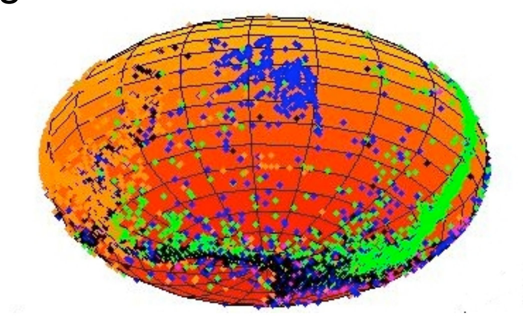
The 3-D structure of the nearby (z<0.5) universe
From the SDSS Archive

Giuseppe Longo[1,2,3]

In collaboration with:

Massimo Brescia[3,2], **Raffaele D'Abrusco[1]**, Elisabetta De Filippis[4,1], Maurizio Paolillo[1,2,3] ,Antonino Staiano[4,3], Roberto Tagliaferri[5]

1 - Department of Physics – University of Napoli Federico II
2 - INFN – Sezione di Napoli
3 - INAF – Sezione di Napoli
4 – INAF – Sezione di Trieste
5 – DMI – Università di Salerno



A paradigm shift has occurred in astronomy

PAST	PRESENT	NEAR FUTURE
Pointed, heterogeneous observations (~ MB - GB) Small samples of objects (~ $10^1 – 10^3$) Few parameters	Large, homogeneous sky surveys multi-TB Archives Large samples of objects (~ $10^6 – 10^9$) Dozen parameters	Federated sky surveys and archives (~PB) Whole Sky Surveys Hundreds of parameters

"**VO** aims to enabling data analysis techniques through a coordinating entity that will provide common standards and state-of-the-art analysis tools."(1)

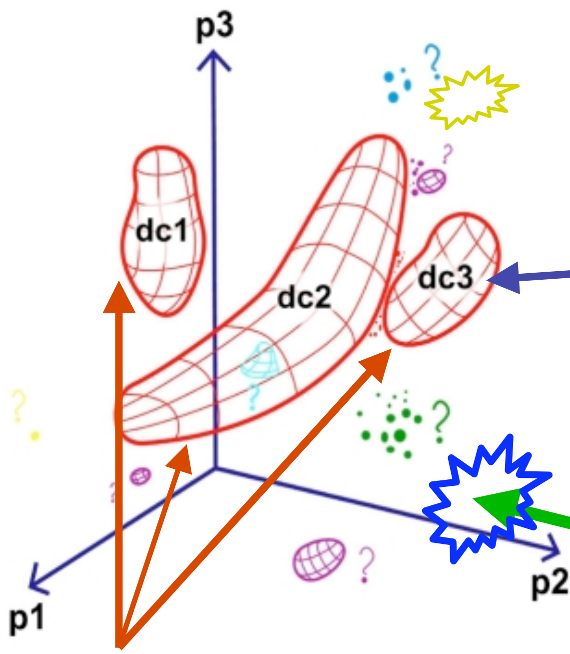
- Interoperability
- Good science cases

Virtual Observatory

Precision cosmology needs accurate data for large (>10^6) samples of galaxies

Why did we focus on data mining?

… one of VO goals is the development of tools for the exploration of parameter space distribution of objects attributes measured in different surveys.



Discovery of trends in multiparametric spaces

Search for rare or even new types of objects, most probably identified as outliers of the P.S. distribution

Classification of properties

$$P_i \qquad i = 1,\dots,N \qquad \mathrm{N} = 100 \text{ and more}$$

Most astronomical work can be traced to either a clustering or a classification problem: Phyical classification of galaxies, star/galaxy separation, search for QSO's, variable objects, K-B objects, etc..

The case of photometric redshifts and photometric classification

The relevance of photometric redshifts

- luminosity functions
- 3-D correlation functions
- mass function through weak lensing
- clustering properties,
- constraints on cosmological parameters
- anisotropy of UHECR's etc.

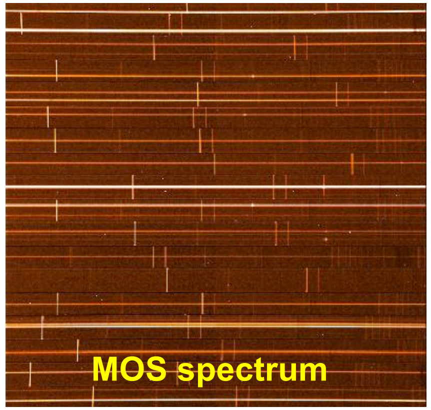
- So far most of our informations on the cosmic structures has come from 2-D data (distribution of points projected on the celestial sphere).
- Modern spectroscopic and photometric surveys are allowing to explore the 3-D structure of the universe, with obvious advantages
- Higher contrast allows to detect structures of lower multiplicity (groups of galaxies instead then clusters)
- Distance allows better understanding of the role played by different types of sources
- Etc…. **IN OTHER WORDS …..**

2 is always better than 3

Spectroscopic redshifts through multiobject spectroscopy (MOS)

High precision ($\Delta z.<.005$)… but

- In one frame at most 400-500 low resolution spectra
- 10^5 spectra require more than 100 nights of observations with 8 m class telescopes, hence: either pencil beam surveys or shallow surveys



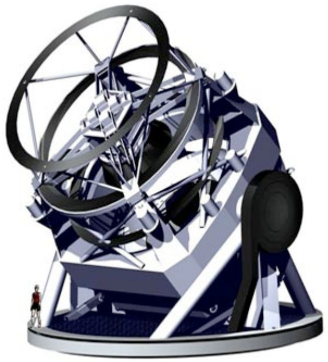
Photometric redshifts through multiband photometry

- Lower precision (but not much…)
- In 100 nights, up to $10^{7\text{-}9}$ phot-z
- Additional information on galaxy types

LSST (USA)

8 m diameter
3.5x3.5 sq deg f.o.v.
80 k x 80 k CCD mosaic (NIR)

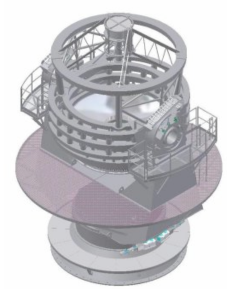
Operational 2013



VST (I,ESO)

2.5 m diameter
1x1 sq deg f.o.v.
16 k x 16k CCD mosaic (optical)

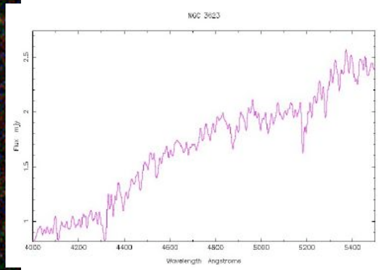
Operational 2007



VISTA (UK, ESO)

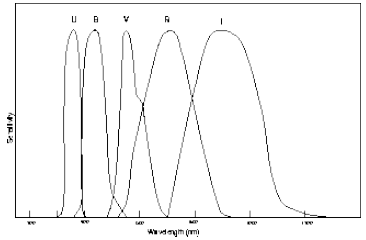
4 m diameter
1.65x1.65 sq deg f.o.v.
8 k x 8 k CCD mosaic (NIR)
Operational 2007

Photometric redshifts



Galaxy spectrum - F(λ)

X



Photometric system - $S_i(\lambda)$

=

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

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

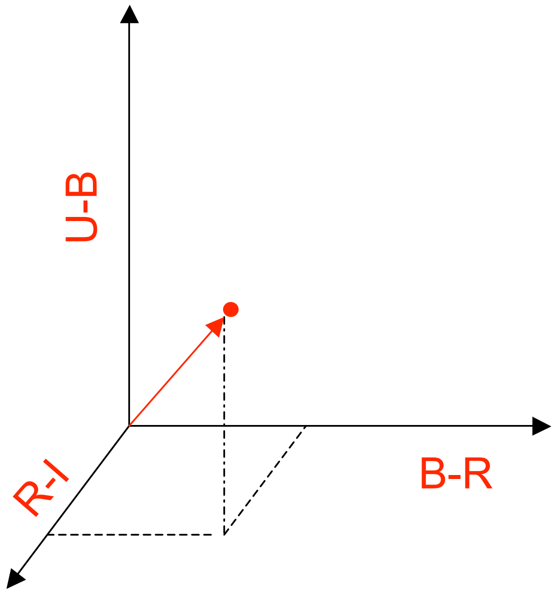
Etc…

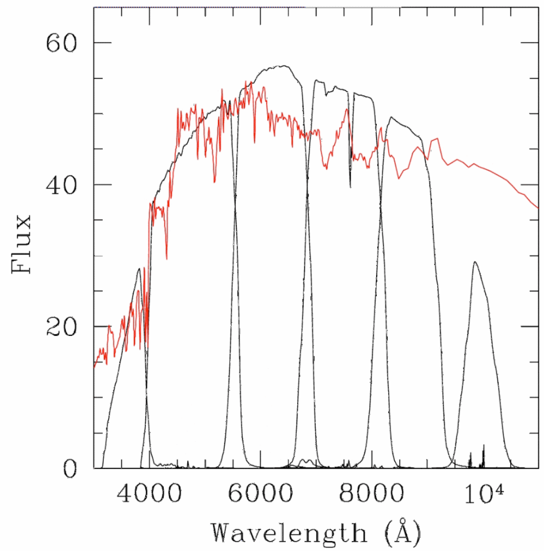
Color indexes

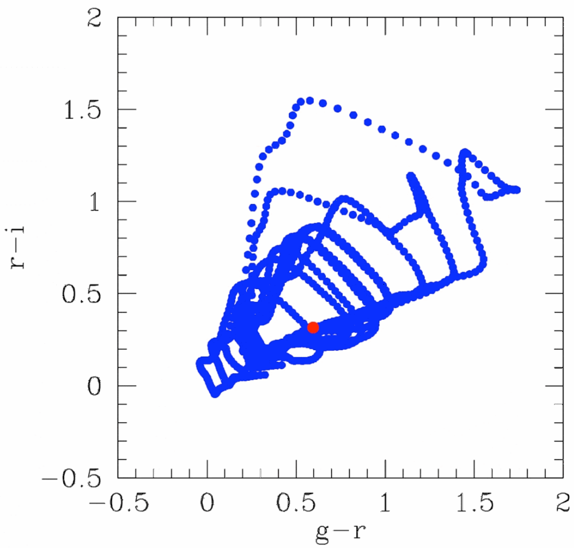
$$U - B \equiv m_U - m_B$$

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

etc.







LSST photometric system animation

- **Different "loops" correspond to different SEDs (morph. Types)**
- **Same colors but different intrinsic luminosities (same apparent magnitude at different redshifts)**
  - Degeneracy can be removed by adding dimensions (colors) but not too many are needed !
  - At any time one broad spectral feature dominates the game (in the optical for low-medium redshifts, it is the Balmer Break)
- **Within a given redshift range three colors (four bands) are the minimum**

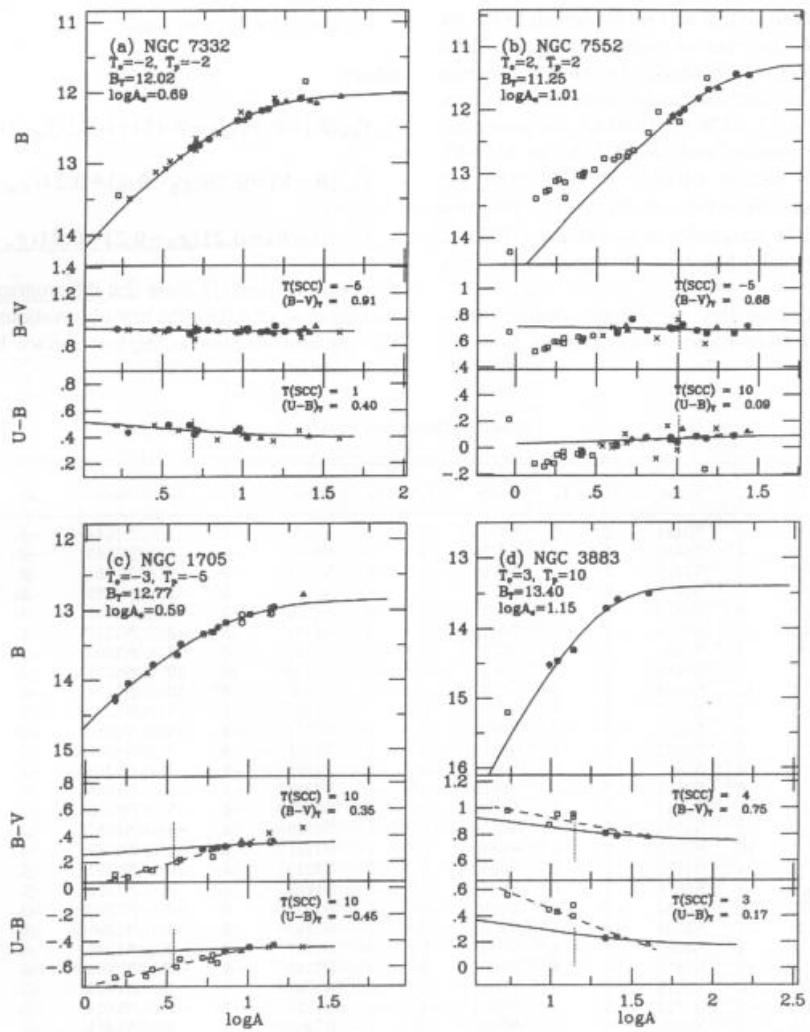
Some general facts have long been known

Connolly et al. 1995.

Distribution in the U,B,R space is almost planar, galaxies occupy less than 4% of the parameter space.

- Galaxies at the same redshift form an iso-redshift slab were the coordinates are the luminosity and the intrinsic color.
  The red edge is shifted due to the fact that red galaxies are on average more luminos than blue ones. As redshift increases galaxies get dimmer and change color (K correction).
- **In broad band photometry the interval between two bands corresponds roughly to a redshift 0.4; this is equivalent to a 90° rotation of the distribution every 0.4 step in redshift.**
- Only one feature at the time plays a role (Balmer break, Ly forest, UV RB). Emission lines negligible. Three magnitudes are usually enough.
- **Optical is confined to z< 1.0, then NIR comes in; UV useful but not crucial.**
- This can be used as a base for physical classification (redshift, SED type, intrinsic luminosity).

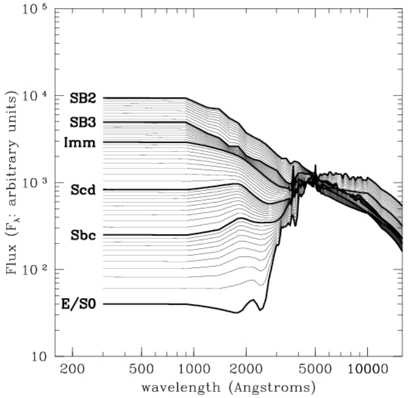
Photometric types have little to do with morphology



Therefore morphology correlates only weakly with SEDs

Buta, de Vaucouleurs and Longo, 1995, AJ

Two types of methods are possible:



SED template fitting

- empirical templates → **Poor coverage of parameter space**
- synthetic templates → **Too many assumptions on underlying physics**

Empirical methods based on interpolating an “a priori base of knowledge”

- are relatively free from possible systematic effects within the photometric calibration.
- As such, they provide a simple measure of the statistical uncertainties with the data and can demonstrate the accuracy to which we should be able to estimate redshifts once we can control the systematic errors.

The best interpolators known so far are neural networks ….

In order to be useful Cosmological redshifts need to have well behaved and well understood error distributions

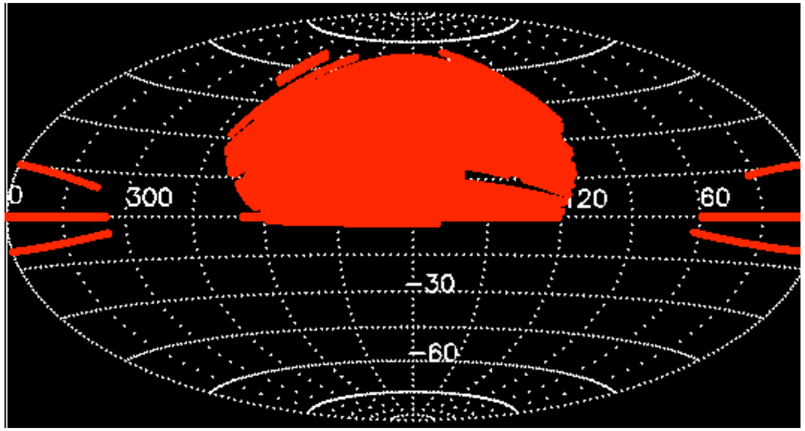
$$\left.\frac{dN}{dz}\right|_{observed} \cong \left.\frac{dN}{dz}\right|_{real} \otimes g(\Delta z)$$

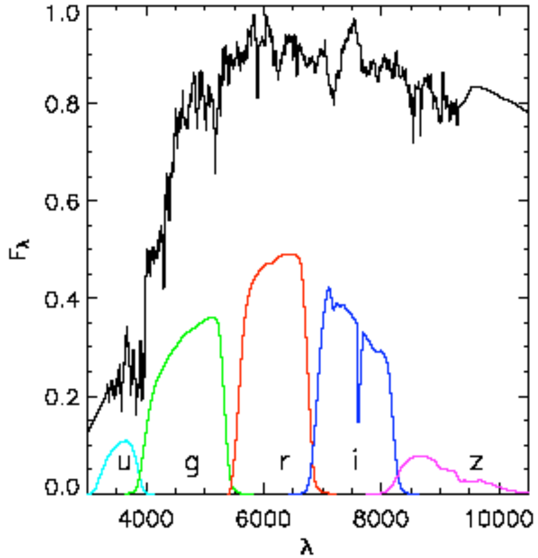
$$\left.\frac{dN}{dz}\right|_{observed}(z_{phot}) = \int_0^\infty \left.\frac{dN}{dz}\right|_{real}(z')\, g(z'-z_{phot}, z')\, dz'$$

- **Which is the well known Fredholm equation (Craig & Brown, 1986, Inverse problems in astronomy. Adam Hilger Ltd, Bristol**

- **And it is ill defined since it describes a smoothing operator that generically loses information, implying that the solution will in general require incorporating some prior knowledge about d*N*/dz.**

The Sloan Digital Sky Survey – Data release 5





8000 sq degrees
>210 million galaxies
data are public

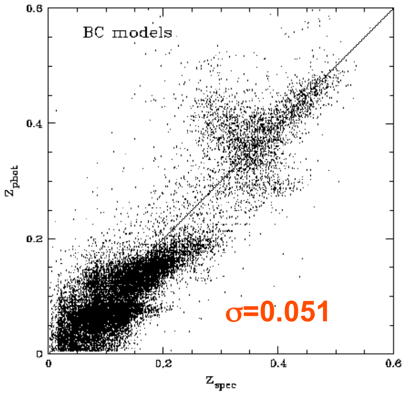
Base of Knowledge:
700.000 galaxy spectra

Benchmark for almost everything in observational cosmology

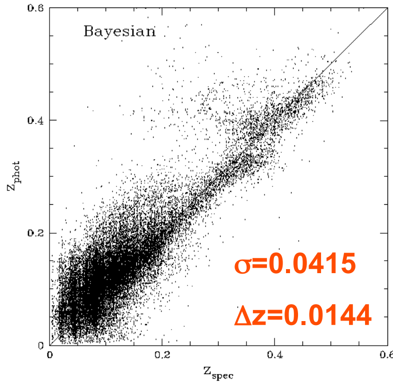
Subsample of about 10^7 Luminous Red Galaxies (LRG)

SED fitting

Csabai et al. 2003



Benitez et al. 2009



Firth et al. 2003

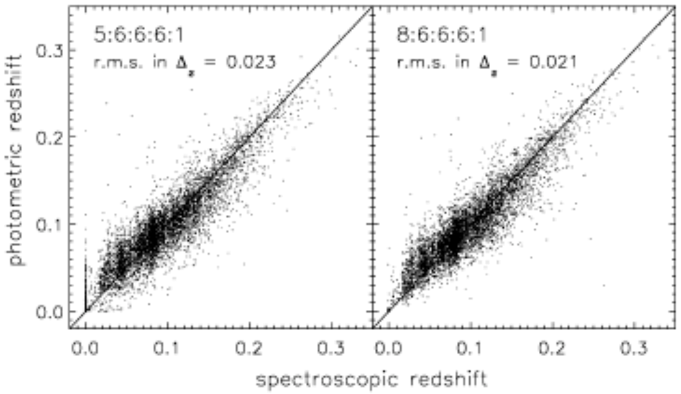


Figure 9. A comparison of photometric and spectroscopic redshifts using SDSS public data. Two ANN architectures were used, taking as input *ugriz* photometry (5:6:6:6:1 architecture, left) and *ugriz* photometry, SDSS star/galaxy classifier and Petrosian 50 and 90 per cent *r*-band flux radii (8:6:6:6:1 architecture, right). A training set of size 10 000 was used. The ANNs were tested on a separate testing set of size 7000 (plotted). In each panel, redshift estimates are medians from a committee of five ANNs.

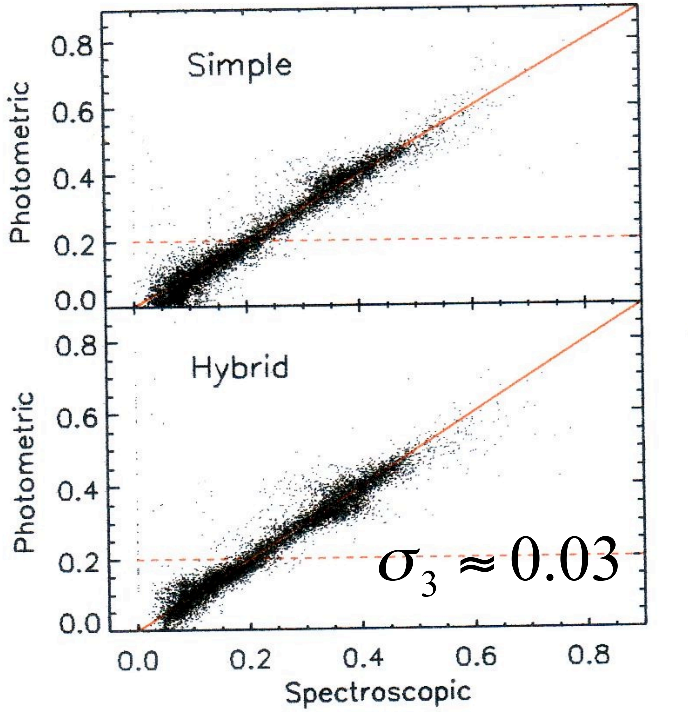
Empirical method (ANN)

ASTRONEURAL

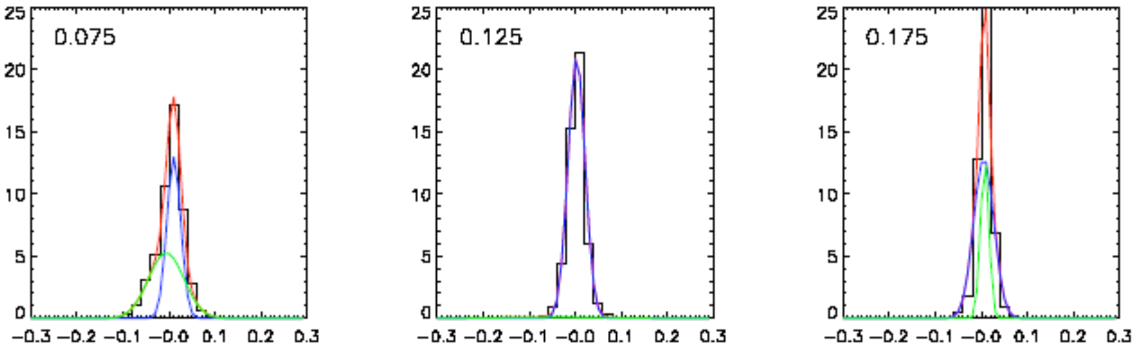
Padmanahaban et al., 2005

Constructing a photometric redshift catalogue involves three steps:

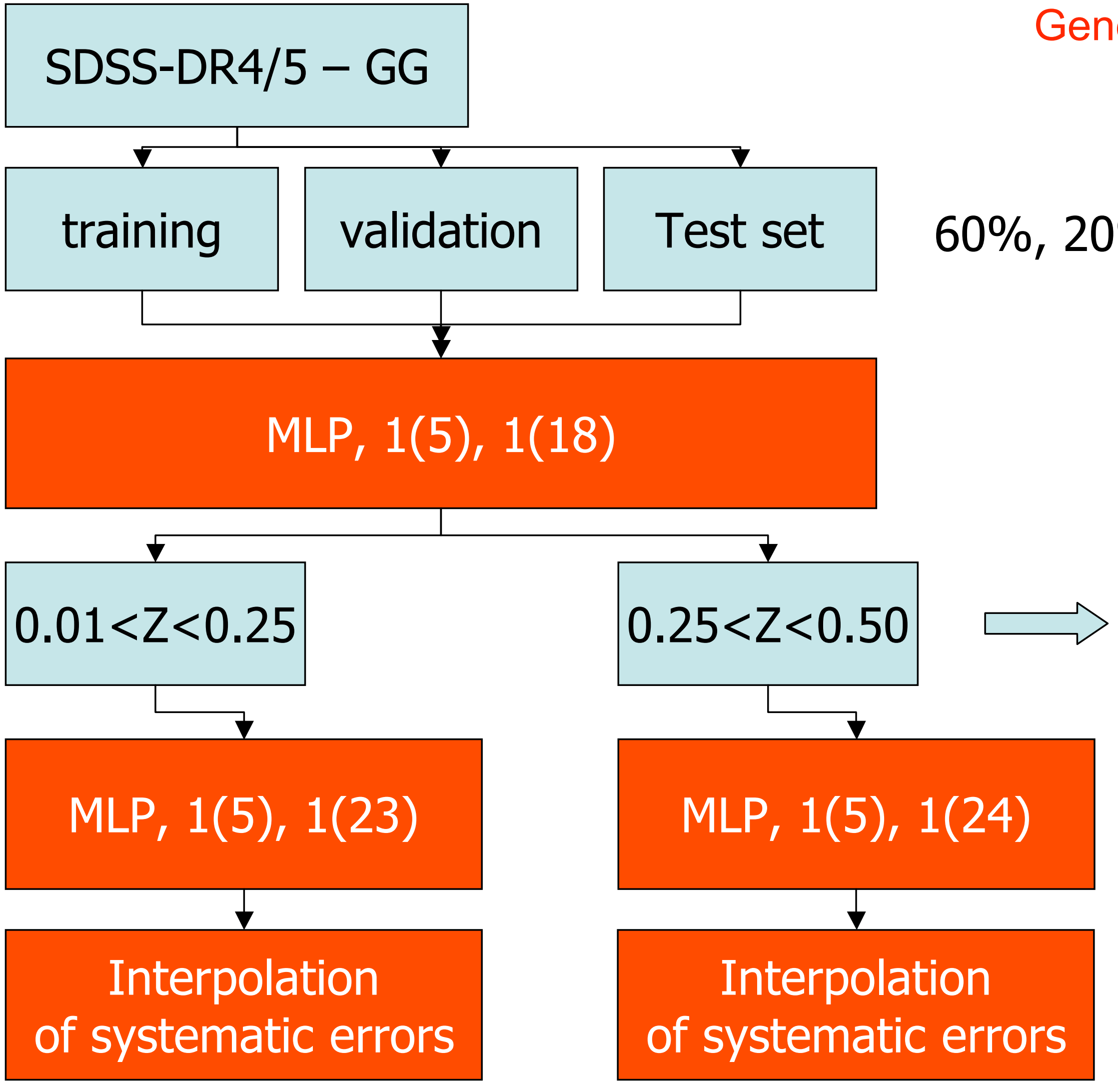
- **Photometrically selecting a sample for which accurate photometric redshifts can be obtained**;
- Measuring the photometric redshift error distribution for the resulting sample;
- estimating the true redshift distribution.



LRG sample



General Galaxy sample



60%, 20%, 20%

99.6 % accuracy

σ rob = 0.206

σ rob = 0.234

Optimization of the network

1 hidden layer

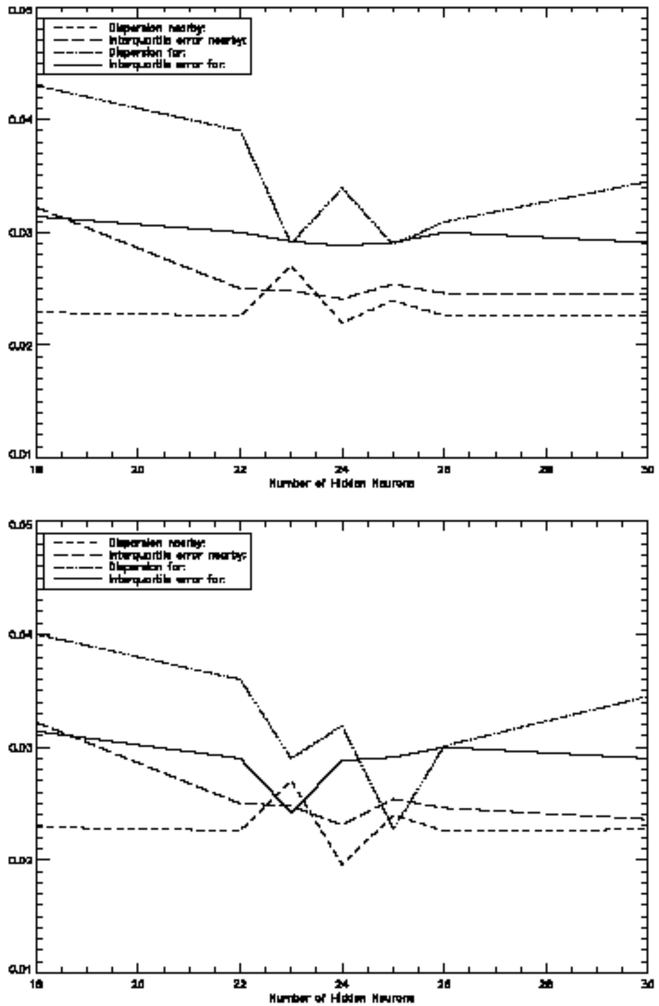
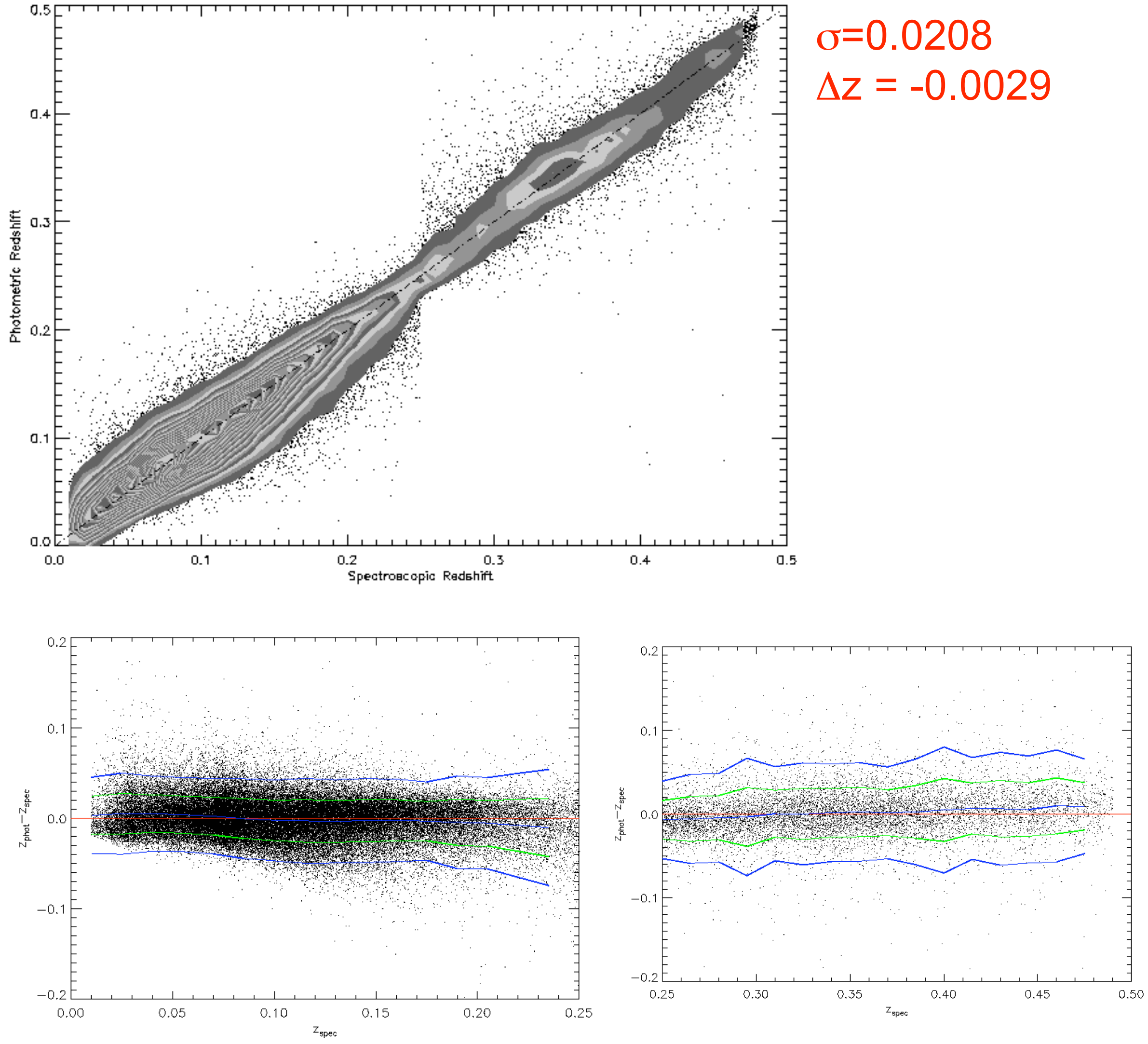


Figure 4. Upper panel: CGS sample, trend of the interquartile error and of the variance as a function of the number N of the neurons in the hidden layer. Notice the minimum at N=24. The nearby and distant samples are plotted separately. Lower panel: the same as above for the LRG sample.

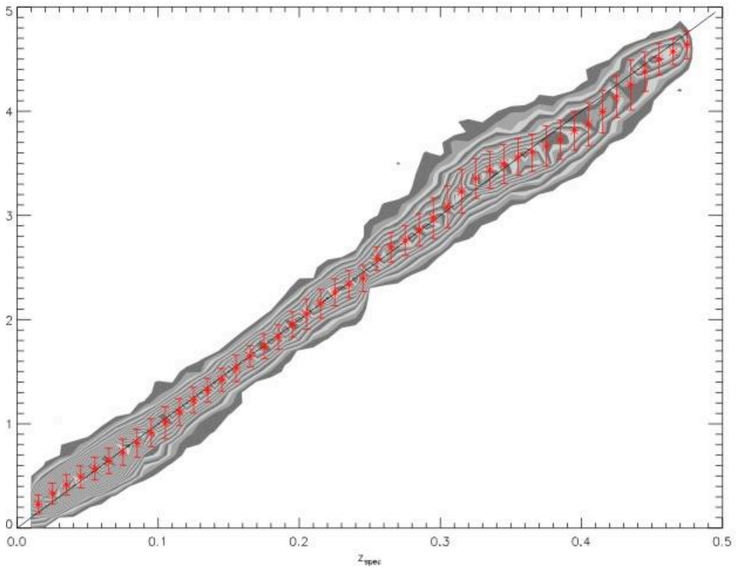
General galaxy sample



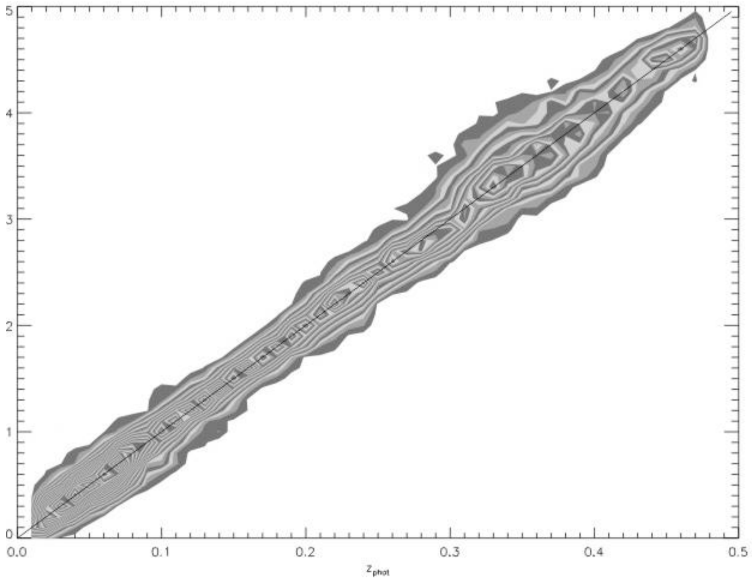
σ=0.0208

Δz = -0.0029

LRG sample



Σ =0.0164
Δz = -0.0006 !

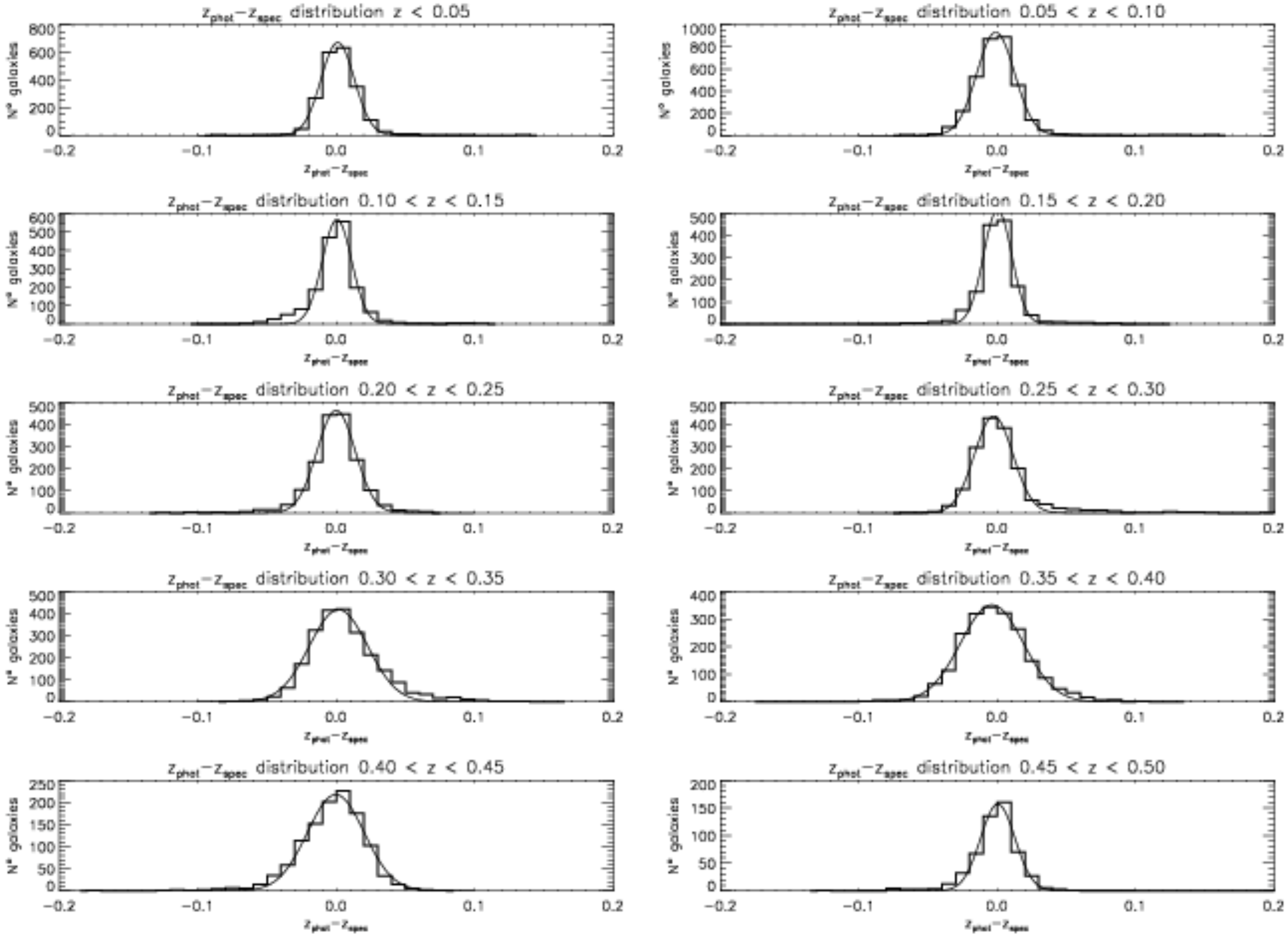


Catalogues of photometric redshifts for:

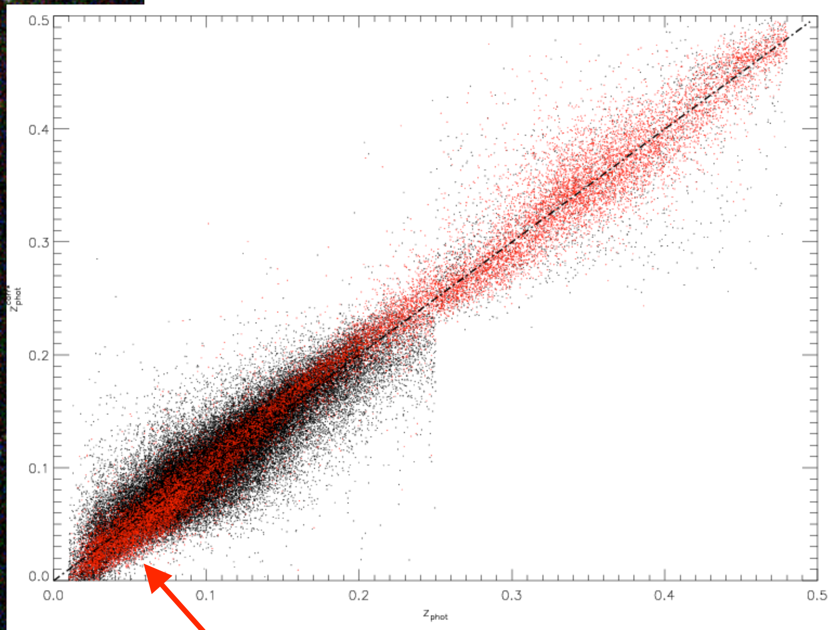
32 million galaxies in the nearby universe (z<0.48)

1.2 million objects in LRG sample

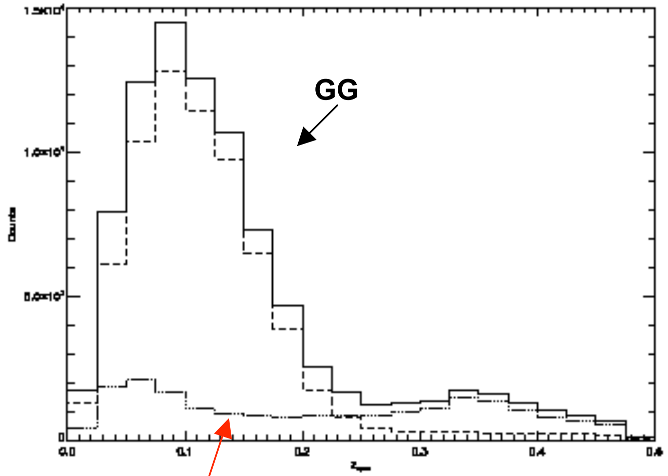
Errors are definitely gaussian



Spectroscopic redshifts characterization – General Galaxies Catalogue



Contamination of LRG at small z due to nearby low luminosity objects



Error in far sample for non-LRG is:

0.0363

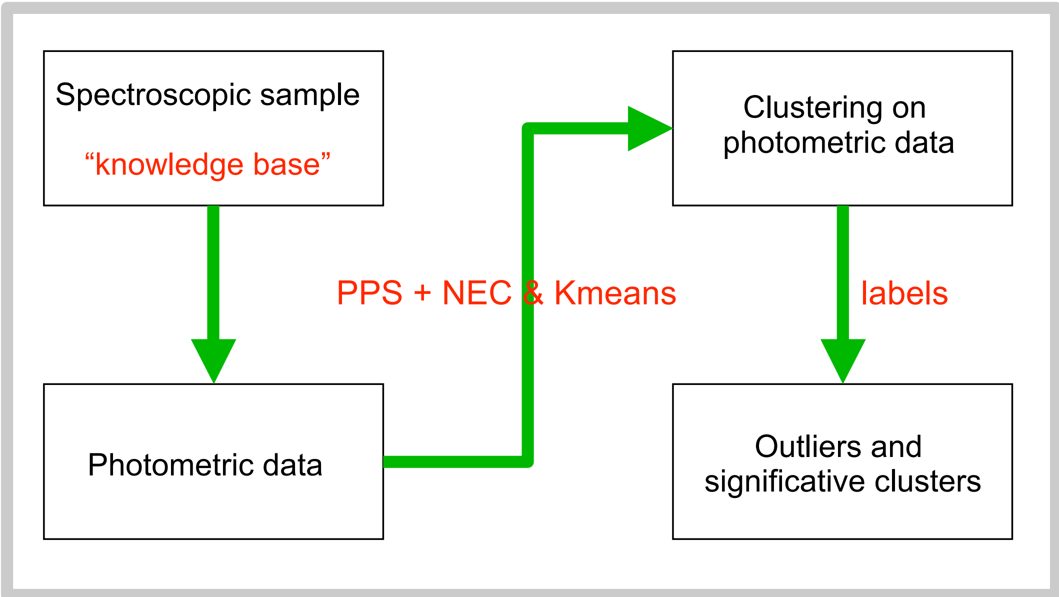
ADA-IV
marseille
2006

Our goal is an objective classifier which can achieve spectroscopic-like classification using only photometric attributes of objects.

Id est, a statistical device aimed at discovering unknown correlation between points (sources) in a photometric only parameter-space using clustering techniques.
Our choice was an unsupervised (no a-priori categories) neural network-based combination of algorithm:

PPS (Probabilistic Principal Surfaces)+NEC (Negentropy Clustering) & Kmeans

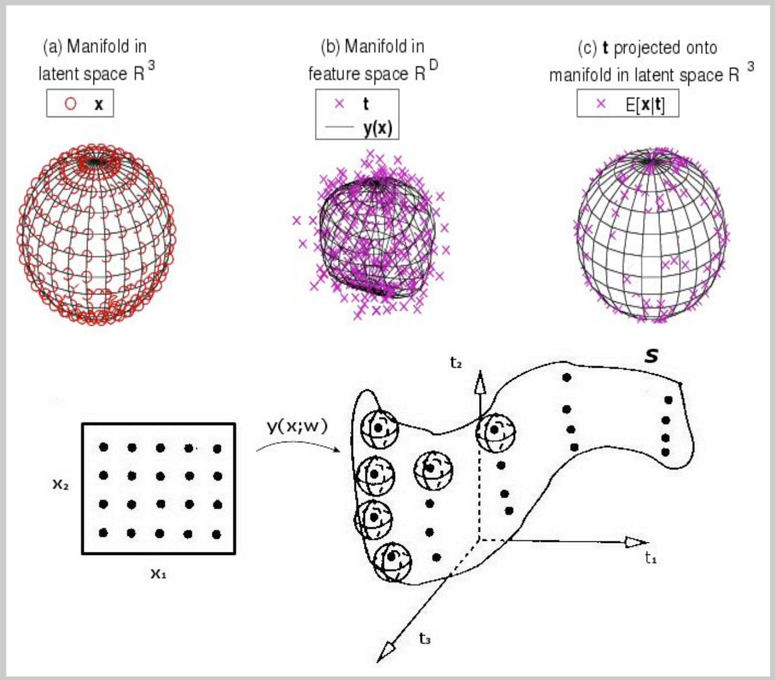
See next slide...



And maybe something interesting...

We need a "knowledge base": spectroscopic measured features (in our case, spectral classification represented by specClass) are needed and will be used as labels, before applying clustering to the only photometric objects.

Brief sketch of PPS and NEC



PPS: the Beauty of Spheres

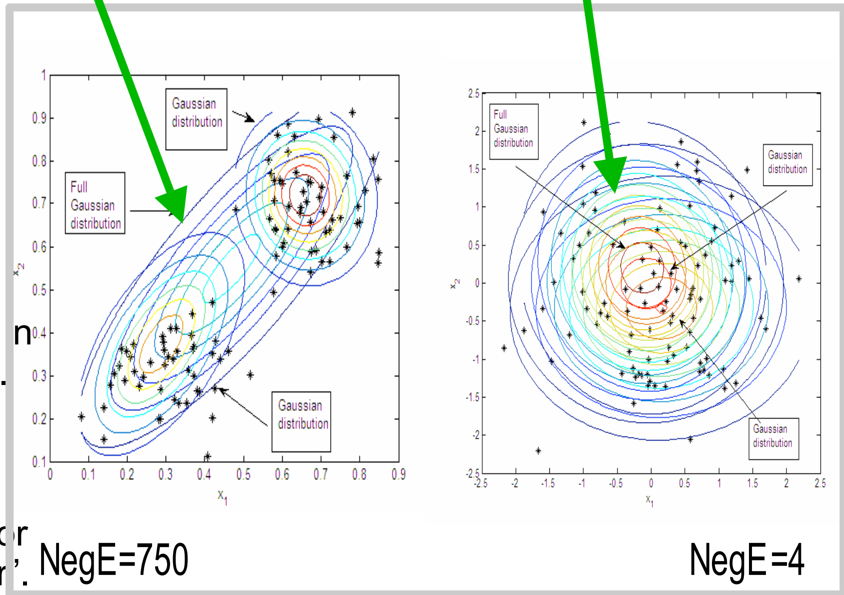
The original *m*-dimensional data space is mapped into a lower *n*-dimensional space, called "latent space". Visualization ease as a spherical manifold is fitted to the data, then projected into the manifold in R^3 and plotted as points on the sphere surface.
Each latent variable on the sphere is responsible for a number of projected points, which form a "cluster".

NEC: a matter of Gaussians

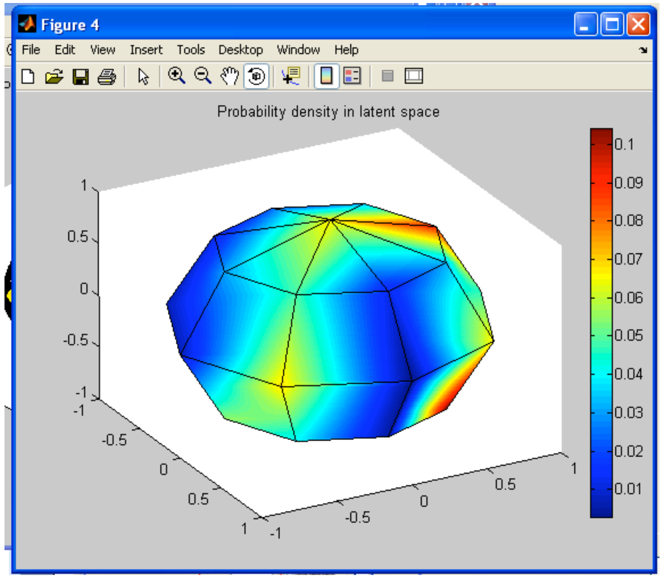
Clustering method based on the "neg-entropy" NegE, a measure of non gaussianity of a variable. If *A* is gaussian, then NegE(*A*) = 0. Given a threshold *d:*

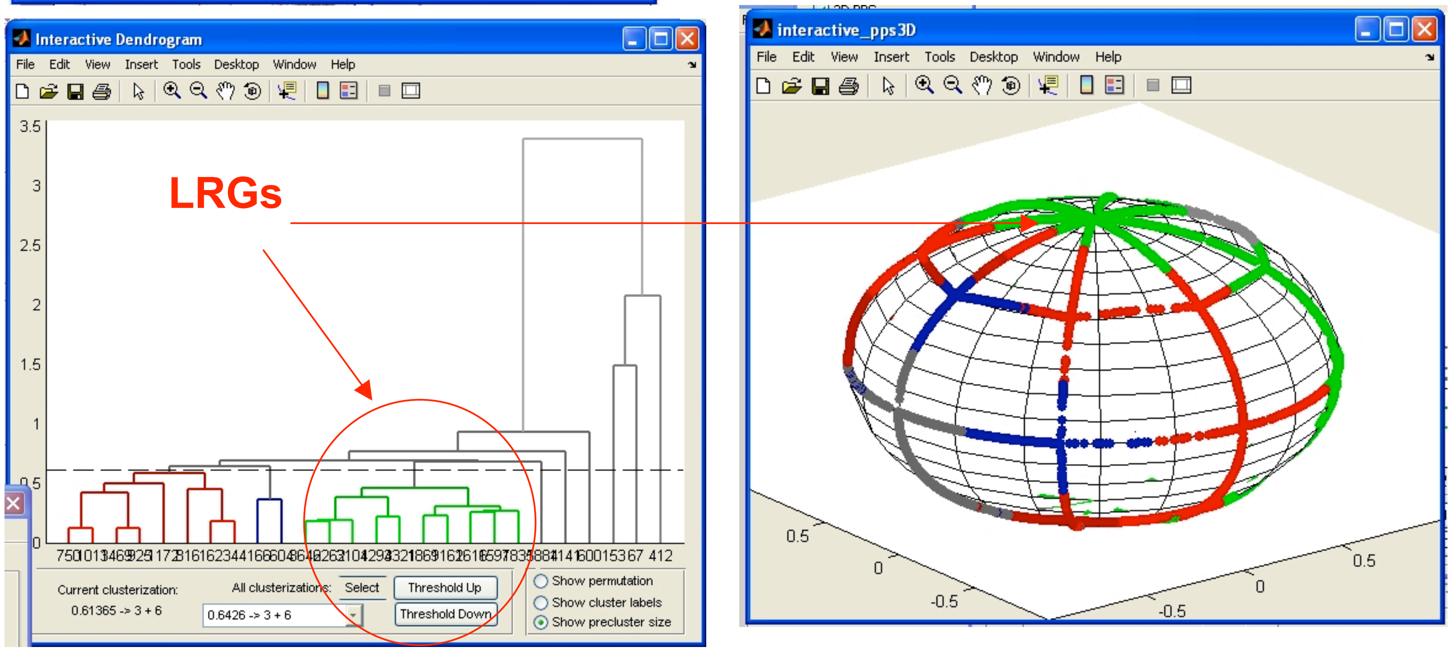
If NegE(*A* U *B*) < *d*, then clusters *A* and *B* are replaced by cluster *A* U *B*

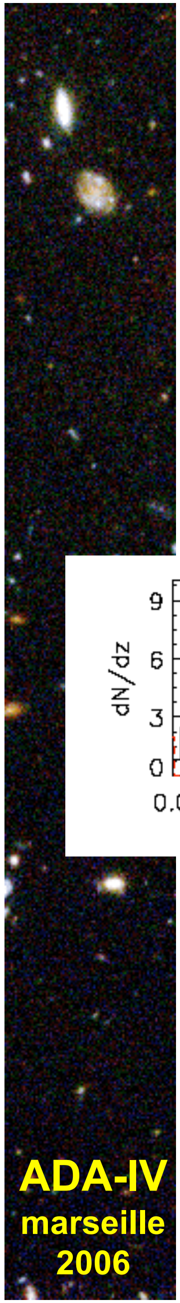
Not replaced! Replaced!



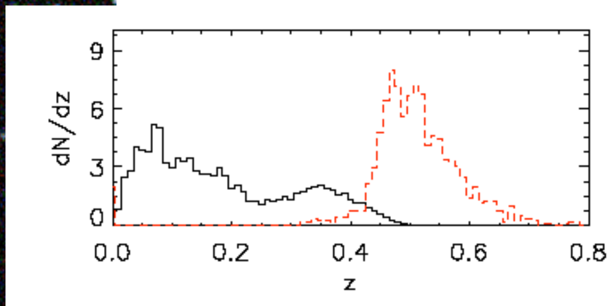
Preliminary result show that using Phot_z removes degeneracy….

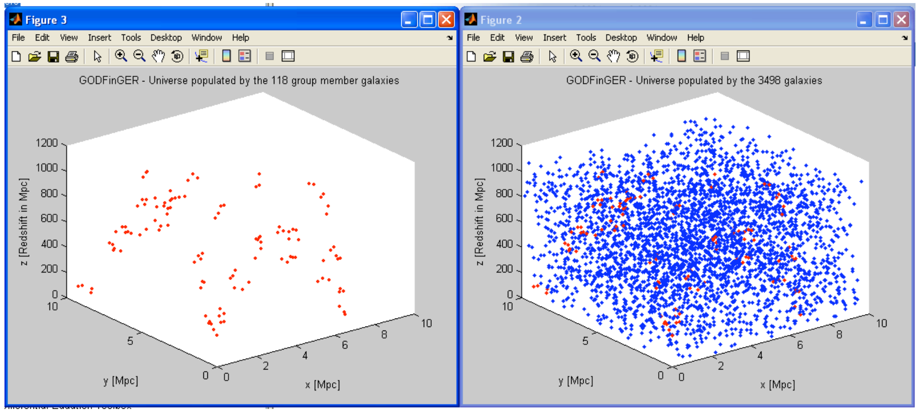






- **Extension of base of knowledge by including other and deeper survey data (e.g. 2dF, AEGIS, etc.)**
- **Extension of photometric base line by including NIR/FIR survey data (e.g. UKIDDS, Spitzer, etc.)**
- **Construction of catalogues of structures is in progress with unsupervised clustering algorithms**

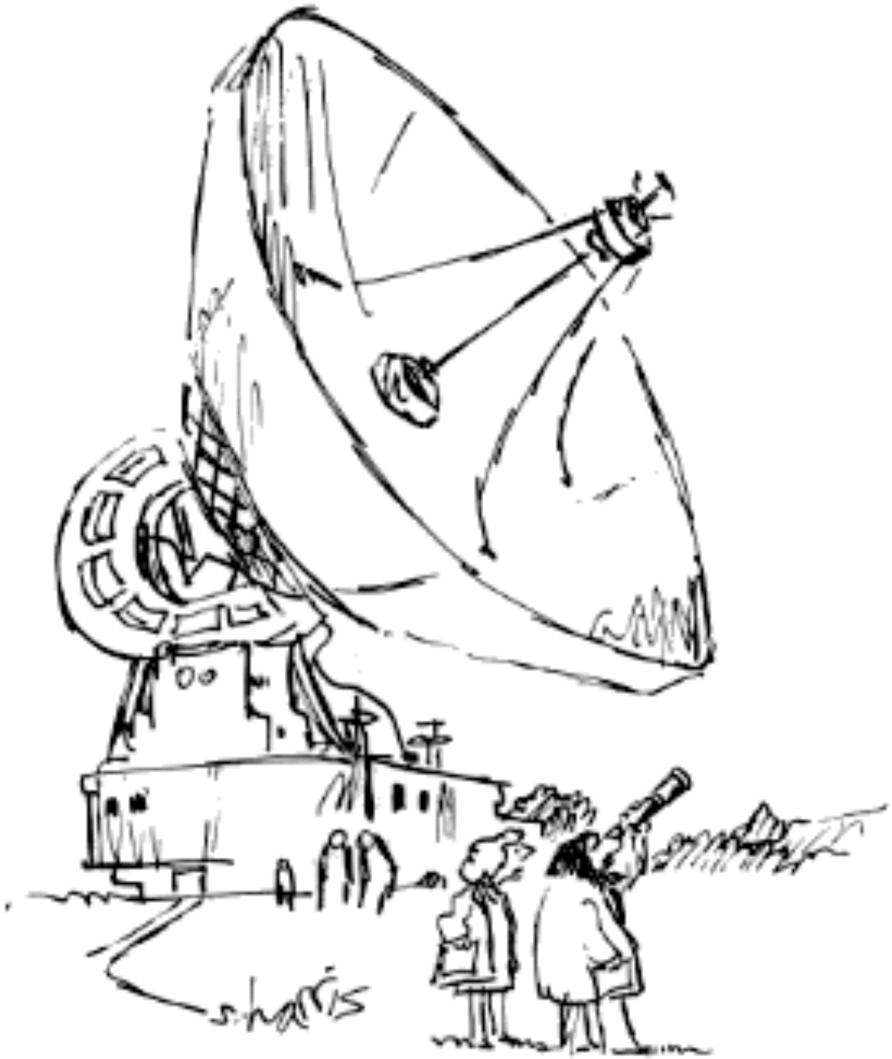




Preliminary tests show that $\sigma \approx 0.010$

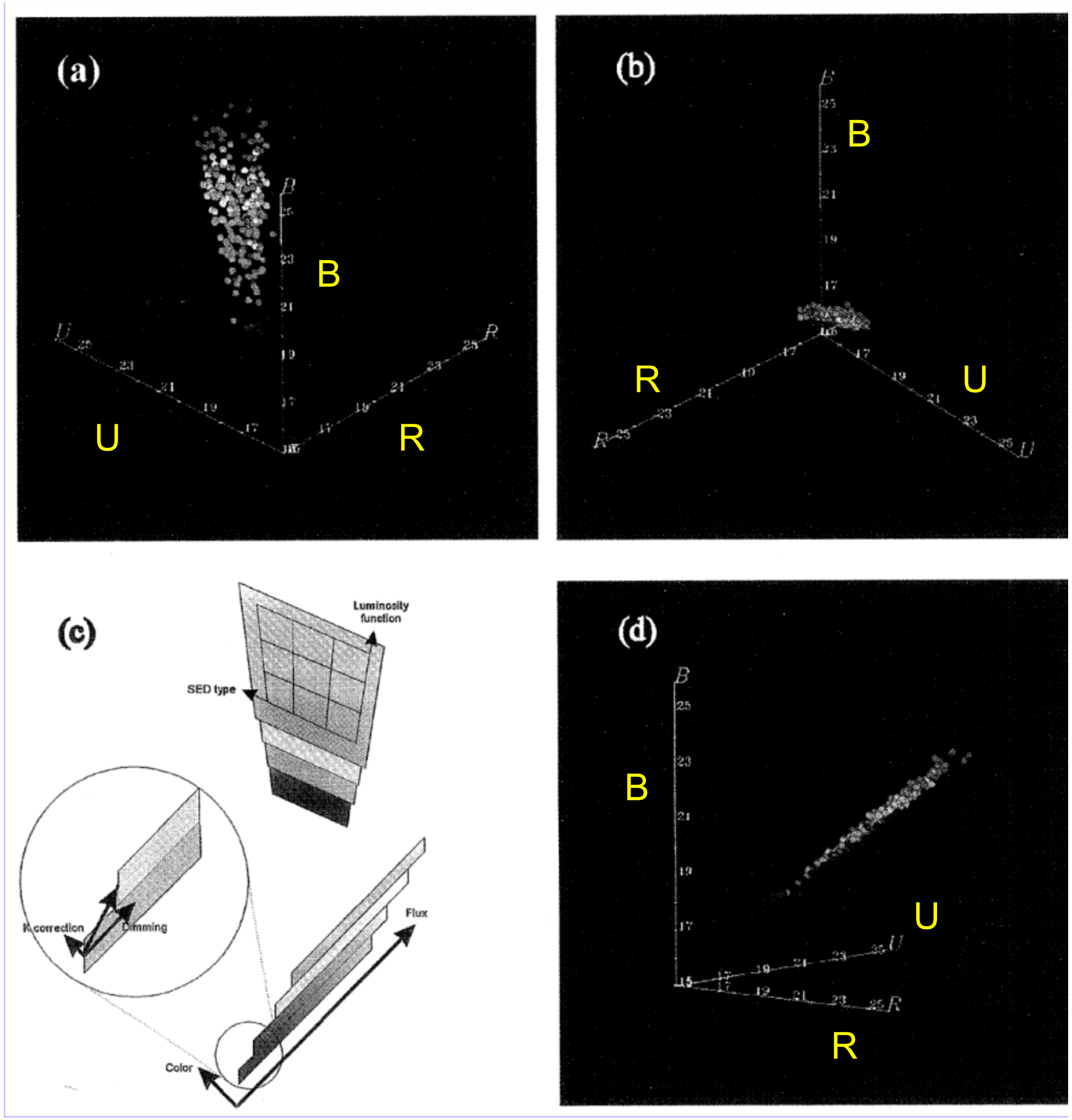
over the range $0.01 < z < 1.5$

Should be achieavable



"Just checking."

THE END


(a)
B
U
R
(b)
B
R
U
(c)
Luminosity function
SED type
K correction
Dimming
Flux
Color
(d)
B
U
R