Mining Digital Surveys … for photo-z’s and other things…

Massimo Brescia & Stefano Cavuoti
INAF – Astr. Obs. of Capodimonte
Napoli, Italy

Giuseppe Longo
Dept of Physics – Univ. Federico II
Napoli, Italy

Results from

C.E. Petrillo (now PhD in Groningen (NL)
V. De Stefano (Photoraptor)

and the DAME collaboration

Fra Luca Pacioli, inventor of Algebra
Capodimonte Museum, Napoli

The tool: DAMEWARE

web-based application (FREE AND OPEN TO THE PUBLIC) for massive data mining based on a suite of machine learning methods on top of a virtualized hybrid computing infrastructure

A joint effort between University Federico II, INAF-OACN & Caltech

http://dame.dsf.unina.it/

- Science and management
- Technical documents
- Template science cases
- Newsletters
- Tutorials

Released in 2013
Brescia et al., 2014 PASP august issue

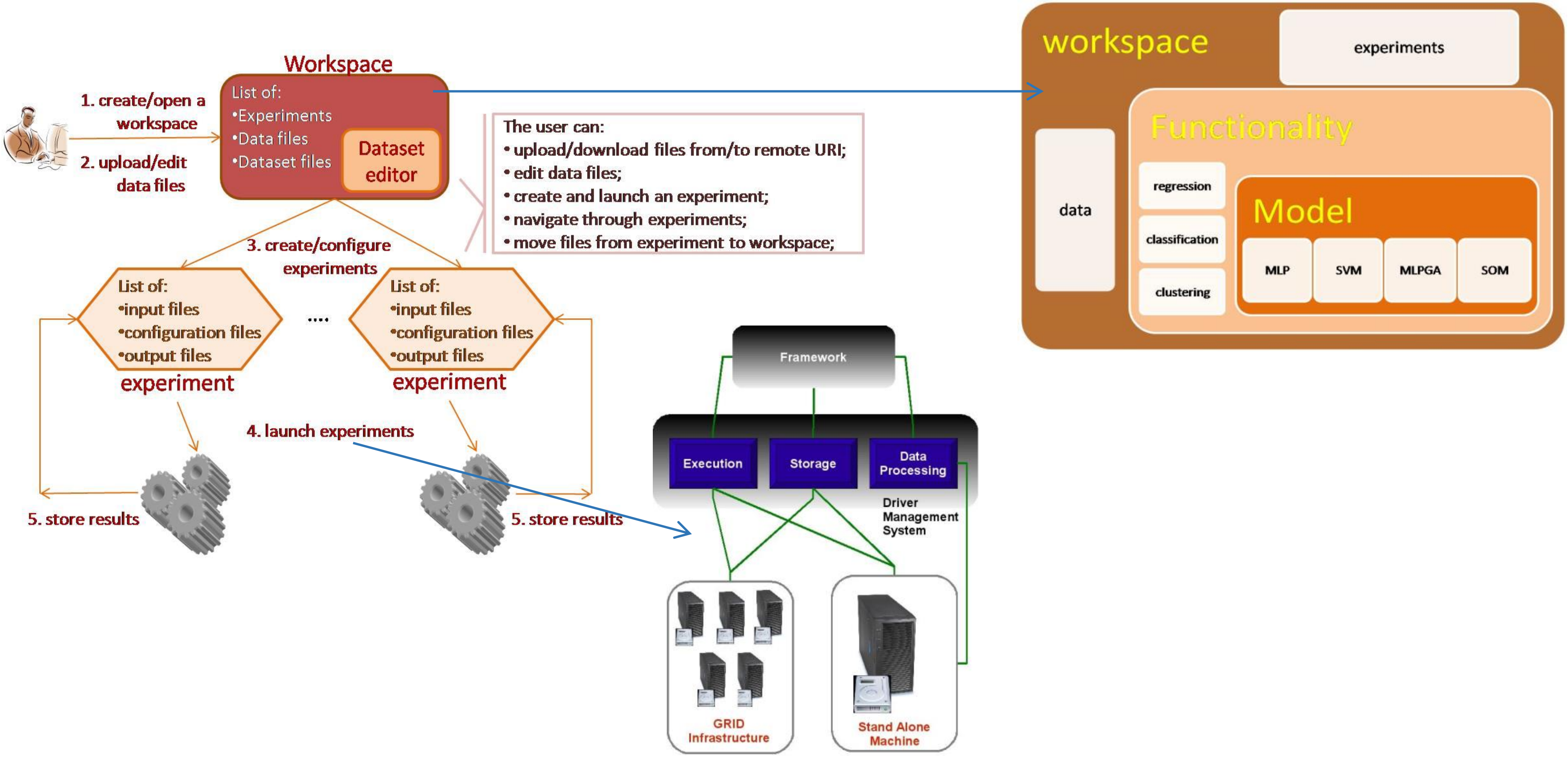
In ca. 18 months 100 groups from 27 countries
Ca. 11.000 independent accesses

DAMEWARE

It is multi-disciplinary (astronomy, geophysics, bioinformatics and medical diagnostics)

End users can remotely exploit high computing and storage power to process massive datasets (in principle they can do data mining on their smartphone…)

User can automatically plug-in his/her own algorithm and launch experiments through the Suite via a simple web browser



Effective DM **REQUIRES** complex work-flows

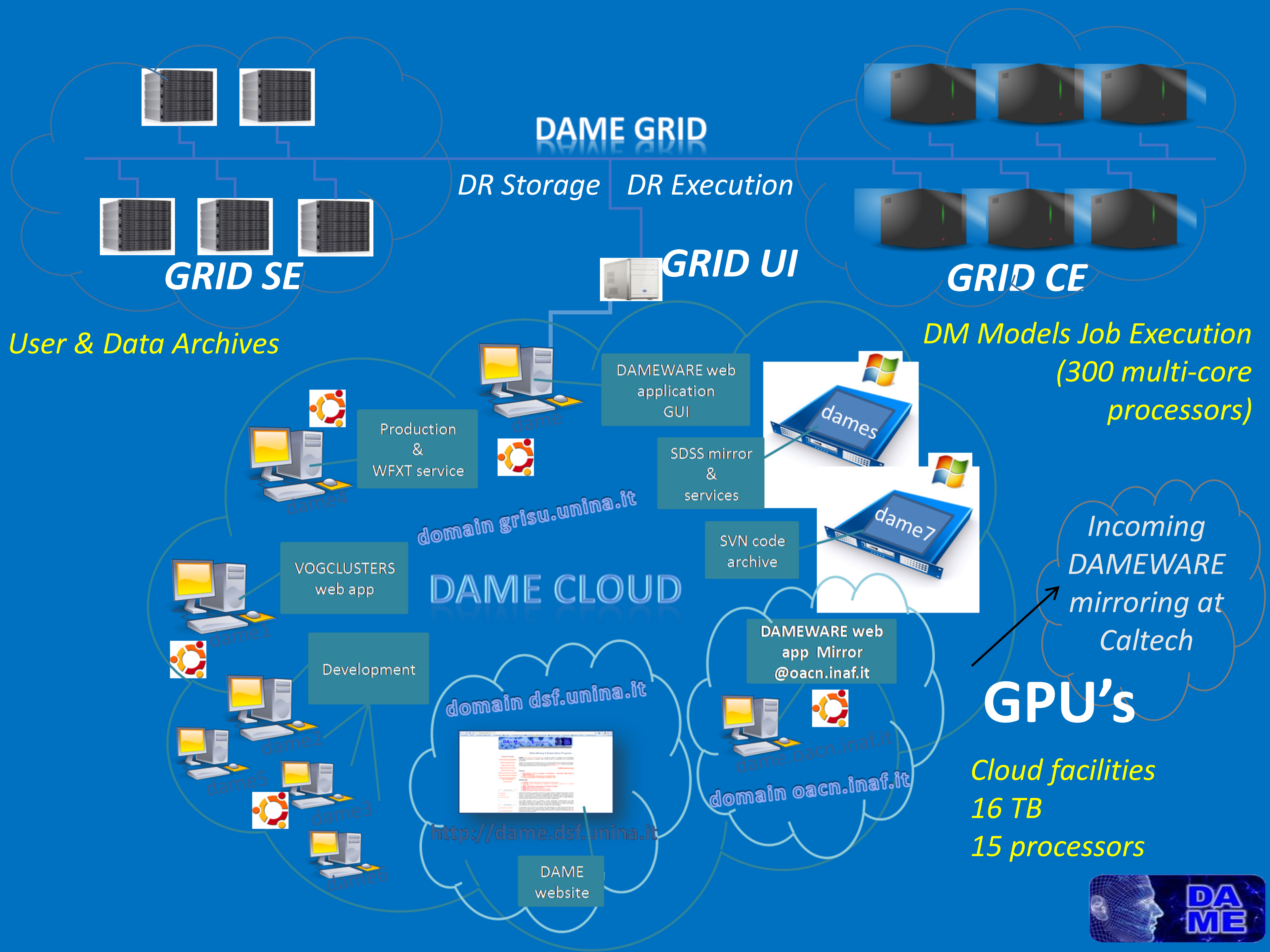
Use case
→ pre-processing
→ feature selection
→ choice of DM model
→ experiments
→ evaluation of results

The logic behind DAMEWARE

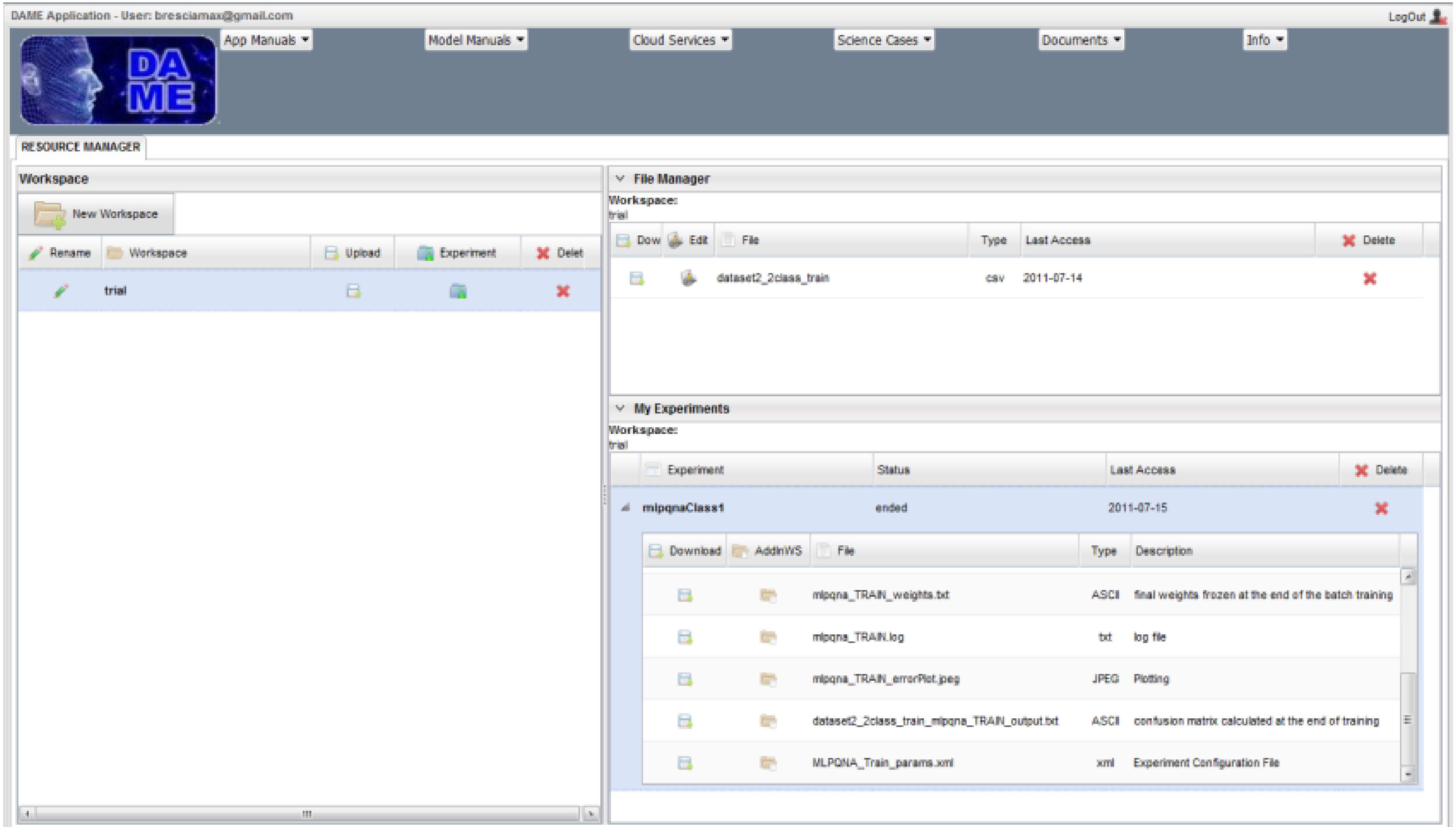
Use case

Functionality	**DM models**	**Experiments**
Classification	GAME S, C,R	1-st
Regression	MLPBP S, C,R	2-nd
Clustering	MLPGA S, C,R	3-rd
Feature selection	MLPQNA S, C,R	4-th
	SVM S, C,R	….
	K-Means U, Cl	N-th
	ESOM U, Cl	
	SOFM U, Cl	
	SOM U, Cl	
	PPS U, Cl, FS	

And … N is very large


DAME GRID
DR Storage
DR Execution
GRID SE
GRID UI
GRID CE
User & Data Archives
DM Models Job Execution
(300 multi-core
processors)
DAMEWARE web
application
GUI
Production
&
WFXT service
SDSS mirror
&
services
dames
dame7
SVN code
archive
domain grisu.unina.it
VOGCLUSTERS
web app
DAME CLOUD
Incoming
DAMEWARE
mirroring at
Caltech
Development
DAMEWARE web
app Mirror
@oacn.inaf.it
domain dsf.unina.it
GPU's
domain oacn.inaf.it
http://dame.dsf.unina.it
Cloud facilities
16 TB
15 processors
DAME
website
DA
ME

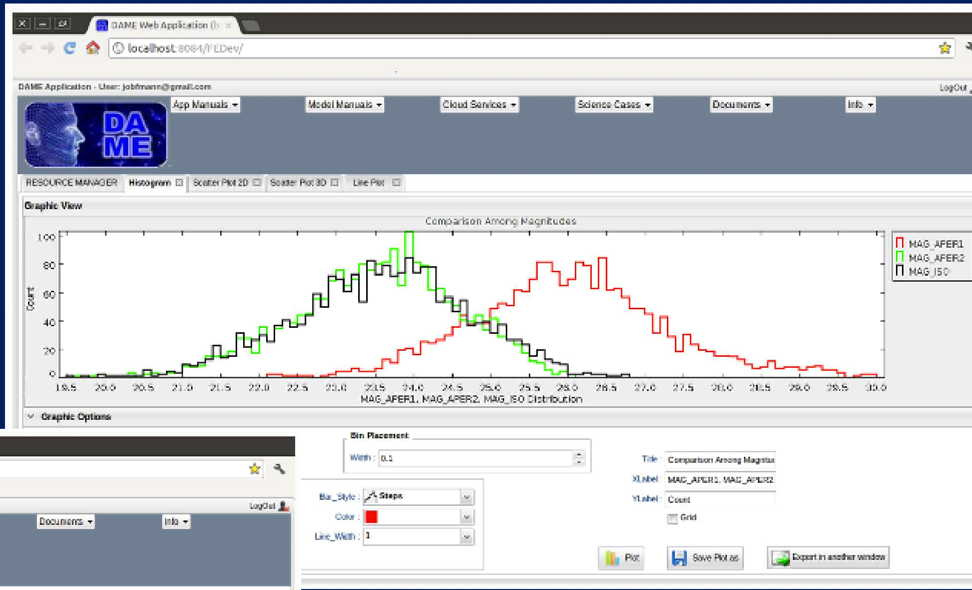
DAMEWARE - the GUI

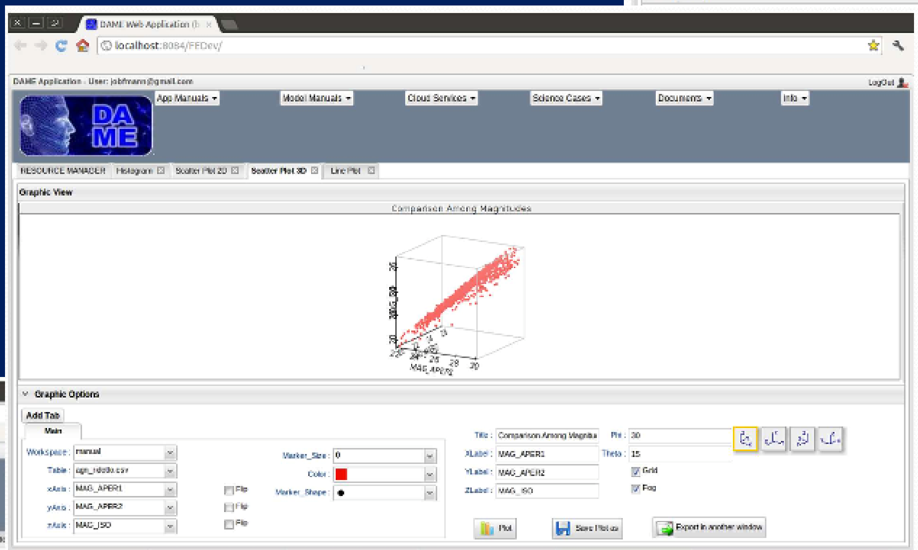


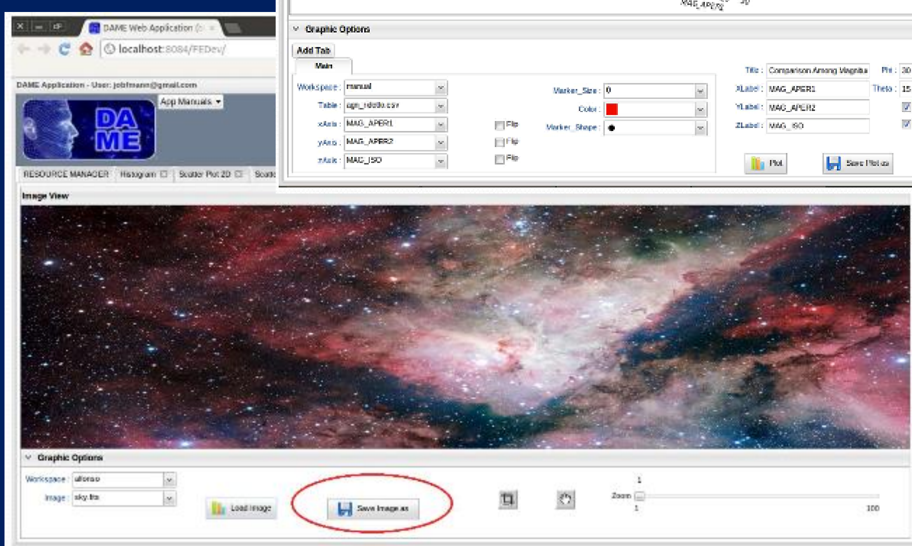
Graphical capabilities in DAMEWARE

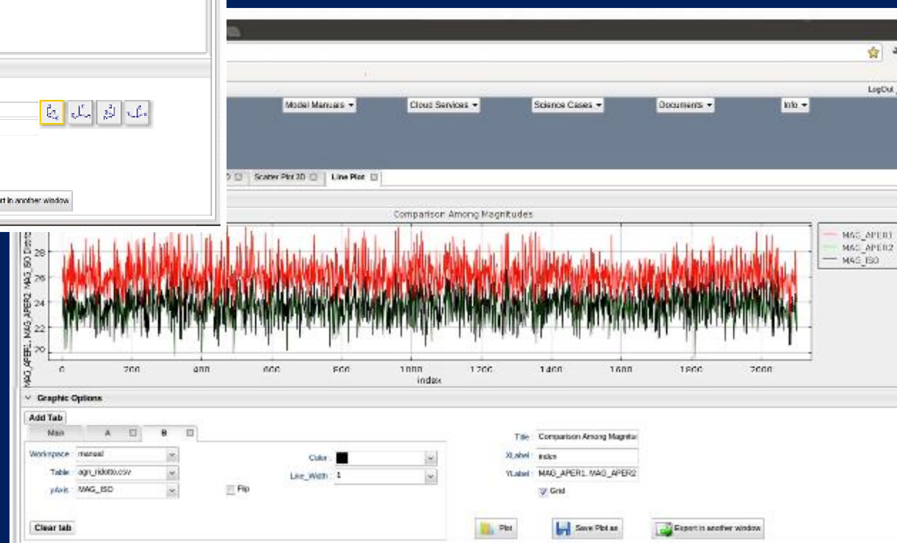
Histograms
2-D & 3-D plots
Line plots
Image visualization

Java client









AGN identification and classification

Photometric classification of emission line galaxies with Machine Learning methods, Cavuoti et al., 2014, MNRAS

Star/Galaxy separation

The detection of globular clusters as a data mining problem, Brescia et al., 2012, MNRAS, 421, 1155-1165 (arXiv:1110.2144)

GPUs for astrophysical data mining. A test on the search for candidate globular clusters in external galaxies. S. Cavuoti, et al., New Astronomy, april 20, 2013, http://dx.doi.org/10.1016/j.newast.2013.04.004 (astro-ph: 1304.0597)

Photometric redshifts

Mining the SDSS archive. I. Photometric redshifts in the nearby universe, D'Abrusco, Logno G., Walton N., 2007, ApJ, 663, 752

Astroinformatics of galaxies and quasars: a new general method for photometric redshifts estimation , O. Laurino, R. D'Abrusco, G. Longo, and G. Riccio, MNRAS, 2011, 418, 2165 (arXiv/1107.3160);

Photometric redshifts with Quasi Newton Algorithm (MLPQNA) Results in the PHAT1 context, Cavuoti et al. 2012, , Astronomy and Astrophysics 546, 13, (ArXiv:1206.0876)

Photometric redshifts for quasars in multiband surveys, M. Brescia et al., 2013, ApJ, 772, 140 (astro-ph: 1305.5641)

Inside catalogs: a comparison of source extraction software, M. Annunziatella, et al., 2012, PASP, 125, 68 (astro-ph:1212.0564).

Other

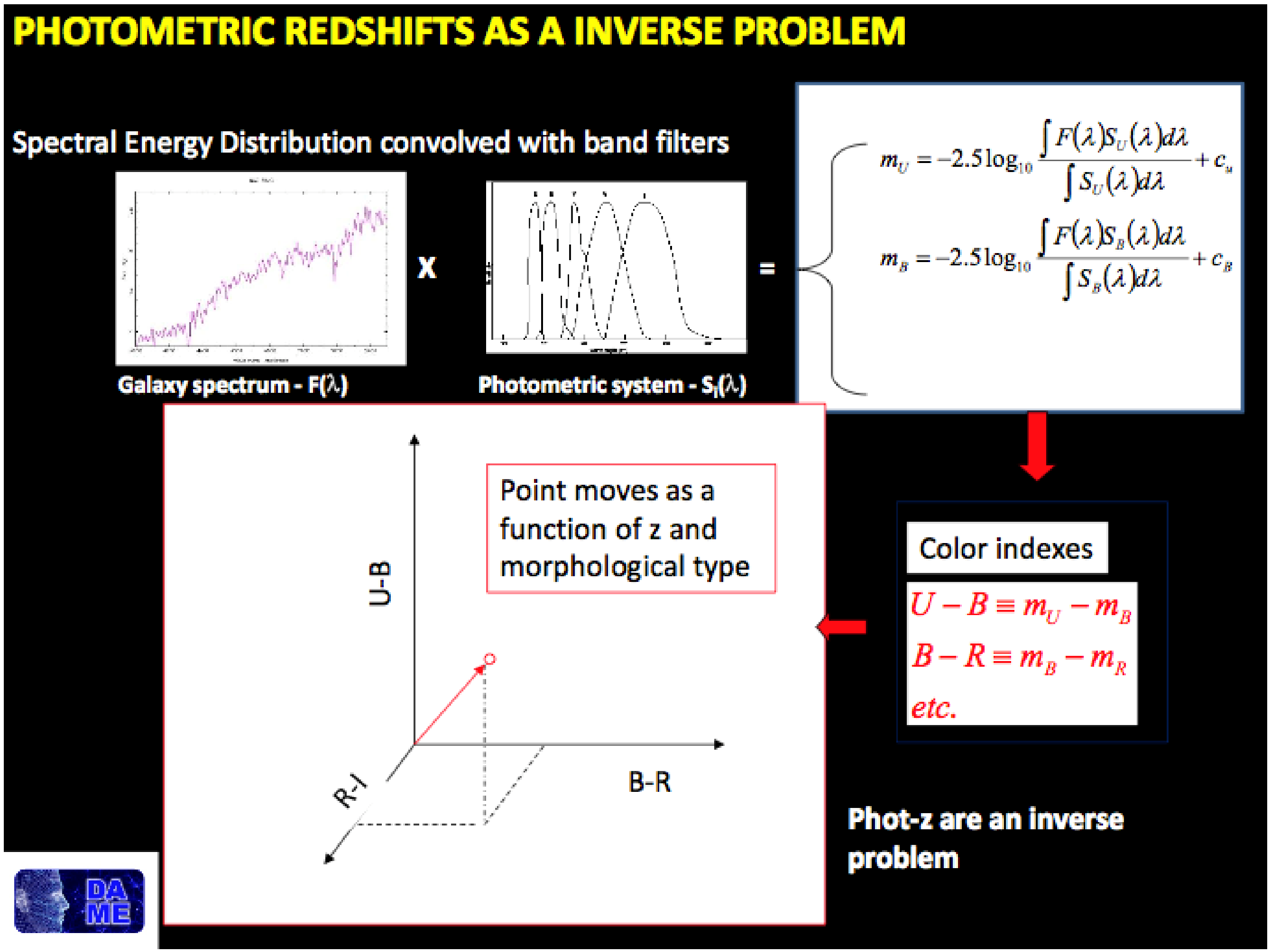
Astroinformatics, data mining and the future of astronomical research, M. Brescia & G. Longo, 2012, invited to appear in proceed. of IFDT2 - 2nd International conference frontiers on diagnostic technologies (arXiv:1201.1867)

CLASPS: a new methodology for knowledge extraction from complex astronomical data sets, R. D'Abrusco, G. Fabbiano, S.G. Djorgovski, C. Donalek, O. Laurino & G. Longo, 2012, ApJ, 755, 92 (ArXiv:1206.2919)

Distances to galaxies usually derived through Hubble's law, hence via redshifts

Spectroscopic measure $$z \equiv \frac{\Delta\lambda}{\lambda} = \frac{v}{c}$$ Accurate Δz ca. 10^{-3} or better

Photometric indirect estimate of z



Less accurate (Δz 10^{-2})

Why are photo-z so crucial….

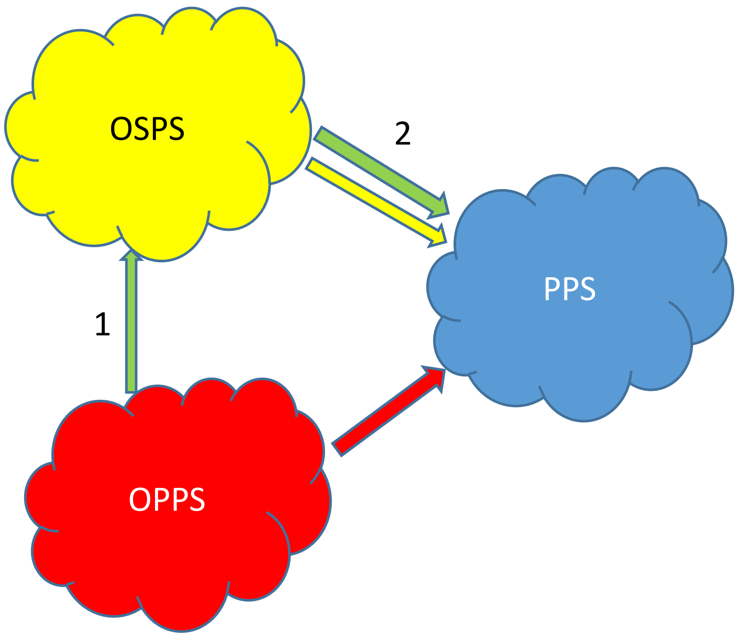
- **Larger and deeper samples with respect to spectroscopic z's**
    - Modern digital surveys produce high accuracy photometric data for hundreds of millions/billions of galaxies
    - Best spectroscopic surveys are bound to sample at most 10^6 galaxies
    - All current and future digital surveys (e.g. DES, VST VOICE, VST KIDS, Euclid, LSST, etc) require accurate photo-z's to achieve their scientific goals

- **Weak lensing (hence mass distribution, DM and DE estimates) requires accurate photo-z's for huge samples**
    - Strong requirements on bias and on controlling the selection effects

- **Large scale structure, galaxy formation and evolution, etc… strongly benefit from them**

In digital photometric surveys most data mining consists in finding the proper mapping functions between 3 hyperspaces....

Observed Spectroscopic Parameter Space - OSPS
defined by the observed spectroscopic properties
(e.g. Continuum gradients, equivalent widths for emission and absorption lines, absolute fluxes .)



Physical Parameter Space - PPS
defined by the physical properties
(e.g. distance, mass, average chemical composition, presence or absence of an AGN, etc.)

Observed Photometric Parameter Space - OSPS
defined by the observed photometric properties
(e.g. fluxes integrated over broad bands, colors, morphology and astrometry.)

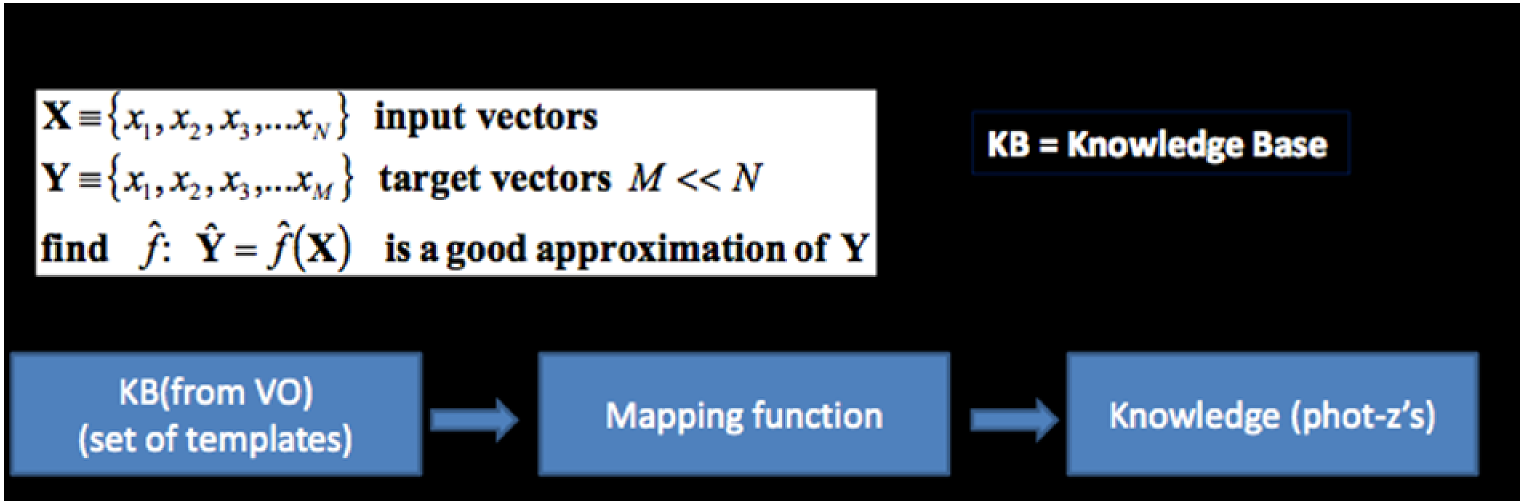
Two families of methods for photo-z's

SED template fitting

observed photometry is compared against a library of template energy distribution (either synthetic or observed) and best fit interpolation is found. Spectra are needed for zero point calibration

ML based methods (supervised learning but not only)

An extensive knowledge base of spectroscopic examples is used to teach the methods how to map OPPS into OSPS and then into PPS



ML based methods

IF

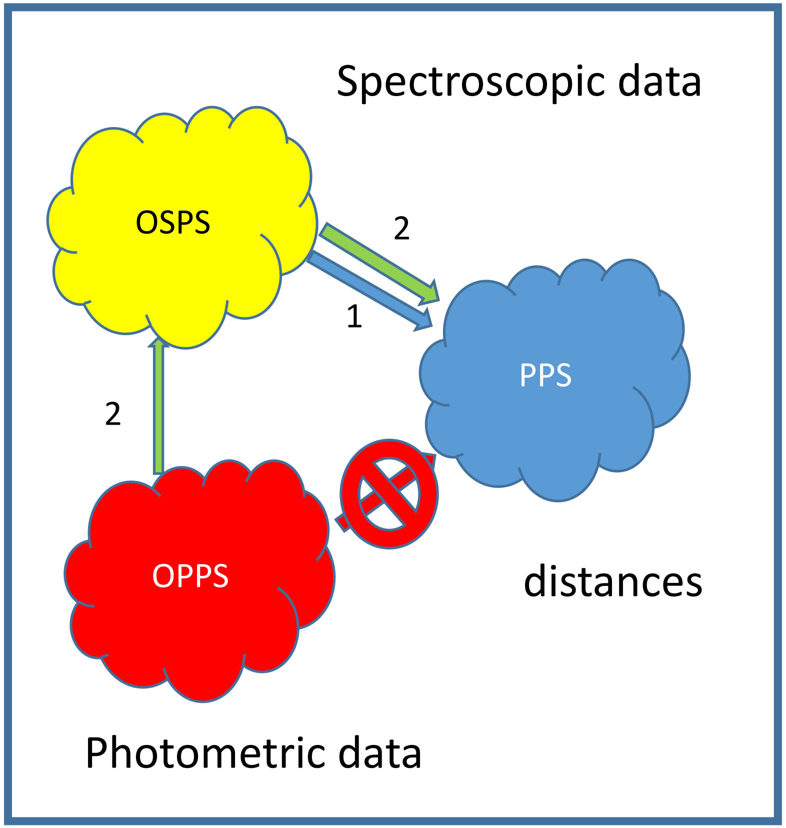
extensive KB and good coverage of
OPPS onto OSPS & **OSPS onto PPS**
are provided …

THEN

ML methods outperform SED fitting

ELSE

SED fitting methods are better



The OSPS KB needs to properly sample BOTH OPPS and PPS !!!!

An example

Use case: Photometric redshifts evaluation for quasars

Functionality: regression

Pre-processing: preparation of KB (10^5 objects), removal of NaN, splitting of train, validation, test sets

Feature selection (>50 experiments)

Selection of best DM model: SVM; MLPBP, MLP-GA, GAME, MLPQNA

Training, Validation, Test

Visualization, comparison & Evaluation of results

& Hundreds of runs are needed to evaluate pdf, characterize data, etc.

Photometric redshifts for SDSS Quasar candidates:

Brescia et al, 2013, ApJ, 772, 140

Survey	Bands	Name of feature	Synthetic description
SDSS	u, g, r, i, z	psfMag_u, psfMag_g, psfMag_r, psfMag_i, psfMag_ z	PSF fitting magnitude in the u g, r, i, z bands.
UKIDSS	Y,J,H,K	yPsfMag, j_1PsfMag, hPsfMag, kPsfMag	PSF fitting magnitude in Y, J, H, K bands
	Y,J, H,K	Es. for y band: yAperMag3, yAperMag4, yAperMag6	aperture photometry through 2, 2.8 & 5.7″ circular aperture
	J,H,K	Es. for J band: jHallMag, JPetroMag	Calibrated magnitude within circular aperture r_hall and Petrosian magnitude
GALEX	NUV	Nuv_mag, Nuv_mag_iso;	Respectively: Near UV total and isop. mags
	NUV	Nuv_mag_Aper_1 Nuv_mag_Aper_2 Nuv_mag_Aper_3	aperture photometry through 2,3 & 5 pxl apertures
	NUV	Nuv_mag_auto and Nuv_kron_radius	magnitudes and Kron radius in units of A or B
	FUV	Fuv_mag, Fuv_mag_iso;	Respectively: Far UV total and isop. mags
	FUV	Fuv_mag_Aper_1 Fuv_mag_Aper_2 Fuv_mag_Aper_3	aperture photometry through 2,3 & 5 pxl apertures
	FUV	Fuv_mag_auto and Fuv_kron_radius	magnitudes and Kron radius in units of A or B
WISE	W1, W2, W3, W4	W1mpro, W2mpro, Wmpro, Wmpro4	W1: 3.4 μm and 6.1″ angular resolution, W2: 4.6 μm and 6.4″ angular resolution. W3 12 μm and 6.5″ W4 22 μm and 12″ angular resolution Magnitudes measured with profile-fitting photometry at the 95% level. Brightness upper limit if the flux measurement has SNR$<$ 2
SDSS	-	zspec	Spectroscopic redshift

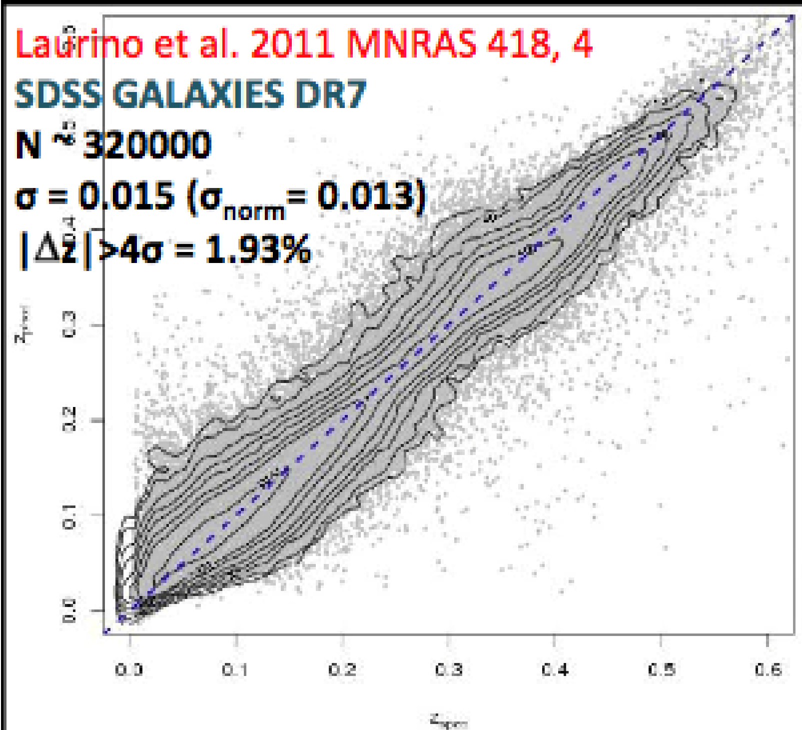
1	2	3	4	5	6	7	8	9	10	11	12
E1	X	X	X	X	All	0,0038	0,174	15,96%	4,75%	2,24%	0,92%
E2	X	X	X	X	UKIDSS: hall GALEX: mag + mag_iso	-0,0001	0,152	19,66%	4,49%	1,85%	0,92%
E3	X	X	X	X	UKIDSS: hall GALEX: Aper 1, 2 ,3	-0,0016	0,165	15,83%	3,96%	1,98%	1,19%
E4	X	X	X	X	UKIDSS: hall GALEX: mag	0,0054	0,151	16,23%	4,75%	1,98%	1,06%
E5	X	X	X	X	UKIDSS: hall GALEX: mag_iso	-0,0026	0,151	18,47%	4,62%	2,37%	0,79%
E6	X	X	X	X	UKIDSS: hall GALEX: mag_auto + kron radius	-0,0008	0,152	17,81%	5,15%	2,64%	0,79%
E7	X	X	X	X	UKIDSS: hall GALEX: mag + mag_iso + Aper 1, 2, 3	0,0041	0,163	19,39%	4,22%	2,51%	0,66%
E8	X	X	X	X	UKIDSS: hall GALEX: mag_iso + Aper 1, 2, 3	-0,0038	0,155	19,26%	5,01%	1,98%	0,92%
E9				X	All	0,0165	0,297	22,16%	5,80%	2,11%	0,53%
E10	X				All	-0,0162	0,338	19,66%	7,26%	2,37%	0,40%
E11		X			UKIDSS: hall + petro	-0,0091	0,299	23,75%	4,88%	1,58%	0,66%
E12			X		GALEX: mag + mag_iso	0,055	0,419	29,68%	4,75%	0,79%	0,26%
E13		X			UKIDSS: petro	0,0111	0,465	34,43%	3,43%	0,40%	0,00%
E14		X			UKIDSS: hall	-0,0081	0,294	22,82%	5,94%	1,85%	0,66%
E15		X		X	UKIDSS: hall	0,0045	0,296	17,94%	4,75%	2,11%	1,06%
E16	X	X	X		UKIDSS: hall GALEX: mag_iso	-0,0046	0,152	21,11%	4,88%	1,98%	0,79%
E17	X		X	X	GALEX: mag_iso	0,0025	0,162	16,23%	3,69%	2,37%	1,06%
E18	X	X		X	UKIDSS: hall	-0,0032	0,179	14,38%	4,49%	2,11%	1,32%
E19		X	X	X	UKIDSS: hall GALEX: mag_iso	0,011	0,203	19,26%	4,88%	1,72%	0,79%
E20			X	X	GALEX: mag_iso	0,0175	0,288	22,96%	4,88%	1,45%	0,53%
E21	X	X			UKIDSS: hall	-0,0027	0,21	15,96%	5,15%	2,24%	1,06%
E22	X			X	All	-0,0039	0,197	13,85%	3,43%	2,37%	1,58%
E23	X		X		GALEX: mag_iso	-0,0055	0,24	17,55%	6,73%	2,51%	0,79%
E24		X	X		UKIDSS: hall GALEX: mag_iso	0,0133	0,288	23,22%	6,20%	1,72%	0,40%

Table 2. List of the experiments performed during the pruning phase in order to evaluate the best possible combination of parameters. Column 1: identification number of the experiments. Column 2 through 5: surveys used for the experiment. The order of the surveys is SDSS, UKIDSS, GALEX and WISE. A cross in a column meaning that all the bands of the survey corresponding to that column were used for the experiment. **Credo che l'ordine delle survey andrebbe cambiato seguendo la lunghezza d'onda (WISE UKIDSS SDSS GALEX)** For bands with multiple types of magnitudes measured, Column 6 gives the type which has been used for a given experiment. Columns 7-8: rms scatter and the bias. Columns 9-12: percentage of outliers at, respectively, 1,2,3 and 4 σ. In order to be as conservative as possible, for all experiments we used 3029 objects in the training set and 758 disjoint objects as test set.

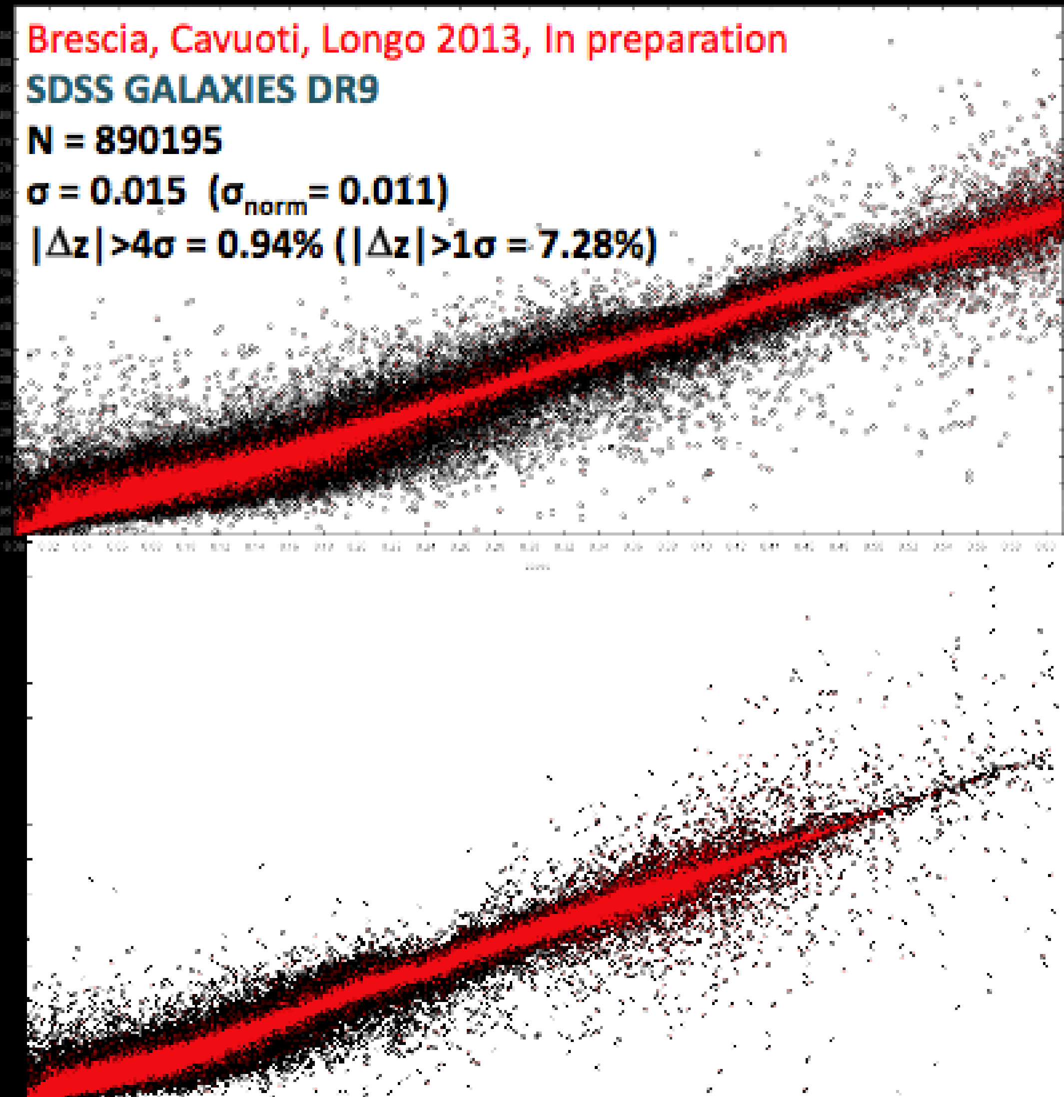
Feature selection phase

WISE substantially useless

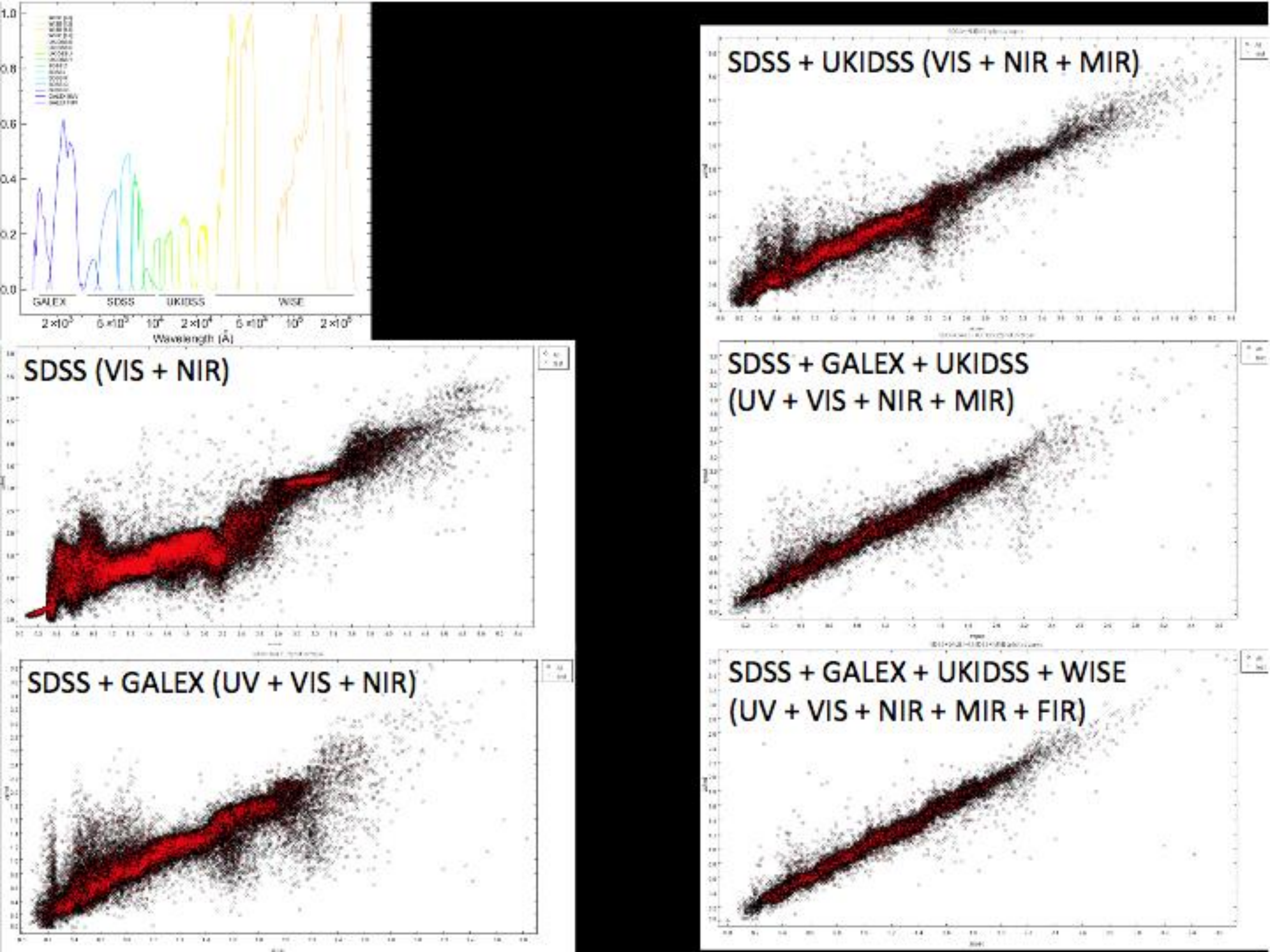
Mag_iso substantially useless



Laurino et al. 2011 MNRAS 418, 4
SDSS GALAXIES DR7
N ~ 320000
σ = 0.015 (σnorm = 0.013)
|Δz|>4σ = 1.93%
0.0
0.1
0.2
0.3
0.4
0.5
0.6



Brescia, Cavuoti, Longo 2013, In preparation
SDSS GALAXIES DR9
N = 890195
σ = 0.015 (σnorm = 0.011)
|Δz|>4σ = 0.94% (|Δz|>1σ = 7.28%)

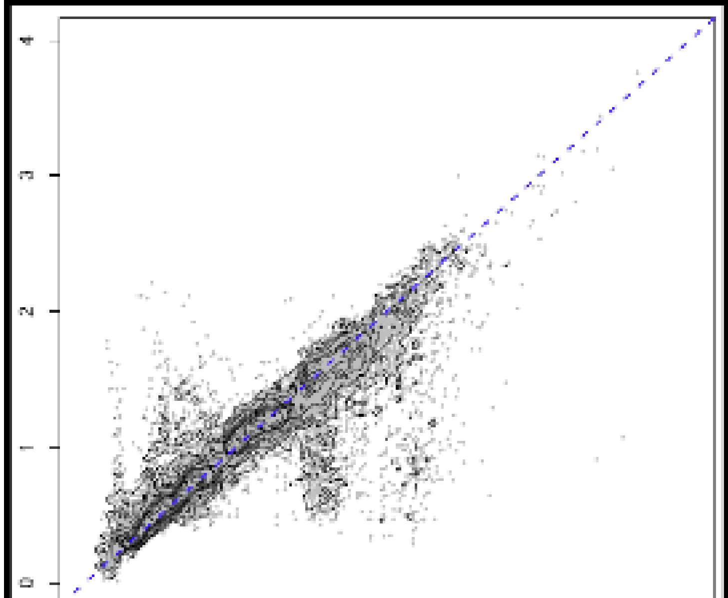


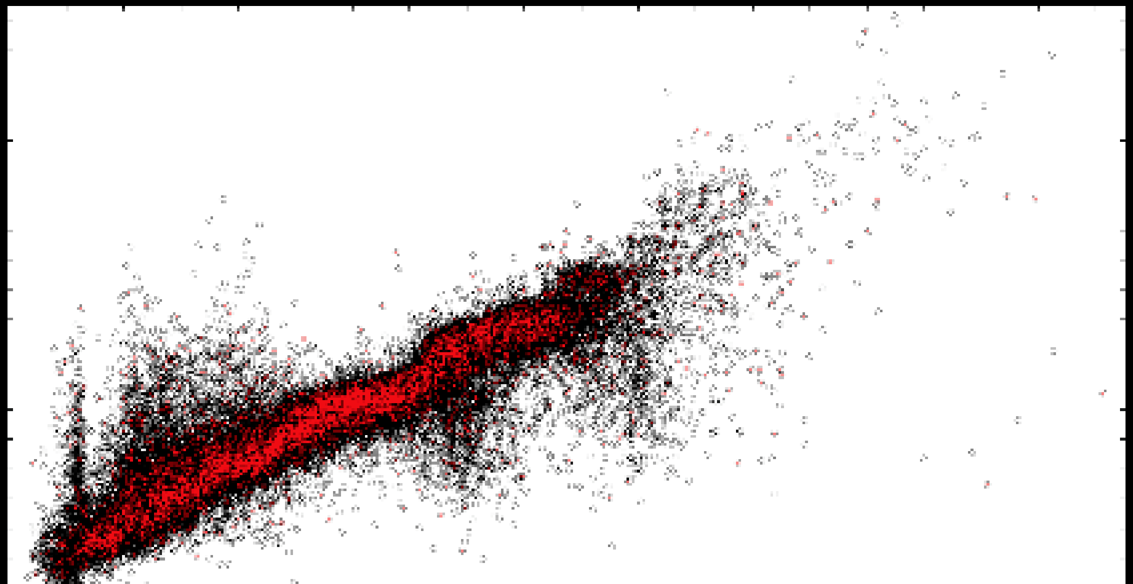
1.0
0.8
0.6
0.4
0.2
0.0
GALEX
SDSS
UKIDSS
WISE
Wavelength (Å)
SDSS + UKIDSS (VIS + NIR + MIR)
SDSS (VIS + NIR)
SDSS + GALEX + UKIDSS
(UV + VIS + NIR + MIR)
SDSS + GALEX (UV + VIS + NIR)
SDSS + GALEX + UKIDSS + WISE
(UV + VIS + NIR + MIR + FIR)

Quasar Photometric redshifts prediction from matched data (SDSS, GALEX, UKIDSS, WISE);

Laurino et al. 2011, MNRAS 418, 4
QSO SDSS+GALEX
N ~ 40000
$\sigma = 0.21$ ($\sigma_{norm} = 0.29$)
$|\Delta z| > 4\sigma = 1.93\%$ ($|\Delta z| > 1\sigma = 19.56\%$)

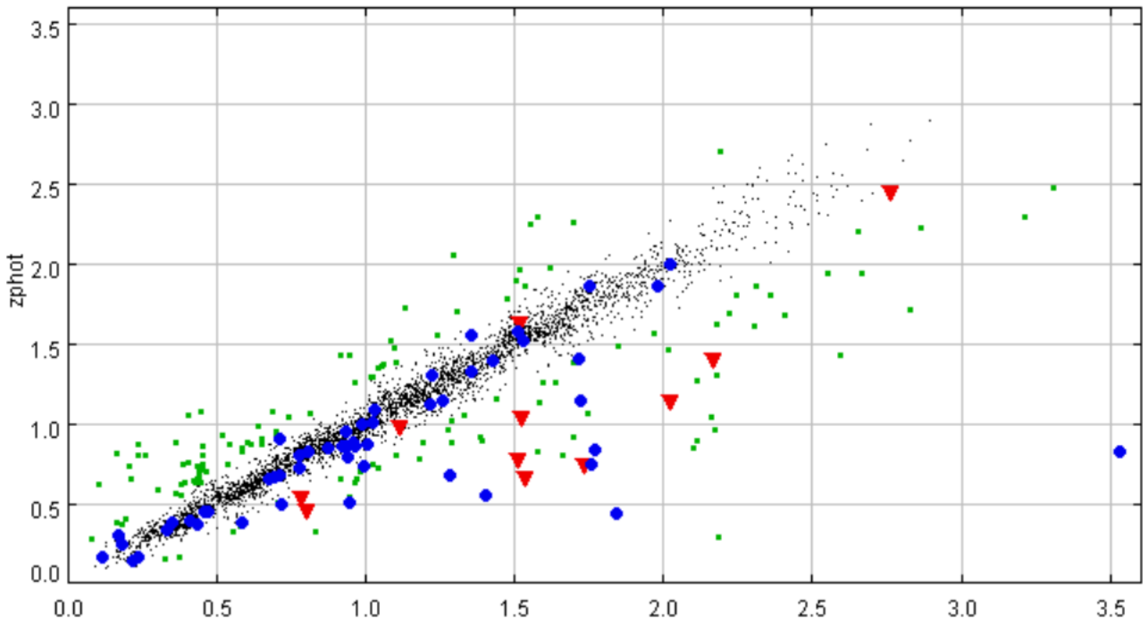
Brescia Cavuoti D'Abrusco Longo Mercurio. 2013, Subm. to MNRAS
QSO SDSS+GALEX
N = 40219
$\sigma = 0.21$ ($\sigma_{norm} = 0.14$)
$|\Delta z| > 4\sigma = 1.08\%$ ($|\Delta z| > 1\sigma = 14.97\%$)





Can we reduce bias and fraction of catastrophic outliers on photometric grounds only?

Doesnot matter how good your pipeline is…. When it comes to confuse ideas… data are always smarter than you …



All points which are not black dots are catastrophic outliers

«Astroinformatician»

By D. Vinkovic

Individual inspection needed …
and even in SDSS DR-9

In spite of galaxy zoo, and thousands of people looking at the data and using the catalogues….

SDSS still contains lots of artifacts !!!!

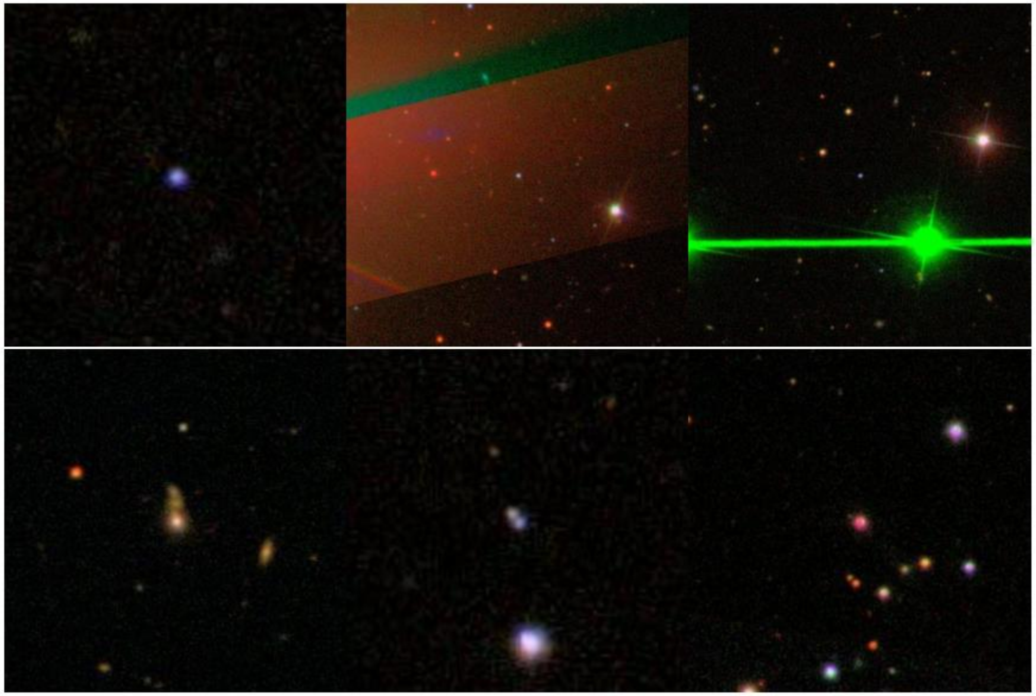
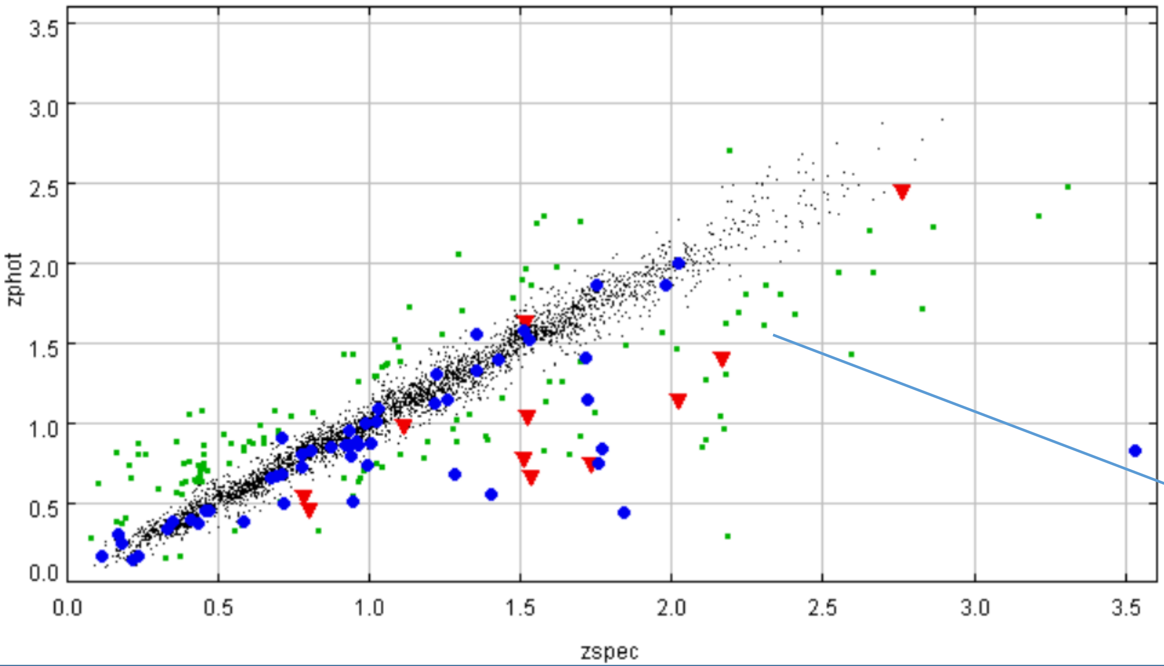


Figure 5.3: A compilation of "strange" objects. From the upper left corner in clockwise direction: A normal quasar; a disturbed image; an artifact; an elongated object; a pair object; a source with an unusual photometry.

Flag	Test-set %	Outliers %
PSF_FLUX_INTERP	8%	21%
INTERP_CENTER	10%	29%
DEBLEND_NOPEAK	0%	3%
science_primary=0	11%	24%
nuv_flags	11%	18%
fuv_artifact	18%	24%

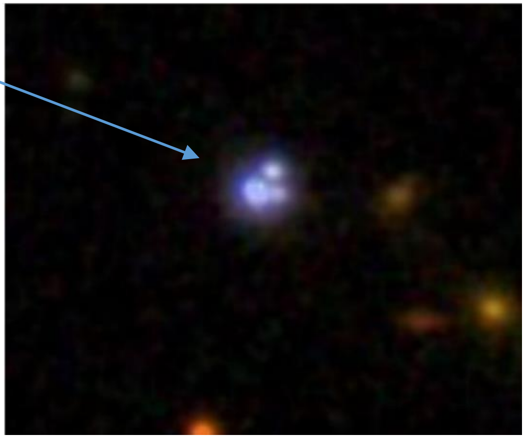
Table 5.5: Percentage of certain quality flags among the entire test-set and the outliers.

Some flags may prove useful even though not decisive



QSO's in DR9
(spectroscopic sample)

Application to
D'Abrusco 's candidates list



Green dots: Cat. Outliers
Blue circles: known blazars
Red triangles: gravitationally lensed quasars

Only 8 outliers are found to be variable in CRTS (out of 600 in our sample)

In contrast with what found by Salvato et al.

Keep the same training set… **run 50 training … get 50 frozen networks and produce 50 photo-z** estimates for all objects in the test set and then check what happens

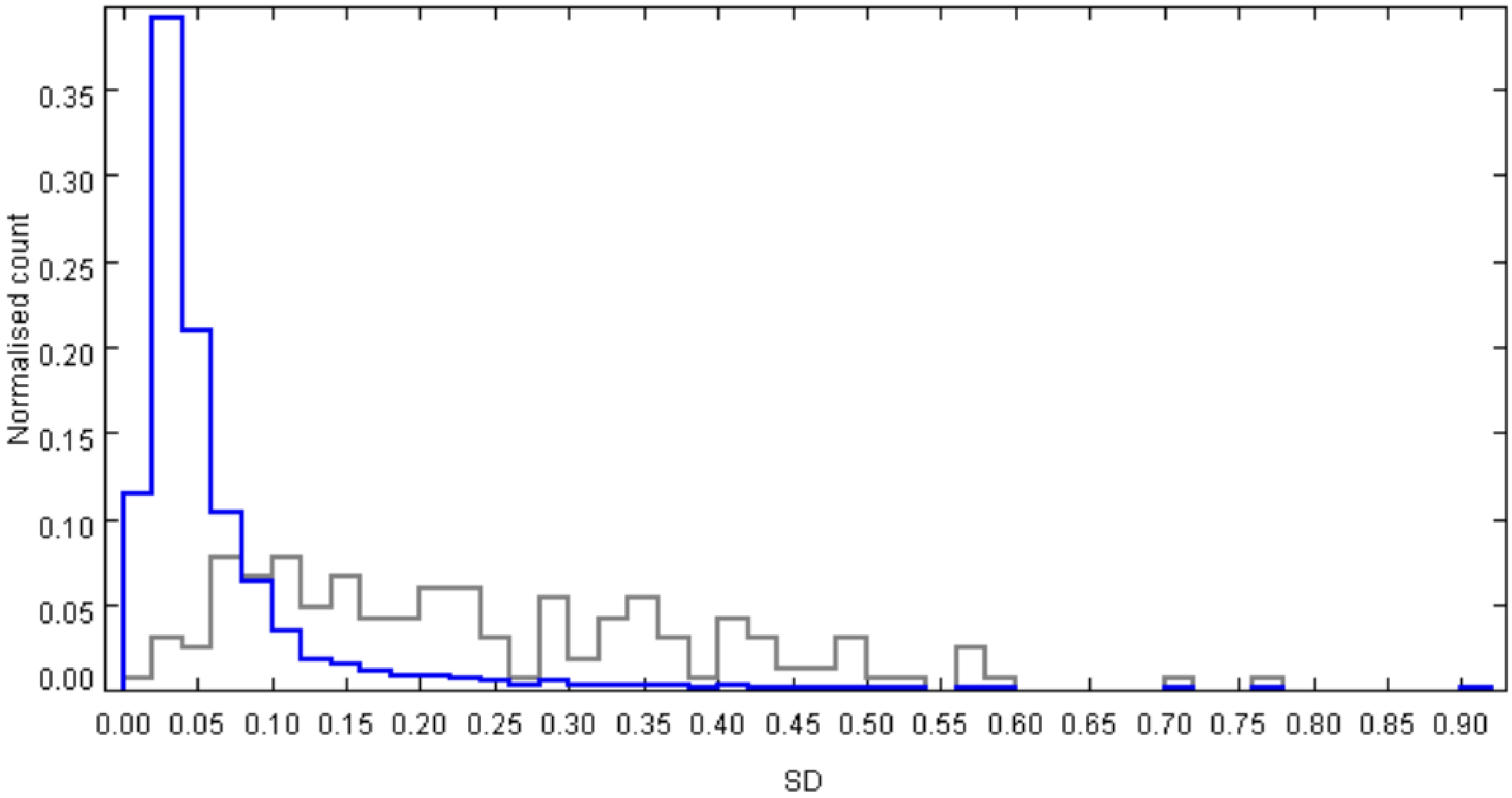


Figure 5.8: Normalized distribution of the standard deviations of the entire test-set (blue line) and of the outliers only (grey line).

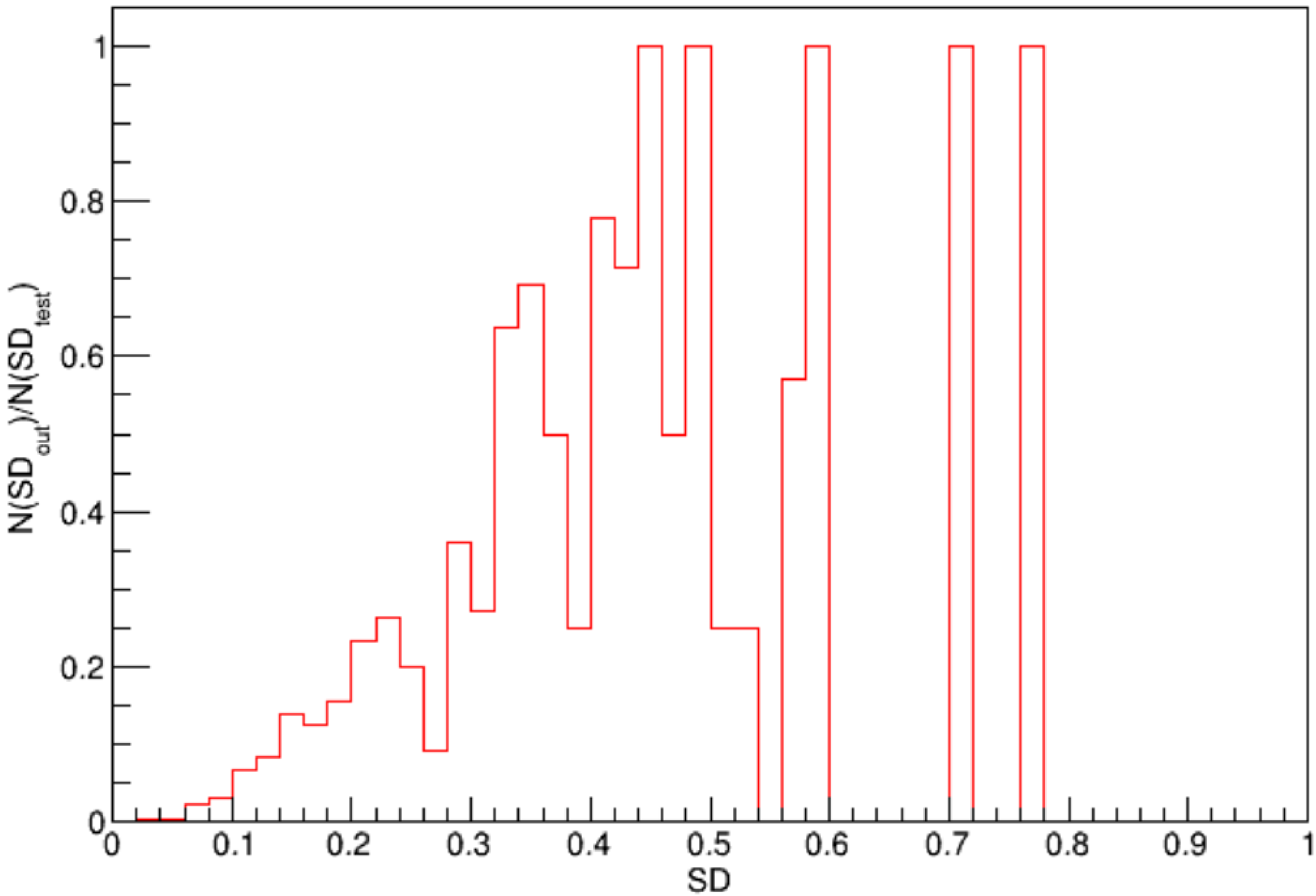


Figure 5.9: Distribution of the ratios between the value of the standard deviation of the outliers and the standard deviation of the entire test-set.

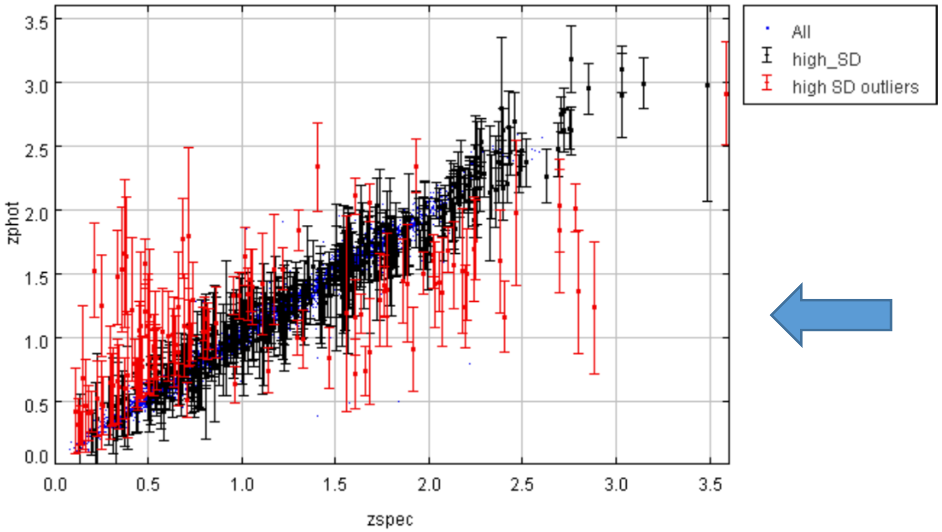


Figure 5.10: Scatter plot (z_{spec} vs. z_{phot}) for the mean experiment. Sources with a standard deviation greater than 0.125 are marked in black or in red (if are outliers). Blue dots indicate the entire test-set.

This can be eliminated on photometric grounds only using flags and SD

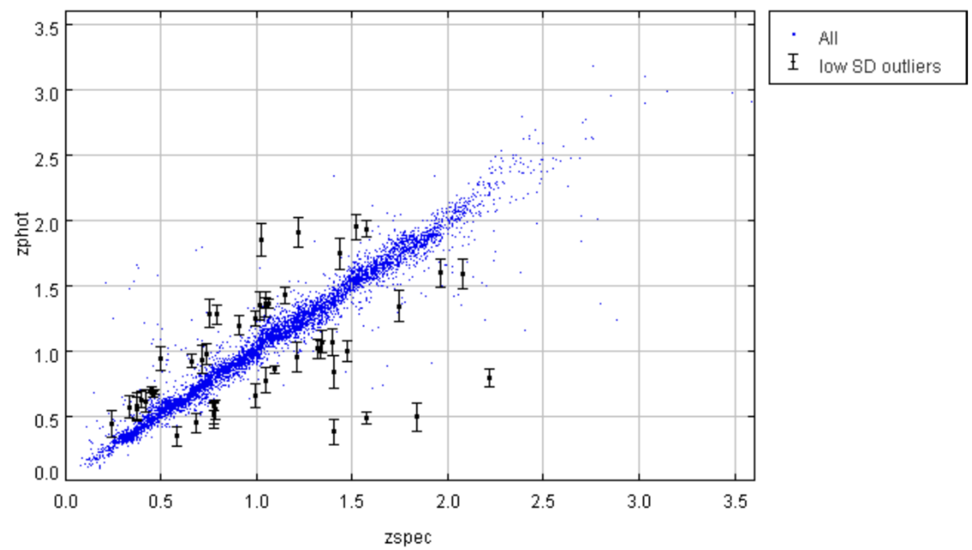
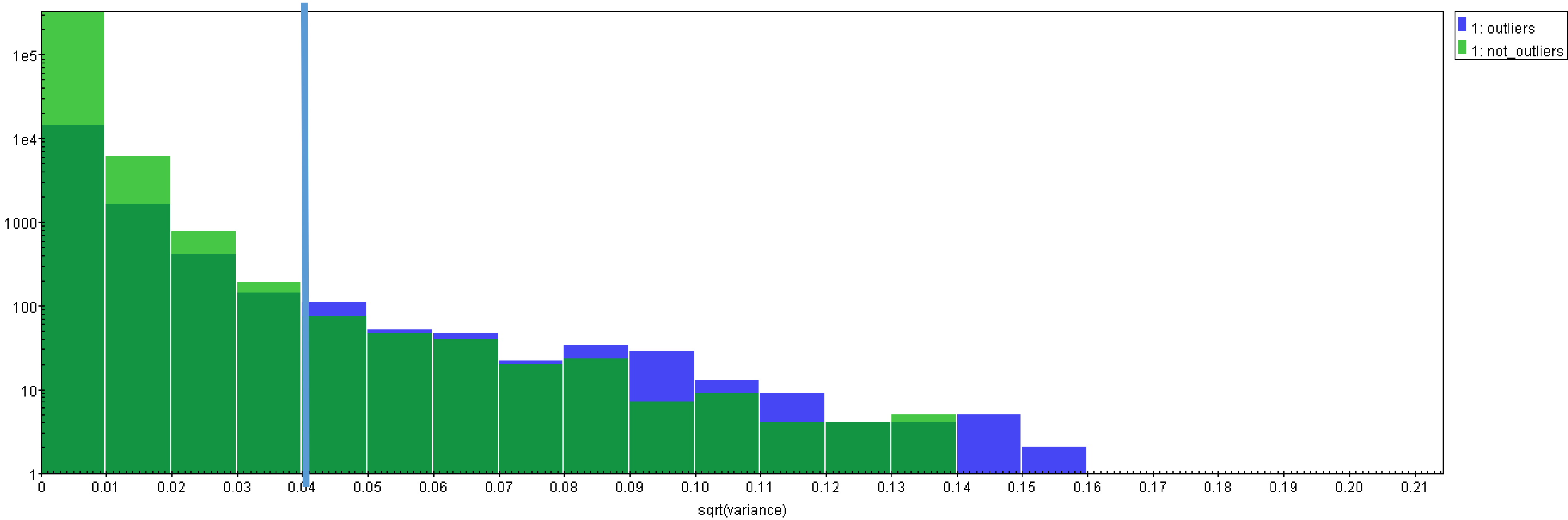


Figure 5.11: Scatter plot (z_{spec} vs. z_{phot}) for the mean experiment. Outliers with a standard deviation less than 0.125 are marked in black. Blue dots indicate the entire test-set.

	Homogenous	*Average*	*Average low SD*
Dataset	14284	14284	(14284 - 367)
$\mathbf{BIAS}(\boldsymbol{\Delta}\mathbf{z})$	0.002	0.0001	0.0007
$\sigma(\boldsymbol{\Delta}\mathbf{z})$	0.14	0.12	0.077
$\mathbf{MAD}(\boldsymbol{\Delta}\mathbf{z})$	0.043	0.036	0.034
$\mathbf{RMS}(\boldsymbol{\Delta}\mathbf{z})$	0.14	0.12	0.077
$\mathbf{NMAD}(\boldsymbol{\Delta}\mathbf{z})$	0.063	0.054	0.050
$> \mathbf{2}\sigma(\boldsymbol{\Delta}\mathbf{z})$	2.94%	3.17%	3.67%
$> \mathbf{4}\sigma(\boldsymbol{\Delta}\mathbf{z})$	1.14%	0.10%	0.40%
$\mathbf{BIAS}(\boldsymbol{\Delta}\mathbf{z}_{\mathbf{norm}})$	0.003	0.003	0.0005
$\sigma(\boldsymbol{\Delta}\mathbf{z}_{\mathbf{norm}})$	0.70	0.059	0.037
$\mathbf{MAD}(\boldsymbol{\Delta}\mathbf{z}_{\mathbf{norm}})$	0.021	0.018	0.017
$\mathbf{RMS}(\boldsymbol{\Delta}\mathbf{z}_{\mathbf{norm}})$	0.070	0.060	0.037
$\mathbf{NMAD}(\boldsymbol{\Delta}\mathbf{z}_{\mathbf{norm}})$	0.031	0.027	0.025
$> \mathbf{2}\sigma(\boldsymbol{\Delta}\mathbf{z}_{\mathbf{norm}})$	2.66%	2.98%	3.87%
$> \mathbf{4}\sigma(\boldsymbol{\Delta}\mathbf{z}_{\mathbf{norm}})$	0.84%	0.89%	0.57%

Table 5.6: The number of sources in the dataset, the statistical indicators and percentages of catastrophic outliers calculated on $\Delta z = (z_{spec} - z_{phot})$ and $z = (z_{spec} - z_{phot})$ for the average experiment (second column), and the homogeneous experiment for comparison (first column). The recomputed quantities after the removal of 367 objects with an high standard deviation (third column).



Results on SDSS DR 10 data galaxy data

16991 objects (KB)
sigma norm: 0.0232
bias 0.000699

Removing (0.36% of the objects)
sigma norm: 0.021 (3.34% impr.)
bias: 0.000639 (5.29% impr.)

Download the app at the DAME web site

PhotoRaptor

Pre-Processing
(on spectroscopic KB)
Feature Selection

Split

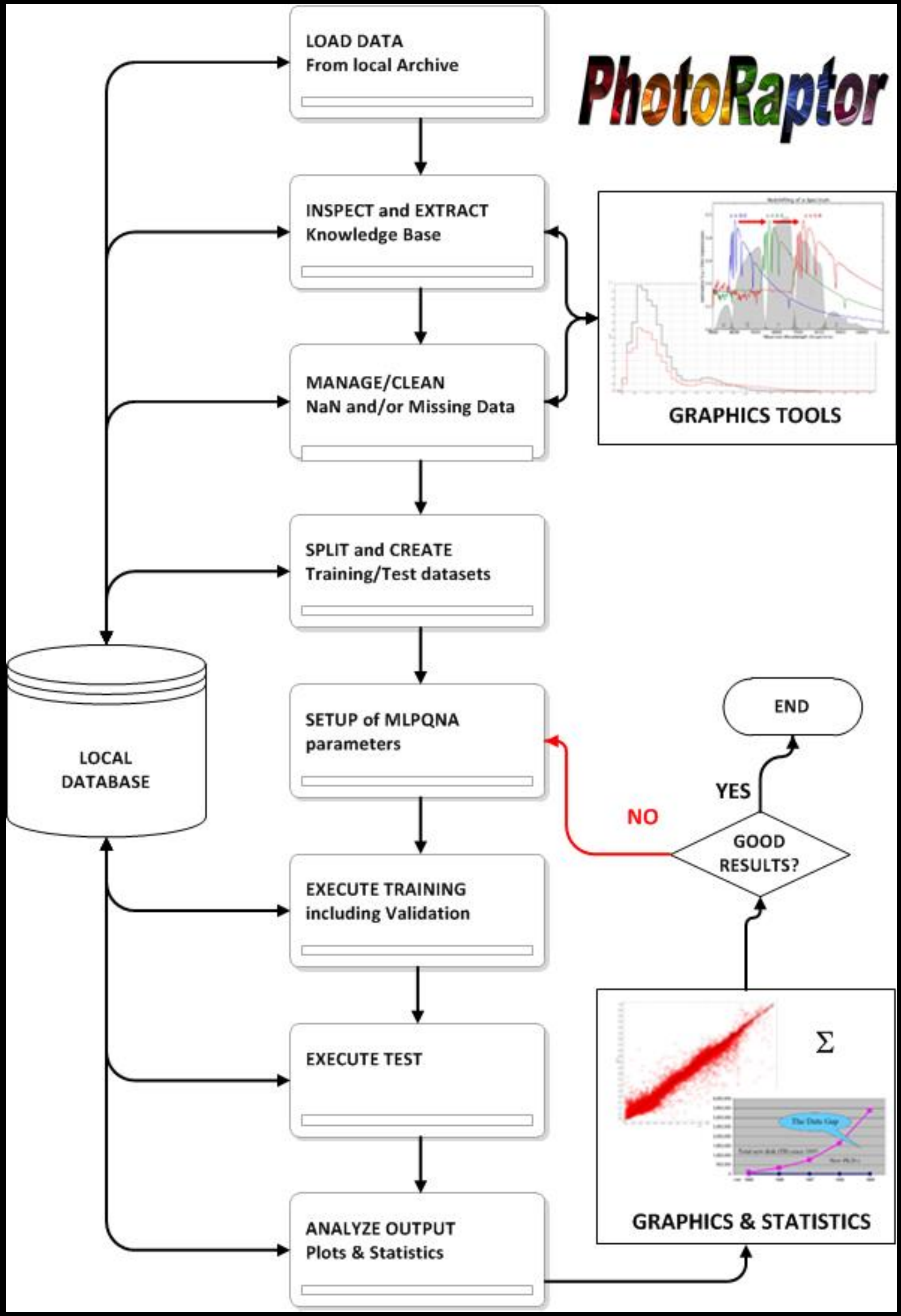
Training Set

Test dataset

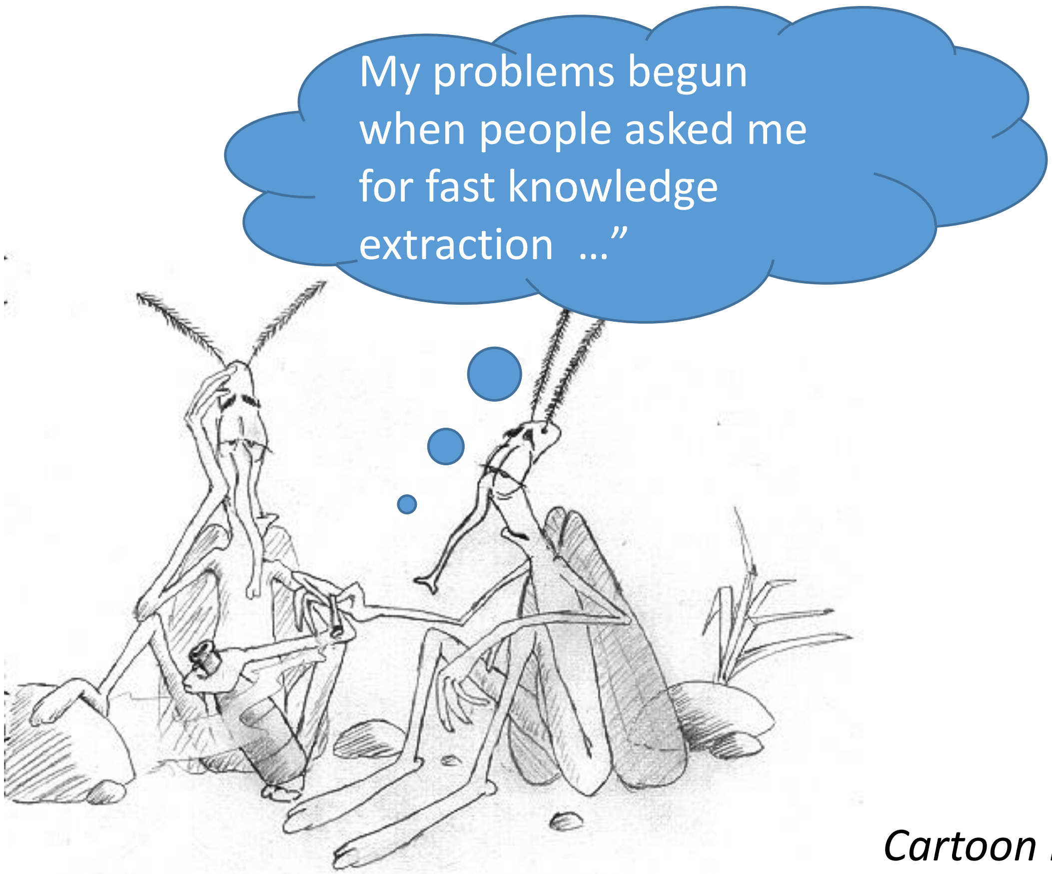
Processing (MLPQNA)

Post-processing

Statistical Analysis
and Generalization



To be continued ….



Cartoon by D. Vinkovic