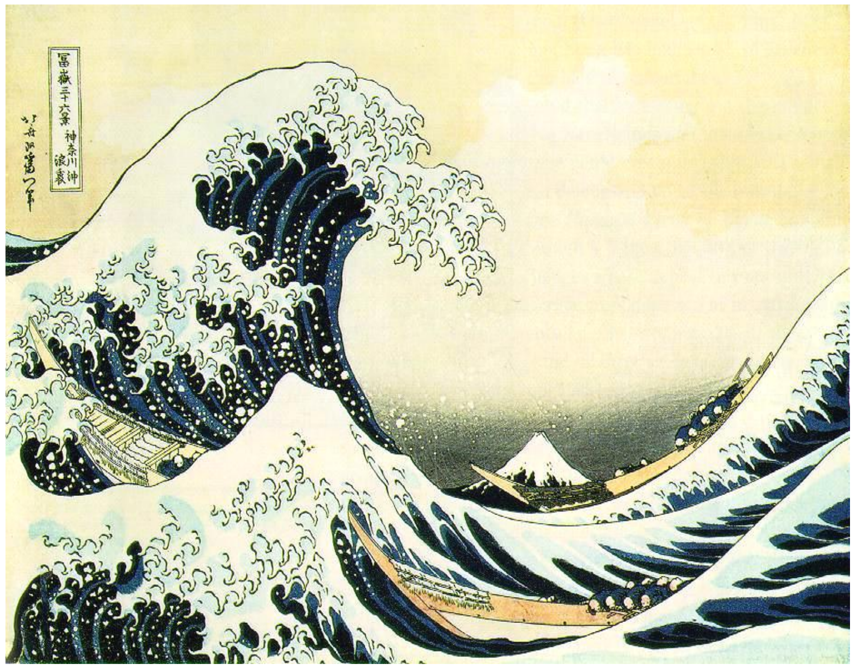
Mining Massive Data Sets in data rich sciences

astrophysics: a study case of how to face the modern data tsunami



M. Brescia[1], G. Longo[2], F. Pasian[3]

1- INAF – Astronomical Observatory of Capodimonte in Napoli (longo@na.infn.it)
2 - Department of Physical Sciences - University Federico II Napoli
3 - INAF – Information systems unit & Astronomical Observatory of Trieste

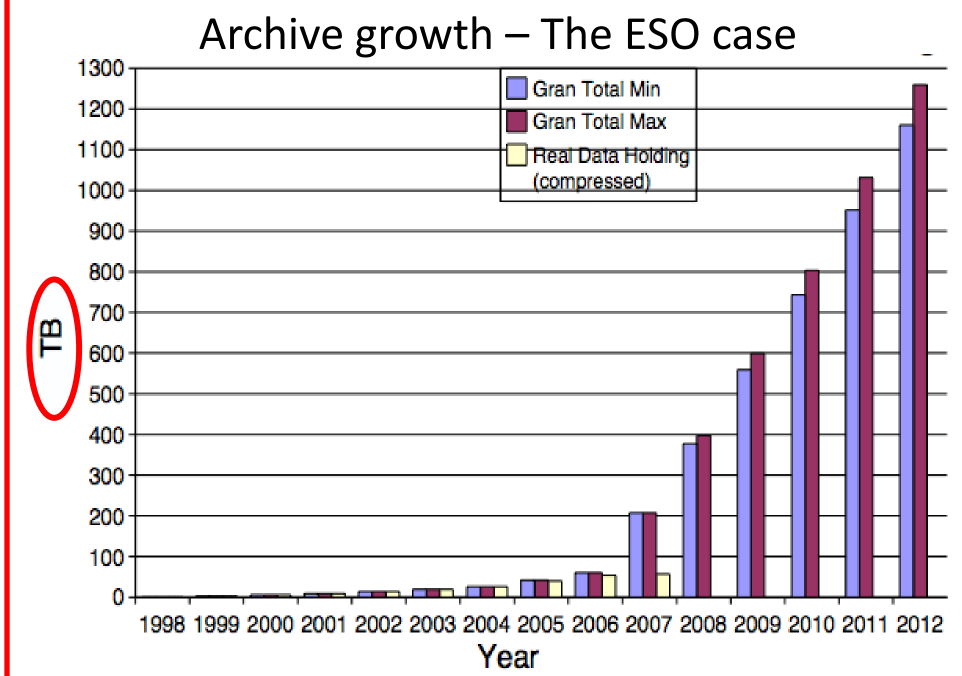
An overview of the topics:

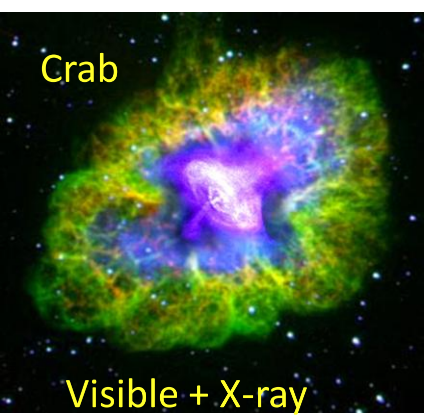
- Information Technology revolution and science in the exponential world: i.e. coping with the data avalanche
  - The Virtual Observatory: a new type of a scientific research environment
    - Massive data sets and a new scientific methodology
      - DAME project: Data Mining and Exploration
        - Some general considerations on the future

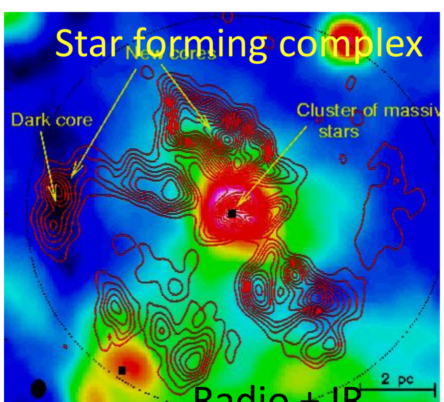
Astrophysics as a data rich science

- Telescopes (ground- and space-based, covering the full electromagnetic spectrum)
- Instruments (telescope/band dependent)

- **Large digital sky surveys** are becoming the dominant source of data in astronomy: ~ 10-100 TB/survey (soon PB), ~ 10^6 - 10^9 sources/survey, many wavelengths…
- **Data sets many orders of magnitude larger, more complex, and more homogeneous than in the past**





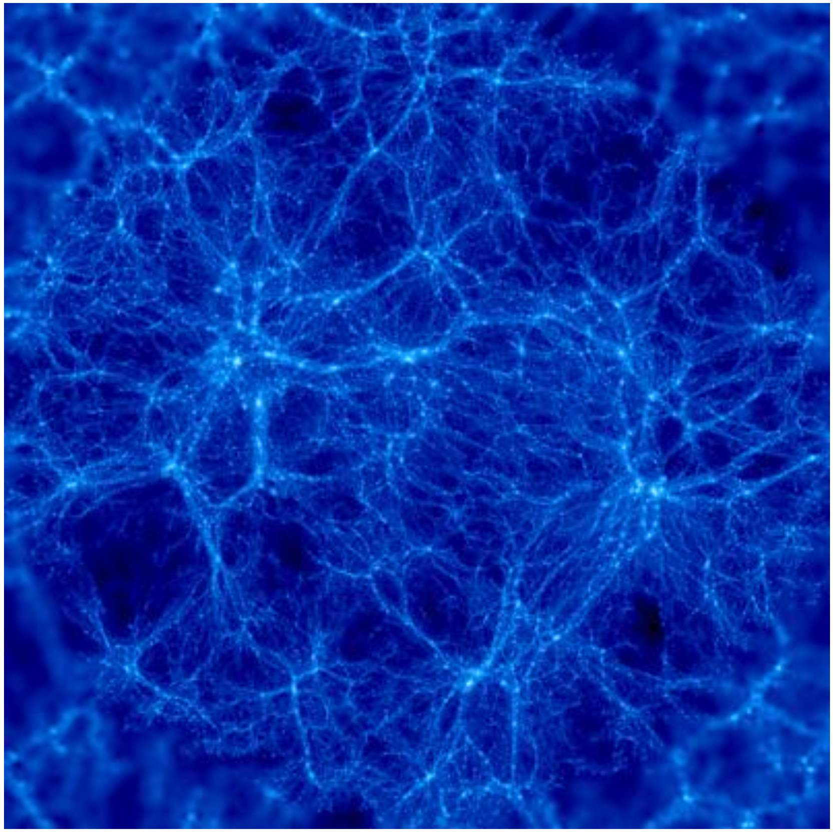


Panchromatic Views of the Universe:
Data Fusion - A More Complete, Less Biased Picture

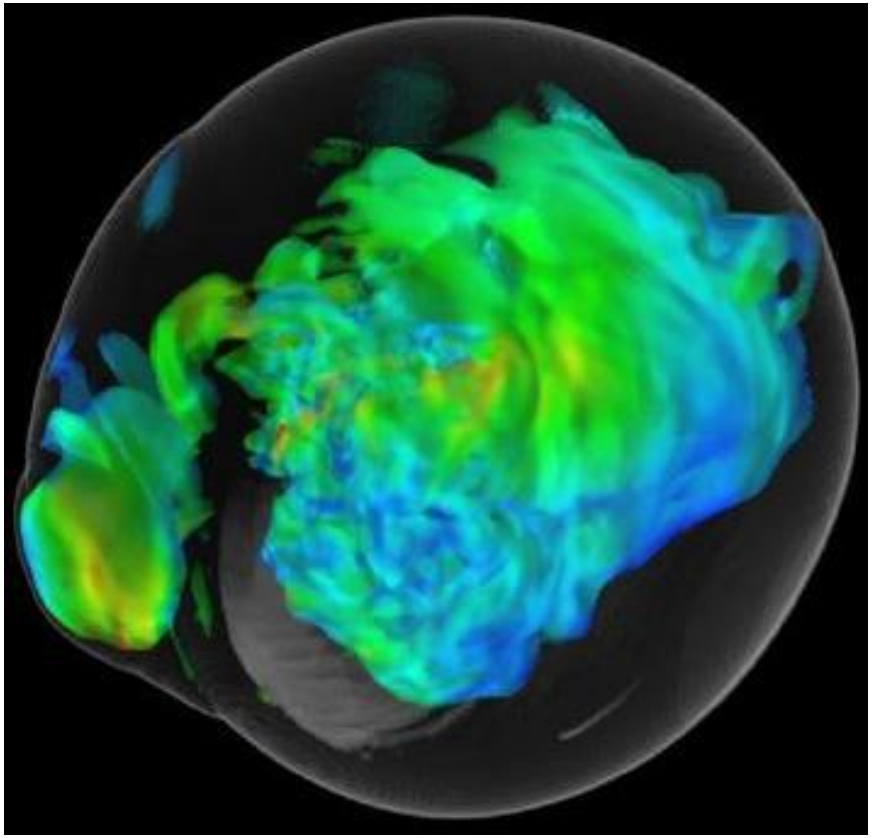
3

2. The astronomical data tsunami:

Theoretical Simulations Are Becoming More Complex and Generate TB's of Data …



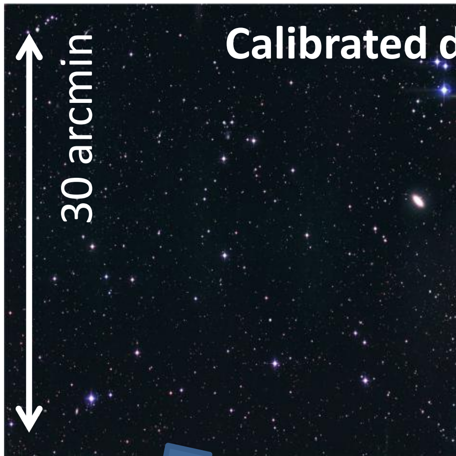
Structure formation in the Universe



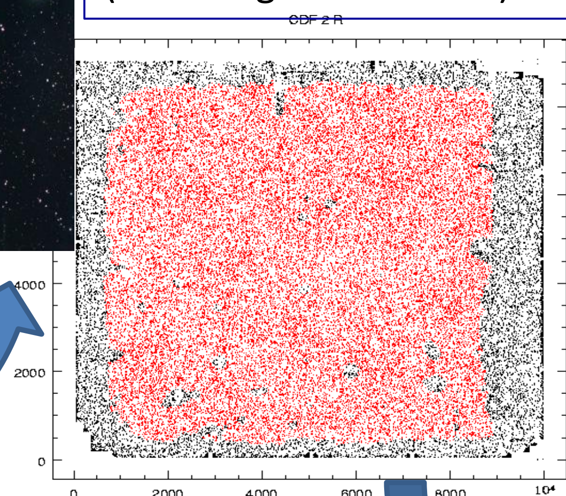
Supernova explosion instabilities

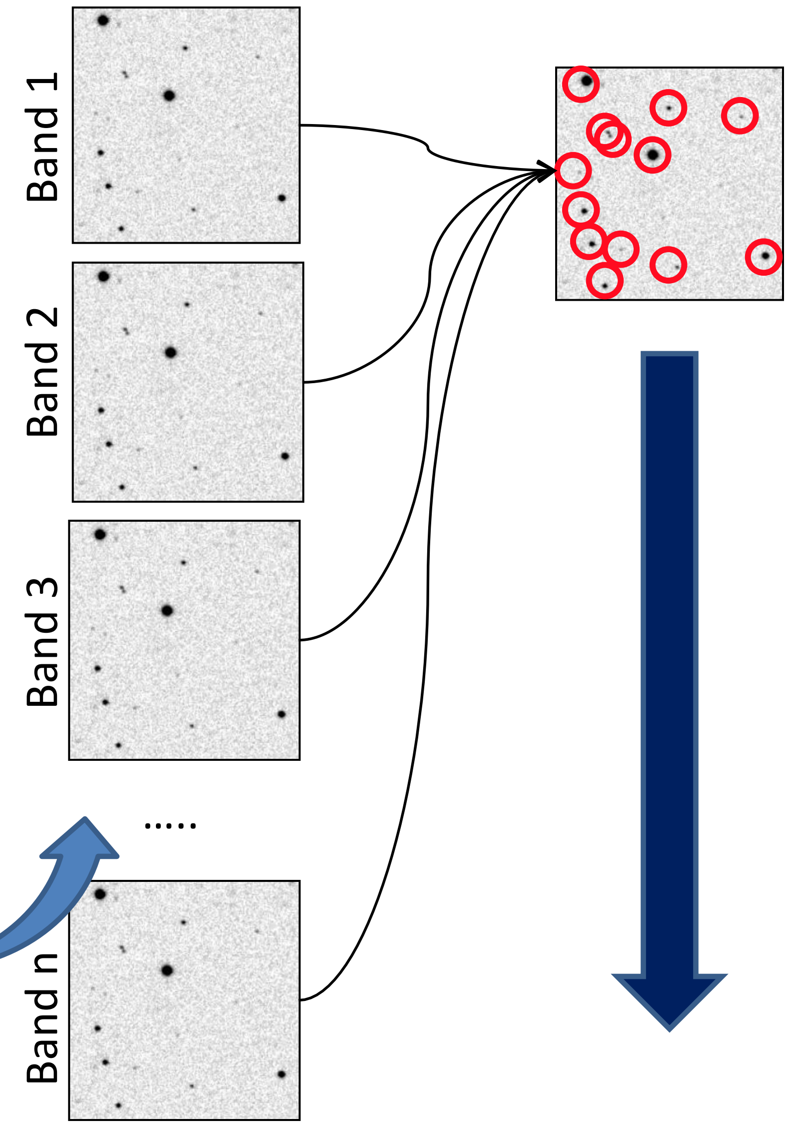
Comparing the massive, complex output of such simulations to equally massive and complex data sets is a non-trivial problem!

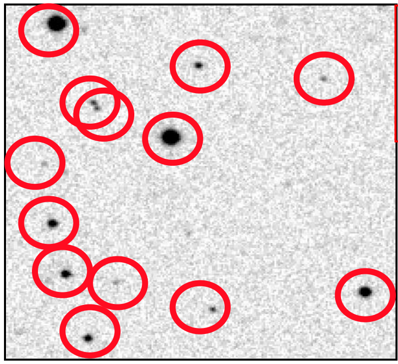
3. The data mining perspective. An example of Data complexity: the parameter space



1/160.000 of the sky, not sodeep (25.0 in r)
55.000 detected sources
(0.75 mag above m lim)







Detect sources and measure their attributes (brightness, position, shapes, etc.)

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

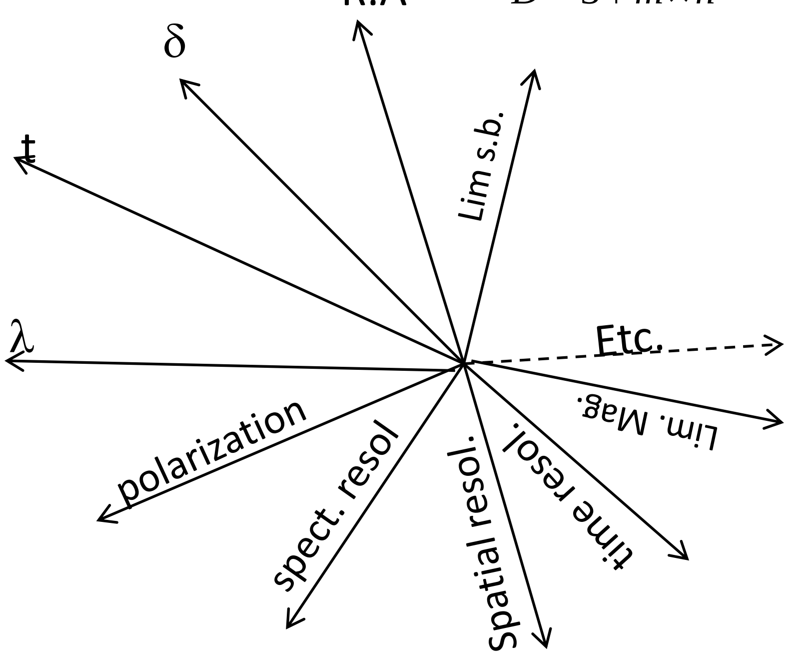
$$p^1 = \{RA^1, \delta^1, t, \{\lambda_1, \Delta\lambda_1, f_1^{1,1}, \Delta f_1^{1,1}, \ldots, f_1^{1,m}, \Delta f_1^{1,m}\}, \ldots, \{\lambda_n, \Delta\lambda_n, f_n^{1,1}, \Delta f_n^{1,1}, \ldots, f_n^{1,m}, \Delta f_n^{1,m}\}\}$$

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

....

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

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



PARAMETER SPACE

From the Data Mining point of view, a**ny observed (simulated) datum *p* defines a point (region) in a subset of R^N.**

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

3. Information Technology & New Science

Due to new instruments and new diagnostic tools, the information volume grows exponentially

Most data will never be seen by humans!

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

Information complexity is also increasing greatly

Most knowledge hidden behind data complexity is lost

Most (all) empirical relationships known so far depend on 3 parameters ….
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

Extracting knowledge

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

MASSIVE, COMPLEX DATA SETS with:

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

The computational cost of Data Mining:

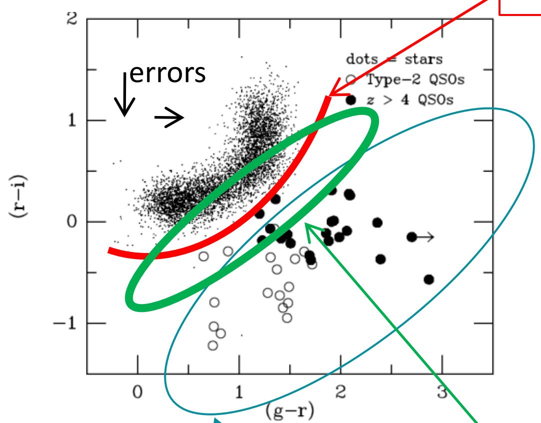
N = no. of data vectors, D = no. of data dimensions
K = no. of clusters chosen, K_{max} = max no. of clusters tried
I = no. of iterations, M = no. of Monte Carlo trials/partitions

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 > ~ $(NxD)^3$

Lots of computing power

More dimensions allow better disentanglement

Traditional way to look for candidate QSO in 3 band survey

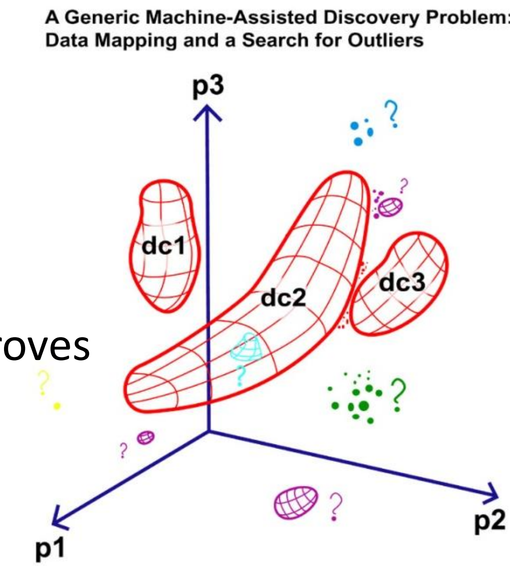


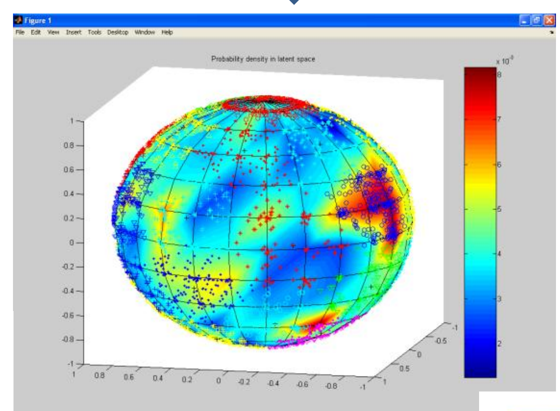
Cutoff line

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

From data to knowledge: KDD

Knowledge Discovery in Databases

Data Gathering (e.g., from sensor networks, telescopes…)

Data Farming:

- Storage/Archiving
- Indexing, Searchability
- Data Fusion, Interoperability, ontologies, etc.

Database technologies

Data Mining (or Knowledge Discovery in Databases):

- Pattern or correlation search
- Clustering analysis, automated classification
- Outlier / anomaly searches
- Hyperdimensional visualization

Key mathematical issues

Data understanding

- Computer aided understanding
- KDD
- Etc.

Ongoing research

New Knowledge

OK, So …

Which was the answer of the astronomical community?

The Virtual Observatory (VObs)

VObs Represents a New Type of a Scientific Organization for the era of information abundance

courtesy of P. Quinn

- It is inherently ***distributed,*** and web-centric
- It is fundamentally based on a ***rapidly developing technology*** (IT/CS)
- ***It transcends the traditional boundaries*** between different wavelength regimes, agency domains, etc.
- It has an ***unusually broad range of constituents*** and interfaces
- It is inherently ***multidisciplinary***

Vobs standards for interoperability: UCD, VO-Table, ontology, etc..

UCD (Unified Content Descriptor): describing in unique & standard way attributes contained in data tables

```
<DATA>
<TABLEDATA>
<TR>
   <TD>010.68</TD><TD>+41.27</TD>
    <TD>N 224</TD><TD>-297</TD>
   </TR>
   <TR>
    <TD>287.43</TD><TD>-63.85</TD>
    <TD>6</TD><TD>10.4</TD>
   </TR>
```

```
  </TABLEDATA>
 </DATA>
```

```
<?xml version="1.0"?>
<VOTABLE version="1.1" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:noNamespaceSchemaLocation="http://www.ivoa.net/xml/VOTable/VOTable/v1.1">
```

```
<RESOURCE name="myFavouriteGalaxies">
<DESCRIPTION>Velocities and Distance estimations</DESCRIPTION>
<PARAM name="Telescope" datatype="float" ucd="phys.size;instr.tel" unit="m" value="3.6"/>
```

```
<FIELD name="RA" ID="col1" ucd="pos.eq.ra;meta.main" ref="J2000" datatype="float"
    width="6" precision="2" unit="deg"/>
<FIELD name="Dec" ID="col2" ucd=" "pos.eq.dec;meta.main" ref="J2000" datatype="float"
    width="6" precision="2" unit="deg"/>
<FIELD name="R" ID="col6" ucd="phys.distance" datatype="float" width="4"
    precision="1" unit="Mpc">
<DESCRIPTION>Distance of Galaxy, assuming  H=75km/s/Mpc</DESCRIPTION>
</FIELD>
```

Data mining is …

There are known knowns,
There are known unknowns, and
There are unknown unknowns

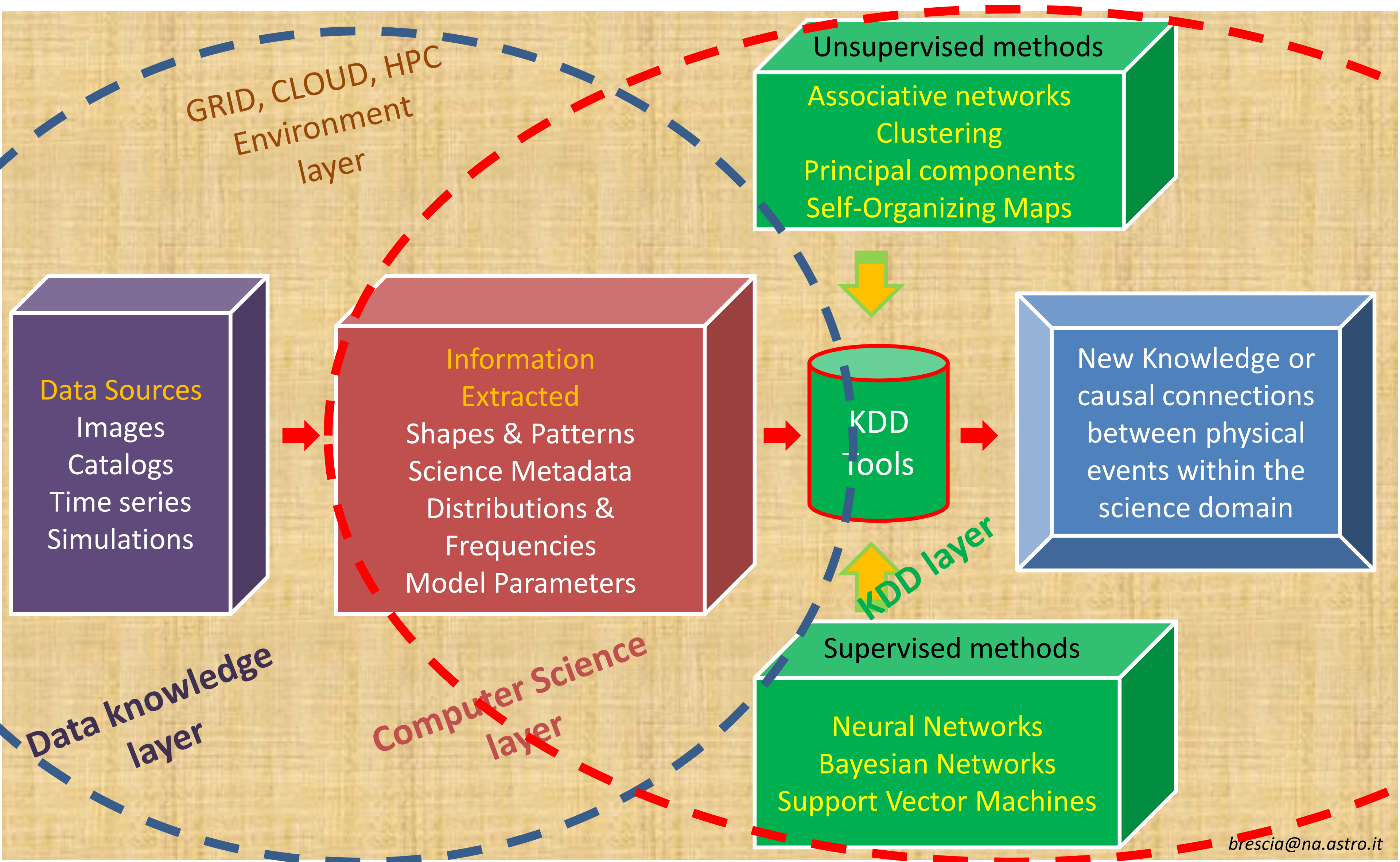
Classification
Morphological classification of galaxies
Star/galaxy separation, etc.

Regression
Photometric redshifts

Clustering
Search for peculiar and rare objects,
Etc.

Vobs standards and infrastructure

Data mining level



What is DAME

DAME is a joint effort between University Federico II, INAF-OACN, and Caltech aimed at implementing (as web application) a scientific gateway for data analysis, exploration, mining and visualization tools, on top of virtualized distributed computing environment.

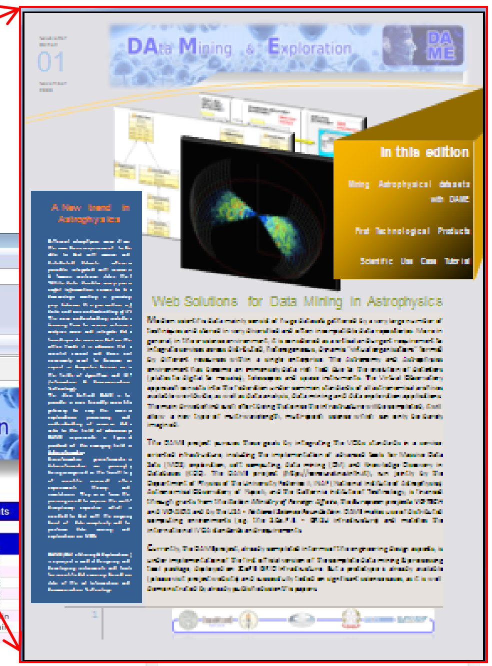
http://voneural.na.infn.it/

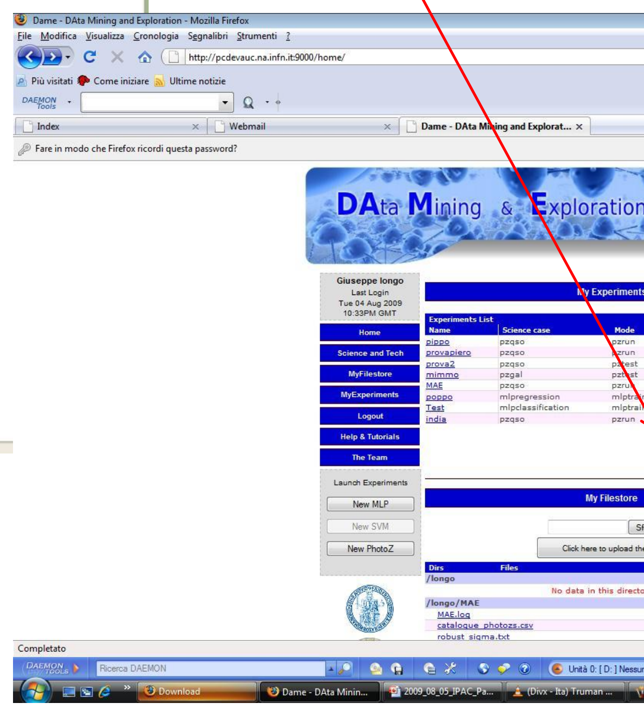
Technical and management info

Documents

Science cases

Newsletter



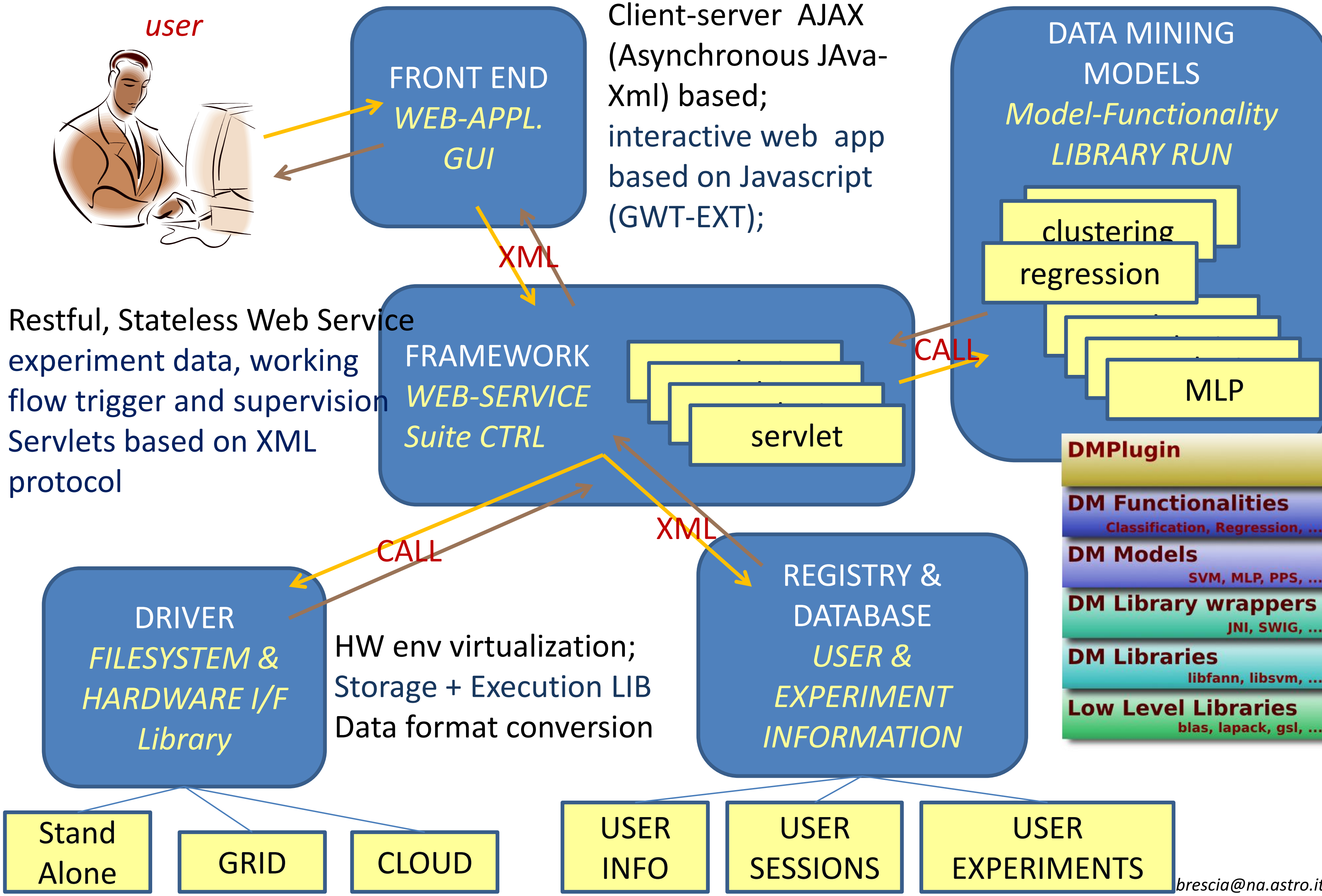


http://dame.na.infn.it/

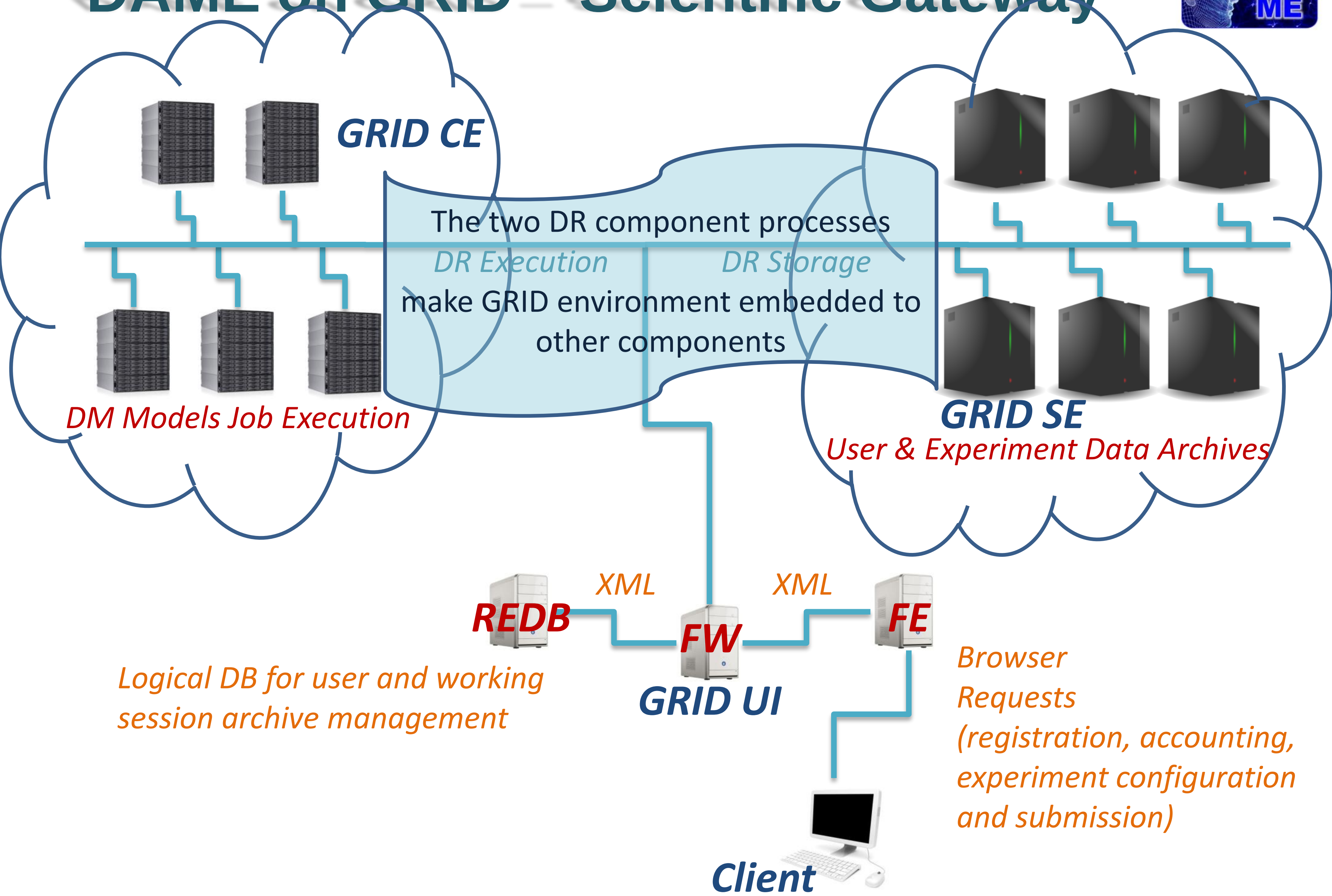
Web application PROTOTYPE

brescia@na.astro.it

The DAME architecture



DAME on GRID – Scientific Gateway



How to spread the word within the community

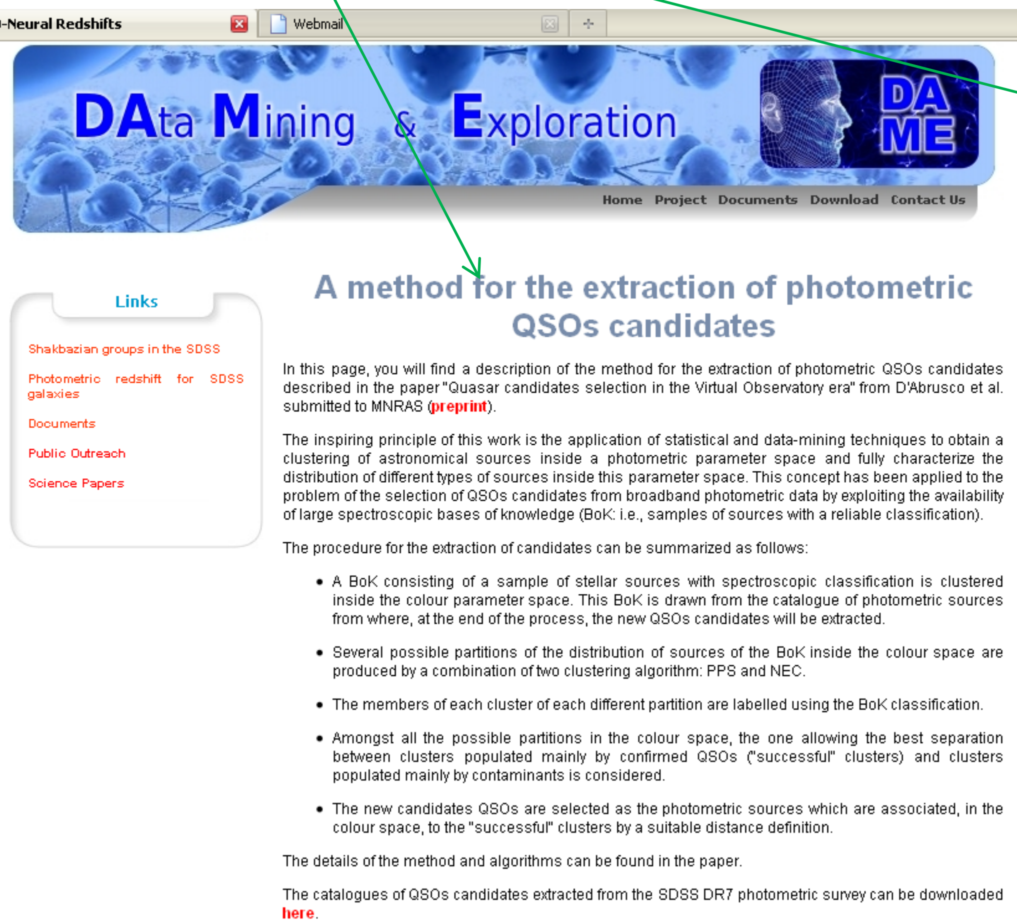
In parallel with the Suite R&D process, all data processing algorithms (foreseen to be plugged in) have been massively tested on real astrophysical cases.

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

Technical and management info

Documents

Science cases



0-Neural Redshifts | Webmail

DAta Mining & Exploration — DA ME

Home Project Documents Download Contact Us

Links

- Shakbazian groups in the SDSS
- Photometric redshift for SDSS galaxies
- Documents
- Public Outreach
- Science Papers

A method for the extraction of photometric QSOs candidates

In this page, you will find a description of the method for the extraction of photometric QSOs candidates described in the paper "Quasar candidates selection in the Virtual Observatory era" from D'Abrusco et al. submitted to MNRAS (**preprint**).

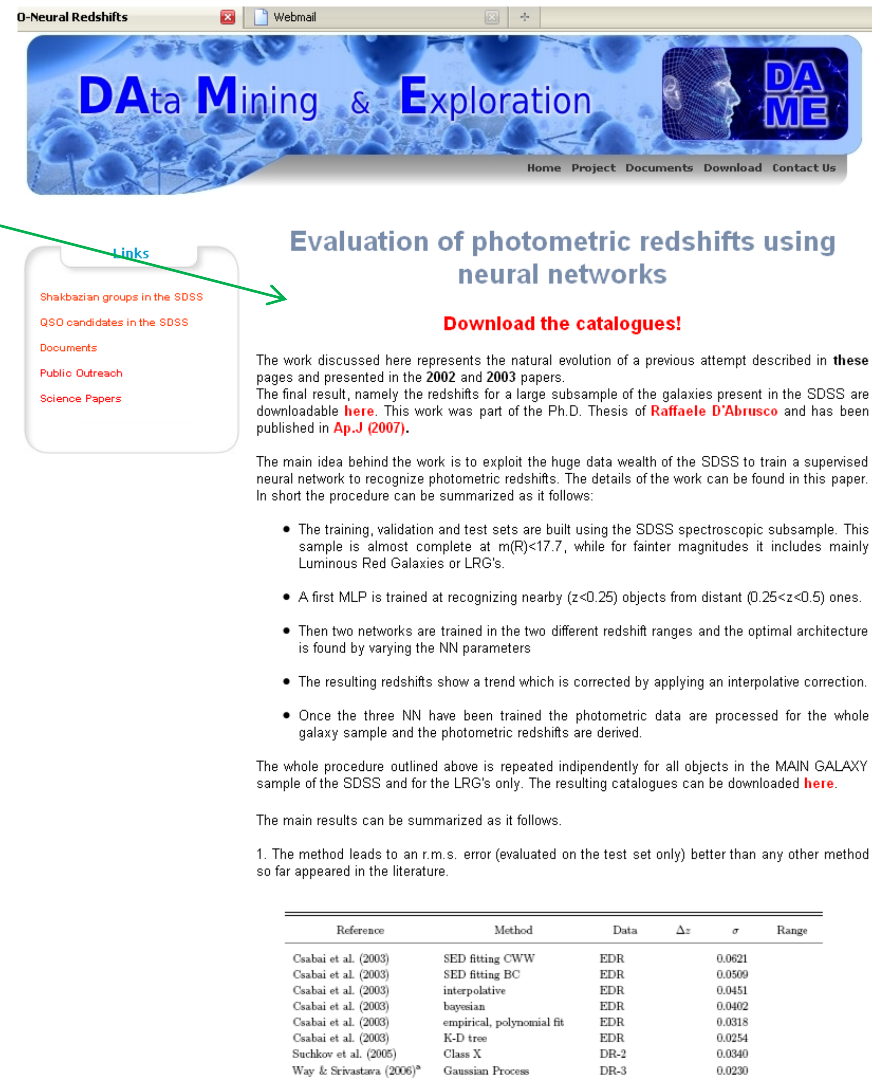
The inspiring principle of this work is the application of statistical and data-mining techniques to obtain a clustering of astronomical sources inside a photometric parameter space and fully characterize the distribution of different types of sources inside this parameter space. This concept has been applied to the problem of the selection of QSOs candidates from broadband photometric data by exploiting the availability of large spectroscopic bases of knowledge (BoK: i.e., samples of sources with a reliable classification).

The procedure for the extraction of candidates can be summarized as follows:

- A BoK consisting of a sample of stellar sources with spectroscopic classification is clustered inside the colour parameter space. This BoK is drawn from the catalogue of photometric sources from where, at the end of the process, the new QSOs candidates will be extracted.
- Several possible partitions of the distribution of sources of the BoK inside the colour space are produced by a combination of two clustering algorithm: PPS and NEC.
- The members of each cluster of each different partition are labelled using the BoK classification.
- Amongst all the possible partitions in the colour space, the one allowing the best separation between clusters populated mainly by confirmed QSOs ("successful" clusters) and clusters populated mainly by contaminants is considered.
- The new candidates QSOs are selected as the photometric sources which are associated, in the colour space, to the "successful" clusters by a suitable distance definition.

The details of the method and algorithms can be found in the paper.

The catalogues of QSOs candidates extracted from the SDSS DR7 photometric survey can be downloaded **here**.



0-Neural Redshifts | Webmail

DAta Mining & Exploration — DA ME

Home Project Documents Download Contact Us

Links

- Shakbazian groups in the SDSS
- QSO candidates in the SDSS
- Documents
- Public Outreach
- Science Papers

Evaluation of photometric redshifts using neural networks

Download the catalogues!

The work discussed here represents the natural evolution of a previous attempt described in **these** pages and presented in the **2002** and **2003** papers.
The final result, namely the redshifts for a large subsample of the galaxies present in the SDSS are downloadable **here**. This work was part of the Ph.D. Thesis of **Raffaele D'Abrusco** and has been published in **Ap.J (2007)**.

The main idea behind the work is to exploit the huge data wealth of the SDSS to train a supervised neural network to recognize photometric redshifts. The details of the work can be found in this paper. In short the procedure can be summarized as it follows:

- The training, validation and test sets are built using the SDSS spectroscopic subsample. This sample is almost complete at m(R)<17.7, while for fainter magnitudes it includes mainly Luminous Red Galaxies or LRG's.
- A first MLP is trained at recognizing nearby (z<0.25) objects from distant (0.25<z<0.5) ones.
- Then two networks are trained in the two different redshift ranges and the optimal architecture is found by varying the NN parameters
- The resulting redshifts show a trend which is corrected by applying an interpolative correction.
- Once the three NN have been trained the photometric data are processed for the whole galaxy sample and the photometric redshifts are derived.

The whole procedure outlined above is repeated indipendently for all objects in the MAIN GALAXY sample of the SDSS and for the LRG's only. The resulting catalogues can be downloaded **here**.

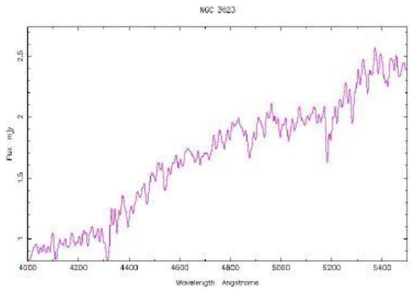
The main results can be summarized as it follows.

1. The method leads to an r.m.s. error (evaluated on the test set only) better than any other method so far appeared in the literature.

Reference	Method	Data	Δz	σ	Range
Csabai et al. (2003)	SED fitting CWW	EDR		0.0621	
Csabai et al. (2003)	SED fitting BC	EDR		0.0509	
Csabai et al. (2003)	interpolative	EDR		0.0451	
Csabai et al. (2003)	bayesian	EDR		0.0402	
Csabai et al. (2003)	empirical, polynomial fit	EDR		0.0318	
Csabai et al. (2003)	K-D tree	EDR		0.0254	
Suchkov et al. (2005)	Class X	DR-2		0.0340	
Way & Srivastava (2006)[a]	Gaussian Process	DR-3		0.0230	

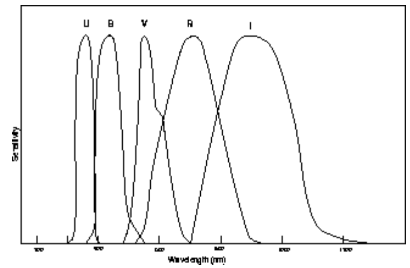
An EXAMPLE: photometric redshifts of SDSS galaxies

$$z \times c \equiv \frac{\Delta\lambda}{\lambda_0}$$



Galaxy spectrum - F(λ)

X



Photometric system - $S_i(\lambda)$

=

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

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

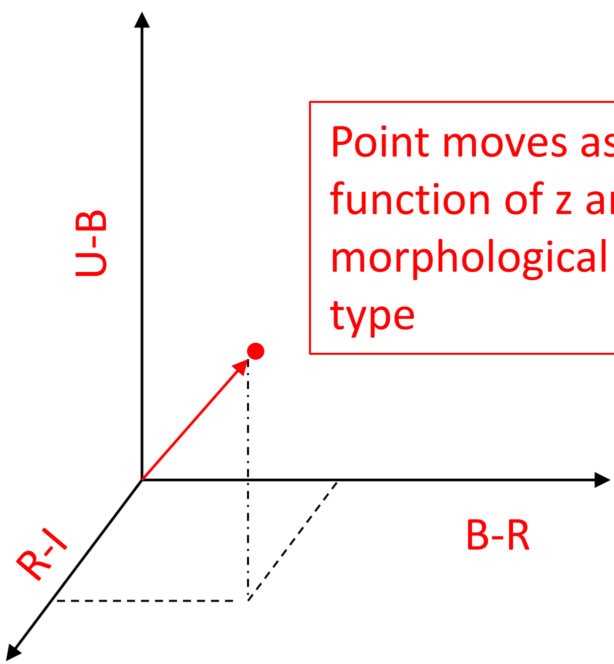
Etc...

Color indexes

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

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

etc.



Point moves as a function of z and morphological type

Phot-z are an inverse problem

Photometric redshifts: the DM approach

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 = Base of Knowledge

BoK (from Vobs)
(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 case:

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

UKIDDS: overlap with SDSS
3 infrared bands.

GALEX: overlap with SDSS
Ultraviolet bands;

SDSS

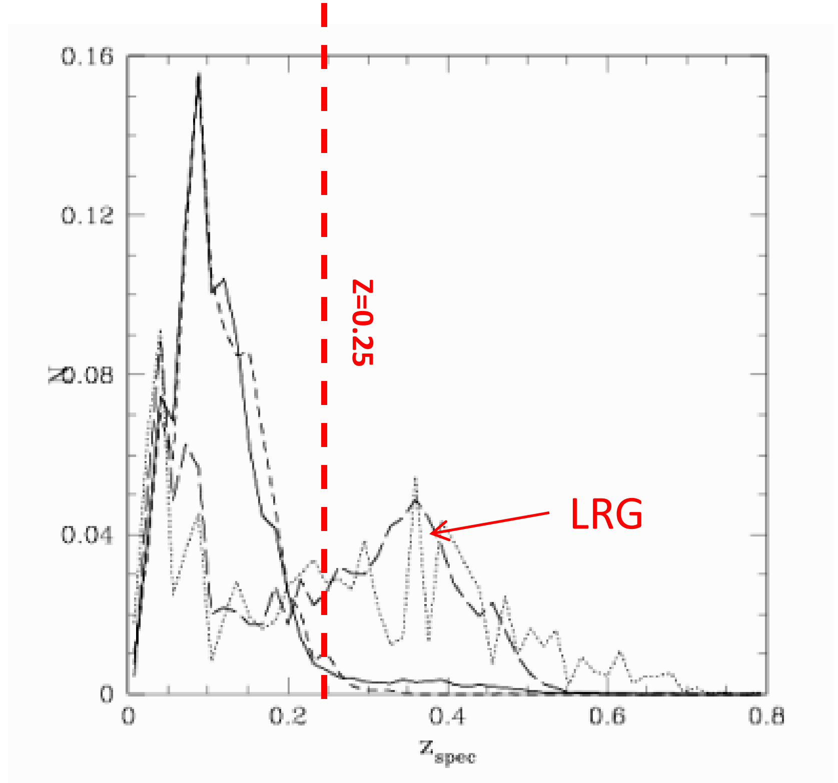


Fig. 1.— The spectroscopic redshift histogram for the SDSS main EDR (solid), the EDR LRG (long dash), the 2dF (short dash) and the CNOC2 sets.

SDSS-DR4/5 - SS

Training 60%

Validation 20%

Test set 20%

MLP, 1(5), 1(18)

0.01<Z<0.25

0.25<Z<0.50

99.6 % accuracy

MLP, 1(5), 1(23)

MLP, 1(5), 1(24)

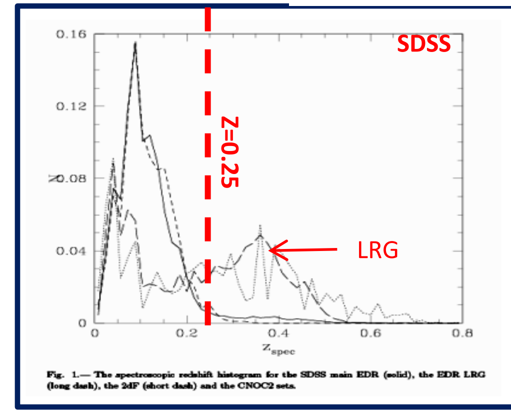
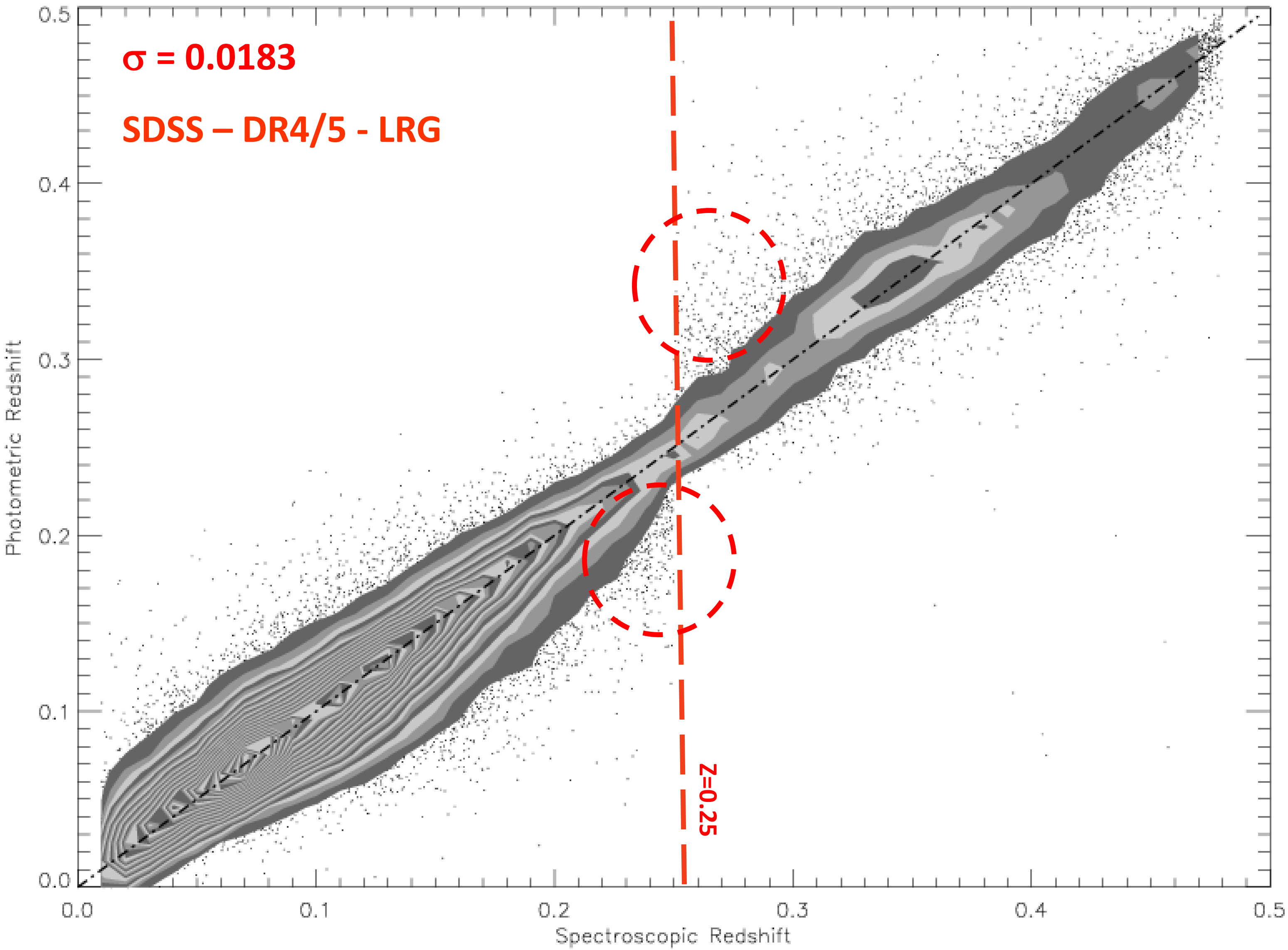
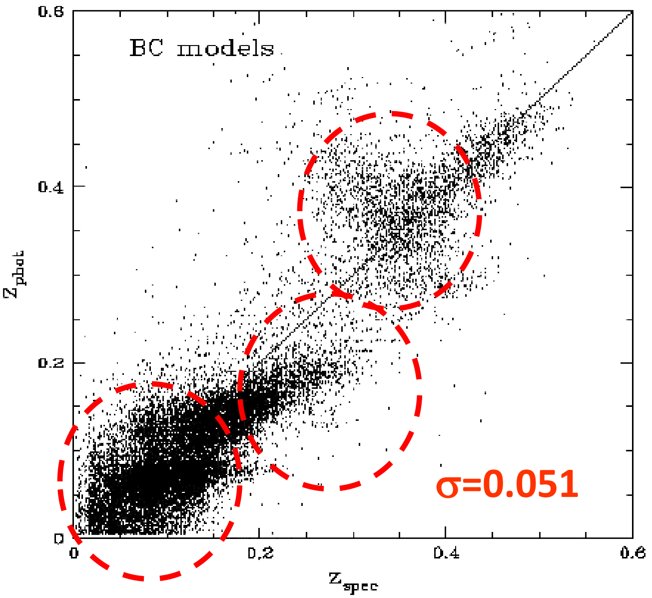


Fig. 1.— The spectroscopic redshift histogram for the SDSS main EDR (solid), the EDR LRG (long dash), the 2dF (short dash) and the CNOC2 sets.



D’Abrusco et al. 2007

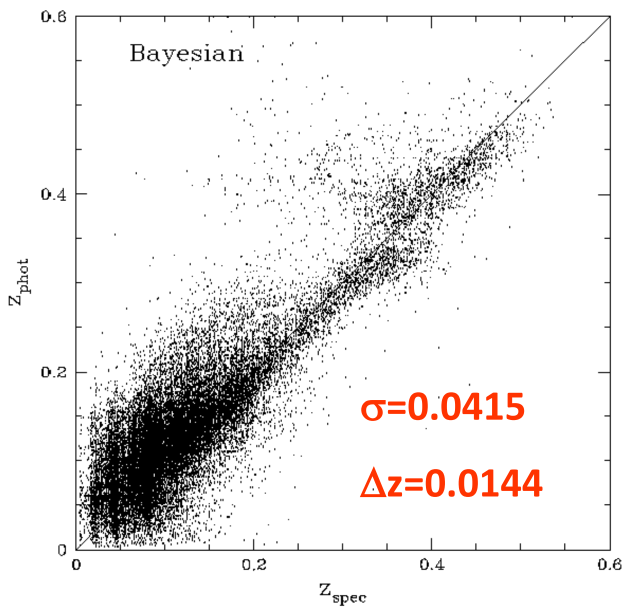
Traditional approaches: interpolation based on BoK



BoK from Spectral Energy Distribution (SED) fitting

Templates from synthetic colors obtained from theoretical SED's

Mapping function from simple interpolation



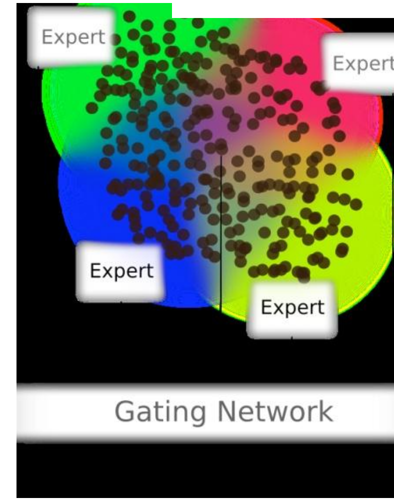
BoK from Spectral Energy Distribution (SED) fitting
Interpolative

Templates from synthetic colors obtained from theoretical SED's

Mapping function from Bayesian inference

What do we learn if the BoK is biased:

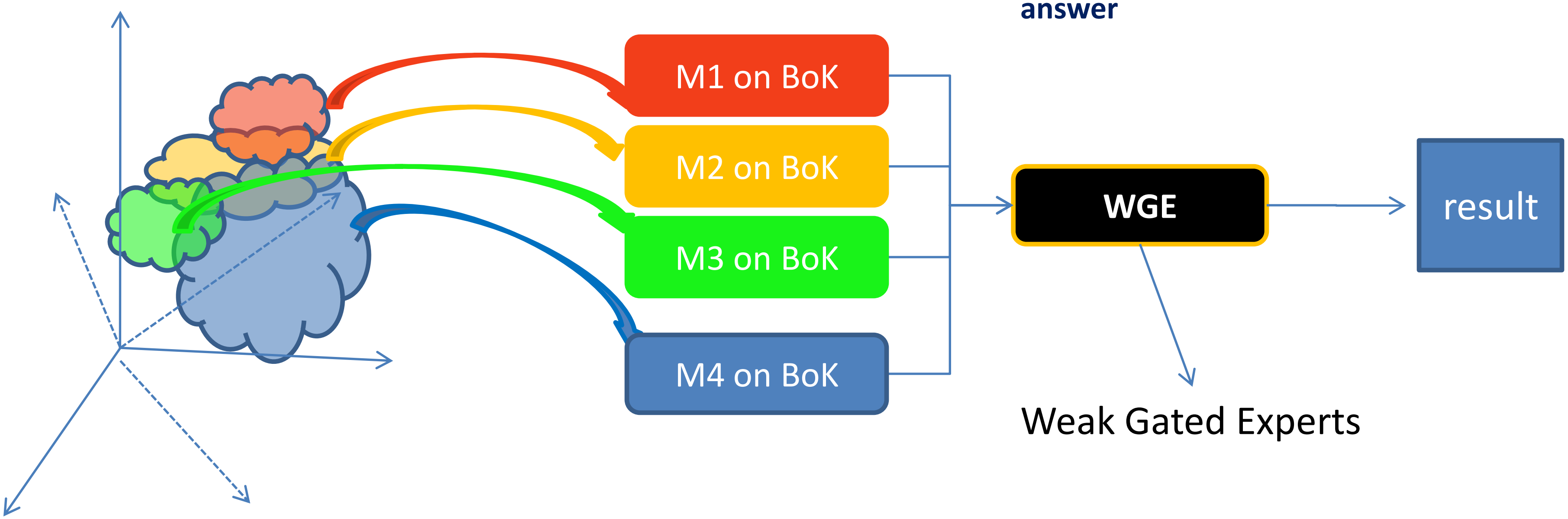
- At high z LRG dominate and interpolative methods are not capable to “generalize” rules
- An unique method optimizes its performances on the parts of the parameter space which are best covered in the BoK



Step 1:
unsupervised clustering in parameter space

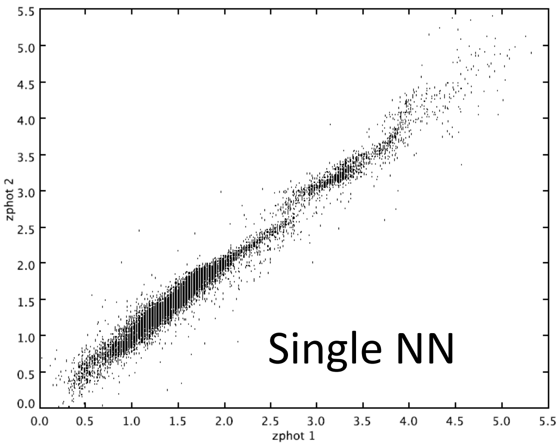
Step 2:
supervised training of different NN for each cluster

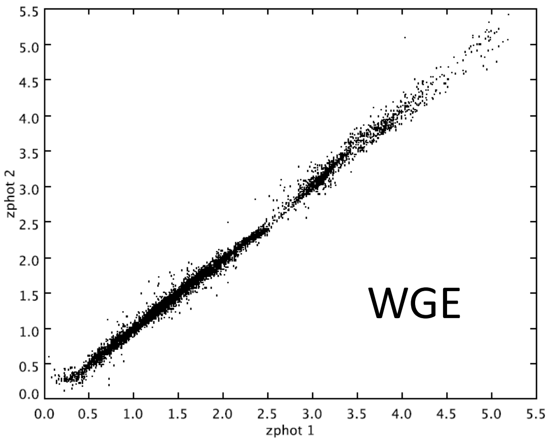
Step 3:
output of all NN go to WGE which learns the correct answer

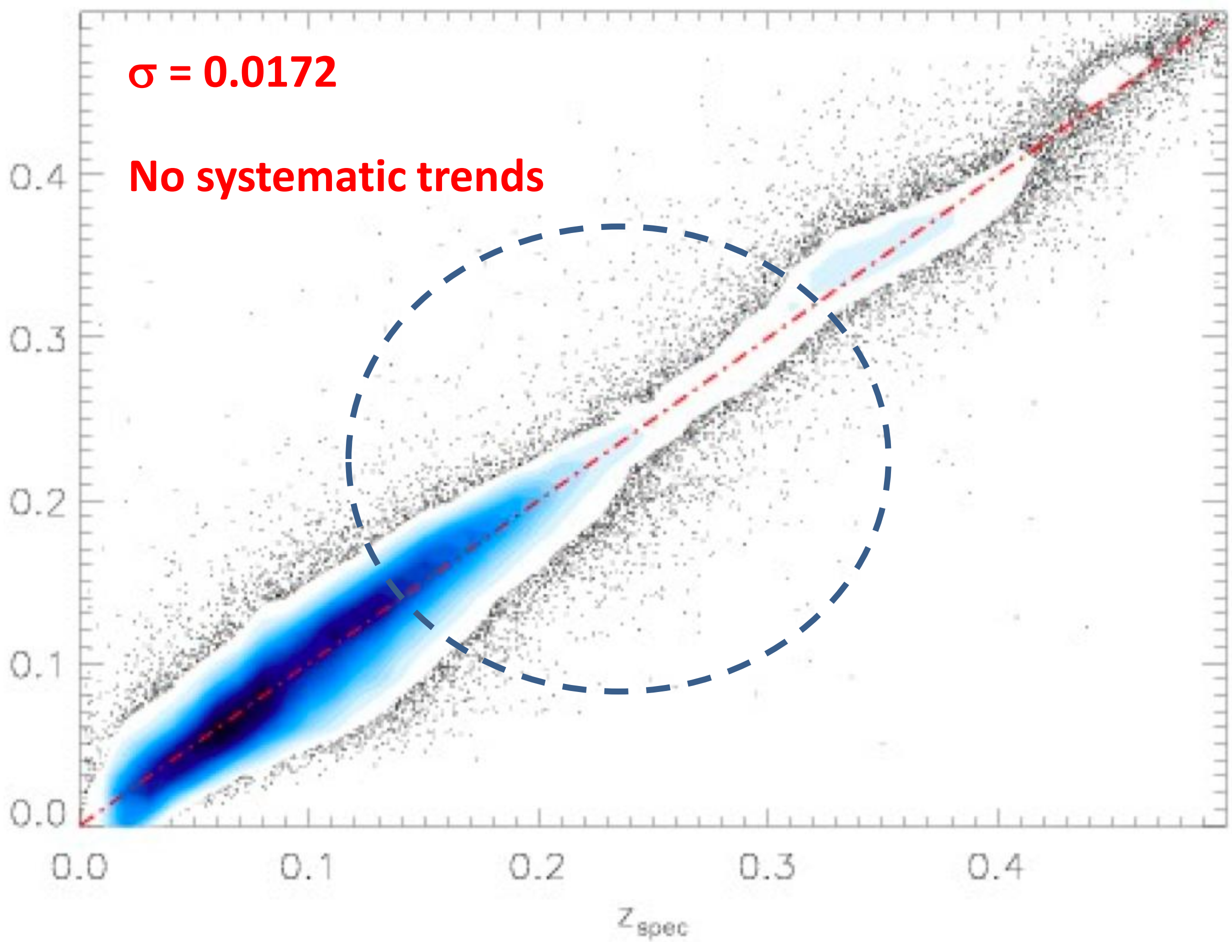


Laurino et al. 2009a,2009b

Laurino et al. 2009a,2009b

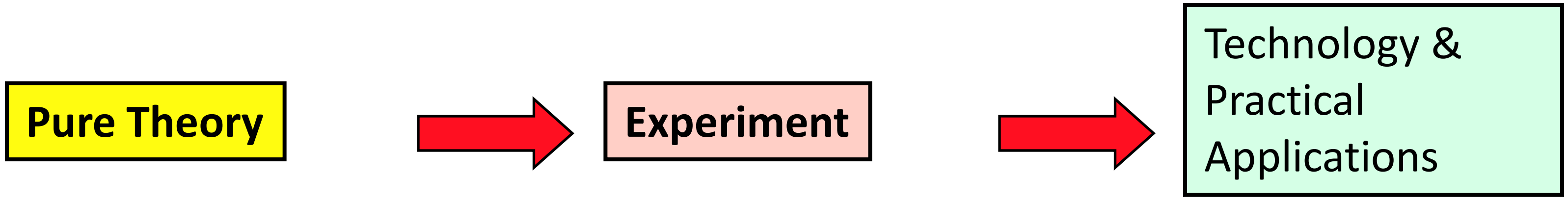




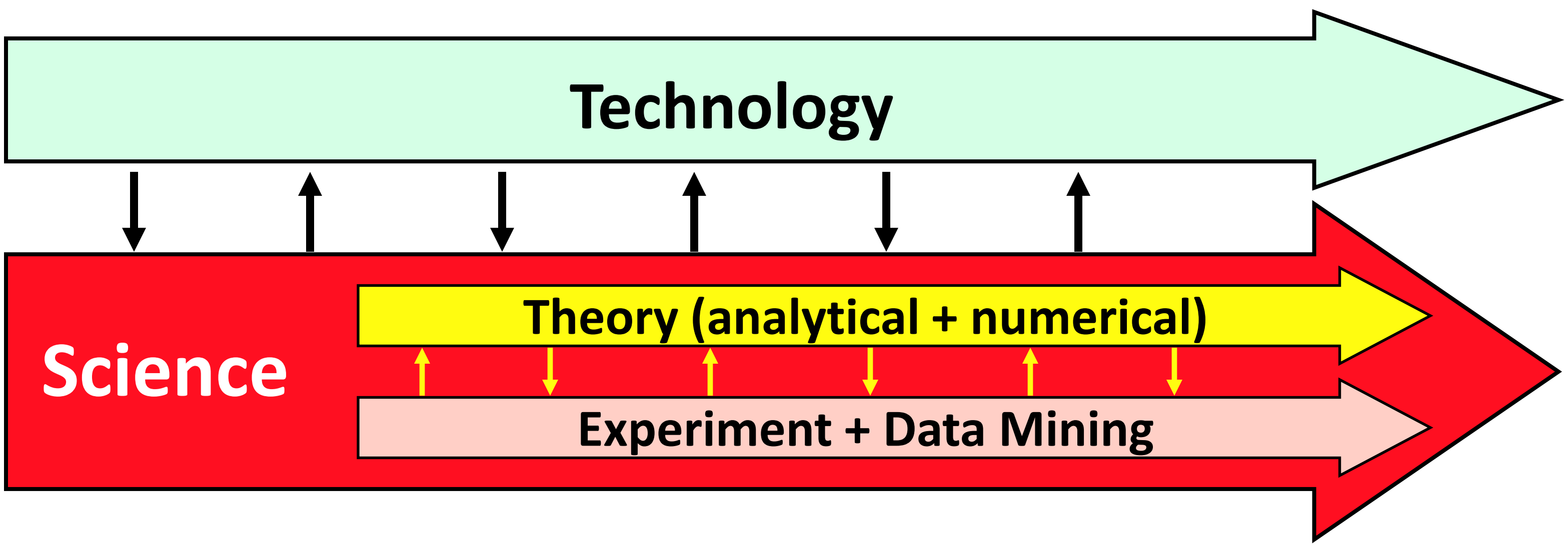


Conclusion I. I.T. is changing the methodology of science

The old traditional, “Platonistic” view:



The modern and realistic view when dealing with complex data sets:



This synergy is stronger than ever and growing

Conclusion I. I.T. is changing the methodology of science

- Standardization of data access is indispensable to ensure data exploitation and to optimize both costs and scintific return
- VObs methodologies even though fine tuned on Astrophysics are general and can be easily exported to other domains
- Data Mining is the “fourth leg of science” (besides theory, experimentation and simulations)
- Sociological issues to be solved (formation, infrastructures, and so on)
- Sinergy between different worlds is required