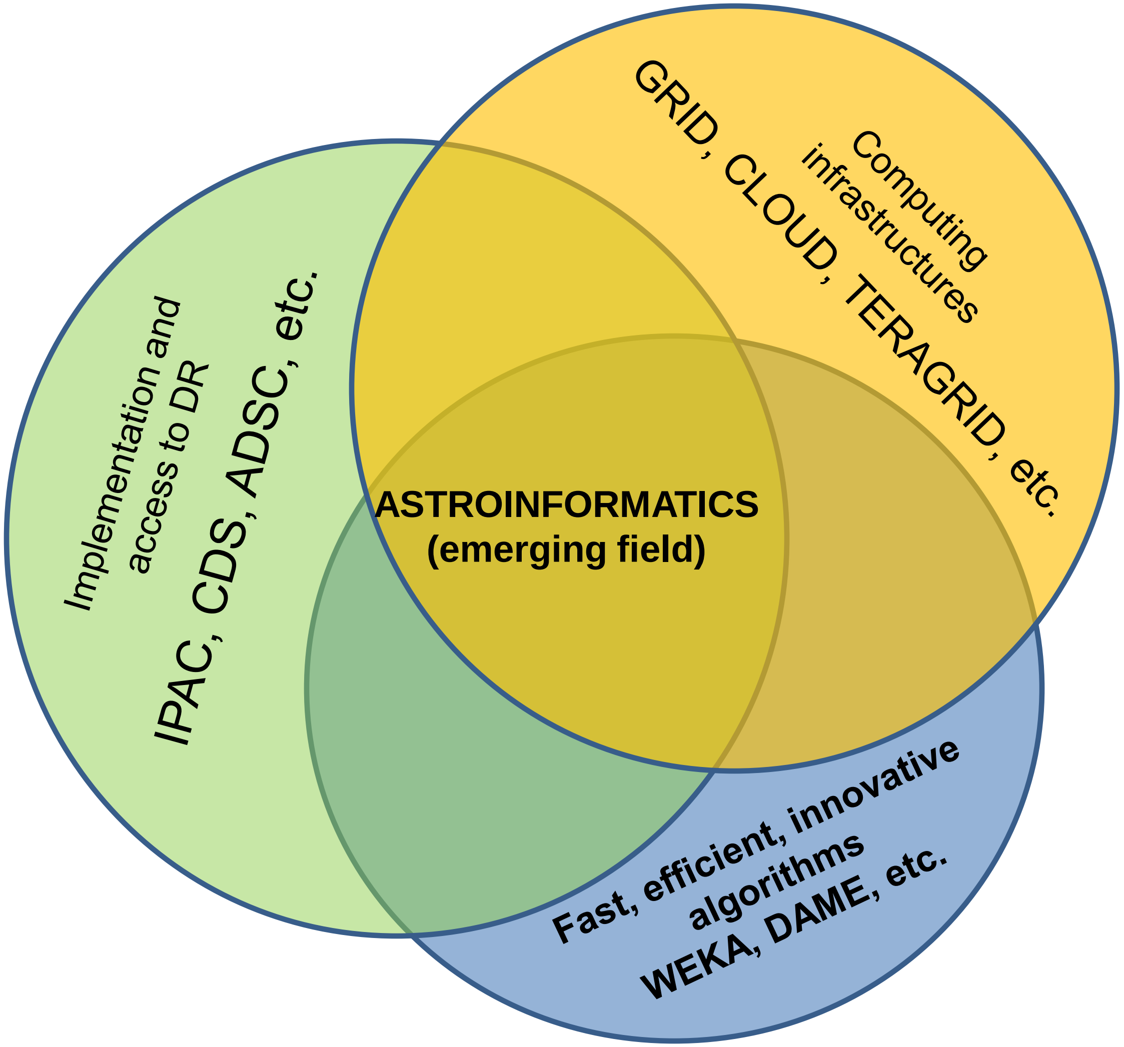
1. Guide the quest for new discoveries (observations can be guided to explore poorly known regions), …
2. Find new physical laws (patterns)
3. Etc,

The computational cost of DM:

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 Maximisation: $K \times N \times I \times \mathbf{D^2}$
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 > ~ $(NxD)^3$

ASTROINFORMATICS



Machine Learning problems as "function approximation"

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

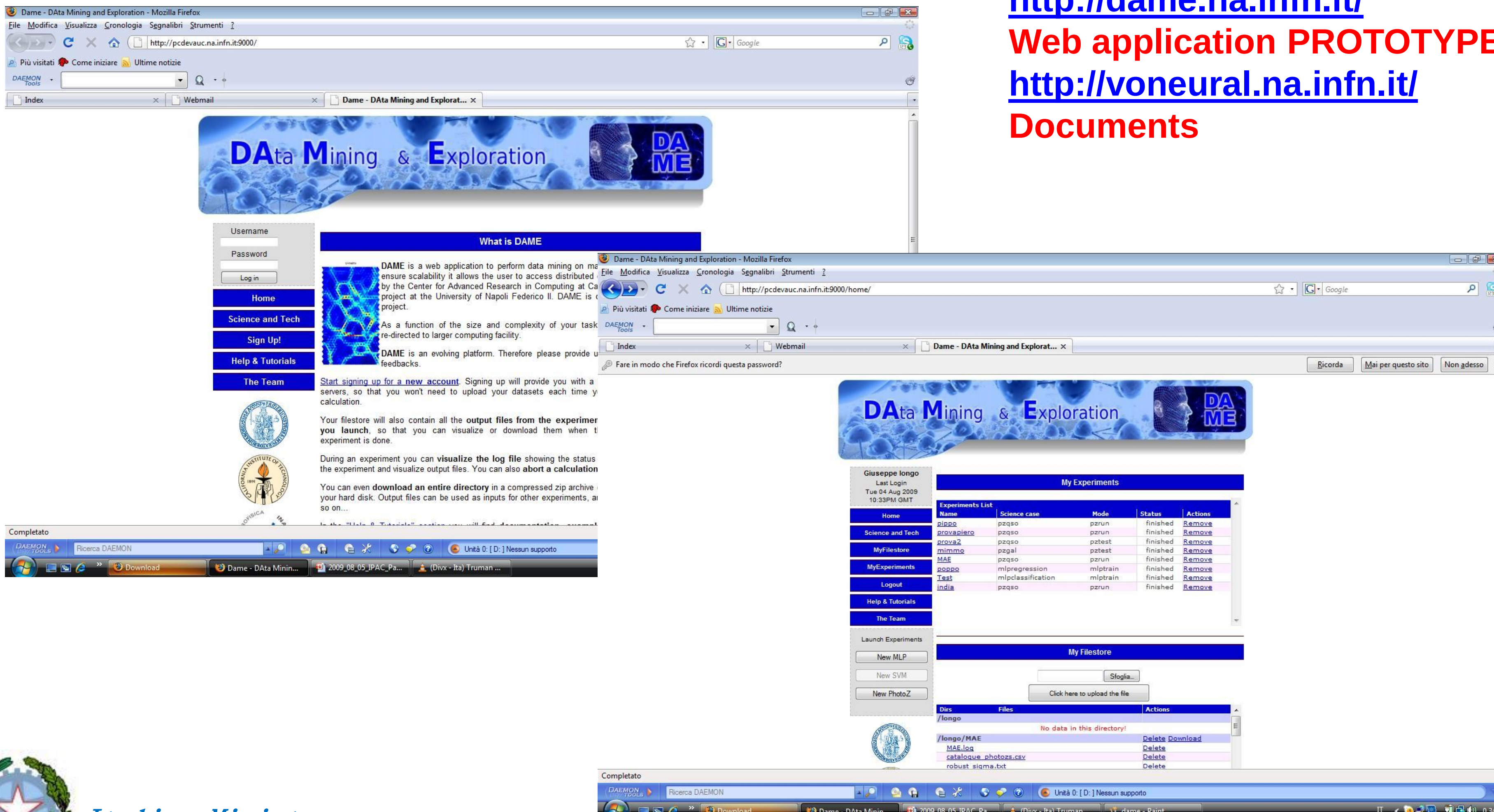
$$\mathbf{Y} \equiv \{x_1, x_2, x_3, \ldots x_M\} \quad \text{target vectors } M << N$$

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

variable	characteristics	Type	Operation/example
Quantitative	Numerical with ordering relationship and possibility to define a metric	Actual measurement	**Regression** Photometric redshifts
Categorical (non ordered)	Membership into a finite umber of classes. No ordering relationship.	Numerical codes (targets) arbitrarily ordered	**Classification** Search for peculiar objects, QSO's, Star/galaxy, etc.
Ordered categorical	Classes ordered by a relationship but there is no metric	Numerical codes non arbitrarily ordered	**Classification** Morphological and physical classification of galaxies, etc.

Prototype by O. Laurino

DAME is a joint effort between University Federico II, INAF and Caltech aimed at: implementing (as web application) a suite of data exploration, data mining and data visualization tools.



http://dame.na.infn.it/

Web application PROTOTYPE

http://voneural.na.infn.it/

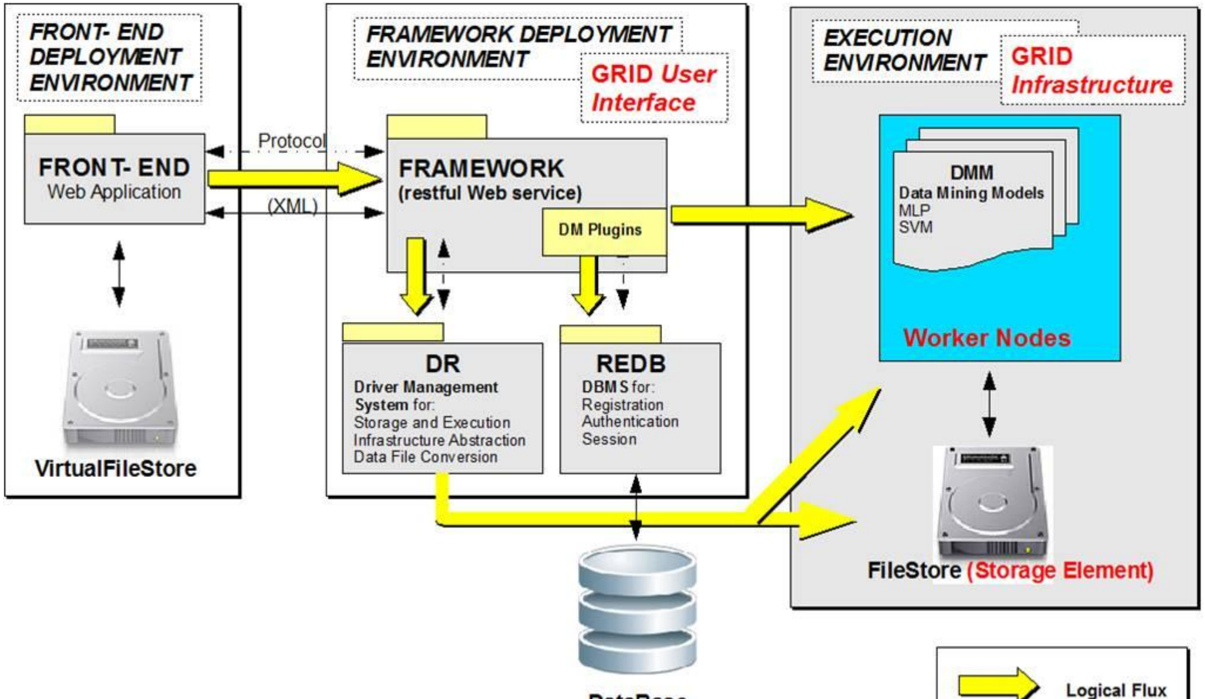
Documents

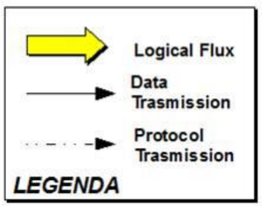
Italian Ministry of Foreign Affairs

What is the real DAME

1. The real thing

P.M. Massimo Brescia



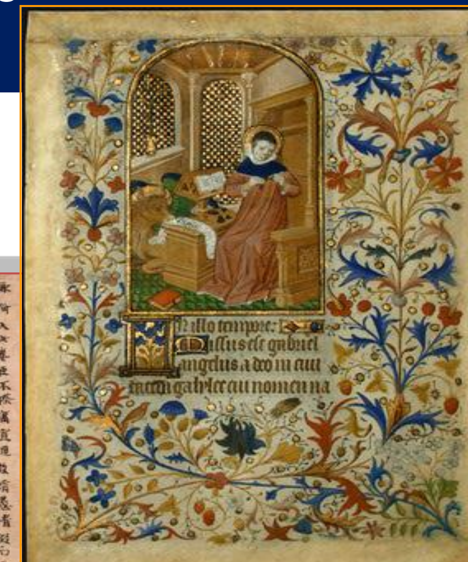


1. **User friendly**
2. **Flexible and expandable**
3. **Running also on HPC or distributed systems**

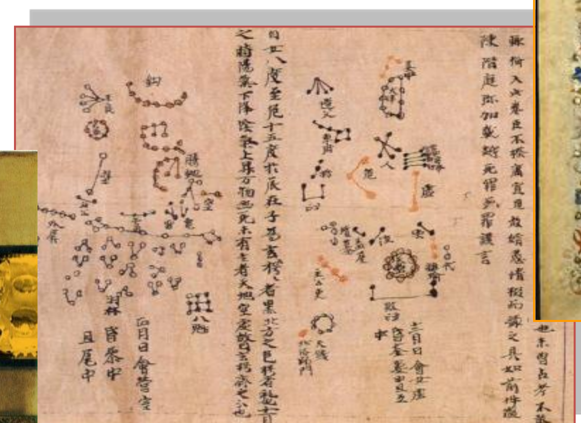
Will substitute the prototype at the end of October 2009

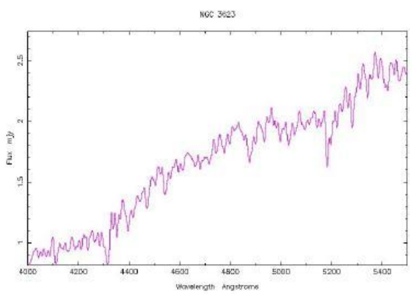
PART II - applications of DAME to observational cosmology

Photometric redshifts of galaxies and QSO's
Selection of candidate quasars



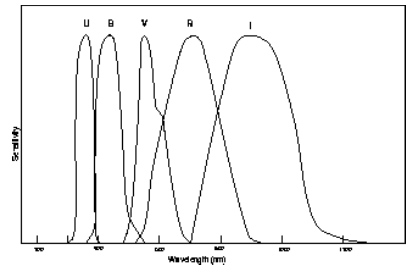
D'Abrusco et al. 2007, ApJ, 663, pp. 752-764
D'Abrusco et al. 2009, MNRAS, 396, 223-262
Laurino et al., 2009, Thesis
Laurino et al., 2009, MNRAS, in preparation





Galaxy spectrum - F(λ)

X



Photometric system - $S_i(\lambda)$

=

$$m_U = -2.5\log_{10}\frac{\int F(\lambda)S_U(\lambda)d\lambda}{\int S_U(\lambda)d\lambda} + c_u$$

$$m_B = -2.5\log_{10}\frac{\int F(\lambda)S_B(\lambda)d\lambda}{\int S_B(\lambda)d\lambda} + c_B$$

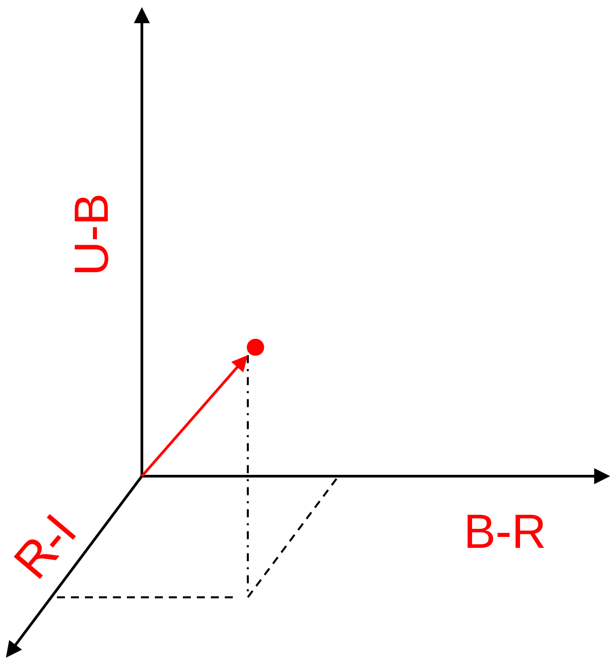
Etc…

Color indexes

$$U - B \equiv m_U - m_B$$

$$B - R \equiv m_B - m_R$$

$$etc.$$



Photometric redshifts are always a function approximation hence a DM problem:

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

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

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

BoK (set of templates) → Mapping function → Knowledge (phot-z's)

Observed Spectroscopic Redshifts

Synthetic colors from theoretical SEDs
Synthetic colors from observed SED's
.....

Knowledge always reflects the biases in the BoK.

Interpolative
Uneven coverage of parameter space

SED fitting
Unknown or oversimplified physics
Unjustified assumptions
.....

Data used in the science cases:

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

UKIDDS: overlap with SDSS

GALEX: overlap with SDSS

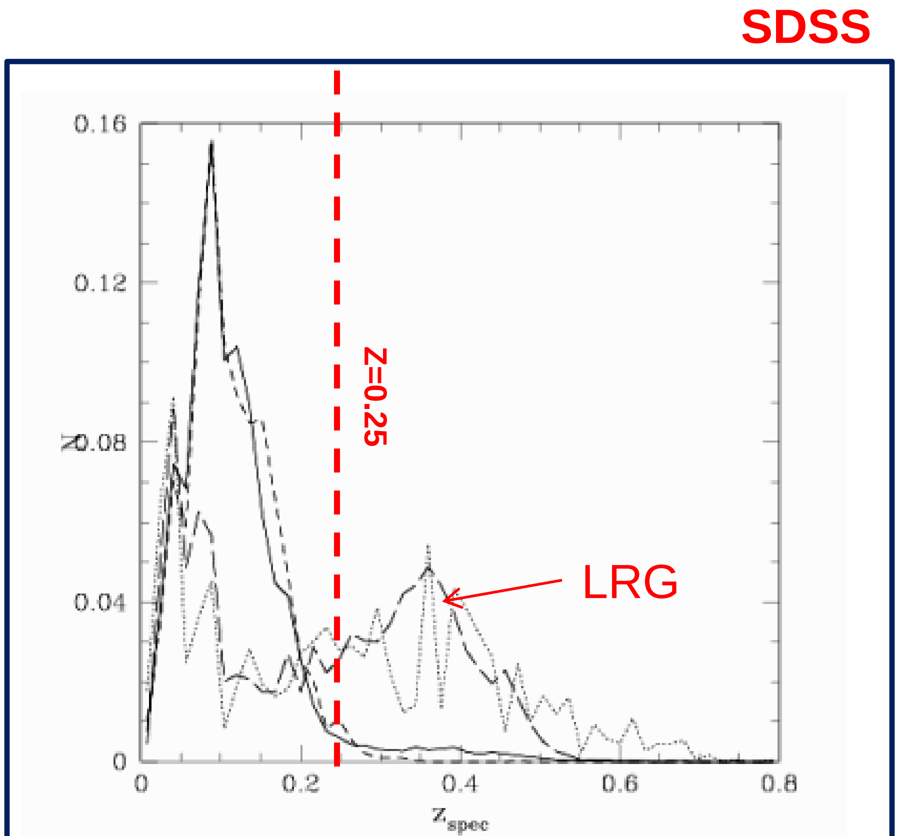
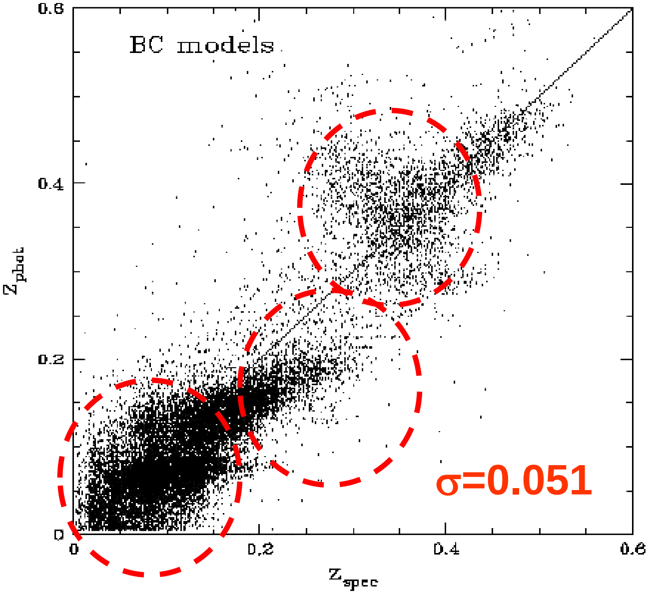
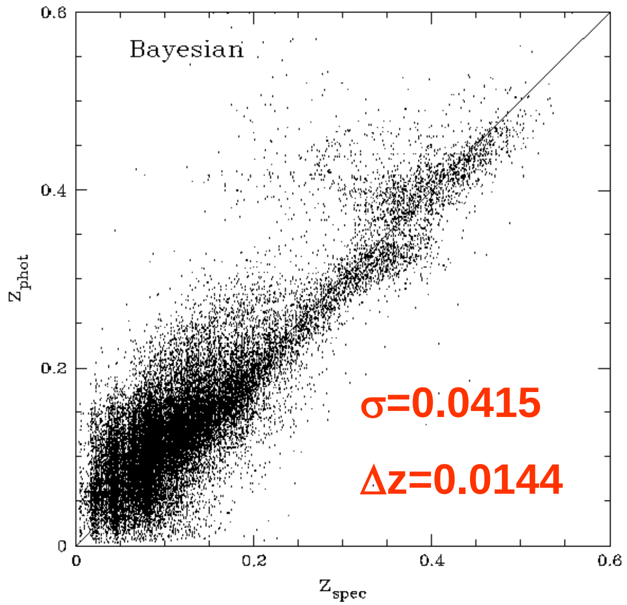


Fig. 1.— The spectroscopic redshift histogram for the SDSS main EDR (solid), the EDR LRG (long dash), the 2dF (short dash) and the CNOC2 sets.



SED fitting

Templates from synthetic colors obtained from theoretical SED's
Mapping function from simple interpolation



Interpolative

Templates from synthetic colors obtained from theoretical SED's
Mapping function from Bayesian inference

- the color space is partitioned (KD-tree - a binary search tree) into cells containing the same number of objects from the training set
- In each cell a second order polynomial is fit to BoK.

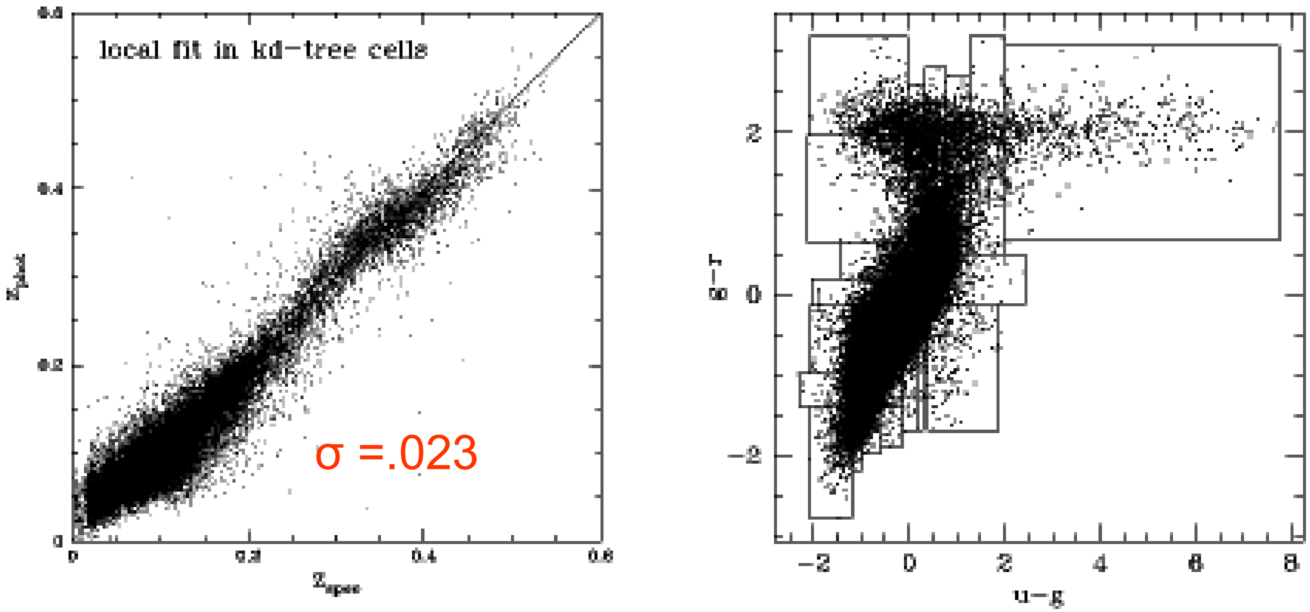
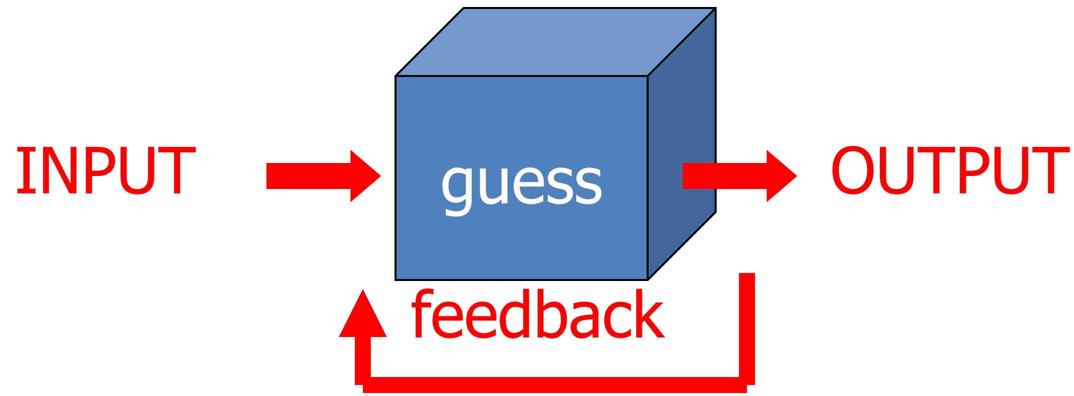
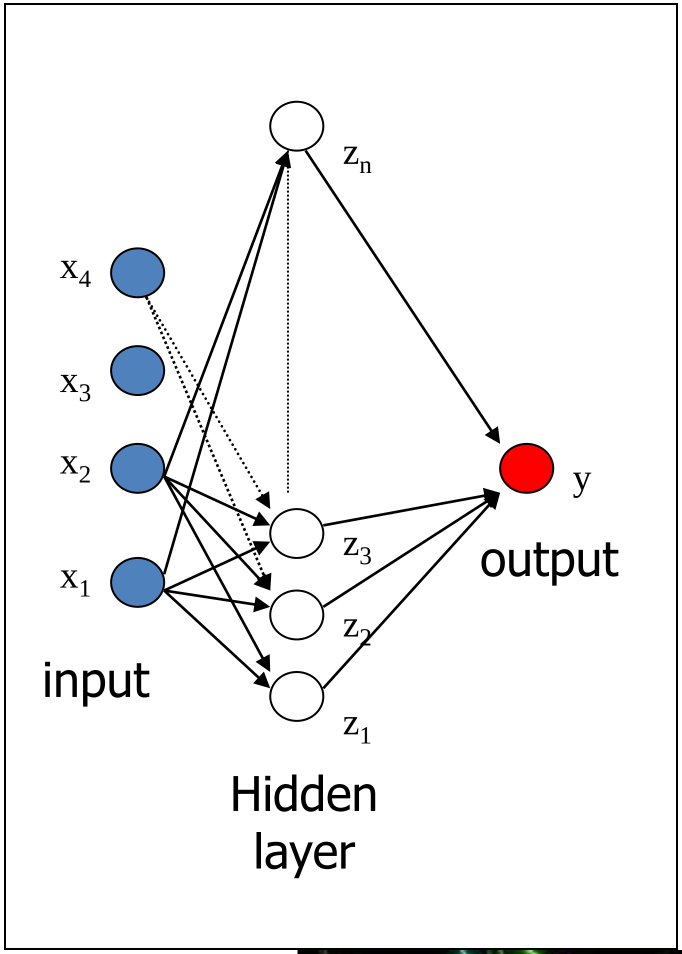


Fig. 4.— On the right we plot a 2 dimensional demonstration of the color space partitioning. In each of these cells we applied the polynomial fitting technique to estimate redshifts. The left figure show the results.

MLP or Multi Layers Perceptron

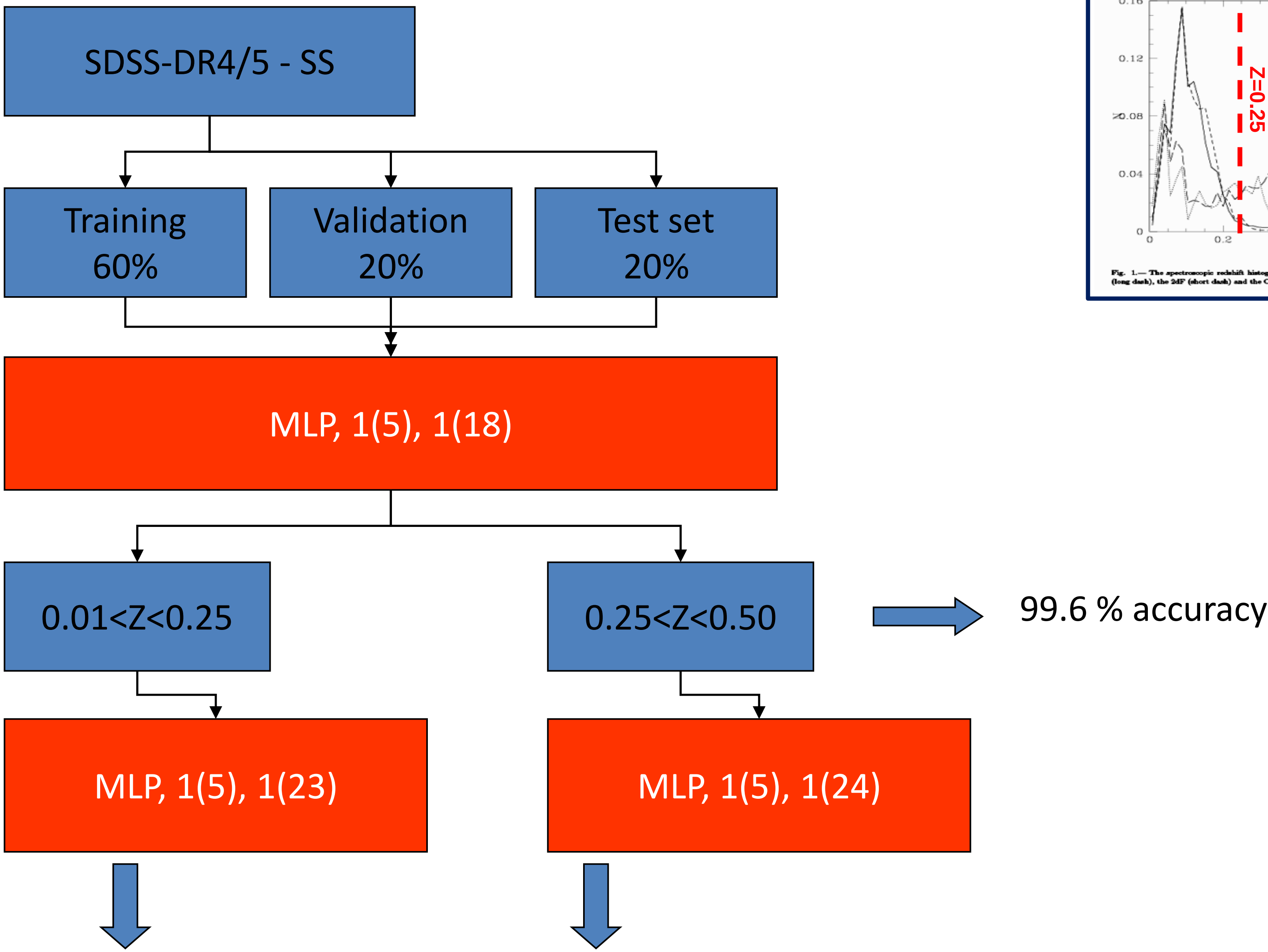




- input layer (n neurons)
- M hidden layer (1 or 2)
- Output layer (n' <n neurons)

Neurons are connected via activation functions

Different NN's given by different topologies, different activation functions, etc.



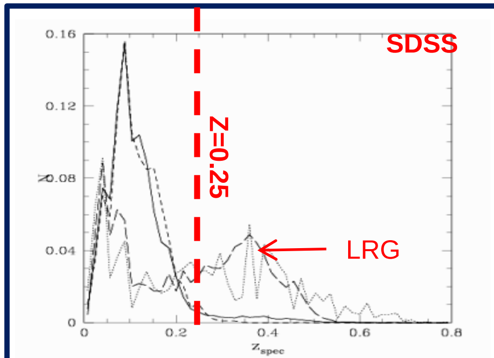
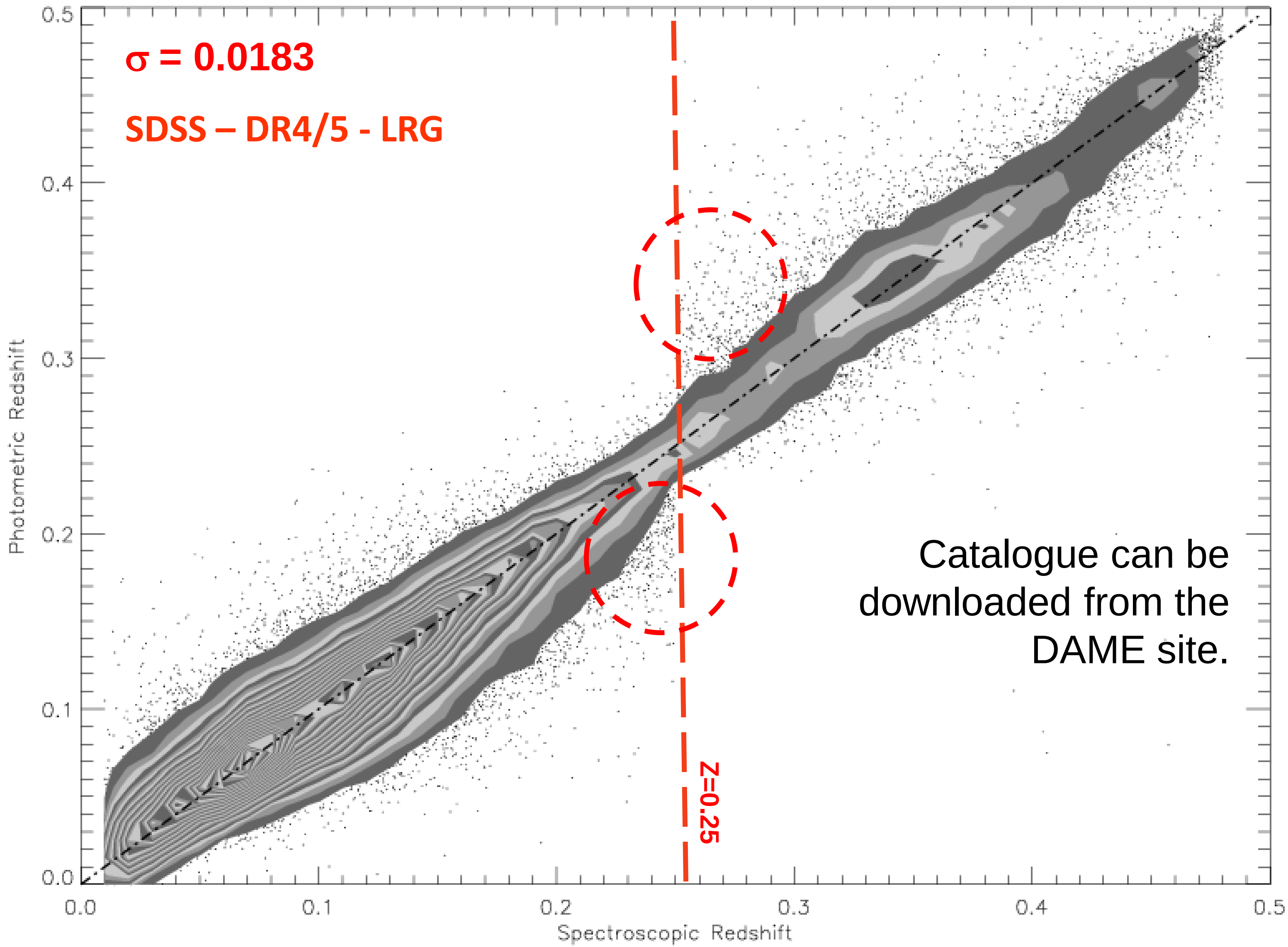


Fig. 1.— The spectroscopic redshift histogram for the SDSS main EDR (solid), the EDR LRG (long dash), the 2dF (short dash) and the CNOC2 sets.



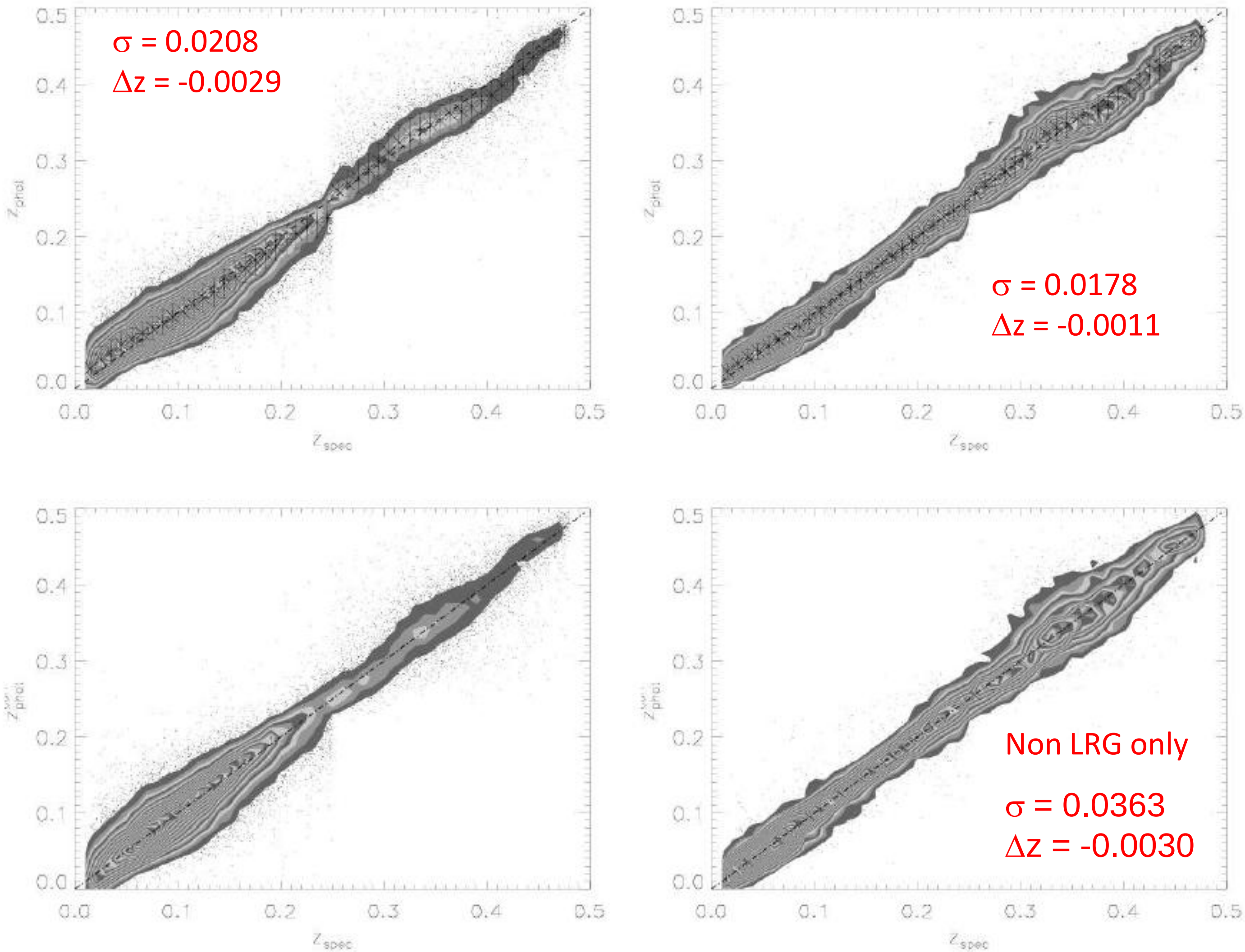
Catalogue can be downloaded from the DAME site.

D'Abrusco et al. 2007

type	method	data	Δz_{rms}	Notes	Reference
	CWW	EDR	0.0666		(Csabai et al. 2003)
SEDF	Bruzual-CHarlot	EDR	0.0552		(Csabai et al. 2003)
	Interpolated	EDR	0.0451		(Csabai et al. 2003)
	Polyomial	EDR	0.0318		(Csabai et al. 2003)
	KD-tree	EDR	0.0254		(Csabai et al. 2003)
	ANNz	EDR	0.0229		(Collister & Lahav 2004)
ML	SVM	EDR	0.027		(Wadadekar 2004)

General galaxy sample

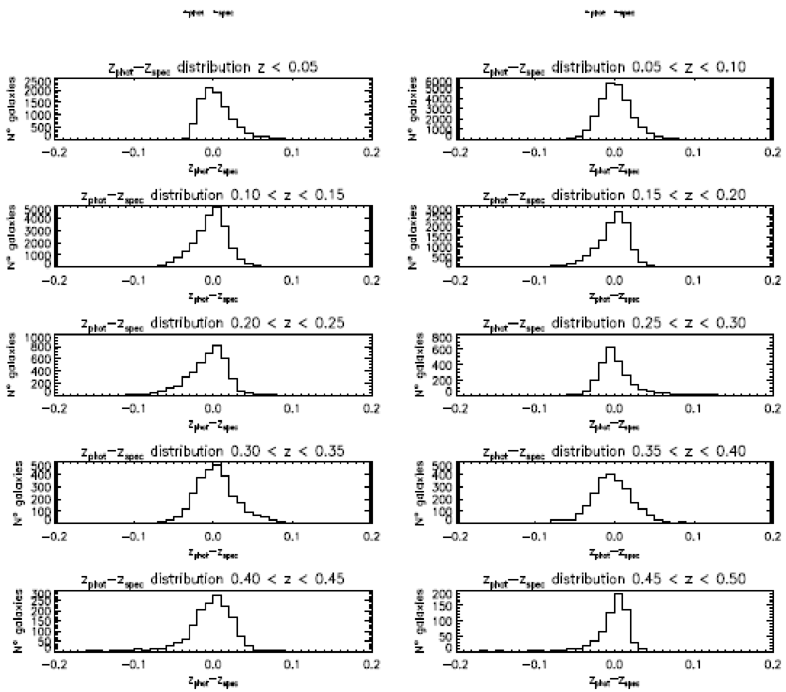
LRG sample



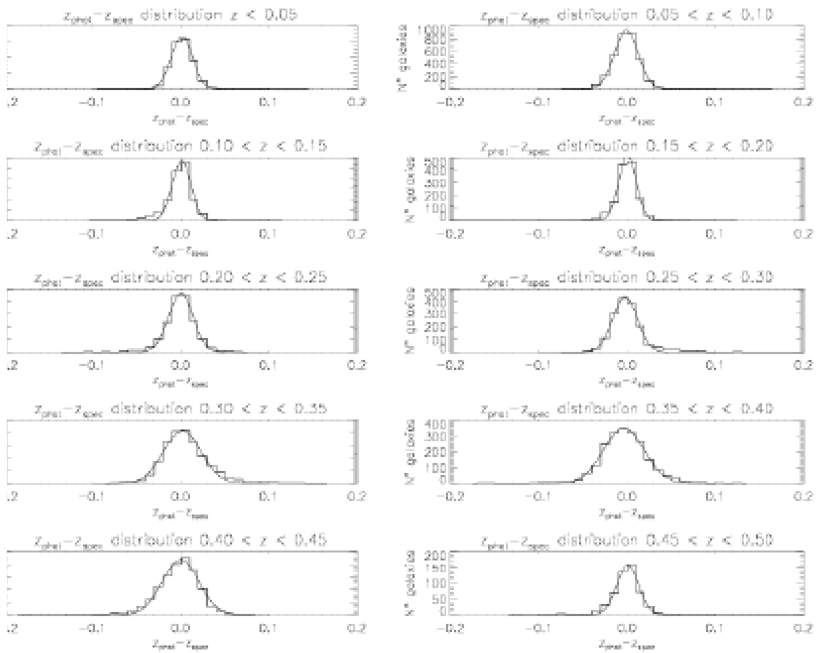
D'Abrusco et al. 2007

Science with Dame

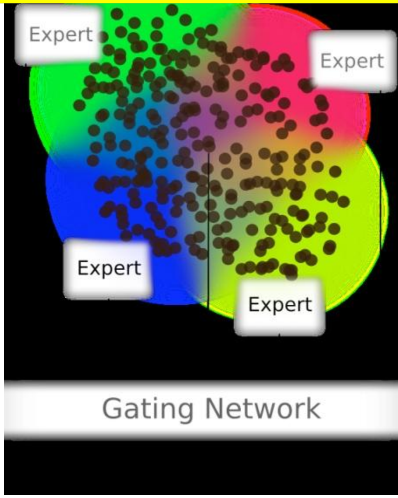
1. Photometric redshifts of galaxies



General galaxy sample



LRG sample



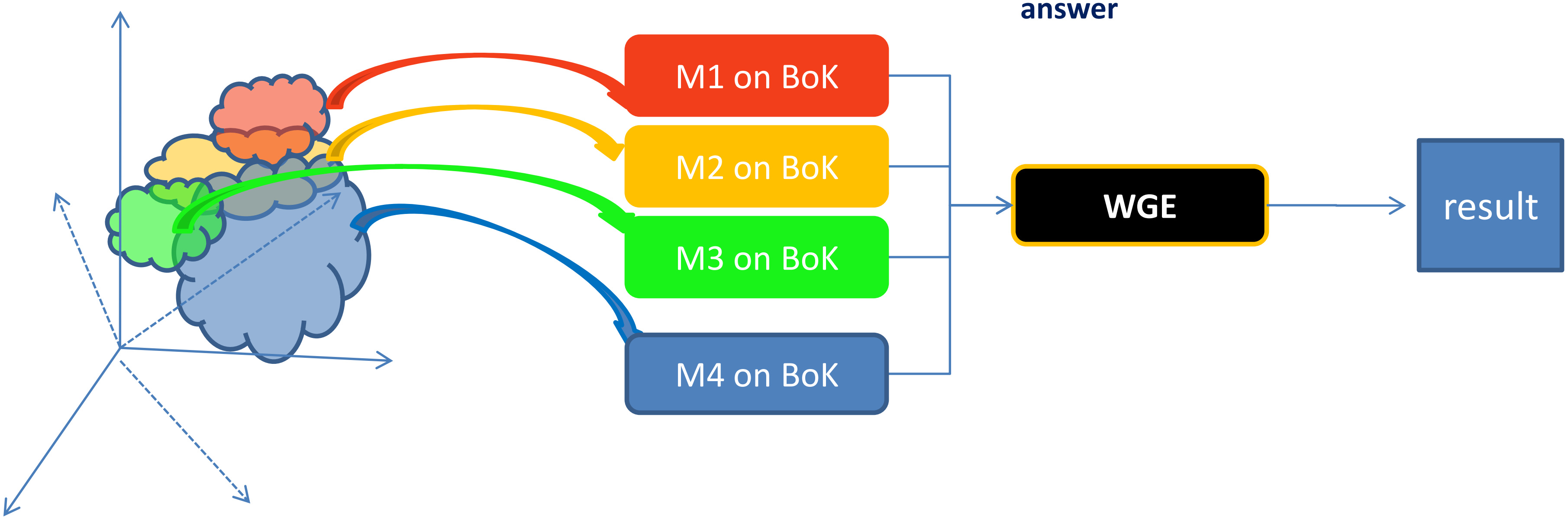
What do we learn if the BoK is biased:

- At high z LRG dominate and interpolative methods are not capable to "generalize" rules
- An unique method optimizes its performances on the parts of the parameter space which are best covered in the BoK

Step 1:
unsupervised clustering in parameter space

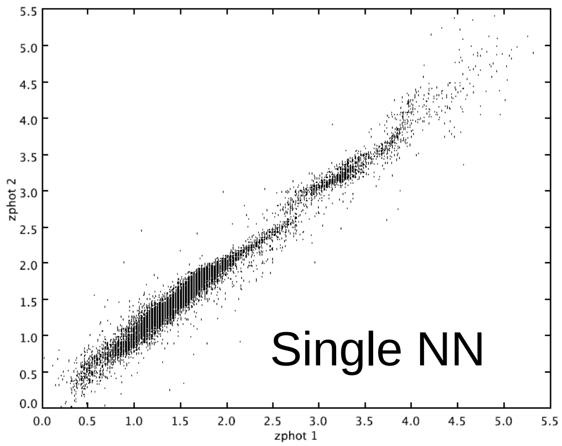
Step 2:
supervised training of different NN for each cluster

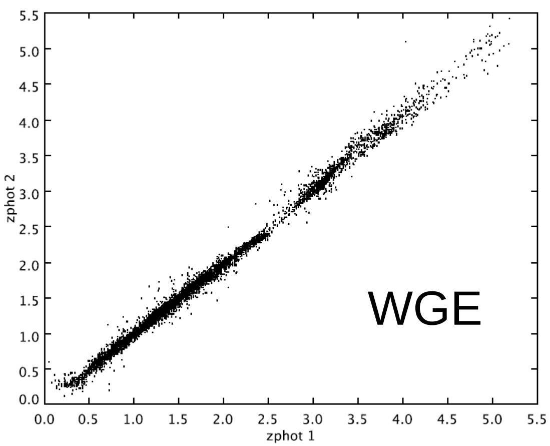
Step 3:
output of all NN go to WGE which learns the correct answer

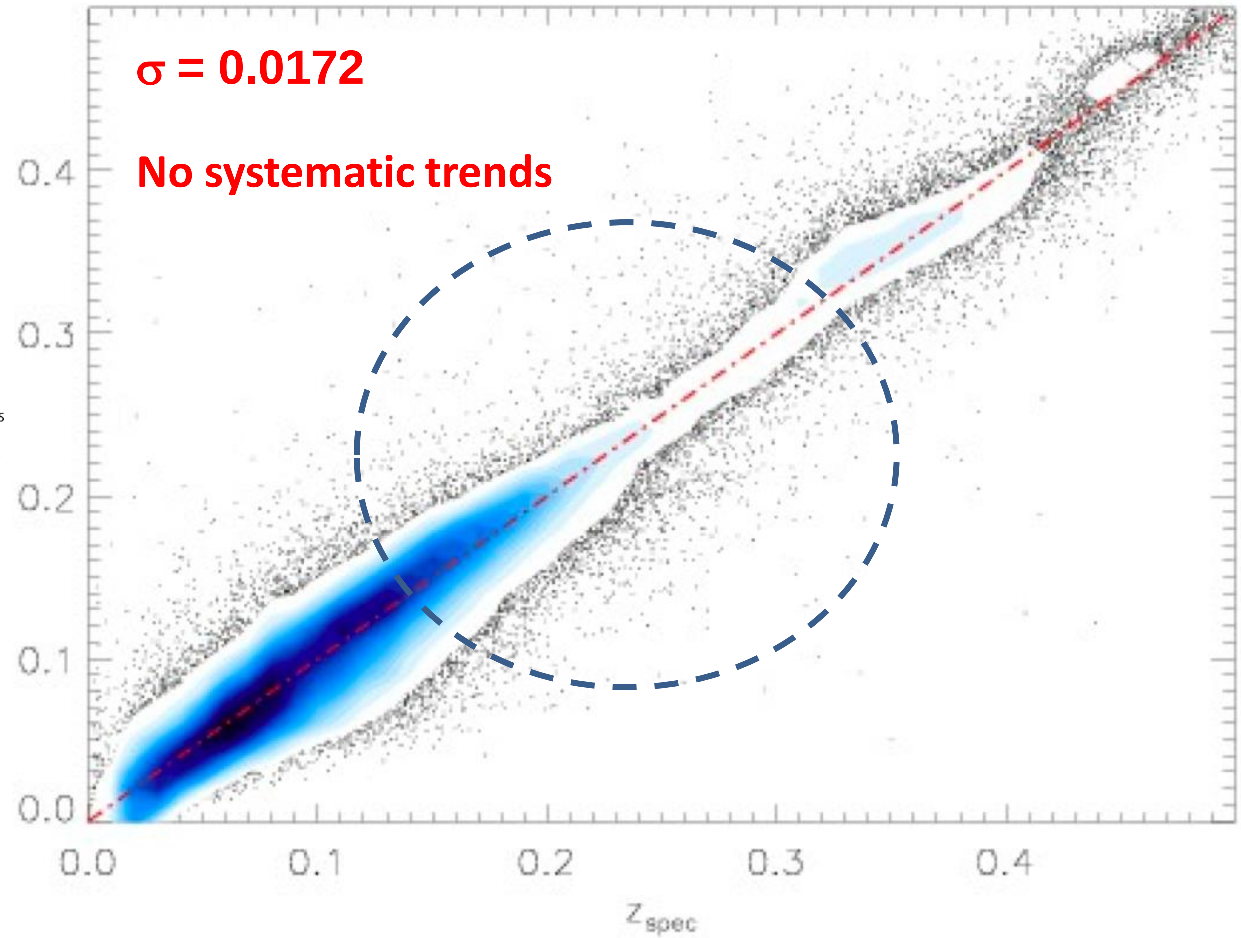


Laurino et al. 2009a,2009b

Laurino et al. 2009a,2009b





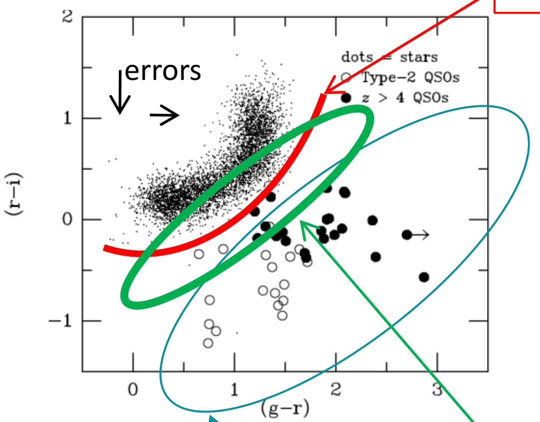


PART II - applications to observational cosmology

Photometric selection of candidate QSO's
(as a clustering problem)

Traditional way to look for candidate QSO in 3 band survey

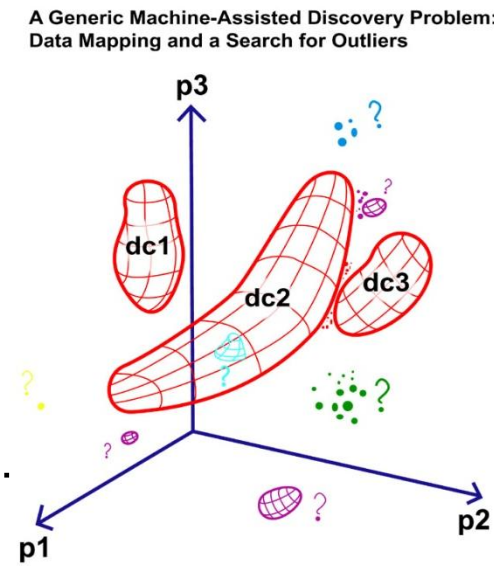
Cutoff line

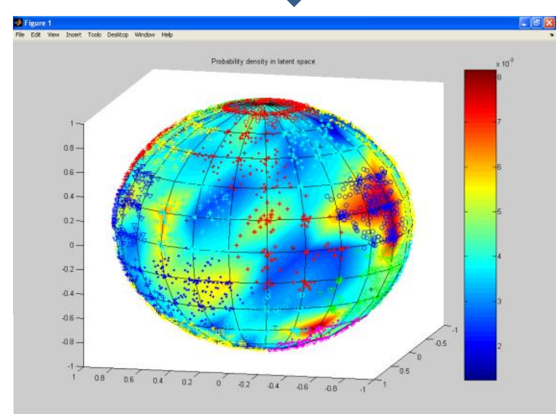


Candidate QSOs for spectroscopic follow-up's

Ambiguity zone

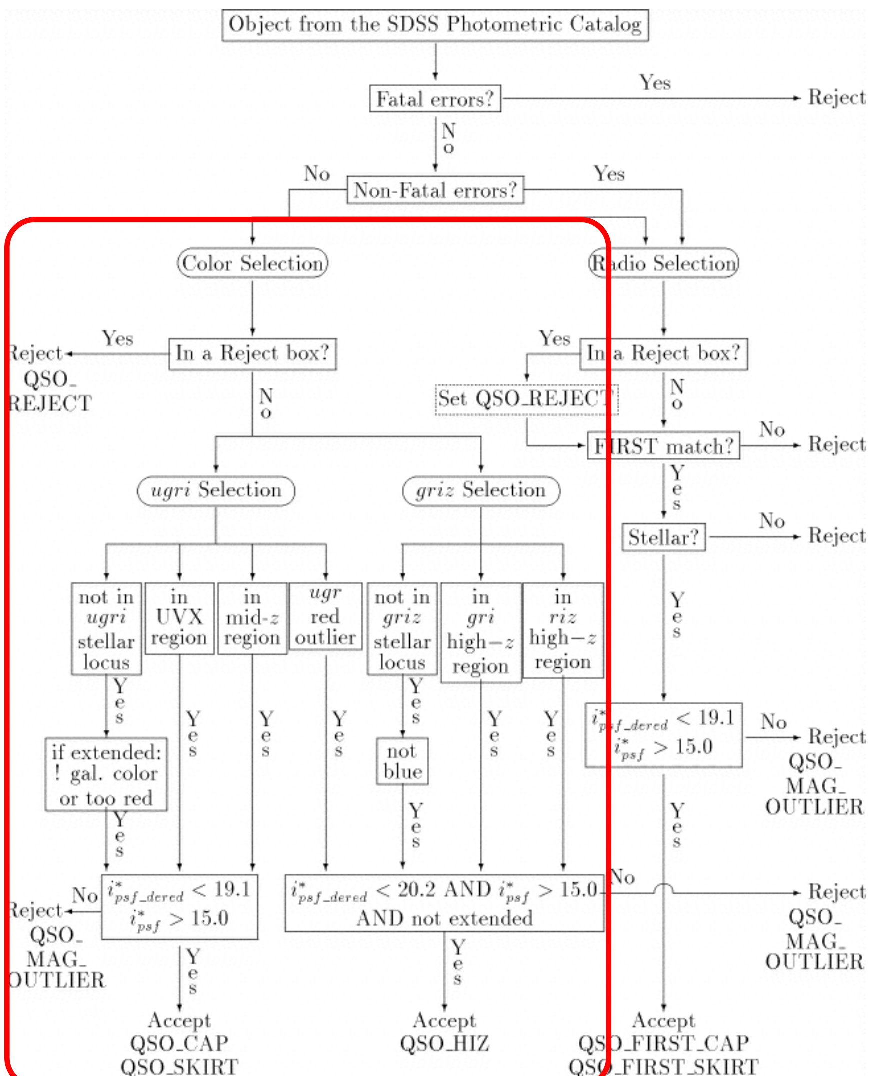
Adding one feature improves separation…





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

SDSS QSO candidate selection algorithm (Richards et al, 2002) targets star-like objects as QSO candidate according to their position in the SDSS colours space (u-g,g-r,r-i,i-z), if one of these requirements is satisfied:

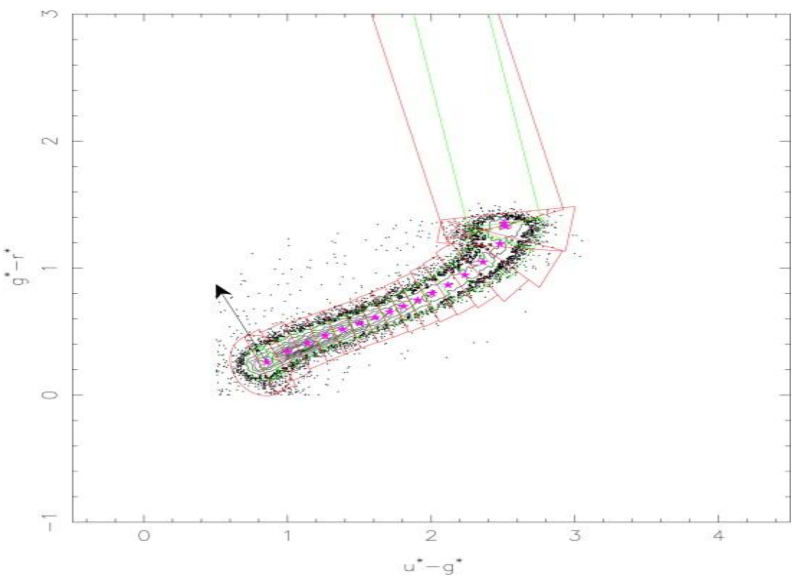


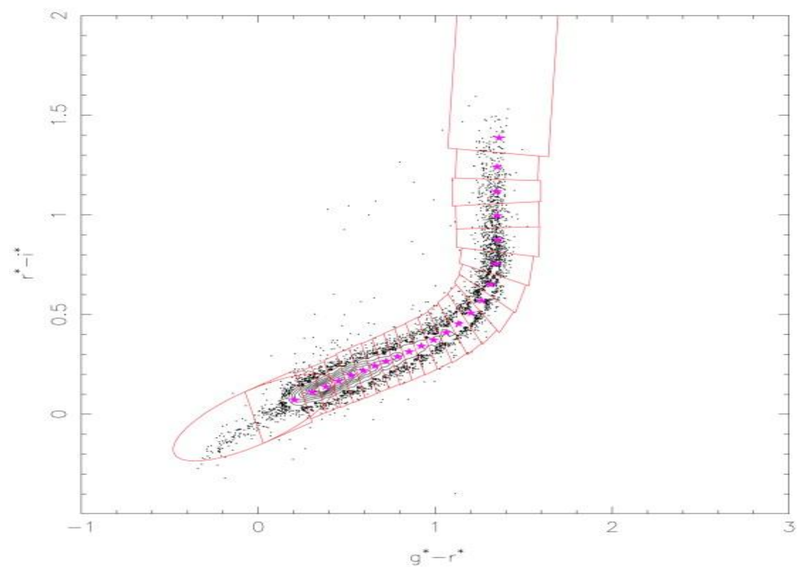
- **QSOs are supposed to be placed >4σ far from a cylindrical region containing the "stellar locus"** (S.L.), where σ depends on photometric errors.

OR

- **QSOs are supposed to be placed inside the inclusion regions**, even if not meeting the previous requirement.

c = 95%, e = 65%
locally less





1. **inclusion regions** are regions where S.L. meets QSO's area (due to absorption from Lyα forest entering the SDSS filters, which changes continuum power spectrum power law spectral index). All objects in these areas are selected so to sample the [2.2, 3.0] redshift range (where QSO density is also declining), but at the cost of a worse efficiency (Richards et al, 2001).
2. **exclusion regions** are those regions outside the main "stellar locus" clearly populated by stars only (usually WDs). All objects in these regions are discarded.

Overall performance of the algorithm: completeness c = 95%, efficiency e = 65%, but locally (in colours and redshift) much less.

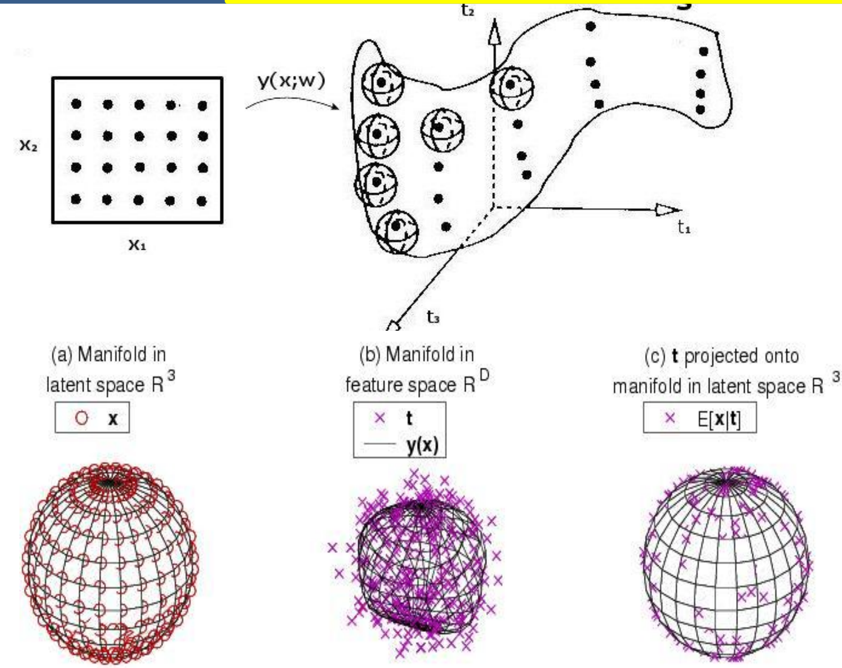
DAta Mining & Exploration

Science with Dame

1. **Probabilistic Principal Surfaces**
2. **Negative Entropy Clustering**

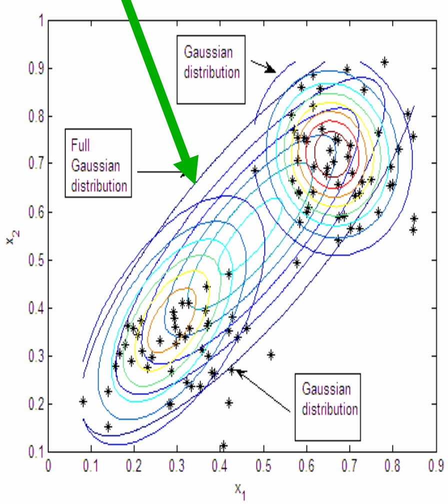
Step 1: Unsupervised clustering

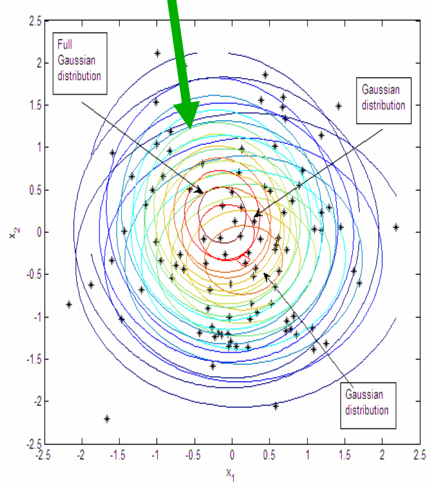
PPS determines a large number of distinct groups of objects: nearby clusters in the colours space are mapped onto the surface of a sphere.



Not replaced!

Replaced!





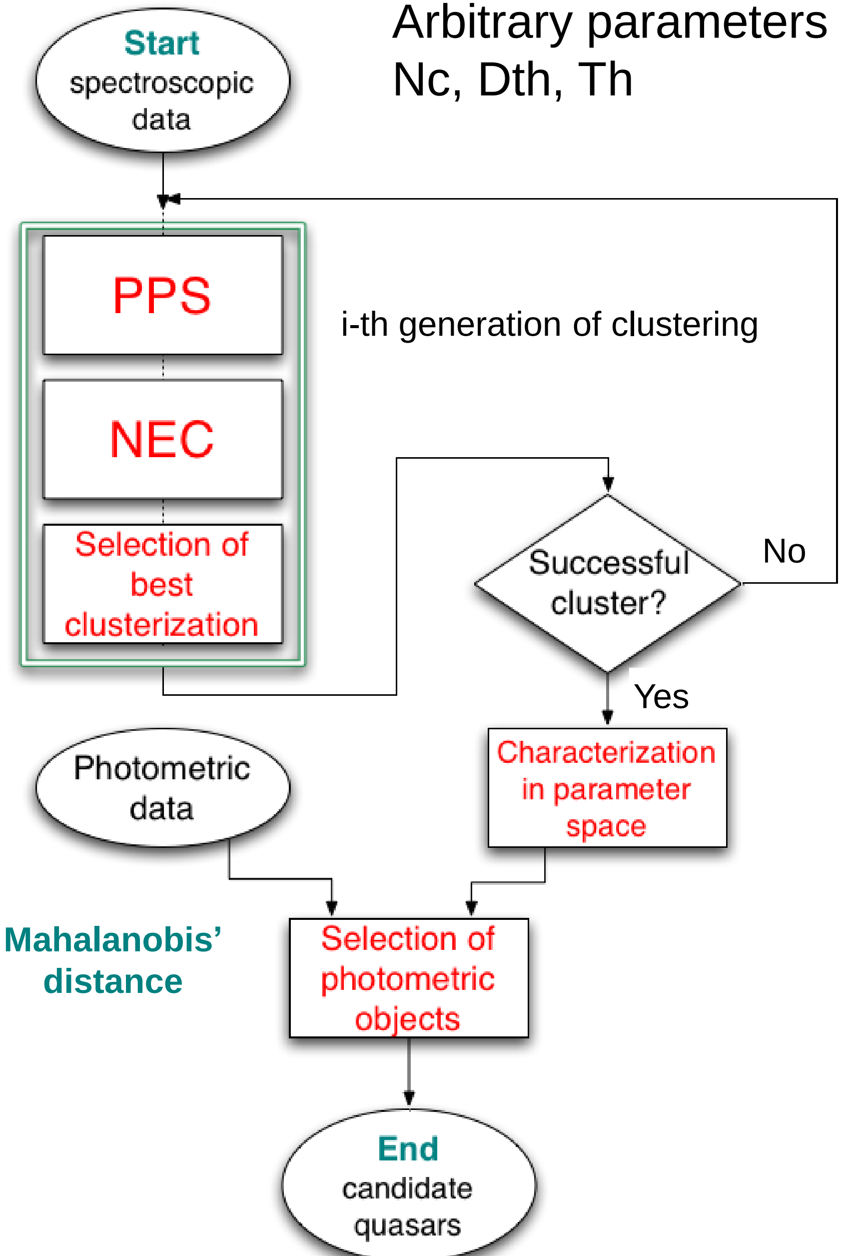
NegE=750

NegE=4

Step 2: Cluster agglomeration

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

1. **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.
2. **Dendrogram analysis**: the stability threshold(s) D_{th} can be determined observing the number of branches at different levels of the graph.



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)

Dth 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.

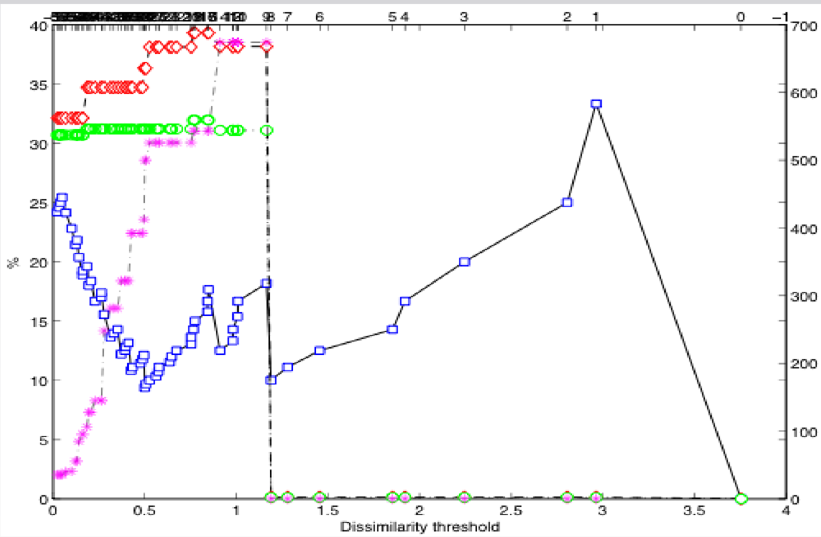
The overall efficiency of the process e_{tot} is the sum of weighed efficiencies e_i for each generation:

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

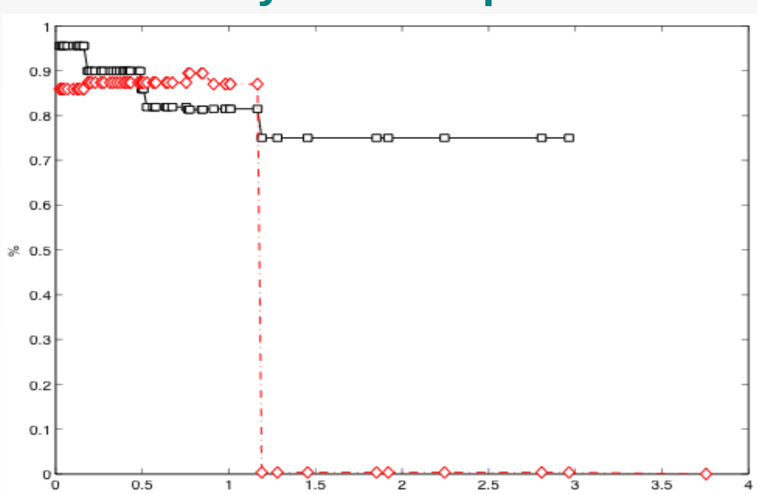
An example of “tuning”

Choice of the clustering

NSR



Efficiency and completeness



e and *c* estimation

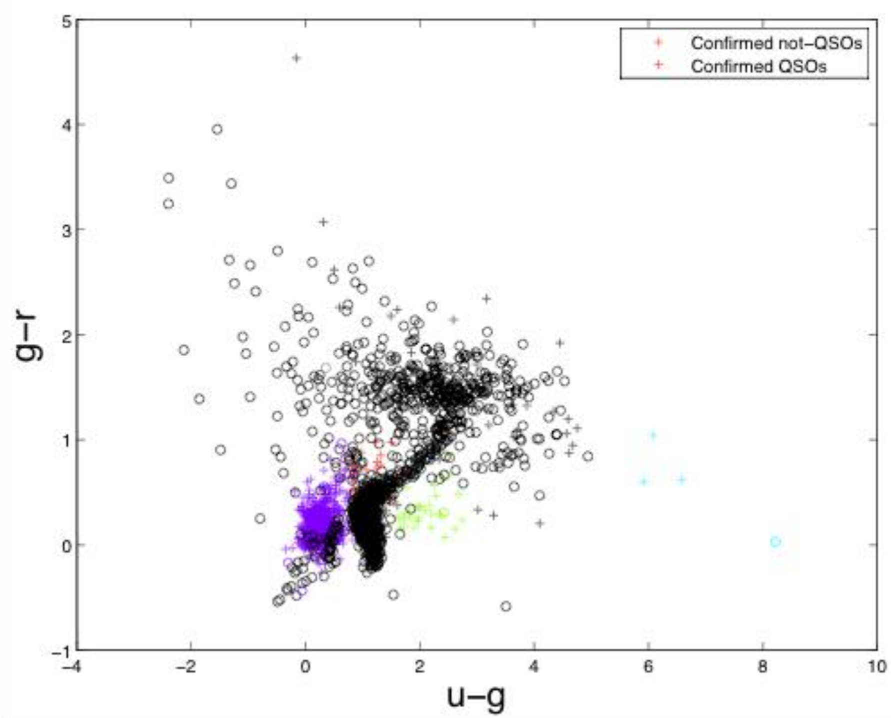
To assess the reliability of the algorithm, the same objects used for the “training” phase have been re-processed using photometric informations only. Results have been compared to the BoK.

labels / algorithm	QSOs	not QSOs
QSOs	759	72
not QSOs	83	1327

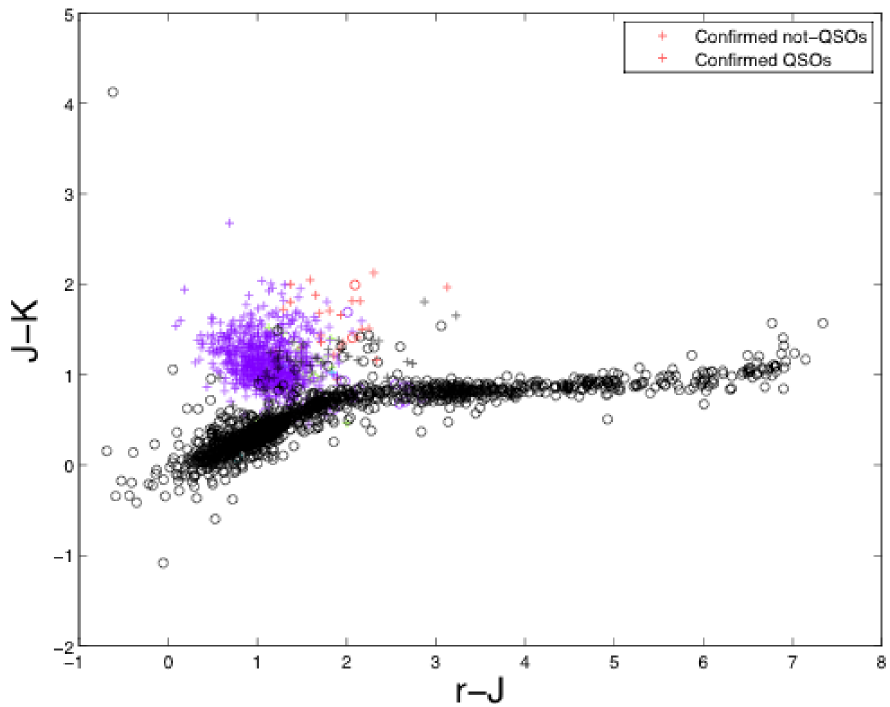
e = 83.4 % c = 89.6 %

Confusion matrix

u - g vs g - r

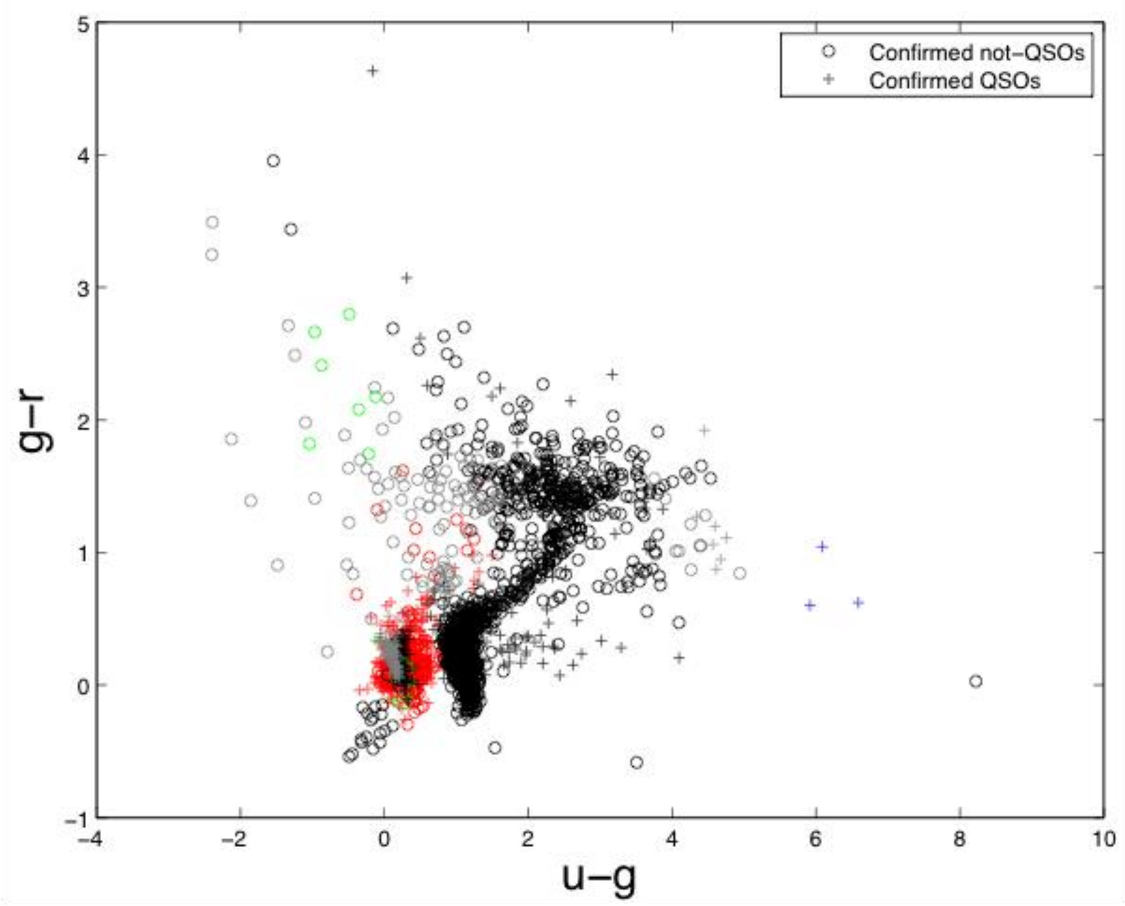


r - J vs J - K

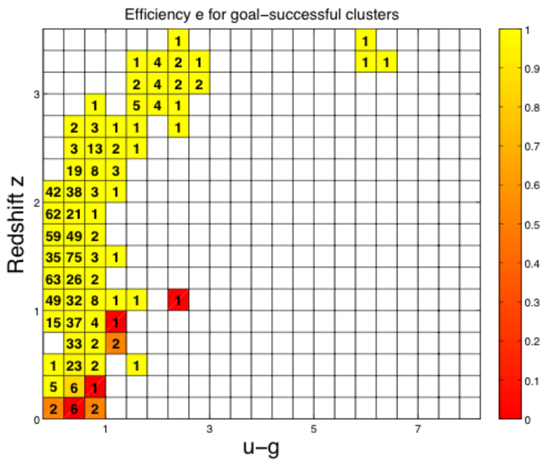


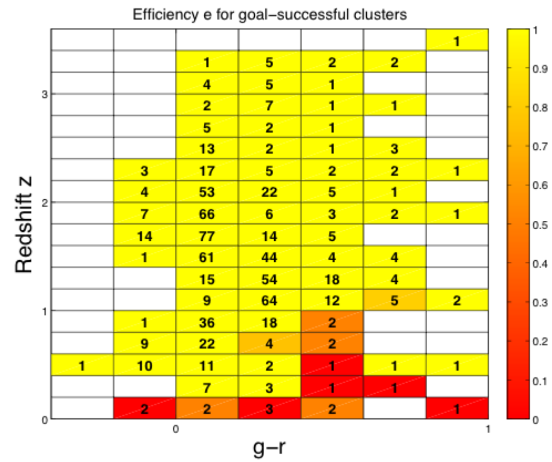
Only a fraction (43%) of these objects have been selected as candidate QSO's by SDSS targeting algorithm in first instance: the remaining sources have been included in the spectroscopic program because they have been selected in other spectroscopic programmes (mainly stars).

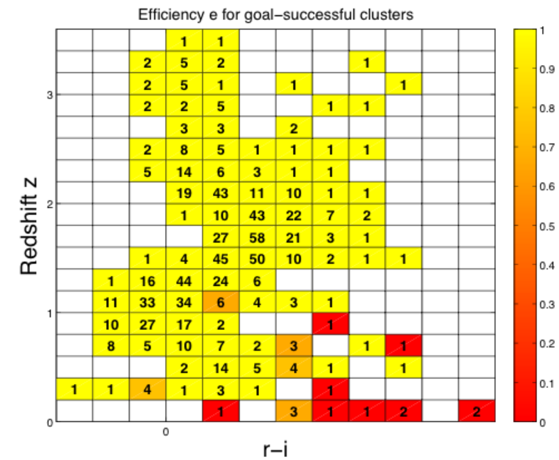
u - g vs g - r

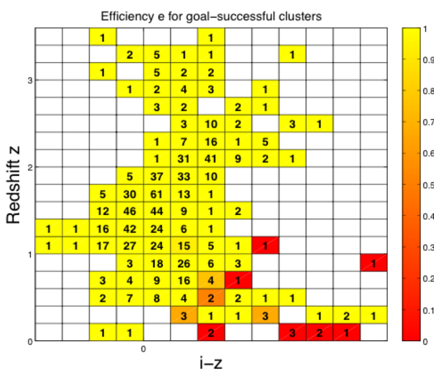


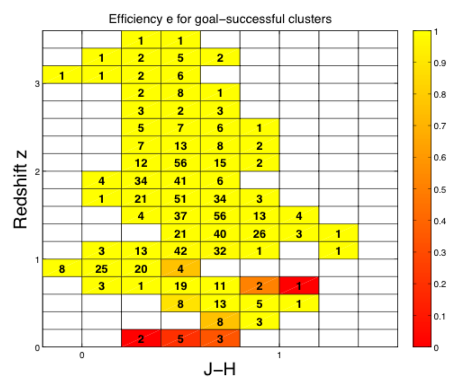
In this experiment the clustering has been performed on the same sample of the previous experiment, using only optical colours.

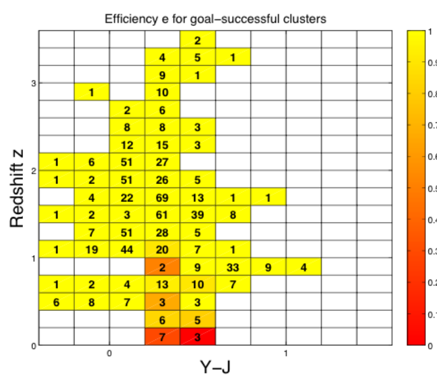


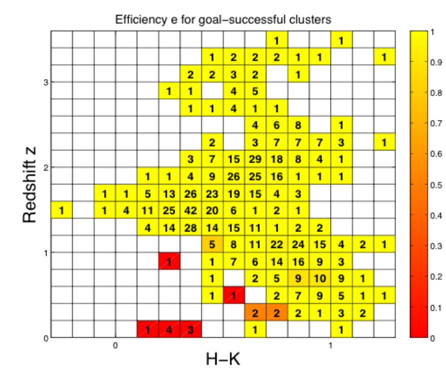






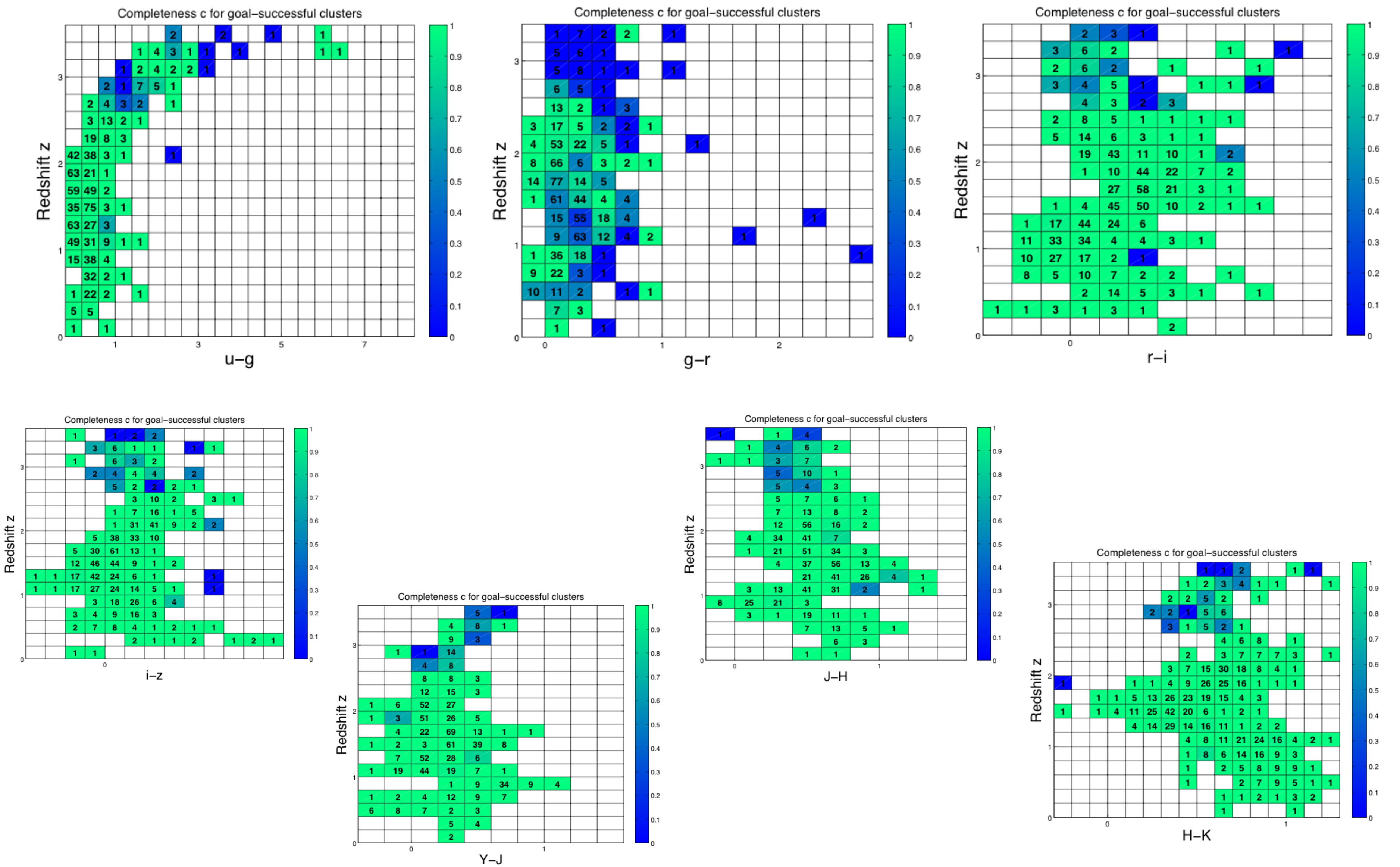






Experiment 2: local values of *e*

Experiment 2: local values of *c*



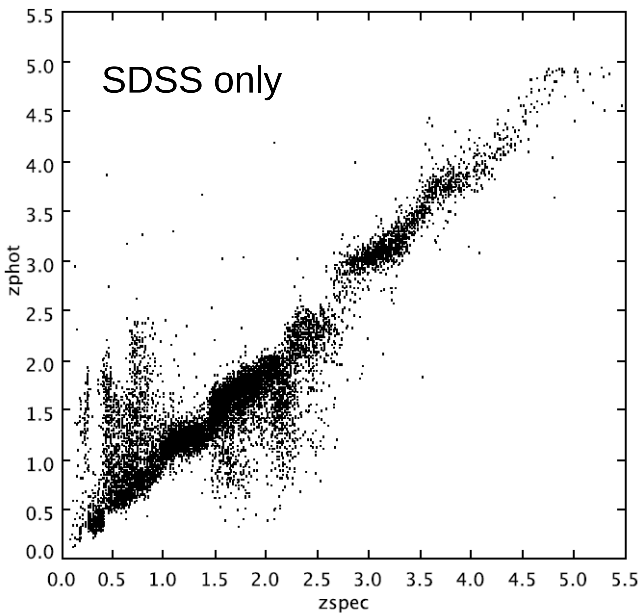
Sample	Parameters	Labels	e_{tot}	C_{tot}	n_{gen}	n_{suc_clus}
Optical QSO candidates (1)	SDSS colours	'specClass'	83.4 % (± 0.3 %)	89.6 % (± 0.6 %)	2	(3,0)
Optical + NIR star-like objects (2)	SDSS colours + UKIDSS colours	'specClass'	91.3 % (± 0.5 %)	90.8 % (± 0.5 %)	3	(3,1,0)
Optical + NIR star-like objects (3)	SDSS colours	'specClass'	92.6 % (± 0.4 %)	91.4 % (± 0.6 %)	3	(3,0,1)

The catalogue of candidate quasars is publicly available at the URL:

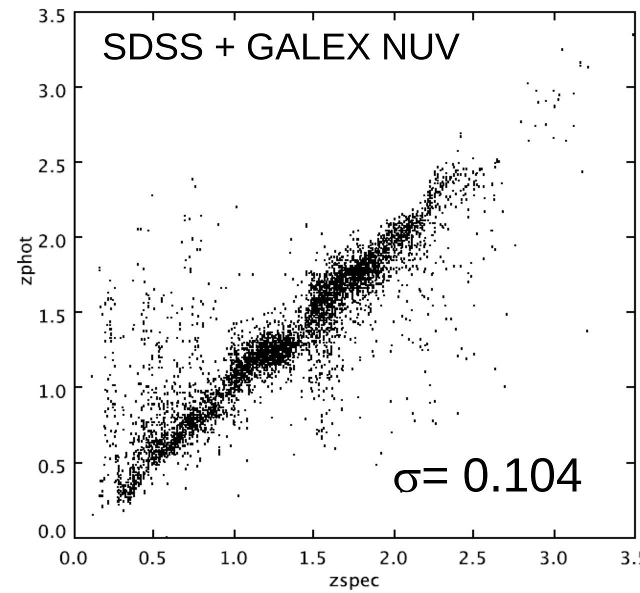
http://voneural.na.infn.it/catalogues_qsos.html

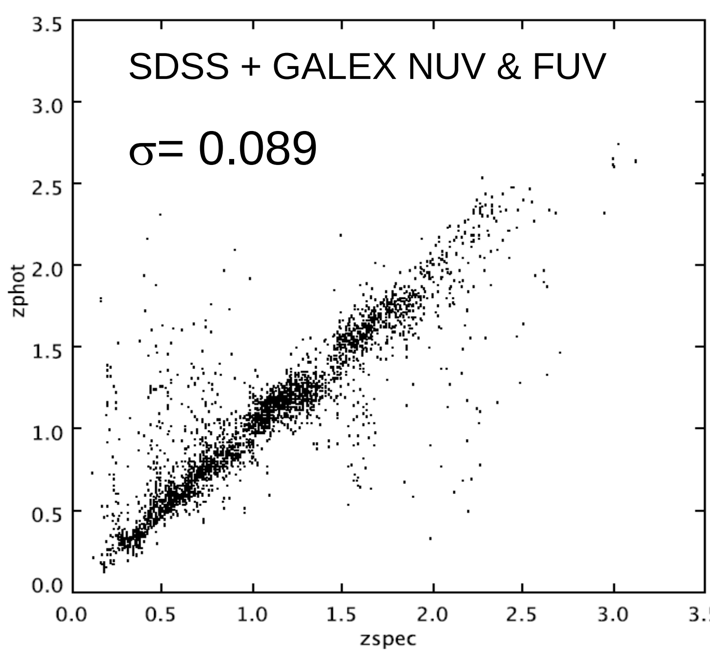
BUT … LET'S GO BACK TO PHOT-Z

1. Photometric redshifts of QSOs



$\sigma = 0.154$

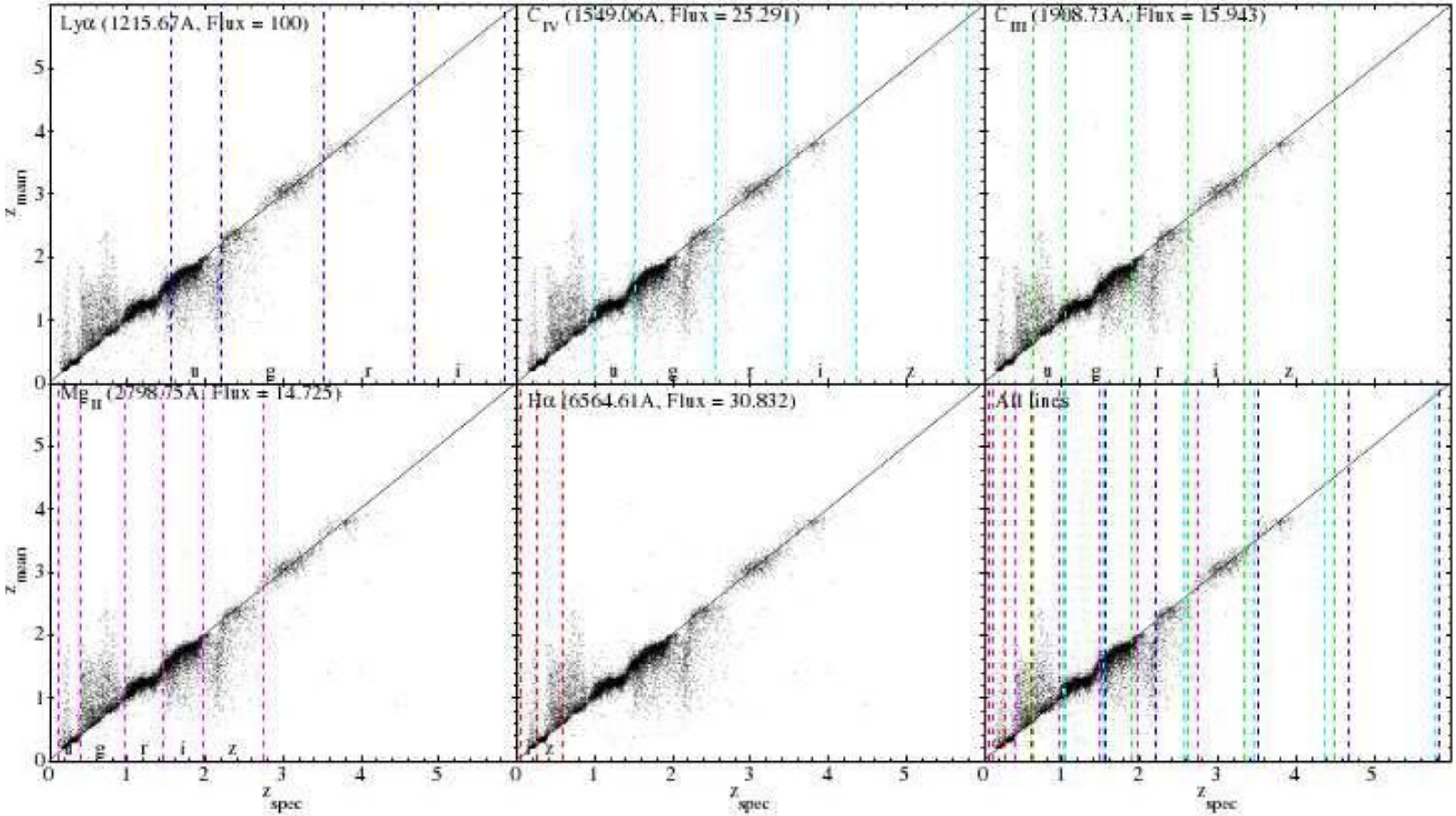




No need for fine tuning !!!

Only New BoK !!!

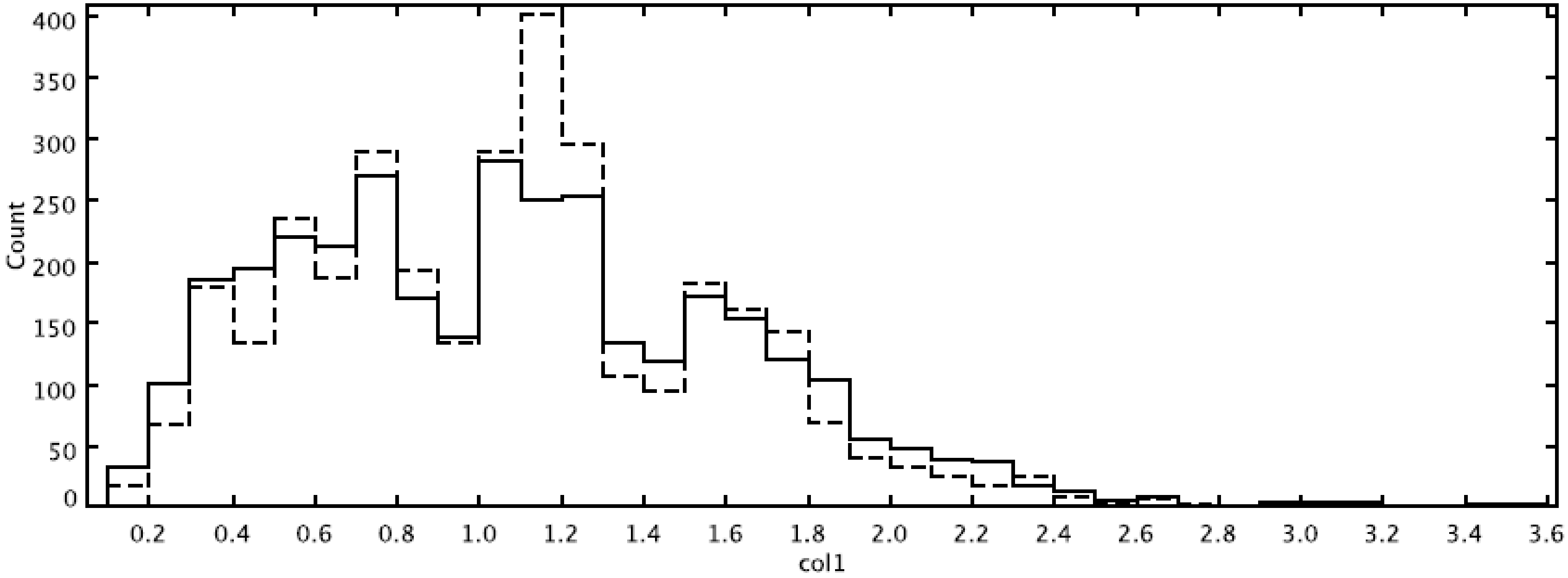
Laurino et al. 2009a,2009b

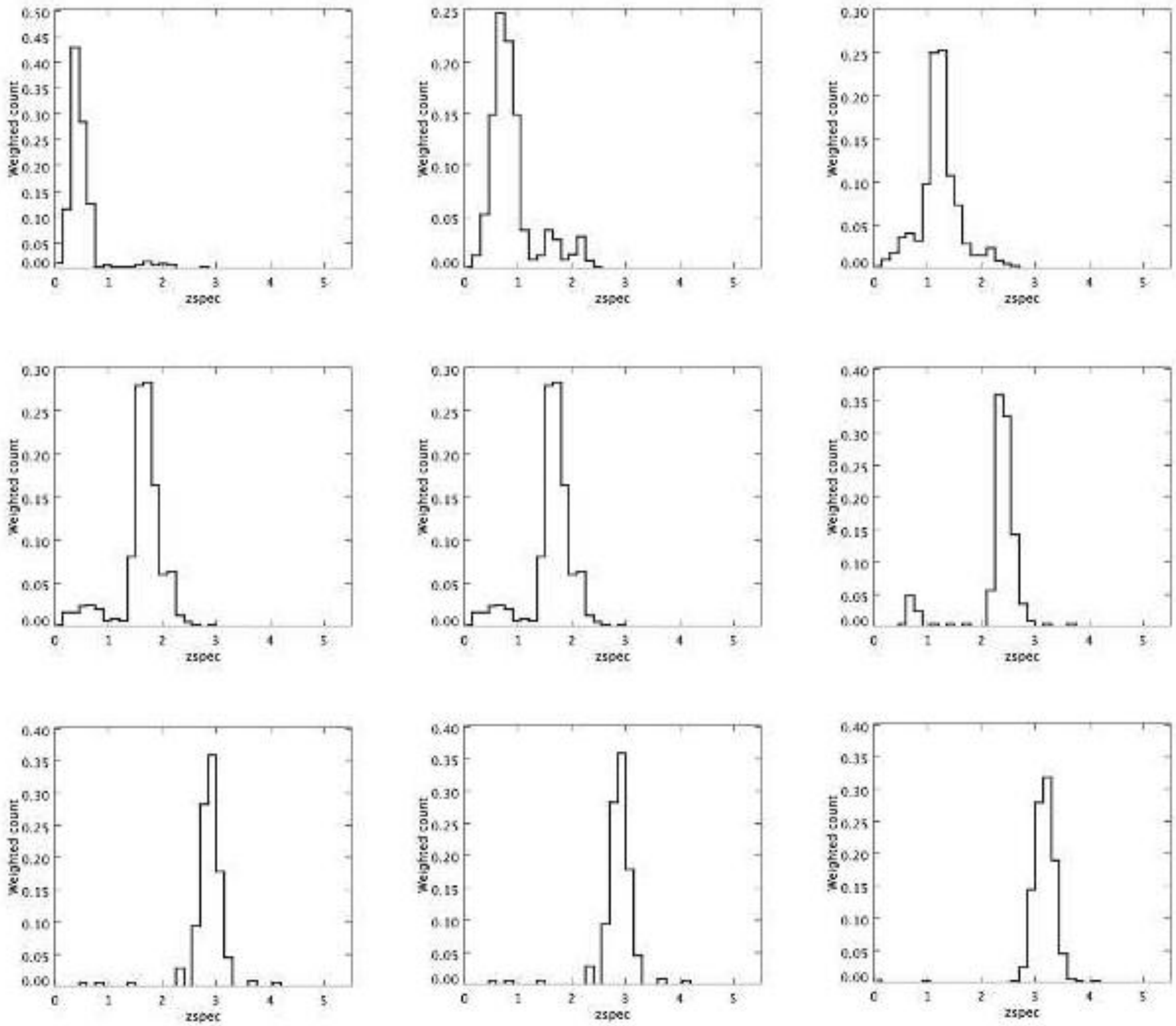


Degeneracy induced by lines exiting photometric bands

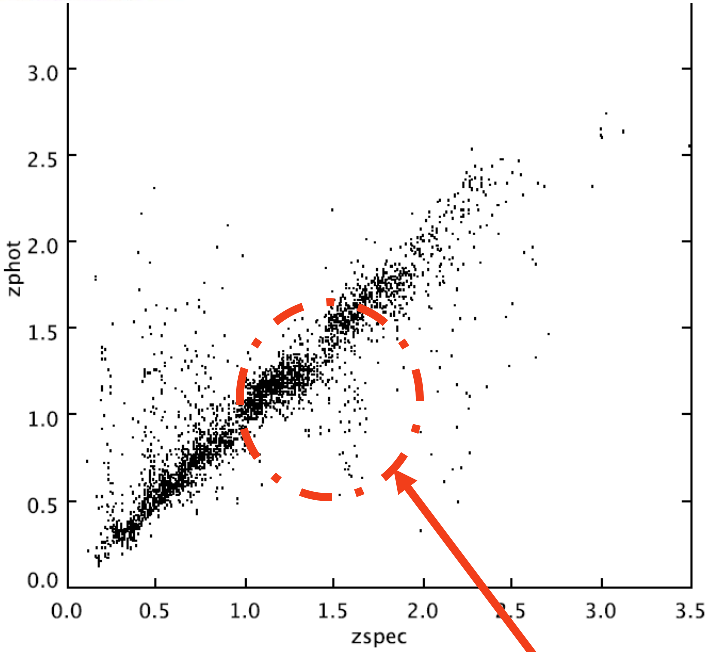
Ball et al. 2008, ApJ, 683, 1221

Distribution of Z_spec (solid) and Z_phot (dashed) for test set !!!!



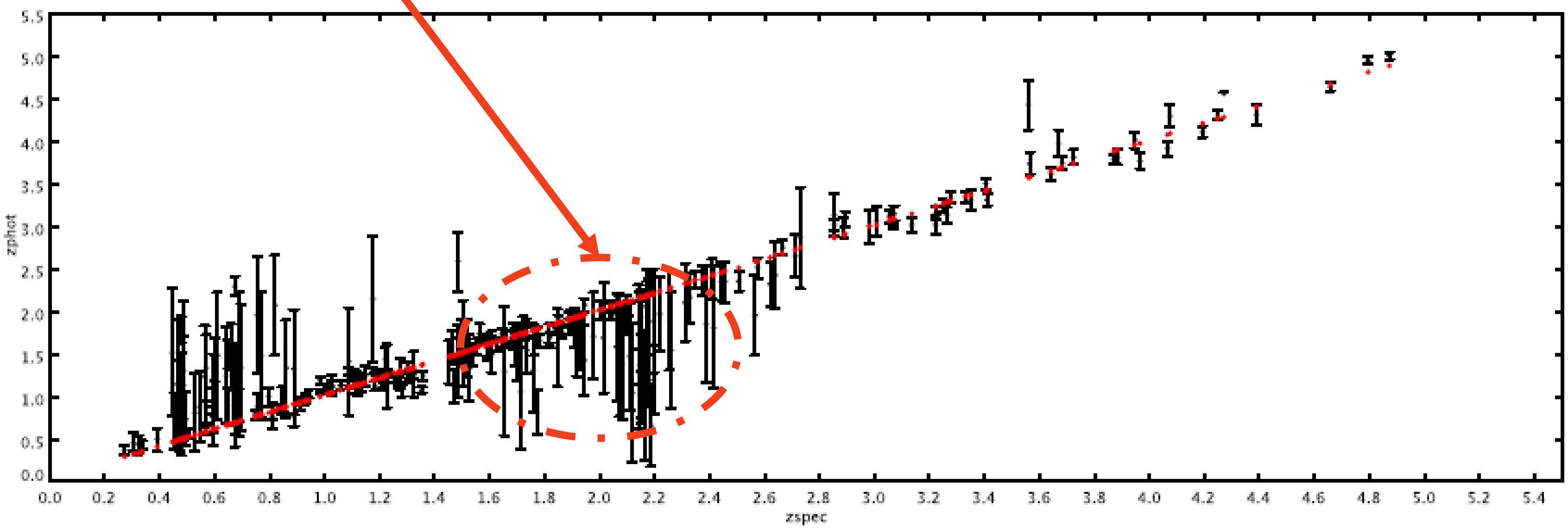


Laurino et al. 2009a,2009b



Errors:

- **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 may be 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***).

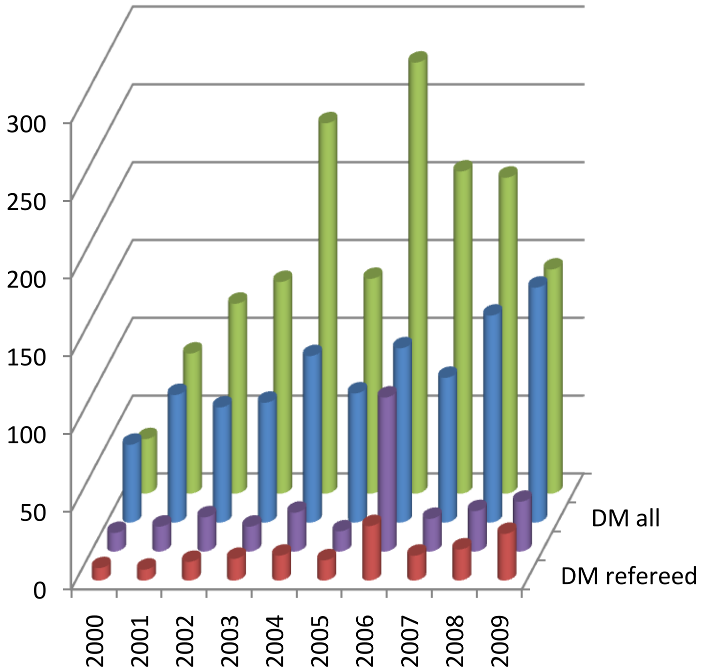


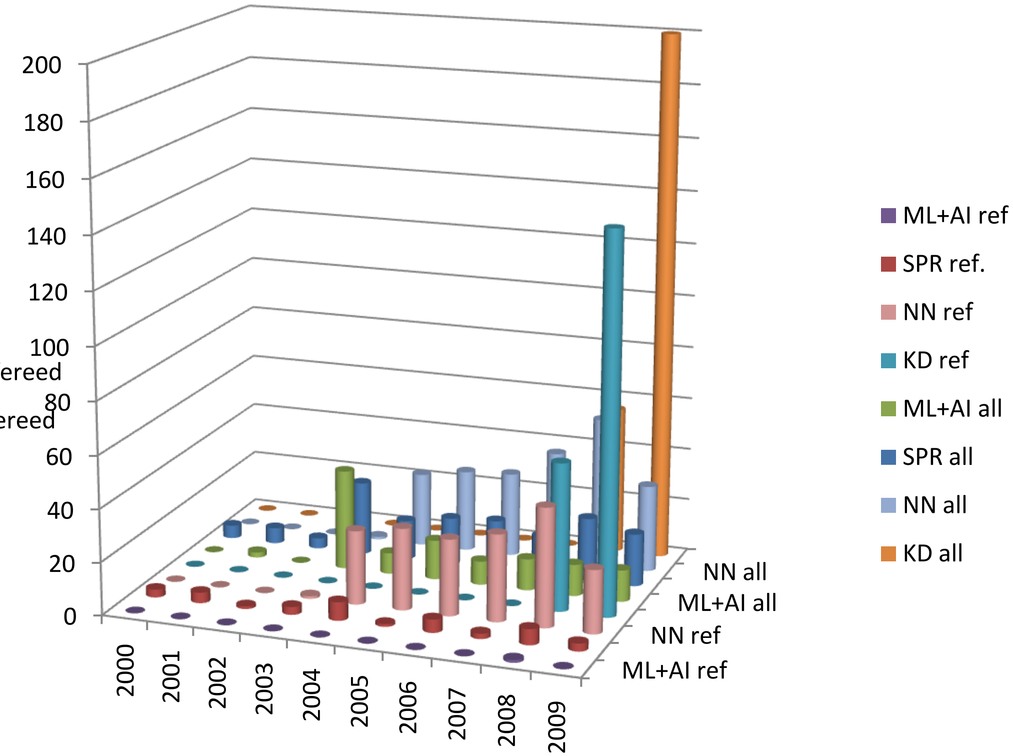
Laurino et al. 2009a,2009b

DAta Mining & Exploration

DAME

Summary and Conclusions II. Sociological issue to be solved.

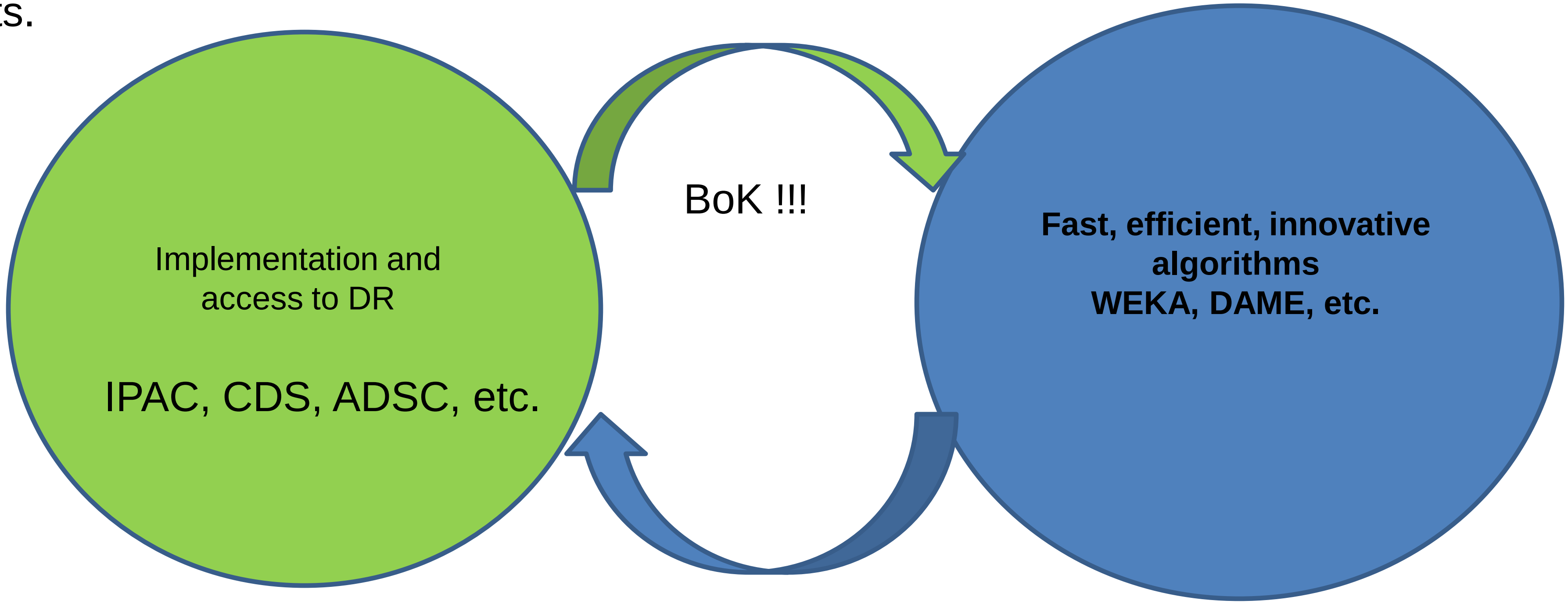




1. Number of technical/algorithmic papers increases with new funding opportunities. Number of refereed papers remains constant.
2. Most of the work, so far, remains at the implementation stage (computer Science and algorithm development) and does not enter the “science production” stage…
3. Out of one thousand papers checked (galaxies, observational cosmology, survey) over the last two years: DM could be applied or involved in at least 30% of them leading to better results

Summary and Conclusions II.

Machine Learning based Data Mining is unavoidable when working on huge data sets.



Quality control
Metadata

Accuracy of results depends on accuracy of BoK !!!!

The extraction of BoK's offers challenges to good data repositories and data archives.

Reliability and completeness of information

(no data is better than bad data)

Compliance with ontologies

Advanced queries in natural language

THANKS


Antonino
Simona
Roberto
Lara
betty
Ciro
Omar
Giuseppe
Giancarlo
Silvia
Marilena
Maurizio
Raffaele
Mauro
Alfonso
Max Brescia
23/01/2006