Università degli studi di Napoli “Federico II”
Laurea Magistrale in Fisica

PhotoRaptor

PHOTOmetric Research APplication TO Redshifts and application to the SDSS-DR9 galaxies

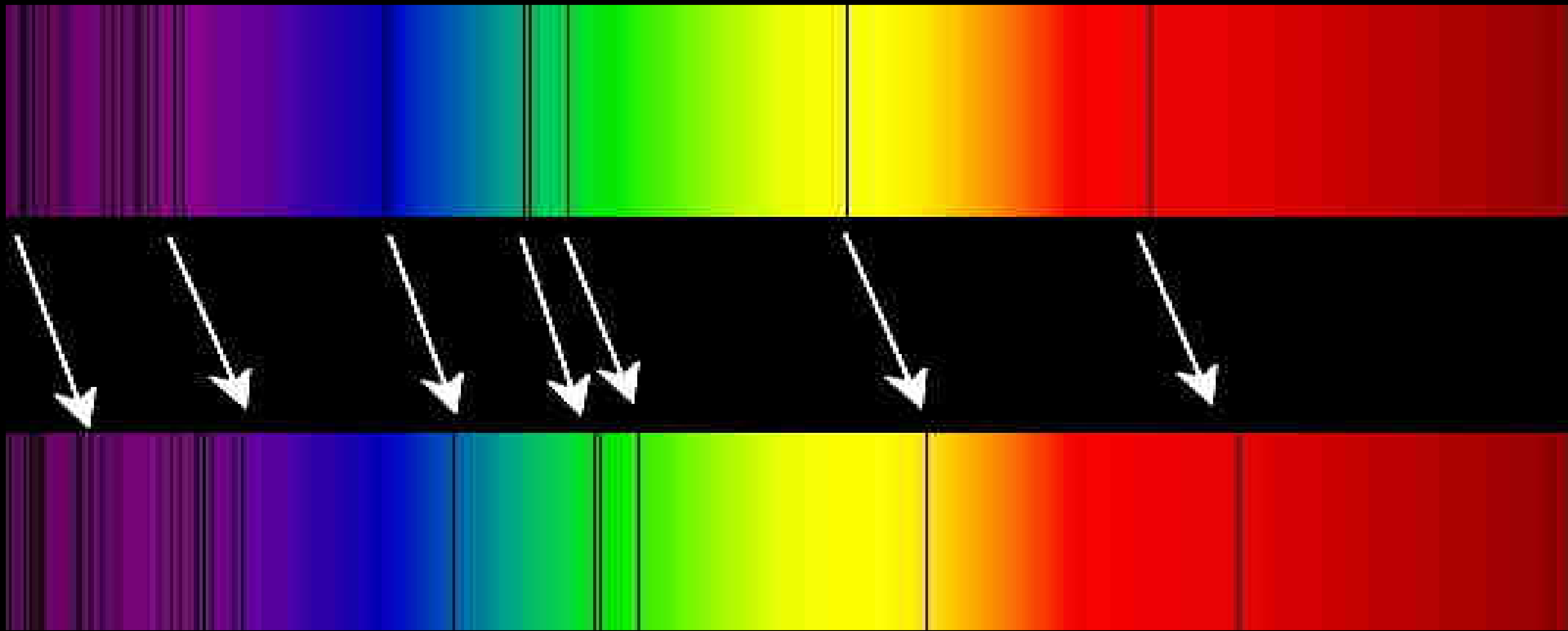
Relatori:
Ch.mo Prof. Giuseppe Longo
Dott. Massimo Brescia

Candidato:
Virgilio De Stefano
Matricola N94/147

Outline

- ✓ What are the Redshifts?
- ✓ Why Photometric Redshifts?
- ✓ Data Rich Astronomy
- ✓ Machine Learning Method: MLPQNA
- ✓ PhotoRApToR Java Tool
- ✓ SDSS-DR9 complete photo-z Catalogue

Spectroscopic Redshift (zspec)



$$z \equiv \frac{\Delta\lambda}{\lambda} = \frac{\lambda_{obs} - \lambda_{rest}}{\lambda_{rest}}$$

Virgilio De Stefano

Doppler Effect

$$z = \frac{v}{c}$$

Redshift expression for v<<c

Observed frequency due to time dilation

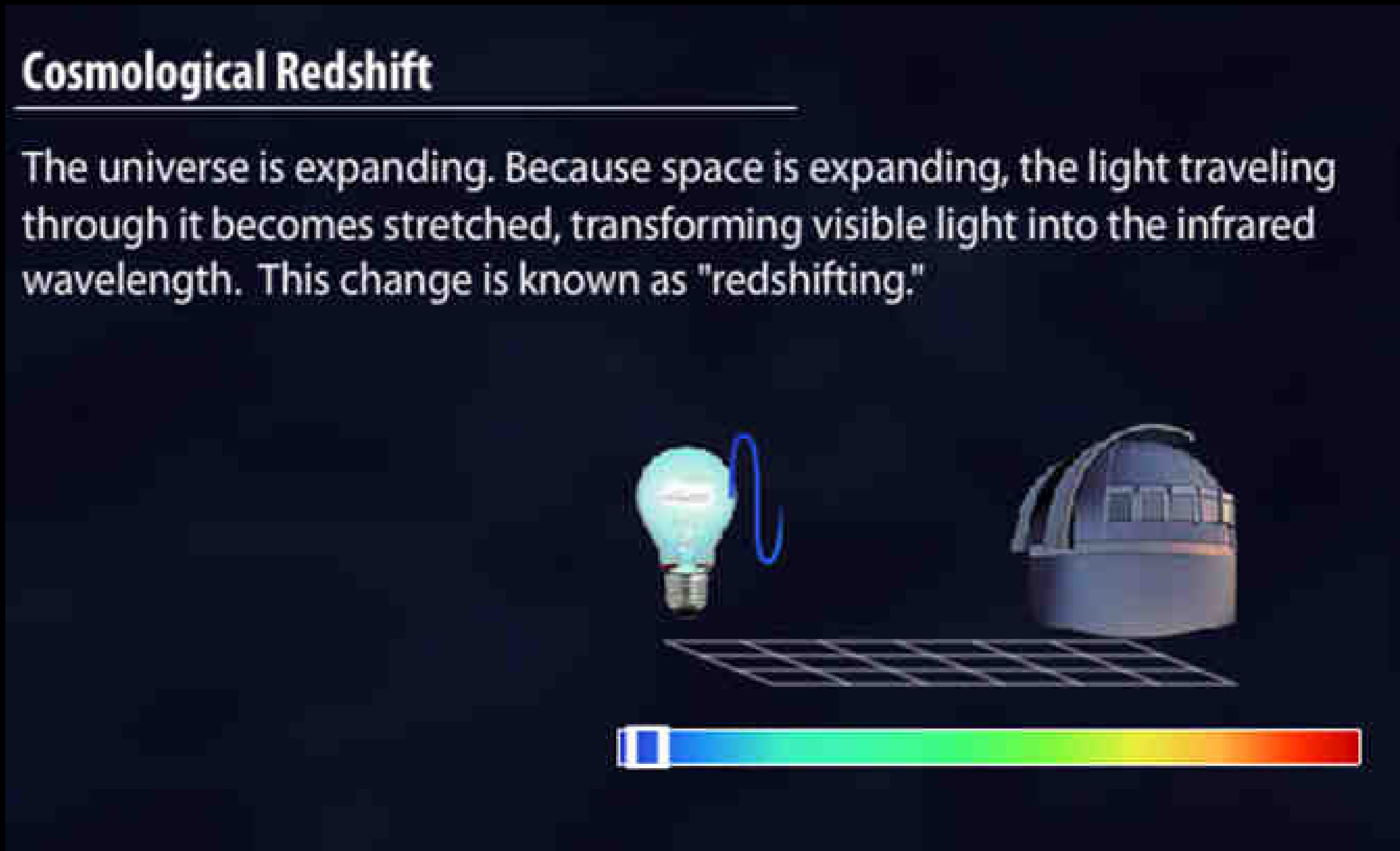
$$\nu_{obs} = \frac{1}{t_{obs}} = \gamma(1-\beta)\nu_{rest} = \sqrt{\frac{1-\beta}{1+\beta}}\nu_{rest}$$

$$z = \sqrt{\frac{c+v}{c-v}} - 1$$

Special relativistic expression for redshift

Cosmological Redshift

There is a correlation between redshift and cosmological expansion



Cosmological Redshift

$$d(t) = R(t)d_0$$

$$c^2dt^2 = R^2(t)\frac{dr^2}{1-kr^2}$$

Friedmann-Lemaitre-Robertson-Walker cosmological model

$$wave1 \qquad \int_{t_1}^{t_0} \frac{cdt}{R(t)} = -\int_0^r \frac{1}{\sqrt{1-kr^2}}dr$$

$$wave2 \qquad \int_{t_1+\Delta t_1}^{t_0+\Delta t_0} \frac{cdt}{R(t)} = -\int_0^r \frac{1}{\sqrt{1-kr^2}}dr$$

Application to two electromagnetic waves separated in time

$$z = \frac{\lambda_0 - \lambda_1}{\lambda_1} = \frac{\Delta t_0 - \Delta t_1}{\Delta t_1}$$

$$z = \frac{R(t_0)}{R(t_1)} - 1$$

Redshift related to the scale factor

Why do we need Redshift

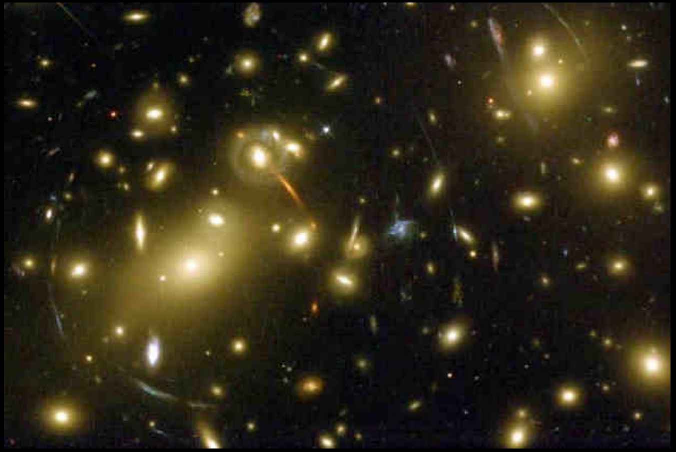
To measure the distance of objects

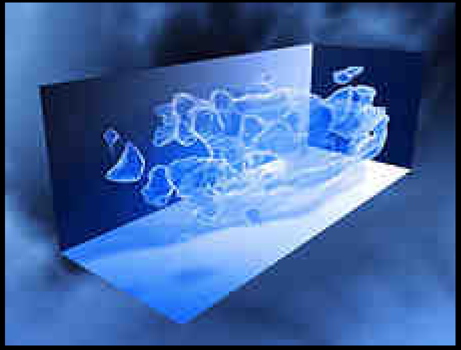
$$d = \frac{v}{H_0} \approx \frac{cz}{H_0}$$

To disentangle the degeneracies in the object classification

Cosmological parameters

Lensing Effects

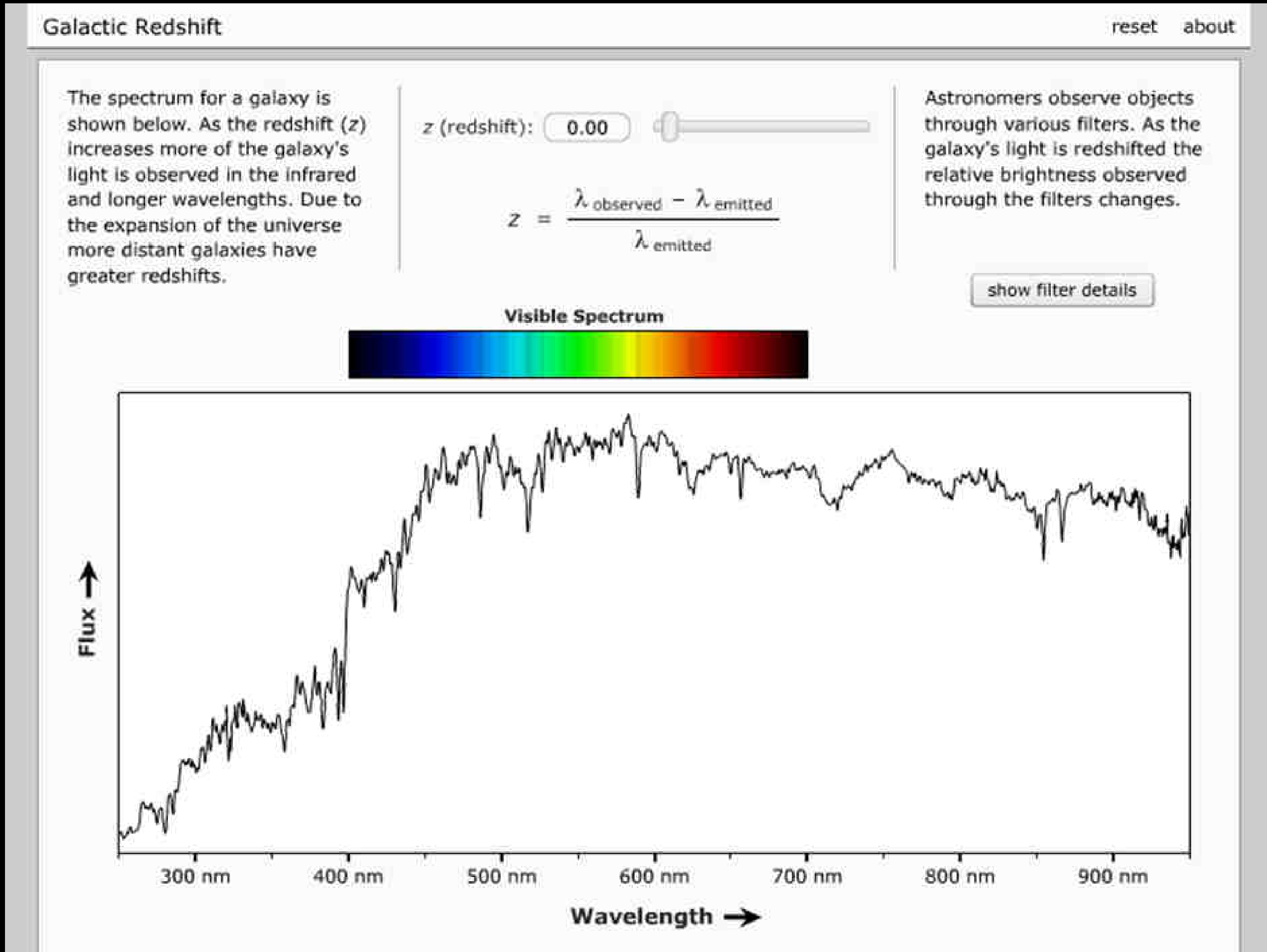




Dark Matter

Dark Energy

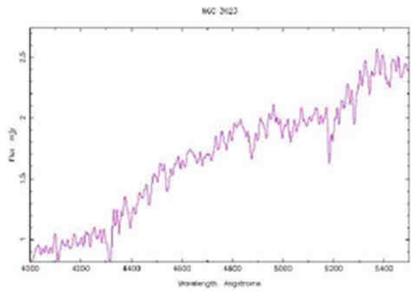
Photometric Redshift



Virgilio De Stefano

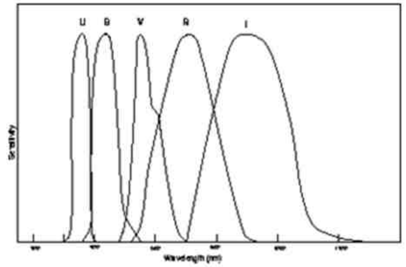
Photo-z as an inverse problem

Spectral Energy Distribution convolved with band filters



Galaxy spectrum - F(λ)

X



Photometric system - $S_i(\lambda)$

=

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

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

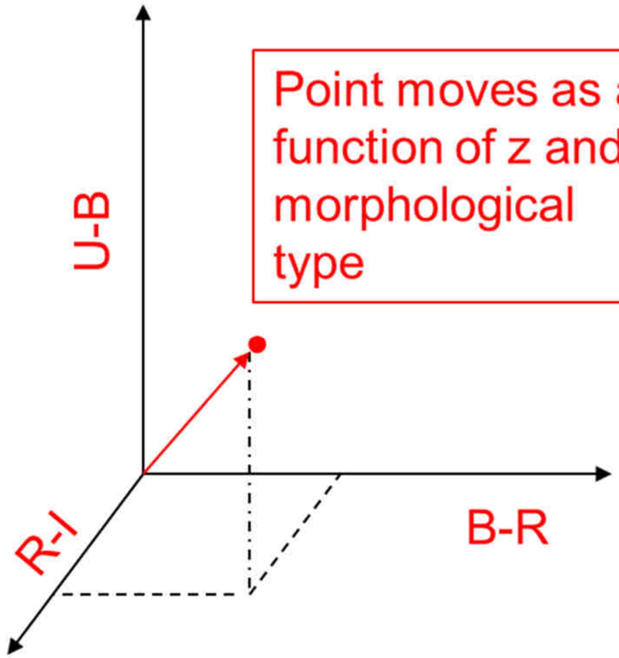
Etc…

Color indexes

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

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

etc.



Point moves as a function of z and morphological type

Phot-z are an inverse problem

Photo-z history

Baum (1962)

Colors of early type 9-band galaxies turned into a low resolution SED to determine inter-cluster distances

Koo (1985)

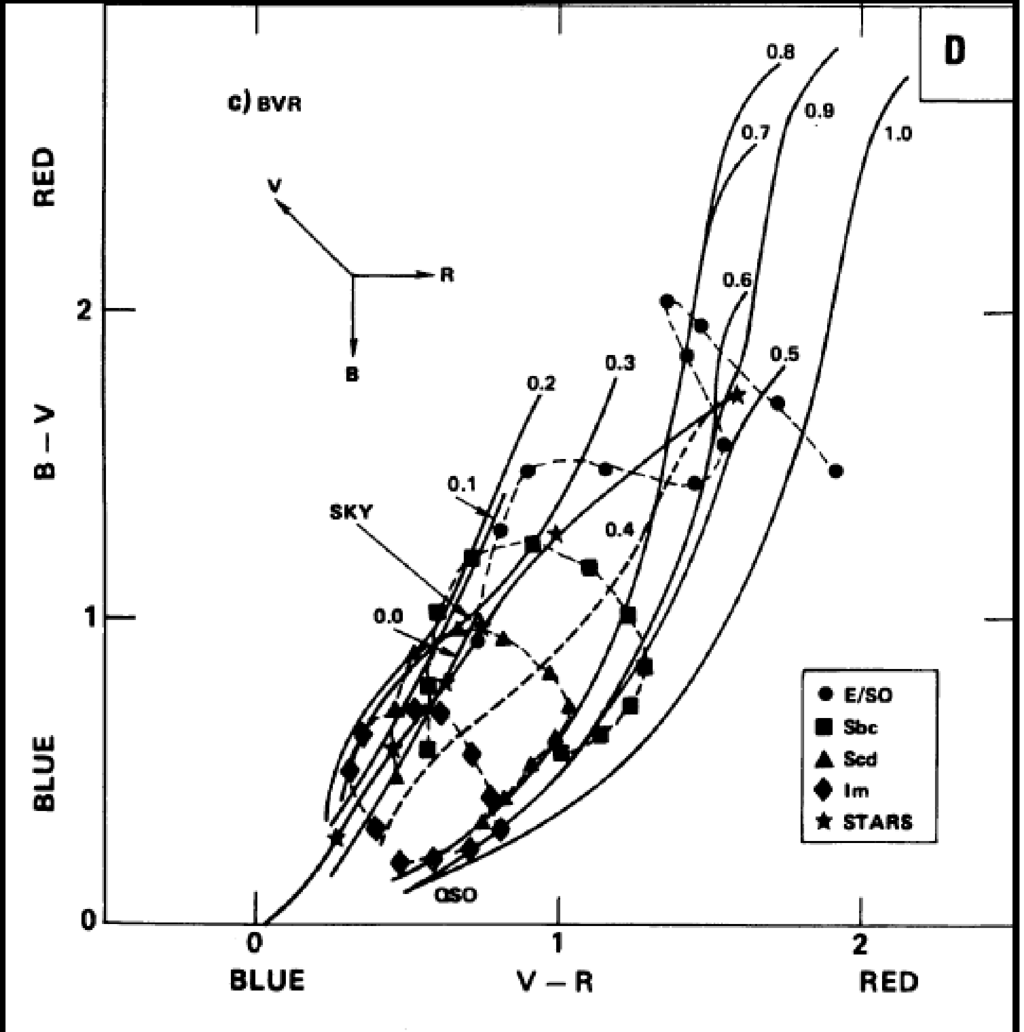
Colors (from photo plate) compared to expected colors for synthetic SEDs. Redshifts estimated from iso-z lines

Loh & Spillar (1986)

χ2-minimization for redshift estimates

Pello and others (1996)

intersection of “permitted” redshift ranges for all galaxy colors defines “the” redshift of a galaxy.



Photometric redshifts have become very popular since the middle of the 1990s

well calibrated, multi-band representative spectroscopic datasets have become available to test methods (Keck, VLT, SDSS...)

Photometric Redshift: Methods

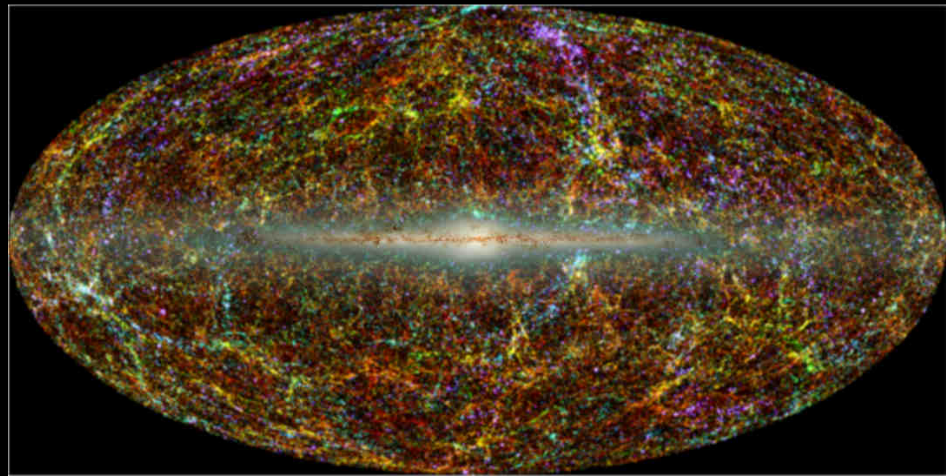
(SED fitting) Template Based

Use SED templates

Convolve with filter transmission curves

Fit object’s fluxes (χ^2 minimization)

Outputs photometric redshift



- *Hyper-z (Bolzonella et al. 2000, A&A 363, 476)*
- *BPZ (Benitez 2000, ApJ 536, 571)*
- *EAZY (Brammer et al. 2008, ApJ 676, 1503)*
- *Le PHARE (Arnouts et al. 2008, MNRAS 329, 355)*
- *LRT (Assef et al. 2008, ApJ 676, 286)*
- *ZEBRA (Feldmann et al. 2006, MNRAS 372, 565)*

They use SED models with priors (IMF, SFR, metallicity, age);
They suffer of mismatched templates

Photometric Redshift: Methods

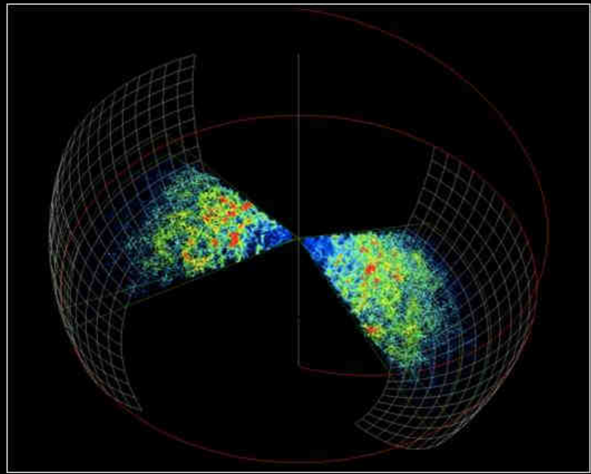
Empirical (data mining) Methods

Use a Knowledge Base of zspec

Learn hidden photometry-zspec correlation

Generalize learning on new photometric objects

Outputs photometric redshift

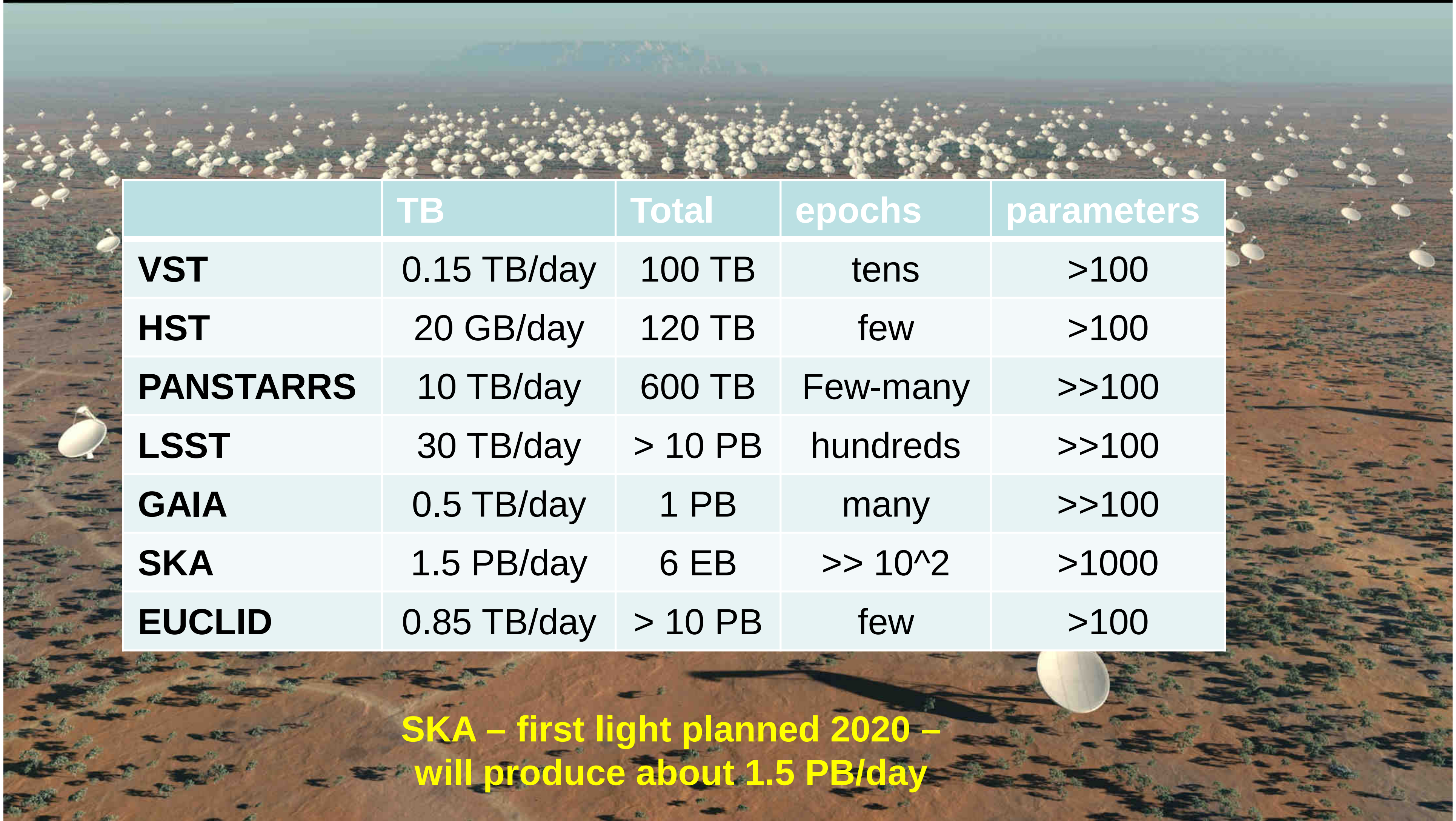


- Nearest Neighbour (*Csabai & al. 2007, A N 328, 852*)
- Decision-tree (*BDT Gerdes et al 2010, ApJ 715, 823*)
- Direct fitting (*Barth 2002, AJ 124, 5*)
- Neural Networks (*ANNz Collister & Lahav 2004, PASP 116, 345*)
- Support Vector Machines (*Wadadekar 2005, PASP 117, 79*)
- Regression Trees & Random Forests (*Carliles et al 2010, ApJ 712, 511*)
- MLPQNA (*Brescia et al. 2013, ApJ, 772, 140*)

NO kind of priors and can find unkown relations;

They require huge Knowledge Base and mismatch outside training spec range

Data Rich Astronomy

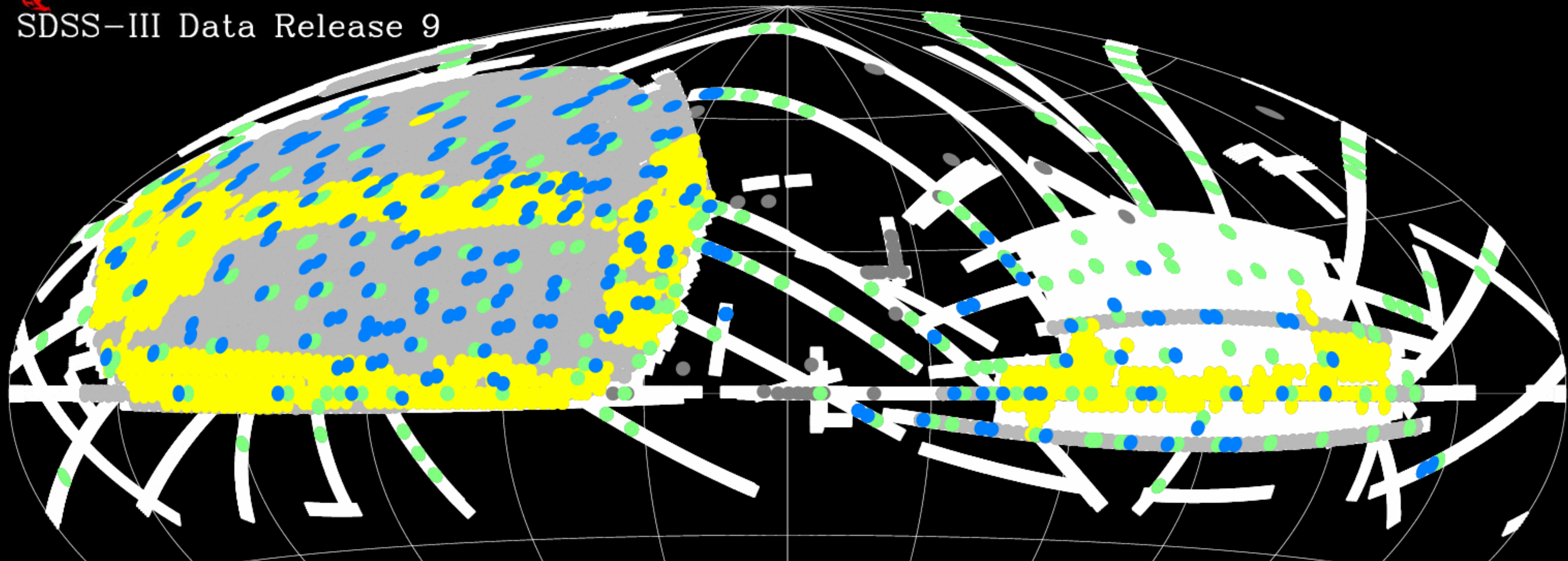


	TB	Total	epochs	parameters
VST	0.15 TB/day	100 TB	tens	>100
HST	20 GB/day	120 TB	few	>100
PANSTARRS	10 TB/day	600 TB	Few-many	>>100
LSST	30 TB/day	> 10 PB	hundreds	>>100
GAIA	0.5 TB/day	1 PB	many	>>100
SKA	1.5 PB/day	6 EB	>> 10^2	>1000
EUCLID	0.85 TB/day	> 10 PB	few	>100

SKA – first light planned 2020 – will produce about 1.5 PB/day

AND THIS IS JUST ONE SURVEY!!!

Sloan Digital Sky Survey (SDSS)

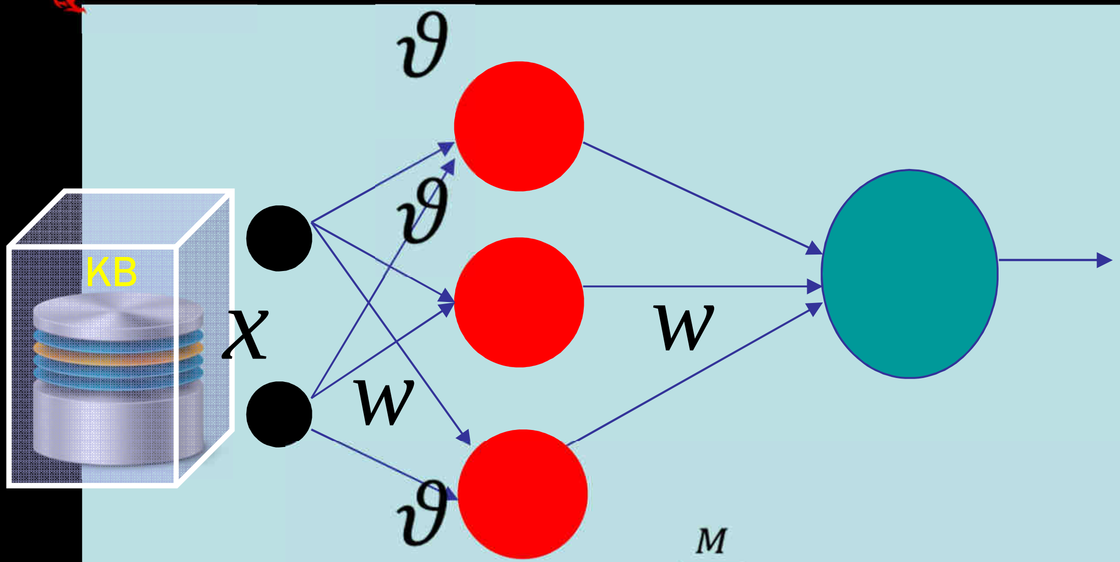


SDSS DR9 Facts	
Sky coverage	**14,555 deg^2**
Catalog objects	**932,891,133**
Galaxy spectra	**1,457,002**
Quasar spectra	**228,468**
Star spectra	**668,054**

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

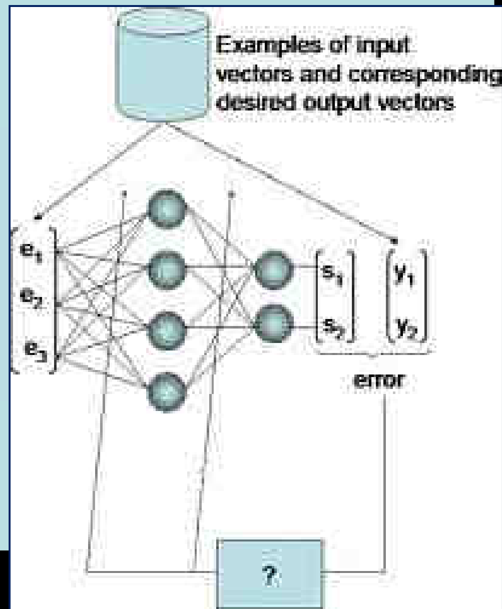
Virgilio De Stefano

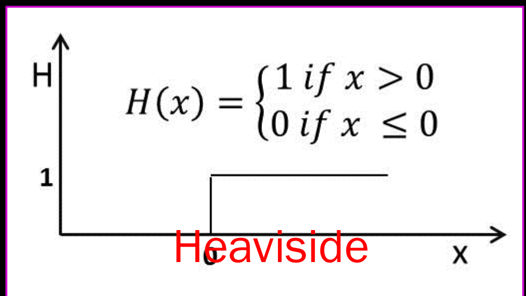
Multi Layer Perceptron

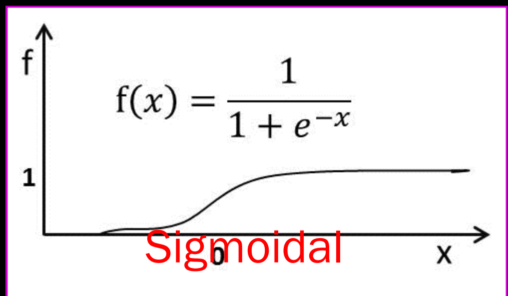


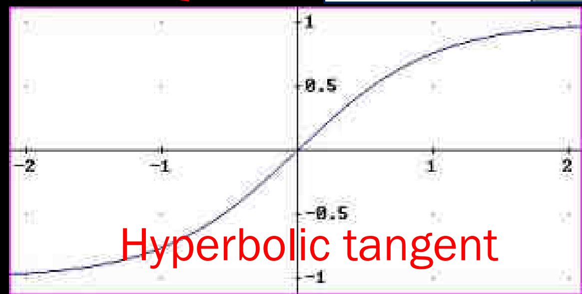
$y(x; w, \theta)$ = photo-z

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







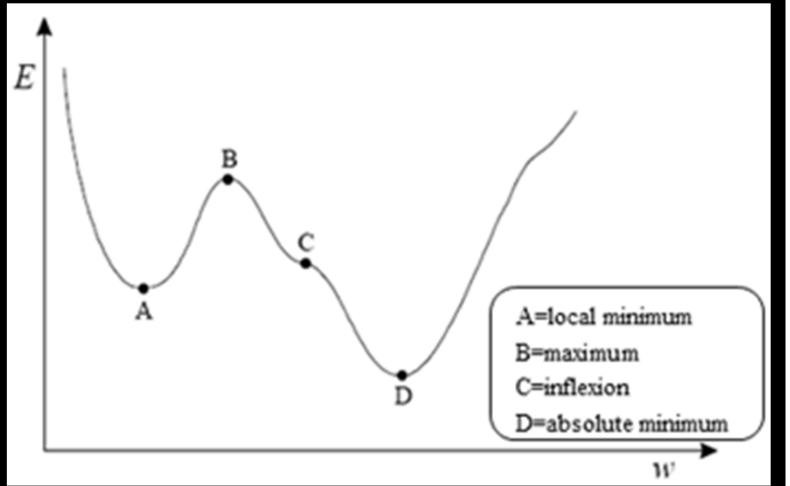


Virgilio De Stefano

Learning Rule

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

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**

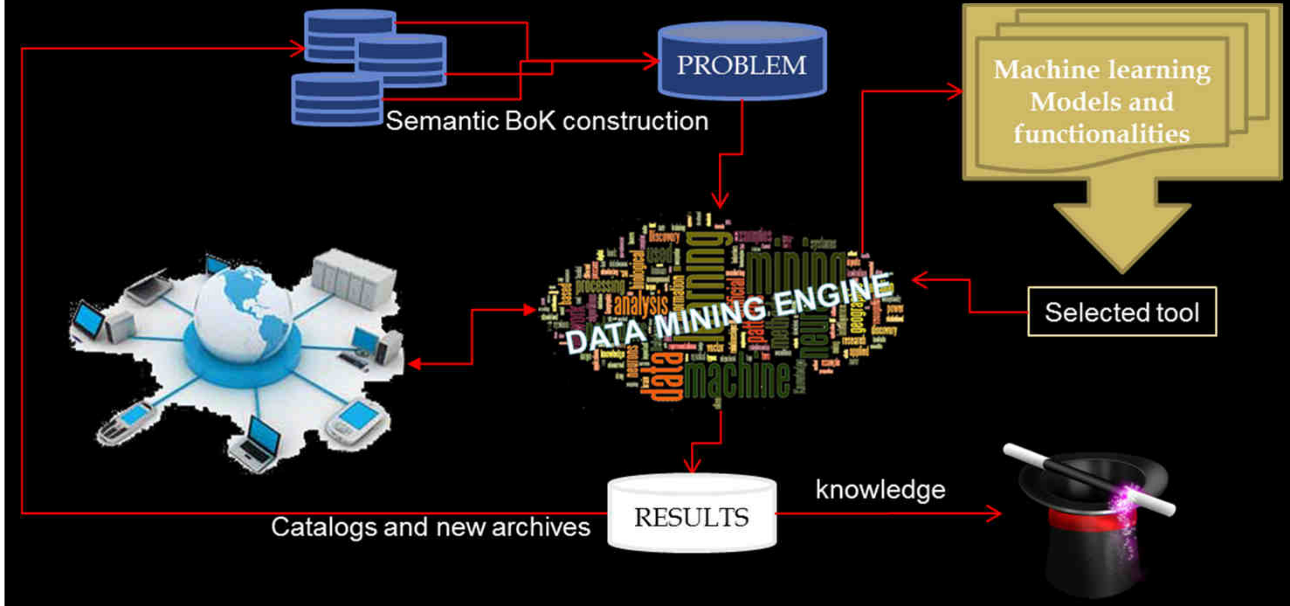
$d^k = -\nabla E(w^k)$ **Descent gradient (BP)**

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

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

DAMEWARE

Inspired by human brain features: high-parallel data flow, generalization, robustness, self-organization, pruning, associative memory, incremental learning, genetic evolution.



Multi Layer Perceptron trained by:
- Back Propagation
- **Quasi Newton**
- Genetic Algorithm

Support Vector Machines

Genetic Algorithms

Self Organizing Maps

K-Means

Multi-layer Clustering

Principal Probabilistic Surface

Classification

Regression

Clustering

Feature Extraction

next …

Bayesian Networks

Random Decision Forest

MLP with Levenberg-Marquardt

Virgilio De Stefano

PhotoRApToR

Photometric Research Application To Redshifts

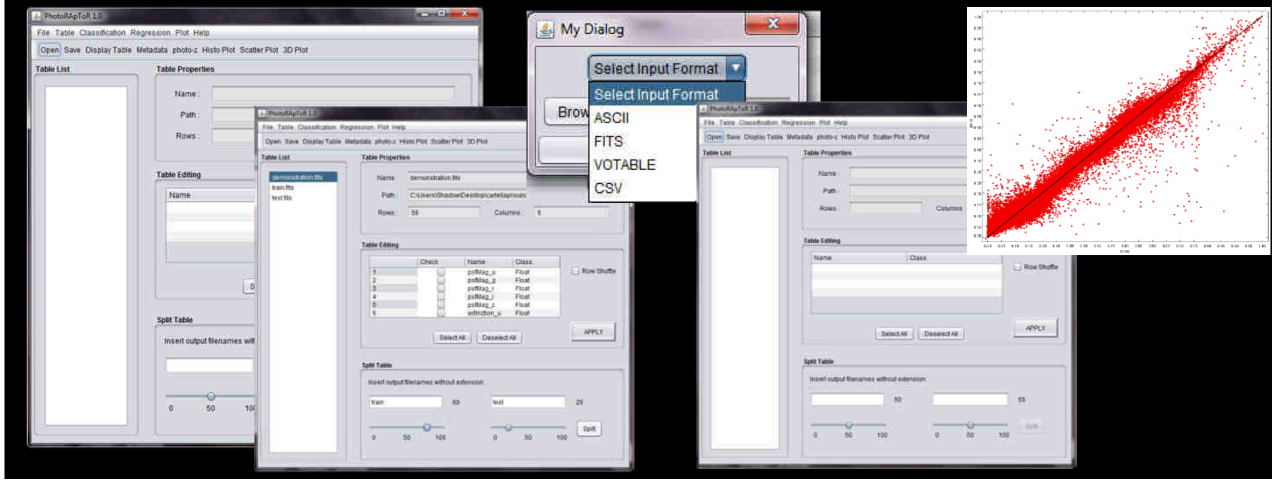
Java Desktop Application (multi-platform, Win7, Linux, Mac)

Dataset manipulation (plotting, editing, split, metadata selection, ordering, shuffling…)

MLPQNA-based photo-z estimation

General-purpose classification/regression problem solving

Post-processing (visualizing, statistical analysis)



Virgilio De Stefano

Photo-z Estimation

Pre-Processing

(on spectroscopic KB)

Feature Selection

↓

Split

Training Set | Test dataset

↓

Processing (MLPQNA)

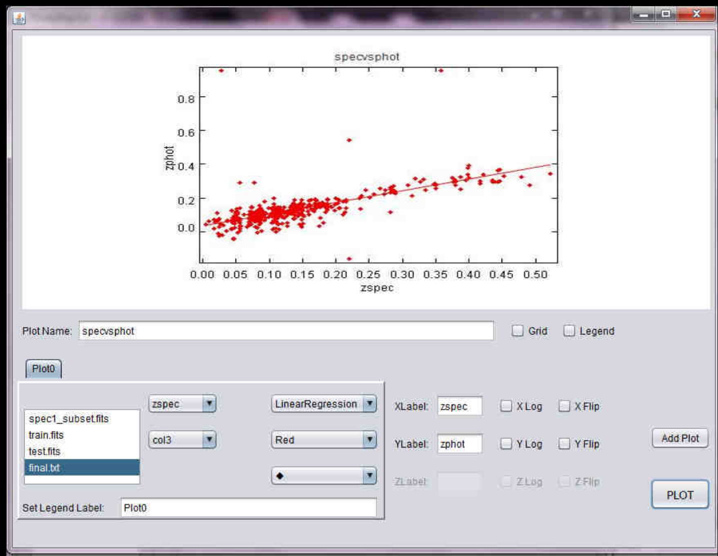
↓

Post-processing

↓

Statistical Analysis
and Generalization

ITEM	EXPERIMENTS	
	MAG	MIXED
TRAIN SET	712022 (80%)	535016 (60%)
TEST SET	178097 (20%)	356103 (40%)
INPUT FEATURES	5 magnitudes (ugriz)	4 colors + R mag
hidden layers	1	2
hidden 1 neurons	11	13
hidden 2 neurons	0	4
learning decay	0.1	0.01
hessian approx. restarts	30	60
error threshold	0.0001	0.0001
max iterations at each restart	4000	30000

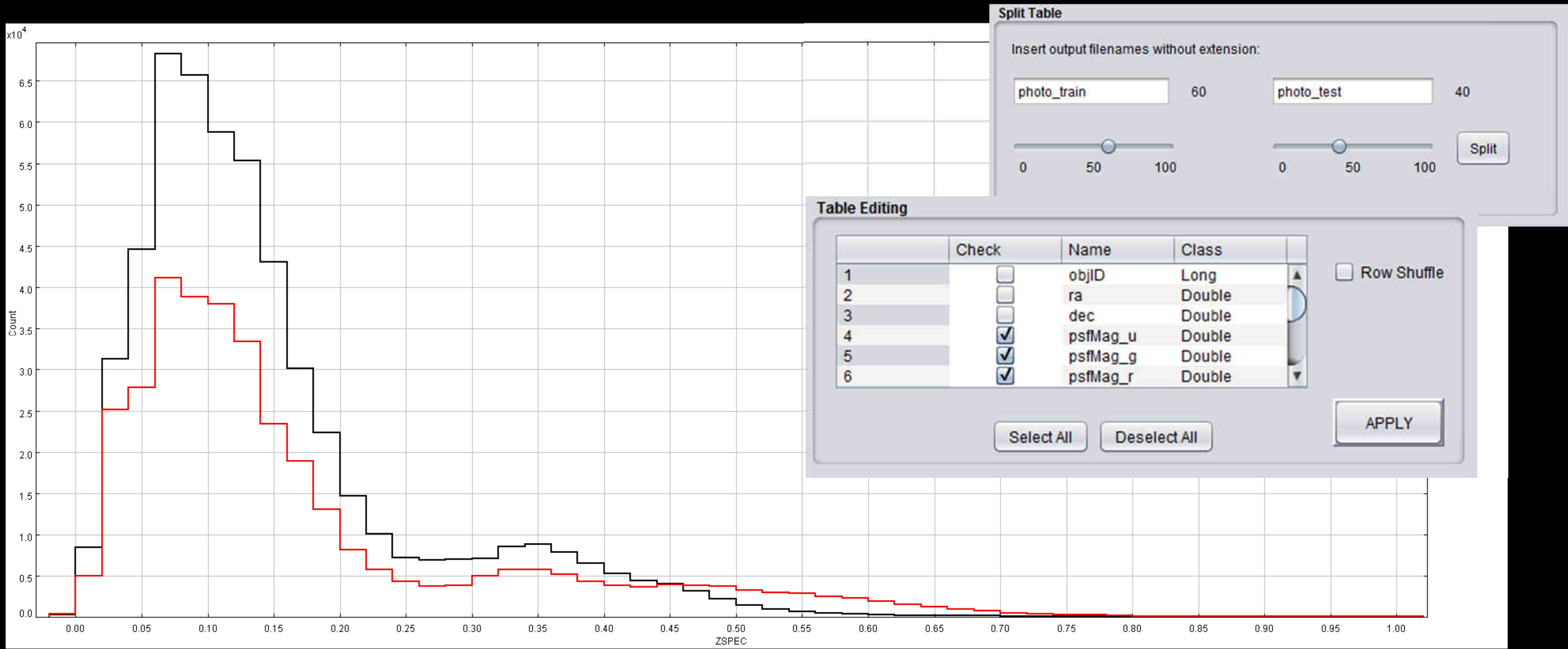


Virgilio De Stefano

Knowledge Base

The Knowledge Base (KB) data were extracted from the spectroscopic subsample of the SDSS-DR9.

The resulting KB (from the MIXED type) consisted of 890,119 objects, submitted as training+test sets to the network.



Post-Processing

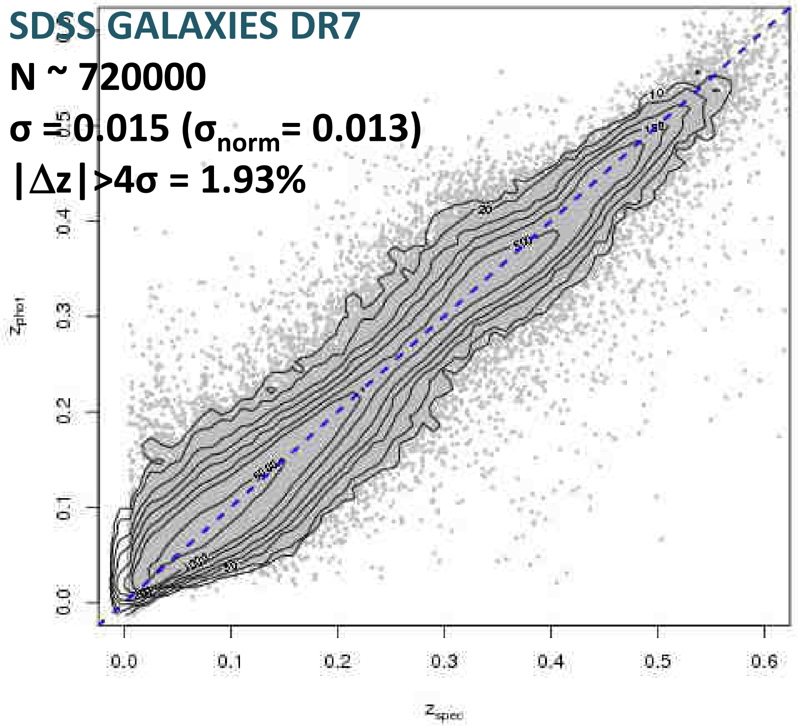
Ref.	|bias|	σ	MAD	RMS	$bias_{norm}$	$σ_{norm}$	MAD_{norm}	RMS_{norm}
TEST DATASET ONLY								
MAG	0.0002	0.016	0.001	0.016	0.0003	0.012	0.0009	0.012
MIXED	0.0004	0.014	0.001	0.014	0.0003	0.011	0.0009	0.011
Laurino et al.	0.015	0.015	0.011	0.021	0.014	0.013	0.009	0.019

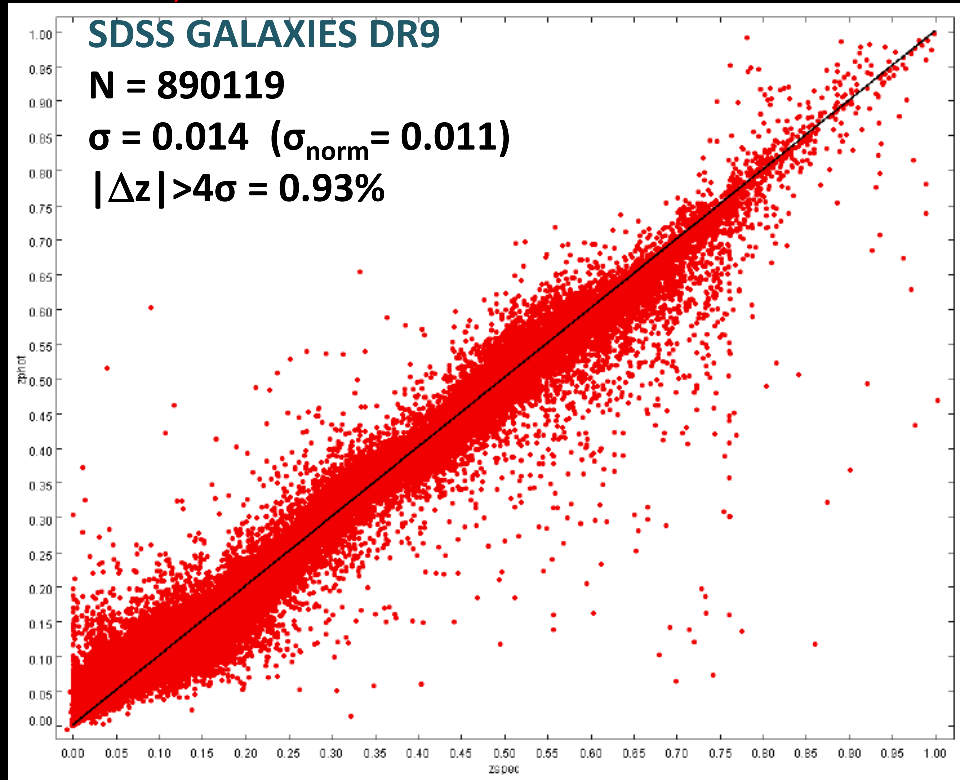
$$\Delta z = z_{spec} - z_{phot}$$

$$\Delta z_{norm} = \frac{z_{spec} - z_{phot}}{1 + z_{spec}}$$

Brescia M., Cavuoti S., De Stefano V., Longo G., 2013, Submitted to A&A

Laurino et al. 2011, MNRAS 418, 4





Virgilio De Stefano

Catastrophic Outliers

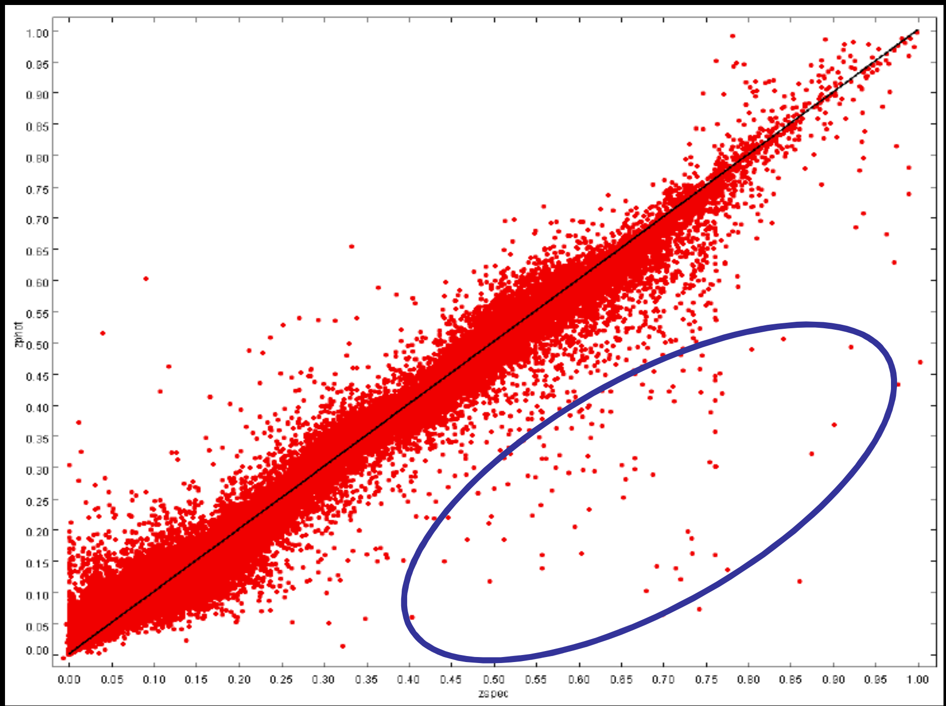
Ref.	$\lvert\Delta z\rvert > 1\sigma$	$\lvert\Delta z\rvert > 2\sigma$	$\lvert\Delta z\rvert > 3\sigma$	$\lvert\Delta z\rvert > 4\sigma$	$\lvert\Delta z_{norm}\rvert > 1\sigma$	$\lvert\Delta z_{norm}\rvert > 2\sigma$	$\lvert\Delta z_{norm}\rvert > 3\sigma$	$\lvert\Delta z_{norm}\rvert > 4\sigma$
TEST DATASET ONLY								
MAG	9.59%	4.01%	1.82%	0.95%	11.54%	4.52%	2.09%	0.92%
MIXED	9.65%	3.76%	1.76%	0.93%	10.62%	4.33%	2.03%	1.05%

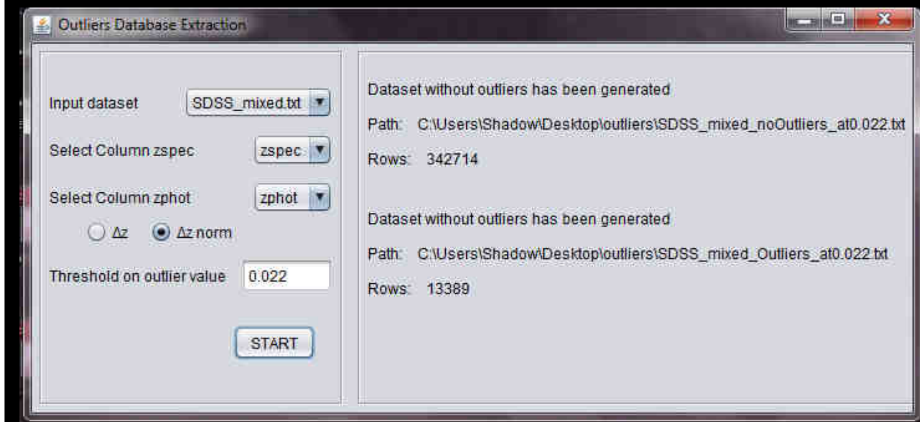
$$\Delta z = z_{spec} - z_{phot}$$

$$\Delta z_{norm} = \frac{z_{spec} - z_{phot}}{1 + z_{spec}}$$

Catastrophic outliers are objects such that $|\Delta z_{norm}| > 2\sigma(\Delta z_{norm})$

By removing catastrophic outliers, the normalized standard deviation decreases from 0.011 to 0.005

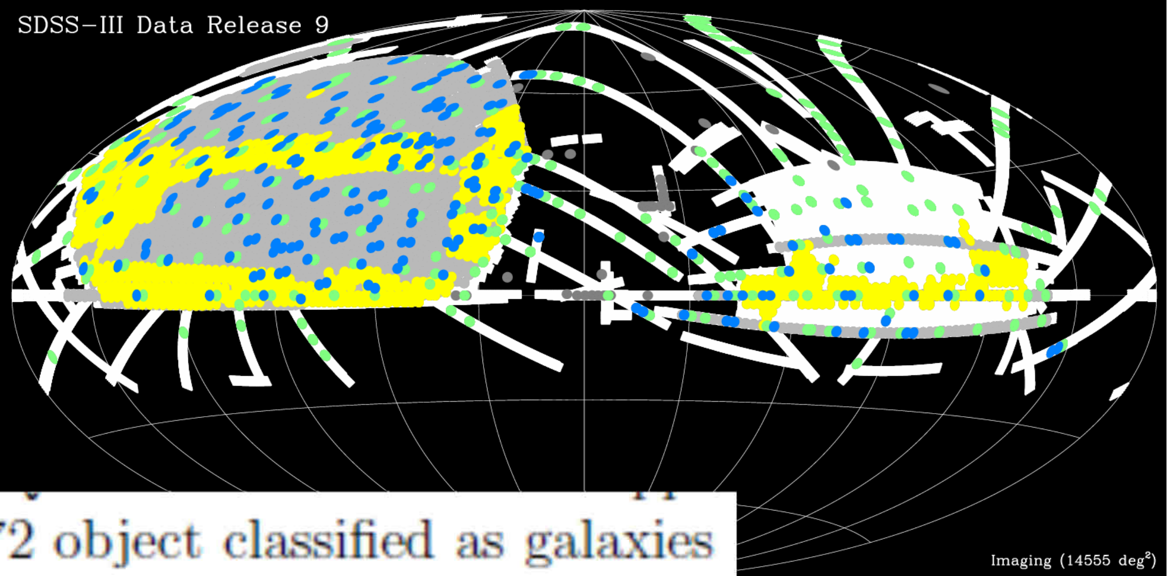




Virgilio De Stefano

SDSS DR9 Complete Catalogue

```
SELECT
  objID, ra, dec,
  psfMag_u, psfMag_g, psfMag_r, psfMag_i, psfMag_z,
  psfMagErr_u, psfMagErr_g, psfMagErr_r, psfMagErr_i,
psfMagErr_z,
  extinction_u, extinction_g, extinction_r, extinction_i,
extinction_z
FROM Galaxy
WHERE
  (dec between -30° and 85°)
  AND clean = 1
  AND psfMag_r < 23.6
```



133,923,672 object classified as galaxies

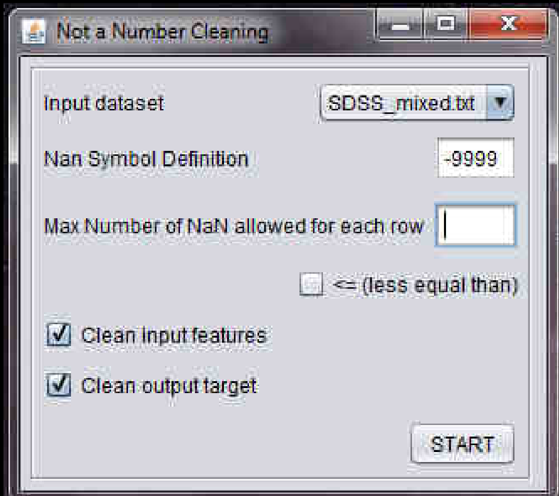
SDSS DR9 Complete Catalogue

The complete catalogue of photo-z has been produced by submitting all DR9 galaxies (with zspec not available) to the trained MLPQNA network

original file (18 cols)	objects	RA range		DEC range		NaN presence						clean objects
						U	G	R	I	Z	affected objects	
DEC(-30_-20)	601938	0.0001	359.9999	-25.04	-20				1		1	601937
DEC(-20_-15)	1481224	0.0001	359.9999	-20	-15				1		1	1481223
DEC(-15_-08)	3680861	0.00002	360	-15	-8		3		1	1	5	3680856
DEC(-08_-05)	3592643	0.00008	359.9999	-8	-5	3	1			2	5	3592638
DEC(-05_-03)	2604477	0.00001	359.9999	-5	-3		10				10	2604467
DEC(-03_-01)	4219721	0.00009	359.9998	-3	-1	69	151		97	116	256	4219465
DEC(-01_00)	2998898	0.000004	359.9997	-1	0	507	575		127	344	716	2998182
DEC(00_01)	3032526	0.00001	359.9999	0	1	453	529		255	282	612	3031914
DEC(01_02)	2452964	0.00016	359.9999	1	2	178	156		80	125	227	2452737
DEC(02_04)	4610038	0.00005	360	2	4	1	2		11	13	27	4610011
DEC(04_06)	4783288	0.00011	359.9999	4	6				7		7	4783281
DEC(06_08)	4859462	0.00005	359.9999	6	8				3	2	5	4859457
DEC(08_09)	2288884	0.00001	359.9999	8	9	1			1		2	2288882
DEC(09_10)	2240853	0.00008	359.9998	9	10					1	1	2240852
DEC(10_11)	2222816	0.00013	359.9997	10	11		1			1	2	2222814
DEC(11_12)	2159483	0.0001	359.9999	11	12		1		1		2	2159481
DEC(12_14)	4466312	0.00002	359.9999	12	14	2			1		3	4466309
DEC(14_16)	4670192	0.00002	359.9999	14	16	1	29		5	1	36	4670156
DEC(16_18)	4586435	0.00011	359.9998	16	18		1		3		4	4586431
DEC(18_20)	4500830	0.00001	359.9999	18	20	2	4		5	3	11	4500819
DEC(20_22)	4362582	0.00014	359.9999	20	22	1	2		2		5	4362577
DEC(22_24)	4428773	0.00009	359.9999	22	24		1		2		3	4428770
DEC(24_26)	4494914	0.000003	359.9999	24	26	1	3		2	1	6	4494908
DEC(26_28)	4314904	0.00012	360	26	28	1	3		3	1	8	4314896
DEC(28_30)	4174238	0.00004	359.9999	28	30		4				4	4174234
DEC(30_32)	3957318	0.00004	359.9998	30	32		1		2	1	3	3957315
DEC(32_34)	3693310	0.00003	359.9999	32	34	3			8	1	11	3693299
DEC(34_37)	5093590	0.000007	359.9999	34	37		2				2	5093588
DEC(37_40)	4894931	0.48362	359.6878	37	40	6	15		4	1	26	4894905
DEC(40_43)	4606853	7.32184	358.7502	40	43				1		1	4606852
DEC(43_46)	4339477	10.39352	357.7354	43	46		1		4		5	4339472
DEC(46_50)	4589912	13.8513	359.6718	46	50		2		2		4	4589908
DEC(50_55)	4766875	21.66491	359.5998	50	55		1		2	1	4	4766871
DEC(55_60)	4013134	22.92468	359.1274	55	60	1	1		1	1	3	4013131
DEC(60_65)	3481079	24.54303	358.5617	60	65	1	2		4		7	3481072
DEC(65_70)	1516317	26.71197	357.8400	65	70		1		1		2	1516315
DEC(70_75)	351332	29.91958	356.7731	70	75						0	351332
DEC(75_85)	792316	35.13998	354.9676	75	84.97				1		1	792315
TOTAL	**133925700**										**2028**	**133923672**

The DR9 galaxy catalogue consists of ~134 million objects, divided into 38 files, ordered on the DEC range

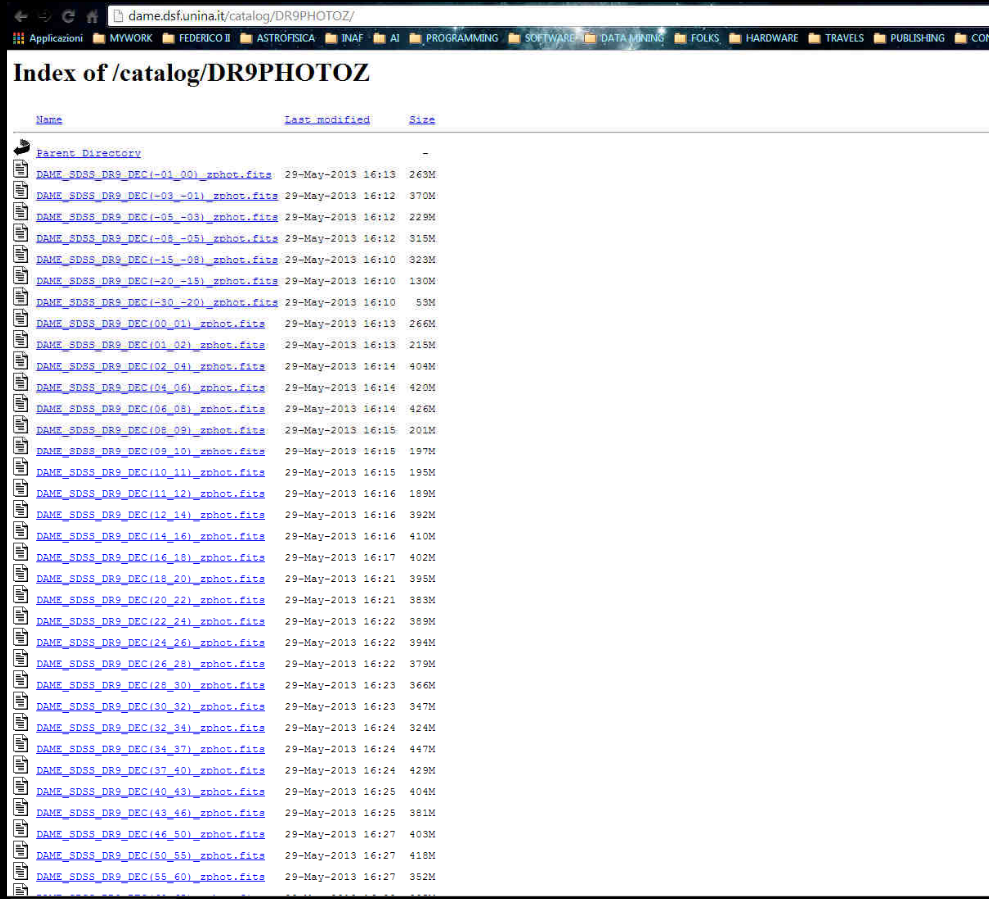
It covers ~25% of the northern sky emisphere



Virgilio De Stefano

SDSS DR9 Photo-z Catalogue

Catalog File (DEC range)	objects	photo-z range	
		MIN	MAX
DEC(-30_-20)	601937	0	0.774
DEC(-20_-15)	1481223	0	0.768
DEC(-15_-08)	3680856	0	0.774
DEC(-08_-05)	3592638	0	0.772
DEC(-05_-03)	2604467	0	0.772
DEC(-03_-01)	4219465	0	0.776
DEC(-01_00)	2998182	0	0.776
DEC(00_01)	3031914	0	0.771
DEC(01_02)	2452737	0	0.776
DEC(02_04)	4610011	0	0.782
DEC(04_06)	4783281	0	0.781
DEC(06_08)	4859457	0	0.779
DEC(08_09)	2288882	0	0.765
DEC(09_10)	2240852	0	0.772
DEC(10_11)	2222814	0	0.779
DEC(11_12)	2159481	0	0.779
DEC(12_14)	4466309	0	0.776
DEC(14_16)	4670156	0	0.779
DEC(16_18)	4586431	0	0.778
DEC(18_20)	4500819	0	0.779
DEC(20_22)	4362577	0	0.780
DEC(22_24)	4428770	0	0.777
DEC(24_26)	4494908	0	0.782
DEC(26_28)	4314896	0	0.775
DEC(28_30)	4174234	0	0.777
DEC(30_32)	3957315	0	0.774
DEC(32_34)	3693299	0	0.775
DEC(34_37)	5093588	0	0.774
DEC(37_40)	4894905	0	0.771
DEC(40_43)	4606852	0	0.770
DEC(43_46)	4339472	0	0.780
DEC(46_50)	4589908	0	0.777
DEC(50_55)	4766871	0	0.775
DEC(55_60)	4013131	0	0.776
DEC(60_65)	3481072	0	0.778
DEC(65_70)	1516315	0	0.769
DEC(70_75)	351332	0	0.767
DEC(75_85)	792315	0	0.764
TOTAL	133923672	0	0.782



http://dame.dsf.unina.it/catalog/DR9PHOTOZ/

Virgilio De Stefano

Conclusions

- **Best photometric redshifts in literature**
  - ✓ **$\sigma = 0.014$ ($\sigma_{norm} = 0.011$)**
  - ✓ **$|\Delta z| > 2\sigma = 3.76\%$**

- **Two refereed papers**
  - ✓ *Brescia M., Cavuoti S., De Stefano V., Longo G., 2013. A Catalogue of photometric redshifts for the SDSS DR9 galaxies, Submitted to A&A*
  - ✓ *De Stefano V., Cavuoti S., Brescia M., Longo G., 2014. Photometric redshift estimation with PhotoRApToR, in preparation, to be submitted to Astronomy and Computing*

- **A catalog of ~134 million photometric redshifts**
  - ✓ *http://dame.dsf.unina.it/catalog/DR9PHOTOZ/*

- **A public application for photo-z estimation**
  - ✓ Multi-platform (Java-based)
  - ✓ Expandable
  - ✓ Scalable on massive data sets
  - ✓ based on empirical machine learning methods

Thanks for your attention