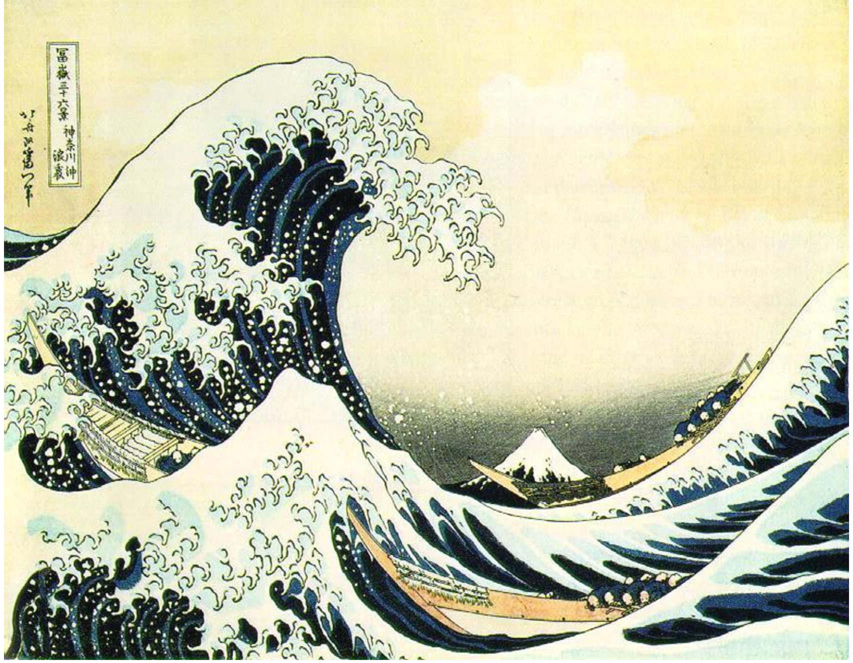
Mining Astronomical Massive Data Sets



Giuseppe Longo

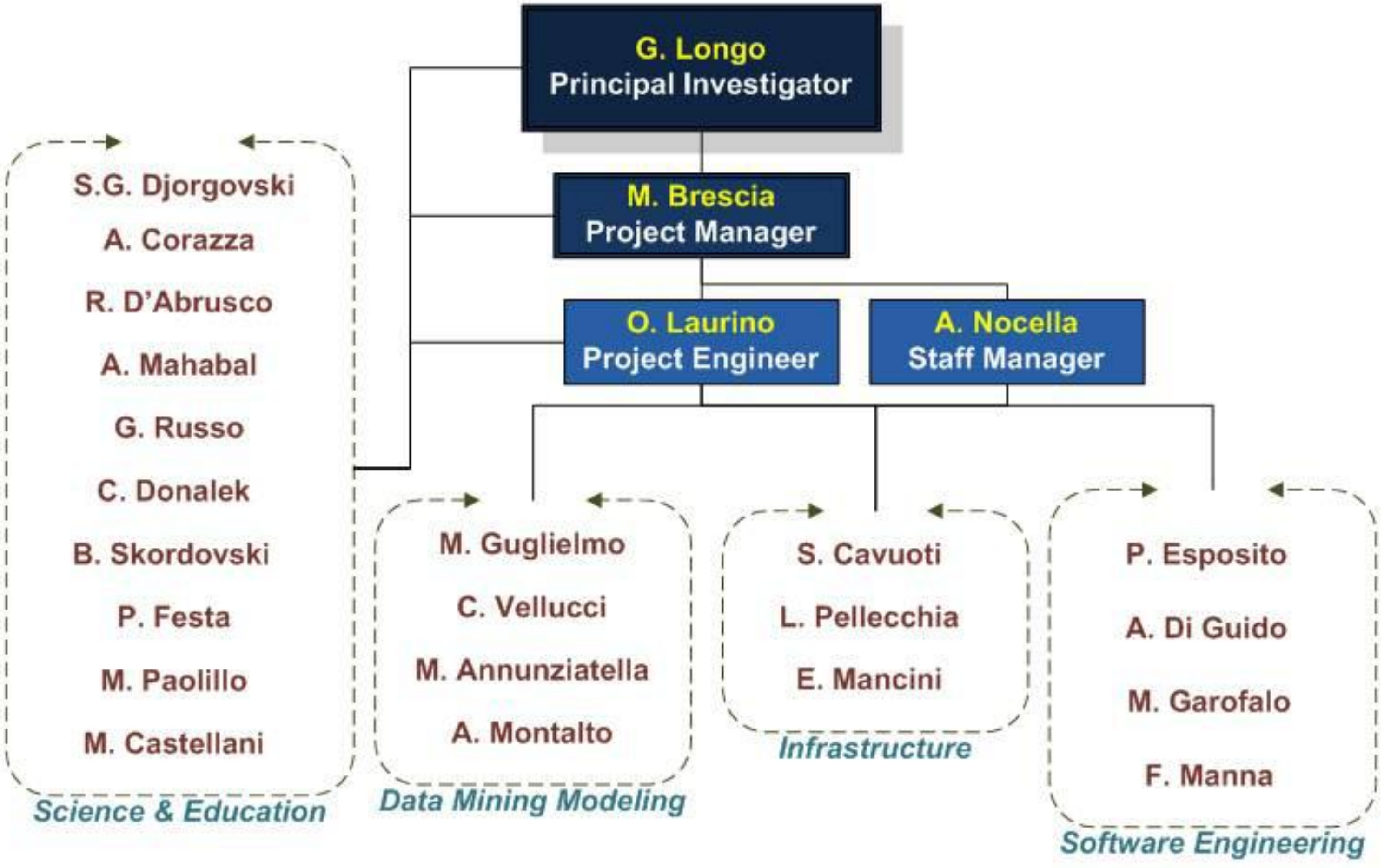
Dept.of Physics, University Federico II in Napoli – ITALY
Dept. of Astronomy, Caltech, Pasadena (USA)
INAF – Italian Institute of Astrophysics
INFN – Italian Institute of Nuclear Physics
longo@na.infn.it

Summary

- methodological introduction on the problems posed by the data tsunami & why DM and SPR are a need !!
- some classification and clustering methods and their applications to some problems in observational cosmology
- possible applications in an evolving scenario

ASI, July 2010

DAME ORGANIZATION CHART



G. Longo
Principal Investigator

M. Brescia
Project Manager

O. Laurino
Project Engineer

A. Nocella
Staff Manager

S.G. Djorgovski
A. Corazza
R. D'Abrusco
A. Mahabal
G. Russo
C. Donalek
B. Skordovski
P. Festa
M. Paolillo
M. Castellani

Science & Education

M. Guglielmo
C. Vellucci
M. Annunziatella
A. Montalto

Data Mining Modeling

S. Cavuoti
L. Pellecchia
E. Mancini

Infrastructure

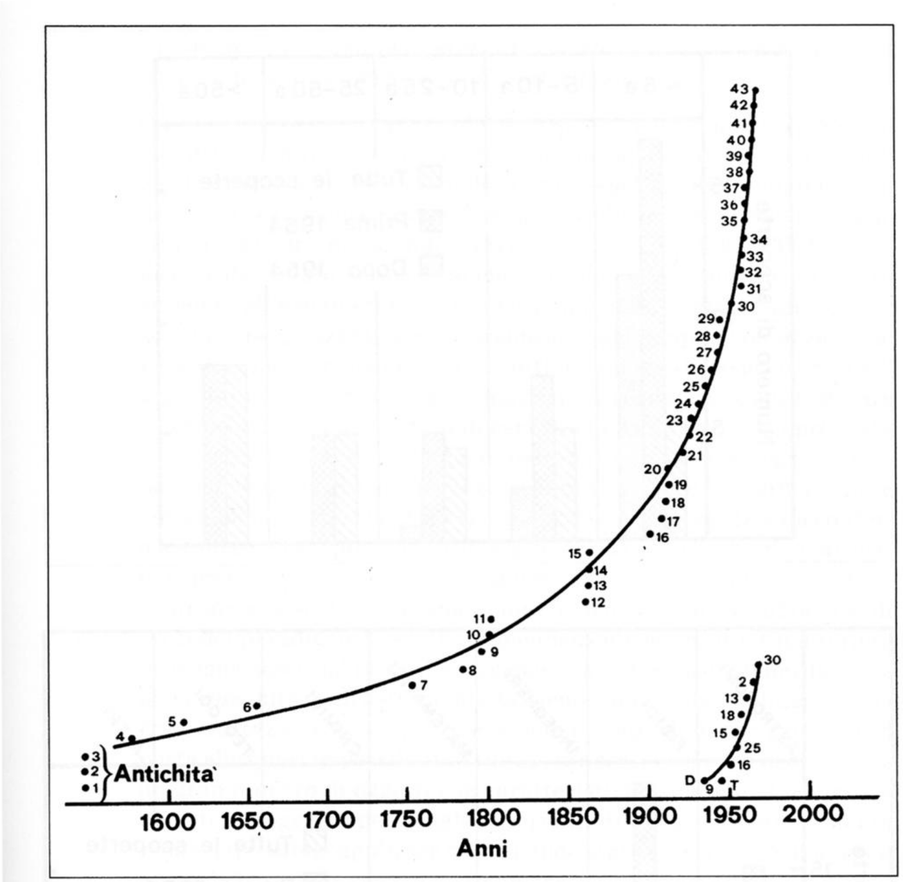
P. Esposito
A. Di Guido
M. Garofalo
F. Manna

Software Engineering

Summary

- methodological introduction on the problems posed by the data tsunami & why DM and SPR are a need !!
- some classification and clustering methods and their applications to two problems in observational cosmology (Photometric redshifts and QSO candidates identification)
- Future developments and possible applications in an evolving scenario

Part I – the scenario

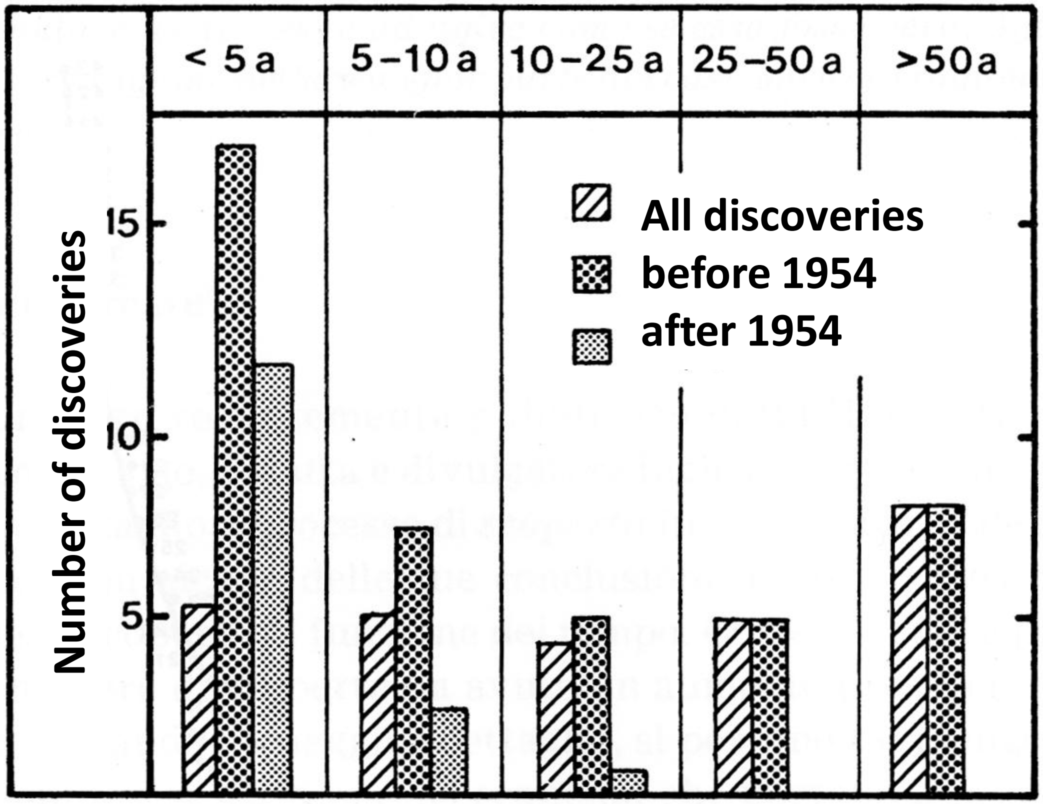


From M.Harwit, Cosmic discoveries

1. Stars
2. Planets
3. Novae
4. Comets
5. Satellites
6. Rings
7. Galactic clusters
8. Galaxy clusters
9. Interplanetary dust
10. Asteroids
11. Binary stars
12. Variable stars
13. Planetary nebulae
14. Globular clusters
15. HII regions
16. Cold ISM
17. Giant stars
18. Cosmic rays
19. Pulsating variables
20. White dwarfs
21. Galaxies
22. Expansion of universe
23. Cosmic dust
24. Supernovae/novae
25. Gas in galaxies
26. SN remnants
27. Radiogalaxies
28. Magnetic variables
29. Flare stars
30. Intergalactic magnetic fields
31. X stars
32. X background
33. Quasar
34. CMB
35. Masers
36. Infrared stars
37. X galaxies
38. Pulsar
39. Gamma background
40. IR galaxies
41. Superluminal sources
42. GRB
43. Unidentified radio sources
44. …
45. ….

The role of technology

Most discoveries take place immediately after a technological breaktrough



And now, the question is….
Where to search …
for the next discoveries?

Next breakthrough will be in data fusion and access

- We have **almost** reached the physical limit of observations (i.e. single photon counting) at **almost** all wavelenght…
- Detectors are linear
- All electromagnetic bands have been opened…

Hence technological breakthrough can be in:

- **Accuracy** (lower flux limits, increased statistics)
- **Sampling** (angular resolution, time domain)
- **Complexity** (data fusion, data mining, modeling, etc.)

New insights will depend mainly on:

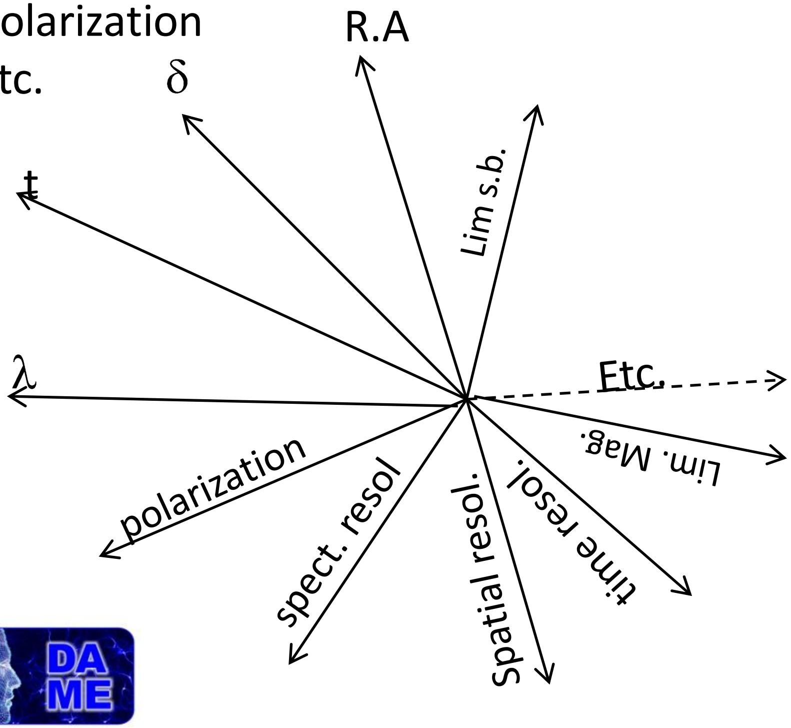
- Capability to ACCESS AND MERGE heterogeneous information (multi-epoch, multi-λ, etc.)
- Capability to recognize patterns or trends in the parameter space (i.e. physical laws) which are not limited to the human 3-D visualization
- Capability to extract patterns from very large multiwavelenght, multiepoch, multi-technique parameter spaces

The parameter space

Vesuvius, now

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.



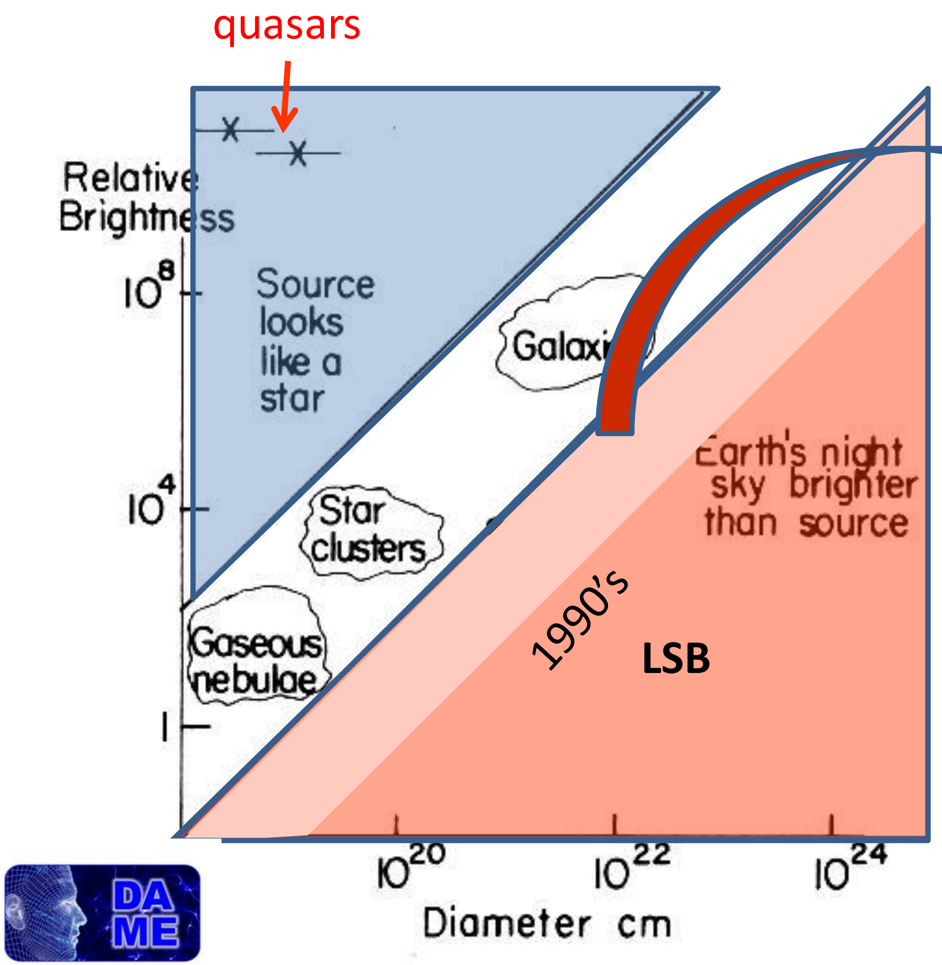
$$p \in \Re^N \qquad N >> 100$$

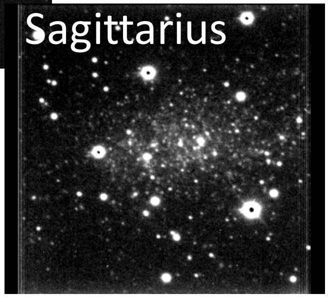
The parameter space concept is crucial to:

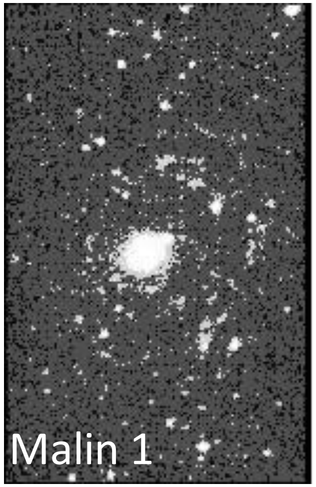
1. Guide the quest for new discoveries (observations can be guided to explore poorly known regions), …
2. Find new physical laws (patterns)
3. Etc,

Every time you improve the coverage of the PS….

Every time a new technology enlarges the parameter space or allows a better sampling of it, new discoveries are bound to take place

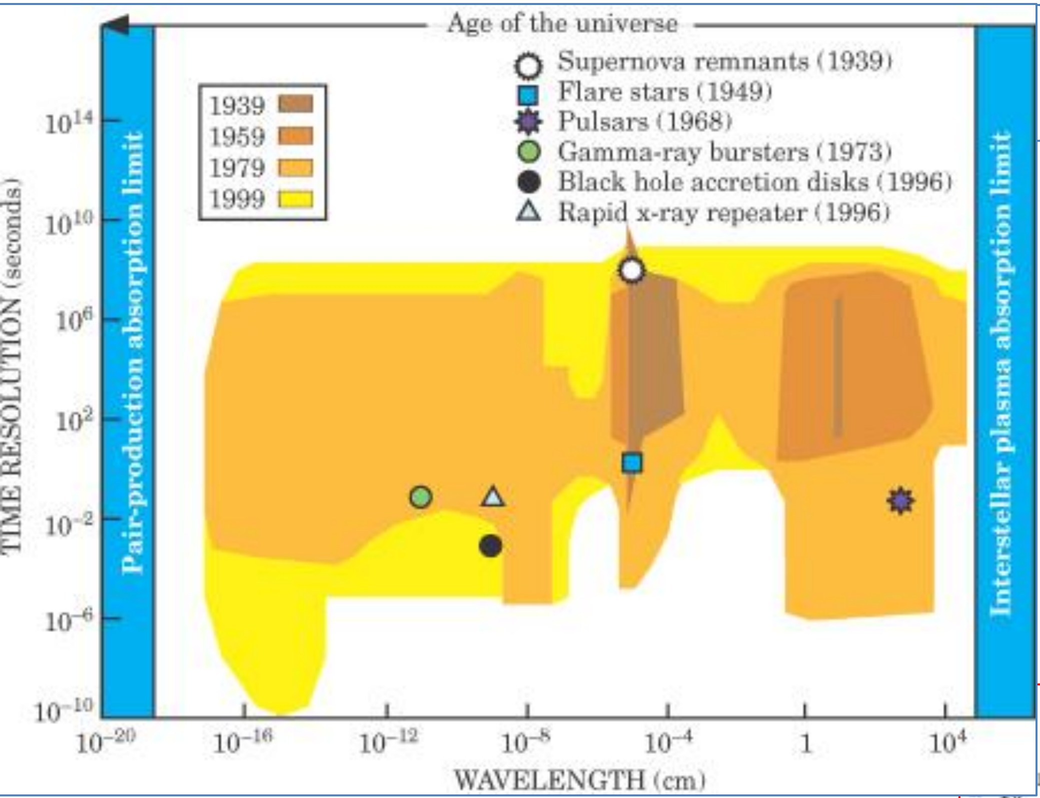




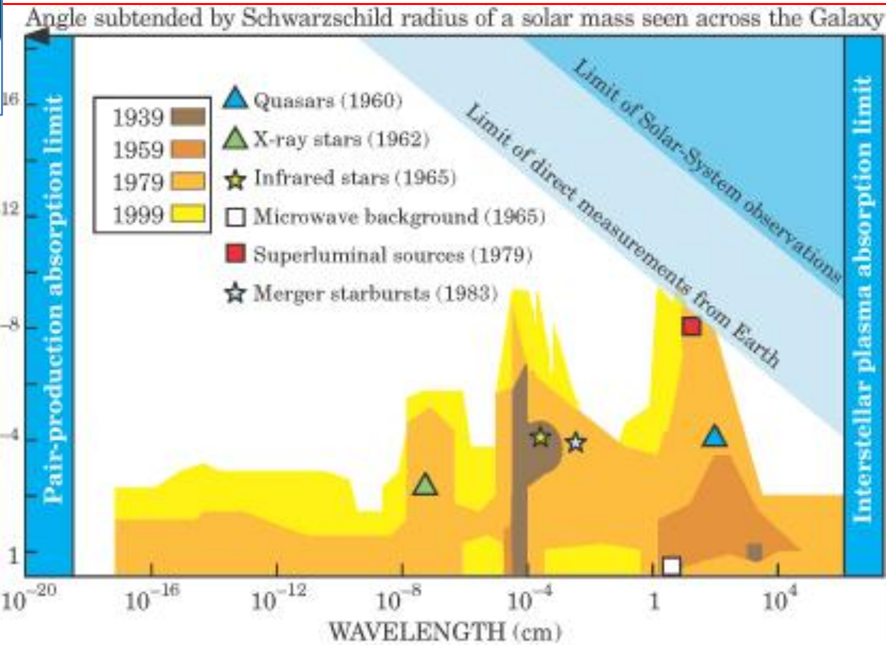


Discovery of
Low surface brightness
Universe

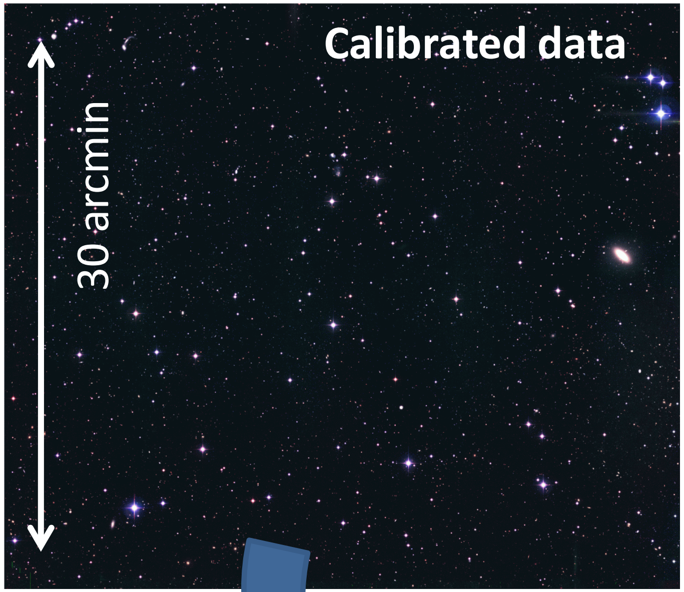
Improving coverage of the Parameter space - II



Projection of parameter space along (time resolution & wavelength)

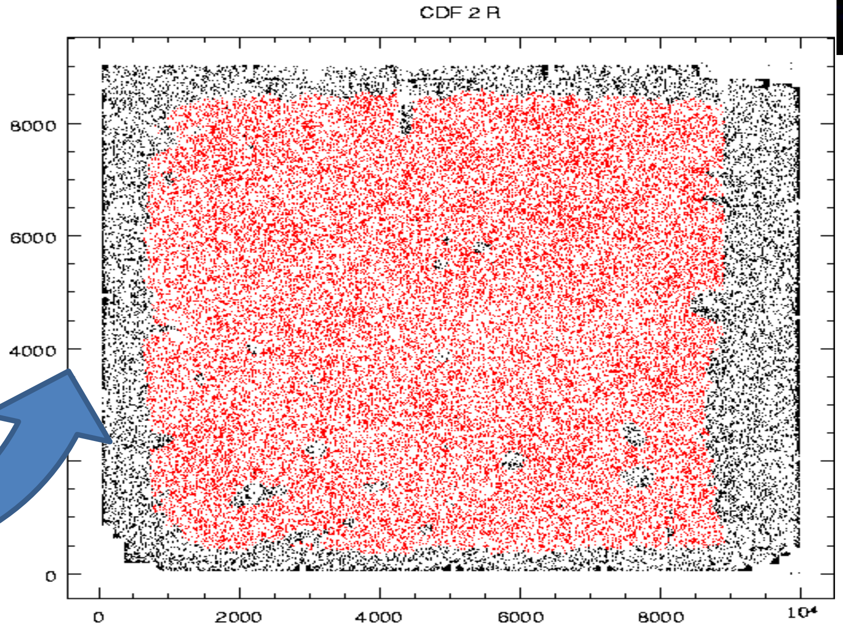


Projection of parameter space along (angular resolution & wavelength)



1/160.000 of the sky, moderately deep (25.0 in r)

55.000 detected sources
(0.75 mag above m lim)



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

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

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

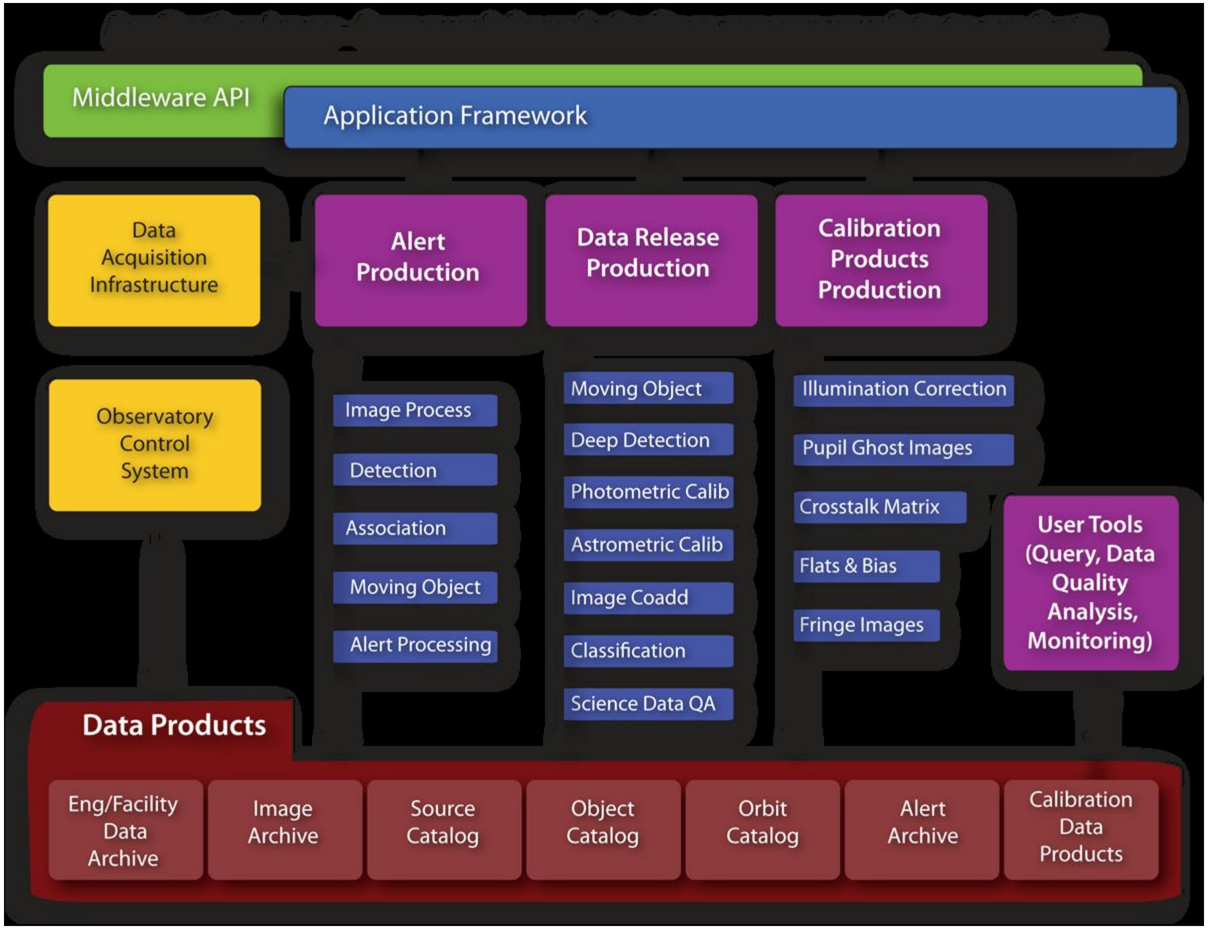
....

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

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

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

Computational (HW+SW) challenges: LSST



Per Night

- 15 TB of images
- 1 TB catalogs
- 60 sec alerts for 10^5-10^6 Objects

Per Year

- 6.5 PB per year of images and catalogs

Lifetime

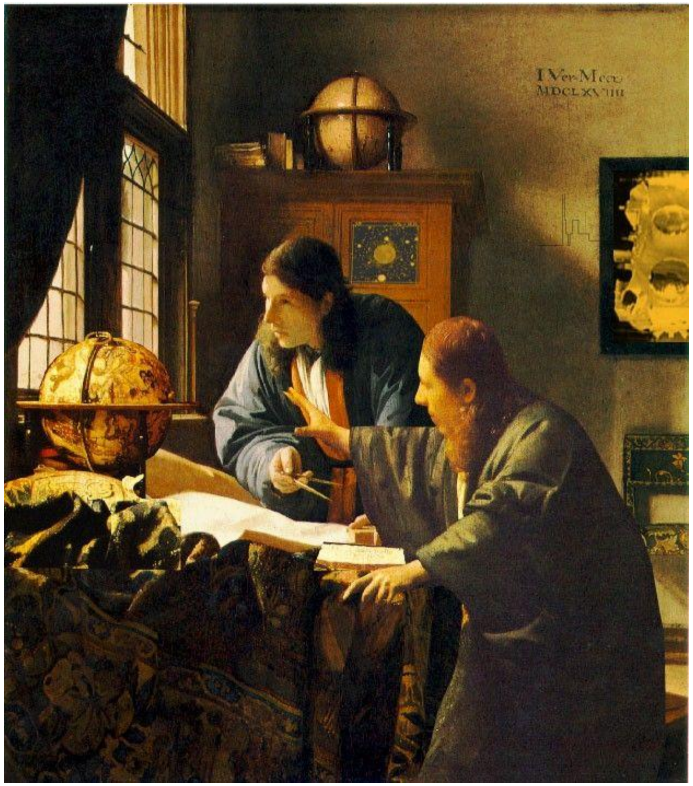
- 10 B Stars and 10 B Galaxies
- 60-70 PB of images

Courtesy of Krughoff– Astroinformatics 2010

Part II
DATA MINING IN ASTRONOMY

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



The astroinformatics domain

Data Gathering (e.g., new generation instruments …)

→ Data Farming:

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

→ Data Mining (or Knowledge Discovery in Databases):

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

→ Data visualization and understanding

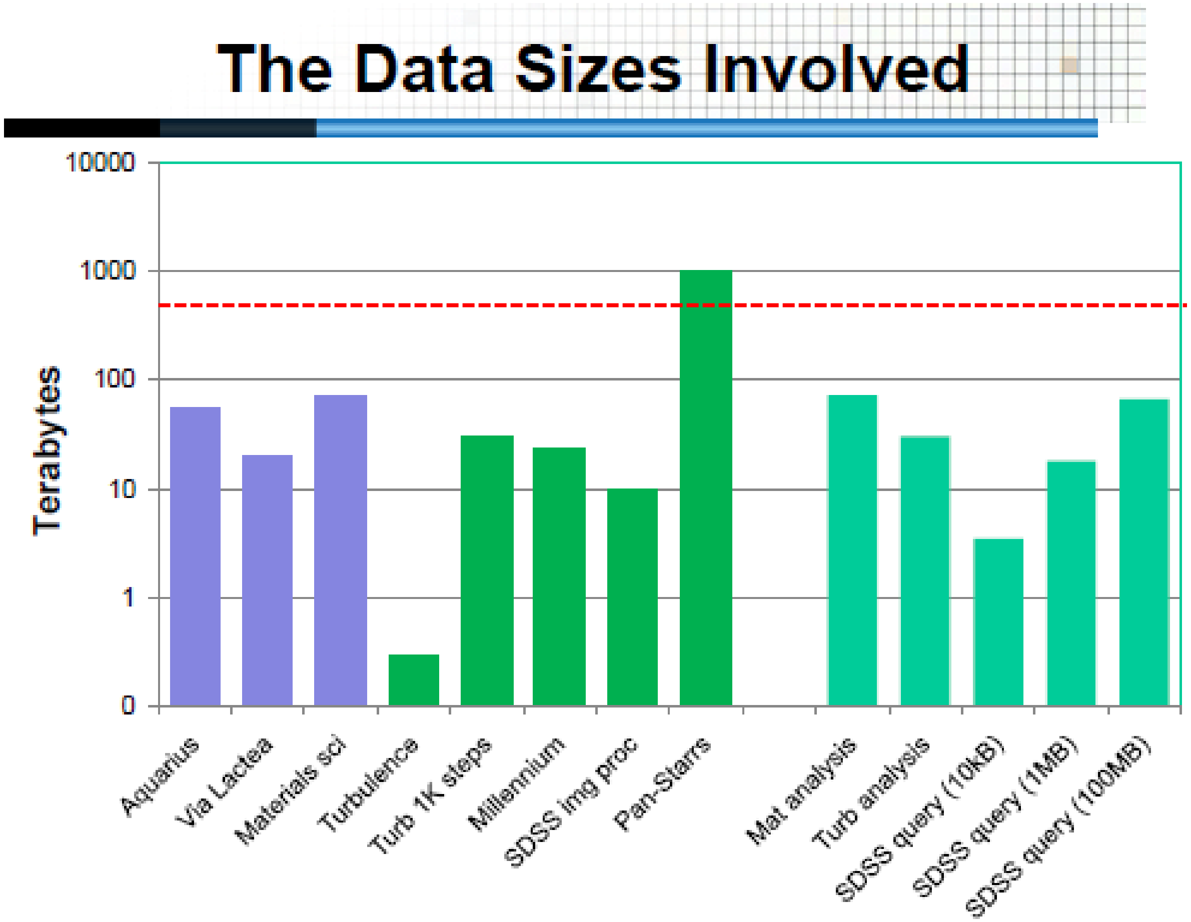
- Computer aided understanding
- KDD
- Etc.

→ New Knowledge

Data storage , Pbytes
Data access >10^3 access

Scalability: Petaflops, Exaflops
Computing power (multicore)
Algorithm: parallelism
Visualization: N-dimensional

Data storage (problem to be solved)



Memory of Today's Biggest System

Elaboration needs to take place where the data are

From Alex Szalay, " Amdahl's Law and Extreme Data-Intensive Computing," 2010 Salishan Conf. on High Speed Computing

Expected growth rates can exceed 1 PB/year for Raw Data - LSST may reach 100 PB!

Donald Rumsfeld's explanation of data mining (but he did not know…)

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.

Donald Rumsfeld's about Iraqi war

Courtesy S.G. Djorgovski

Scalability of most relevant astronomical algorithms

- **Querying: spherical range-search O(N), orthogonal range-search O(N),** spatial join **O(N2), nearest-neighbor O(N), all-nearest-neighbors O(N^2)**
- **Density estimation: mixture of Gaussians, kernel density estimation O(N^2),** kernel conditional density estimation **O(N^3)**
- **Regression: linear regression, kernel regression O(N^2), Gaussian process** regression **O(N^3)**
- **Classification: decision tree, nearest-neighbor classifier O(N^2),** nonparametric Bayes classifier **O(N^2), support vector machine O(N^3)**
- **Dimension reduction: principal component analysis, non-negative matrix** factorization, kernel PCA **O(N^3), maximum variance unfolding O(N^3)**
- **Outlier detection: by density estimation or dimension reduction**
- **Clustering: by density estimation or dimension reduction, k-means, meanshift** segmentation **O(N^2), hierarchical (FoF) clustering O(N^3)**
- **Time series analysis: Kalman filter, hidden Markov model, trajectory** tracking **O(N^n)**
- **Feature selection and causality: LASSO, L1 SVM, Gaussian graphical** models, discrete graphical models
- **2-sample testing and testing and matching: bipartite matching O(N^3), n-point** correlation **O(N^n)**

Courtesy of A. Gray – Astroinformatics 2010

Brute force is not a solution

Exascale (needed for crosscorrelation on archives like LSST: Exascale = *1,000X capability of Today*

- **Exascale != Exaflops but**

Exascale at the data center size => ***Exaflops***
Exascale at the "rack" size => ***Petaflops for*** departmental systems
Exascale embedded => ***Teraflops in a cube***

It took 14+ years to get from
1st Petaflops workshop: 1994, thru NSF studies, HTMT, HPCS … to give us to Petaflops *in 2009*

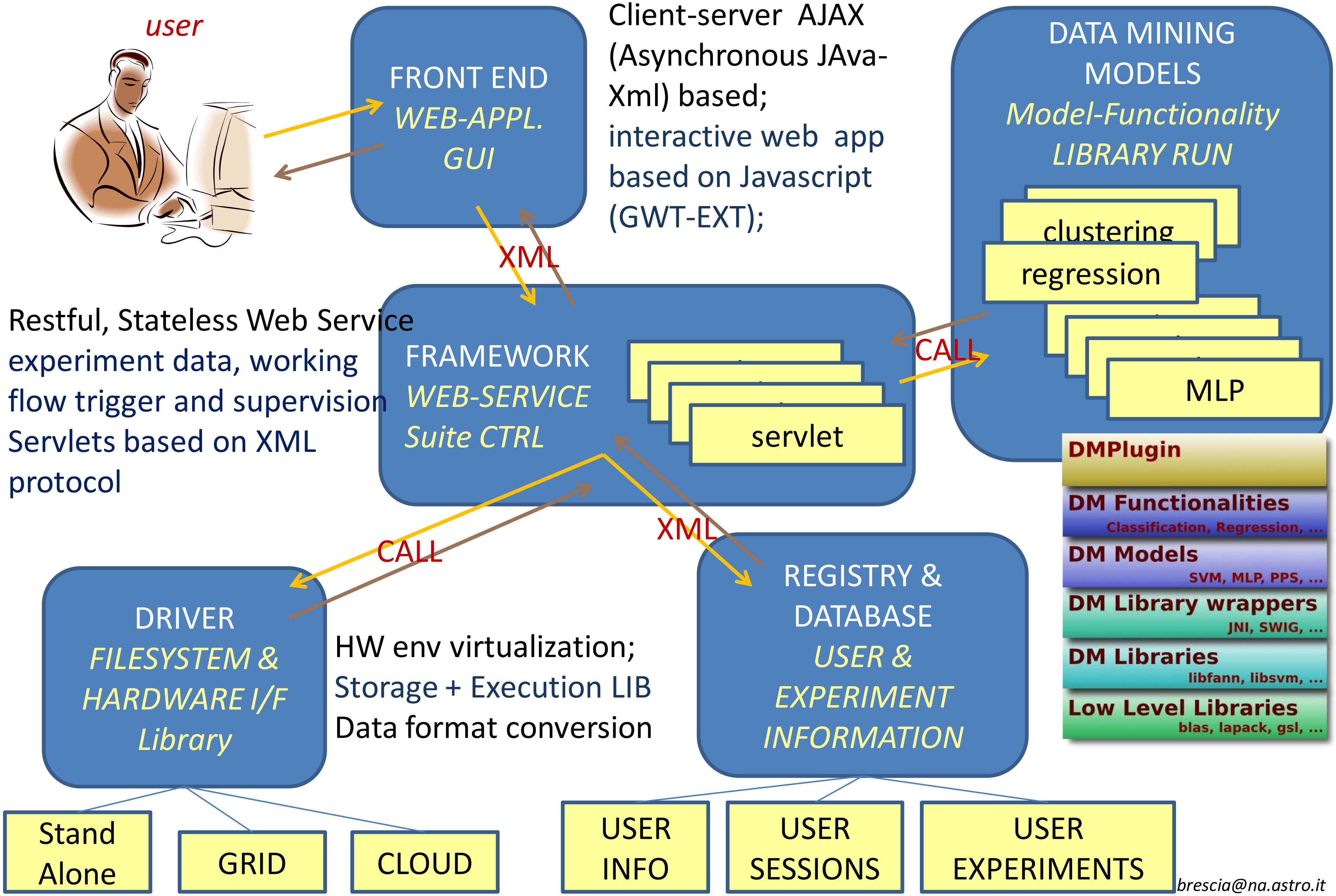
We should be OVERJOYED if all You need is:

- ***JUST a Million cores***
- ***ONLY 1 Nuclear Power Plant***
- ***MINIMAL programming support***

Better algorithms are needed

Adapted from Kogge– Astroinformatics 2010

The DAME architecture



user

FRONT END
WEB-APPL. GUI

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

DATA MINING MODELS
Model-Functionality LIBRARY RUN

clustering

regression

MLP

XML

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

FRAMEWORK
WEB-SERVICE Suite CTRL

servlet

CALL

DMPlugin

DM Functionalities
Classification, Regression, …

DM Models
SVM, MLP, PPS, …

DM Library wrappers
JNI, SWIG, …

DM Libraries
libfann, libsvm, …

Low Level Libraries
blas, lapack, gsl, …

CALL

XML

DRIVER
FILESYSTEM & HARDWARE I/F Library

HW env virtualization;
Storage + Execution LIB
Data format conversion

REGISTRY & DATABASE
USER & EXPERIMENT INFORMATION

Stand Alone

GRID

CLOUD

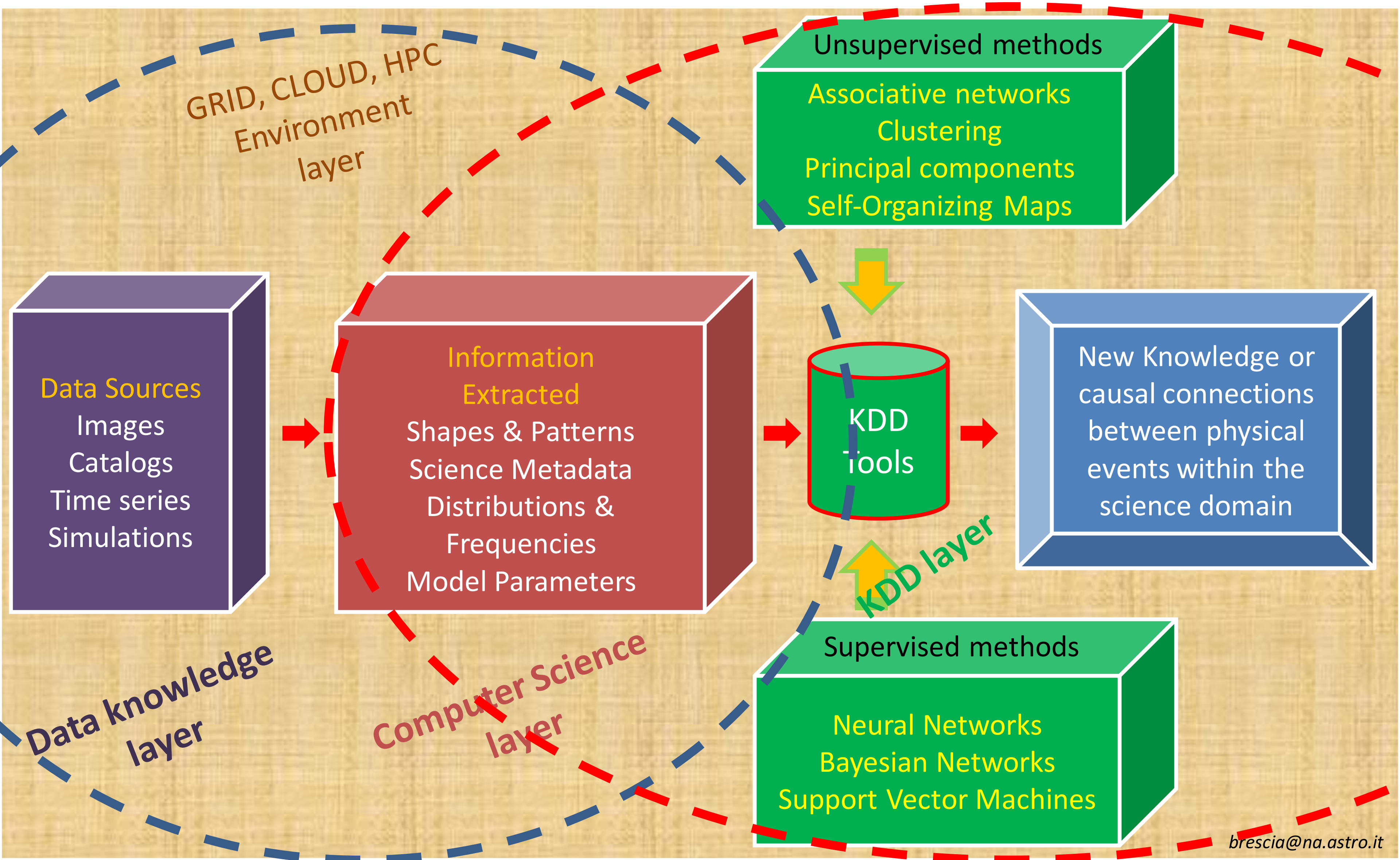
USER INFO

USER SESSIONS

USER EXPERIMENTS

Vobs standards and infrastructure

Data mining level



GRID, CLOUD, HPC Environment layer

Data Sources
Images
Catalogs
Time series
Simulations

Information Extracted
Shapes & Patterns
Science Metadata
Distributions & Frequencies
Model Parameters

Unsupervised methods
Associative networks
Clustering
Principal components
Self-Organizing Maps

KDD Tools

Supervised methods
Neural Networks
Bayesian Networks
Support Vector Machines

New Knowledge or causal connections between physical events within the science domain

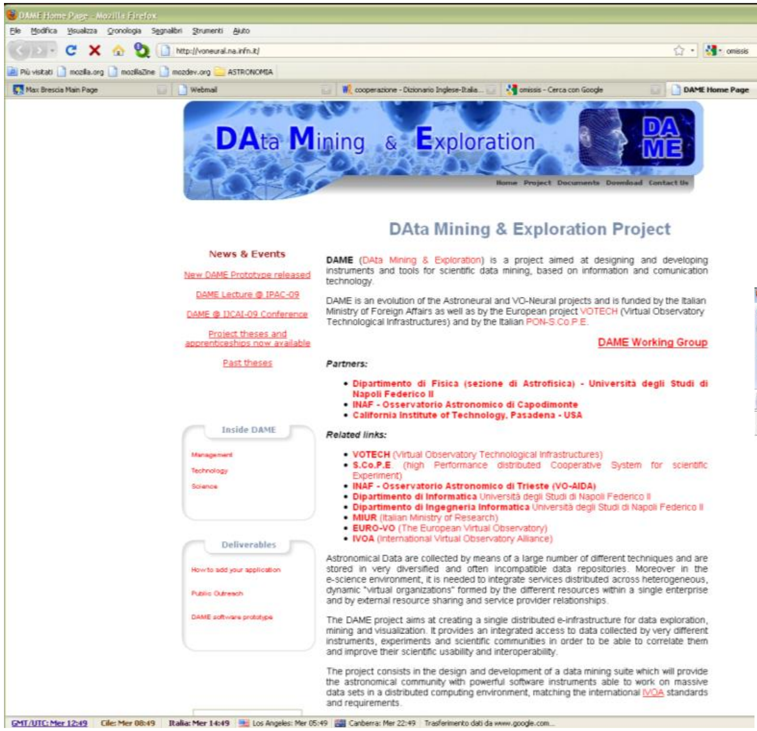
KDD layer

Data knowledge layer

Computer Science layer

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.



DAta Mining & Exploration

DAta Mining & Exploration Project

News & Events

- New DAME Prototype released
- DAME Lecture @ IPAC-09
- DAME @ IJCAI-09 Conference
- Project theses and apprenticeships now available
- Past theses

Inside DAME

- Management
- Technology
- Science

Deliverables

- How to add your application
- Public Outreach
- DAME software prototype

DAME (DAta Mining & Exploration) is a project aimed at designing and developing instruments and tools for scientific data mining, based on information and comunication technology.

DAME is an evolution of the Astroneural and VO-Neural projects and is funded by the Italian Ministry of Foreign Affairs as well as by the European project VOTECH (Virtual Observatory Technological Infrastructures) and by the Italian PON-S.Co.P.E.

DAME Working Group

Partners:

- Dipartimento di Fisica (sezione di Astrofisica) - Università degli Studi di Napoli Federico II
- INAF - Osservatorio Astronomico di Capodimonte
- California Institute of Technology. Pasadena - USA

Related links:

- **VOTECH** (Virtual Observatory Technological Infrastructures)
- **S.Co.P.E.** (high Performance distributed Cooperative System for scientific Experiment)
- **INAF - Osservatorio Astronomico di Trieste (VO-AIDA)**
- **Dipartimento di Informatica** Università degli Studi di Napoli Federico II
- **Dipartimento di Ingegneria Informatica** Università degli Studi di Napoli Federico II
- **MIUR** (Italian Ministry of Research)
- **EURO-VO** (The European Virtual Observatory)
- **IVOA** (International Virtual Observatory Alliance)

Astronomical Data are collected by means of a large number of different techniques and are stored in very diversified and often incompatible data repositories. Moreover in the e-science environment, it is needed to integrate services distributed across heterogeneous, dynamic "virtual organizations" formed by the different resources within a single enterprise and by external resource sharing and service provider relationships.

The DAME project aims at creating a single distributed e-infrastructure for data exploration, mining and visualization. It provides an integrated access to data collected by very different instruments, experiments and scientific communities in order to be able to correlate them and improve their scientific usability and interoperability.

The project consists in the design and development of a data mining suite which will provide the astronomical community with powerful software instruments able to work on massive data sets in a distributed computing environment, matching the international IVOA standards and requirements.

DAME Home Page - Mozilla Firefox
http://voneural.na.infn.it/
GMT/UTC: Mer 12:49
Cile: Mer 08:49
Italia: Mer 14:49
Los Angeles: Mer 05:49
Canberra: Mer 22:49
Trasferimento dati da www.google.com...

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

Technical and management info

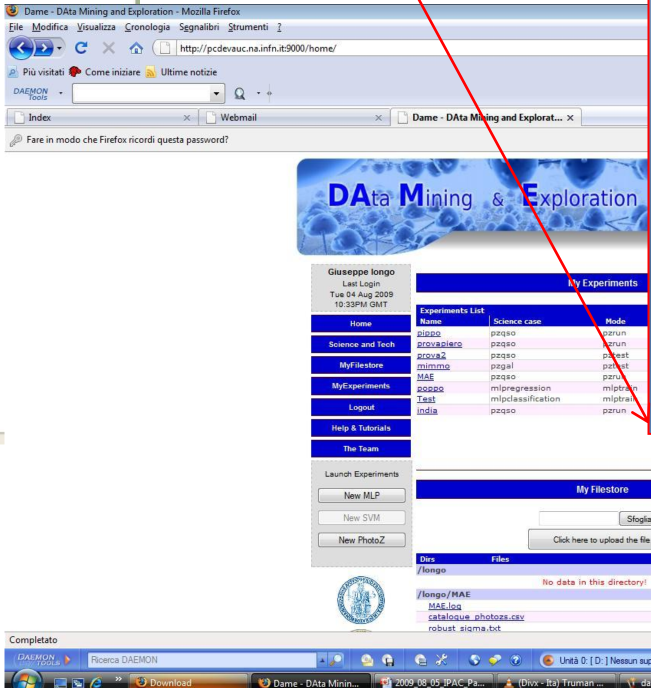
Documents

Science cases

Newsletter

01
DAta Mining & Exploration
DA ME
In this edition
Mining Astrophysical datasets with DAME
First Technological Products
Scientific Use Case Tutorial
A New trend in Astrophysics

Web Solutions for Data Mining in Astrophysics



Dame - DAta Mining and Exploration - Mozilla Firefox
File Modifica Visualizza Cronologia Segnalibri Strumenti ?
http://pcdevauc.na.infn.it:9000/home/
Più visitati
Come iniziare
Ultime notizie
Index
Webmail
Dame - DAta Mining and Explorat...
Fare in modo che Firefox ricordi questa password?
DAta Mining & Exploration
Completato
Ricerca DAEMON
Unità 0: [D:] Nessun supporto
Download
Dame - DAta Minin...
2009_08_05_IPAC_Pa...
(Divx - Ita) Truman ...
dame - Paint
0:34

Giuseppe longo
Last Login
Tue 04 Aug 2009
10:33PM GMT

- Home
- Science and Tech
- MyFilestore
- MyExperiments
- Logout
- Help & Tutorials
- The Team

Launch Experiments

- New MLP
- New SVM
- New PhotoZ

My Experiments

Experiments List

Name	Science case	Mode
pippo	pzqso	pzrun
provapiero	pzqso	pzrun
prova2	pzqso	pztest
mimmo	pzgal	pztest
MAE	pzqso	pzrun
poppo	mlpregression	mlptrain
Test	mlpclassification	mlptrain
india	pzqso	pzrun

My Filestore

Sfoglia...

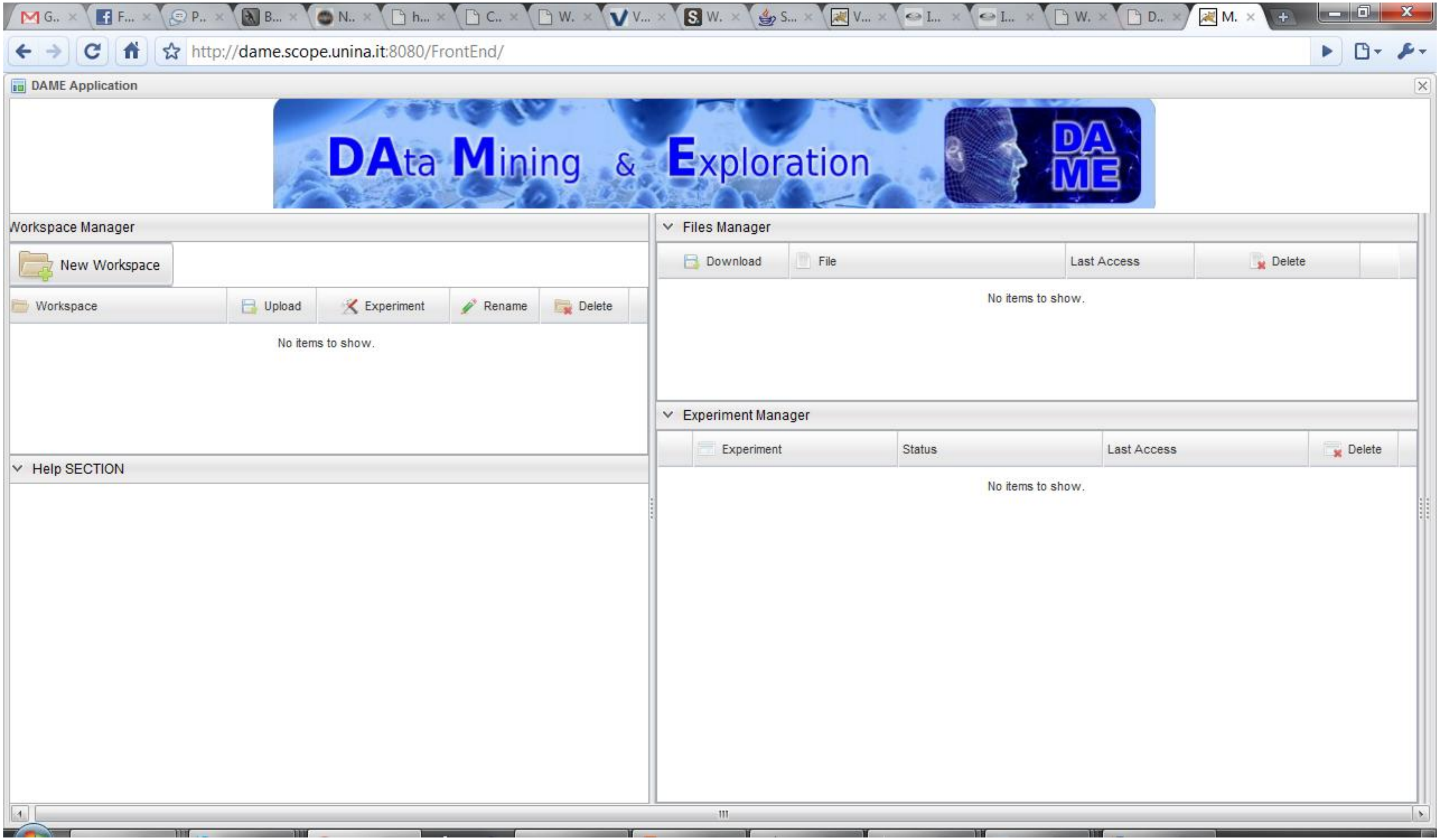
Click here to upload the file

Dirs	Files	Actions
/longo		
	No data in this directory!	
/longo/MAE		Delete Download
	MAE.log	Delete
	catalogue_photozs.csv	Delete
	robust_sigma.txt	Delete

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

Web application PROTOTYPE

DAME front-end



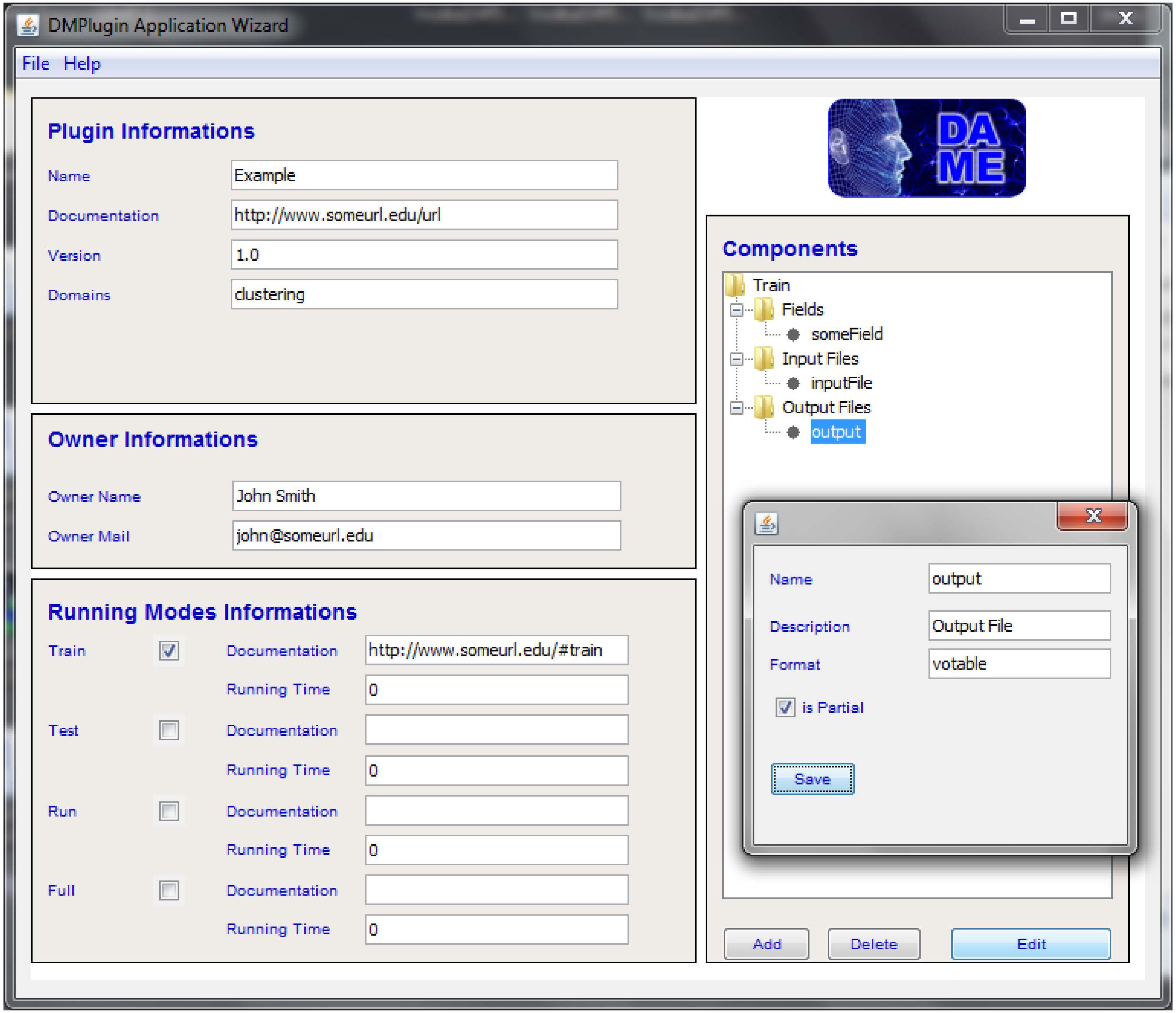
No items to show.

No items to show.

No items to show.

DAME plugin wizard

DAta Mining & Exploration



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	☑	Documentation	http://www.someurl.edu/#train
		Running Time	0
Test	□	Documentation	
		Running Time	0
Run	□	Documentation	
		Running Time	0
Full	□	Documentation	
		Running Time	0

Components

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

Name: output

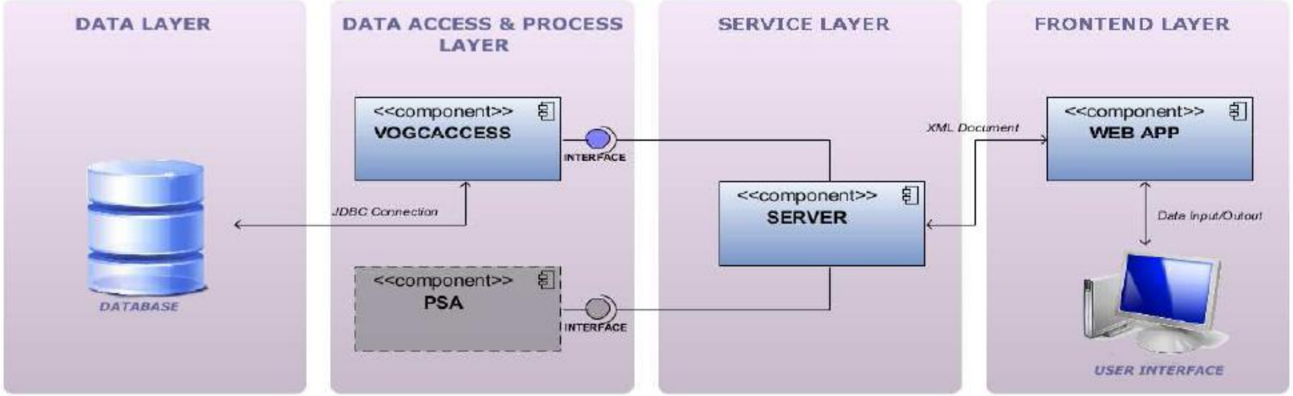
Description: Output File

Format: votable

☑ is Partial

Other DAME based WEB applications

VO-GClusters – Web application for globular clusters (in coll. with M. Castellano, INAF-OAR)



VOGClusters is a sub-framework within DAME for the exploration and mining of VObs data archives for anything related to Globular clusters

Functionalities

- Cross-correlation of complex and bibliographic data
- Interoperability of distributed archives

Part III

DAME APPLICATIONS TO ASTRONOMY

Supervised methods

They learn how to partition the parameter space by means of a training phase based on examples.

Neural Networks such as the Multi Layer Perceptron (MLP), Support Vector Machines (SVM), etc.

Pro’s & Con’s

- They are good for interpolation of data, very bad for extrapolations
- They need extensive bases of knowledge (i.e. uniformously sampling the parameter space) which are difficult to obtain;
- Errors are easy to evaluate
- Relatively easy to use

- **They reproduce all biases and preconceived ideas present in the BoK**

Unsupervised (clustering) methods

They cluster the data relying on their statistical properties only
Understanding takes place through labeling (very limited BoK).

Generative Topographic Mapping (GTM), Self Organizing Maps (SOM), Probabilistic Principal Surfaces (PPS), Support Vector Machines (SVM), etc.

Pro’s & Con’s

- In theory they need little or none knowledge a-priori
- Do not reproduce biases present in the BoK

- Evaluation of errors more complex (through complex statistics)
- They are computationally intensive
- They are not user friendly (… more an art than a science; i.e. lot of experience required)

The Astrophysical Journal, 663:752–764, 2007 July 10

© 2007. The American Astronomical Society. All rights reserved. Printed in U.S.A.

MINING THE SDSS ARCHIVE. I. PHOTOMETRIC REDSHIFTS IN THE NEARBY UNIVERSE

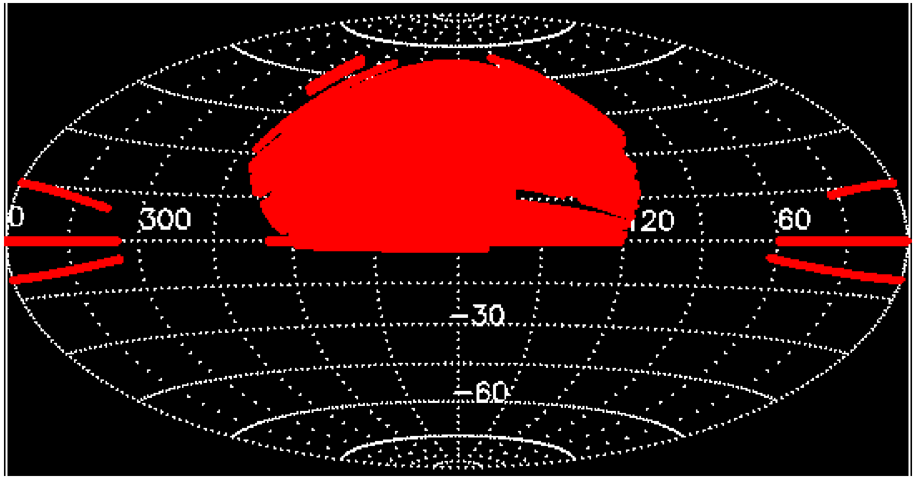
Raffaele D'Abrusco,[1,2] Antonino Staiano,[3] Giuseppe Longo,[1,4,5] Massimo Brescia,[5,4] Maurizio Paolillo,[1,4]
Elisabetta De Filippis,[5,1] and Roberto Tagliaferri[6,4]

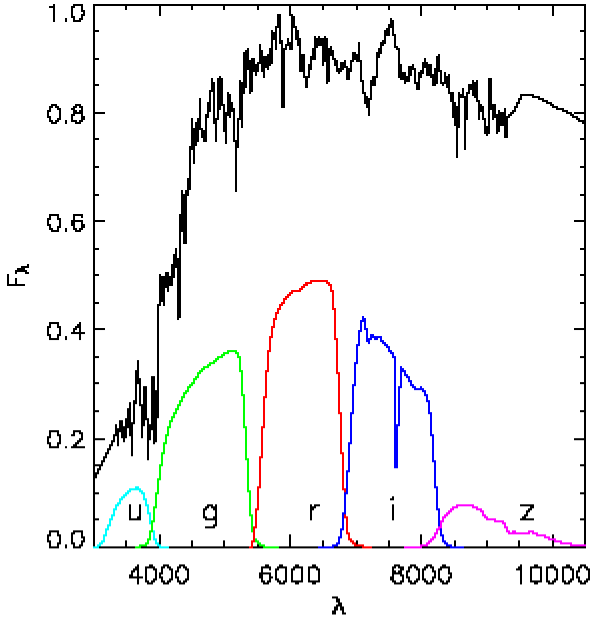
Received 2006 October 11; accepted 2007 March 2

ABSTRACT

We present a supervised neural network approach to the determination of photometric redshifts. The method was fine-tuned to match the characteristics of the Sloan Digital Sky Survey, and as base of "a priori" knowledge, it exploits the rich wealth of spectroscopic redshifts provided by this survey. In order to train, validate, and test the networks, we used two galaxy samples drawn from the SDSS spectroscopic data set, namely, the general galaxy sample (GG) and the luminous red galaxy subsample (LRG). The method consists of a two-step approach. In the first step objects are classified as nearby ($z < 0.25$) and distant ($0.25 < z < 0.50$), with an accuracy estimated as 97.52%. In the second step, two different networks are separately trained on objects belonging to the two redshift ranges. Using a standard multilayer perceptron operated in a Bayesian framework, the optimal architectures were found to require one hidden layer of 24 (24) and 24 (25) neurons for the GG (LRG) sample. The final results on the GG data set give a robust $\sigma_z \simeq 0.0208$ over the redshift range $[0.01, 0.48]$ and $\sigma_z \simeq 0.0197$ and $\simeq 0.0238$ for the nearby and distant samples, respectively. For the LRG subsample we find instead a robust $\sigma_z \simeq 0.0164$ over the whole range, and $\sigma_z \simeq 0.0160$ and $\simeq 0.0183$ for the nearby and distant samples, respectively. After training, the networks have been applied to all objects in the SDSS table GALAXY matching the same selection criteria adopted to build the base of knowledge, and photometric redshifts for circa 30 million galaxies having $z < 0.5$ were derived. A catalog containing redshifts for the LRG subsample was also produced.

The Sloan Digital Sky Survey (SDSS) data set & BoK





8000 sq degrees
>210 million galaxies
data are public

Benchmark for almost everything in observational cosmology

Extensive but biased spectroscopic BoK:
700.000 galaxy spectra

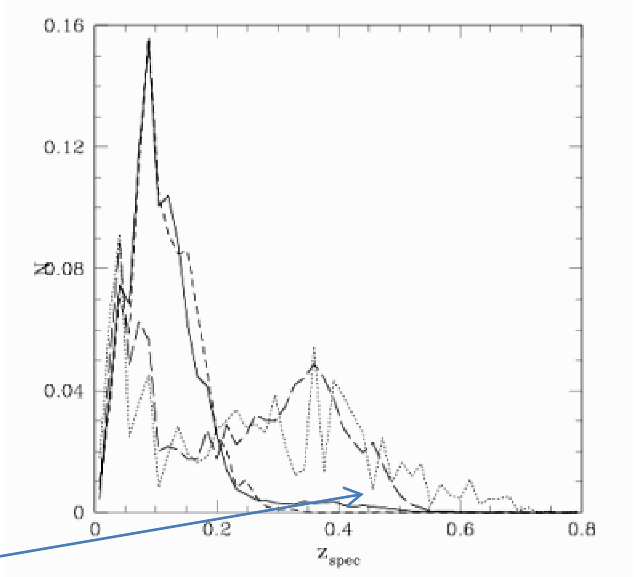
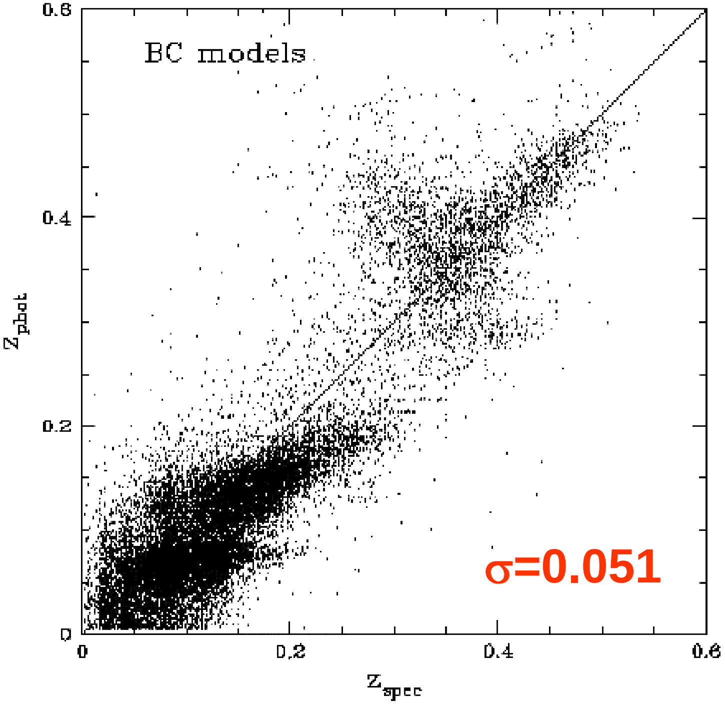


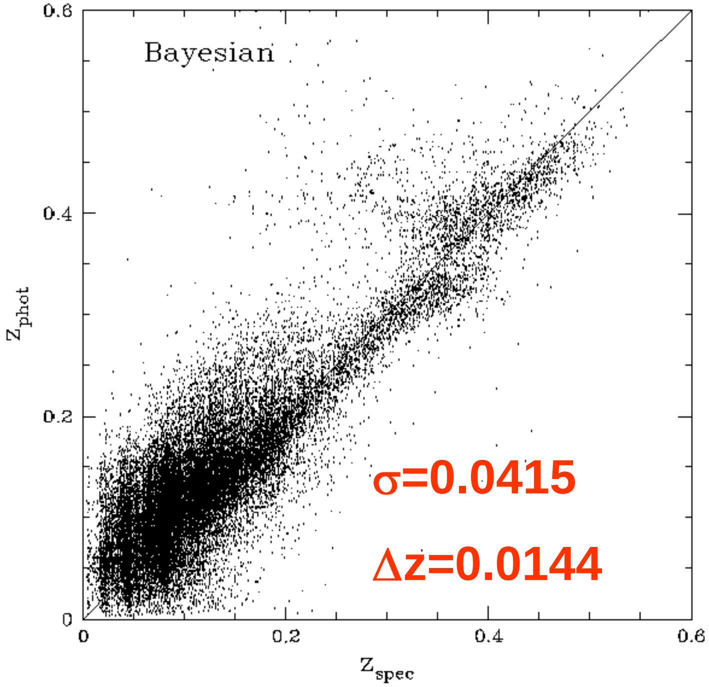
Fig. 1.— The spectroscopic redshift histogram for the SDSS main EDR (solid), the EDR LRG (long dash), the 2dF (short dash) and the CNOC2 sets.

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

Some results

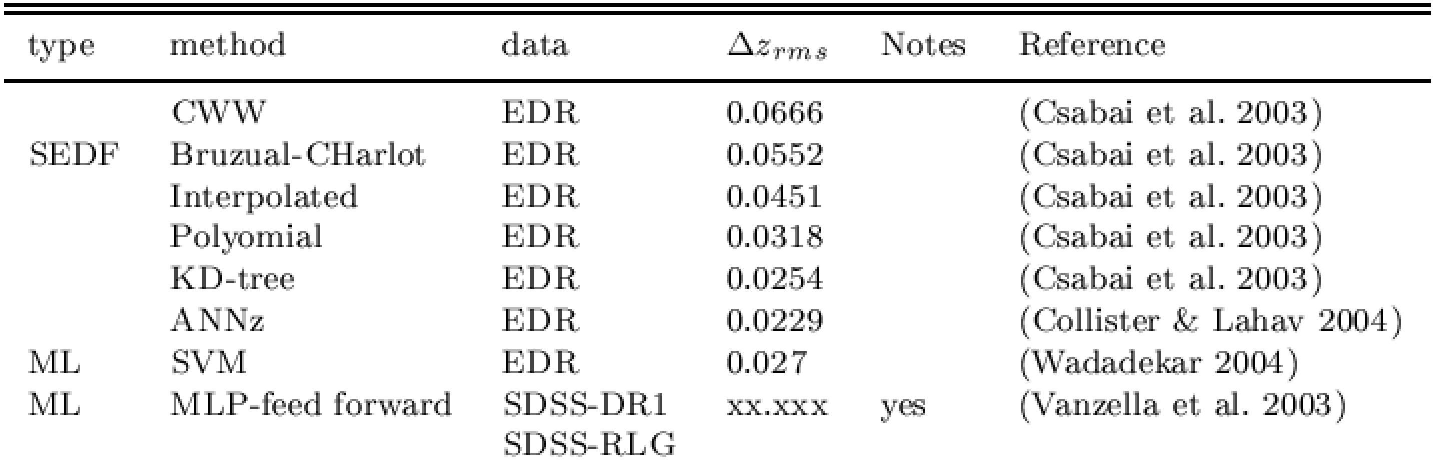


σ=0.051



σ=0.0415

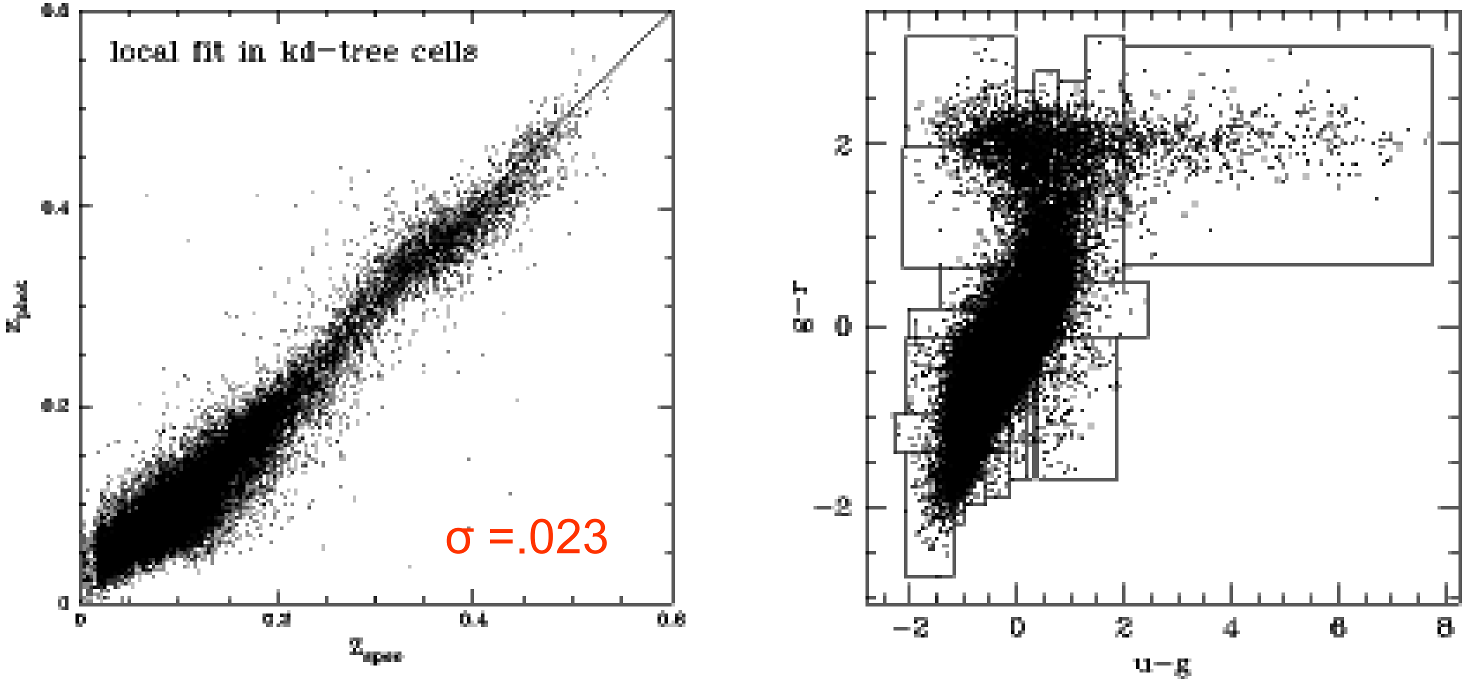
Δz=0.0144



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)
ML	MLP-feed forward	SDSS-DR1 SDSS-RLG	xx.xxx	yes	(Vanzella et al. 2003)

hybrid interpolation+nearest neighbor

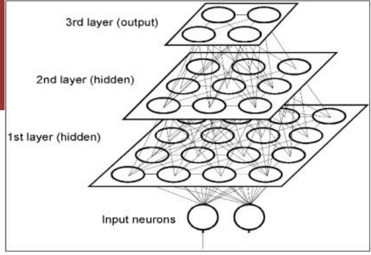
- the color space is partitioned (KD-tree - a binary search tree) into cells containing the same number of objects from the training set
- In each cell fit a second order polynomial.

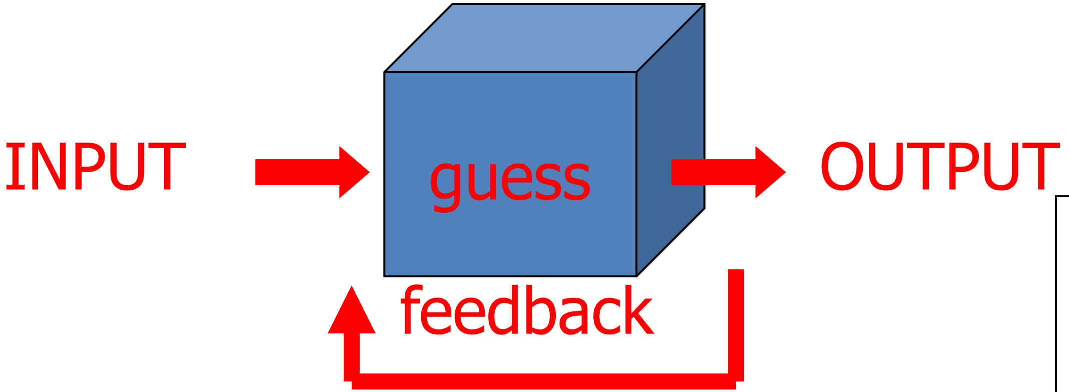


σ =.023

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.

Multi Layer Perceptron





INPUT

guess

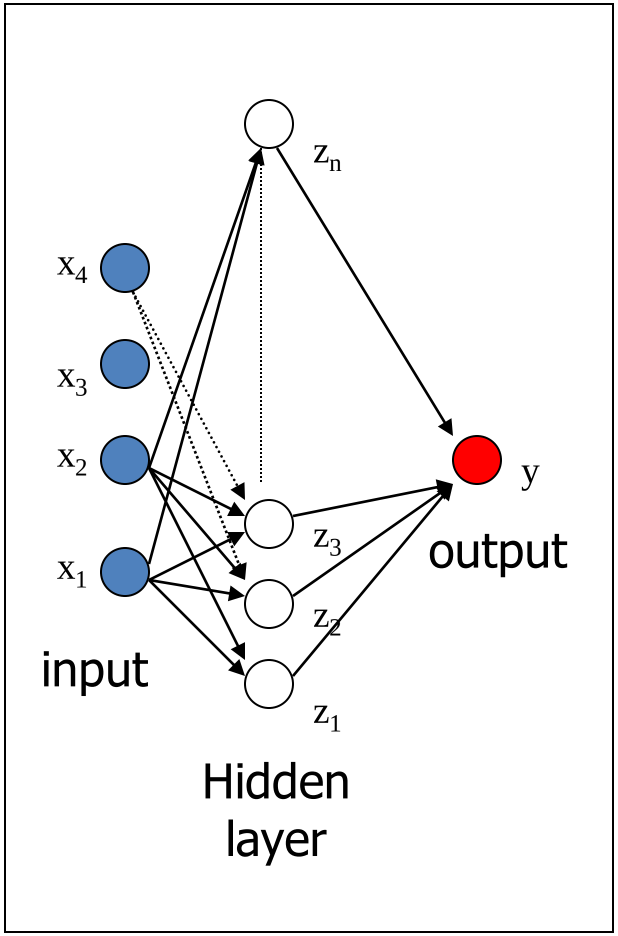
OUTPUT

feedback

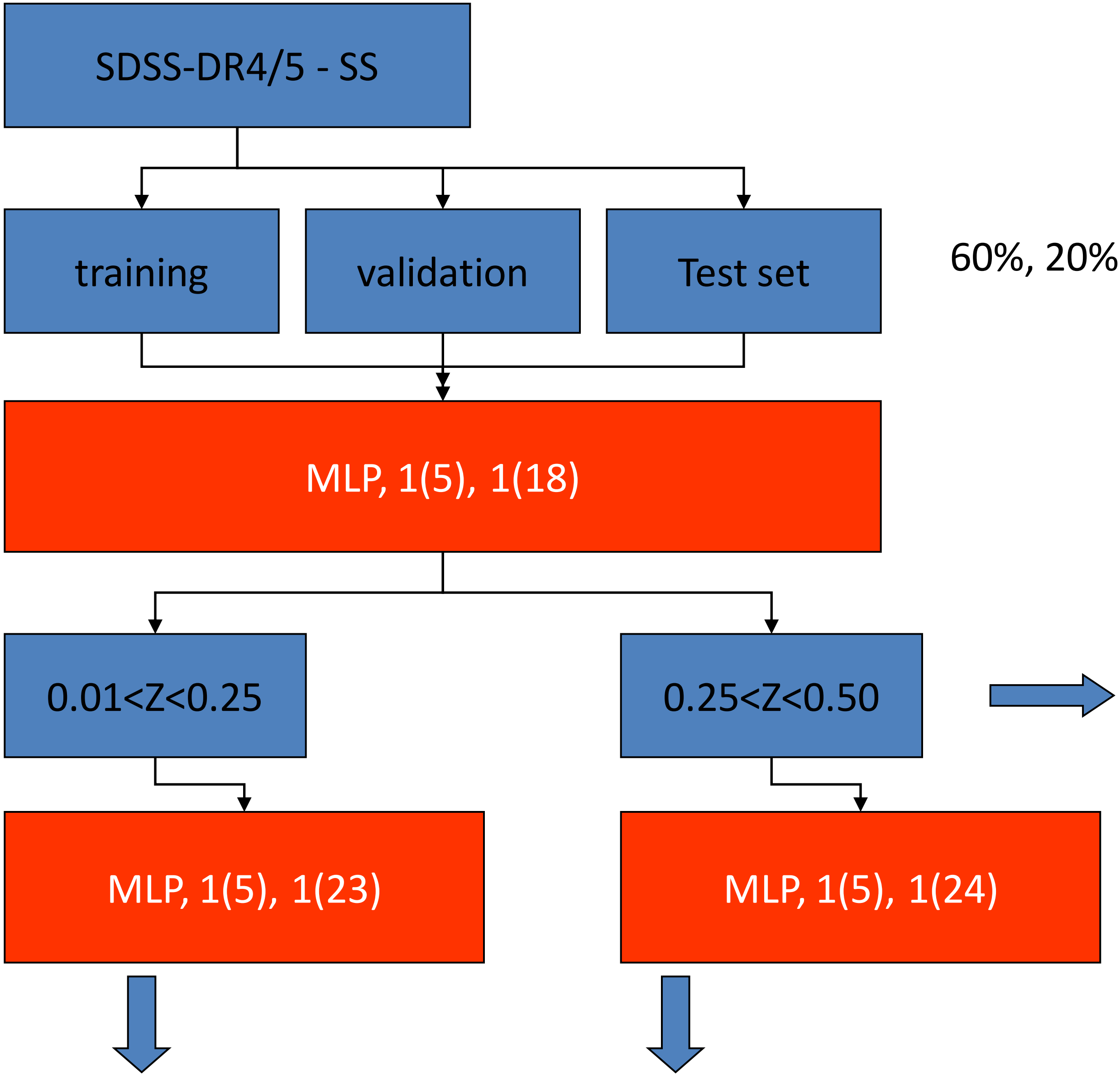
- input layer (n neurons)
- M hidden layer (1 or 2)
- Output layer (n' <n neurons)

Neurons are connected via activation functions

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



VO-Neural approach



SDSS-DR4/5 - SS

training

validation

Test set

60%, 20%, 20%

MLP, 1(5), 1(18)

0.01<Z<0.25

0.25<Z<0.50

99.6 % accuracy

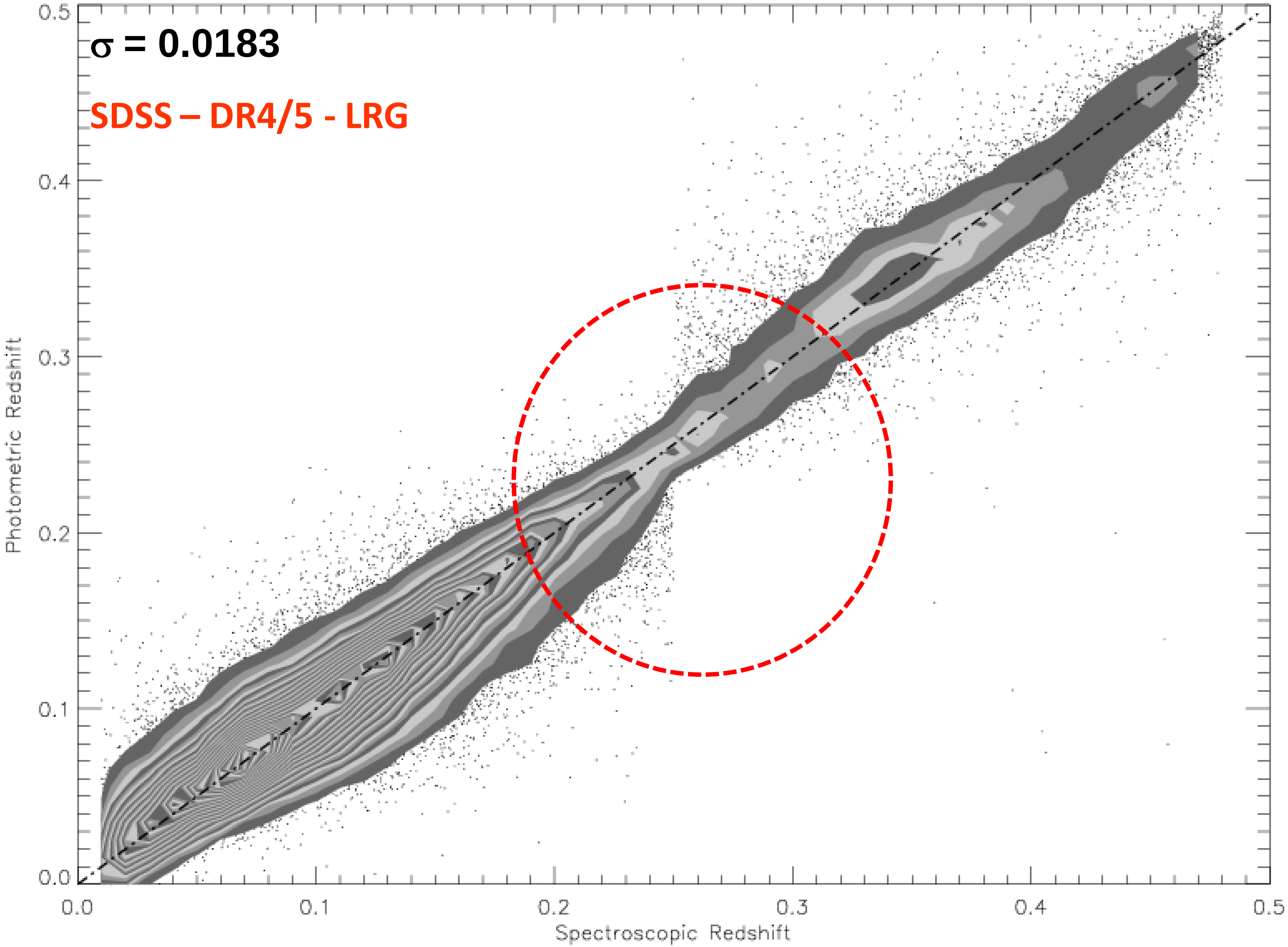
MLP, 1(5), 1(23)

MLP, 1(5), 1(24)

σ rob = 0.196

σ rob = 0.201

VO-Neural results



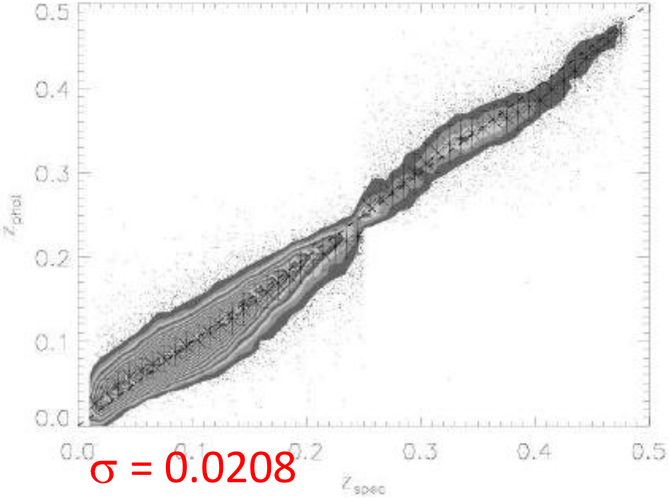
σ = 0.0183

SDSS – DR4/5 - LRG

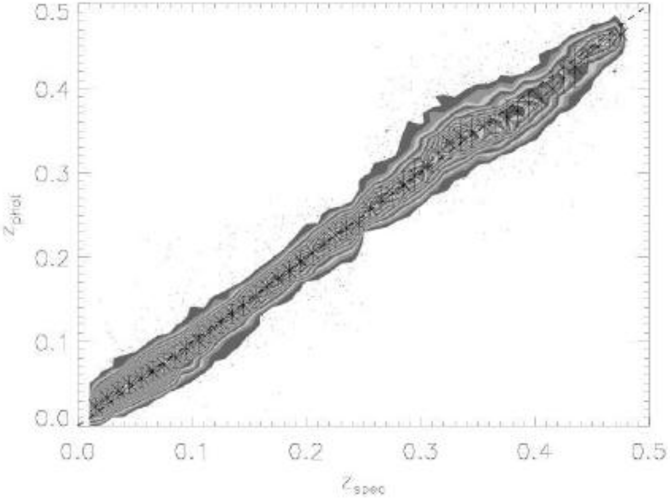
Uneven coverage of parameter space:

General galaxy sample

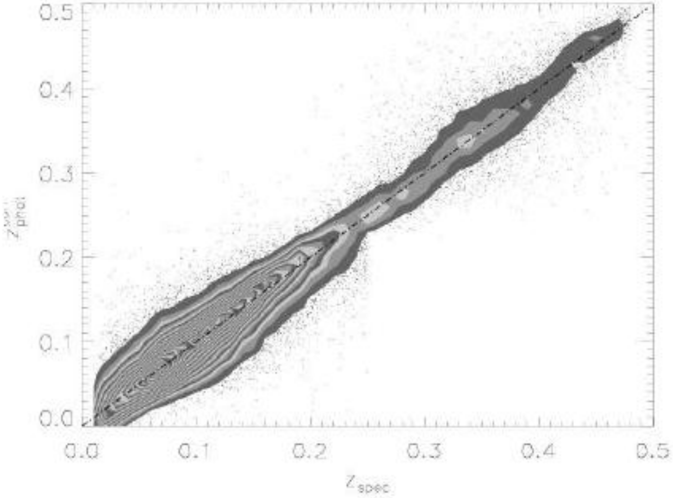
LRG sample

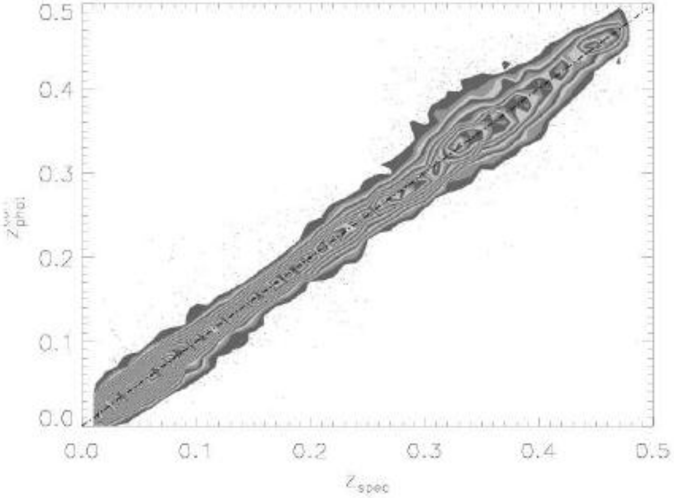


σ = 0.0208
Δz = -0.0029



σ = 0.0178
Δz = -0.0011

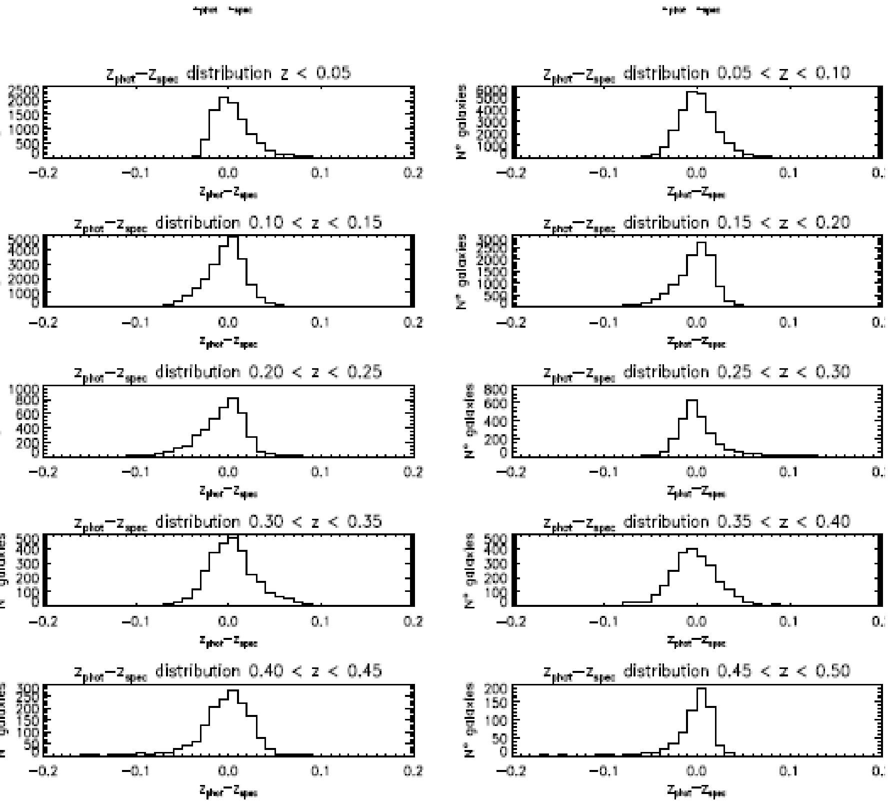




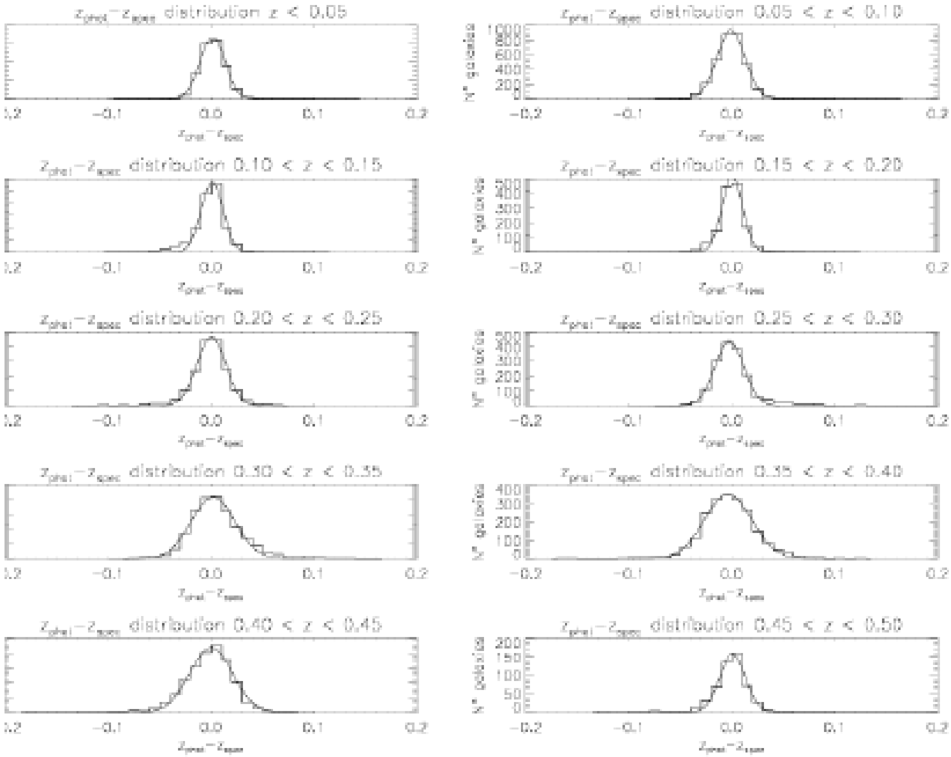
Non LRG only

σ = 0.0363
Δz = -0.0030

Errors can be easily evaluated



General galaxy sample

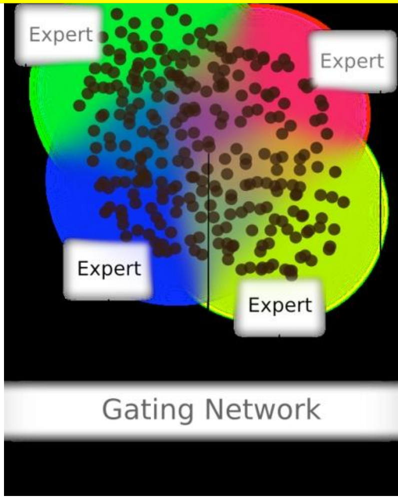


LRG sample

And are, on average, well behaved….

What do we learn if the BoK is biased:

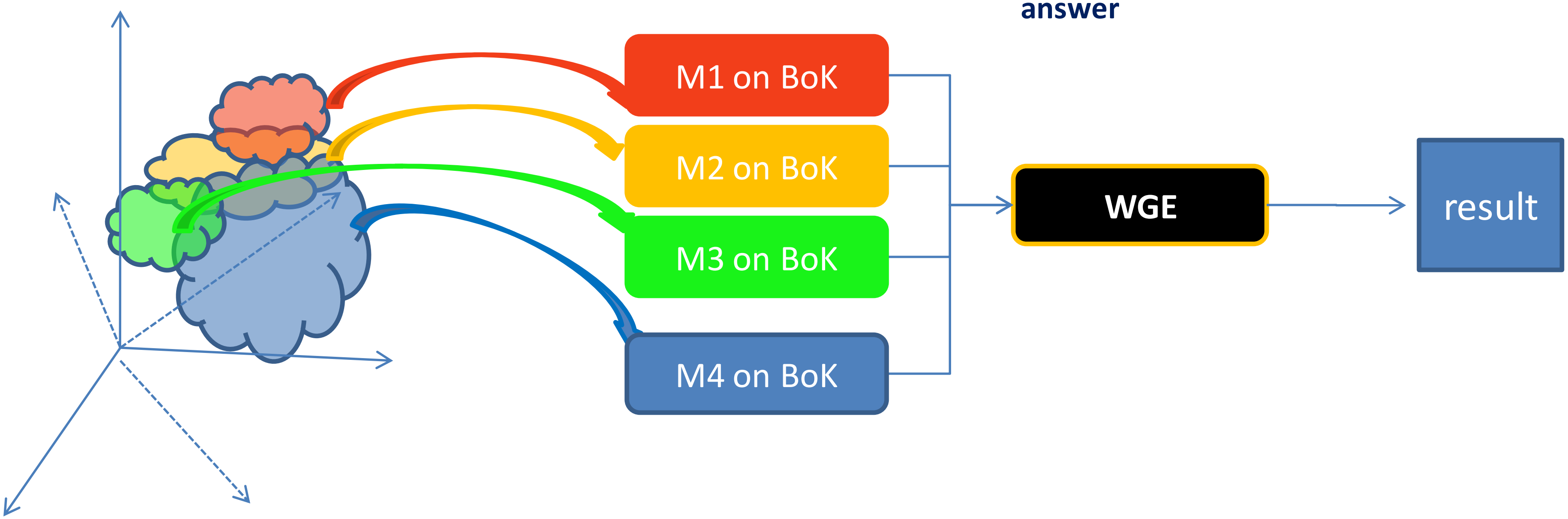
- At high z LRG dominate and interpolative methods are not capable to “generalize” rules
- An unique method optimizes its performances on the parts of the parameter space which are best covered in the BoK



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

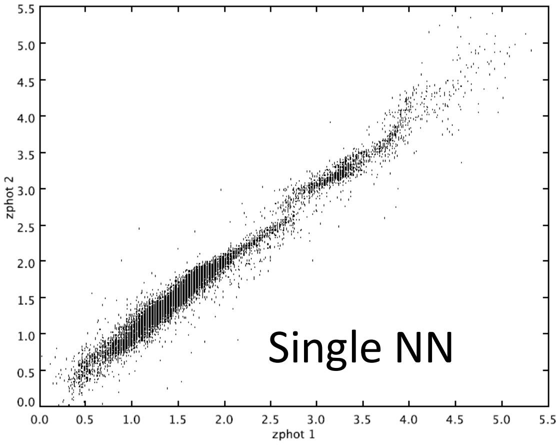


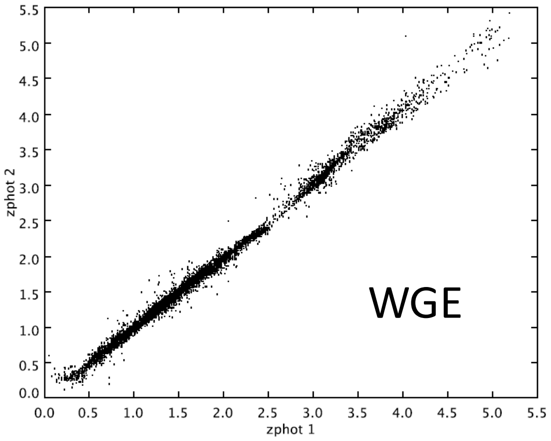
Laurino et al. 2009a,2009b

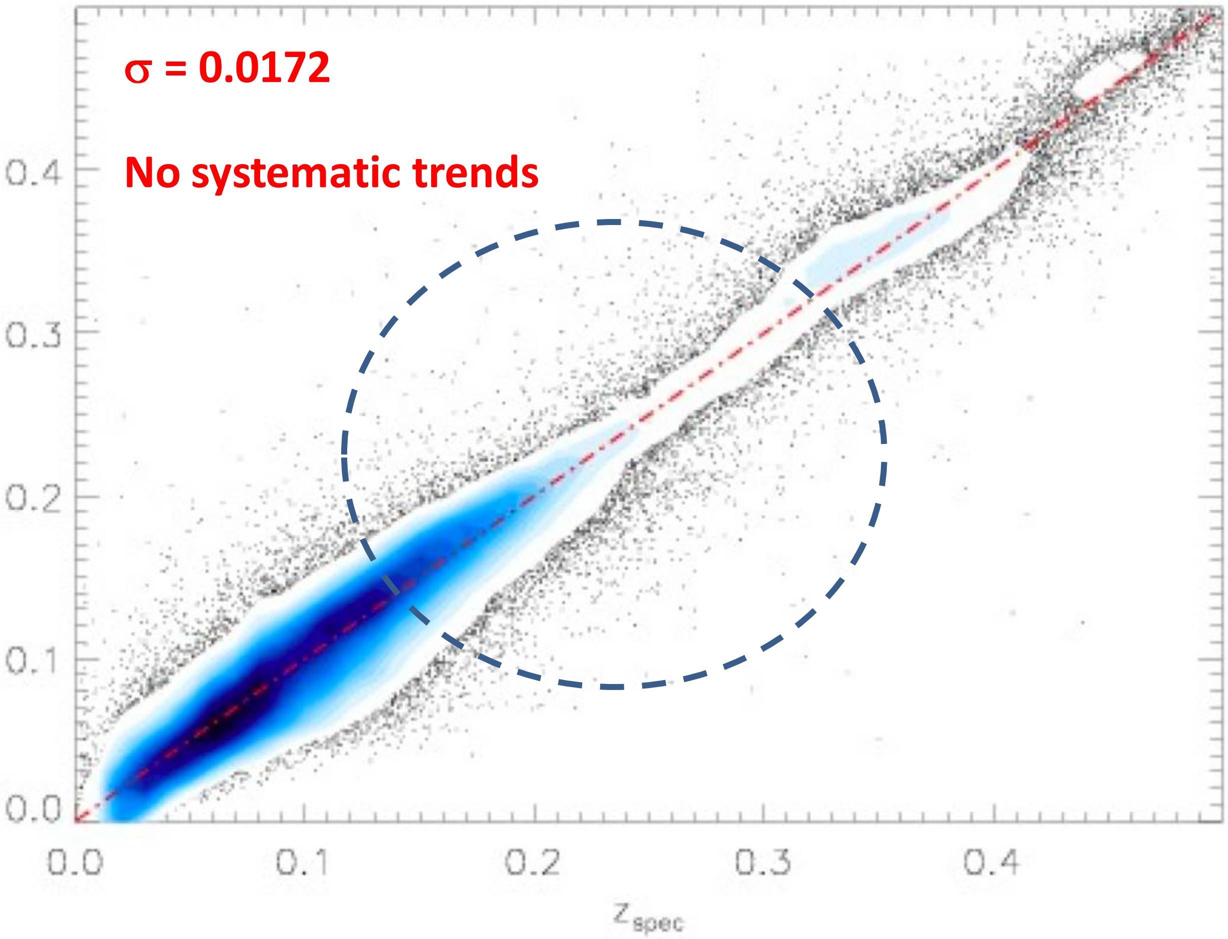
Science with Dame

1. Photometric redshifts of galaxies

Laurino et al. 2009a,2009b







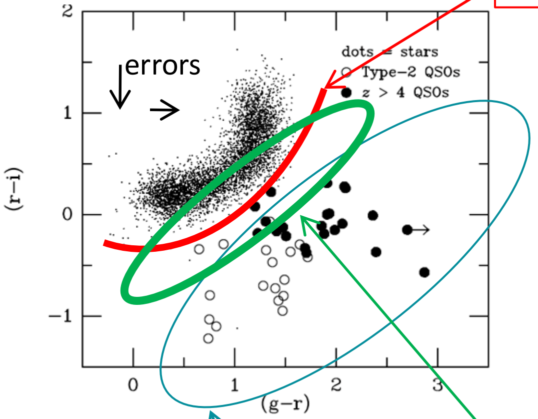
σ = 0.0172

No systematic trends

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

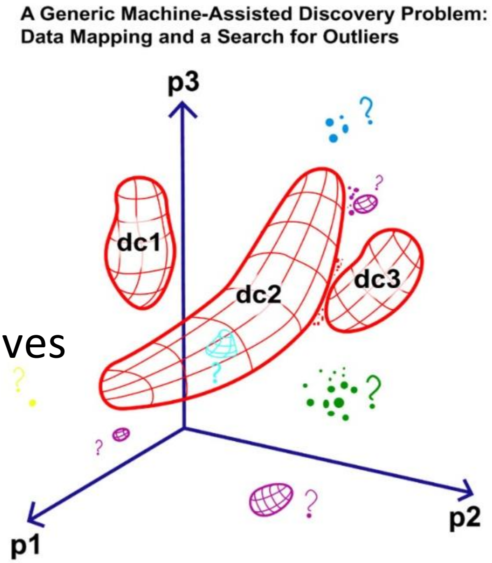


Cutoff line

Candidate QSOs for spectroscopic follow-up's

Ambiguity zone

Adding one feature improves separation…


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

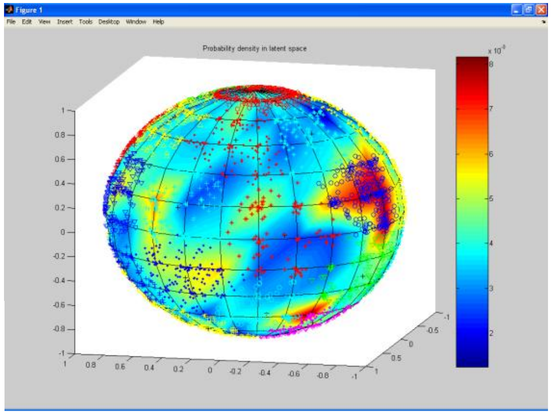
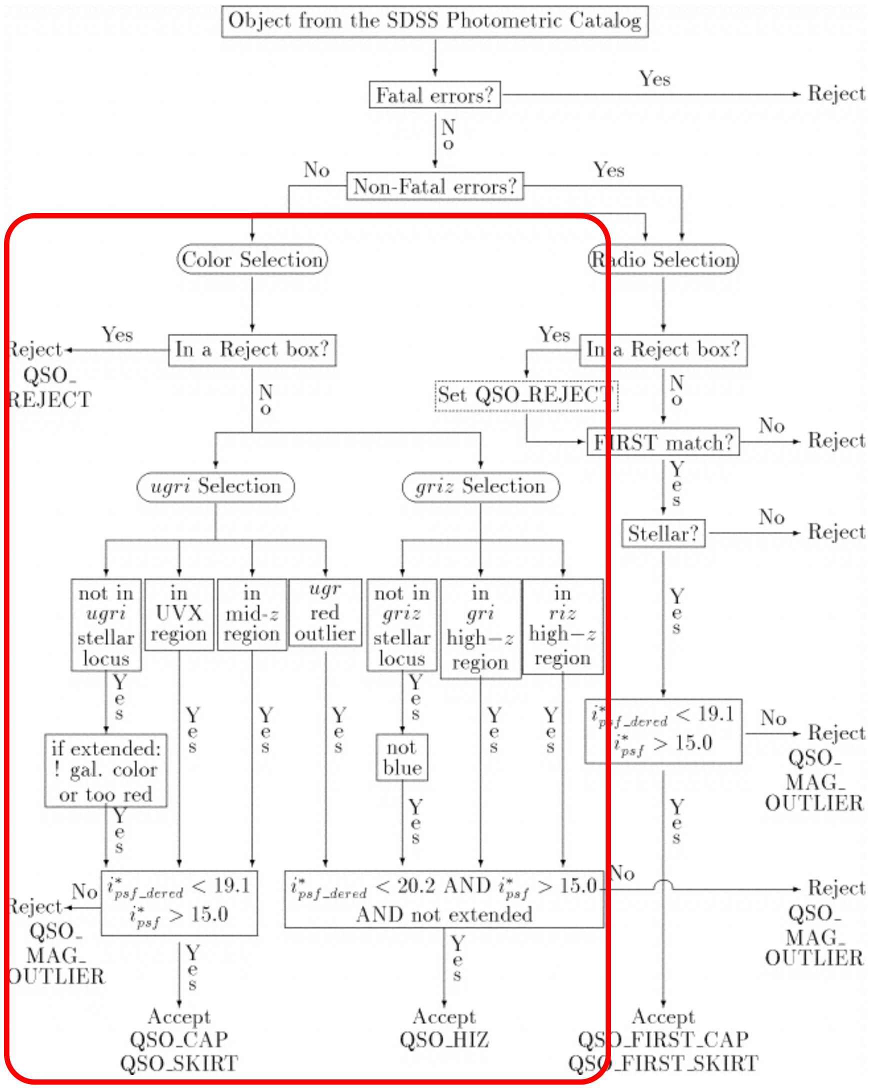


Figure 1
File Edit View Insert Tools Desktop Window Help
Probability density in latent space

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

1. SDSS selection algorithm

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:



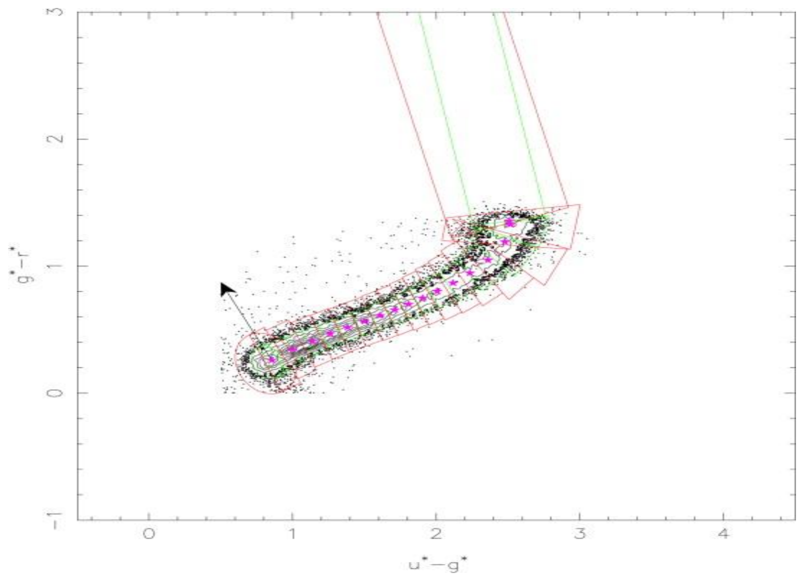
- **QSOs are supposed to be placed >4σ far from a cylindrical region containing the “stellar locus”** (S.L.), where σ depends on photometric errors.

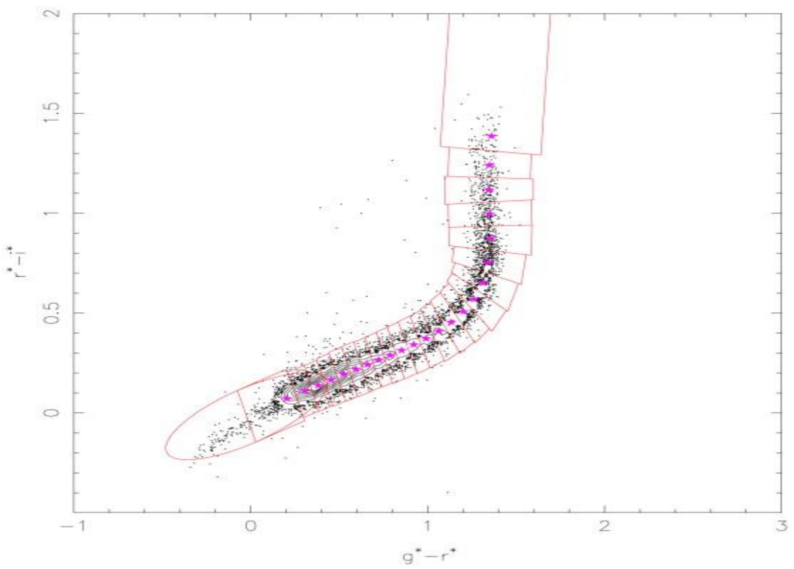
OR

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

c = 95%, e = 65%
locally less

1. SDSS selection algorithm





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.

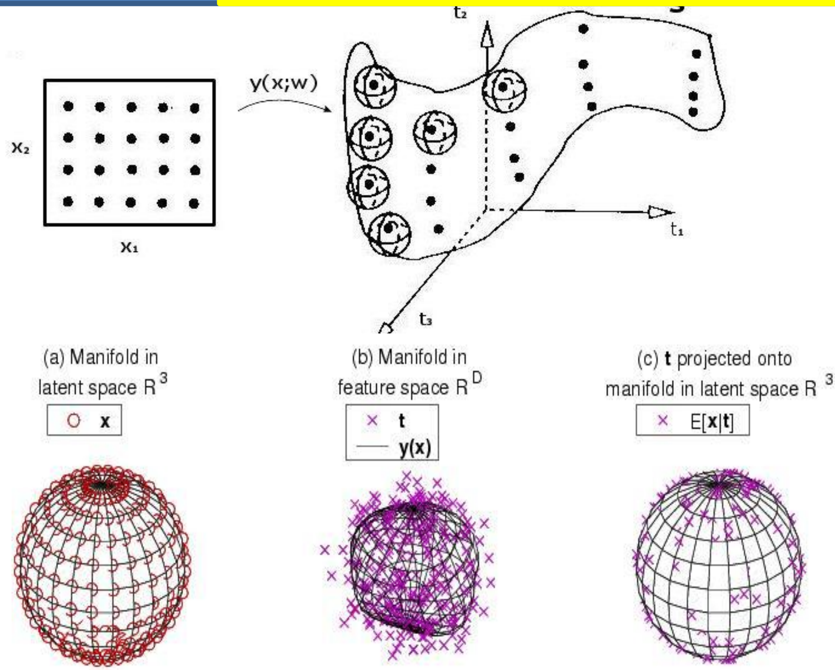
DAta Mining & Exploration

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.



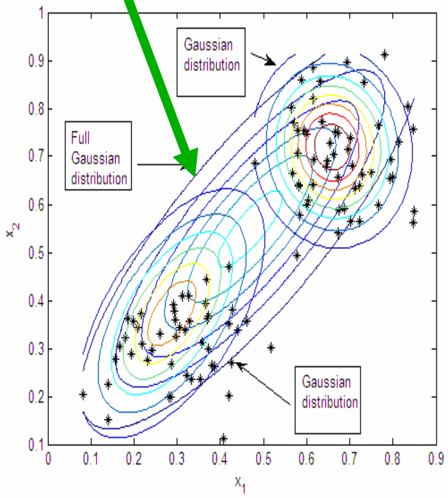
(a) Manifold in latent space R^3

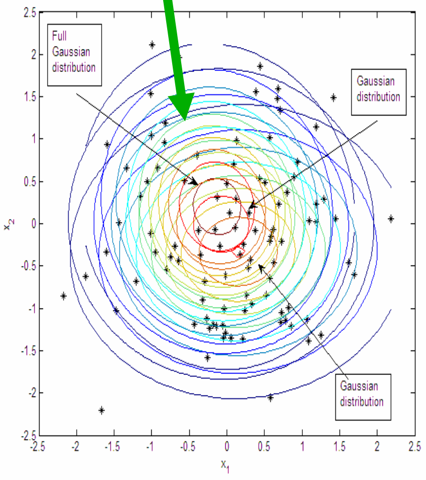
(b) Manifold in feature space R^D

(c) **t** projected onto manifold in latent space R^3

Not replaced!

Replaced!





NegE=750

NegE=4

Step 2: Cluster agglomeration

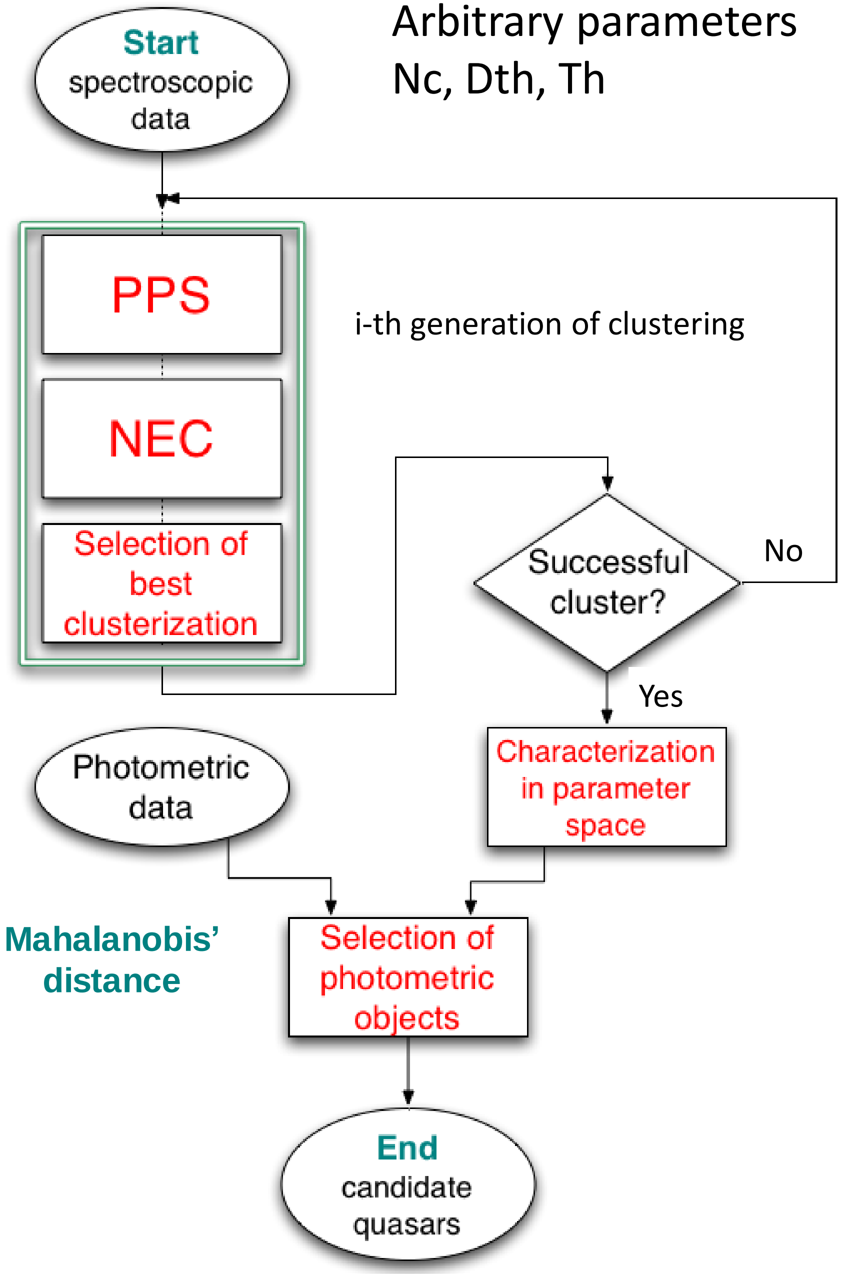
NEC aggregates clusters from PPS to a (a-priori unknown) number of final clusters.

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.

Science with Dame

1. DAME Selection Algorithm



Arbitrary parameters
Nc, Dth, Th

Start
spectroscopic data

PPS

NEC

Selection of best clusterization

i-th generation of clustering

Successful cluster?

No

Yes

Characterization in parameter space

Photometric data

Mahalanobis' distance

Selection of photometric objects

End
candidate quasars

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.

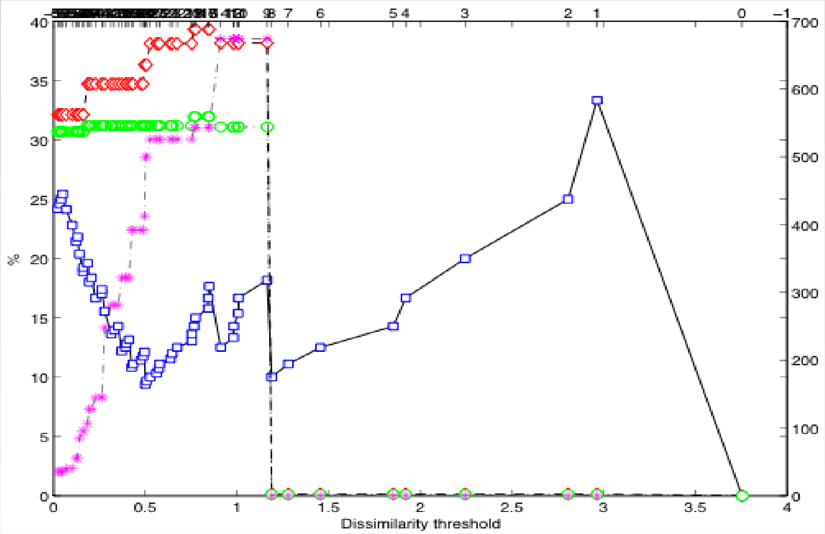
The overall efficiency of the process e_{tot} is the sum of weighed efficiencies e_i for each generation:

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

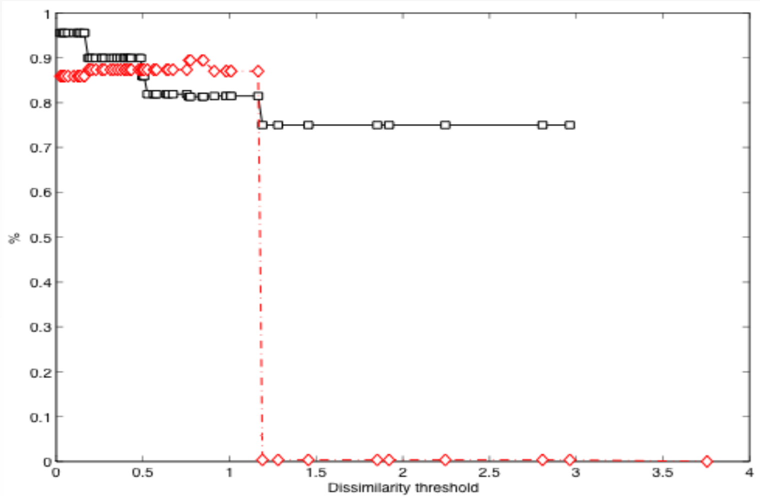
An example of “tuning”

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 %

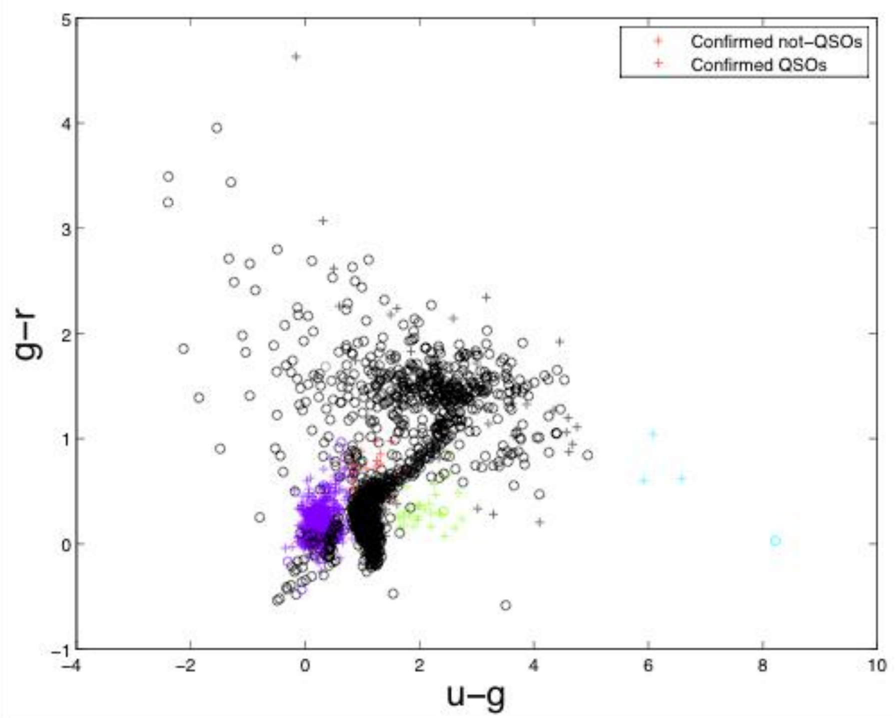
c = 89.6 %

Confusion matrix

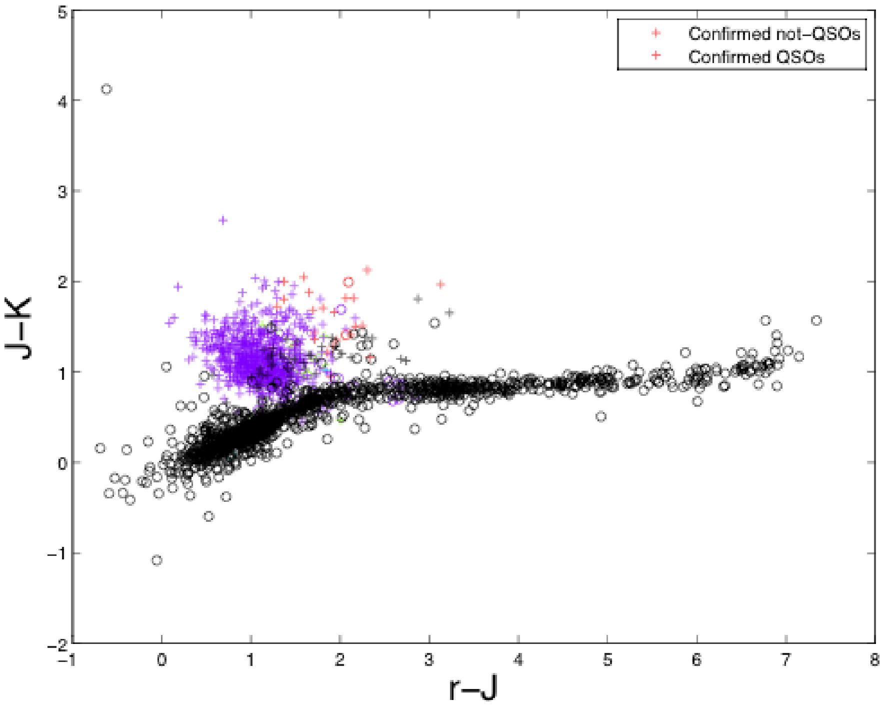
Science with Dame

1. Experiment 2

u - g vs g - r



r - J vs J - K

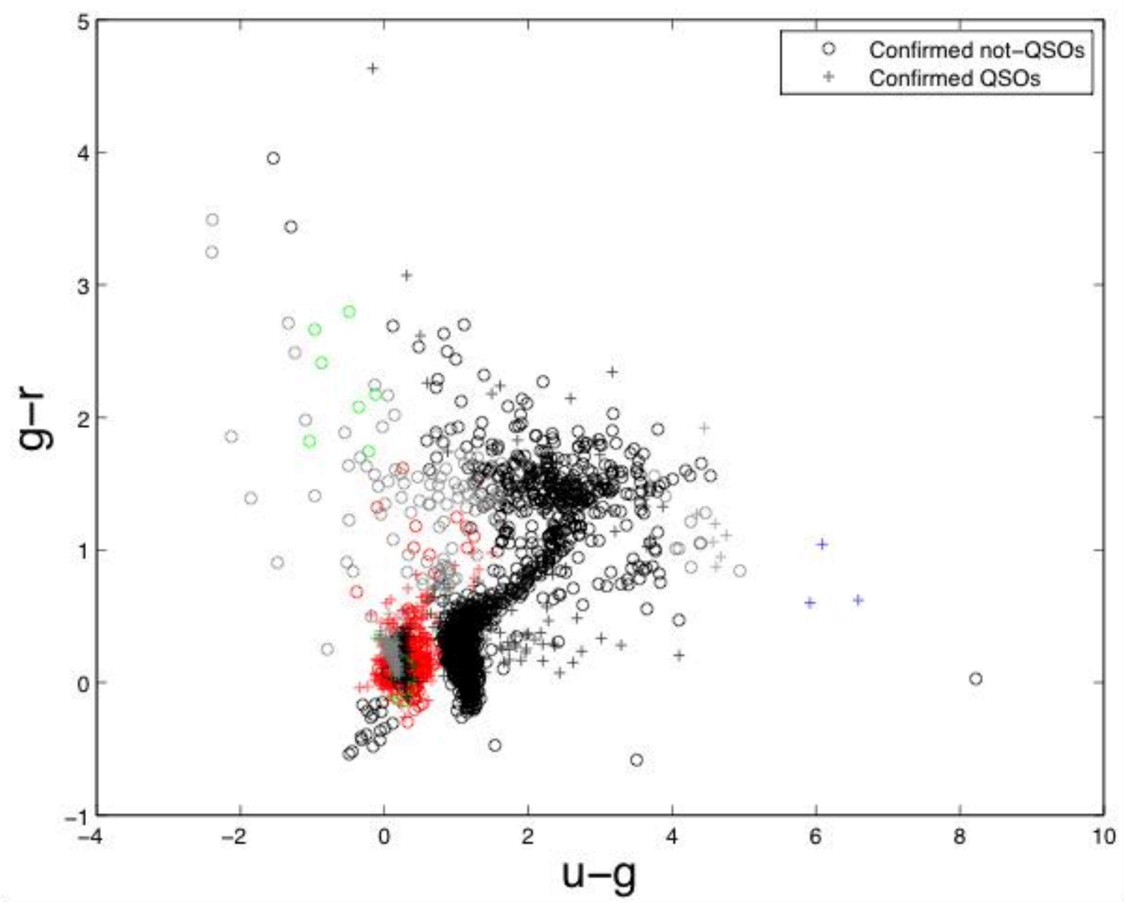


Only a fraction (43%) of these objects have been selected as candidate QSO's by SDSS targeting algorithm in first instance: the remaining sources have been included in the spectroscopic program because they have been selected in other spectroscopic programmes (mainly stars).

Science with Dame

1. **Experiment 2**

u - g vs **g - r**

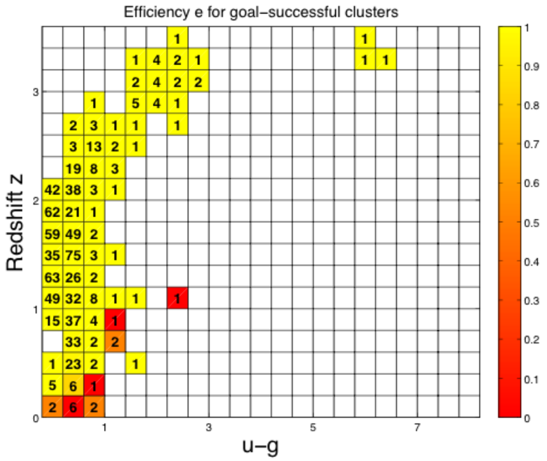


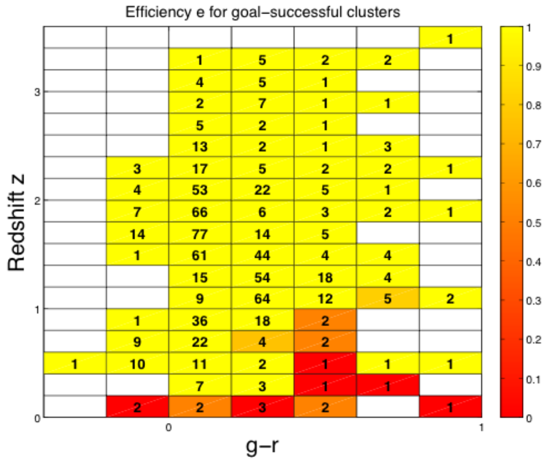
In this experiment the clustering has been performed on the same sample of the previous experiment, using only optical colours.

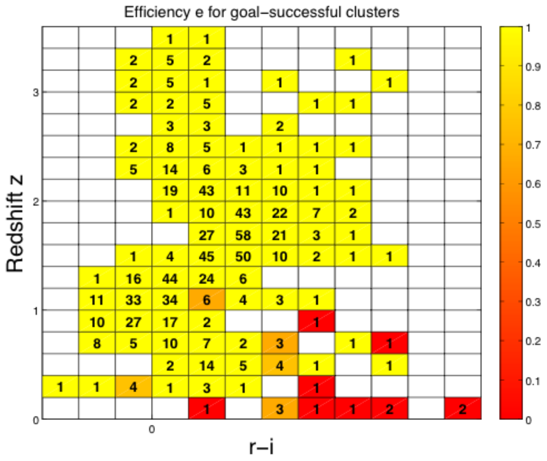
DAta Mining & Exploration

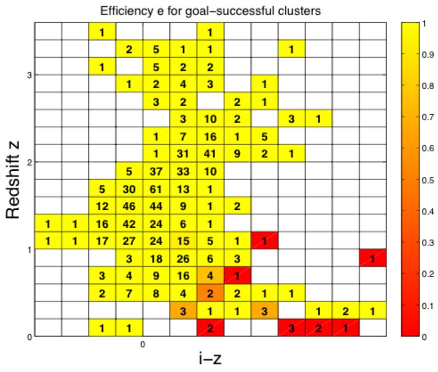
Science with Dame

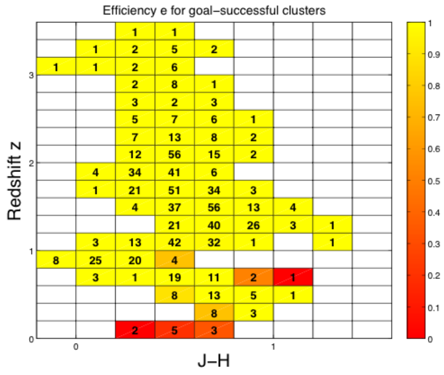
1. **Experiment 2**



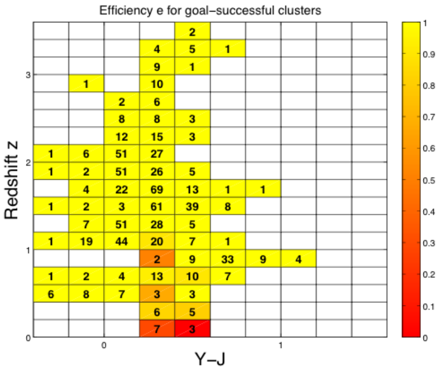


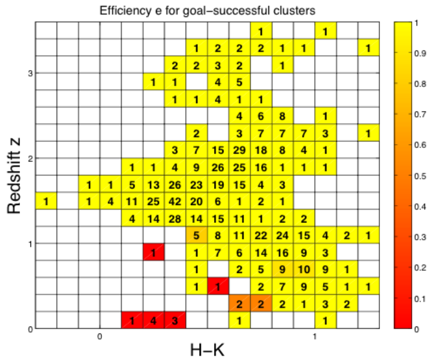




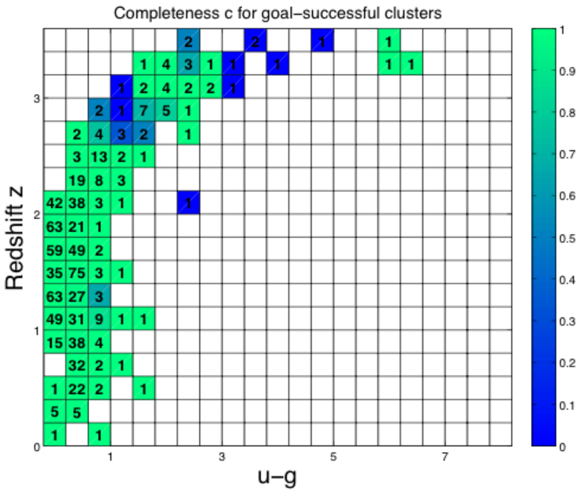


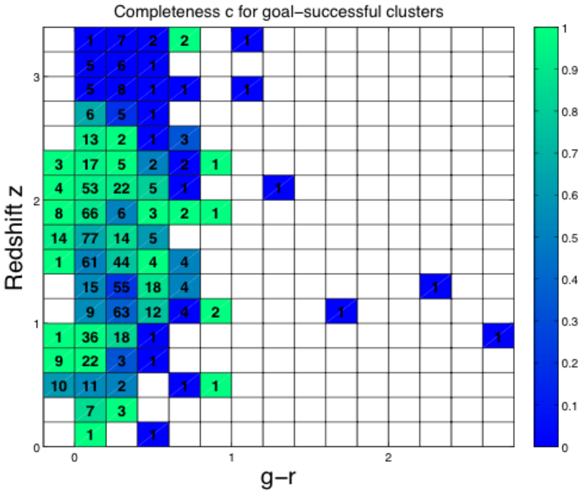
Experiment 2: local values of *e*

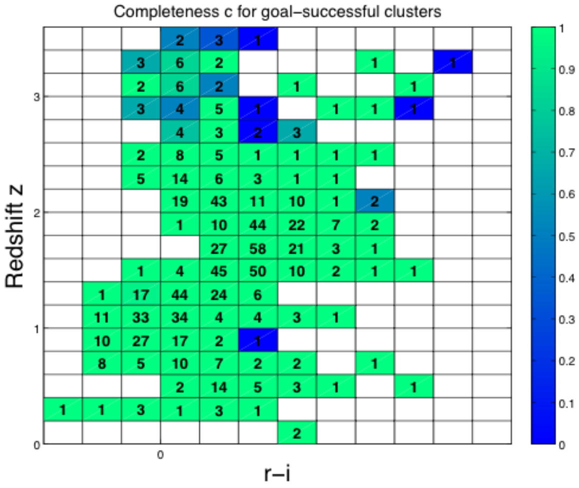


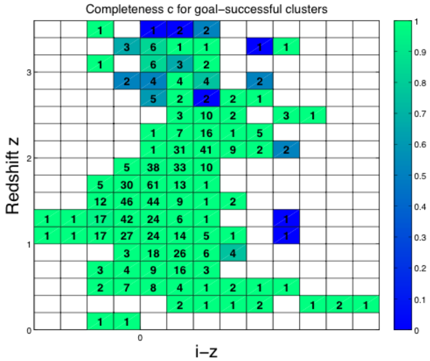


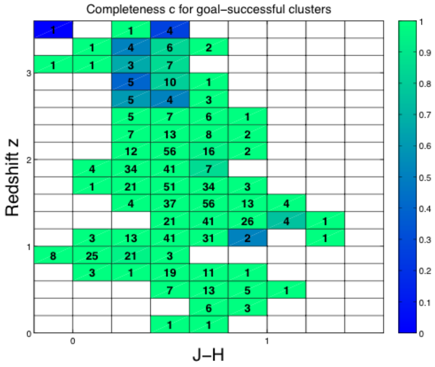
Experiment 2: local values of *c*

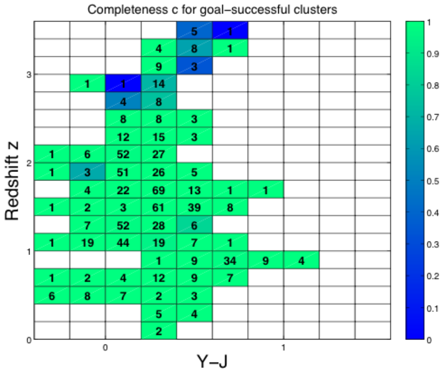


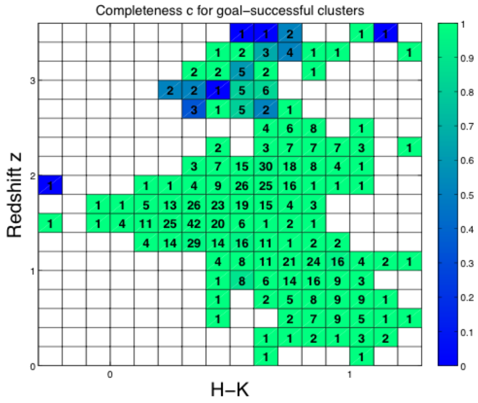












Science with Dame

1. Experiment summary

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)

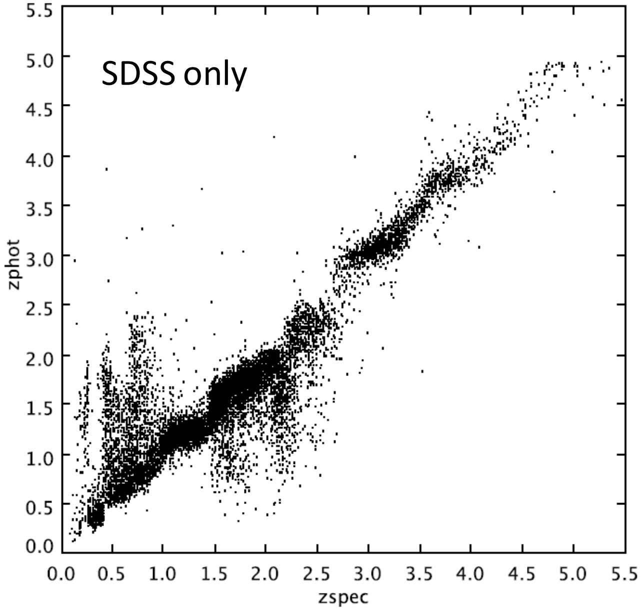
The catalogue of candidate quasars is publicly available at the URL:

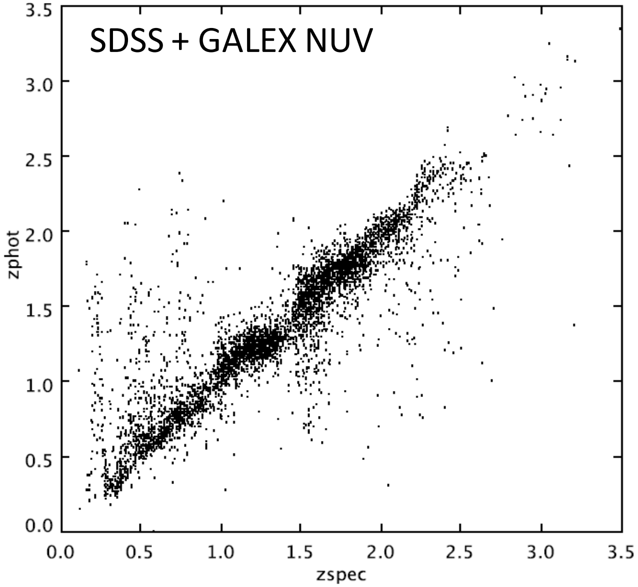
http://voneural.na.infn.it/catalogues_qsos.html

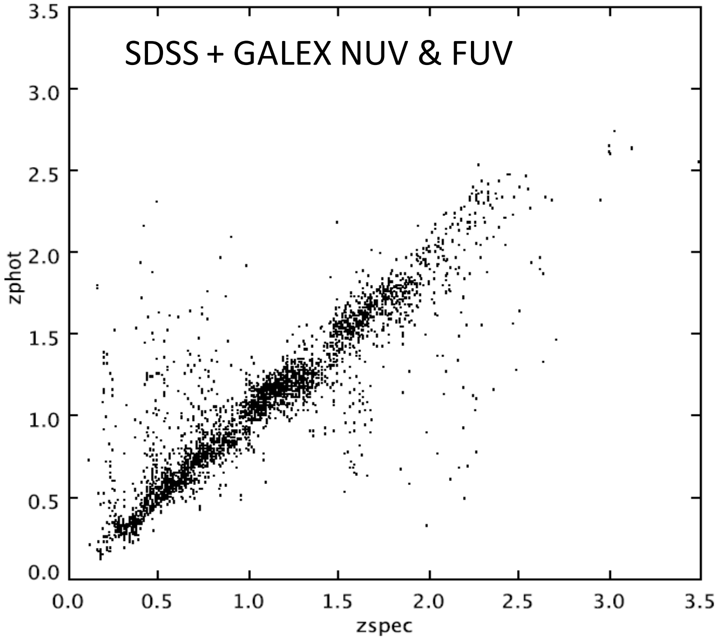
BUT … LET’S GO BACK TO PHOT-Z

Science with Dame

1. Photometric redshifts of QSOs







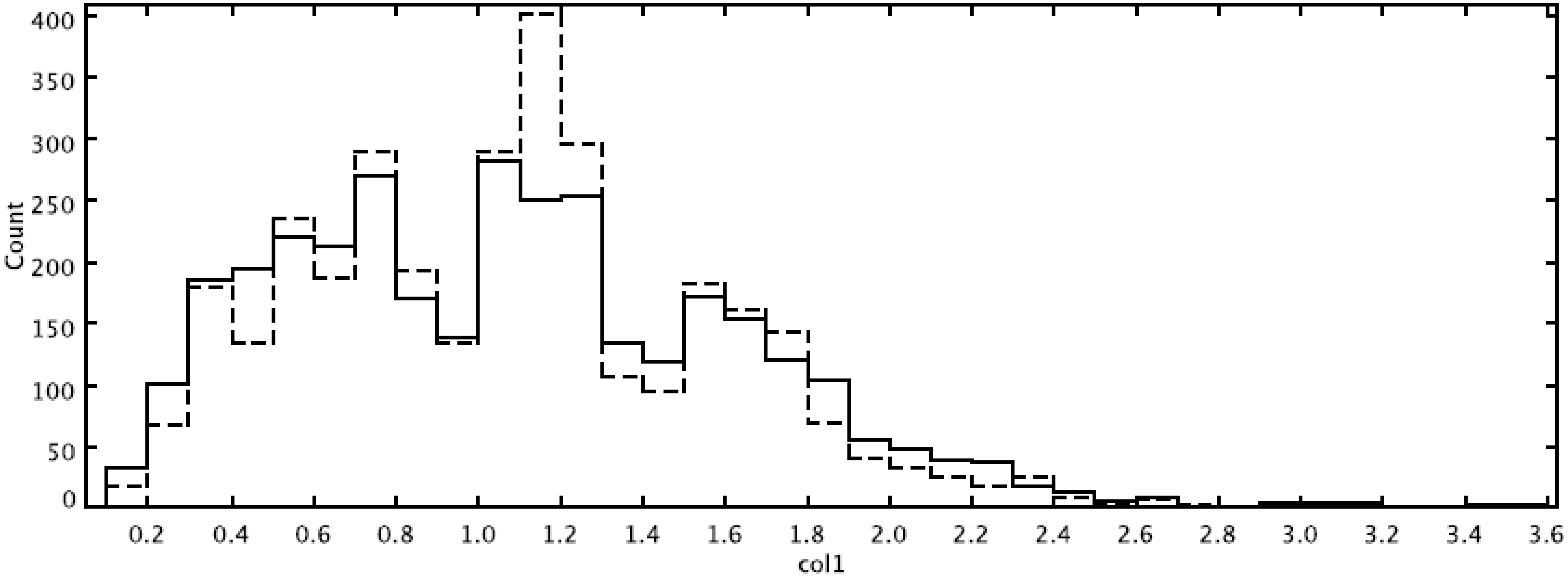
No need for fine tuning !!!

Only New BoK !!!

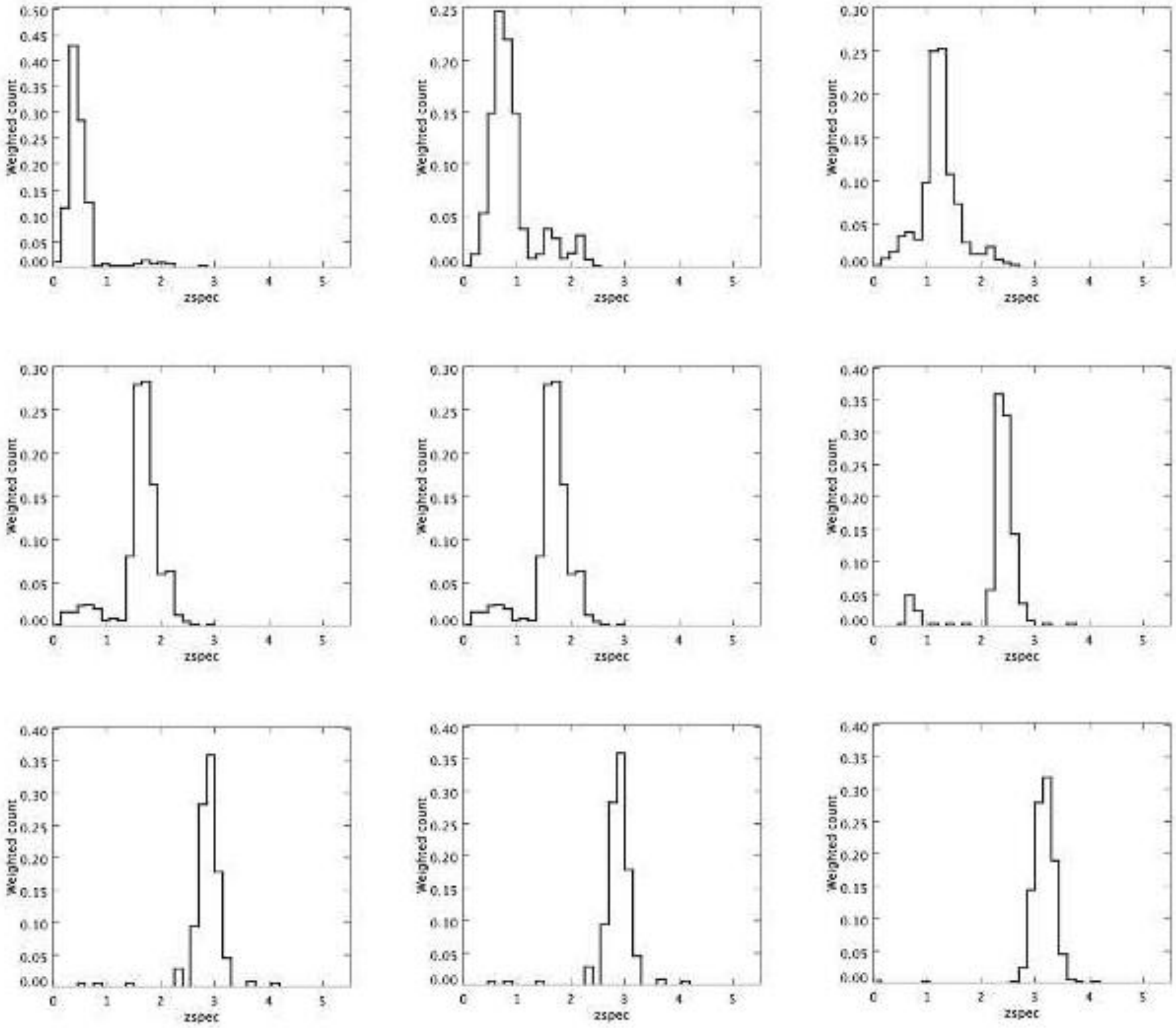
IPAC-Pasadena, August 5 2009

Laurino et al. 2009a,2009b

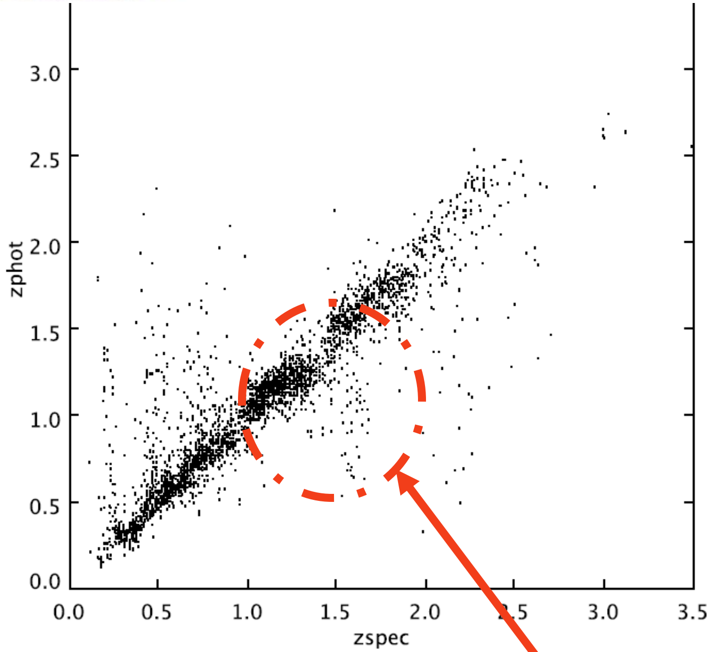
Distribution of Z_spec (solid) and Z_phot (dashed) for test set !!!!



1. Photometric redshifts of QSOs

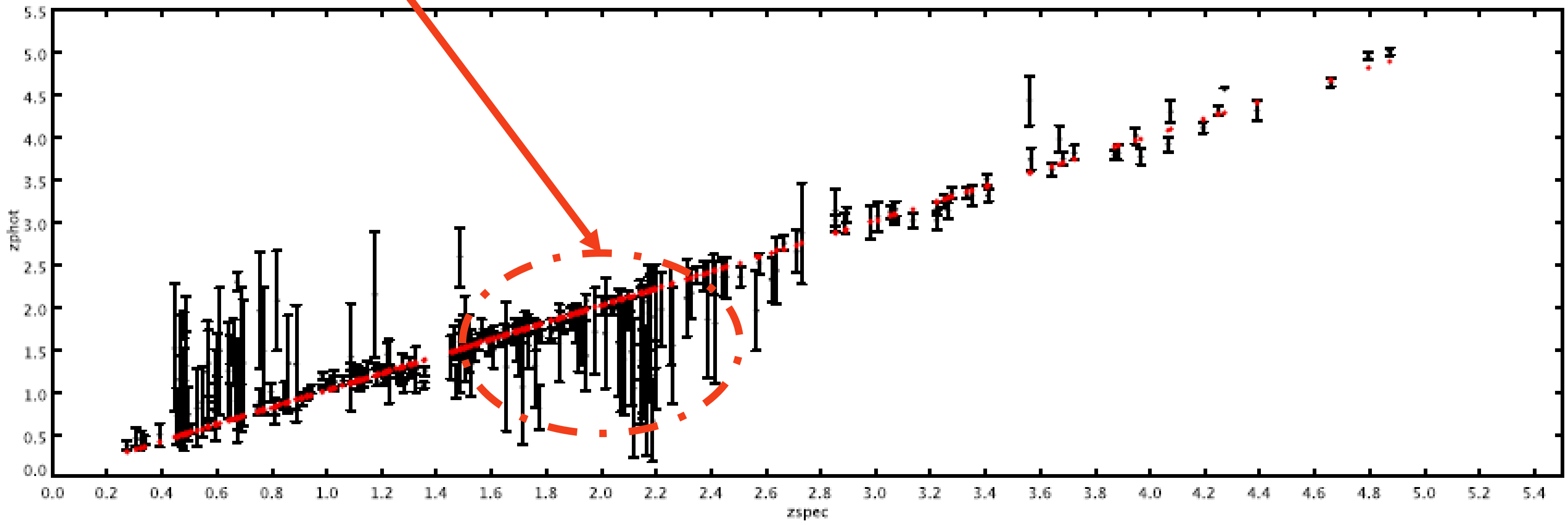


Laurino et al. 2009a,2009b



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

So far restricted choice of problems

Tagliaferri et al. 2003	Ball & Brunner 2009	BoK
S/G separation	S/G separation	Y
Morphological classification of galaxies *(shapes, spectra)*	Morphological classification of galaxies *(shapes, spectra)*	Y
Spectral classification of stars	Spectral classification of stars	Y
Image segmentation	-----	
Noise removal *(grav. waves, pixel lensing, images)*	-----	
Photometric redshifts *(galaxies)*	Photometric redshifts *(galaxies, QSO's)*	Y
Search for AGN	Search for AGN and QSO	Y
Variable objects	Time domain	
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
Interstellar magnetic fields	----	
Stellar evolution models	----	

Limited number of problems due to limited number of reliable BoKs

Bases of knowledge

(set of well known templates for supervised (training) or unsupervised (labeling) methods

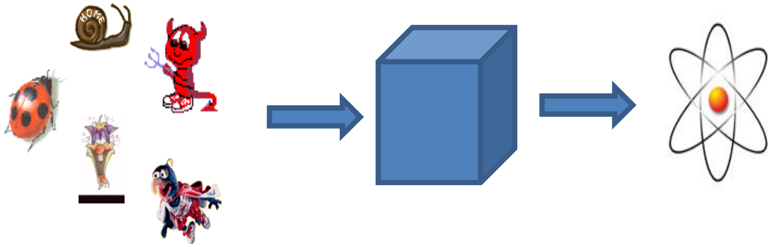
So far

- Limited number of BoK (and of limited scope) available
- Painstaking work for each application (es. spectroscopic redshifts for photometric redshifts training).
- Fine tuning on specific data sets needed (e.g., if you add a band you need to re-train the methods)

Bases of knowledge need to be built automatically from Vobs Data repositories

Community believes AI/DM methods are black boxes

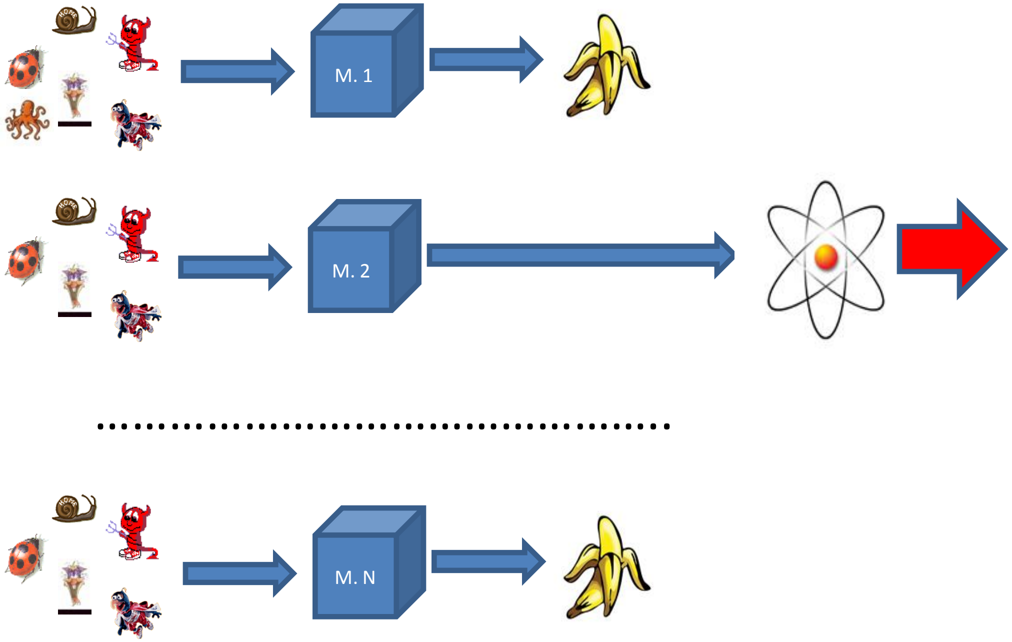
You feed in something, and obtain patters, trends, i.e. knowledge….



Exposed to a wide choice of algorithms to solve a problem, the r.m.s. astronomer usually panics and is not willing to make an effort to learn them ….

The r.m.s astronomer doesn’t want to become a computer scientist or a mathematician (large survey projects overcome the problem)

Tools must run without knowledge of GRID/Cloud no personal certificates, no deep understanding of the DM tool etc.)



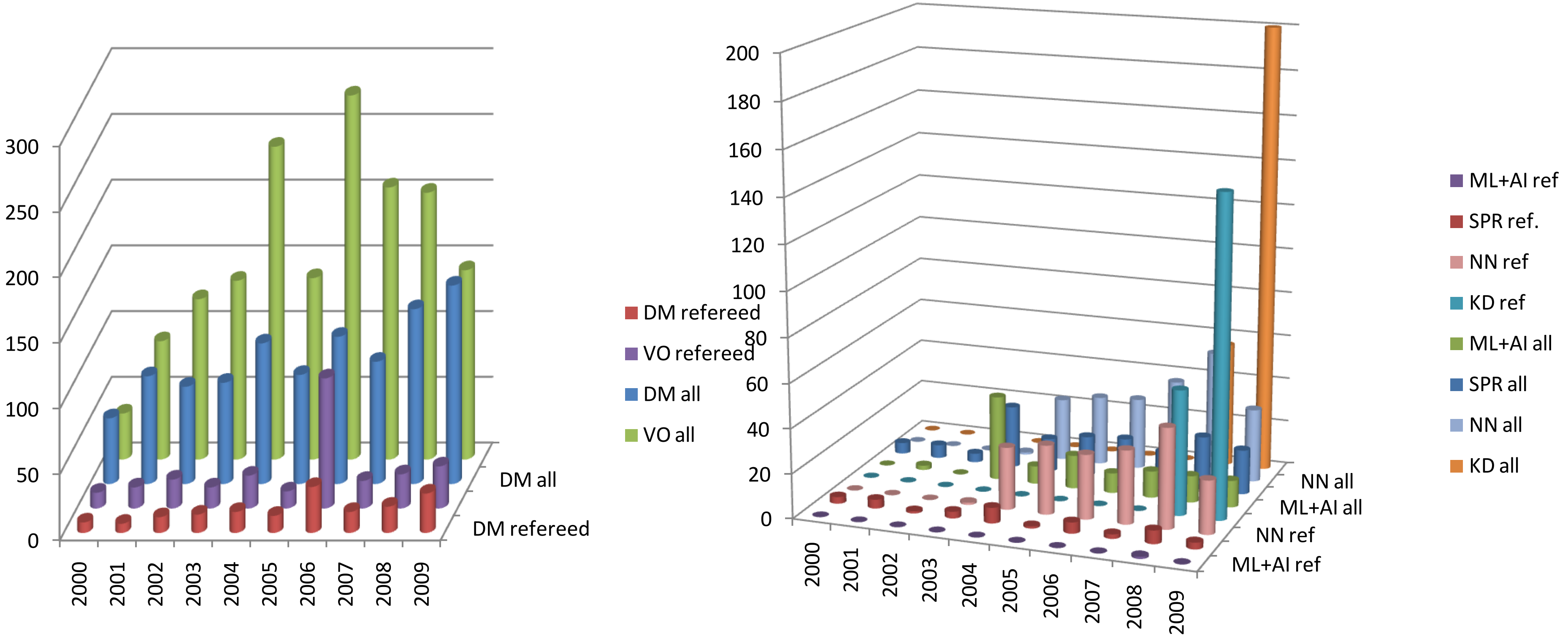
Formation of a new generation of experts

(…… suggesting the solutions)

AND/OR

Implementation of a second generation of tools

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

Recent past	Now	Near Future
Separated archives and data centers *(few TB)*	**Federated archives and data centers** *(10 – 100 Tbyte)*	**Virtual Observatory, LSST, SKA** *(1-1000 Pbyte)*
No common standards (*.fits)	Common standards (*.fits, *.vot, etc.)	Common standards (*.fits, *.vot, etc.)
Little bandwith (10/50 Kb s^{-1})	**Larger bandwith (100-1000 Kb s^{-1})** *(last mile problem)*	Largerbandwith (> 1-10 Gb s^{-1})
Single CPU processing	Still single CPU processing	GRID/Cloud computing/Multicore

Research praxis

Few objects , few information *(parameter space ~ 10 features)*	**Many objects , much information** *(parameter space > 100 features)*	**Whole sky, multi-λ, multi epoch catalogues** *(parameter space > 100 features)*
Traditional statistics	**Multi variate statistics**	**Statistical Pattern Recognition (DM and ML)**
		This is only a part of the game **(*size and not complexity driven)***

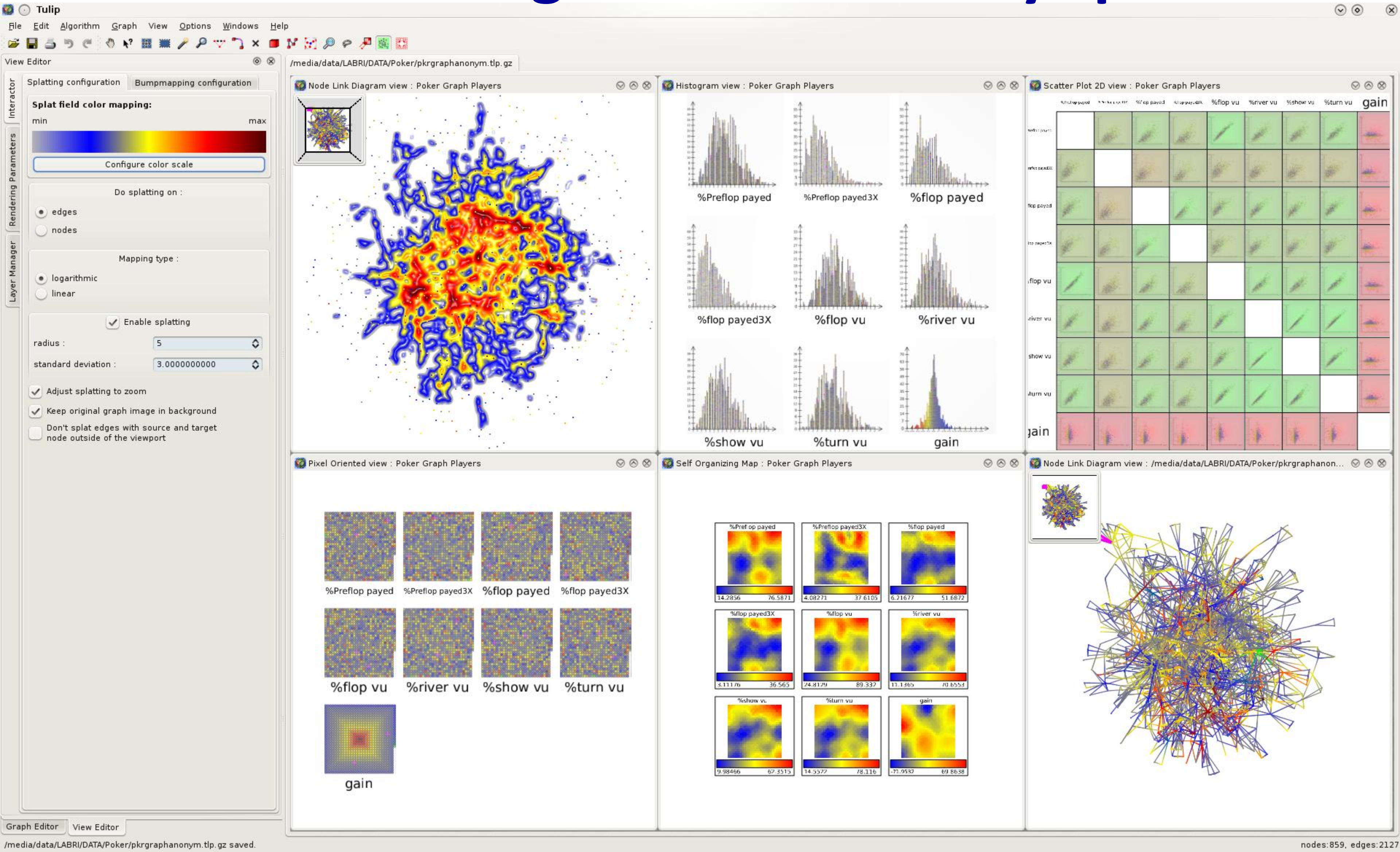
Future developments and some conclusions

- **Better visualization tools for high dimensionality data**

- More machine learning methods

- Parallelization of some codes

Visualization of high dimensionality spaces



Tulip

File Edit Algorithm Graph View Options Windows Help

View Editor

Splatting configuration | Bumpmapping configuration

Splat field color mapping:

min max

Configure color scale

Do splatting on :

● edges

○ nodes

Mapping type :

● logarithmic

○ linear

☑ Enable splatting

radius : 5

standard deviation : 3.0000000000

☑ Adjust splatting to zoom

☑ Keep original graph image in background

☐ Don't splat edges with source and target node outside of the viewport

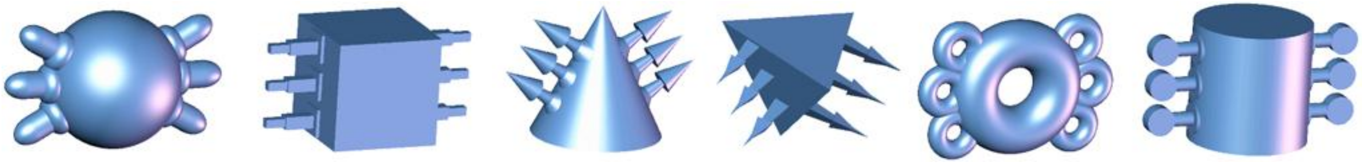
Graph Editor | View Editor

/media/data/LABRI/DATA/Poker/pkrgraphanonym.tlp.gz saved.

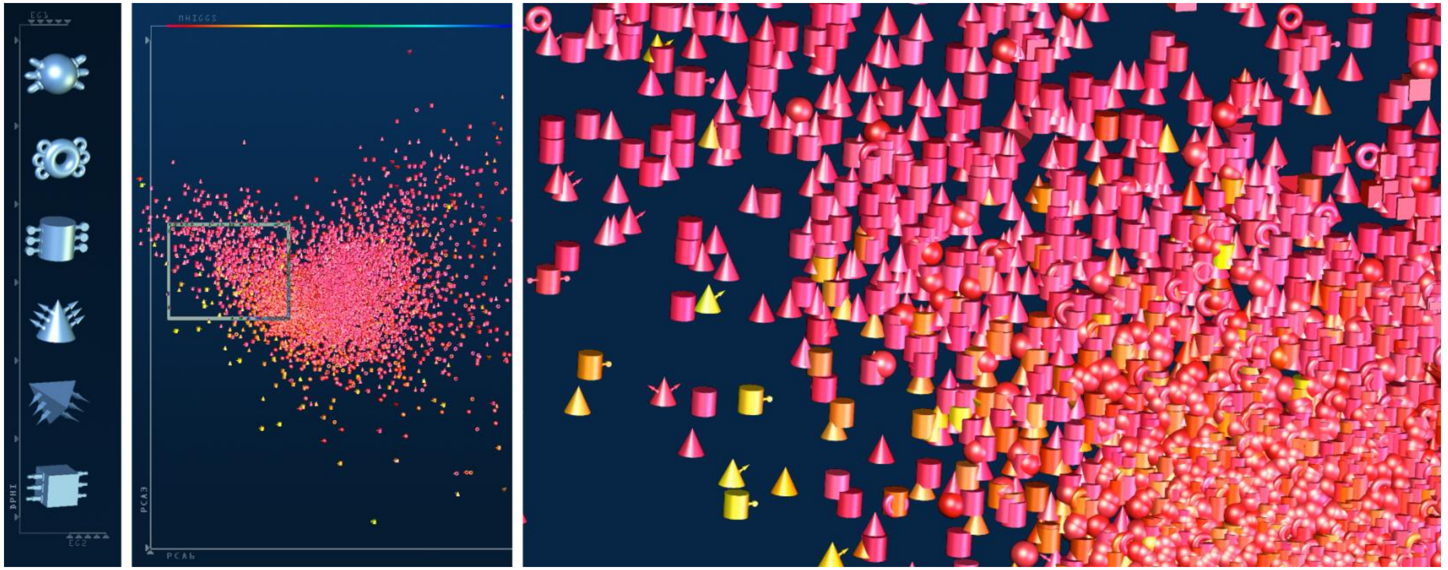
nodes:859, edges:2127

Courtesy of S. V. Lombeyda– Astroinformatics 2010

S. V. Lombeyda



(read order)	(attribute)
1,2	*position (x,y)*
3	*shape*
4	*hue*
5	*left features* }
6	*right features* }
…	*vibration, sound, etc…*



Useful links

DAME: http://voneural.na.inf.it/

IVOA: http://www.ivoa.org/

MICA (Meta Institute for Computational Astrophysics) in Second Life:
http://www.mica.org/

MICA Amphitheater

Thanks