DATA DRIVEN DISCOVERY IN ASTROPHYSICS

CENTER FOR DATA-DRIVEN DISCOVERY

M. Brescia[1], S. Cavuoti[1], S.G. Djorgovski[2], C. Donalek[2], **G. Longo[2,3]**

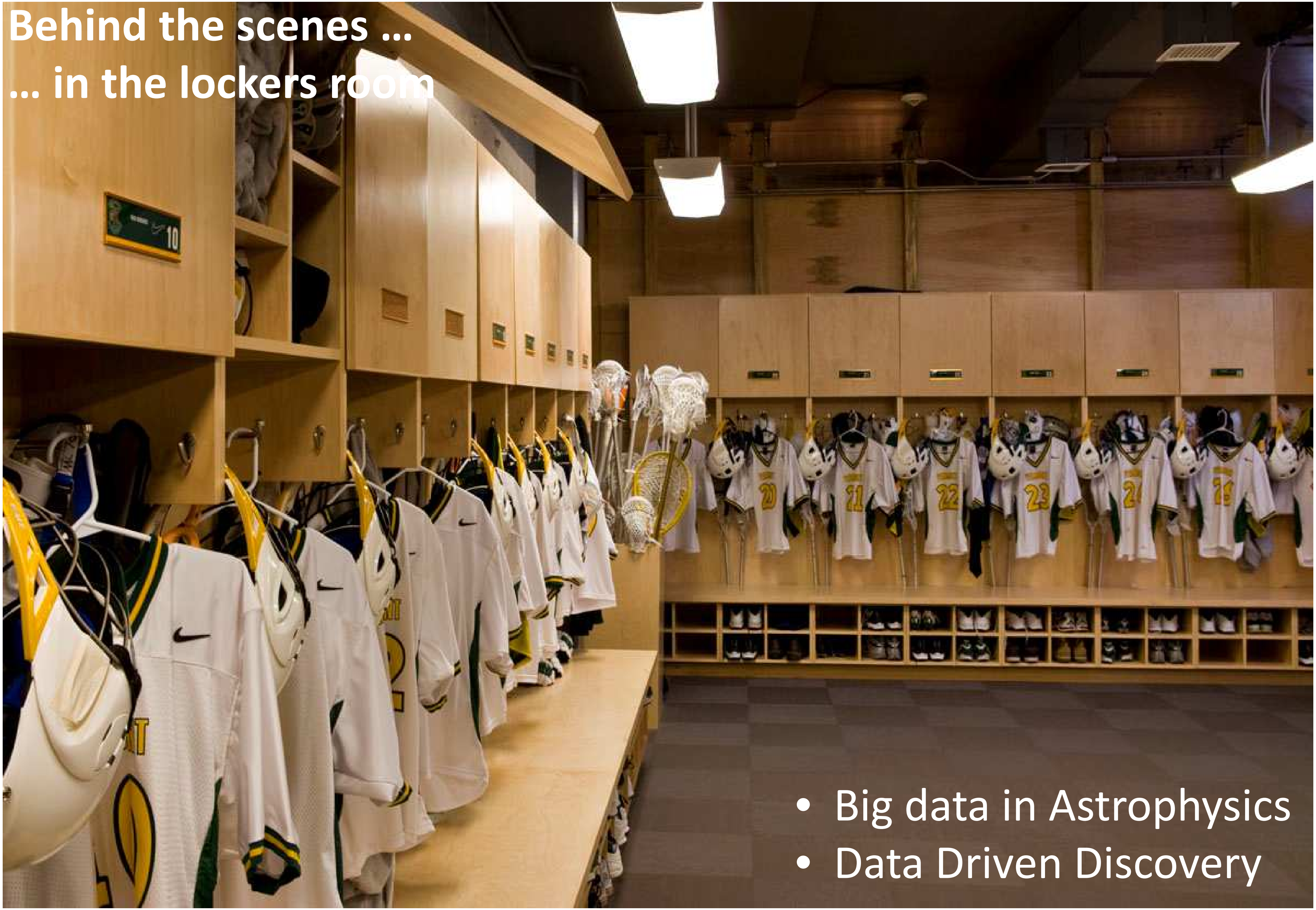
1. *INAF – Astronomical Observatory of Capodimonte (I)*
2. *CD³ – Center for Data Driven Discovery – Caltech (USA)*
3. *Department of Physics, University Federico II in Naples (I)*

longo@na.infn.it

cost TD-403

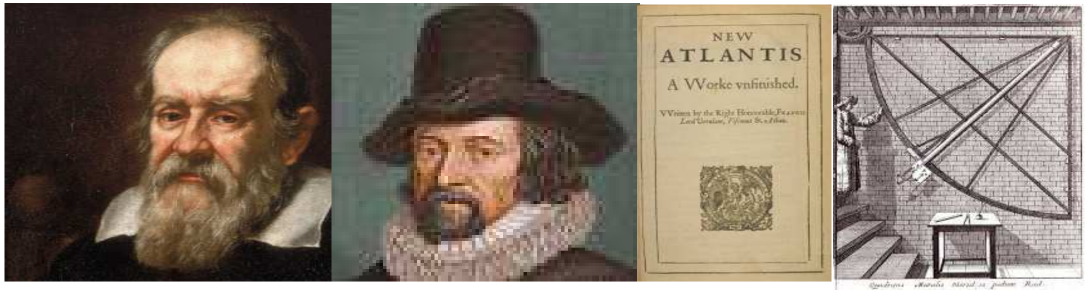
EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

DA ME


Behind the scenes …
… in the lockers room
• Big data in Astrophysics
• Data Driven Discovery

The evolving paths to knowledge

(Jim Gray)



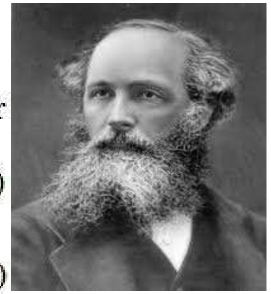
The First Paradigm
Experiments/measurements
(XVII century)

$$\nabla \cdot \mathbf{E} = \frac{\rho_v}{\varepsilon} \quad \text{(Gauss' Law)}$$

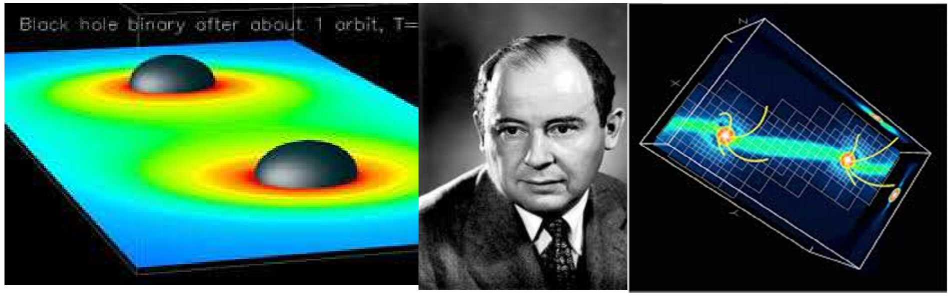
$$\nabla \cdot \mathbf{H} = 0 \quad \text{(Gauss' Law for}$$

$$\nabla \times \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t} \quad \text{(Faraday's Law)}$$

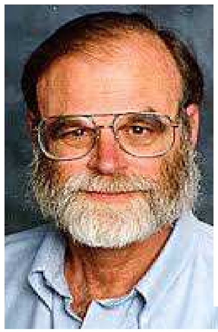
$$\nabla \times \mathbf{H} = \mathbf{J} + \varepsilon \frac{\partial \mathbf{E}}{\partial t} \quad \text{(Ampere's Law)}$$

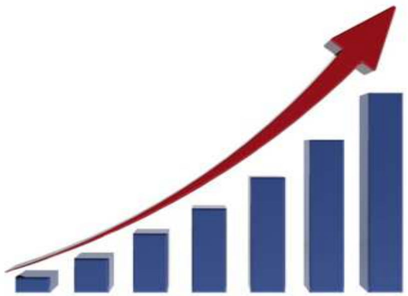


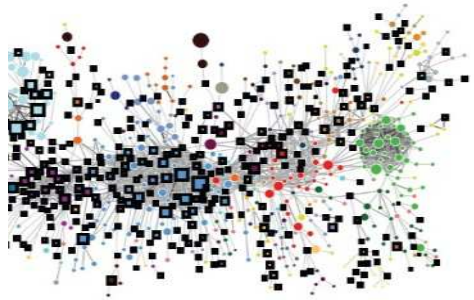
The Second Paradigm
Analytical theory
(XVIII century)



The Third Paradigm
Numerical simulations
(early 40's)

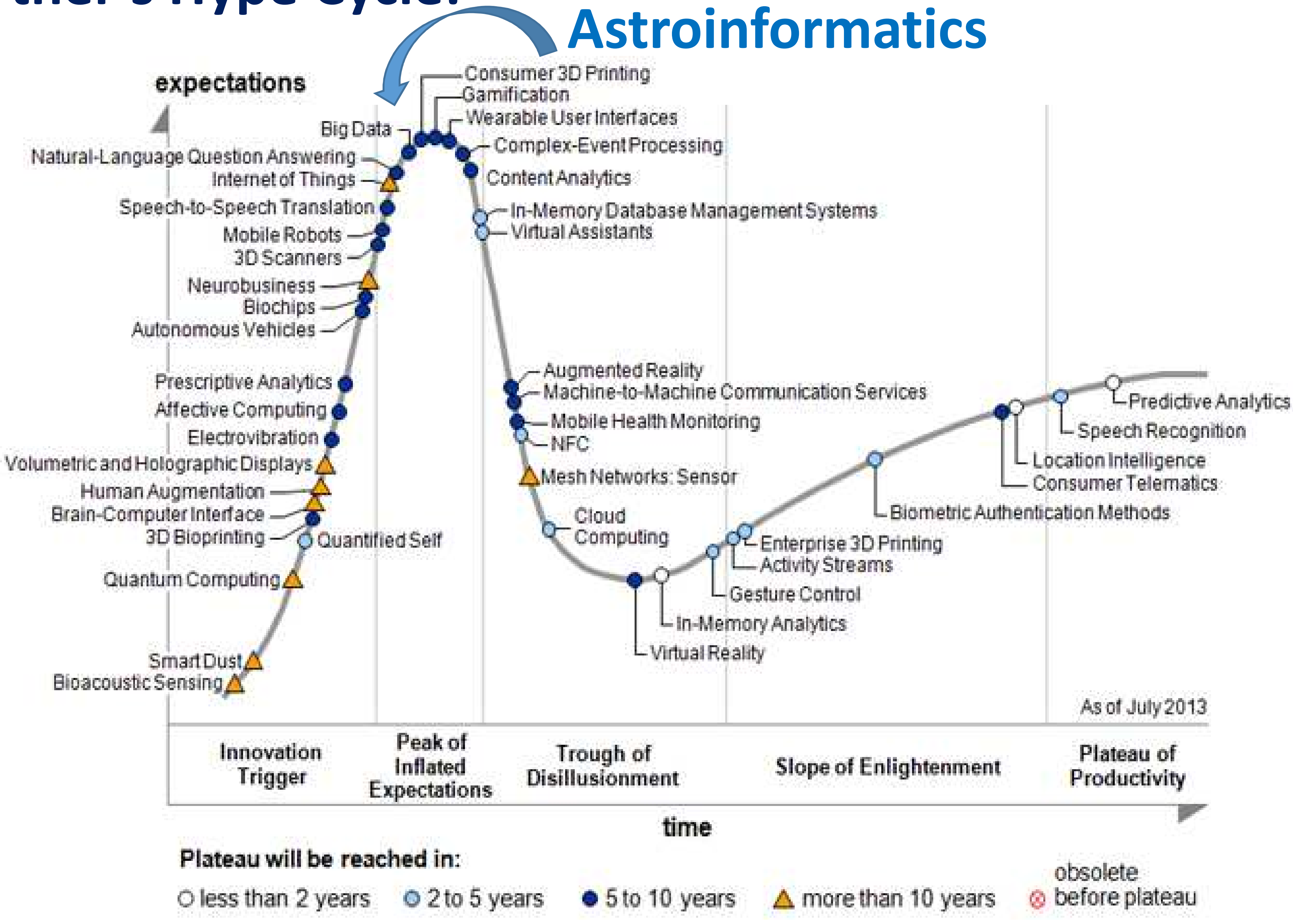






The Fourth Paradigm
Data Driven Discovery
(Now)

Gartner's Hype Cycle:



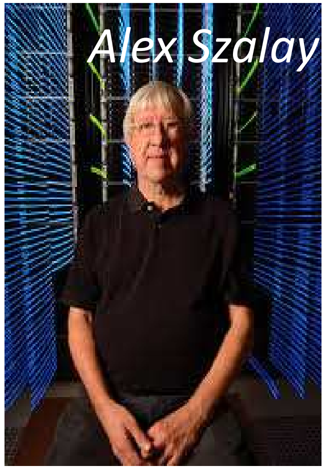
So what are «Big Data» in Astrophysics?

Big Data is like teenage sex:
Everyone talks about it,
Nobody really knows about it,
Everyone thinks everyone else is doing it,
So everyone claims they are doing it ….

But astronomers definetely do it ….

Dan Ariely

The Sloan Digital Sky Survey *(in its various incarnations)*



Sloan Digital Sky Survey – Sky Server

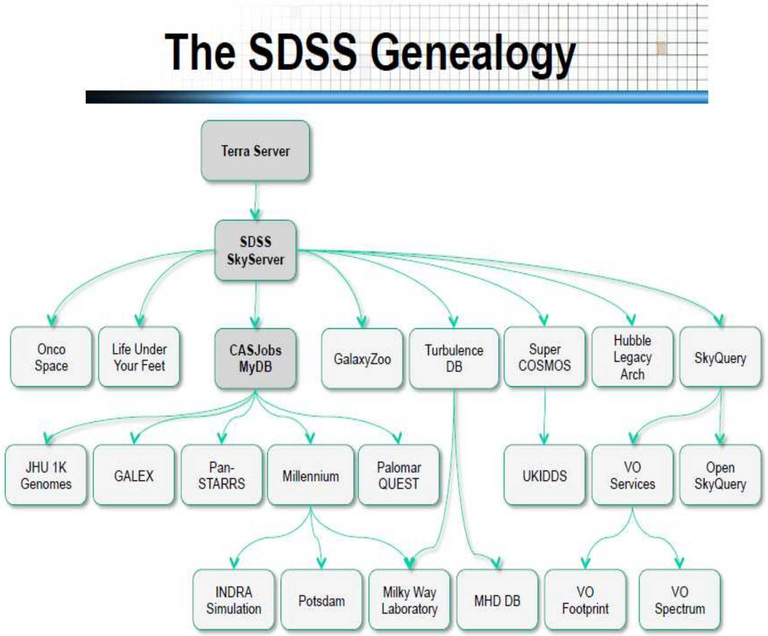
– 2.5 Terapixels of images => 5 Tpx of sky; 10 TB of raw data => 400TB processed; 0.5 TB catalogs => 35TB final

… a Prototype in 21st Century data access

– 1.2B web hits in 12 years; 200M external SQL queries; 4,000,000 distinct users vs. 15,000 astronomers

Data products (e.g. **SPECTROSCOPIC and PHOTOMETRIC** catalogues) and raw data were «immediately» made available to the community

Courtesy of Alex Szalay



The right data set at the right moment

Pioneeristic yet manageable with available technology (10 TB of data products); general in purpose, flexible enough to be useful for a large variety of existing problems, yet capable to rise new ones

The SDSS data set

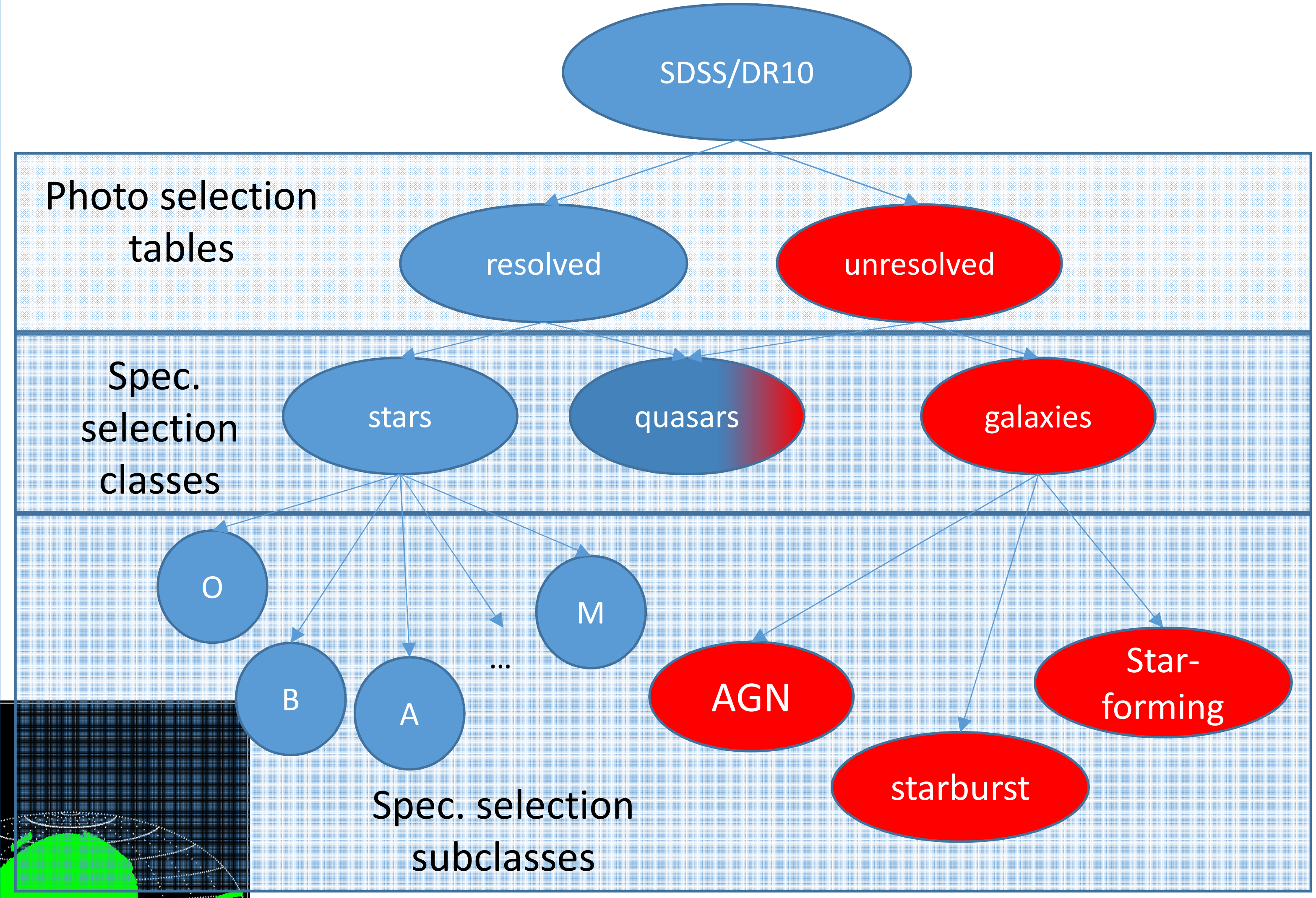
Photometric

Hundreds of features for
300M galaxies and stars
Quality flags

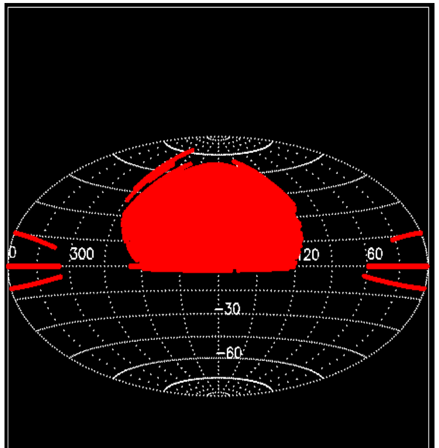
Spectr. subsample

(ca. 3 Mobjects)

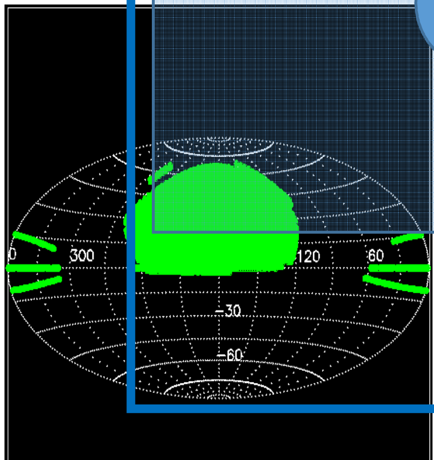
Equivalent widths
Spectroscopic redshifts
Spectral ckassification in
classes and subclasses



Spectroscopic Knowledge Base



DR7 – photo coverage



DR7 – spectral coverage

Name	bands	Area (sq. Deg)	KB's	epochs	Size/access
SDSS	Optical (5)	25.000	yes	1	20/2 Tbyte yes
KIDS	Visible (4)	1.500	Yes /no	1	20/2 TB Yes del.
VIKING	IR (5)	1.500	Yes/no	1	20/2 TB Yes. Del.
CRTS	Optical (1)	33.000 (1)	Yes	>100	100 TB growing yes
EUCLID	Optical/NIR	10.000	Yes	1	>150 PB Yes del
LSST	Optical	15.000	Yes	>>100	15 TB/night >100 PB
SKA	Radio		Yes	>100	1.5 PB/sec

Automatic processing

Hundreds of parameters

- Morphological
- Photometric
- Epoch
- …

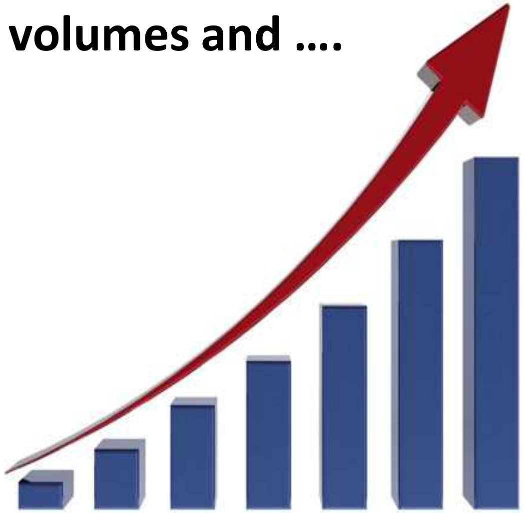
- Public access
- Real time processing
- Needs for real time automatic follow-up scheduling

Hundreds of different groups running hundreds of vastly different research projects

Technological challenges of big data:

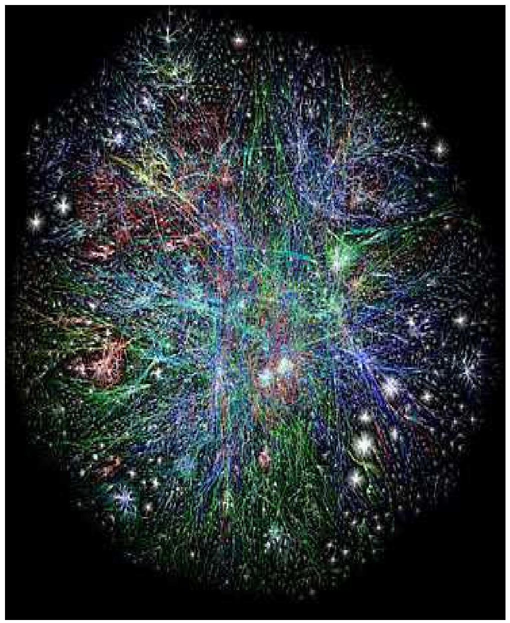
Standards, interoperability, etc, … Taken care by Virtual Observatory projects around the world

Exponential growth of Data volumes and ....



In less than a decade astronomy has moved from

... and data complexity



- From data poverty to data glut
- From data sets to data streams
- From static to dynamic, evolving data
- From anytime to real time analysis and discovery
- From centralized to distributed resources
- From ownership of data to ownership of expertise

These data sets are so large and rich that:

- No single researcher or group can exploit them *(public access)*
- It is impossible to transfer them from the data centers to the final user *(move programs and not the data)*
- Their value increases with time *(data re-use)*
- They impose an entirely different methodological approach *(Data Mining,* and, eventually

The astronomical community **needs *D³*** to scientifically exploit otherwise unmanageable datasets

But …

Does the community understand what ***D³*** is truly about?

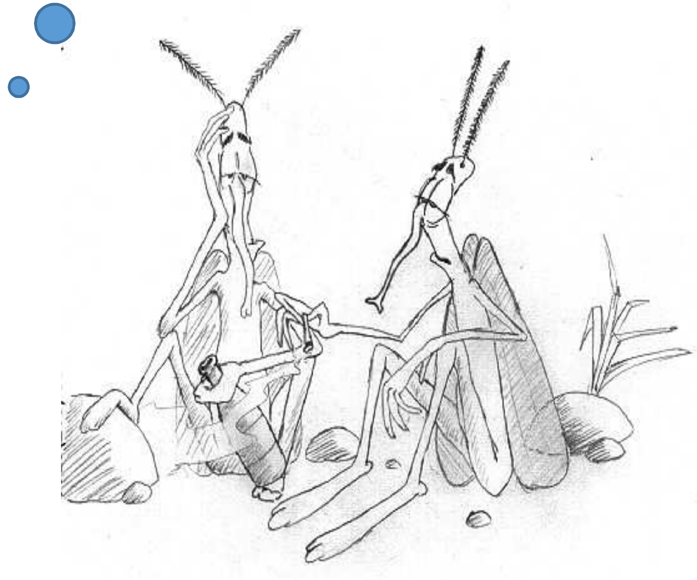
And…

Is the community ready to abandon old ways of thinking and traditional methods (*faster horses*)?

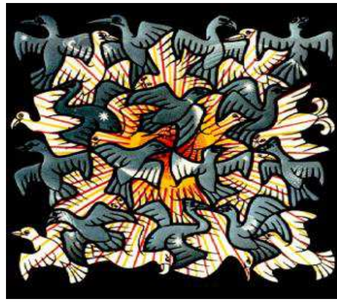
If I had asked people what they wanted, they would have said faster horses…”

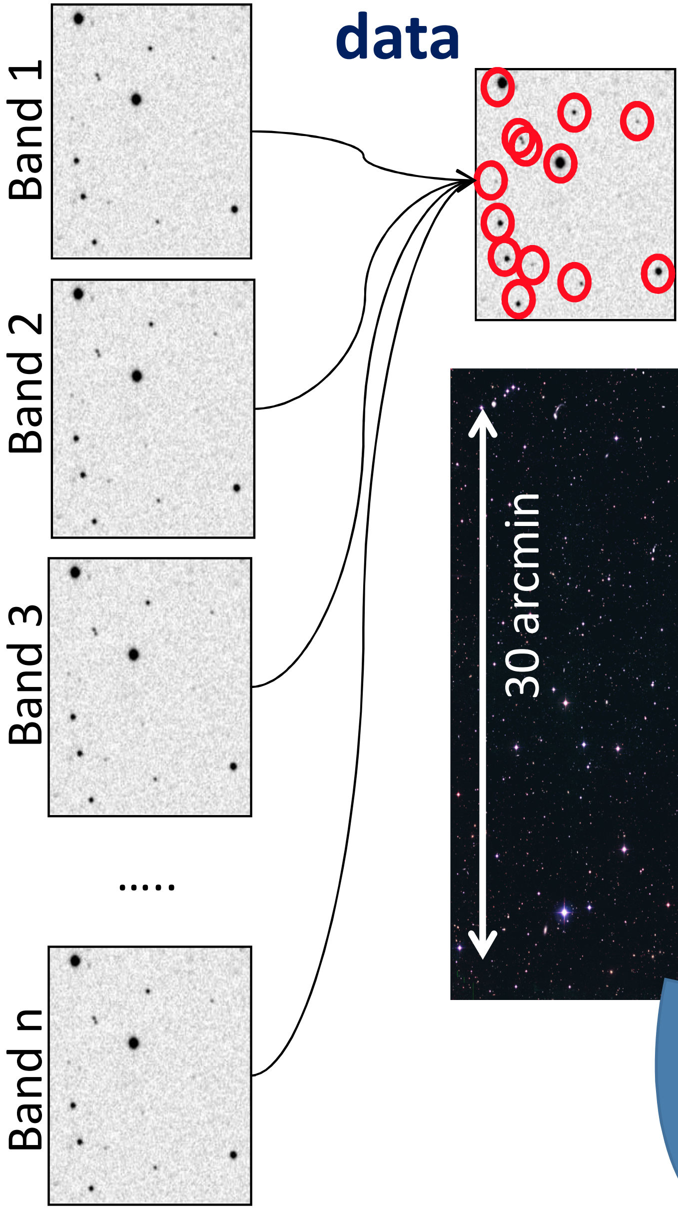
—Henry Ford

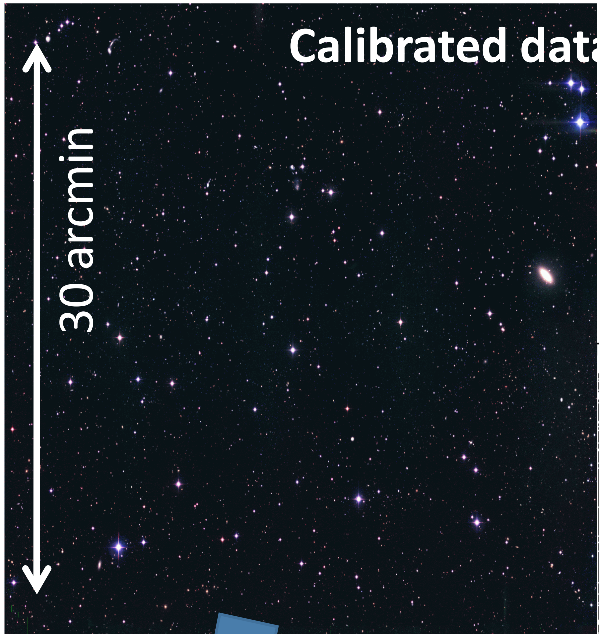
Cartoon by D. Vinkovic



From raw images to data

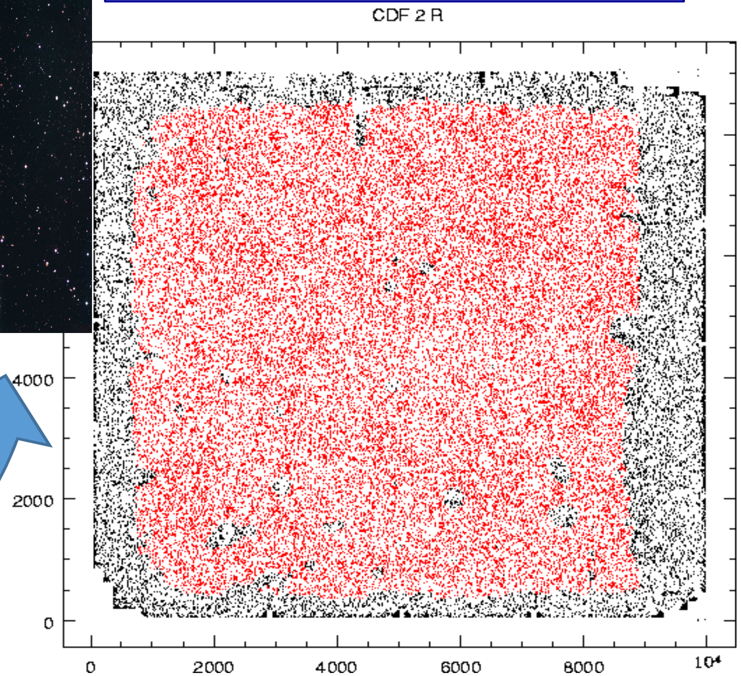




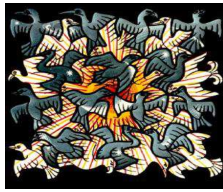


1/160.000 of the sky, moderately deep (25.0 in r)

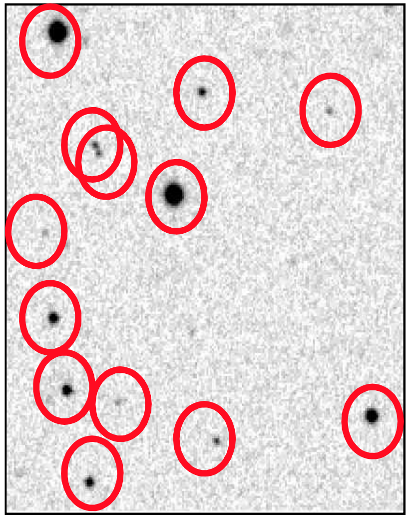
55.000 detected sources (0.75 mag above m lim)



The Data Tsunami: complexity



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



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

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

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

....

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

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

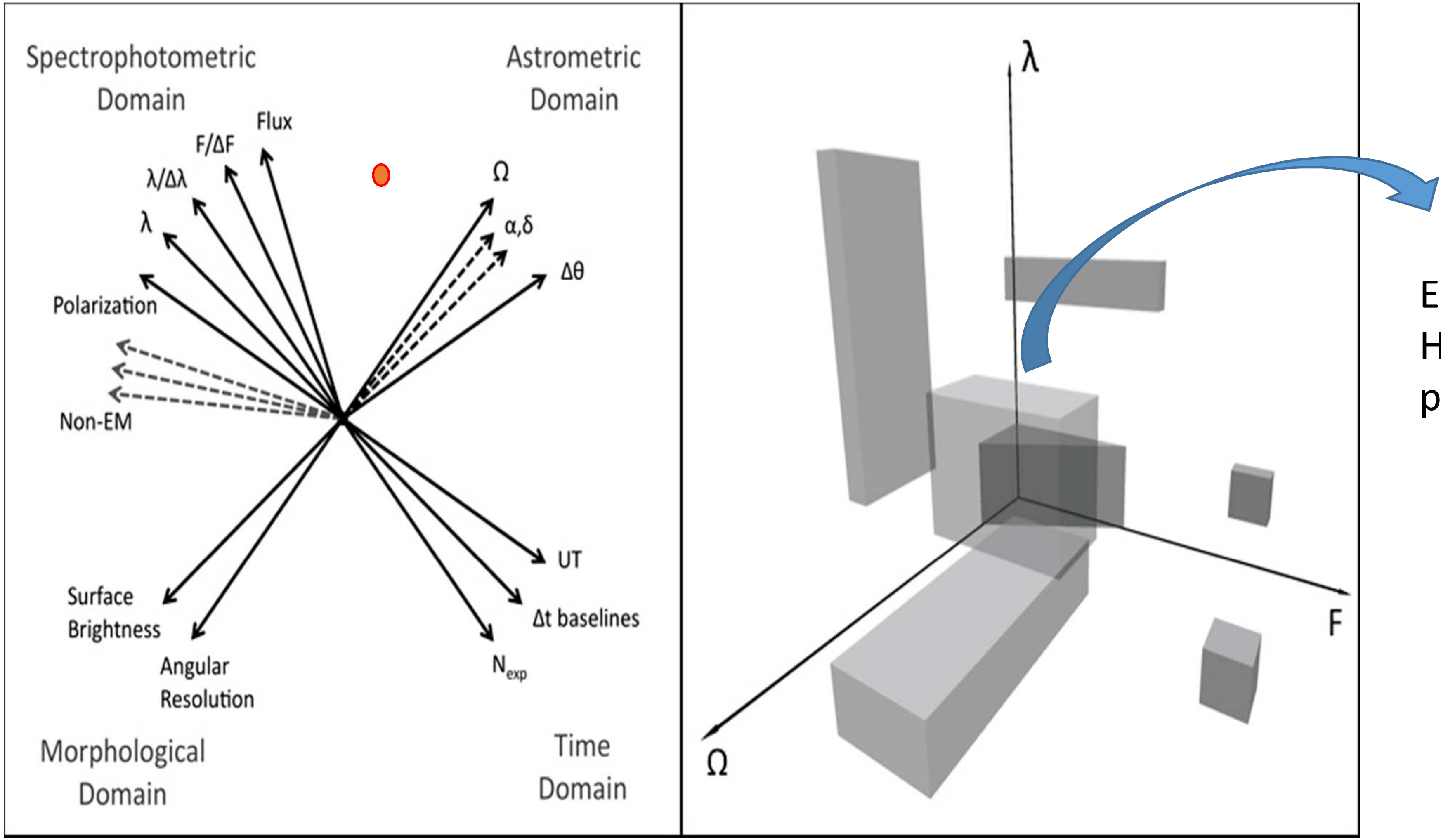
The scientific exploitation of a multi band, multiepoch (K epochs) universe implies to search for patterns, trends, etc. among N points in a DxK dimensional parameter space:

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

X-parameter spaces of very high dimensionality

$\mathbb{R}^n$

Each observation defines a point $p\{x_1, \dots, x_n\} \in \mathbb{R}^n$



Each survey carves an Hypervolume in the parameter space

DATA Mining is about rediscovering/discovering known (unknown) useful patterns in the data

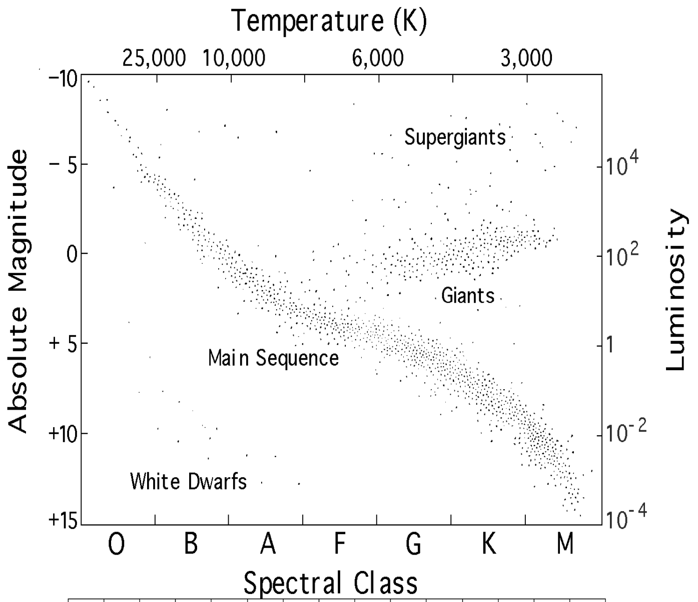
DATA DRIVEN DISCOVERY is not «simply» about machine learning

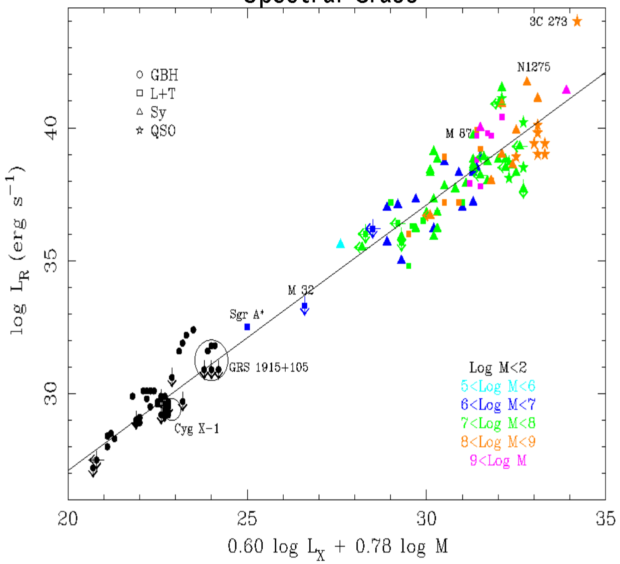
***D*^{3}** is a methodological and paradigmatic shift

$$\boldsymbol{D^3} \equiv \{data\ mining, statistical\ pattern\ recognition, visualization\}$$

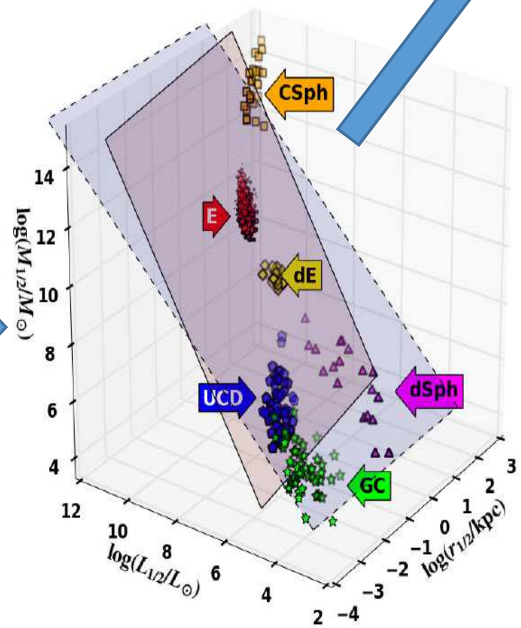
D*^{3}** is about ***letting the data to speak for themselves with minimum use of a-priori assumed models and hypothesis

3-D is an intrinsic human limitations





2-d diagnostics



3-d diagnostics

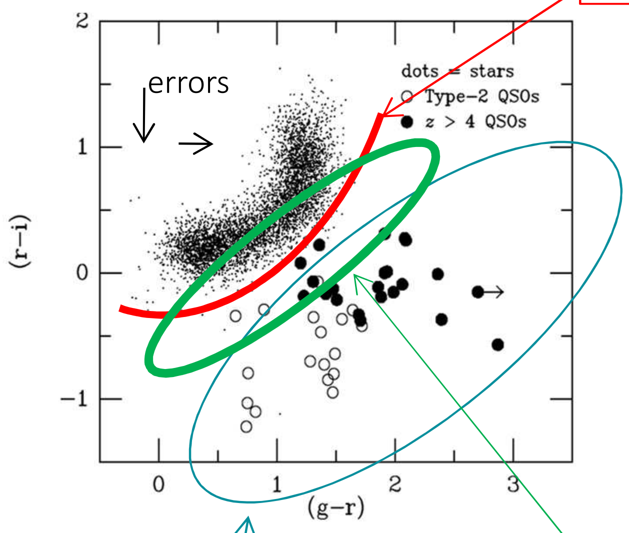
A simple universe

or rather …

… a limitation of human brain?

What should we do to extract patterns (i.e. laws ordering relationships) in a R^n space (n>>100) ?

Traditional way to look for candidate QSO in 3 band survey

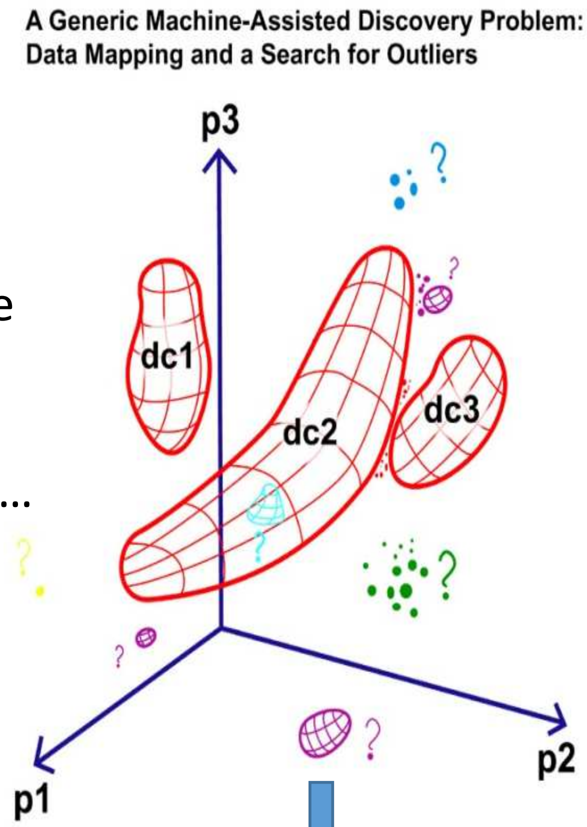


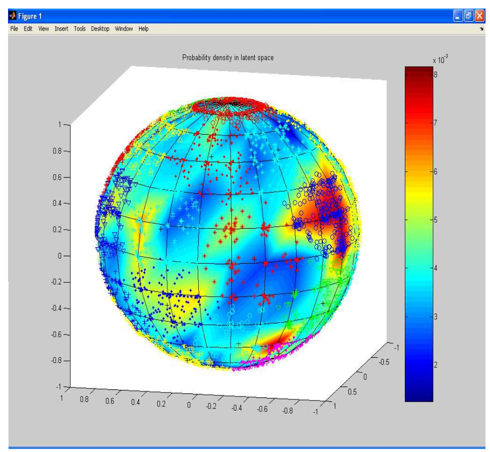
Cutoff line

Candidate QSOs for spectroscopic follow-up's

Ambiguity zone

Adding one feature improves separation…





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

And now… our playing stadium … and the team

DAMEWARE

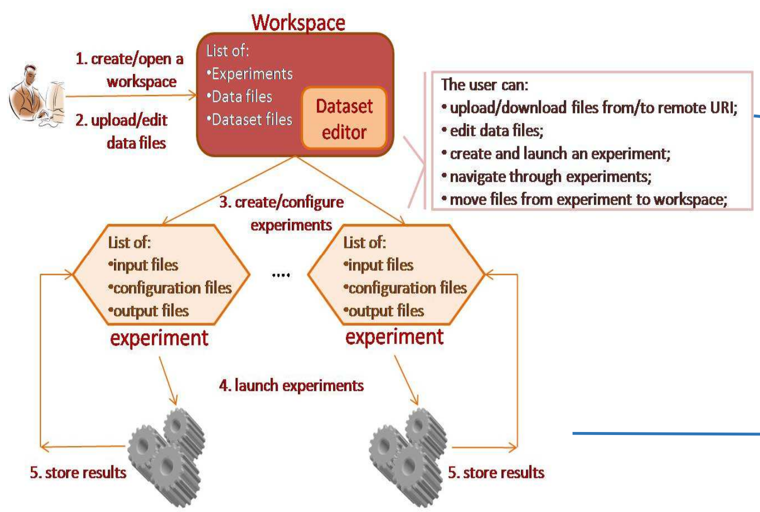
(Data Mining & Exploration
Web Application Resource)

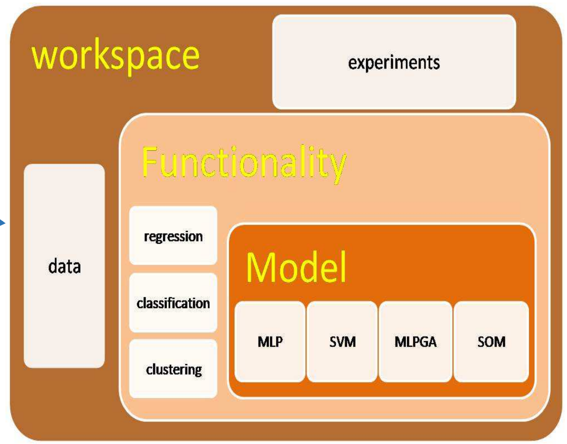
DAMEWARE (DAME Web Application REsource) v 1.0

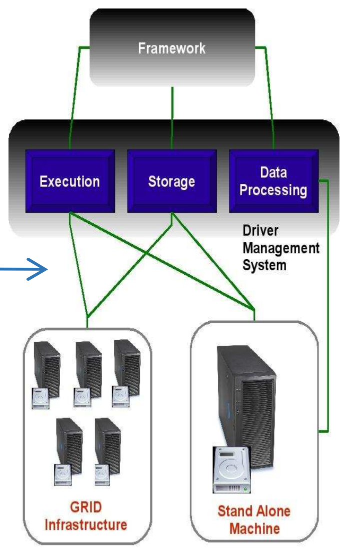
A University Federico II, INAF-OACN & Caltech effort, recently joined by ITHS of Heidelberg, aimed at implementing a science gateway for data exploration on top of a virtualized distributed computing environment. **It is multi-disciplinary platform (astronomy, bioinformatics and medical diagnostics)**

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







First phase ended in 2012

DAMEWARE is a part of the DAME project

Is a 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, recently joined by ITHS of Heidelberg, aimed at implementing (as web 2.0 apps and services) a science gateway for data exploration on top of a virtualized distributed computing environment

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

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

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

Classification

Regression

Clustering

Feature selection

DM models

GAME S, C,R
MLPBP S, C,R
MLPGA S, C,R
MLPQNA S, C,R
SVM S, C,R
K-Means U, Cl
ESOM U, Cl
SOFM U, Cl
SOM U, Cl
PPS U, Cl, FS

Experiments

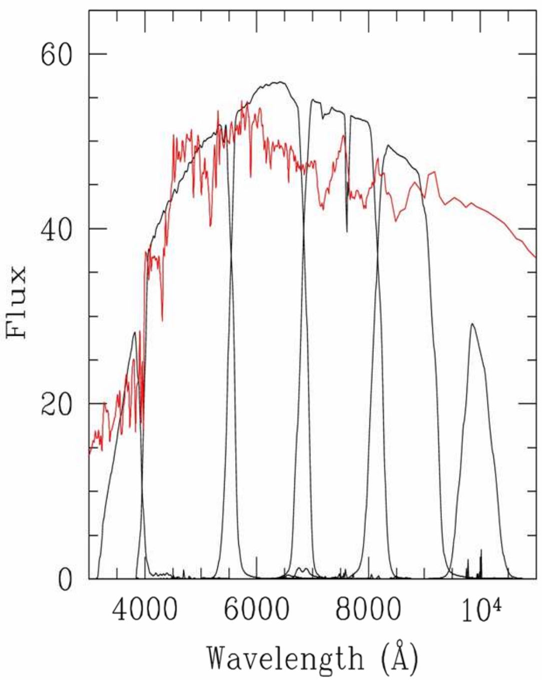
1-st
2-nd
3-rd
4-th
....

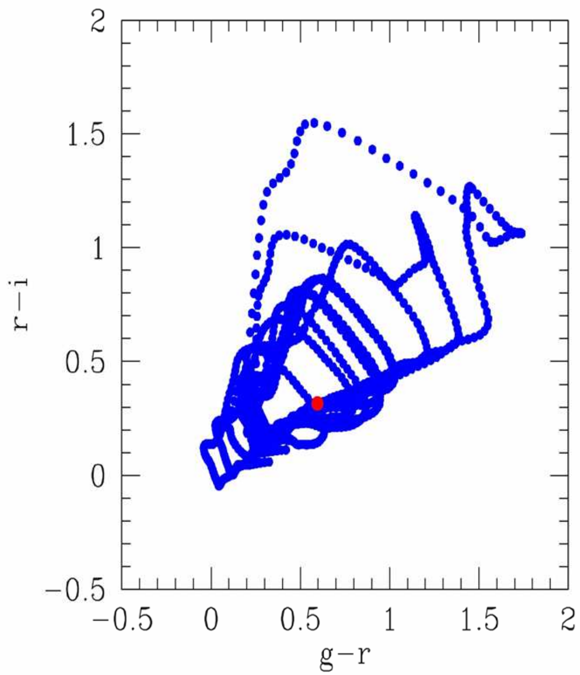
N-th

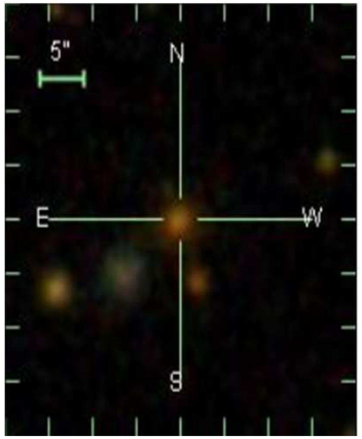

Then … let's play a game
…
Photometric redshifts vs Spectroscopic redshifts
GETTING READY FOR EUCLID

A template case of …. machine learning vs «pure» *D*³
Photometric redshifts for quasars and galaxies

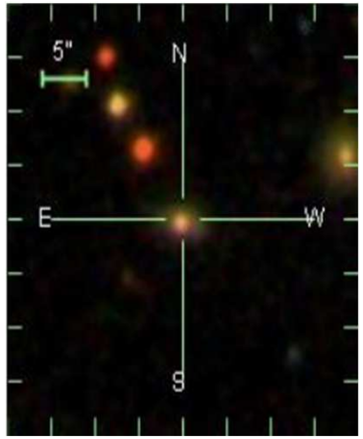
$$1 + z = \frac{\lambda_{obs}}{\lambda_0} \approx \frac{v}{c}$$







QSO; z=3.81



QSO; z=5.31

Only viable way to obtain distance info's for large samples of galaxies

Crucial cosmological probe

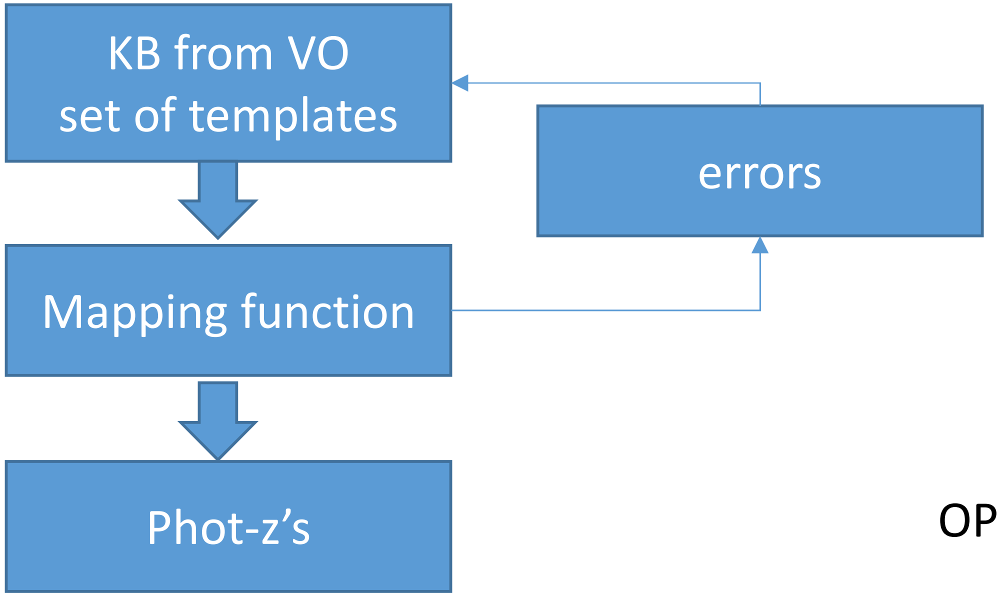
- Large scale structure
- Weak lensing
- Tests of cosmological models

Mathematically simple: to find the mapping function

Input vector $\left(\bar{X}_j\{x_1, \dots, x_n\} j = 1, \dots m\right) \in OPPS \subset \mathbb{R}^n$

Target vector $\bar{Y}\{y_1, \dots, y_m\} \in OPPS \subset \mathbb{R}^n$

Physical redshift $\bar{Y}\{y_1, \dots, y_m\} \rightarrow \bar{Z} \in PPS \subset \mathbb{R}^n$



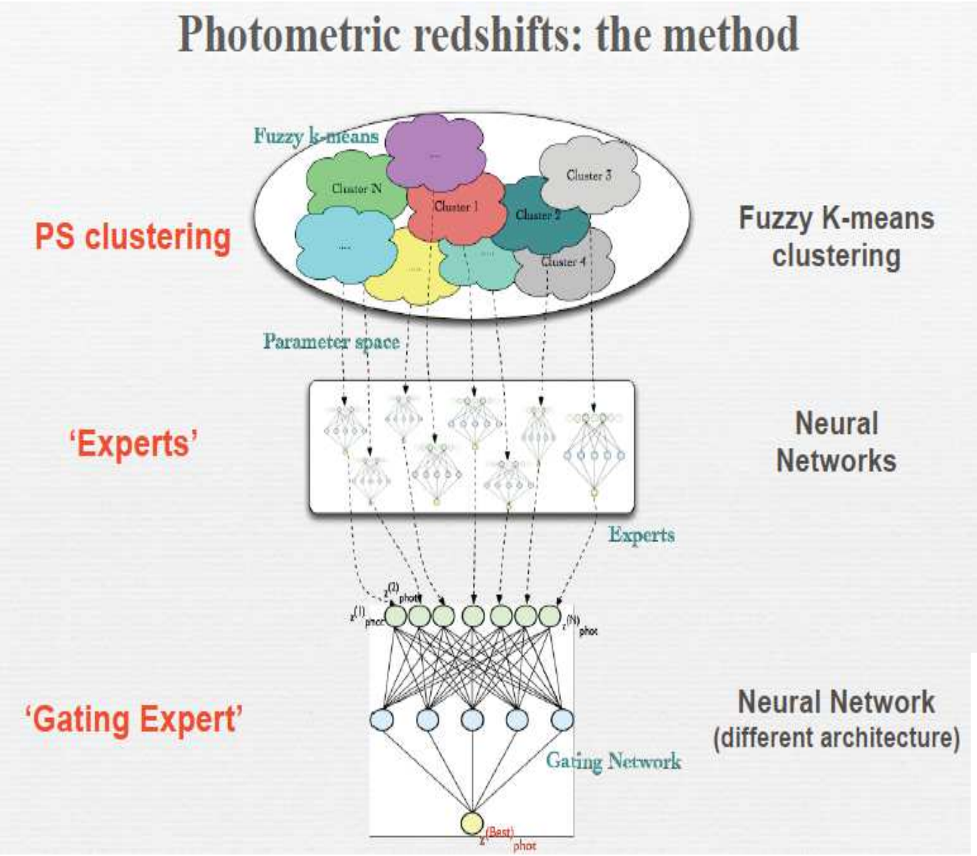
$$f(\bar{x}) \longrightarrow y, where, \bar{x} \epsilon \mathbb{R}^n, y \epsilon \mathbb{R}$$

OPPS = Observable Photometric Parameter Space

OSPS = Observable Spectroscopic Parameter Space

PPS = Physical Parameter Space

Photo-z for Quasars: first attempt (by us)

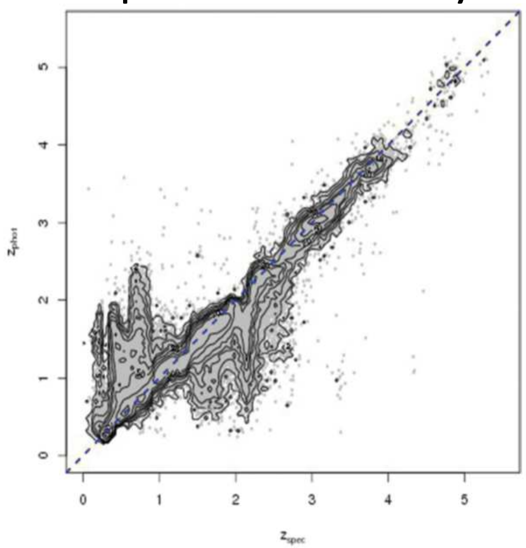


[Astroinformatics of galaxies and quasars: a new general method for photometric redshifts estimation](), O. Laurino, R. D'Abrusco, G. Longo, and G. Riccio, MNRAS, 2011, 418, 2165 (arXiv/1107.3160);

WGE: Weak Gated Expert

Data from the unresolved objects SDSS catalogue

Optical bands only



Optical + UV bands

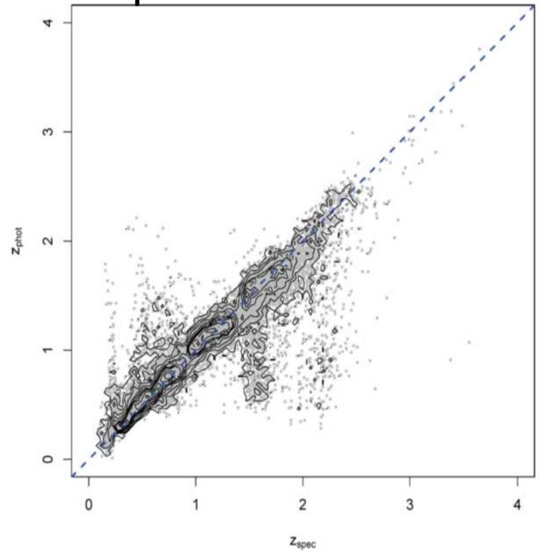


Table 5. Statistical diagnostics of photometric redshifts reconstruction for all the experiments discussed in this paper and for relevant papers in the literature. Column 'Exp. 1' contains the diagnostics for the experiment for the determination of the photometric redshifts of the optical galaxies from the SDSS catalogue described in Section 6.1, while columns 'Exp. 2' and 'Exp. 3' describe the diagnostics for the experiments concerning the determination of the photometric redshifts for quasars with optical and optical+UV photometry, respectively (the details can be found in Sections 6.2 and 6.3). The same statistical diagnostics are shown for some papers from the literature, respectively, D'Abrusco et al. (2007) for optical galaxies in column (1) and both Ball et al. (2008) and Richards et al. (2009) for optical and optical+UV quasars in the columns (2) and (3), respectively (as reported in Ball et al. 2008). The definitions of the statistical diagnostics and other relevant results of the literature are discussed in Section 8.

Diagnostic	Exp. 1	(1)	Exp. 2	(2)	(3)	Exp. 3	(2)	(3)
$\langle \Delta z \rangle$	0.015	0.021	0.21	–	–	0.13	–	–
$\mathrm{rms}(\Delta z)$	0.021	0.074	0.35	–	–	0.25	–	–
$\sigma^2(\Delta z)$	2.3×10^{-4}	5.0×10^{-4}	0.08	0.123	0.27	0.044	0.054	0.136
$\mathrm{MAD}(\Delta z)$	0.011	0.012	0.11	–	–	0.061	–	–
$\mathrm{MAD}'(\Delta z)$	0.012	–	0.098	–	–	0.062	–	–
Per cent(Δz_1)	43.4	41.1	50.7	54.9	63.9	68.1	70.8	74.9
Per cent(Δz_2)	72.4	68.4	72.3	73.3	80.2	86.5	85.8	86.9
Per cent(Δz_3)	86.9	83.4	80.5	80.7	85.7	91.4	90.8	91.0
$\sigma^2(\Delta z_1)$	8.2×10^{-6}	8.2×10^{-6}	7.9×10^{-4}	–	–	7.6×10^{-4}	–	–
$\sigma^2(\Delta z_2)$	3.0×10^{-5}	3.1×10^{-5}	0.003	–	–	0.023	–	–
$\sigma^2(\Delta z_3)$	6.1×10^{-5}	6.3×10^{-5}	0.005	–	–	0.039	–	–
$\langle \Delta z_{\mathrm{norm}} \rangle$	0.014	0.017	0.095	0.095	0.115	0.058	0.06	0.071
$\mathrm{rms}(\Delta_{\mathrm{norm}})$	0.019	0.037	0.19	–	–	0.11	–	–
$\sigma^2(\Delta z_{\mathrm{norm}})$	1.8×10^{-4}	1.1×10^{-3}	0.025	0.034	0.079	0.086	0.014	0.031
$\mathrm{MAD}(\Delta z_{\mathrm{norm}})$	0.009	0.011	0.041	–	–	0.029	–	–
$\mathrm{MAD}'(\Delta z_{\mathrm{norm}})$	0.010	–	0.040	–	–	0.031	–	–
Per cent$(\Delta z_{\mathrm{norm},1})$	48.3	45.6	77.3	–	–	87.4	–	–
Per cent$(\Delta z_{\mathrm{norm},2})$	77.2	73.5	87.3	–	–	94.0	–	–
Per cent$(\Delta z_{\mathrm{norm},3})$	90.1	87.0	91.8	–	–	96.4	–	–
$\sigma^2(\Delta z_{\mathrm{norm},1})$	8.3×10^{-6}	8.2×10^{-6}	6.2×10^{-4}	–	–	5.6×10^{-4}	–	–
$\sigma^2(\Delta z_{\mathrm{norm},2})$	3×10^{-5}	3.0×10^{-5}	0.002	–	–	0.001	–	–
$\sigma^2(\Delta z_{\mathrm{norm},2})$	5.8×10^{-5}	6.0×10^{-5}	0.004	–	–	0.002	–	–

Catalogues for both experiments available on Vizier.

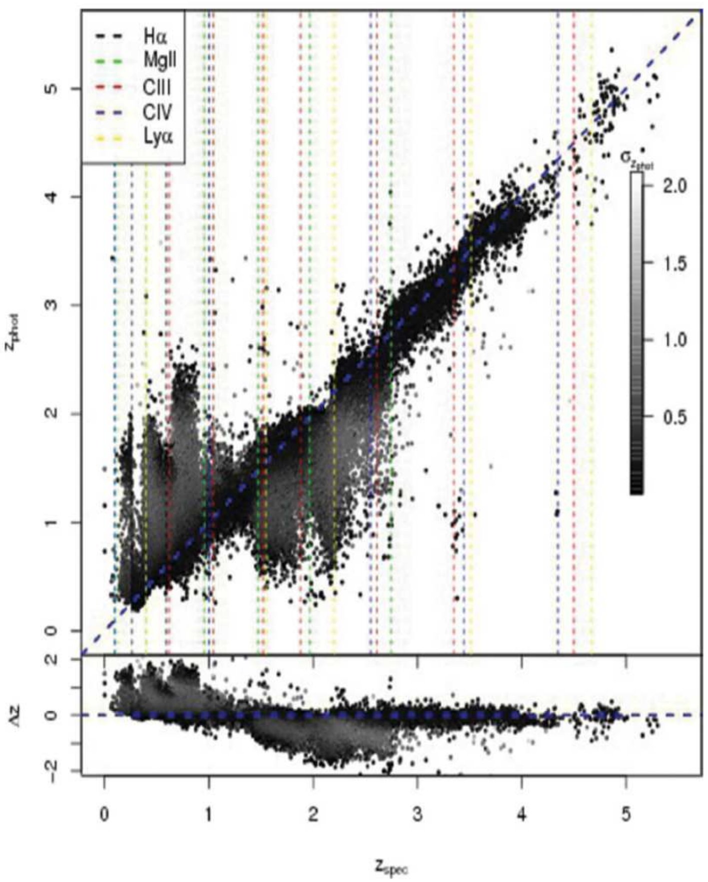


Figure 15. In the upper panel, it is shown the scatter plot of the spectroscopic versus photometric redshifts evaluated with the WGE method for the members of the KB of the experiment for the quasars extracted from the SDSS catalogue with optical photometry, while in the lower panel the scatter plot of the spectroscopic redshift $z_{\rm spec}$ versus Δz variable is shown for the same sources. All points are colour coded according to the value of the errors $\sigma_{z_{\rm phot}}$ as evaluated but the WGE. The vertical dashed lines represent the redshift at which the most luminous emission lines characterizing quasars spectra shift off the SDSS photometric filters due to redshift. Most of the features of the plot are associated to one or more of these lines.

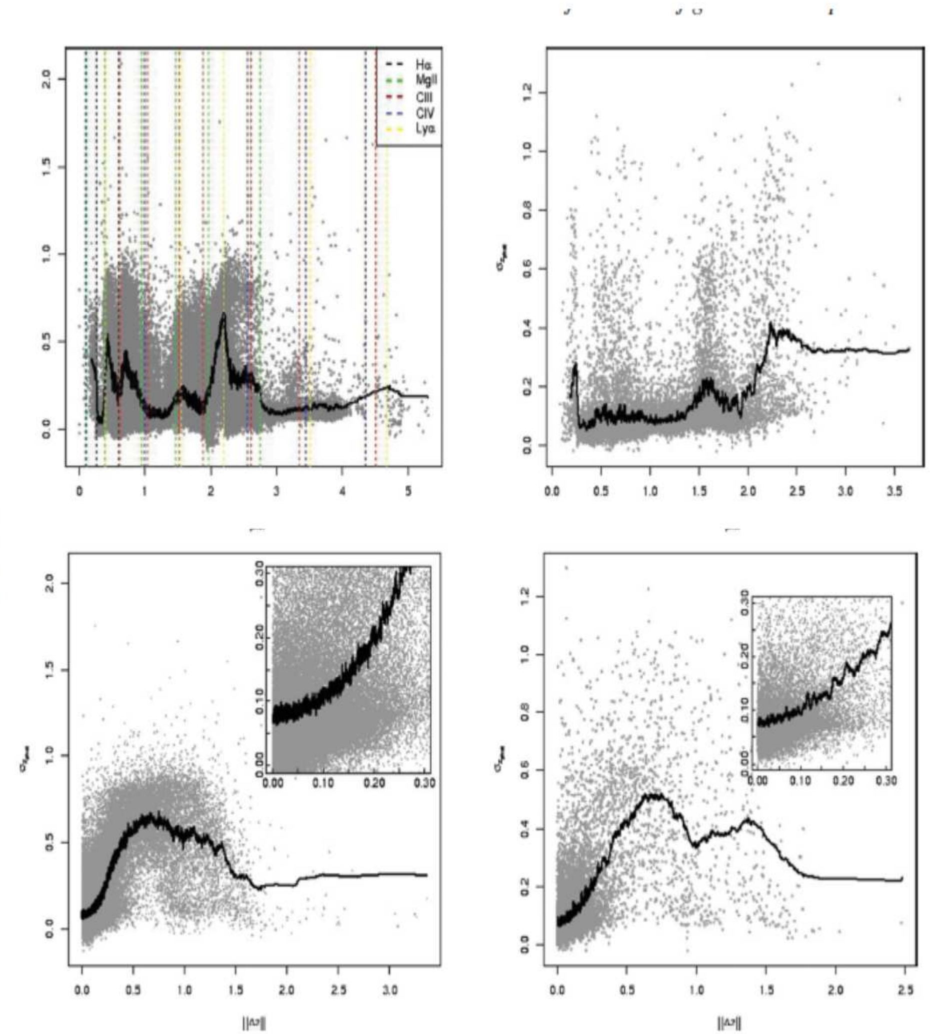


Photo-z for SDSS QSOs with MLPQNA

[*Photometric redshifts for quasars in multiband surveys*](), *M. Brescia, S. Cavuoti, R. D'Abrusco, A. Mercurio, G. Longo 2013, ApJ, 772, 140*

Lenghty feature selection procedure

Survey	Bands	Name of feature	Synthetic description
GALEX	nuv, fuv	mag, mag_iso mag_Aper_1 mag_Aper_2 mag_Aper_3 mag_auto and kron_radius	Near and Far UV total and isophotal mags phot. through 3, 4.5 and 7.5 arcsec apertures magnitudes and Kron radius in units of A or B
SDSS	u, g, r, i, z	psfMag	PSF fitting magnitude in the u g, r, i, z bands.
UKIDSS	Y, J, H, K	PsfMag AperMag3, AperMag4, AperMag6 HallMag, PetroMag	PSF fitting magnitude in Y, J, H, K bands aperture photometry through 2, 2.8 & 5.7″ circular aperture in each band Calibrated magnitude within circular aperture r_hall and Petrosian magnitude in Y, J, H, K bands
WISE	W1, W2, W3, W4	W1mpro, W2mpro, W3mpro, W4mpro	W1: 3.4 μm and 6.1″ angular resolution; W2: 4.6 μm and 6.4″ angular resolution; W3: 12 μm and 6.5″ angular resolution; 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	-	z_{spec}	Spectroscopic redshift

Table 6. Catastrophic outliers evaluation and comparison between the residual $\sigma_{clean}(\Delta z_{norm})$ and $NMAD(\Delta z_{norm})$. The reported number of objects, for each cross-matched catalog, is referred to the test sets only. Catastrophic outliers are defined as objects where $|\Delta z_{norm}| > 2\sigma\,(\Delta z_{norm})$. The standard deviation $\sigma_{clean}(\Delta z_{norm})$ is calculated after having removed the catastrophic outliers, i.e. on the data sample for which $|\Delta z_{norm}| \leq 2\sigma\,(\Delta z_{norm})$

Exp	n. obj.	$\sigma\,(\Delta z_{norm})$	% catas. outliers	$\sigma_{clean}\,(\Delta z_{norm})$	$NMAD\,(\Delta z_{norm})$
SDSS	41431	0.15	6.53	0.062	0.058
SDSS + GALEX	17876	0.11	4.57	0.045	0.043
SDSS+UKIDSS	12438	0.11	3.82	0.041	0.040
SDSS+GALEX+UKIDSS	5836	0.087	3.05	0.040	0.032
SDSS+GALEX+UKIDSS+WISE	5716	0.069	2.88	0.035	0.029

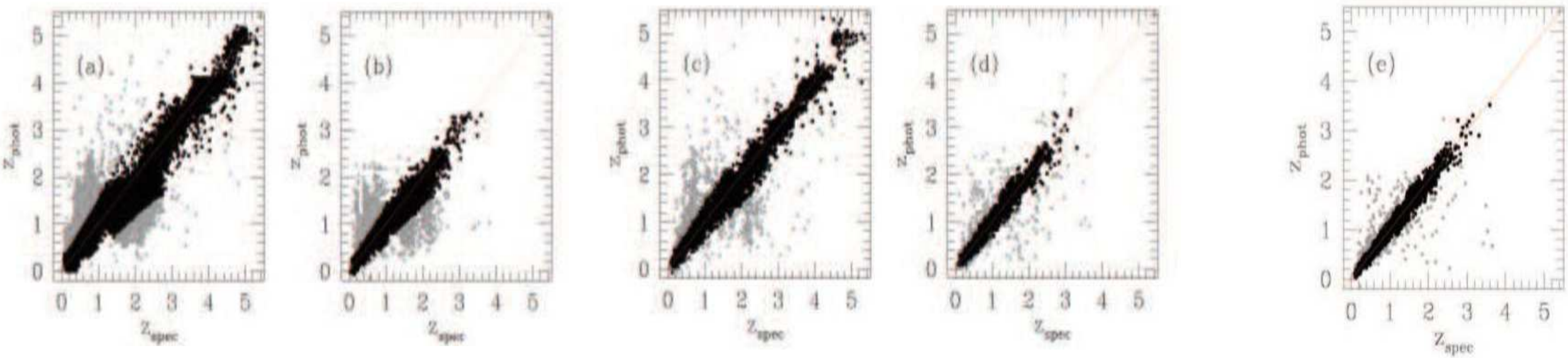


Table 4. Comparison among the performances of the different references. MLPQNA is related to our experiments, based on a four-layers network, trained on the mixed (colors + reference magnitudes) datasets. In some cases the comparison references are not reported, due to the missing statistics. Column 1: reference; columns 2-6, respectively: bias, standard deviation, MAD, RMS and NMAD calculated on $\Delta z_{norm} = (z_{spec} - z_{phot}) / (1 + z_{spec})$ related to the test sets. For the definition of the parameters and for discussion see text.

Exp	$BIAS(\Delta z_{norm})$	$\sigma(\Delta z_{norm})$	$MAD(\Delta z_{norm})$	$RMS(\Delta z_{norm})$	$NMAD(\Delta z_{norm})$
			SDSS		
MLPQNA	0.032	0.15	0.039	0.17	0.058
Laurino et al.	0.095	0.16	0.041	0.19	-
Ball et al.	0.095	0.18	-	-	-
Richards et al.	0.115	0.28	-	-	-
			SDSS + GALEX		
MLPQNA	0.012	0.11	0.029	0.11	0.043
Laurino et al.	0.058	0.29	0.029	0.11	-
Ball et al.	0.06	0.12	-	-	-
Richards et al.	0.071	0.18	-	-	-
			SDSS + UKIDSS		
MLPQNA	0.008	0.11	0.027	0.11	0.040
			SDSS + GALEX + UKIDSS		
MLPQNA	0.005	0.087	0.022	0.088	0.032
			SDSS + GALEX + UKIDSS + WISE		
MLPQNA	0.004	0.069	0.020	0.069	0.029

Table 5. Comparison in terms of outliers percentages among the different references. In some cases the comparison references are not reported, due to the missing statistics. Column 1: reference; Column 2-3 are fractions of outliers at different σ based on $\Delta z = (z_{spec} - z_{phot})$; Column 4-5 are the fractions of outliers at different σ based on $\Delta z_{norm} = (z_{spec} - z_{phot}) / (1 + z_{spec})$. The column 4 reports our catastrophic outliers, defined as $|\Delta z_{norm}| > 2\sigma(\Delta z_{norm})$.

Exp	Outliers ($\|\Delta z\|$)		Outliers ($\|\Delta z_{norm}\|$)	
	$> 2\sigma(\Delta z)$	$> 4\sigma(\Delta z)$	$> 2\sigma(\Delta z_{norm})$	$> 4\sigma(\Delta z_{norm})$
	SDSS			
MLPQNA	7.68	0.38	6.53	1.24
Bovy et al.		0.51		
	SDSS + GALEX			
MLPQNA	4.88	1.61	4.57	1.37
Bovy et al.		1.86		
	SDSS + UKIDSS			
MLPQNA	4.00	1.73	3.82	1.38
Bovy et al.		1.92		
	SDSS + GALEX + UKIDSS			
MLPQNA	2.86	1.47	3.05	0.23
Bovy et al.		1.13		
	SDSS + GALEX + UKIDSS + WISE			
MLPQNA	2.57	0.87	2.88	0.91

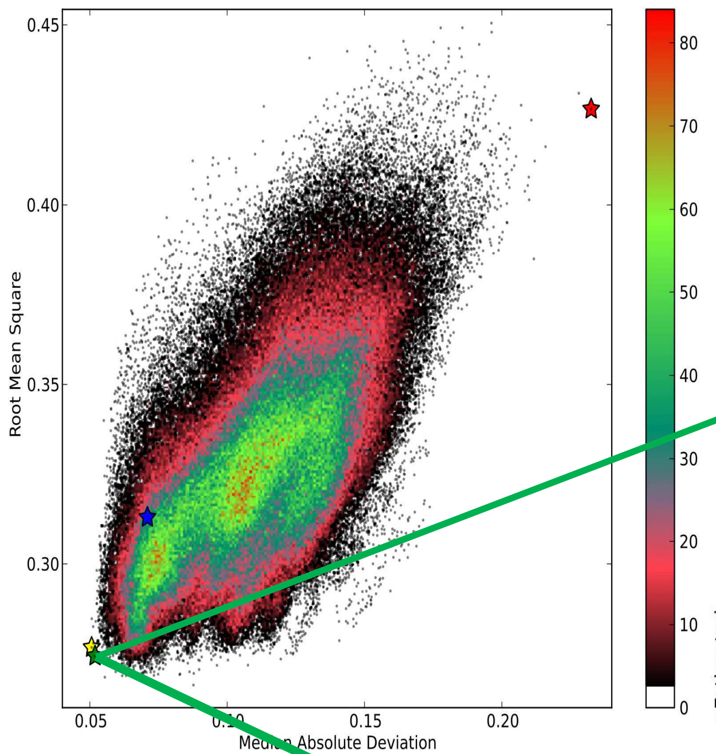
Different Machine Learning methods of different complexity (MLPQNA is conceptually simpler than WGE) lead to similar results with a slight edge for MLPQNA

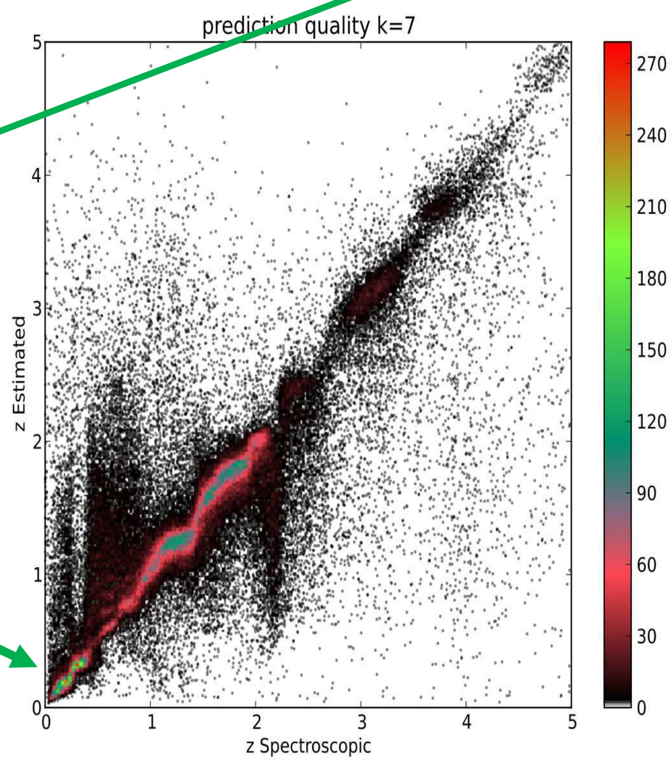
Photometric redshifts for QSO's … a data driven approach (from K. Polsterer, Heidelberg)

One does not know a-priori which features are the most relevant

$$\frac{n!}{(n-r)!\,r!}, \text{with}\; n=55, r=4 \rightarrow 341{,}055\; combinations$$

Use all 55 significant photometric features to select the most significant 4





Best combination

$u_{model} - g_{model}$

$g_{psf} - r_{model}$

$z_{psf} - r_{model}$

$i_{psf} - z_{model}$

Results comparable to Brescia et al. 2014

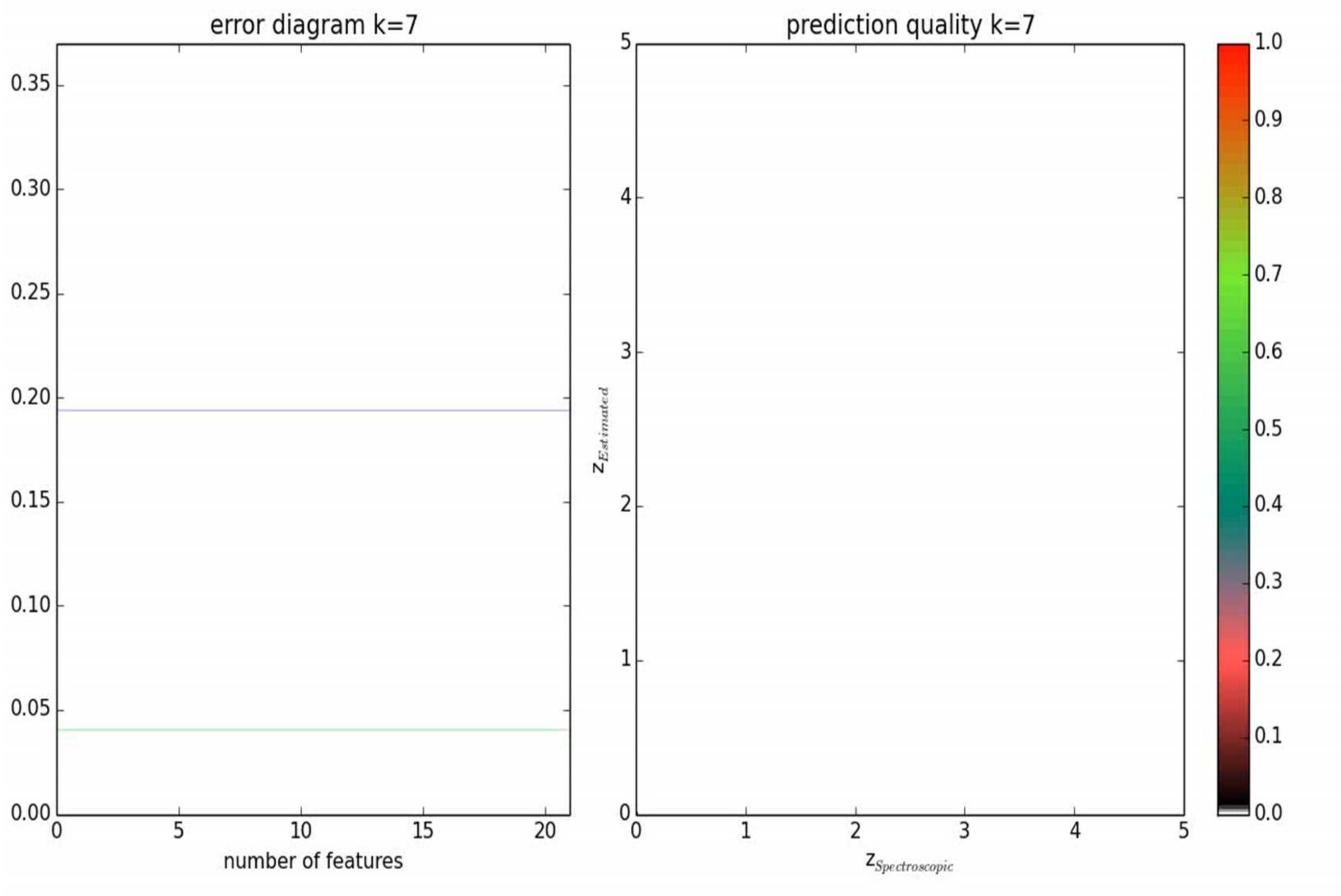
Is it possible to do better ?

Photometric redshifts for SDSS QSO

PSF, Petrosian, Total magnitudes + extinction + errors ….. 585 features….
1,197,308,441,345,108,200,000 combinazioni

1.2 sextilions of combinations

Hence features addition…..



You hit a plateau at 10 features.

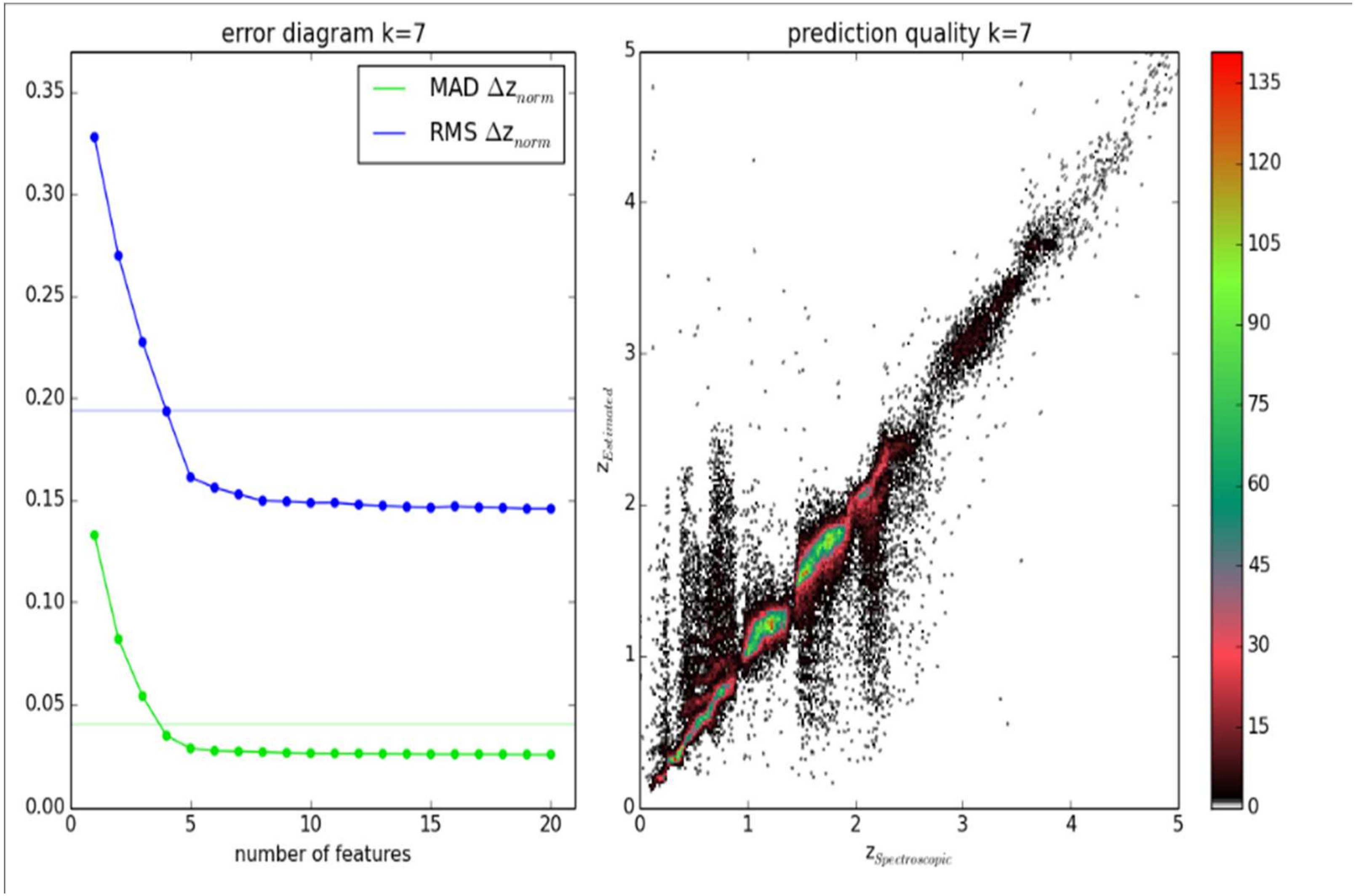
$$u_{psf} - g_{petrosian}$$
$$dered(z_{psf}) - dered(i_{petrosian})$$
$$dered(g_{psf}) - dered(r_{model})$$
$$\frac{dered(r_{psf}) - dered(z_{model})}{\sqrt{\sigma^2_{g_{model}} + \sigma^2_{r_{model}}}}$$
$$dered(r_{model}) - dered(i_{model})$$
$$i_{psf} - i_{petrosian}$$
$$dered(z_{psf}) - dered(r_{petrosian})$$
$$\frac{g_{model} - g_{petrosian}}{\sqrt{\sigma^2_{g_{petrosian}} + \sigma^2_{r_{petrosian}}}}$$

Photometric redshifts for SDSS QSO

PSF, Petrosian, Total magnitudes + extinction + errors ….. 585 features….
1,197,308,441,345,108,200,000 combinazioni

1.2 sextilions of combinations

Hence features addition…..



You hit a plateau at 10 features.

$$u_{psf} - g_{petrosian}$$
$$dered(z_{psf}) - dered(i_{petrosian})$$
$$dered(g_{psf}) - dered(r_{model})$$
$$\frac{dered(r_{psf}) - dered(z_{model})}{\sqrt{\sigma^2_{g_{model}} + \sigma^2_{model}}}$$
$$dered(r_{model}) - dered(i_{model})$$
$$i_{psf} - i_{petrosian}$$
$$dered(z_{psf}) - dered(r_{petrosian})$$
$$\frac{g_{model} - g_{petrosian}}{\sqrt{\sigma^2_{g_{petrosian}} + \sigma^2_{r_{petrosian}}}}$$

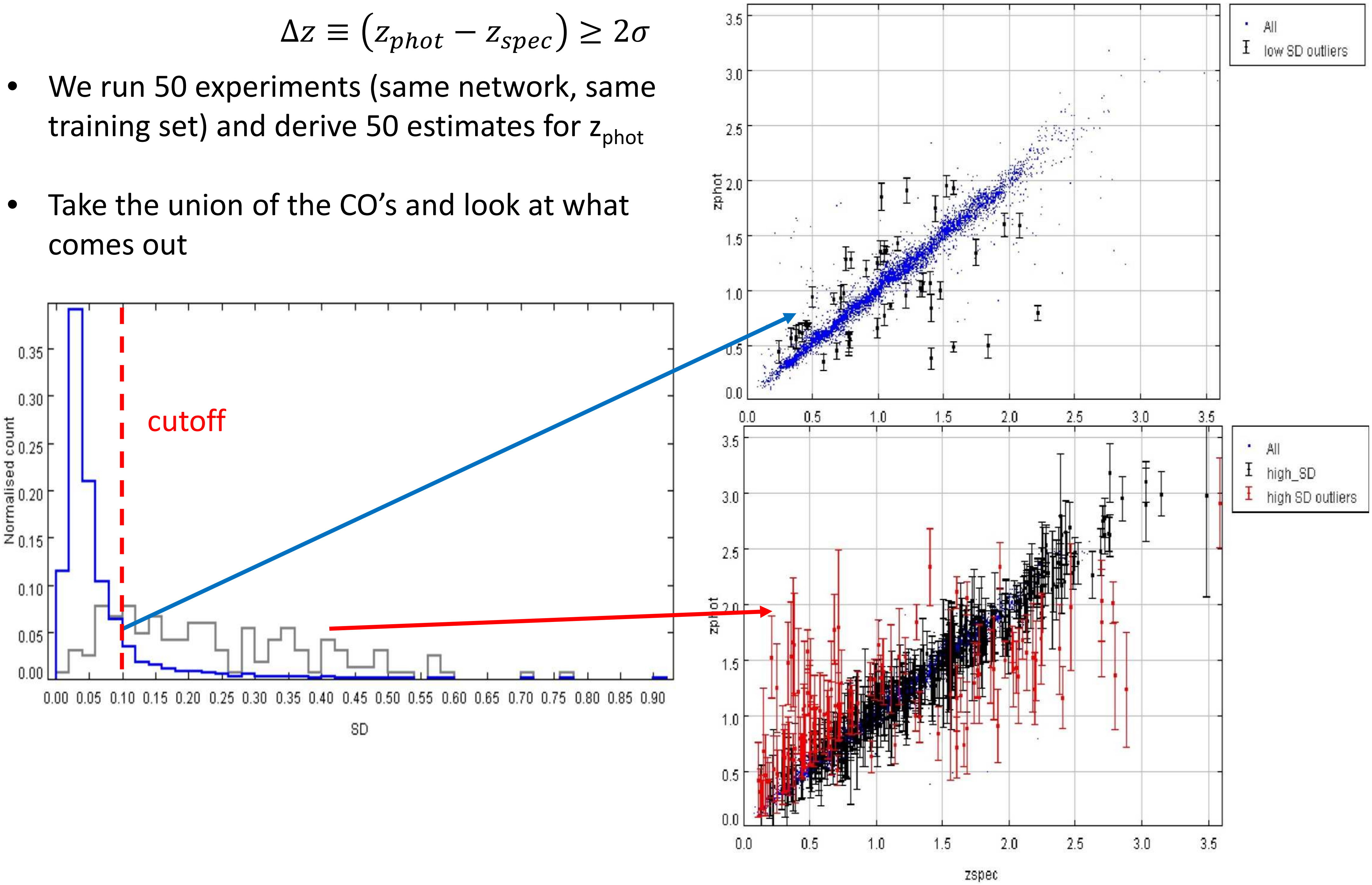
Catastrophic outliers

MLPQNA, same KB and same features as in Brescia et al. 2013

$$\Delta z \equiv (z_{phot} - z_{spec}) \geq 2\sigma$$

- We run 50 experiments (same network, same training set) and derive 50 estimates for z_{phot}

- Take the union of the CO's and look at what comes out



How about quality flags?

SDSS provides a complete set of quality flags extrapolated from astronomers expertise

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%

Inspection of flags for CO's shows that these flag are practically useless to discriminate CO's

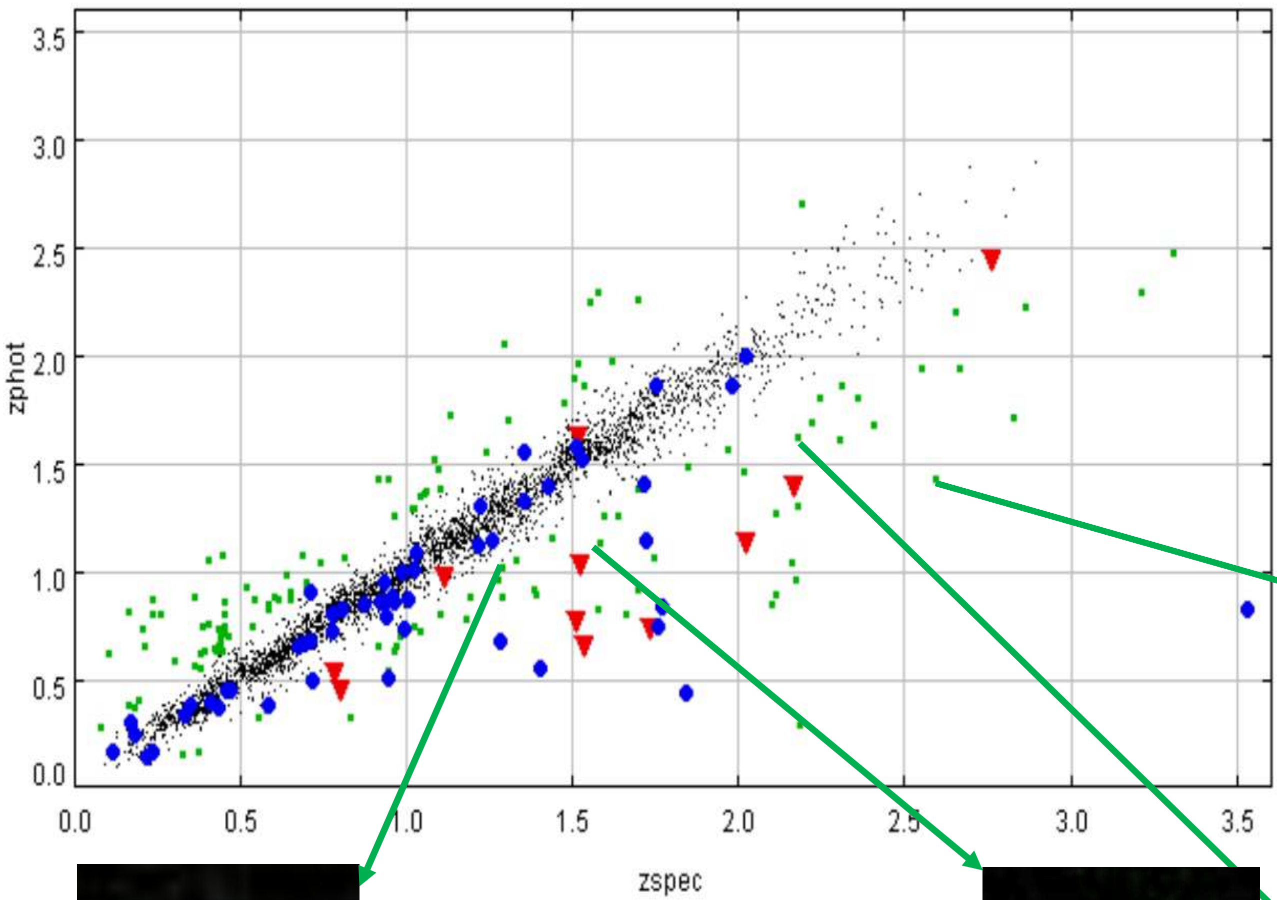
Crosscorrelation with other catalogues to check for variability (e.g. CRTS)

NO clear effect on CO's induced by variability of sources.

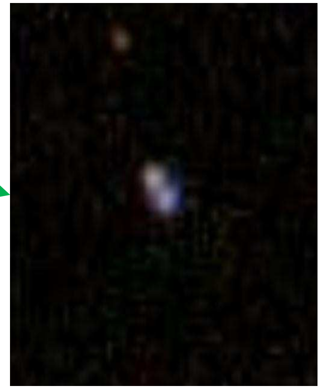
SDSS is almost simultaeous in all optical bands but other surveys are not

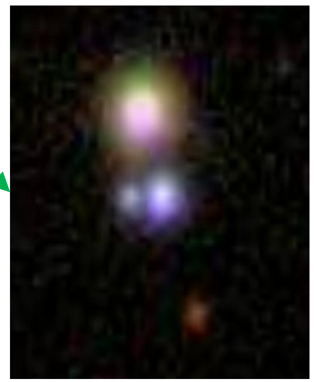
What are these Catastrophic outliers?

Petrillo C.E., Longo G., Brescia M., Cavuoti G., in preparation;
Petrillo Laurea Thesis 2013, University of Naples)

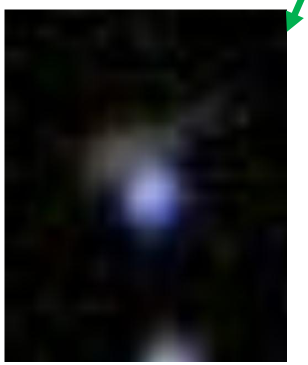


- **Blu dots: blazars**
- **Green dots: unknown CO's**
- **Red triangles: gravitationally lensed quasars**

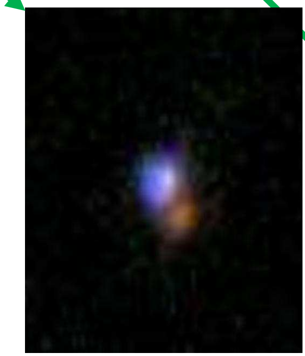




Gravitational lens candidates



Peculiar objects



So, if you apply a rejection criteria based on the σ of the different predictions.....

	initial	average	low sd
Dataset	14284	14284	(14284 - 487)
BIAS(Δz)	0.002	0.0001	0.0007
$\sigma\,(\Delta z)$	0.14	0.12	0.077
MAD(Δz)	0.043	0.036	0.034
RMS(Δz)	0.14	0.12	0.077
NMAD(Δz)	0.063	0.054	0.050
$> 2\sigma\,(\Delta z)$	2.94%	3.17%	3.67%
$> 4\sigma\,(\Delta z)$	1.14%	0.10%	0.40%
BIAS (Δz_{norm})	0.003	0.003	0.0005
$\sigma\,(\Delta z_{norm})$	0.070	0.059	0.037
MAD(Δz_{norm})	0.021	0.018	0.017
RMS(Δz_{norm})	0.070	0.060	0.037
NMAD(Δz_{norm})	0.031	0.027	0.025
$> 2\sigma\,(\Delta z_{norm})$	2.66%	2.98%	3.87%
$> 4\sigma\,(\Delta z_{norm})$	0.84%	0.89%	0.57%

$$\sigma \geq \sigma_{treshold} = 0.125$$

By rejecting all objects which have

Loss in completeness $\sim 5\%$

Gain in $\sigma \sim 2$

Drastic reduction in number of catastrophic outliers

Selection biases

SDSS – Data Release 10

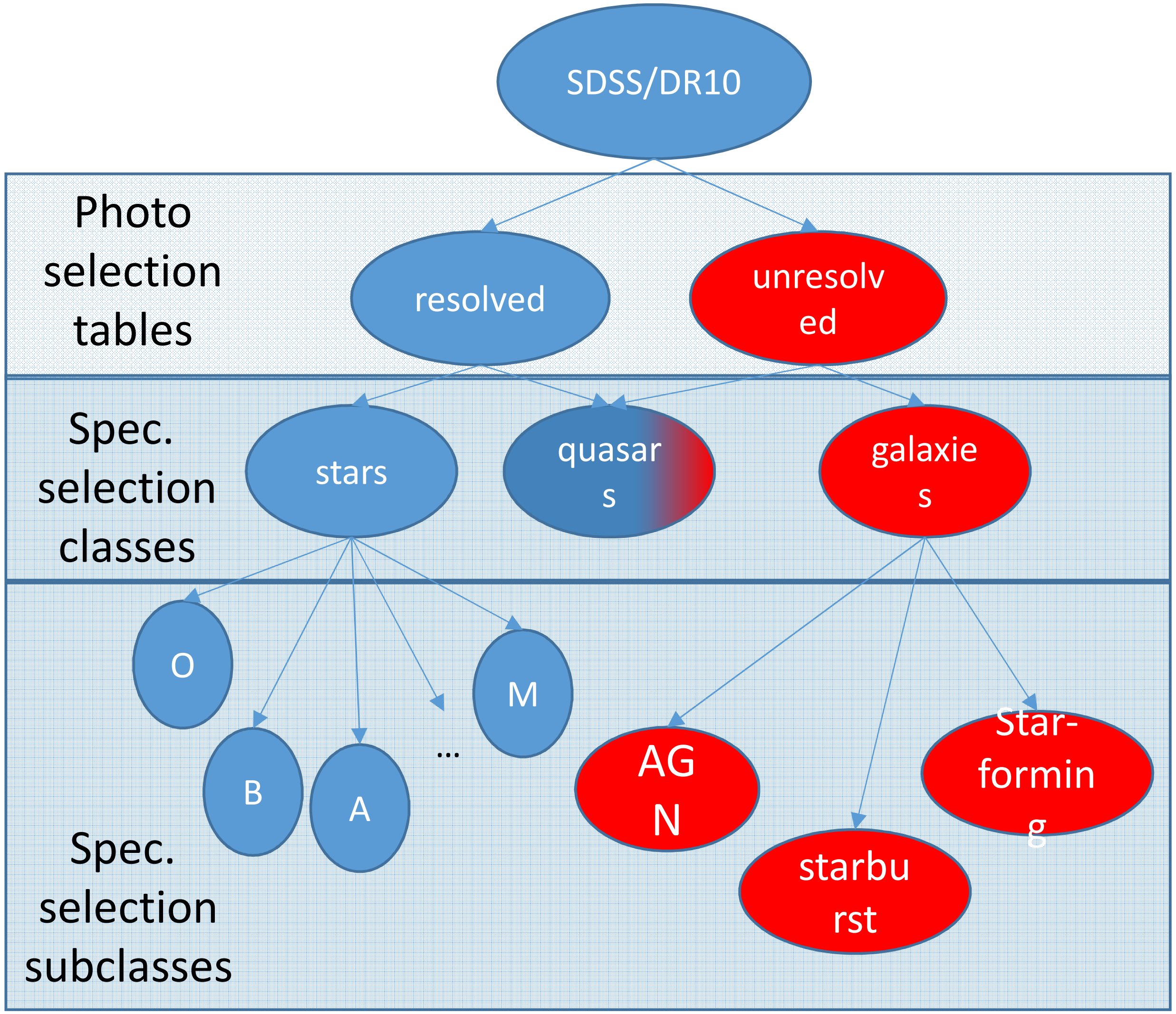
OPPS		OSPS	
$3x10^8$	objects	$3x10^6$	objects
> 100	features	>50	features
> 100	flags	>50	flags

Problem:

To evaluate Photo-z for all SDSS objects using the spectroscopic z's in the KB

The KB is the result of selection criterias and is biased

Not all selections and biases can be mapped in the OPPS

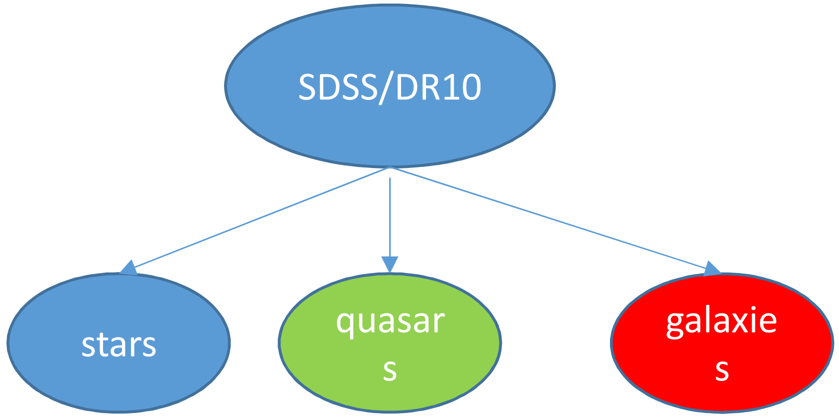


Spectroscopic Knowledge Base

A less biased approach: 3 class classification

Brescia M, Cavuoti S., Longo G., 2014 submitted

Model: MLPQNA



Features:

type	parameters
Identification	objID, specObjID, RA, DEC
psfMag	ugriz mag and *mag_err*
fiberMag	ugriz mag and *mag_err*
modelMag	ugriz mag and *mag_err*
cmodelMag	ugriz mag and *mag_err*
deredMag	ugriz mag
extinction	ugriz
spec redshift	z, zWarning
classification	type, class, subclass, flags

Table 2. Description of the parameters (features and targets) used in this work. The first part of the table lists the photometric parameters in the OPPS, the second one the spectroscopic data and targets. Column 1: collective name of a given set of parameter; column 2: the corresponding SDSS parameters.

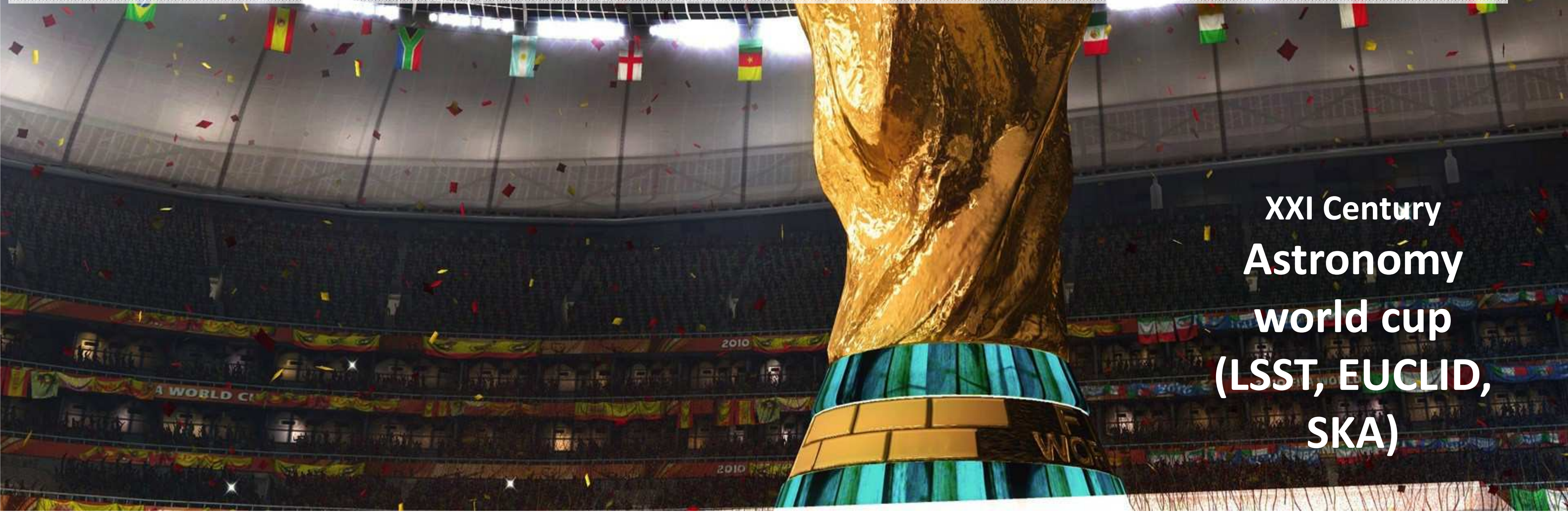
Results:

CLASS	$\%e_{tot}$	$\%C_A$	$\%P_A$	$\%Co_A$
GALAXY		97.02	93.49	6.51
STAR	91.31	86.40	93.82	6.18
QSO		90.49	86.90	13.10

Table 5. Summary of the results of the best three-class experiment (3*a*), referred to the parameter space composed by the 10 magnitudes (*psfMag* and *magModel*) for each object. The training and test sets are respectively 12% and 88% of the given dataset. The columns are referred to equations from 1 to 4. All the quantities reported in the table are percentages.

Conclusions

- Large (Big) data are coming…
- Slow but steady adoption of advanced tools
- Computing infrastructures are only a part (small) of the history
- Most of the work so far consisted in extracting known information
  Using existing data models with automatic techniques
- New set of features specifically designed for ML need to be adopted
- Data Driven Discovery is still (and rightly) in its infancy
- A change in methodology is taking place



XXI Century
Astronomy world cup (LSST, EUCLID, SKA)