Surfing the Digital Universe

Stefano Cavuoti

Department of Physics – University Federico II – Napoli

INAF – Capodimonte Astronomical Observatory – Napoli

Supervisors:

Giuseppe Longo

Department of Physics – University Federico II – Napoli

Massimo Brescia

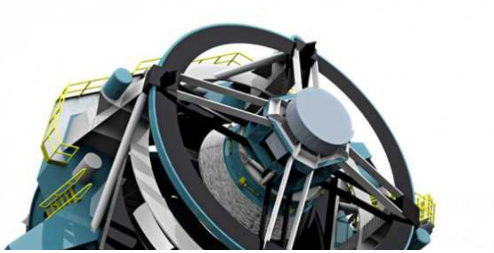
INAF – Capodimonte Astronomical Observatory – Napoli

Astroinformatics: a new era for Astronomy?

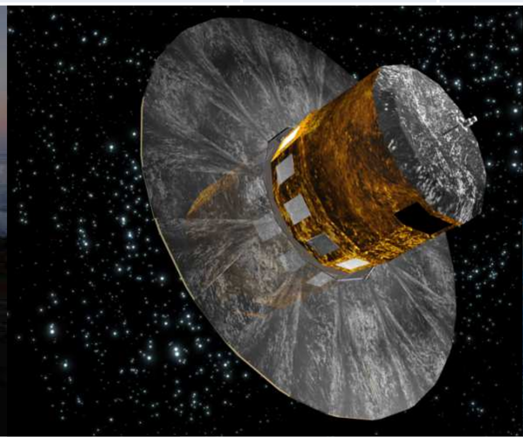
You take the **Blue Pill**,
The story ends. You wake up in your bed and believe whatever you want to believe.
You take the **Red Pill**,
You stay in Wonderland and I show You how deep the rabbit hole goes

I'm only offering You the **TRUTH**...
Nothing more.

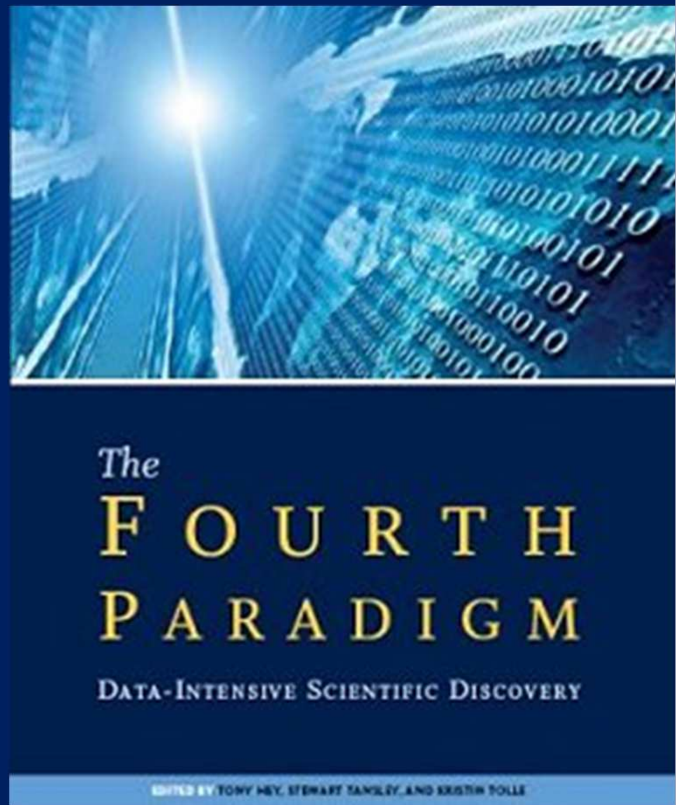
Data quantity and complexity



	TB	Total	epochs	parameters
VST	0.15 TB/day	100 TB	tens	>100
HST		120 TB	few	>100
PANSTARRS		600 TB	Few-many	>>100
LSST	30 TB/day	> 10 PB	hundreds	>>100
GAIA		1 PB	many	>>100 heterogeneous
SKA	1.5 PB/day		>> 10^2	hundreds



"One of the greatest challenges for 21st-century science is ***how we respond to this new era of data intensive science***.
This is recognized as a new paradigm beyond experimental and theoretical research and computer simulations of natural phenomena—one that requires new tools, techniques, and ways of working." — **Douglas Kell**, University of Manchester

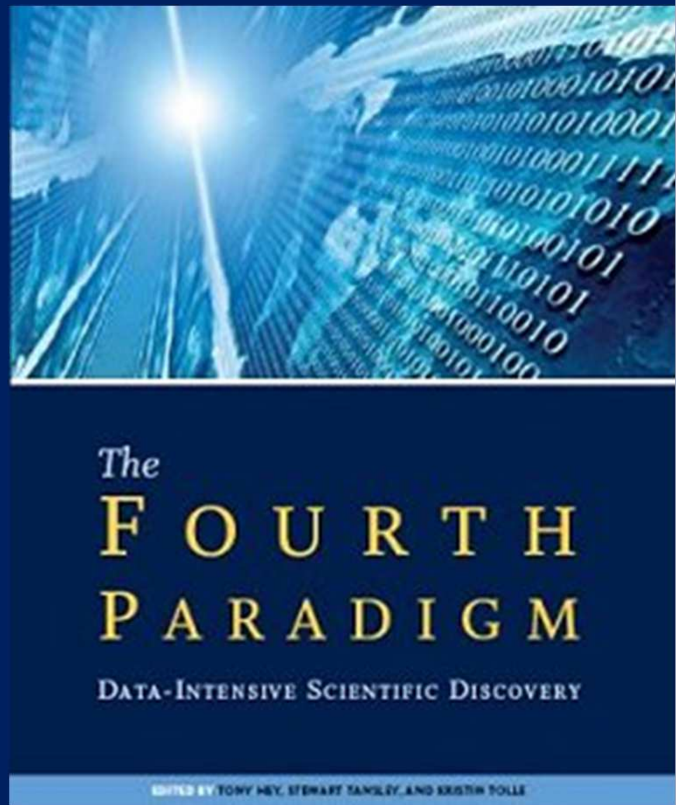


http://research.microsoft.com/fourthparadigm/

"One of the greatest challenges for 21st-century science is ***how we respond to this new era of data intensive science***.

This is recognized as a new paradigm beyond experimental and theoretical research and computer simulations of natural phenomena—one that requires new tools, techniques, and ways of working." — **Douglas Kell**, University of Manchester

1. **Experiment** (ca. 3000 years)

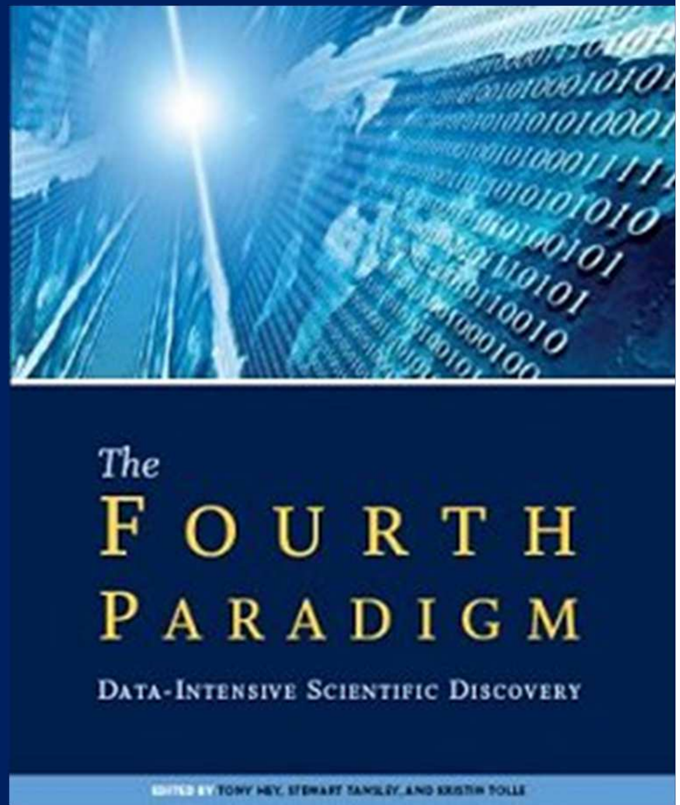


http://research.microsoft.com/fourthparadigm/

"One of the greatest challenges for 21st-century science is ***how we respond to this new era of data intensive science***.
This is recognized as a new paradigm beyond experimental and theoretical research and computer simulations of natural phenomena—one that requires new tools, techniques, and ways of working." — **Douglas Kell**, University of Manchester

1. **Experiment** (ca. 3000 years)

2. **Theory** (few hundreds years)
   mathematical description, theoretical models, analytical laws (e.g. Newton, Maxwell, etc.)



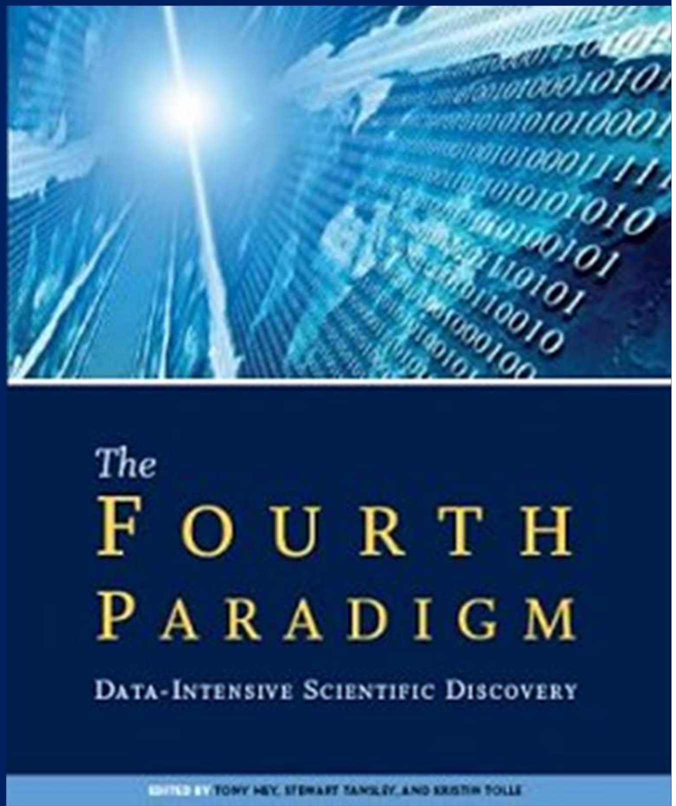
http://research.microsoft.com/fourthparadigm/

"One of the greatest challenges for 21st-century science is ***how we respond to this new era of data intensive science***.
This is recognized as a new paradigm beyond experimental and theoretical research and computer simulations of natural phenomena—one that requires new tools, techniques, and ways of working." — **Douglas Kell**, University of Manchester

1. **Experiment** (ca. 3000 years)

2. **Theory** (few hundreds years)
   mathematical description, theoretical models, analytical laws (e.g. Newton, Maxwell, etc.)

3. **Simulations** (few tens of years) Complex phenomena



http://research.microsoft.com/fourthparadigm/

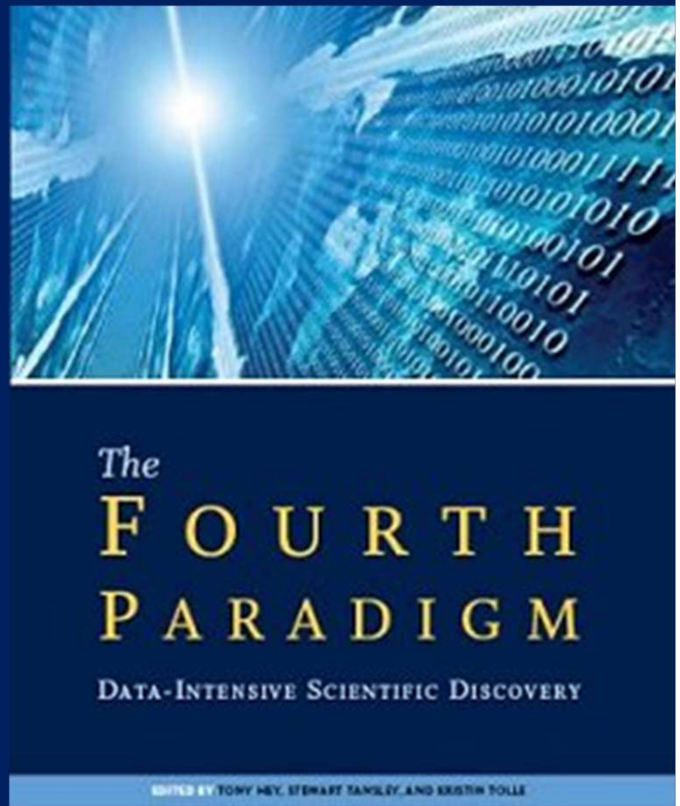
"One of the greatest challenges for 21st-century science is ***how we respond to this new era of data intensive science***.
This is recognized as a new paradigm beyond experimental and theoretical research and computer simulations of natural phenomena—one that requires new tools, techniques, and ways of working." — **Douglas Kell**, University of Manchester

1. **Experiment** (ca. 3000 years)

2. **Theory** (few hundreds years)
   mathematical description, theoretical models, analytical laws (e.g. Newton, Maxwell, etc.)

3. **Simulations** (few tens of years) Complex phenomena

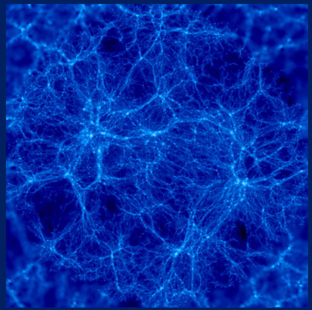
4. **Data-Intensive science** (**now!!!**)



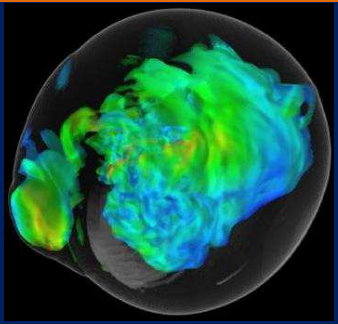
http://research.microsoft.com/fourthparadigm/

The fourth paradigm relies upon….

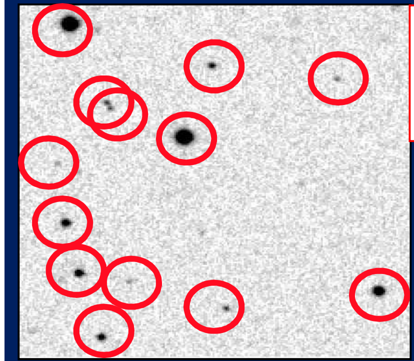
1. Most data will never be seen by human → Need for ML, KDD ecc.



2. Complex correlations *(precursors of physical laws)* cannot be visualized and recognized by the human brain → Most if not all empirical correlations depend on three parameters only: … Simple universe or rather human bias?



3. Real world physics is too complex. Validation of models requires *accurate simulations, tools to compare simulations and data,* and better ways to deal with complex & massive data sets → Need to increase computational and algorithmic capabilities beyond current and expected technological trends



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

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

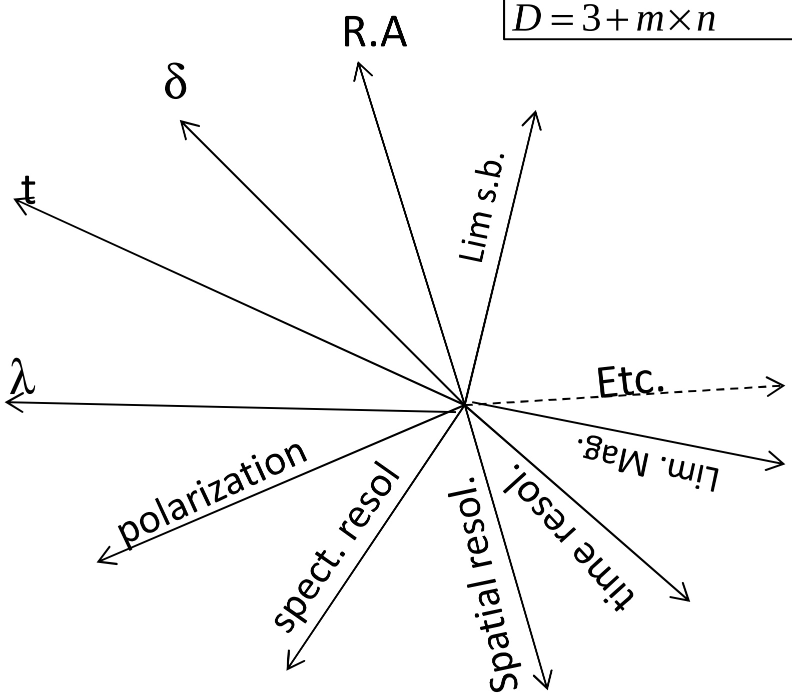
$$p^1 = \{RA^1, \delta^1, t, \{\lambda_1, \Delta\lambda_1, f_1^{1,1}, \Delta f_1^{1,1}, \ldots, f_1^{1,m}, \Delta f_1^{1,m}\}, \ldots, \{\lambda_n, \Delta\lambda_n, f_n^{1,1}, \Delta f_n^{1,1}, \ldots, f_n^{1,m}, \Delta f_n^{1,m}\}\}$$

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

....

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

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

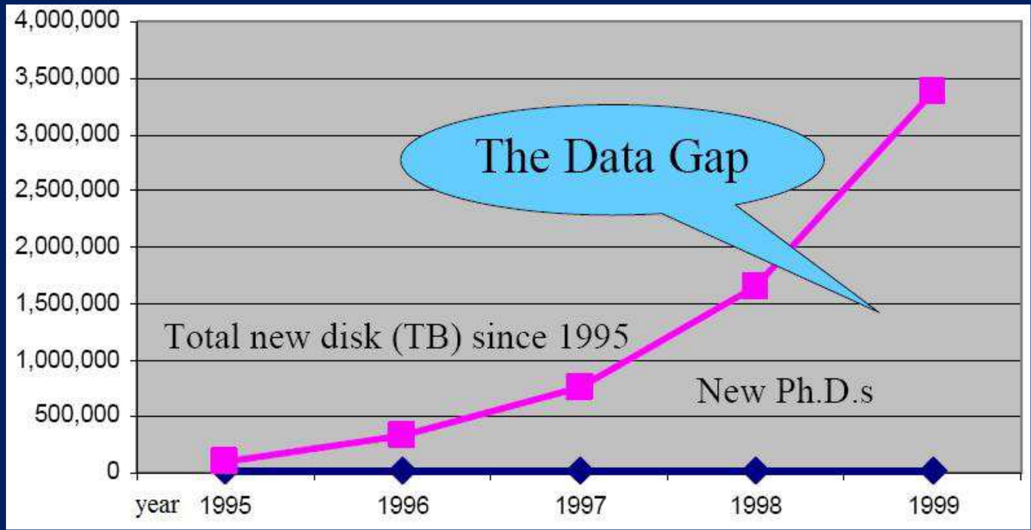


PARAMETER SPACE

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

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

The Data Gap…



4,000,000
3,500,000
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
The Data Gap
Total new disk (TB) since 1995
New Ph.D.s
year
1995
1996
1997
1998
1999

Data Intensive Science

$$\oint \mathbf{E}\cdot d\mathbf{A} = \frac{q_{enc}}{\varepsilon_0}$$
$$\oint \mathbf{B}\cdot d\mathbf{A} = 0$$
$$\oint \mathbf{E}\cdot d\mathbf{s} = -\frac{d\Phi_B}{dt}$$
$$\oint \mathbf{B}\cdot d\mathbf{s} = \mu_0\varepsilon_0\frac{d\Phi_E}{dt} + \mu_0 i_{enc}$$

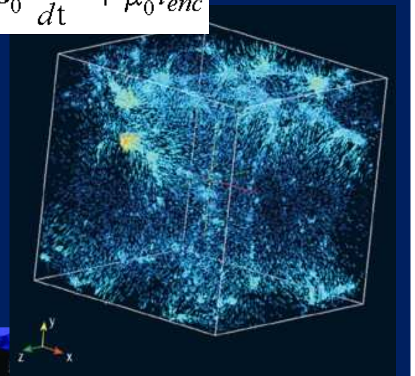
Data Gathering (e.g., from sensor networks, telescopes…)

→ Data Farming:

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

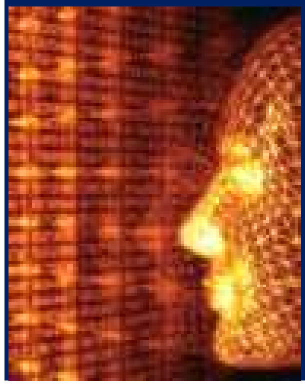
→ Data Mining:

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



→ Data understanding

- Computer aided understanding
- KDD
- Etc.



→ New Knowledge

Data Intensive Science

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

Data Farming:

Storage/Archiving

$$\oint \mathbf{E}\cdot d\mathbf{A} = \frac{q_{enc}}{\varepsilon_0}$$

X - INFORMATICS

Clustering analysis, automated classification

Outlier / anomaly searches

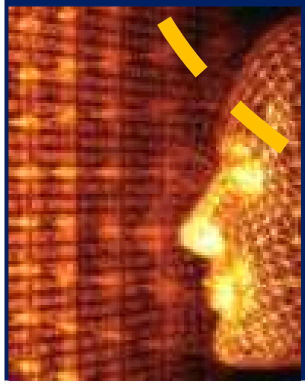
Hyperdimensional visualization

Data understanding

Computer aided understanding

KDD

Etc.



New Knowledge

My Thesis Work

I tried to use the Astroinformatics tools to several problems…

…well sometimes I needed to create that tool…

Algorithmic Aspects:

- GAME
- MLPQNA
- SVM

Technological Aspects

- DAMEWARE
- STraDiWa

Scientific Aspects:

- AGN classification
- Comparison of catalogue extracting methods
- EUCLID Mission
- Globular Cluster classification
- Photometric Redshifts
- Transients detection and modellization

This talk is focused on the Yellow Points

Photometric Redshift

When a spectrum can be obtained, determining the redshift is rather straight-forward: if you can localize the spectral fingerprint of a common element, such as hydrogen, then the redshift can be computed using simple arithmetic. But similarly to the case of Star/Quasar classification, the task becomes much more difficult when only photometric observations are available.
Because of the spectrum shift, an identical source at different redshifts will have a different color through each pair of filters.

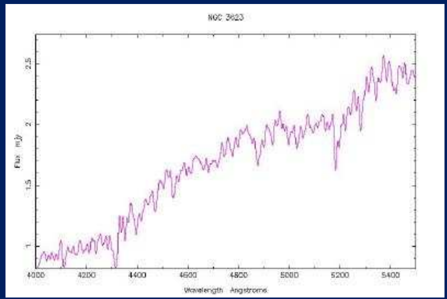
OK, but why we need photometric redshift?

SDSS DR9 Facts	
Sky coverage	14,555 square degrees
Catalog objects	932,891,133
Galaxy spectra	1,457,002
Quasar spectra	228,468
Star spectra	668,054

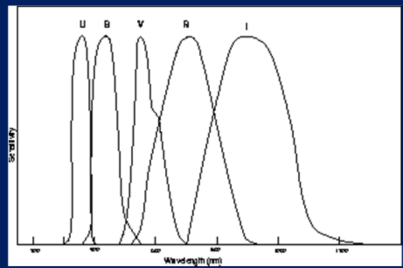
932,891,133 PHOTOMETRIC OBJECTS
2,353,524 SPETTROSCOPIC OBJECTS
~ 400 times more objects!!!

PHOTOMETRIC REDSHIFTS AS A INVERSE PROBLEM

Spectral Energy Distribution convolved with band filters



X



=

$$\begin{cases} m_U = -2.5\log_{10}\dfrac{\int F(\lambda)S_U(\lambda)d\lambda}{\int S_U(\lambda)d\lambda} + c_u \\ m_B = -2.5\log_{10}\dfrac{\int F(\lambda)S_B(\lambda)d\lambda}{\int S_B(\lambda)d\lambda} + c_B \end{cases}$$

Galaxy spectrum - F(λ)

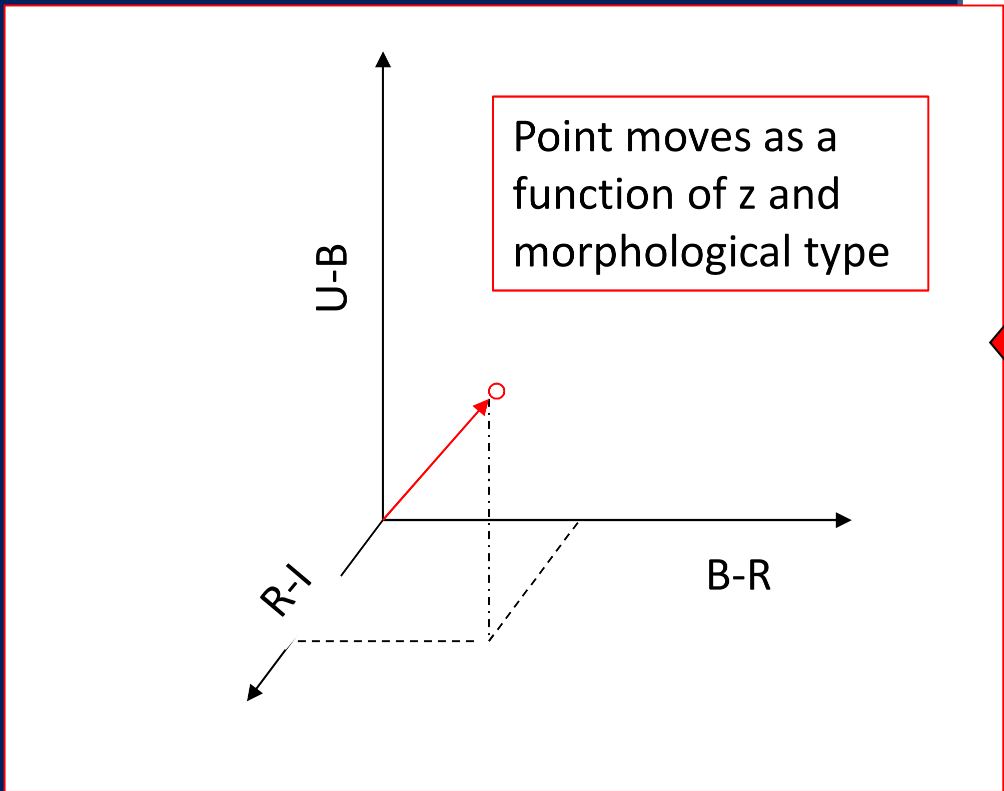
Photometric system - $S_i(\lambda)$

Color indexes

$U - B \equiv m_U - m_B$

$B - R \equiv m_B - m_R$

etc.



Phot-z are an inverse problem

A short History: (see e.g. Yee 1998 for a review)

- Baum (1962)
  Colors of early type galaxies measured from 9 bands with a photometer were turned into a low resolution SED to determine distances of galaxy clusters relative to other clusters of galaxies.

- Koo (1985)
  Colors (from photographic plate material) were compared to colors expected for synthetic Bruzual-Charlot SEDs. Redshifts were estimated from iso-z lines in colorcolor diagrams.

- Loh & Spillar (1986)
  used χ2-minimization for redshift estimates

- Pello and others
  developed a method of `permitted' redshifts; the intersection of the permitted redshift intervalls for all galaxy colors measured defines `the' redshift of a galaxy.

- Photometric redshifts have become very popular since the middle of the 1990s
  - well calibrated, deep multi-waveband data (HDF, other deep fields, SDSS)
  - representative spectroscopic data sets available to test method (Keck, VLT, SDSS...)
  - better cost efficiency if only approximate redshift is needed

Photometric Redshifts: Methods

Template based:

color-space tessellation, $\chi 2$-minimization, maximum likelihood, Bayesian ...

uses physical information: SED's (sizes, compactness...),
... and therefore biased
extrapolates reasonably ok into unknown territory

Learning based:

Nearest Neighbour, Kd-tree, Direct fitting, Neural Networks, Support Vector Machines, Kernel Regression, Regression Trees & Random Forests...

ignores physical information: and therefore unbiased,
can uncover unknown dependencies
requires large training set, bad in extrapolation

Photometric redshifts: the Data Mining approach

Photometric redshifts are treated as a regression problem (i.e. function approximation) , hence a DM problem:

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

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

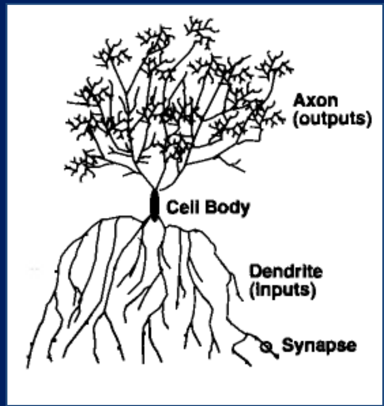
$$\textbf{find} \quad \hat{f}: \quad \hat{\mathbf{Y}} = \hat{f}(\mathbf{X}) \quad \textbf{is a good approximation of } \mathbf{Y}$$

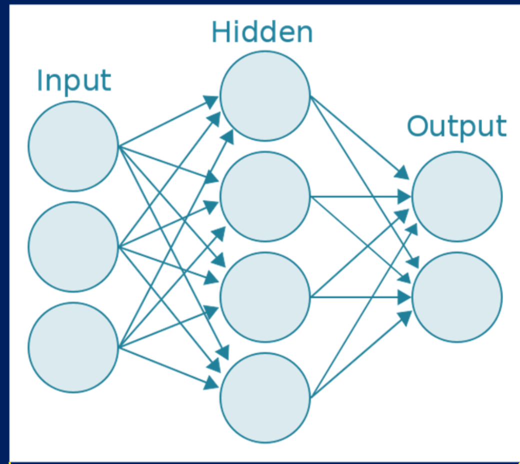
KB = Knowledge Base

KB(from VO)
(set of templates) → Mapping function → Knowledge (phot-z's)

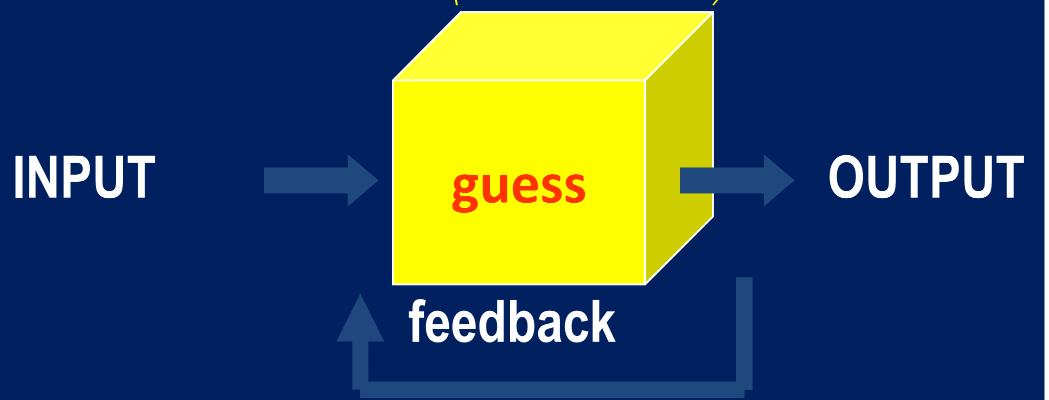
Our Photometric Redshift Method - MLP

A Multi Layer Perceptron is a mathematical operator that mimics the brain behavior:





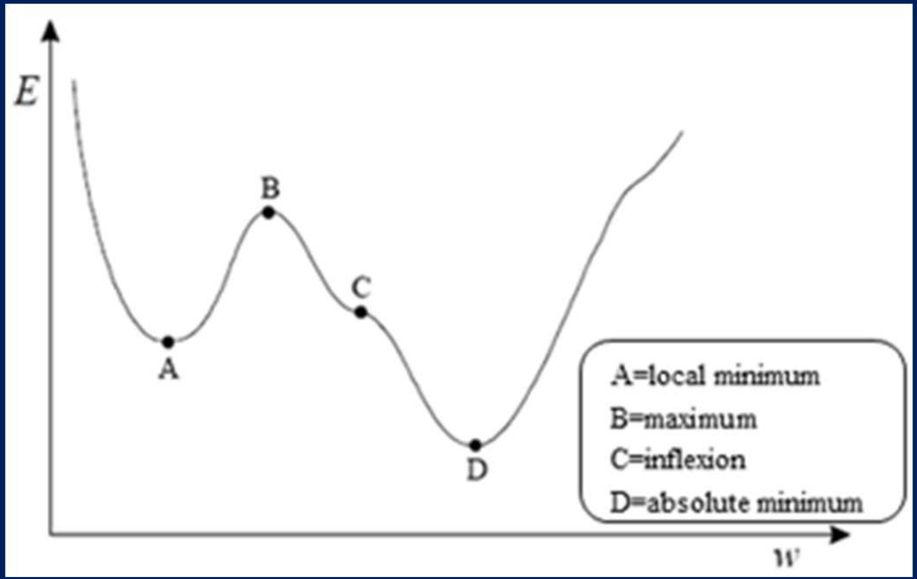
Neurons are connected by «activaction functions» we have different kind of MLP changing the way with they found the best solution



Our Photometric Redshift Method - MLPQNA

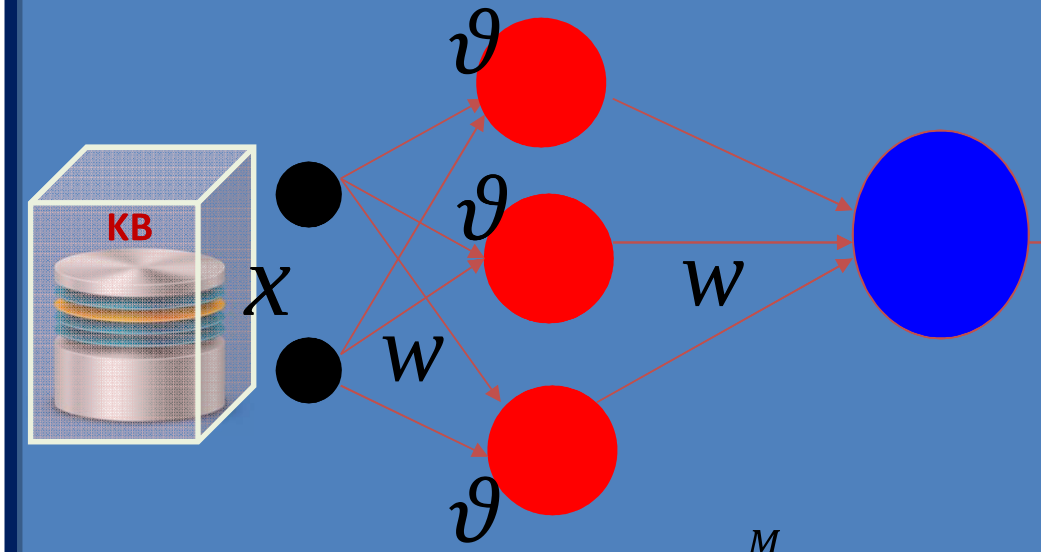
MLP may be trained in several ways, we implement and tested some of them (Back Propagation, Genetic Algorithm and Quasi Newton Algorithm).

QNA are based on Newton's method to find the stationary point of a function, where the gradient is 0. Newton's method assumes that the function can be locally approximated as a quadratic in the region around the optimum, and use the first and second derivatives (gradient and Hessian) to find the stationary point.



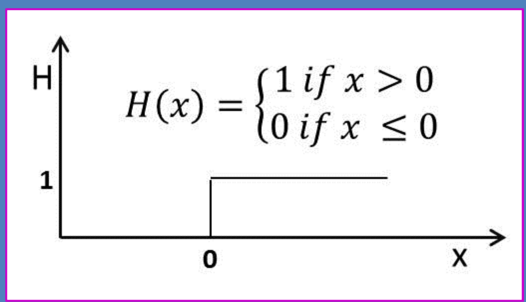
We used MLPQNA with great results both in regression and classification cases, the redshift estimation that follows are the regression use cases.

Multi Layer Perceptron

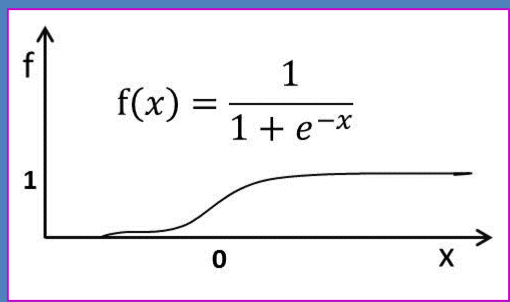


$$y(x; w, \vartheta) = GC/noGC$$

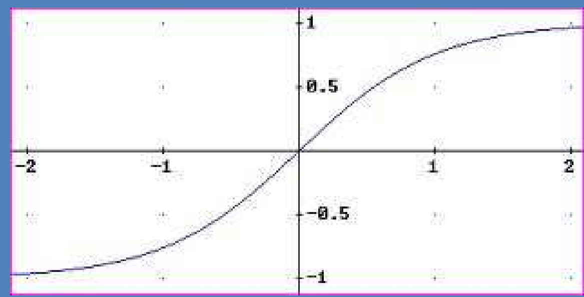
$$y(x; w, \vartheta) = \sum_{i=1}^{M} activ\_func(W_i^T x - \vartheta_i)$$



Heaviside



Sigmoidal

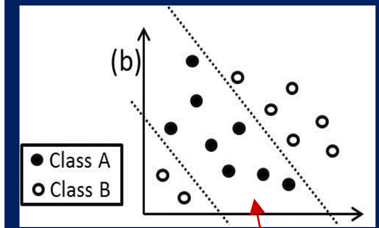


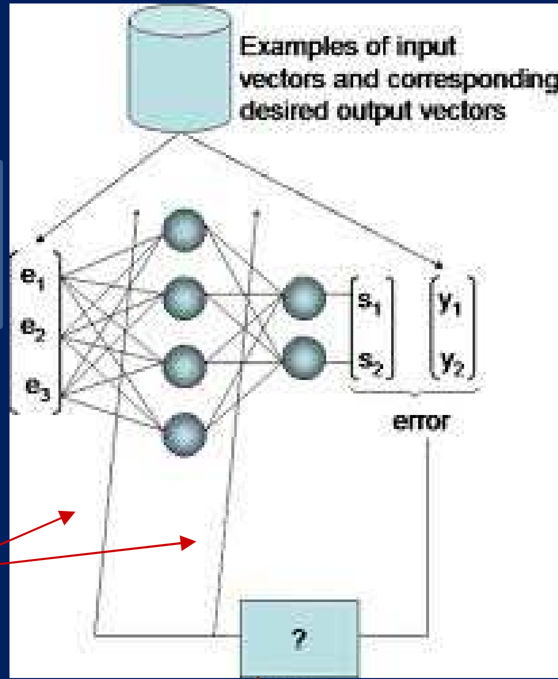
Hyperbolic tangent

MLP learning phase

$$\min_{w} E(w) = \frac{1}{2P}\sum_{p=1}^{P} E_p(w) = \frac{1}{2P}\sum_{p=1}^{P}(y(x^p;w) - d^p)^2$$

E_p is a measure of the **error related to the *p*-th pattern**





$$w^{k+1} = w^k + \alpha^k d^k$$

$d^k \in R^N$ **DIRECTION OF SEARCH**

$\alpha^k \in R$ **STEP**

$$d^k = -\nabla E(w^k)$$ **Descent gradient (BP)**

$$d^k = genetic\ operators$$ **Genetic Algorithms (GA)**

$$\nabla^2 E(w^k) d^k = -\nabla E(w^k)$$ **Hessian approx. (QNA)**

Our Photometric Redshift Environment - DAME Program

DAME Program is a joint effort between University Federico II, Caltech and INAF-OACN, aimed at implementing (as web 2.0 apps and services) a scientific gateway for data exploration on top of a virtualized distributed computing environment.

Multi-purpose data mining with machine learning Web App REsource

Extensions
- DAME-KNIME
- ML Model plugin

Specialized web apps for:
- text mining (VOGCLUSTERS)
- Transient classification (STraDiWA)
- EUCLID Mission Data Quality

Web Services:
- SDSS mirror
- WFXT Time Calculator
- GAME (GPU+CUDA ML model)

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

Science and management
Documents
Science cases
Newsletters

http://www.youtube.com/user/DAMEmedia

DAMEWARE Web Application media channel

PHoto-z Accuracy Testing – PHAT1 CONTEST

The PHAT consists of a **competition** engaged by involving several worldwide groups with the aim at evaluate different (theoretical/empirical) methods to extract photo-z from an ensemble of ground-based and space observation catalogues in several bands, composed to perform photometric redshift prediction evaluation tests of several models, both theoretical and empirical, based on the training/statistics of given spectroscopic redshifts. The imaging dataset is obtained basically on the **GOODS-North** (Great Observatories Origins Deep Survey Northern field.
The total features of object (**1984**) **patterns** are indeed based on **18 bands**.

In this contest, in fact, **only 515 objects** were made available with the corresponding spectroscopic redshift, while for the remaining 1469 objects the related spectroscopic redshift has been hidden from all participants.

PHoto-z Accuracy Testing – PHAT1 CONTEST

Filter	Instrument	$m_{\mathrm{lim.;AB}}$
U	MOSAIC@KPNO-4 m	27.1[a]
B	SUPRIMECAM@Subaru	26.9[a]
V	SUPRIMECAM@Subaru	26.8[a]
R	SUPRIMECAM@Subaru	26.6[a]
I	SUPRIMECAM@Subaru	25.6[a]
Z	SUPRIMECAM@Subaru	25.4[a]
$F435W$	ACS@HST	27.8[b]
$F606W$	ACS@HST	27.8[b]
$F775W$	ACS@HST	27.1[b]
$F850LP$	ACS@HST	26.6[b]
J	ULBCAM@UH-2.2 m	24.1[c]
H	ULBCAM@UH-2.2 m	23.1[c]
HK	QUIRC@UH-2.2 m	22.1[c]
K	WIRC@Hale-5 m	22.5[d]
3.6 μm	IRAC@Spitzer	25.8[e]
4.5 μm	IRAC@Spitzer	25.8[e]
5.8 μm	IRAC@Spitzer	23.0[e]
8.0 μm	IRAC@Spitzer	23.0[e]

18 bands (near UV → mid IR)

Astronomy & Astrophysics manuscript no. aa19755-12
September 5, 2012

Photometric redshifts with the quasi Newton algorithm (MLPQNA). Results in the PHAT1 contest

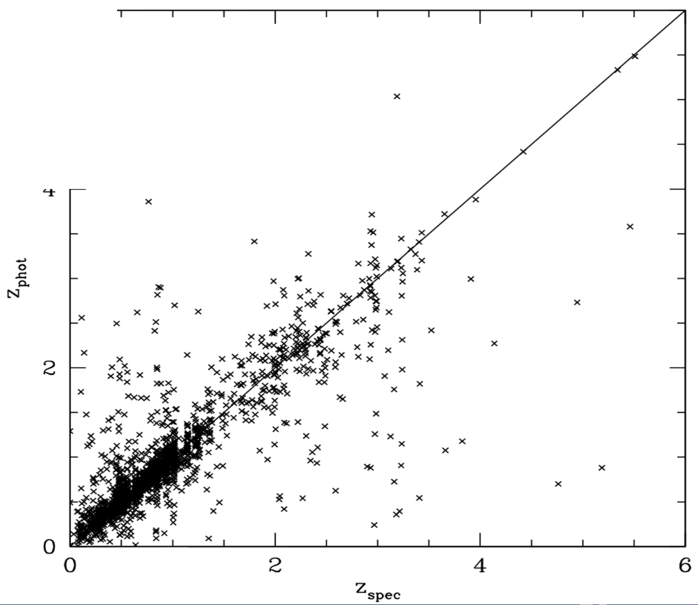
S. Cavuoti[1,2], M. Brescia[2,1], G. Longo[1,2,3], and A. Mercurio[2]

Best among 13 empirical methods

bias ~ 0,0006

σ_{norm}= 0.05

$|\Delta z|>1\sigma$ = 16.33%



A&A 523, A31 (2010)
DOI: 10.1051/0004-6361/201014885

Astronomy & Astrophysics

PHAT: PHoto-*z* Accuracy Testing★

H. Hildebrandt[1], S. Arnouts[2], P. Capak[3], L. A. Moustakas[4], C. Wolf[5], F. B. Abdalla[6], R. J. Assef[7], M. Banerji[8], N. Benítez[9], G. B. Brammer[10], T. Budavári[11], S. Carliles[12], D. Coe[4], T. Dahlen[13], R. Feldmann[14], D. Gerdes[15], B. Gillis[16], O. Ilbert[17], R. Kotulla[18,19], O. Lahav[6], I. H. Li[20], J.-M. Miralles[21], N. Purger[22], S. Schmidt[23], and J. Singal[24]

PHAT1 CONTEST - RESULTS

A	18-band; $\lvert\Delta z\rvert \le 0.15$			14-band; $\lvert\Delta z\rvert \le 0.15$			18-band; $R < 24$; $\lvert\Delta z\rvert \le 0.15$			14-band; $R < 24$; $\lvert\Delta z\rvert \le 0.15$		
Code	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %
QNA	0.0006	0.056	16.3	0.0028	0.063	19.3	0.0002	0.053	11.7	0.0016	0.060	13.7
AN-e	–0.010	0.074	31.0	–0.006	0.078	38.5	–0.013	0.071	24.4	–0.007	0.076	32.8
EC-e	–0.001	0.067	18.4	0.002	0.066	16.7	–0.006	0.064	14.5	–0.003	0.064	13.5
PO-e	–0.009	0.052	18.0	–0.007	0.051	13.7	–0.009	0.047	10.7	–0.008	0.046	7.1
RT-e	–0.009	0.066	21.4	–0.008	0.067	24.2	–0.012	0.063	16.4	–0.012	0.064	18.4
B	18-band; $\lvert\Delta z\rvert \le 0.5$			14-band; $\lvert\Delta z\rvert \le 0.5$			18-band; $R < 24$; $\lvert\Delta z\rvert \le 0.5$			14-band; $R < 24$; $\lvert\Delta z\rvert \le 0.5$		
Code	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %
QNA	–0.0028	0.114	3.8	–0.0046	0.125	3.8	–0.0039	0.101	1.7	–0.0039	0.101	1.7
AN-e	–0.036	0.151	3.1	–0.035	0.173	4.2	–0.047	0.130	1.4	–0.047	0.130	1.4
EC-e	–0.007	0.120	3.6	–0.003	0.114	3.6	–0.015	0.106	1.9	–0.015	0.106	1.9
PO-e	–0.013	0.124	3.1	0.001	0.107	2.3	–0.020	0.098	1.2	–0.020	0.098	1.2
RT-e	–0.031	0.126	3.2	–0.028	0.137	3.6	–0.034	0.111	1.4	–0.034	0.111	1.4
C	18-band; $z_{sp} \le 1.5$, $\lvert\Delta z\rvert \le 0.15$			14-band; $z_{sp} \le 1.5$, $\lvert\Delta z\rvert \le 0.15$			18-band; $z_{sp} > 1.5$, $\lvert\Delta z\rvert \le 0.15$			14-band; $z_{sp} > 1.5$, $\lvert\Delta z\rvert \le 0.15$		
Code	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %	bias	scatter	outliers %
QNA	–0.0004	0.053	14.6	0.0001	0.061	16.6	0.0074	0.072	26.3	0.0222	0.070	35.0
AN-e	–0.017	0.070	27.6	–0.010	0.076	33.6	0.051	0.078	50.7	0.045	0.077	66.4
EC-e	–0.003	0.065	16.1	–0.000	0.064	14.5	0.015	0.077	32.3	0.015	0.077	29.5
PO-e	–0.012	0.049	12.6	–0.011	0.047	9.4	0.019	0.075	48.3	0.026	0.074	37.7
RT-e	–0.016	0.062	19.6	–0.014	0.064	21.1	0.040	0.072	31.8	0.039	0.071	41.9

Statistical Indicators

$$\Delta z = (zspec - zphot)$$

$$\text{bias} = \frac{\sum_{i=1}^{N} \Delta z_i}{N}$$

$$\text{MAD} = Median(|\,\Delta z - Median(\Delta z)\,|)$$

$$\text{standard deviation } \sigma = \sqrt{\frac{\sum_{i=1}^{N}\left[\Delta z_i - \left(\frac{\sum_{i=1}^{N}\Delta z_i}{N}\right)\right]^2}{N}}$$

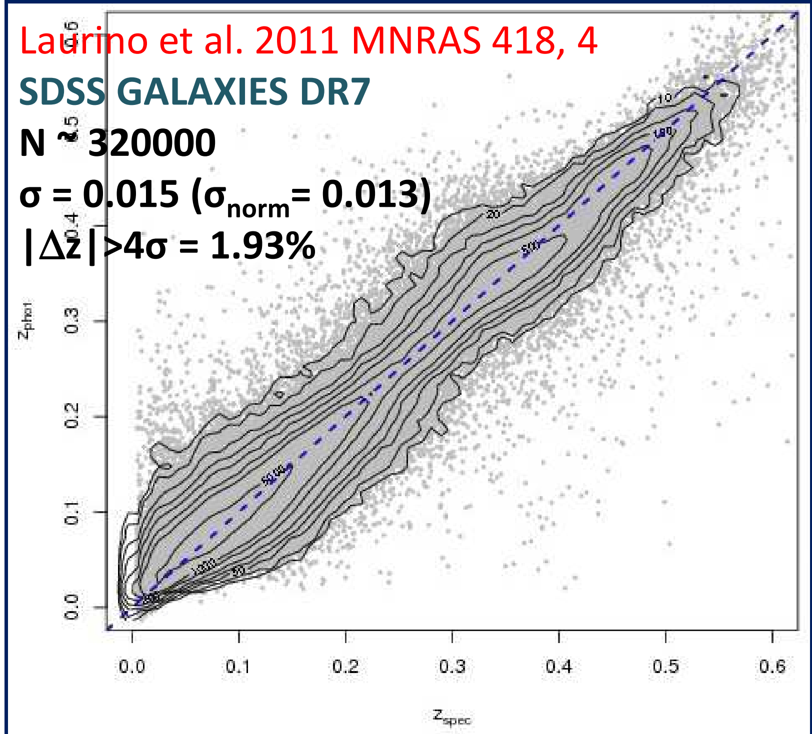
$$\Delta z' = (zspec - zphot)/(1 + zspec)$$

$$\text{bias}_{norm} = \frac{\sum_{i=1}^{N} \Delta z'_i}{N}$$

$$\text{MAD}_{norm} = Median(|\,\Delta z' - Median(\Delta z')\,|)$$

$$\sigma_{norm} = \sqrt{\frac{\sum_{i=1}^{N}\left[\Delta z'_i - \left(\frac{\sum_{i=1}^{N}\Delta z'_i}{N}\right)\right]^2}{N}}$$

Galaxy Photometric redshifts prediction from SDSS DR9 archive;



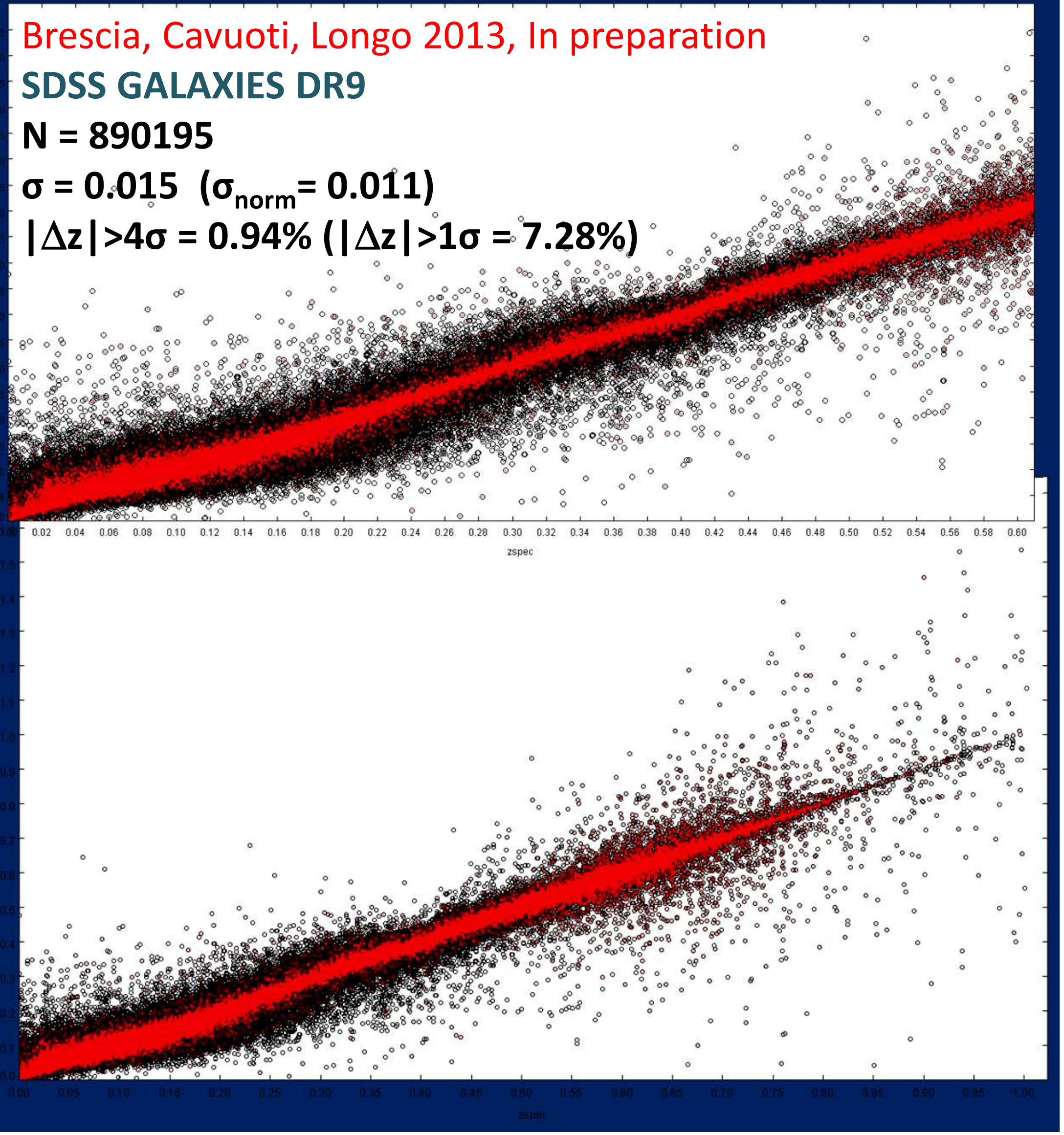
Laurino et al. 2011 MNRAS 418, 4

SDSS GALAXIES DR7

N ~ 320000

σ = 0.015 (σ_{norm}= 0.013)

$|\Delta z|$>4σ = 1.93%



Brescia, Cavuoti, Longo 2013, In preparation

SDSS GALAXIES DR9

N = 890195

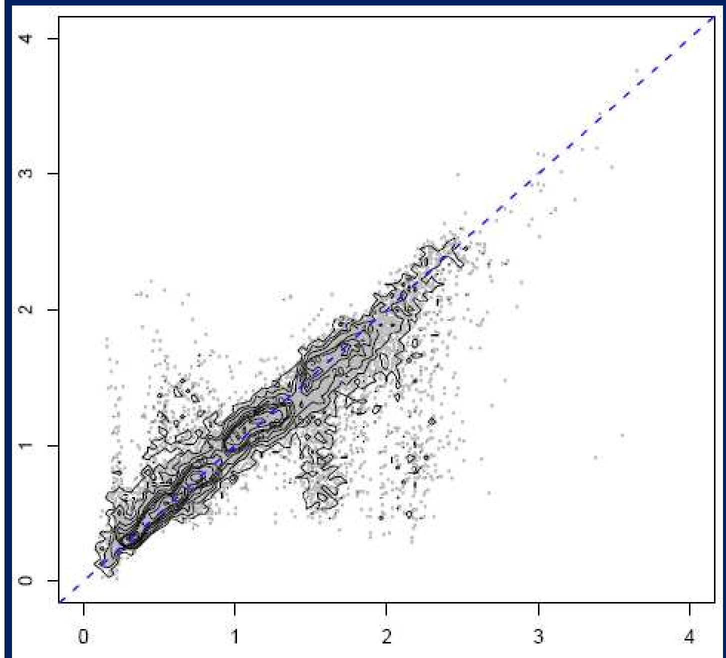
σ = 0.015 (σ_{norm}= 0.011)

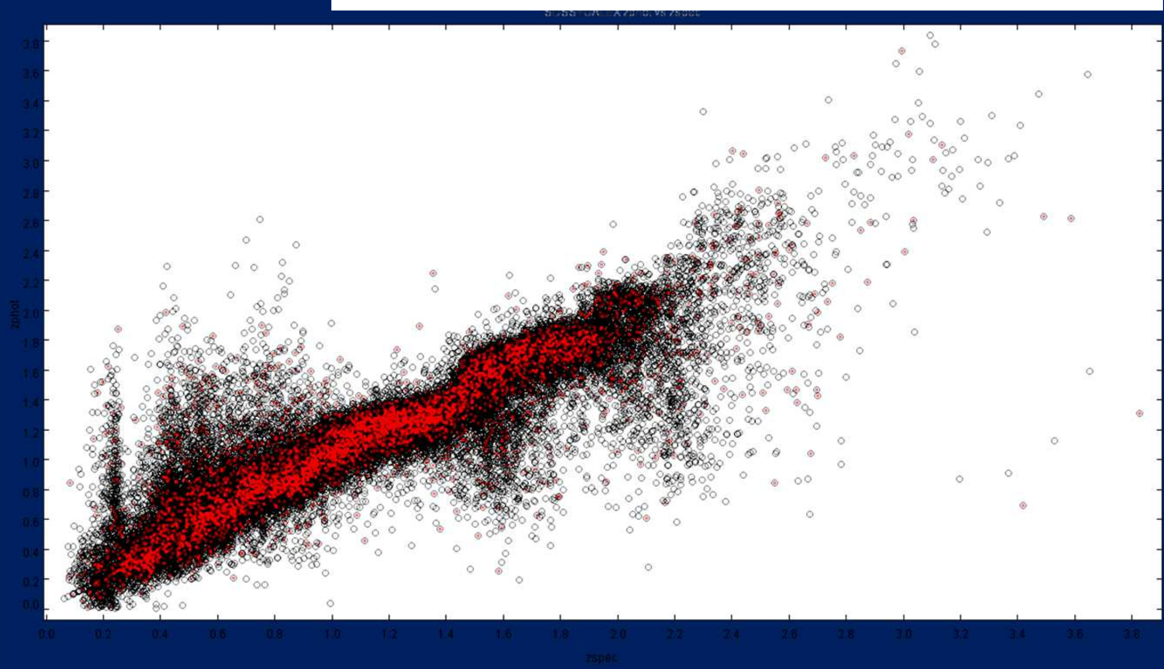
$|\Delta z|$>4σ = 0.94% ($|\Delta z|$>1σ = 7.28%)

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
$\sigma = 0.21$ ($\sigma_{norm} = 0.14$)
$|\Delta z| > 4\sigma = 1.08\%$ ($|\Delta z| > 1\sigma = 14.97\%$)

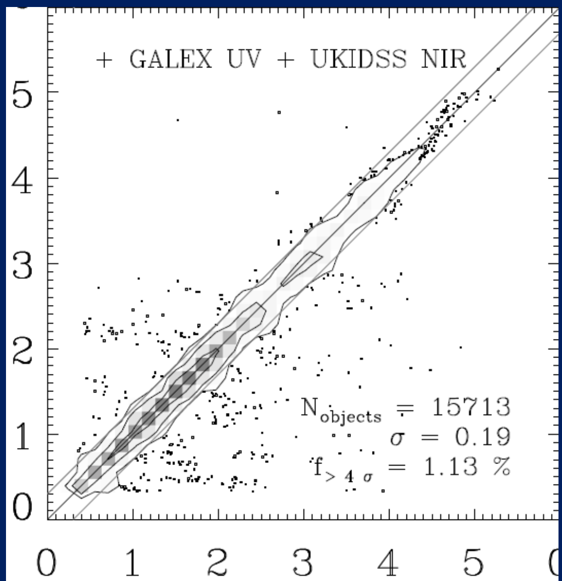




Quasar Photometric redshifts prediction from matched data (SDSS, GALEX, UKIDSS, WISE);

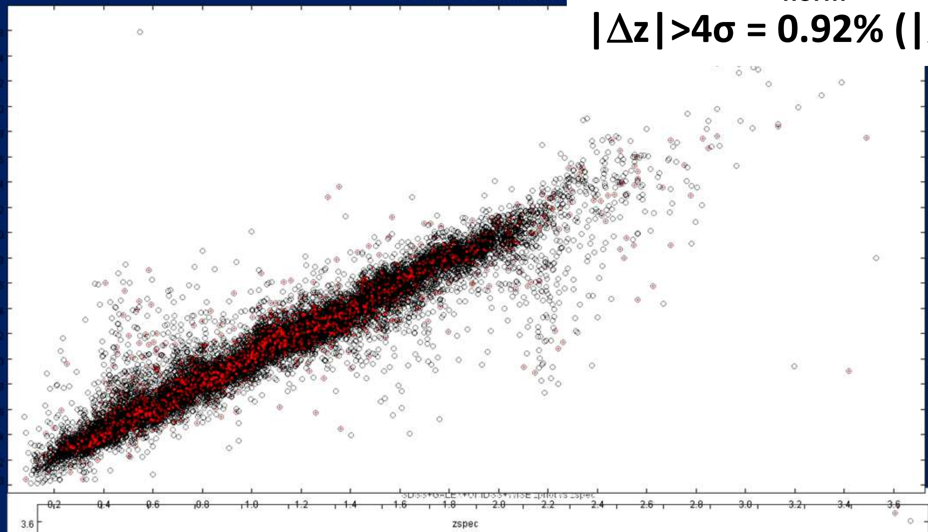
Brescia Cavuoti D'Abrusco Longo Mercurio. 2013, Subm. to MNRAS
, Subm. to MNRAS

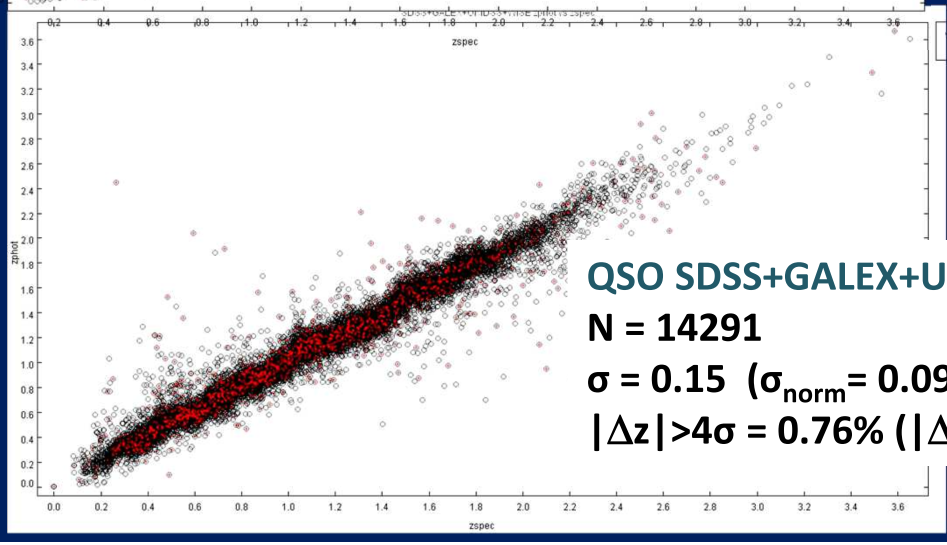
QSO SDSS+GALEX+UKIDSS
N = 14588
$\sigma = 0.15$ ($\sigma_{norm} = 0.096$)
$|\Delta z| > 4\sigma = 0.92\%$ ($|\Delta z| > 1\sigma = 14.81\%$)



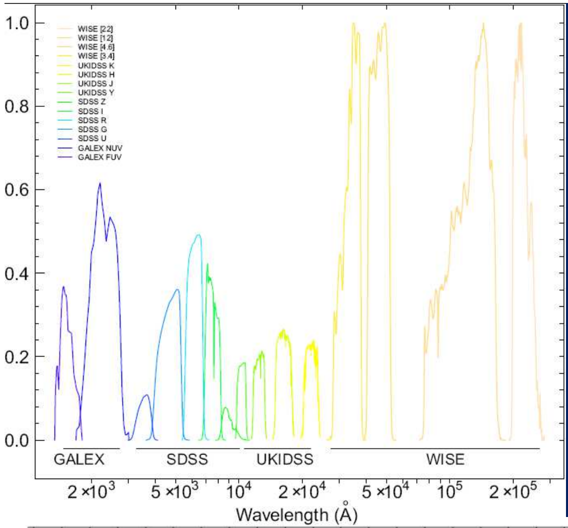
Bovy et al. 2012,
ApJ 749, 41

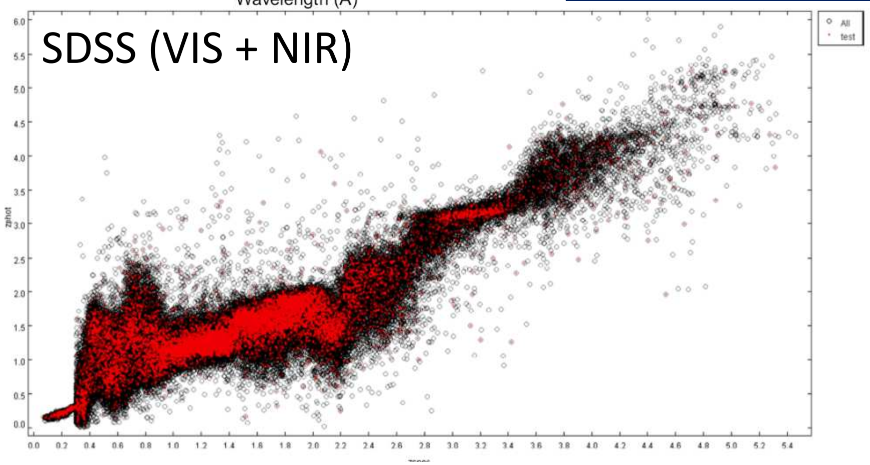
QSO SDSS+GALEX+UKIDSS
N = 15713
$\sigma = 0.19$ ($\sigma_{norm} = 0.21$)
$|\Delta z| > 4\sigma = 1.13\%$
($|\Delta z| > 1\sigma = 19.43\%$)

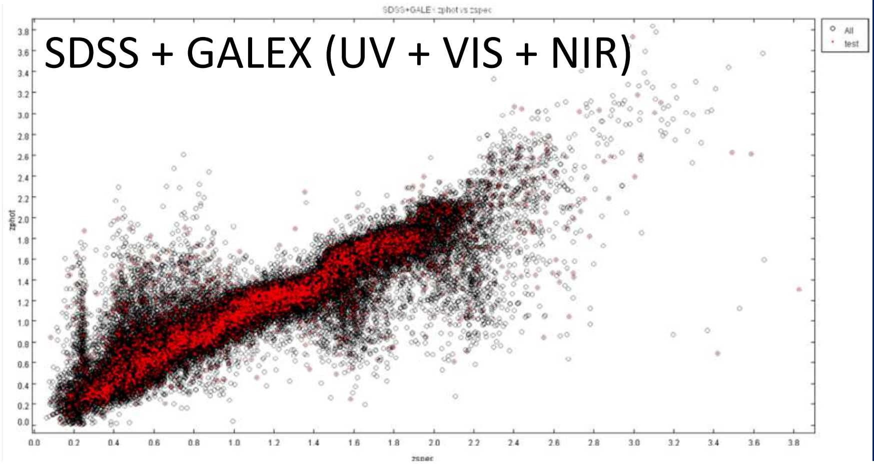


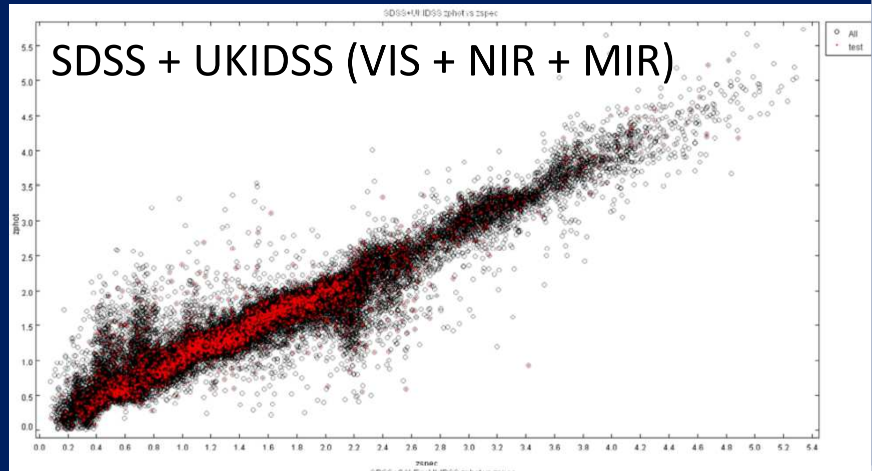


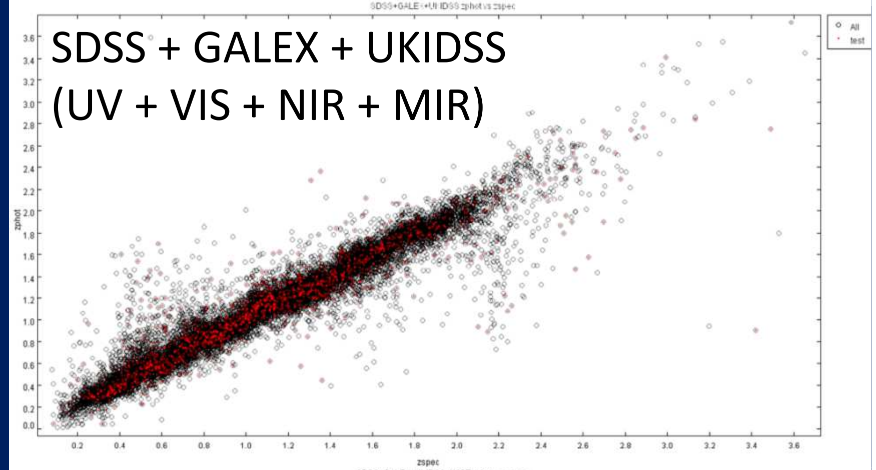
QSO SDSS+GALEX+UKIDSS+WISE
N = 14291
$\sigma = 0.15$ ($\sigma_{norm} = 0.092$)
$|\Delta z| > 4\sigma = 0.76\%$ ($|\Delta z| > 1\sigma = 12.31\%$)

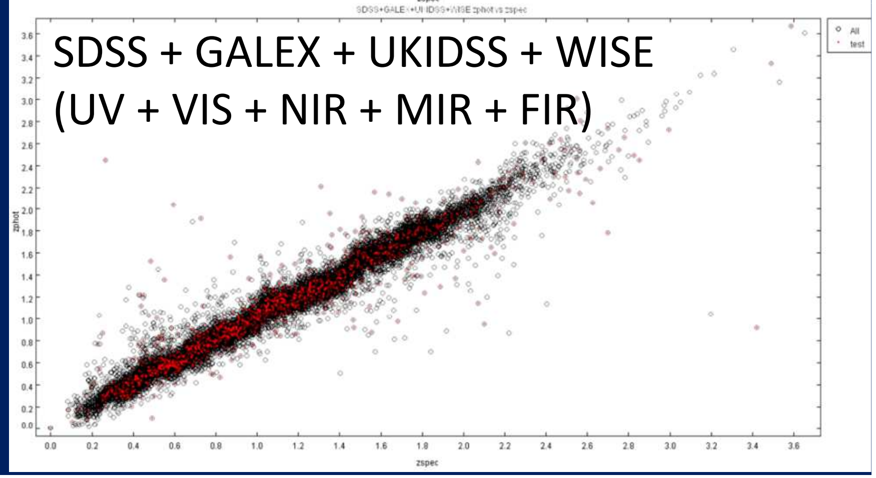

GALEX
SDSS
UKIDSS
WISE
Wavelength (Å)


SDSS (VIS + NIR)

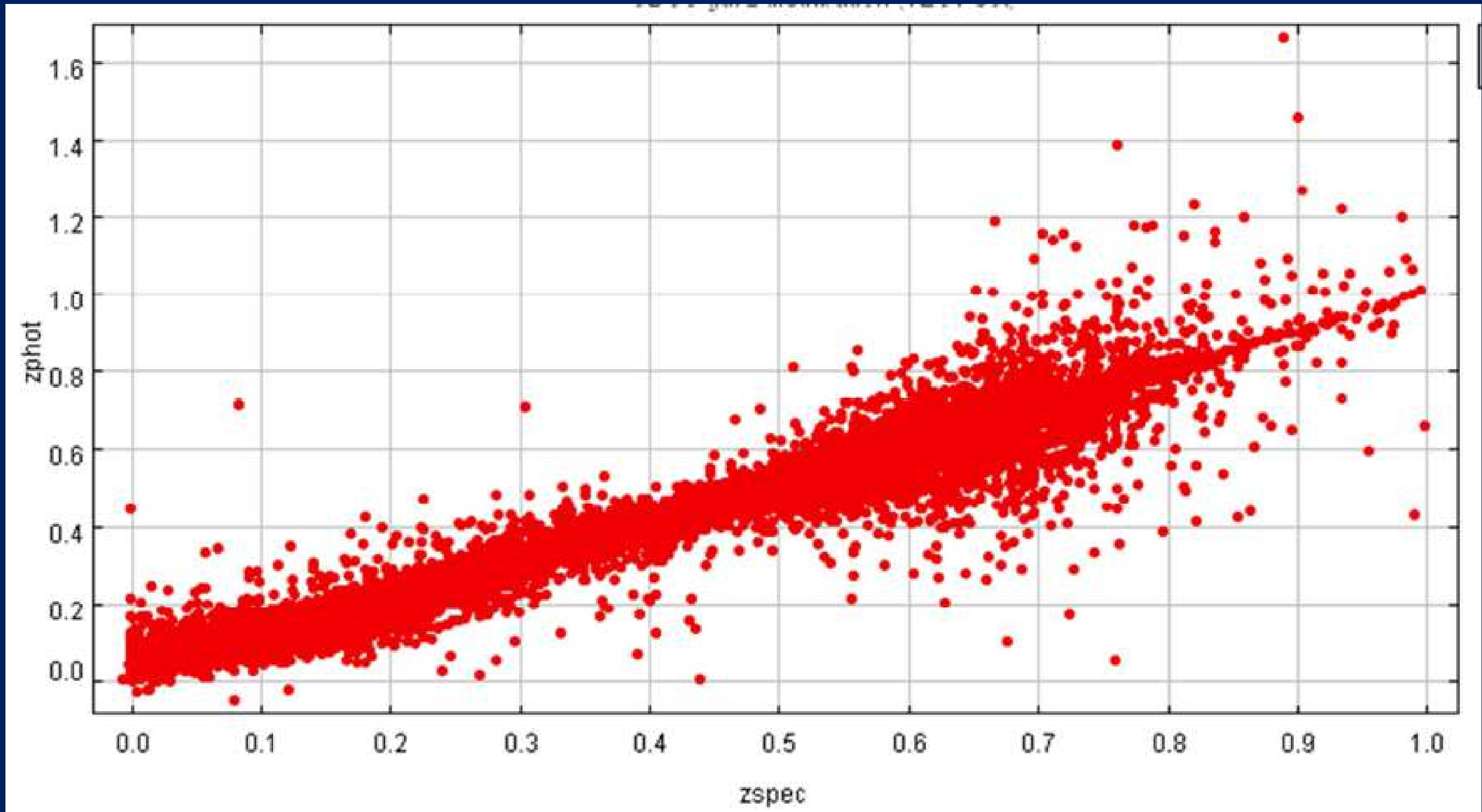

SDSS + GALEX (UV + VIS + NIR)


SDSS + UKIDSS (VIS + NIR + MIR)


SDSS + GALEX + UKIDSS
(UV + VIS + NIR + MIR)


SDSS + GALEX + UKIDSS + WISE
(UV + VIS + NIR + MIR + FIR)

Galaxy Redshift SDSS



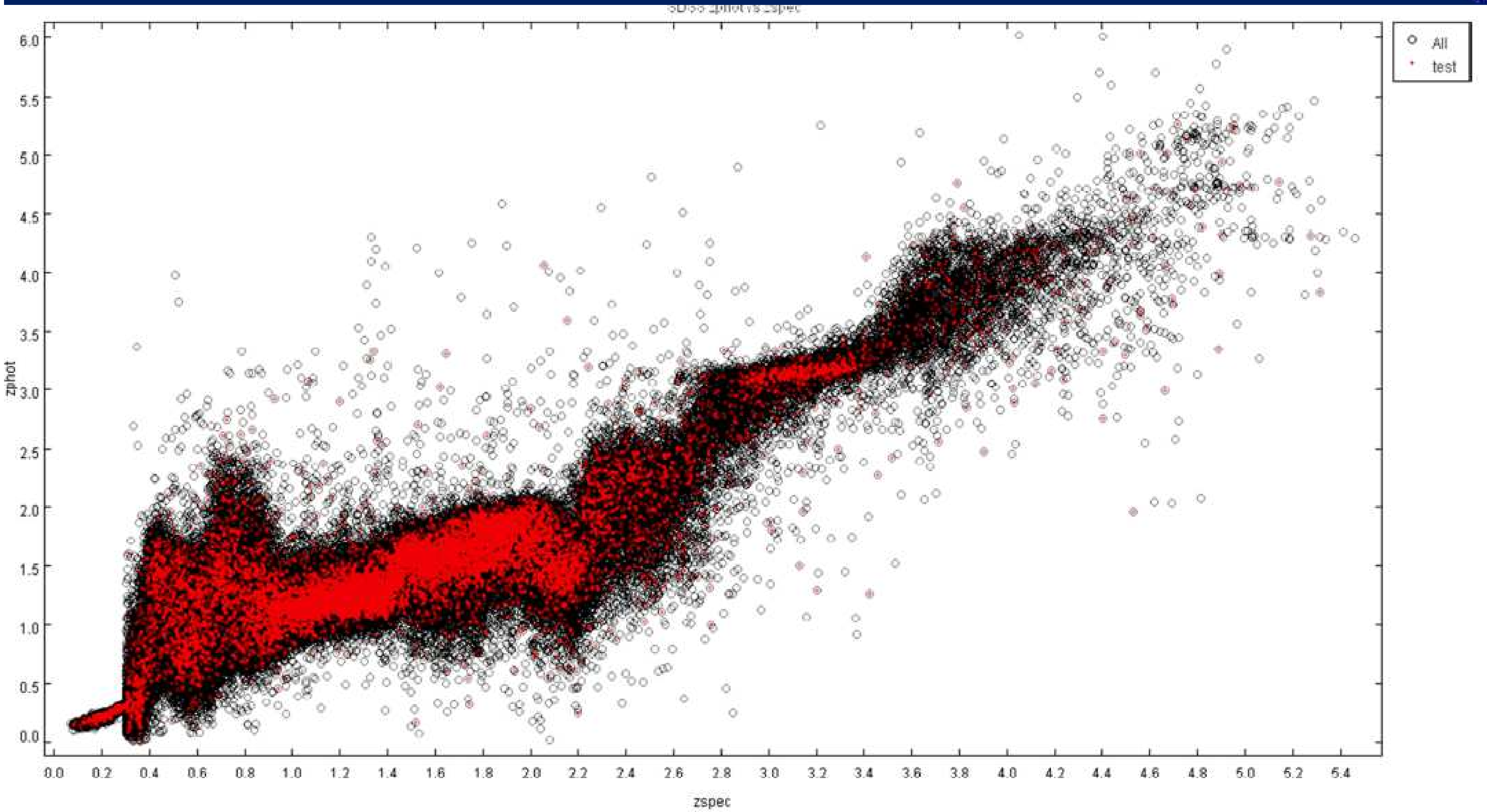
Exp	Bias	sigma	MAD	RMS	biasnorm	snorm	MADnorm	RMSnorm
MLPQNA	-0.0002	0.016	0.001	0.016	-0.0003	0.012	0.0009	0.012
Laurino 2011	0.015	0.015	0.011	0.021	0.014	0.013	0.009	0.019

890119 Objects

QSO Redshift

ID	GALEX	SDSS	UKIDSS	WISE	*BIAS*	σ	*MAD*	out. 1σ	out. 2σ	out. 3σ	out. 4σ
E1	X	X	X	X	0.0033	0.174	0.071	15.96%	4.75%	2.24%	0.92%
E2	$X^{1,2}$	X	X^{1}	X	-0.0001	0.152	0.071	19.66%	4.49%	1.85%	0.92%
E3	X^{3}	X	X^{1}	X	-0.0016	0.165	0.071	15.83%	3.96%	1.98%	1.19%
E4	X^{1}	X	X^{1}	X	0.0054	0.151	0.064	16.23%	4.75%	1.98%	1.06%
E5	X^{2}	X	X^{1}	X	-0.0026	0.151	0.063	18.47%	4.62%	2.37%	0.79%
E6	$X^{4,5}$	X	X^{1}	X	-0.0008	0.152	0.066	17.81%	5.15%	2.64%	0.79%
E7	$X^{1,2,3}$	X	X^{1}	X	0.0041	0.163	0.072	19.39%	4.22%	2.51%	0.66%
E8	$X^{2,3}$	X	X^{1}	X	-0.0033	0.155	0.070	19.26%	5.01%	1.98%	0.92%
E9				X	0.0165	0.297	0.148	22.16%	5.80%	2.11%	0.53%
E10		X			-0.0162	0.338	0.124	19.66%	7.26%	2.37%	0.40%
E11			$X^{1,2}$		-0.0091	0.299	0.144	23.75%	4.88%	1.58%	0.66%
E12	$X^{1,2}$				0.0550	0.419	0.265	29.68%	4.75%	0.79%	0.26%
E13			X^{2}		0.0111	0.465	0.325	34.43%	3.43%	0.40%	0.00%
E14			X^{1}		-0.0081	0.294	0.139	22.82%	5.94%	1.85%	0.66%
E15			X^{1}	X	0.0045	0.236	0.107	17.94%	4.75%	2.11%	1.06%
E16	X^{2}	X	X^{1}		-0.0046	0.152	0.071	21.11%	4.88%	1.98%	0.79%
E17	X^{2}	X		X	0.0025	0.162	0.069	16.23%	3.69%	2.37%	1.06%
E18		X	X^{1}	X	-0.0032	0.179	0.064	14.38%	4.49%	2.11%	1.32%
E19	X^{2}		X^{1}	X	0.0110	0.203	0.091	19.26%	4.88%	1.72%	0.79%
E20	X^{2}			X	0.0175	0.288	0.144	22.96%	4.88%	1.45%	0.53%
E21		X	X^{1}		-0.0027	0.210	0.084	15.96%	5.15%	2.24%	1.06%
E22		X		X	-0.0039	0.197	0.072	13.85%	3.43%	2.37%	1.58%
E23	X^{2}	X			-0.0055	0.240	0.091	17.55%	6.73%	2.51%	0.79%
E24	X^{2}		X^{1}		0.0133	0.238	0.113	23.22%	6.20%	1.72%	0.40%

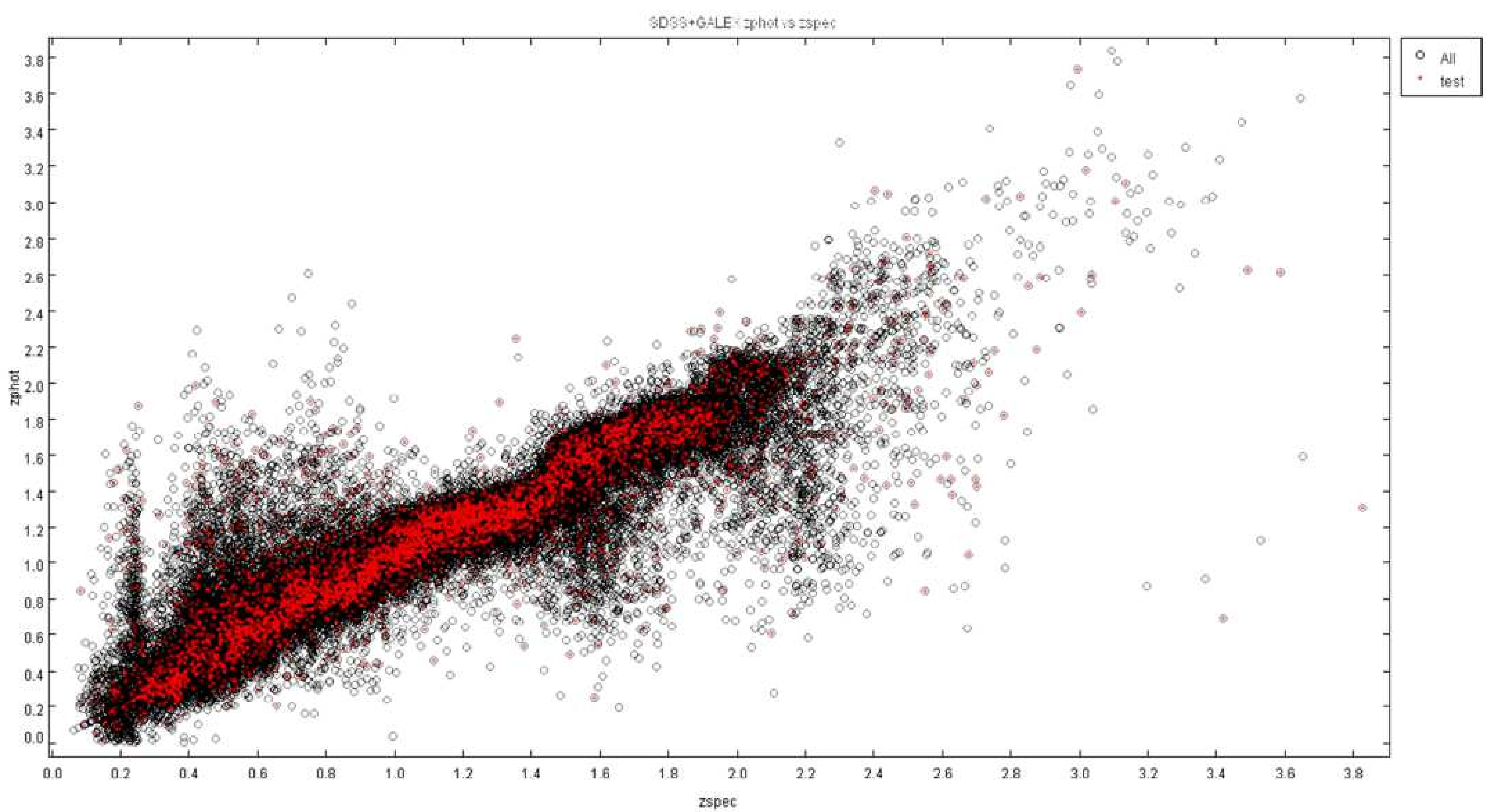
QSO Redshift – SDSS



Ref.	|bias|	sigma	MAD	RMS	|biasnorm|	snorm	MADnorm	RMSnorm
SDSS	0.016	0.34	0,083	0.34	0.034	0.19	0.060	0.19
Bovy 2012		0.46						

105759 objects

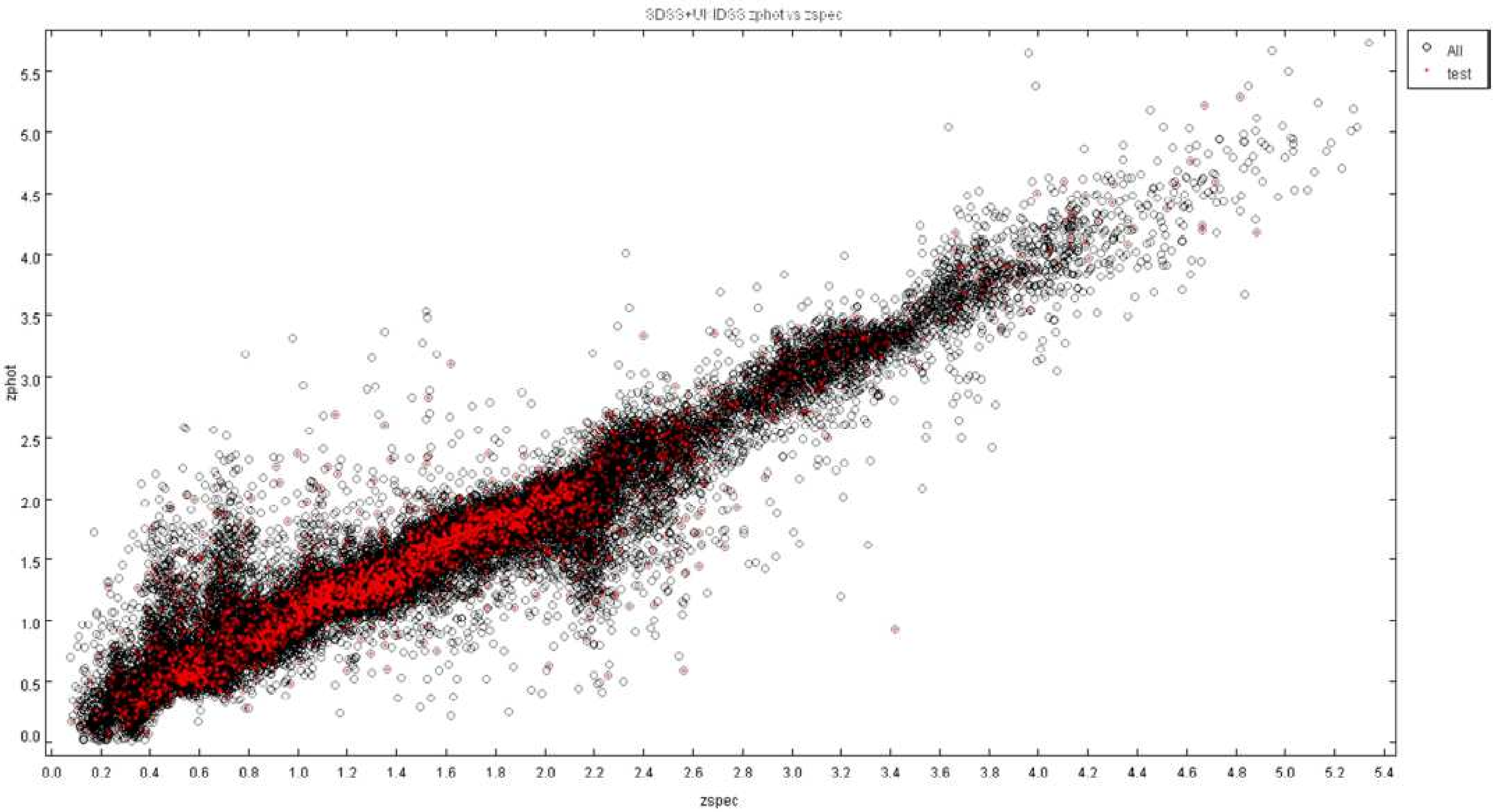
QSO Redshift – SDSS + GALEX



Ref.	|bias|	sigma	MAD	RMS	|biasnorm|	snorm	MADnorm	RMSnorm
SDSS+GALEX	0.005	0.24	0.091	0.24	0.017	0.13	0.046	0.13
Bovy 2012		0.26						

44688 objects

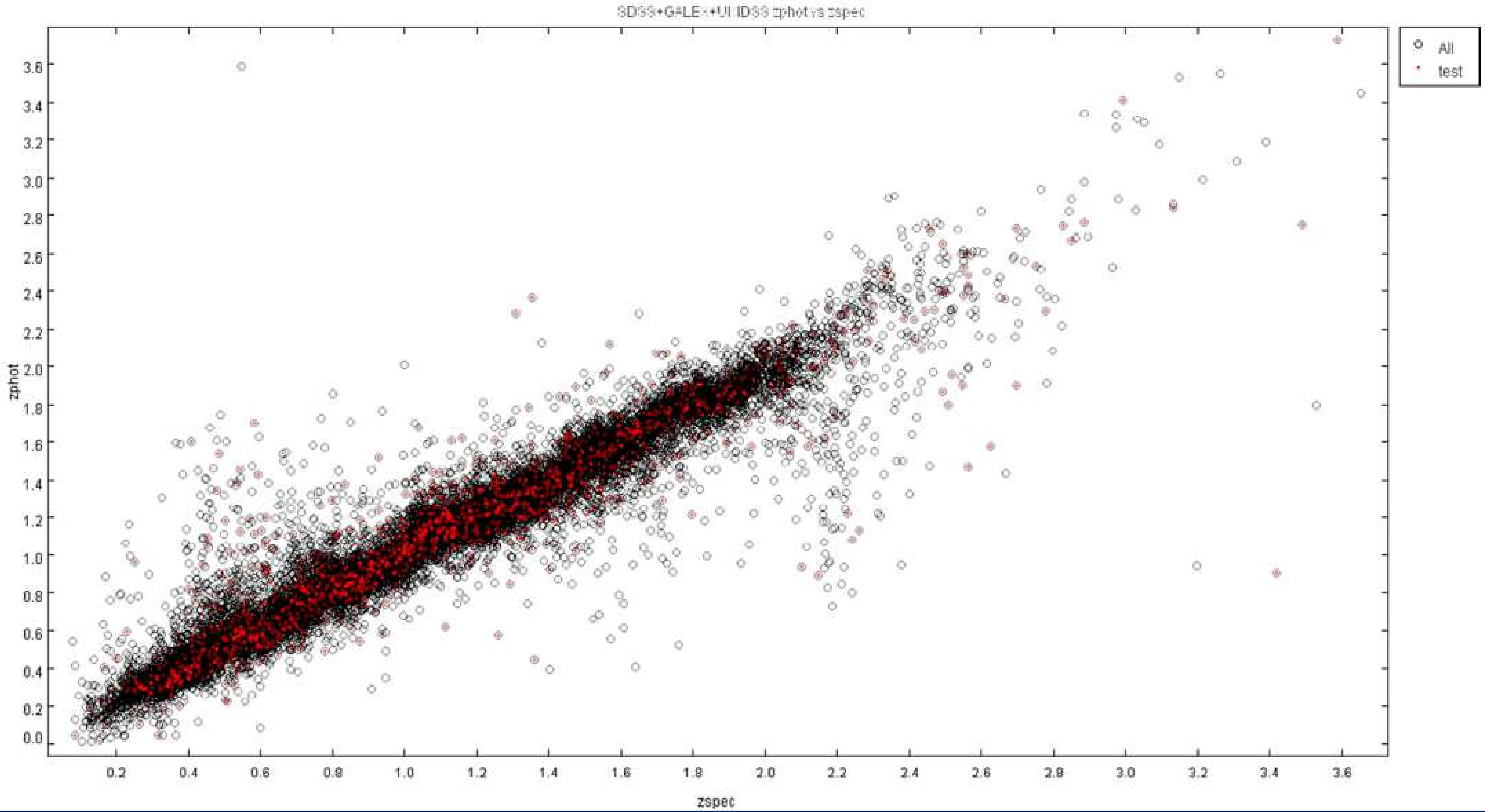
QSO Redshift – SDSS + UKIDSS



Ref.	|bias|	sigma	MAD	RMS	|biasnorm|	snorm	MADnorm	RMSnorm
SDSS+UKIDSS	0.003	0.21	0.084	0.21	0.010	0.11	0.040	0.11
Bovy 2012		0.28						

31094 objects

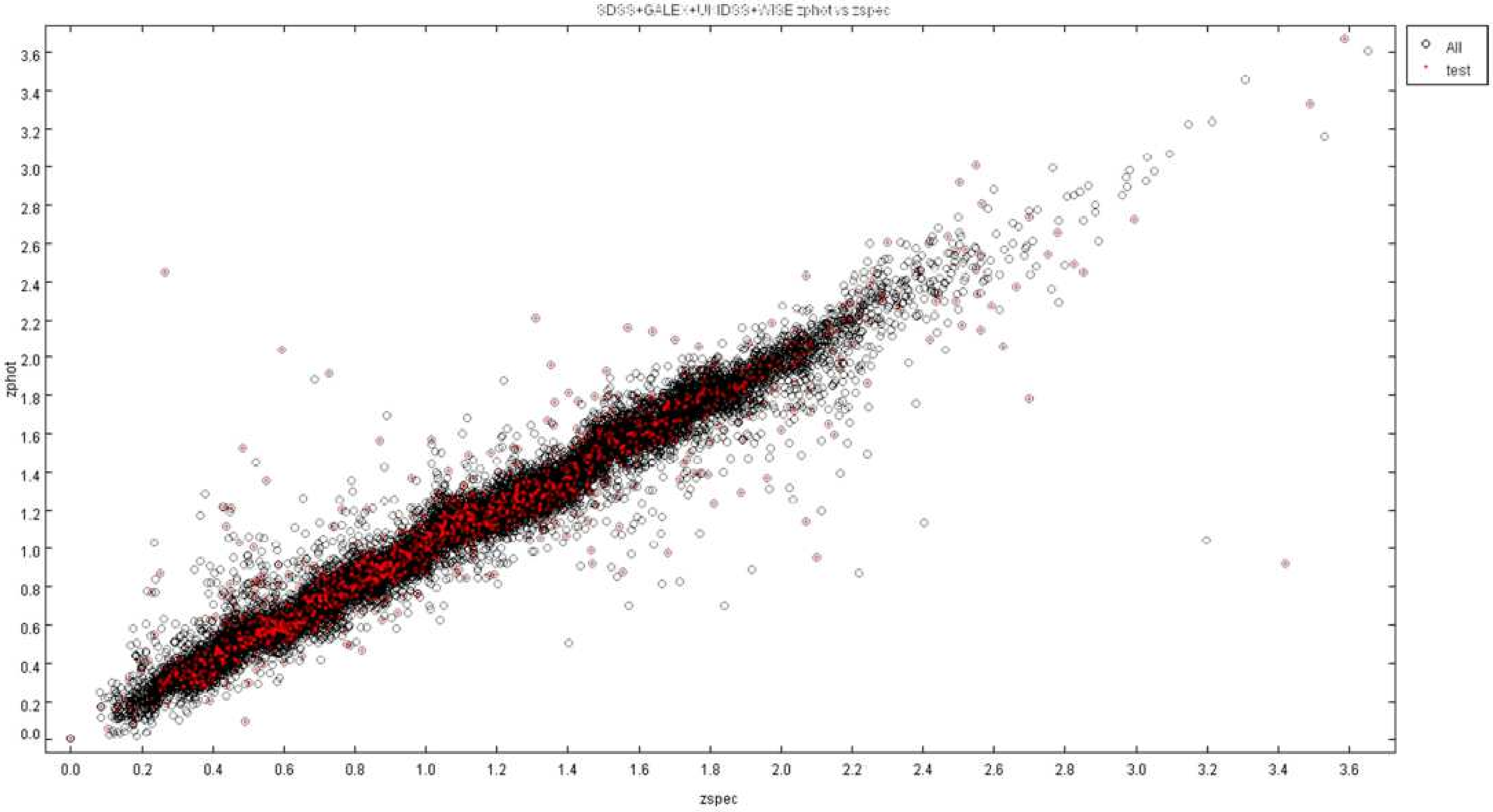
QSO Redshift – SDSS + UKIDSS + GALEX



Ref.	|bias|	sigma	MAD	RMS	|biasnorm|	snorm	MADnorm	RMSnorm
SDSS+GALEX+UKIDSS	0.005	0.15	0.072	0.15	0.006	0.075	0.036	0.075
Bovy 2012		0.21						

14588 objects

QSO Redshift – SDSS + UKIDSS + GALEX - WISE



Ref.	|bias|	sigma	MAD	RMS	|biasnorm|	snorm	MADnorm	RMSnorm
SDSS+GALEX+UKIDSS+WISE	0.003	0.15	0.063	0.15	0.005	0.15	0.063	0.15

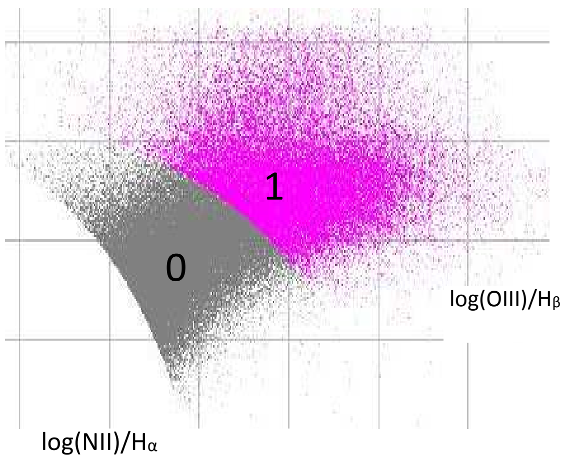
14291 objects

AGN CLASSIFICATION

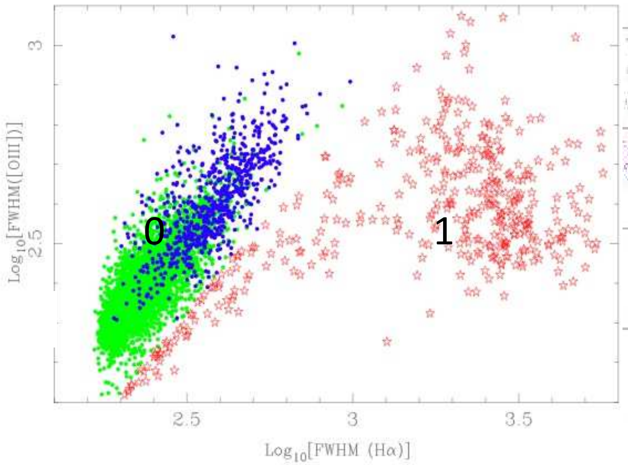
Photometric parameters used for training of the NNs and SVMs:

petroR50_u, petroR50_g, petroR50_r, petroR50_i, petroR50_z
concentration_index_r
fibermag_r
$(u-g)_{dered}$, $(g-r)_{dered}$, $(r-i)_{dered}$, $(i-z)_{dered}$
dered_r
photo_z_corr

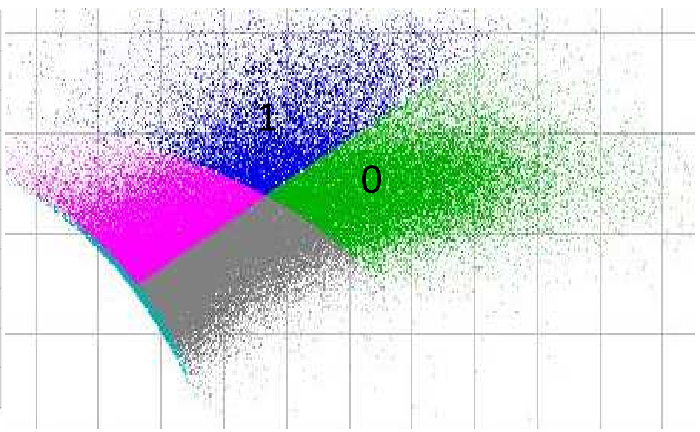
1° Experiment:
AGN -> 1, Mixed-> 0



2° Experiment:
Type 1 -> 1, Type 2 -> 0



3° Experiment:
Seyfert -> 1, LINERs ->0



Cavuoti, S.; Brescia, M.; D'Abrusco, R.; Longo, G.; Photometric AGN Classification in the SDSS with Machine Learning Methods **to be Submitted to MNRAS**

AGN CLASSIFICATION RESULTS

Sample	Parameters	BoK	Algorithm	e_{tot}	C(MLP)
Experiment (1) AGN detection	SDSS photometric parameters + photo redshift	BPT plot +Kewley's line	SVM / MLP	~74% / ~76%	AGN~55 % / Not AGN ~87%
Experiment (2) Type 1 vs. Type 2	SDSS photometric parameters + photo redshift	Catalogue of Sorrentino et al.+Kewley's line	SVM / MLP	e_{typ1}~82% e_{typ2}~86% / etyp2~99% e_{typ1}~98%	Type1 ~99% / Type2 ~100%
Experiment (3) Seyfert Vs. LINERs	SDSS photometric parameters + photo redshift	BPT plot+Heckman's+Kewley's lines	SVM / MLP	Sey~78% / LIN~80%	Sey~53% / LIN~92%

- Checking the trained NN with a dataset of sure not AGN just 12.6% are false positive
- False positive surely not AGN (according BoK) are 0.89%

Globular Cluster Classification

NGC1399 Dataset

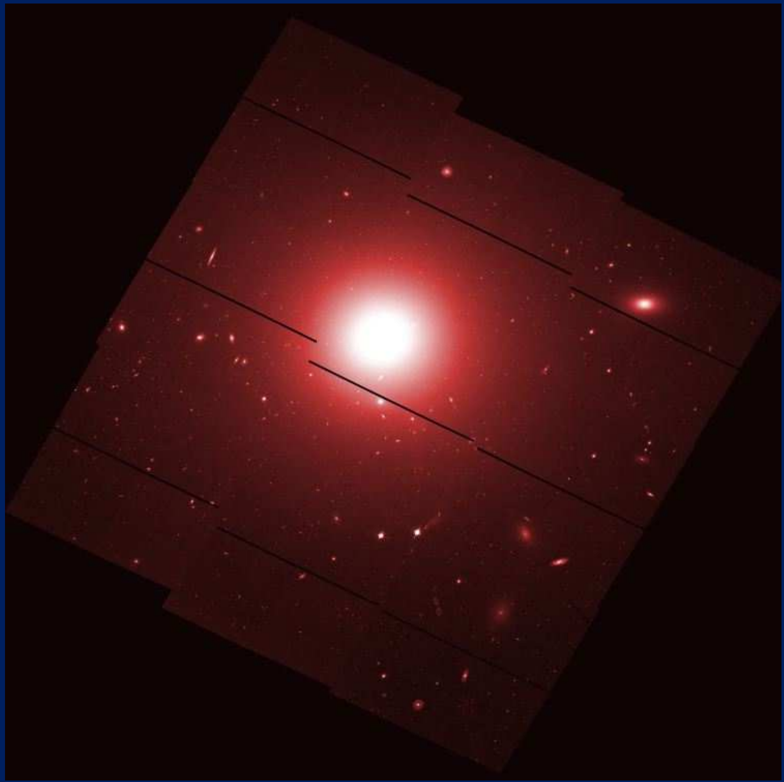
NGC1399 (~20 Mpc) is an ideal target because allows to probe a large fraction of the galaxy and still resolve GC sizes.

9 HST V-band (f606w) observations, drizzled to super-Nyquist sampling the ACS PSF (2.9 pc/pix).

Chandra ACIS-I + ACIS-S

ACS *g-z* colors for central region

Ground-based *C-R* photometry for part of the sources over the whole field



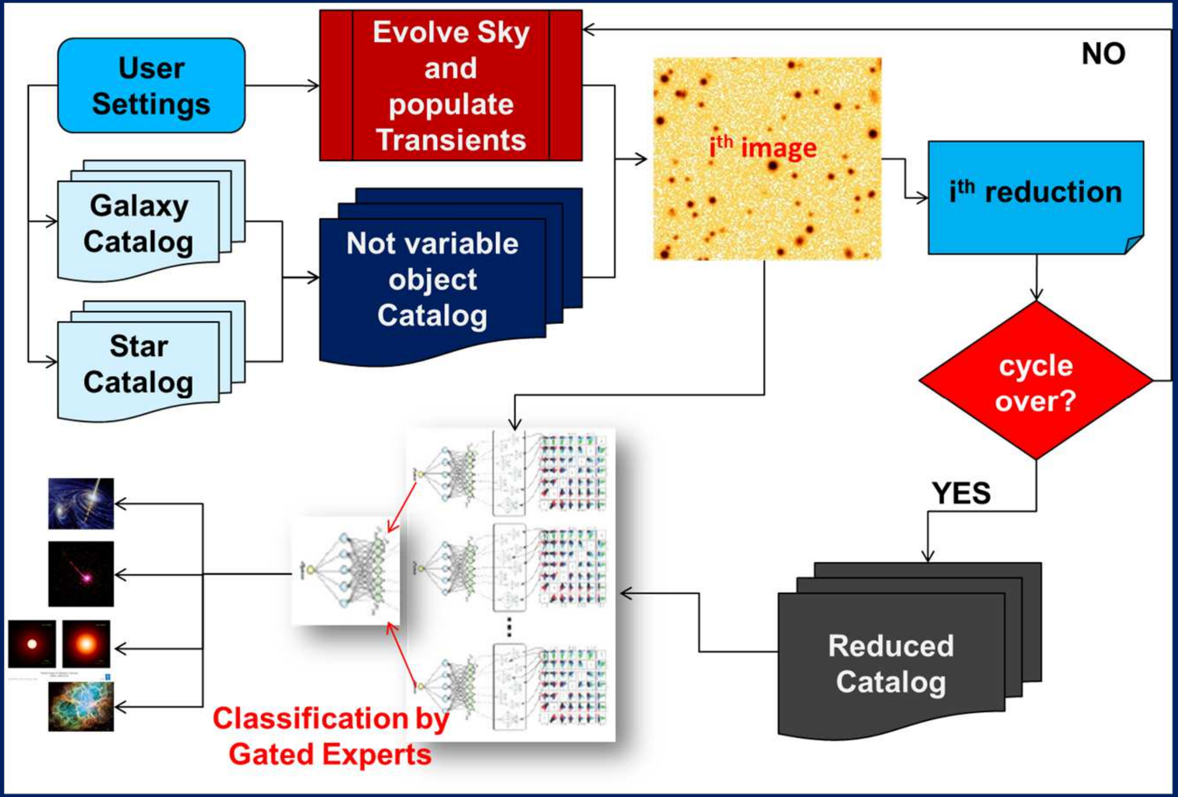
Brescia, M.; Cavuoti, S.; Paolillo, M.; Longo, G.; Puzia, T.; 2012, The detection of Globular Clusters in galaxies as a data mining problem, **MNRAS, Volume 421, Issue 2, pp. 1155-1165, available at arXiv:1110.2144v1**

Quality and pruning results

Type of experiment	Missing features	Figure of merit	MLPQNA	GAME	SVM	MLPBP	MLPGA
Complete patterns	–						
		class.accuracy	**98.3**	82.1	90.5	59.9	66.2
		completeness	**97.8**	73.3	89.1	54.1	61.4
		contamination	**1.8**	18.7	7.7	42.2	35.1
No par. 11	11						
		class.accuracy	**98.0**	81.9	90.5	59.0	62.4
		completeness	**97.6**	79.3	88.9	56.1	62.2
		contamination	**1.6**	19.6	7.9	43.1	38.8
Only optical	8, 9, 10, 11						
		class.accuracy	93.9	86.4	90.9	70.3	76.2
		completeness	91.4	78.9	88.7	54.0	65.1
		contamination	5.9	13.9	8.0	33.2	24.6
Mixed	5, 8, 9, 10, 11						
		class.accuracy	94.7	86.7	89.1	68.6	71.5
		completeness	92.3	81.5	88.6	52.8	63.8
		contamination	5.0	16.6	8.1	37.6	30.1

- **isophotal magnitude** (feature 1);
- **3 aperture magnitudes** (features 2–4) obtained through **circular apertures of radii 2**, **6** and **20 arcsec**, respectively;
- **Kron radius**, **ellipticity** and the **FWHM** of the image (features 5–7);
- **4 structural parameters** (features 8–11) which are, respectively, the **central surface brightness**, the **core radius**, the **effective radius** and the **tidal radius**;

STraDiWA



Prototipation of a web tool (**STraDiWA**, *Sky Transient Discovery Web Application*) for detection and classification of transients from simulated images.

The pipeline includes an automatic system for the extraction of the catalogues from syntetic images.

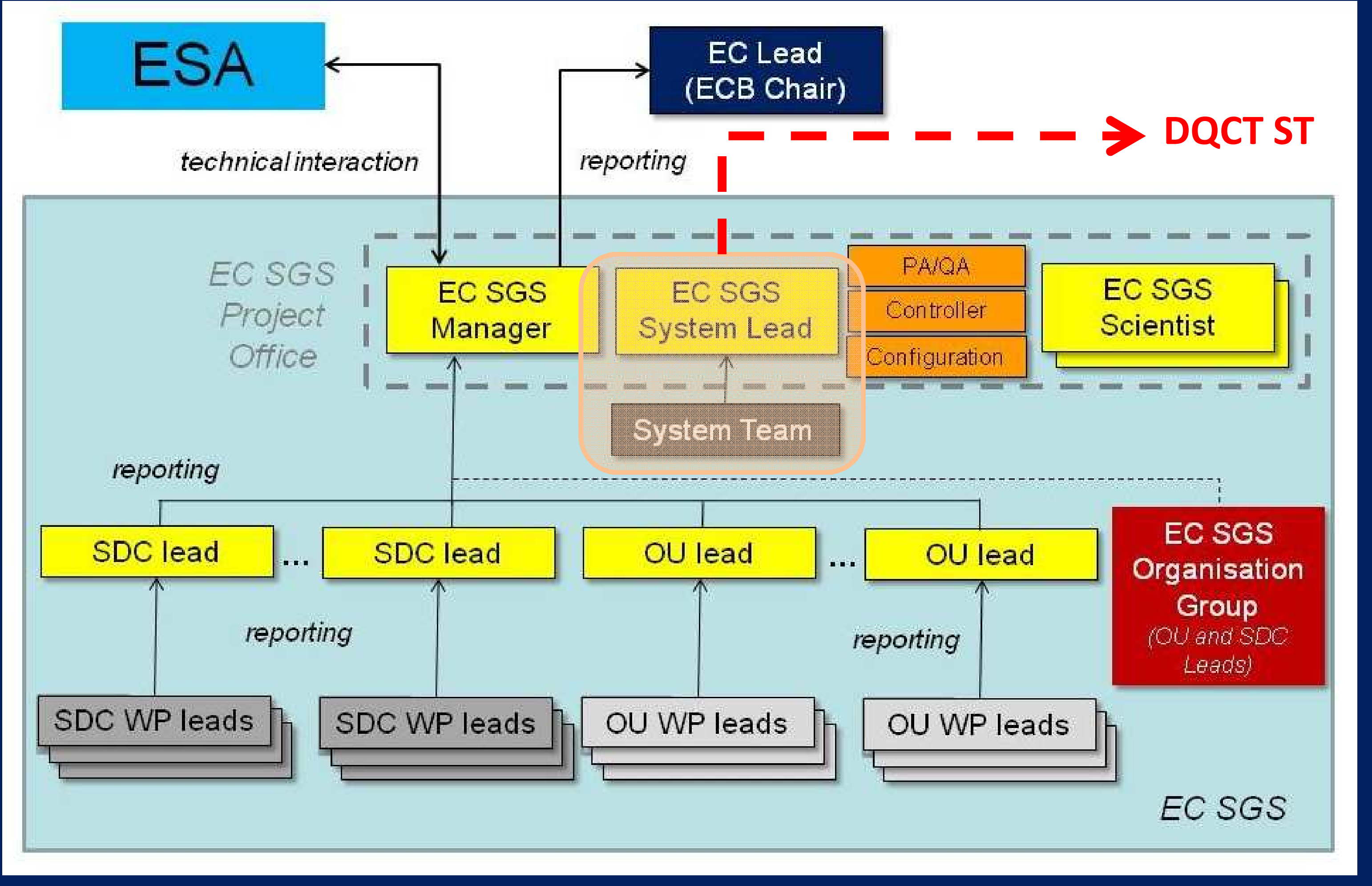
Modeling of transients, Cepheids and Supernovae Ia

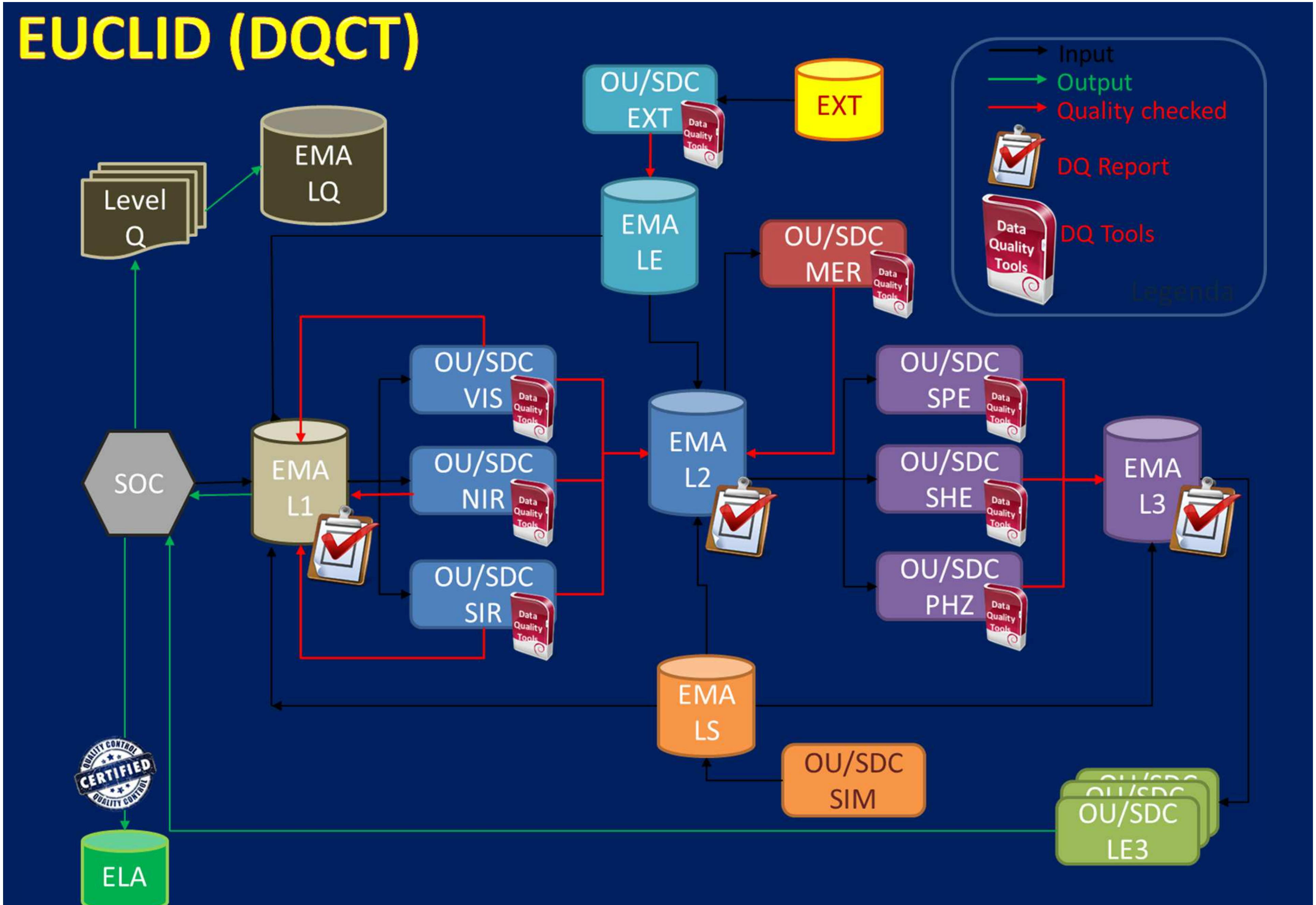
Annunziatella, M.; Mercurio, A.; Brescia, M.; **Cavuoti, S.**; Longo, G, "*Inside catalogs: a comparison of source extraction software*", **2013**, **Accepted by PASP (in press)**, p. 20

EUCLID (DQCT)

In the Euclid project, I'm involved, since Jan 2012 in two tasks:

Science Team (Italy, Norway and Finland) for the design and development of Data Quality Common Tools

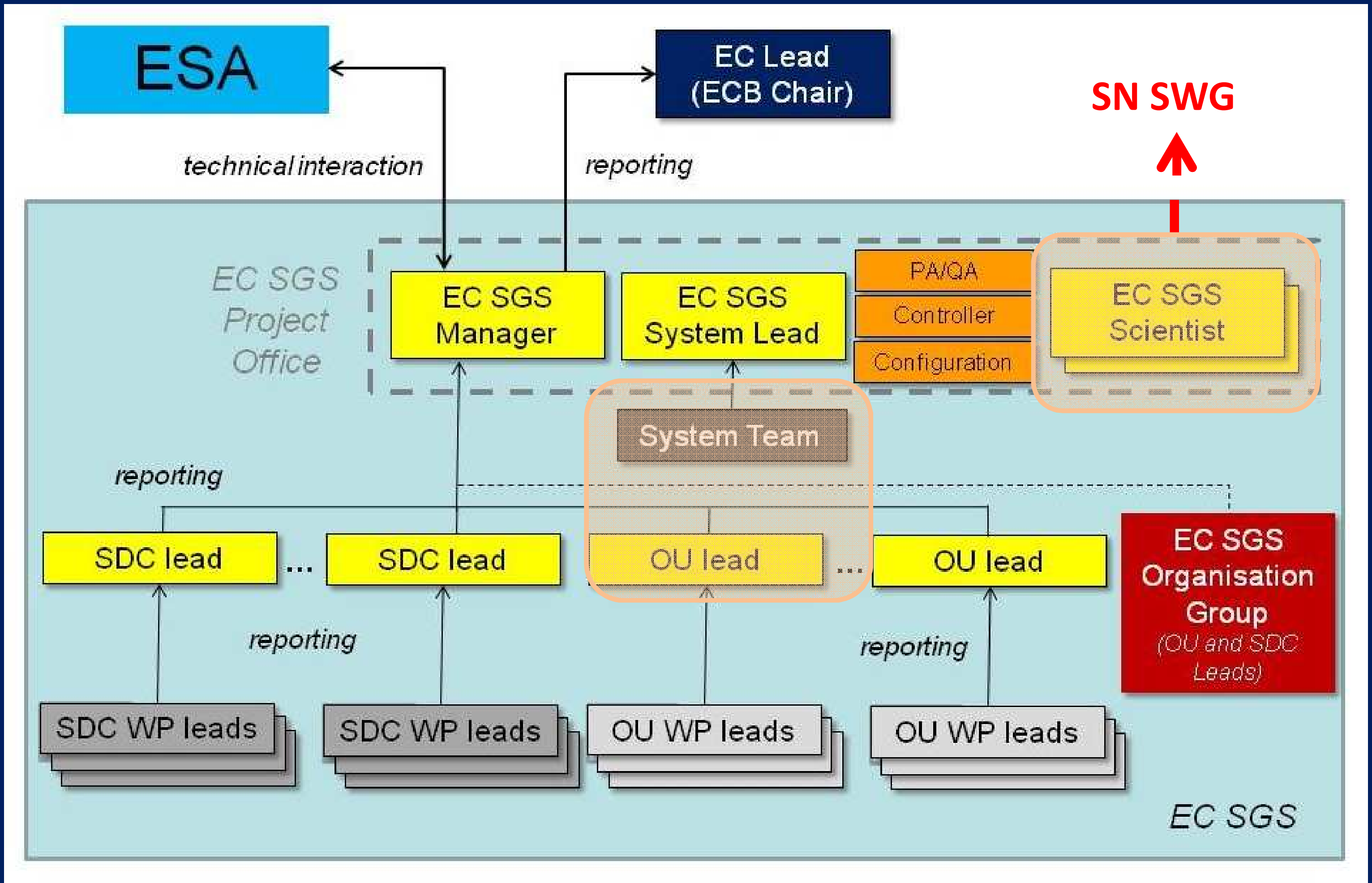




"*Data Quality in Euclid*", Euclid Consortium Scientific Ground Segment document code EUCL-OAC-SGS-TN-00085 (ESA EUCLID Official Archive)

EUCLID (SN)

Science Working group for the Legacy Science requirements definitions dedicated to transient objects detection and classification.



"*Requirements for Supernovae and Transients*", chapter of Euclid Legacy Requirements Document, Euclid Consortium Scientific Ground Segment
document code EUCL-LEI-SGS-REQ-00269 (ESA EUCLID Official Archive)

Publications I - Refeered Papers

Technological

1. Brescia, M.; Cavuoti, S.; Garofalo, M.; Guglielmo, M.; Longo, G.; Nocella, A.; Riccardi, S.; Vellucci, C.; Djorgovski, G.S.; Donalek, C.; Mahabal, A. Data Mining in Astronomy with DAME. **to be Submitted to PASP**

Algorithmic

2. Cavuoti, S.; Garofalo, M.; Brescia , M.; Paolillo, M.; Pescape', A.; Longo, G.; Ventre, G.; GPUs for astrophysical data mining. A test on the search for candidate globular clusters in external galaxies, **New Astronomy (Accepted, in press)**

Scientific

3. Cavuoti, S.; Brescia, M.; D'Abrusco, R.; Longo, G.; Photometric AGN Classification in the SDSS with Machine Learning Methods **to be Submitted to MNRAS**
4. Brescia, M.; Cavuoti, S.; D'Abrusco, R.; Longo, G.; Mercurio, A.; 2012, Photo-z prediction on WISE-GALEX-UKIDSS-SDSS Quasar Catalogue, based on the MLPQNA model, **to be Submitted to MNRAS**
5. Annunziatella, M.; Mercurio, A.; Brescia, M.; Cavuoti, S.; Longo, G.; 2012, Inside catalogs: a comparison of source extraction software, **PASP (Accepted, in Press)**
6. Cavuoti, S.; Brescia, M.; Longo, G.; Mercurio, A.; 2012, Photometric Redshifts with Quasi Newton Algorithm (MLPQNA). Results in the PHAT1 Contest, **A&A, Vol. 546, A13, pp. 1-8**
7. Brescia, M.; Cavuoti, S.; Paolillo, M.; Longo, G.; Puzia, T.; 2012, The detection of Globular Clusters in galaxies as a data mining problem, **MNRAS, Volume 421, Issue 2, pp. 1155-1165, available at arXiv:1110.2144v1**
8. Brescia, M.; Cavuoti, S.; Longo, G., Photometric Redshifts for all galaxies in the SDSS DR9 with the MLPQNA method", in preparation, **to be submitted to A&A**

Publications II - Proceedings

1. **Cavuoti, S.**; Brescia, M.; Longo, G., 2012, Data mining and Knowledge Discovery Resources for Astronomy in the Web 2.0 Age, Proceedings of SPIE Astronomical Telescopes and Instrumentation 2012, Software and Cyberinfrastructure for Astronomy II, Ed.(s): N. M. Radziwill and G. Chiozzi, Volume 8451, RAI Amsterdam, Netherlands, July 1-4 **refeered proceeding**
2. **Cavuoti, S.**; Garofalo, M.; Brescia, M.; Pescape', A.; Longo, G.; Ventre, G., Genetic Algorithm Modeling with GPU Parallel Computing Technology" in "Neural Nets and Surroundings, Smart Innovation, Systems and Technologies", Vol. 19, p. 11, Springer **refeered proceeding**
3. Brescia, M., Cavuoti, S., Djorgovski, G.S., ,Donalek, C., Longo, G.,,Paolillo, M., "Extracting knowledge from massive astronomical data sets", 2012, in "Astrostatistics and Data Mining", Springer Series in Astrostatistics, Volume 2, Springer Media New York, ISBN 978-1-4614-3322-4 **volume contribute**
4. Brescia M., Cavuoti S., D'Abrusco R., Laurino O., Longo G. "DAME: A distributed data mining and exploration framework within the Virtual Observatory",, 2011, in "Remote Instrumentation for eScience and Related Aspects", F. Davoli et al. (eds.), Springer Science+Business Media, LLC 2011, ISBN 978-1-4614-0508- **volume contribute**
5. Brescia M., **Cavuoti, S.**, Djorgovski, G.S., ,Donalek, C., Longo, G., Paolillo, M., 2011, Extracting knowledge from massive astronomical data sets, arXiv:1109.2840, to appear in Astrostatistics and data mining in large astronomical databases, L.M. Barrosaro et al. eds, Springer Series on Astrostatistics, 15 pages **invited review**.
6. **Cavuoti, S.**; Brescia, M.; Longo, G.; Garofalo, M.; Nocella, A.; 2012,DAME: A Web Oriented Infrastructure for Scientific Data Mining and Exploration,Science - Image in Action. Edited by Bertrand Zavidovique (Universite' Paris-Sud XI, France) and Giosue' Lo Bosco (University of Palermo, Italy) . Published by World Scientific Publishing Co. Pte. Ltd., 2012. ISBN 9789814383295, pp. 241-247
7. Djorgovski, S. G.; Longo, G., Brescia, M., Donalek, C., **Cavuoti, S.**, Paolillo, M., D'Abrusco, R., Laurino, O., Mahabal, A., Graham, M., DAta Mining and Exploration (DAME): New Tools for Knowledge Discovery in Astronomy. American Astronomical Society, AAS Meeting #219, #145.12, Tucson, USA, January 08-12
8. Brescia, M.; **Cavuoti, S.**; D'Abrusco, R.; Laurino, O.; Longo, G.; 2010, DAME: A Distributed Data Mining & Exploration Framework within the Virtual Observatory, INGRID 2010 Workshop on Instrumenting the GRID, Poznan, Poland, in Remote Instrumentation for eScience and Related Aspects, F. Davoli et al. (eds.), Springer Science+Business Media, LLC 2011, DOI 10.1007/978-1-4614-0508-5 17
9. Brescia, M.; Longo, G.; Castellani, M.; **Cavuoti, S.**; D'Abrusco, R.; Laurino, O., 2012, DAME: A DistributedWeb Based Framework for Knowledge Discovery in Databases, 54th SAIT Conference, Astronomical Observatory of Capodimonte, Napoli, Italy, May 6, Mem. S.A.It. Suppl. Vol. 19, 324

MDS with: $N > 10^9$, $D \gg 100$, $K > 10$

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 Maximization: $K \times N \times I \times \mathbf{D^2}$
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**

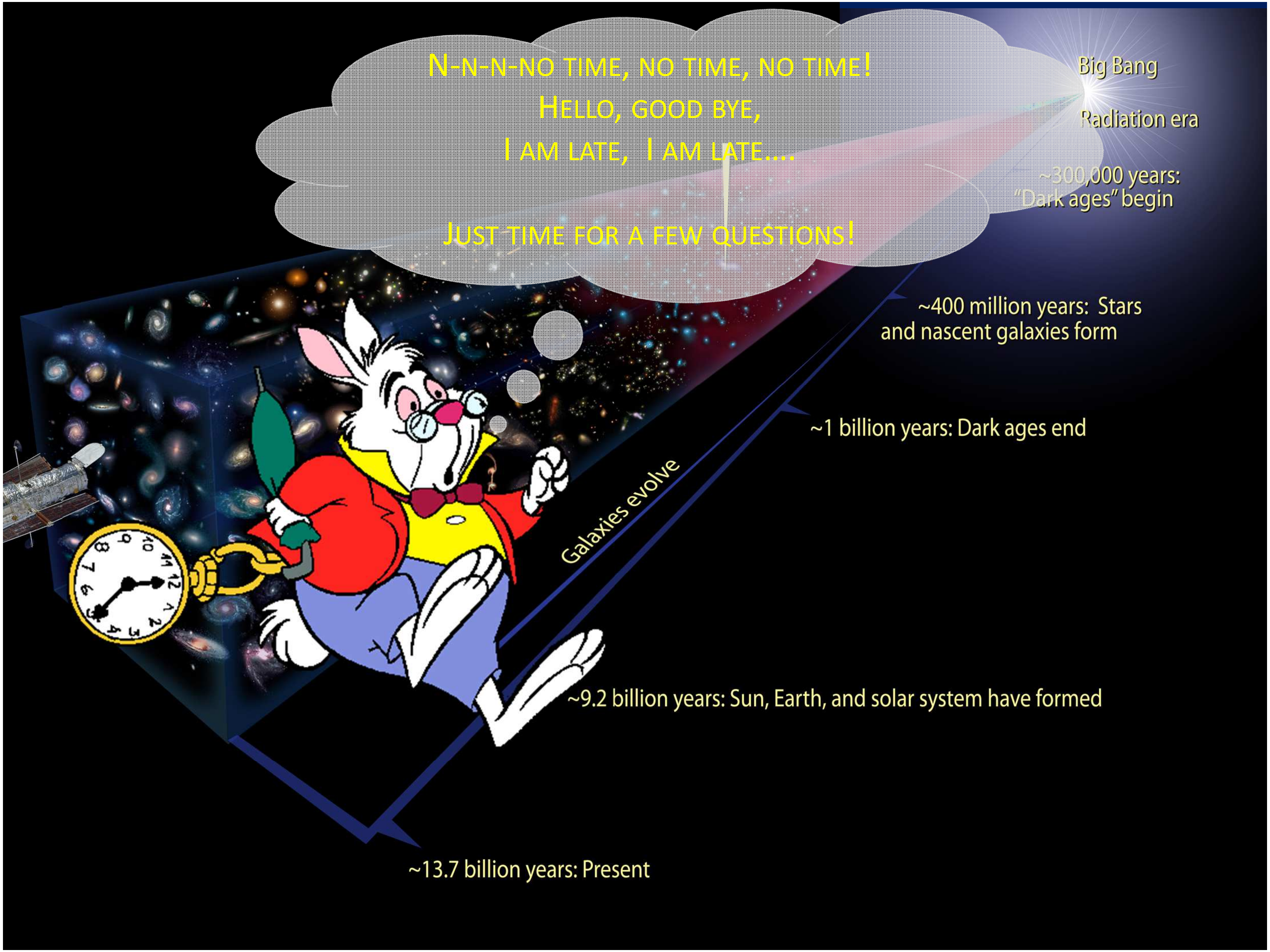
Conclusions, in the middle of the white Rabbit Hole…

Well, in conclusion we have not yet concluded, actually just started…

We obtained a lot of great results about redshifts and about the other issue, but this is not the core of this talk.

For the Red Pills consumers: YES

Astroinformatics is opening a new wide and encouraging door, and a new era of observational Astronomy has started


N-N-N-NO TIME, NO TIME, NO TIME!
HELLO, GOOD BYE,
I AM LATE, I AM LATE....
JUST TIME FOR A FEW QUESTIONS!
Big Bang
Radiation era
~300,000 years:
"Dark ages" begin
~400 million years: Stars
and nascent galaxies form
~1 billion years: Dark ages end
Galaxies evolve
~9.2 billion years: Sun, Earth, and solar system have formed
~13.7 billion years: Present