Photometric Redshift Estimation of Quasars by Machine Learning

Yanxia Zhang

Collaborators: Jingyi Zhang, Bo Han, Lina Qiao, Yongheng Zhao

National Astronomical Observatories, CAS

ADASS October 8, 2019

Outline

- Quasars
- Quasars in large sky surveys
- Photometric redshifts
- Application of ML in photo-z
- Conclusions

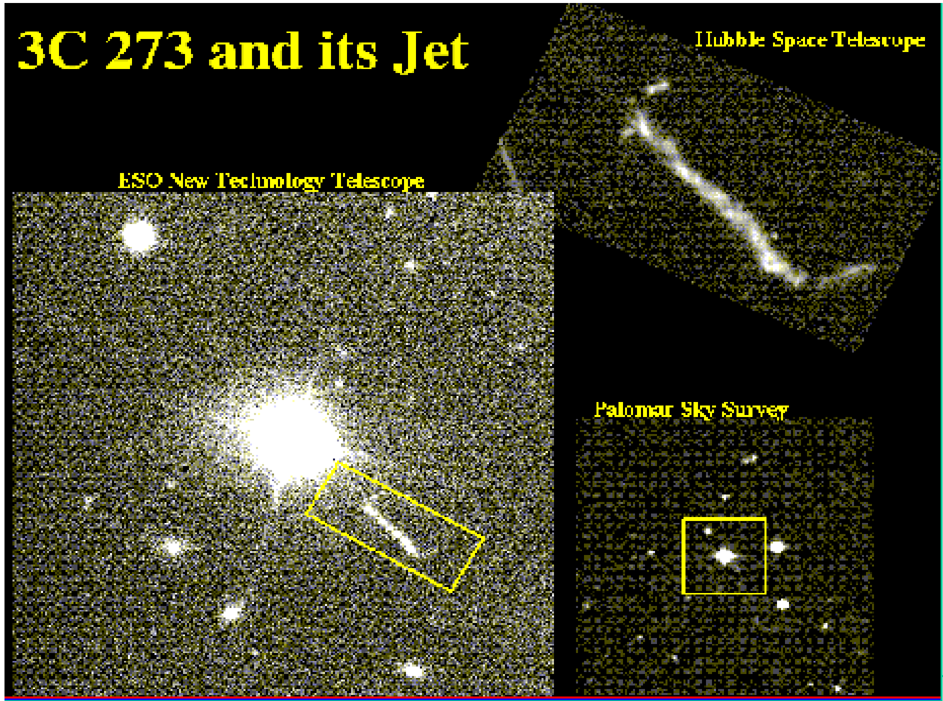
Quasars

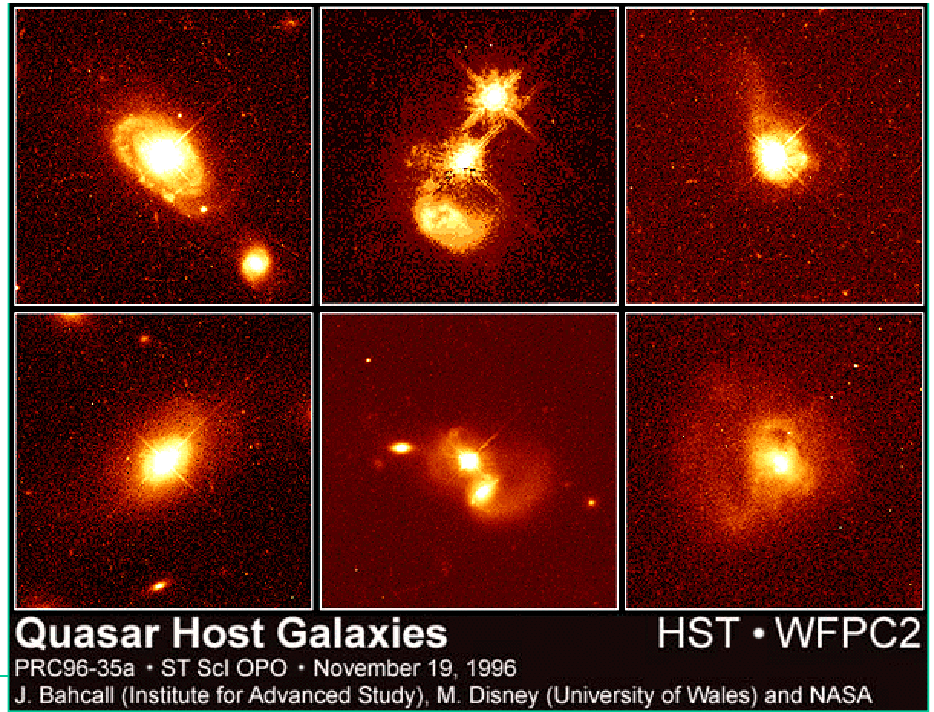
- Originally discovered in radio survey (3C) in 1950s; first identified as star-like optical sources with emission lines in 1963; Maarten Schmidt (1963) realized the redshift of 3C 273 (z=0.158)
- First named simply as "quasi-stellar radio sources", shorten to "quasars" by Hong-Yi Chiu (1964), accepted by ApJ in 1970
- the most distant known quasar is at redshift z=7.085 (Mortlock, D. J.; et al. 2011,Nature)

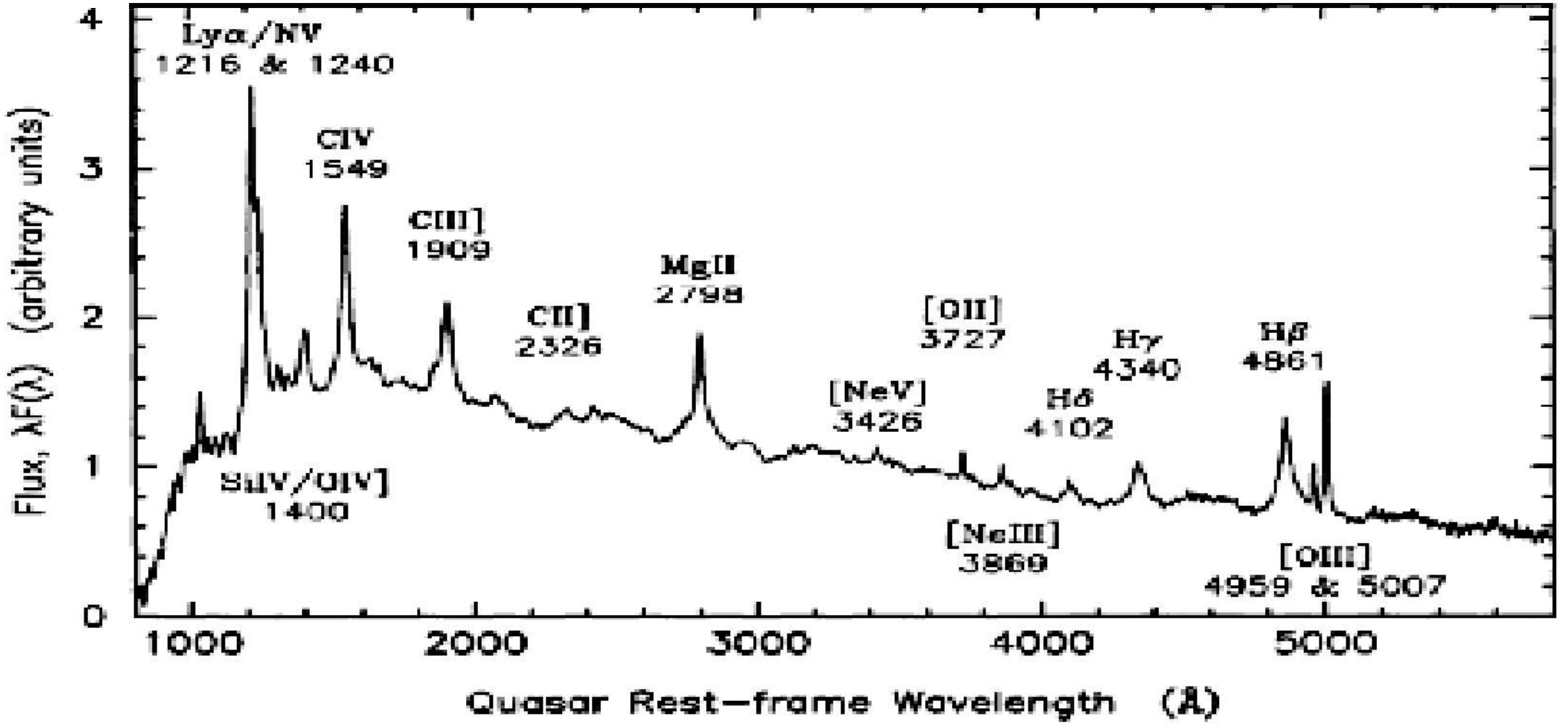
Quasars

- AGN
- Powered by central black hole
- High redshift (0.1~7)
- High luminosity~10^42-10^46erg/s
- Luminosity variability
- Full spectrum emission
- Pointed sources
- Strong, broad emission line spectra

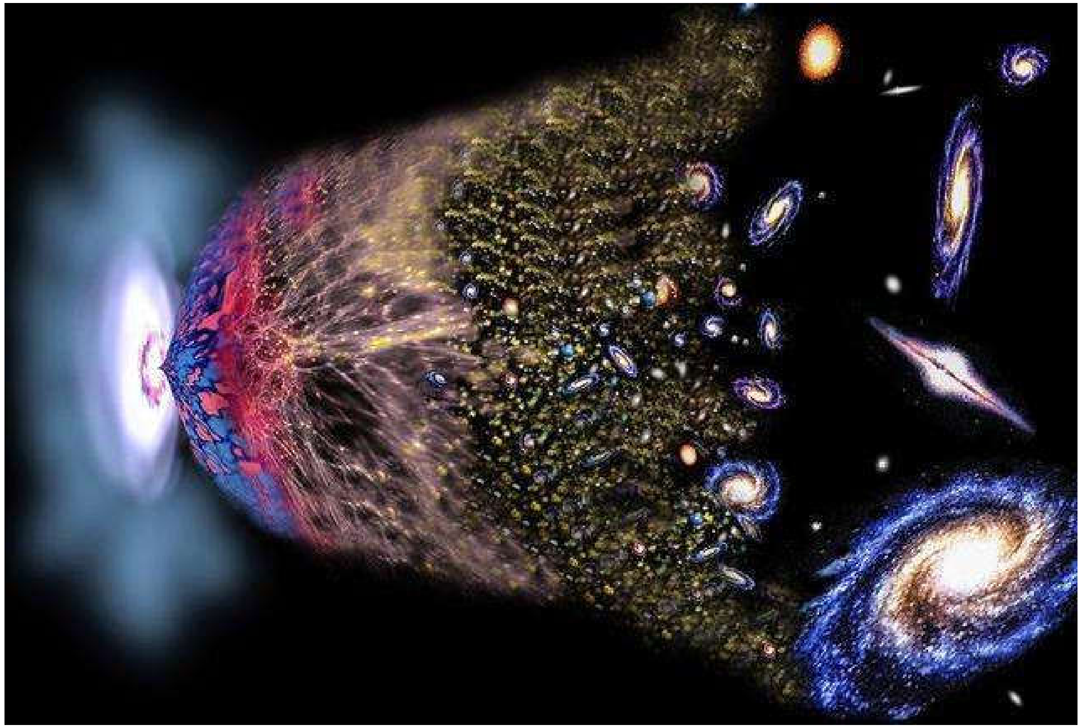
Credit: ESO/M. Kornmesser

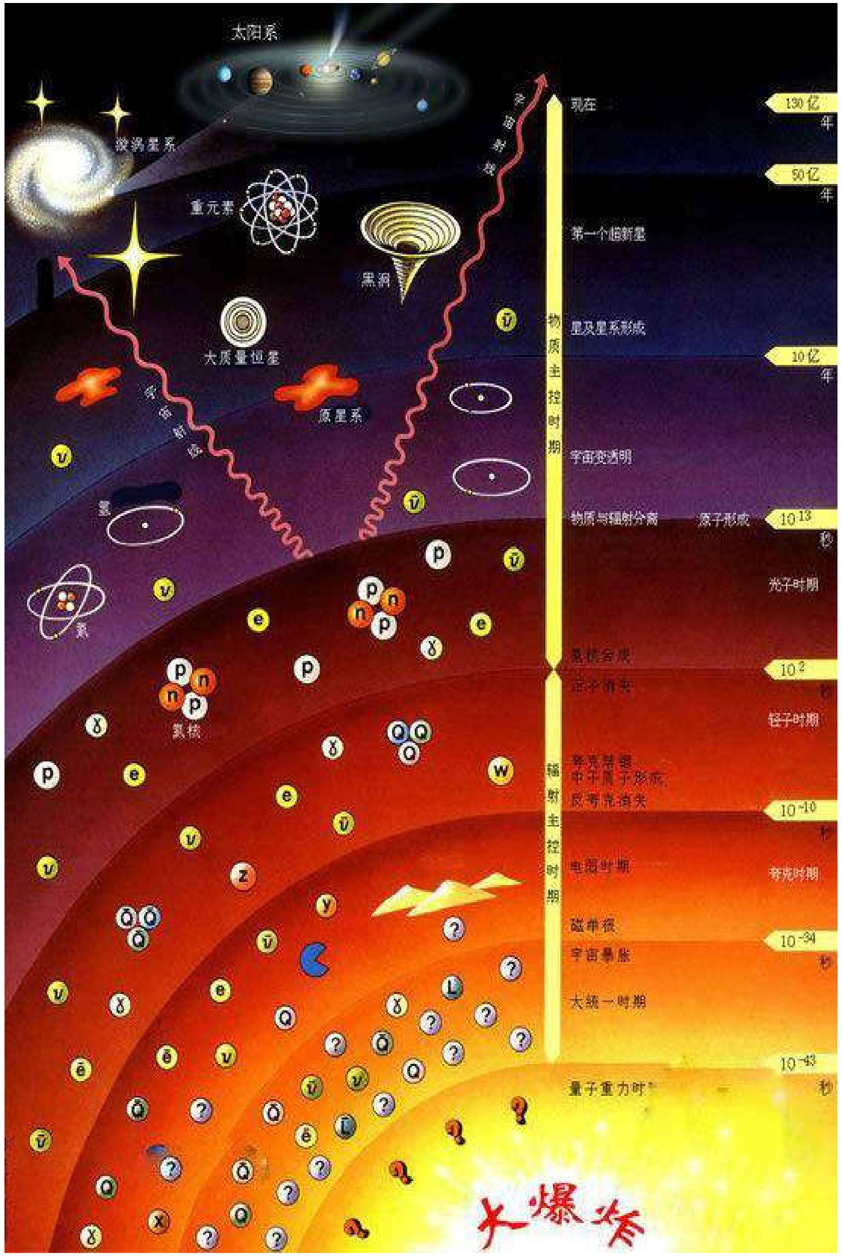

3C 273 and its Jet
Hubble Space Telescope
ESO New Technology Telescope
Palomar Sky Survey


Quasar Host Galaxies
HST • WFPC2
PRC96-35a • ST ScI OPO • November 19, 1996
J. Bahcall (Institute for Advanced Study), M. Disney (University of Wales) and NASA


Lyα/NV
1216 & 1240
CIV
1549
C[III]
1909
C[II]
2326
MgII
2798
[OII]
3727
[NeV]
3426
Hδ
4102
Hγ
4340
Hβ
4861
SiIV/OIV]
1400
[NeIII]
3869
[OIII]
4959 & 5007
Flux, λF(λ) (arbitrary units)
Quasar Rest-frame Wavelength (Å)
0
1
2
3
4
1000
2000
3000
4000
5000

The first quasar and the cosmic reionization

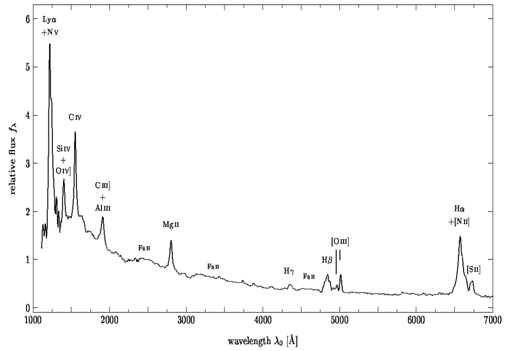


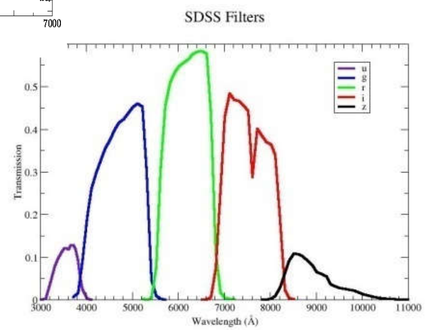


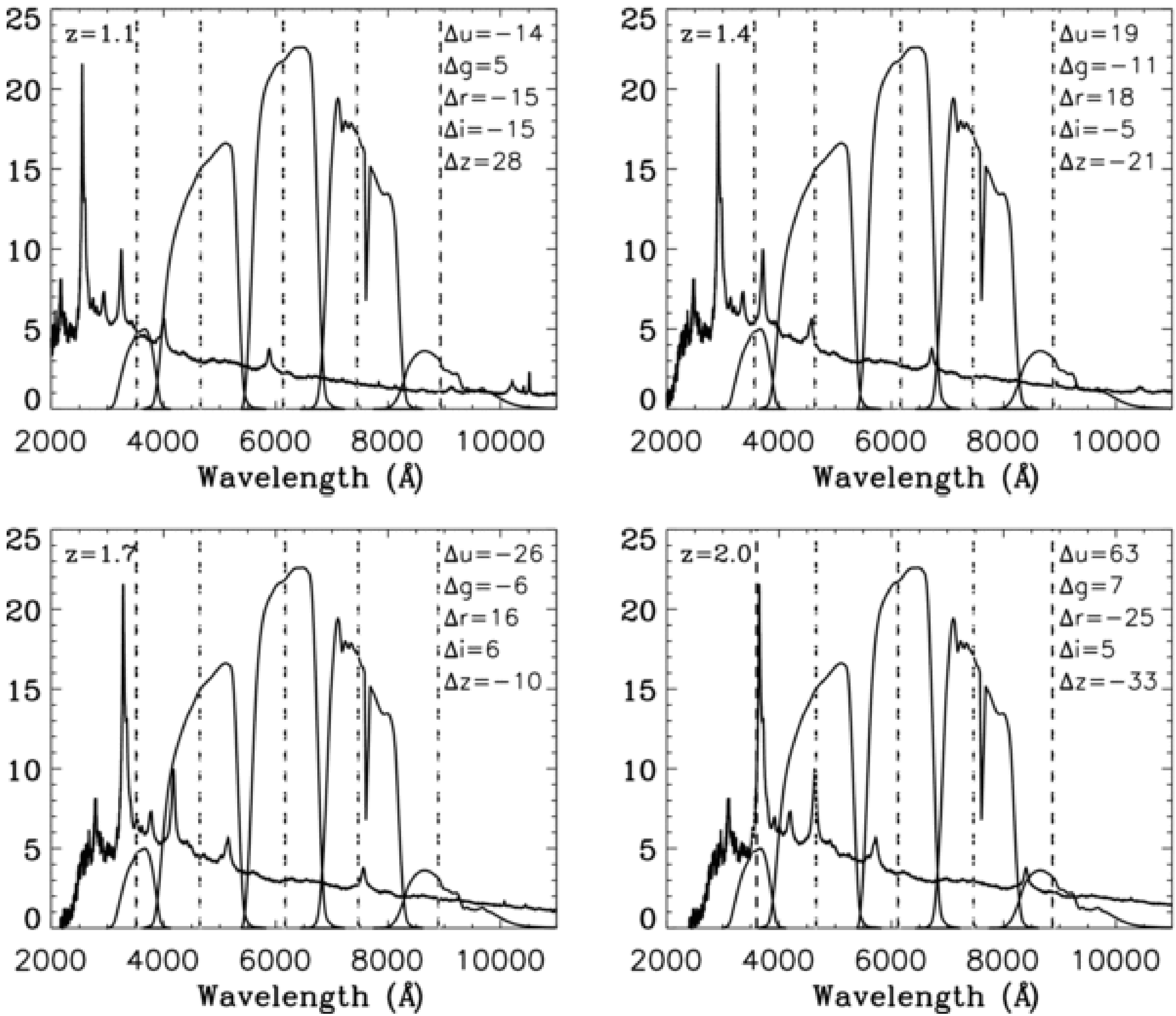
Photometric redshifts (photo-zs)

- **Photo-zs are determined from the fluxes (or magnitudes or colors) of galaxies through a set of filters**
- **May be thought of as redshifts from (very) low-resolution spectroscopy**
- **Photo-z' are needed in particular when it's too observationally expensive to get spectroscopic redshifts (e.g., if galaxies are too many or too faint)**
- **Well-calibrated photo-z's are a key ingredient to obtaining cosmological constraints in large photometric surveys like DES and LSST**

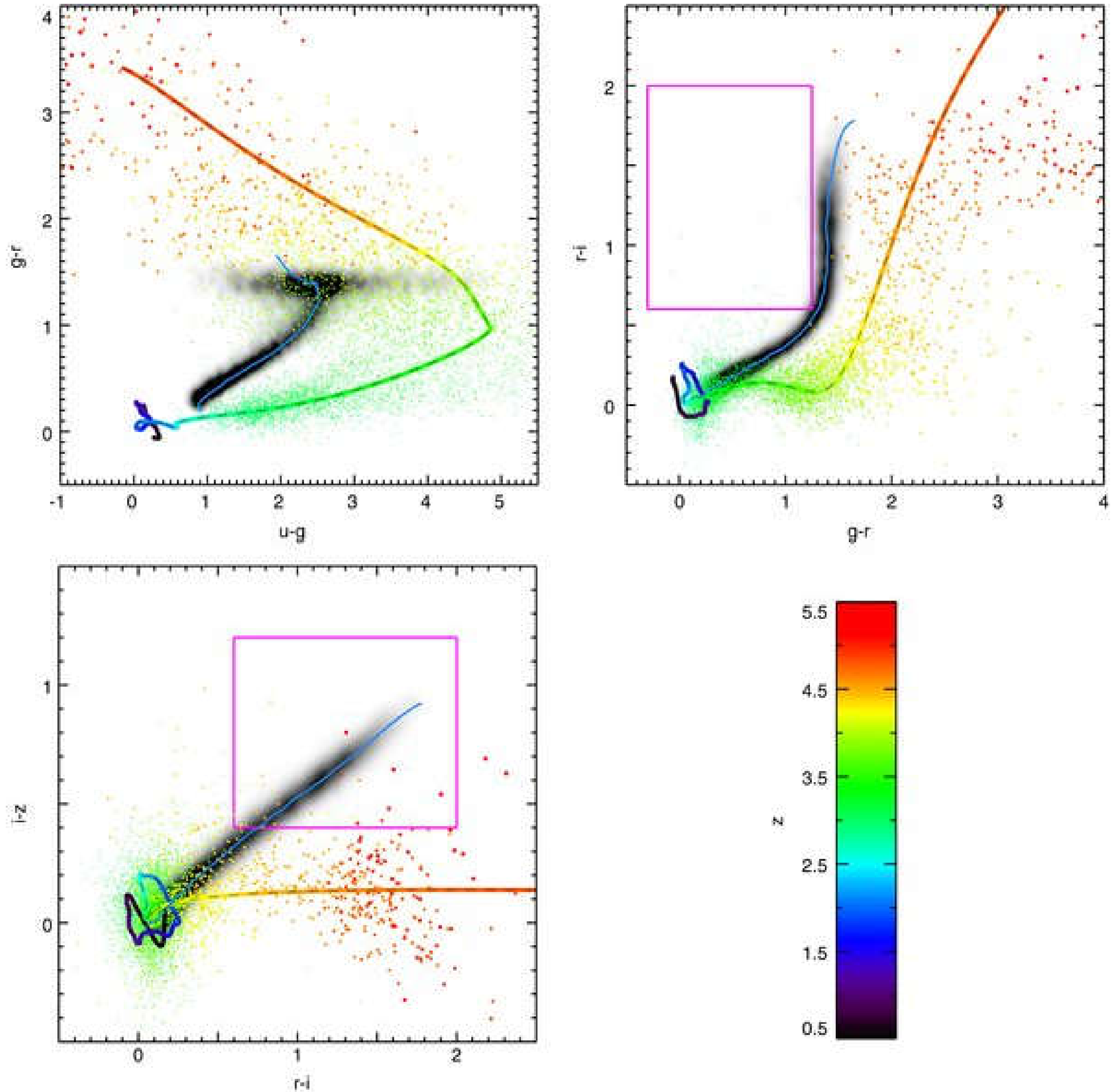
- For example, SDSS
  - Spectra: $r < 17.7$ 1.6M sources
  - Photo: $r < 21+$ 360M sources







z=1.1
Δu=−14
Δg=5
Δr=−15
Δi=−15
Δz=28
z=1.4
Δu=19
Δg=−11
Δr=18
Δi=−5
Δz=−21
z=1.7
Δu=−26
Δg=−6
Δr=16
Δi=6
Δz=−10
z=2.0
Δu=63
Δg=7
Δr=−25
Δi=5
Δz=−33
Wavelength (Å)



g-r
u-g
r-i
g-r
i-z
r-i
z
5.5
4.5
3.5
2.5
1.5
0.5

Large Optical Sky Quasar Surveys

- ***Palomar-Green (PG) Bright Quasar Survey (BQS):*** **B<16, 10000 deg^2, ~120 Quasars (~7%)**
- ***Large Bright Quasar Survey (LBQS):*** **B<17.5, ~10^3 quasars**
- ***2dF:*** **200 deg^2, U-V<-0.3, ~2.6x10^4 quasars**
- ***Sloan Digital Sky Survey(SDSS):*** **~5.3x10^5 quasars**
- **LAMOST survey:~5x10^4 quasars**

Photometric survey

UV: *GALEX* at 1530 A˚(FUV) and 2310 A˚ (NUV)

Optical: SDSS (*u* 3551Å, *g* 4686Å, *r* 6165Å, i 7481Å, z 8931Å)

Pan-STARRS (g 4866 A˚, r 6215 A˚, i 7545 A˚, z 8679 A˚, y 9633 A˚)

LSST(u, g, r, i, z, and y)

Infrared: 2MASS at J-band (1.235 μm); H-band (1.662 μm); Ks-band (2.159 μm),

WISE at 3.4, 4.6, 12, and 22 μm (W1, W2, W3, W4) (W1,W2,W3,W4)

UKIDSS (y,j,h,k)

----etc.

Methods for Photometric redshifts

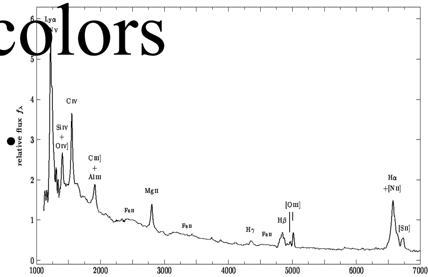
- Template fitting
- Machine learning /training set/empirical methods

In the era of astronomical big data, ML is a must!

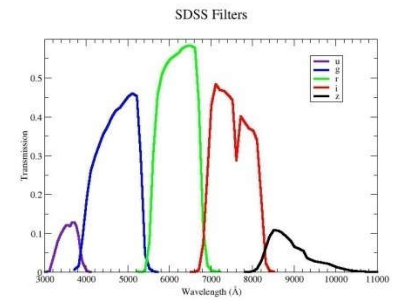
Template fitting

- Model SEDs are being redshifted by

  λ(z) = λrest(1 + z) for various values of z.
- Then the spectrum is projected through the filter throughputs to obtain a simulated photometric observation of a galaxy with that SED.
- Searches for the minimum value of the difference between observed colors and synthetic colors derived from model (or template) SEDs.



$$\chi^2 = \sum_{k=1}^{N\,filters} \left(\frac{F_{obs,k} - p \cdot \mathrm{SED}_k(z)}{\sigma_k} \right),$$



- Template: synthetic or observed

Pros and cons of Template fitting

- Physical meaning is obvious.
- Easy to explain.
- Go very deep, well beyond the spec-z limit.
- No training set needed.

- Arbitrary choice of template, lots of assumptions on physics, strong dependence on zero points
- Not too accurate.

Machine learning

- ML is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.
- Need to learn or train and obtain a relation between photometric observables of a galaxy and its spec-z.
- A subset of the objects as training set, their spec-zs are known.
- Training set, test set, validation set.

Pros and cons of ML

- More accurate
- No assumptions on physics, almost independent on zero points, photometric calibration, etc.

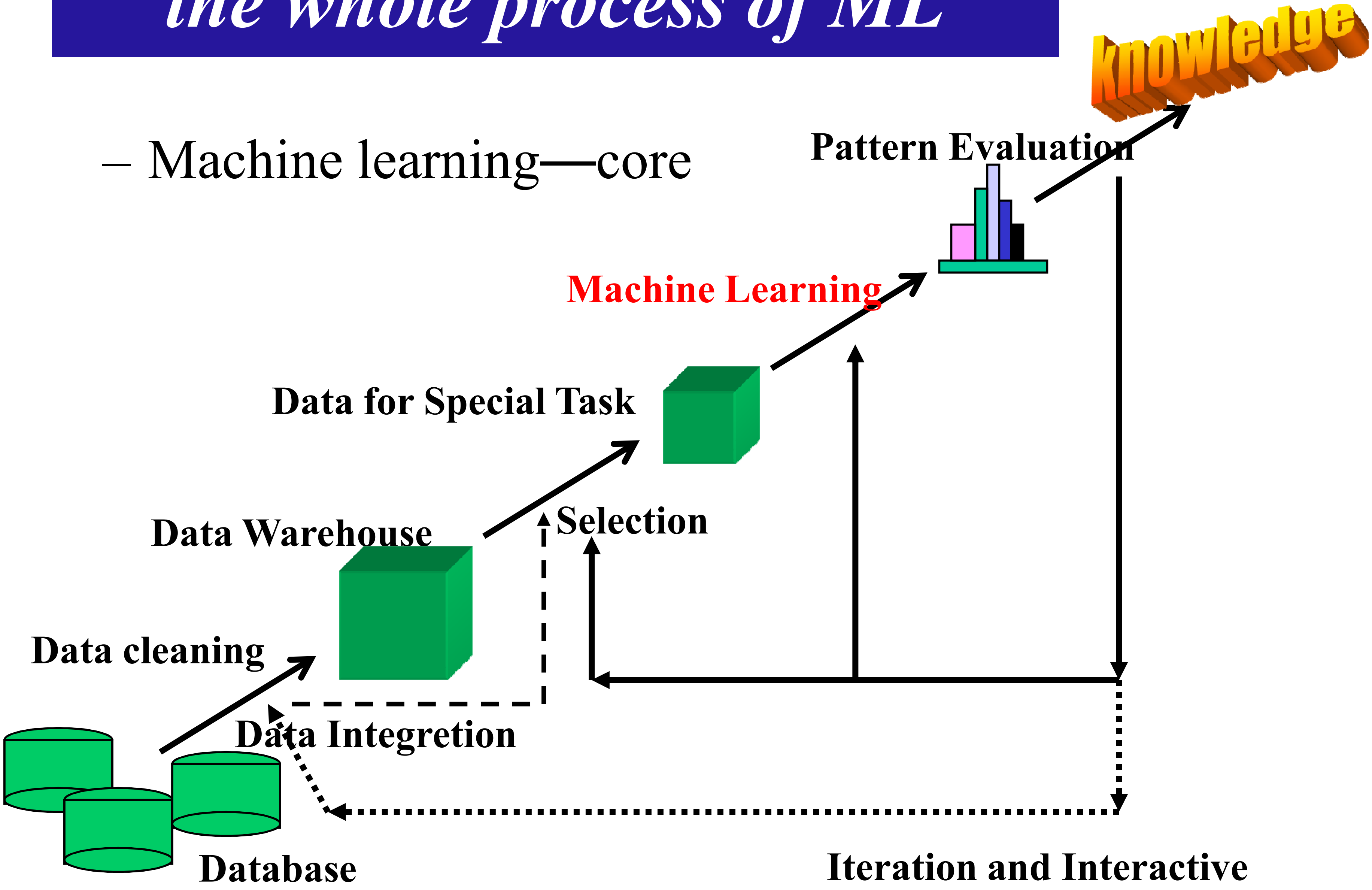
- Difficult to understand
- Bounded by the spec-z limit
- Unreliable extrapolation
- Retraining for every survey

Commonly Used ML

- **Polynomial fitting**, e.g. Connolly et al. (1995)
- **Piece-wise fitting of 2nd order polynomials**, e.g. Brunner et al. (1997)
- **a linear function** of three photometric colors ,e.g. Wang et al. (1998)
- **k nearest neighbors**, e.g. Csabai et al. (2003) ; Han et al. (2016)
- **Kernel regression**, e.g. Wang et al. (2006)
- **Support Vector Machines**, e.g., Wadadekar (2005) ;Zheng & Zhang (2012)
- **Relevance Vector Machines**, e.g., Sanchez et al. (2014)
- **Boosted Decision Trees**, e.g., Gerdes (2009, ArborZ)
- **Gaussian Processes**, e.g., Way et al. (2009)
- **Diffusion Maps**, e.g., Richards et al. (2009) and Freeman et al. (2009)
- **Random Forests**, e.g., Carliles et al. (2010)
- **Self Organizing Maps**, e.g., Carrasco Kind & Brunner (2014)
- **Artificial Neural Network** , e.g. Li et al. (2007), Zhang et al. (2008), Collister & Lahav, 2004
- et al.

the whole process of ML

– Machine learning—core



Our work

The factors to influence performance of Photo-Z estimation!

- **Data preprocessing (feature engineering).**
- **Data quality.**
- **Algorithms.**
- **Separation of training sample.**

Samples

The DR14Q catalog contains 526,356 unique quasars

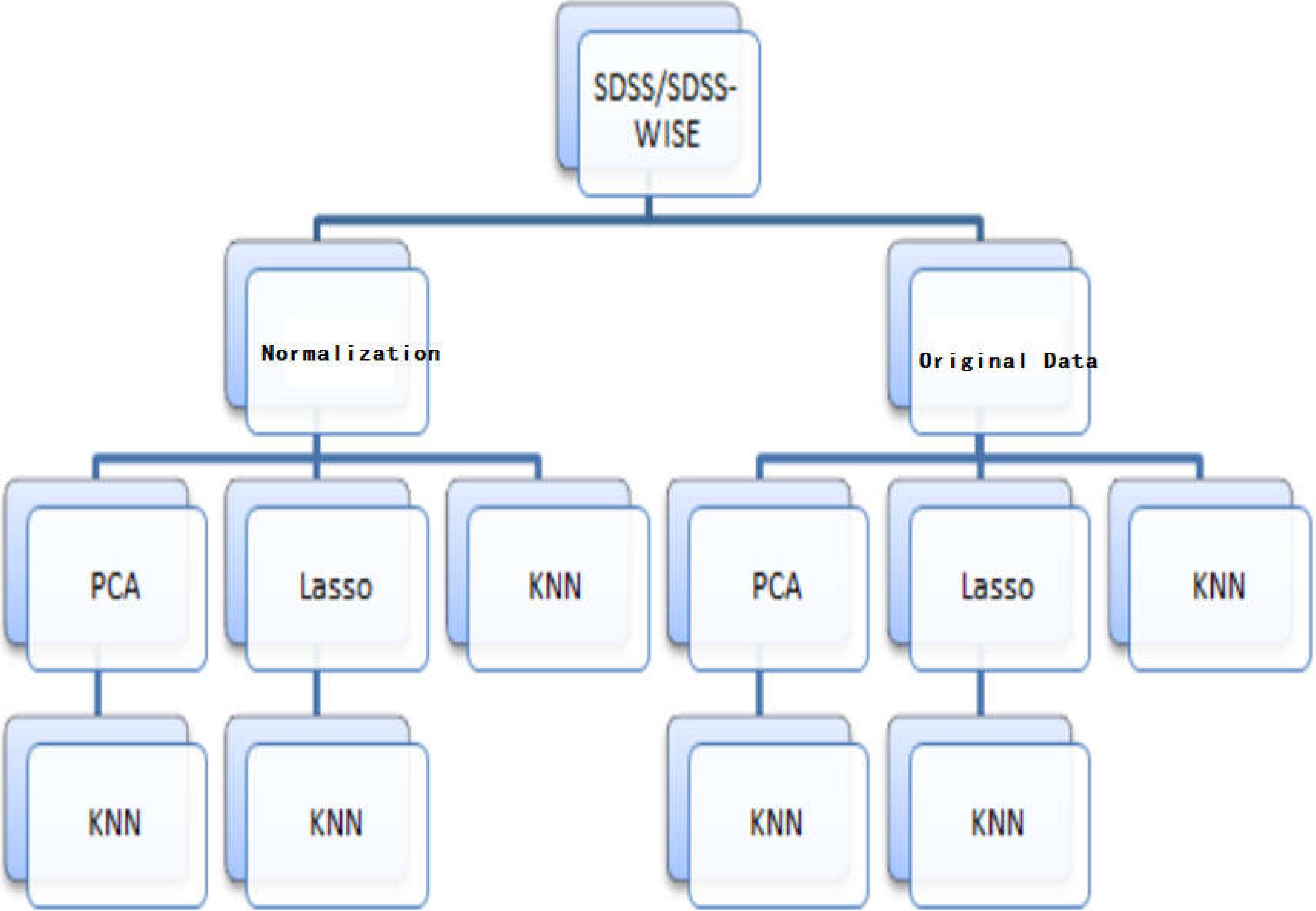
Better quality

Sample	Number
SDSS	336357
SDSS + WISE	261705

Good quality

Sample	Number
SDSS	445958
SDSS + WISE	324333

1 data preparation



Good quality sample

Table 1. Performance of KNN with the SDSS sample by different data preprocessing.

Normalization	PCA	LASSO	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ_{rms}
			62.18	**80.00**	**86.99**	**0.3344**
☑			62.03	79.98	86.95	0.3349
☑	☑		62.03	79.98	86.95	0.3349
	☑		62.09	79.98	86.94	0.3349
		☑	40.57	72.30	84.54	0.4594

Table 2. Performance of KNN with the SDSS-WISE sample by different data preprocessing.

Normalization	PCA	LASSO	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ_{rms}
			78.57	**91.10**	**95.20**	**0.1983**
☑			77.70	90.94	95.20	0.2031
☑	☑		77.70	90.94	95.20	0.2031
	☑		78.44	91.02	95.18	0.1991
		☑	63.16	85.06	93.07	0.2790

2 data quality

Table 3. Performance of various algorithms for the SDSS-WISE sample with r_6c

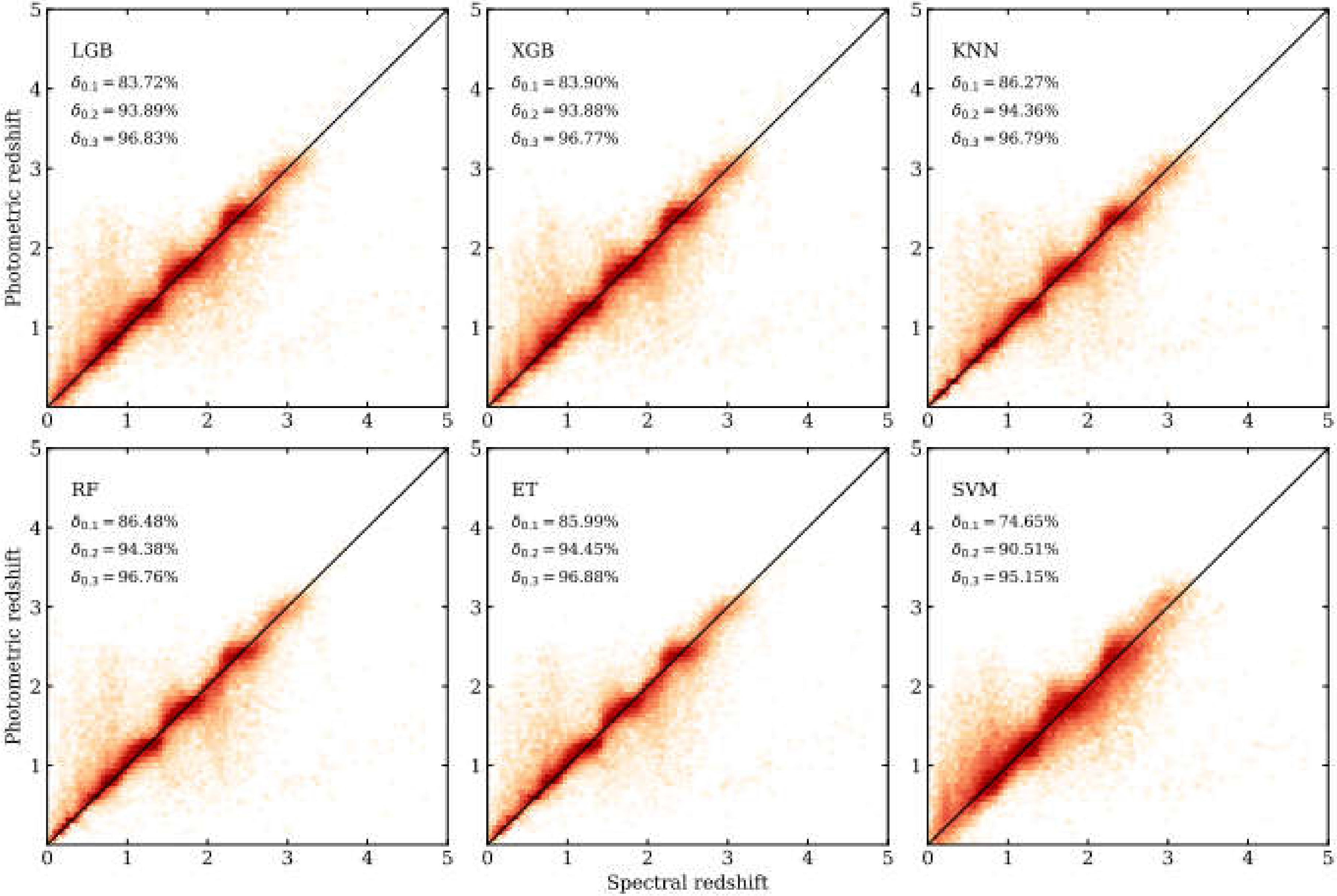
Algorithm	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ_{rms}
SVR	69.74	88.15	93.81	0.2384
XGBoost	77.43	90.81	95.27	0.2007
KNN	79.40	91.37	95.28	0.1931
RF	79.87	91.37	95.23	0.1907
better quality data				
SVR	72.40	90.35	95.15	0.016
XGBoost	83.90	93.88	96.77	0.013
KNN	86.27	94.36	96.79	0.013
RF	86.48	94.38	96.76	0.012

3 algorithms

Better quality sample

Method	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	MAE	MSE	R^2	Time(S)
SVM	74.65	90.51	95.15	0.193	0.092	0.827	3587.45
LGB	83.72	93.89	96.83	0.088	0.013	0.867	3.24
XGB	83.90	93.88	96.77	0.088	0.013	0.865	56.24
KNN	86.27	94.36	96.79	0.078	0.013	0.862	10.87
RF	86.48	94.38	96.76	0.075	0.012	0.866	106.36
ET	85.99	94.45	96.88	0.078	0.012	0.871	17.45

XGB: XGBoost; LGB: LightGBM;NN: k-nearest neighbor;
RF: random forest; ET: Extremely randomized trees
SVM: support vector machine



LGB
$\delta_{0.1} = 83.72\%$
$\delta_{0.2} = 93.89\%$
$\delta_{0.3} = 96.83\%$
XGB
$\delta_{0.1} = 83.90\%$
$\delta_{0.2} = 93.88\%$
$\delta_{0.3} = 96.77\%$
KNN
$\delta_{0.1} = 86.27\%$
$\delta_{0.2} = 94.36\%$
$\delta_{0.3} = 96.79\%$
RF
$\delta_{0.1} = 86.48\%$
$\delta_{0.2} = 94.38\%$
$\delta_{0.3} = 96.76\%$
ET
$\delta_{0.1} = 85.99\%$
$\delta_{0.2} = 94.45\%$
$\delta_{0.3} = 96.88\%$
SVM
$\delta_{0.1} = 74.65\%$
$\delta_{0.2} = 90.51\%$
$\delta_{0.3} = 95.15\%$
Photometric redshift
Spectral redshift

3 Seperation of training sample

- random forest

Original Scheme

Data Set	Algorithm	Model Parameters	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ	Time(s)
SDSS	RF	*n_estimators*=300 *max_depth*=15	63.34	80.48	87.34	0.3271	37628
SDSS-WISE	RF	*n_estimators*=300 *max_depth*=20	79.87	91.37	95.23	0.1907	36762

Other two Schemes

- Firstly, classification of the sample into two subsamples according to redshift range ($0 < z \leq 2.2$, $z > 2.2$) or into four subsamples according to redshift range ($0 < z \leq 1.5$, $1.5 < z \leq 2.2$, $2.2 < z \leq 3.5$, $z > 3.5$) by random forest
- Secondly, create regressors to predict the unknown samples.

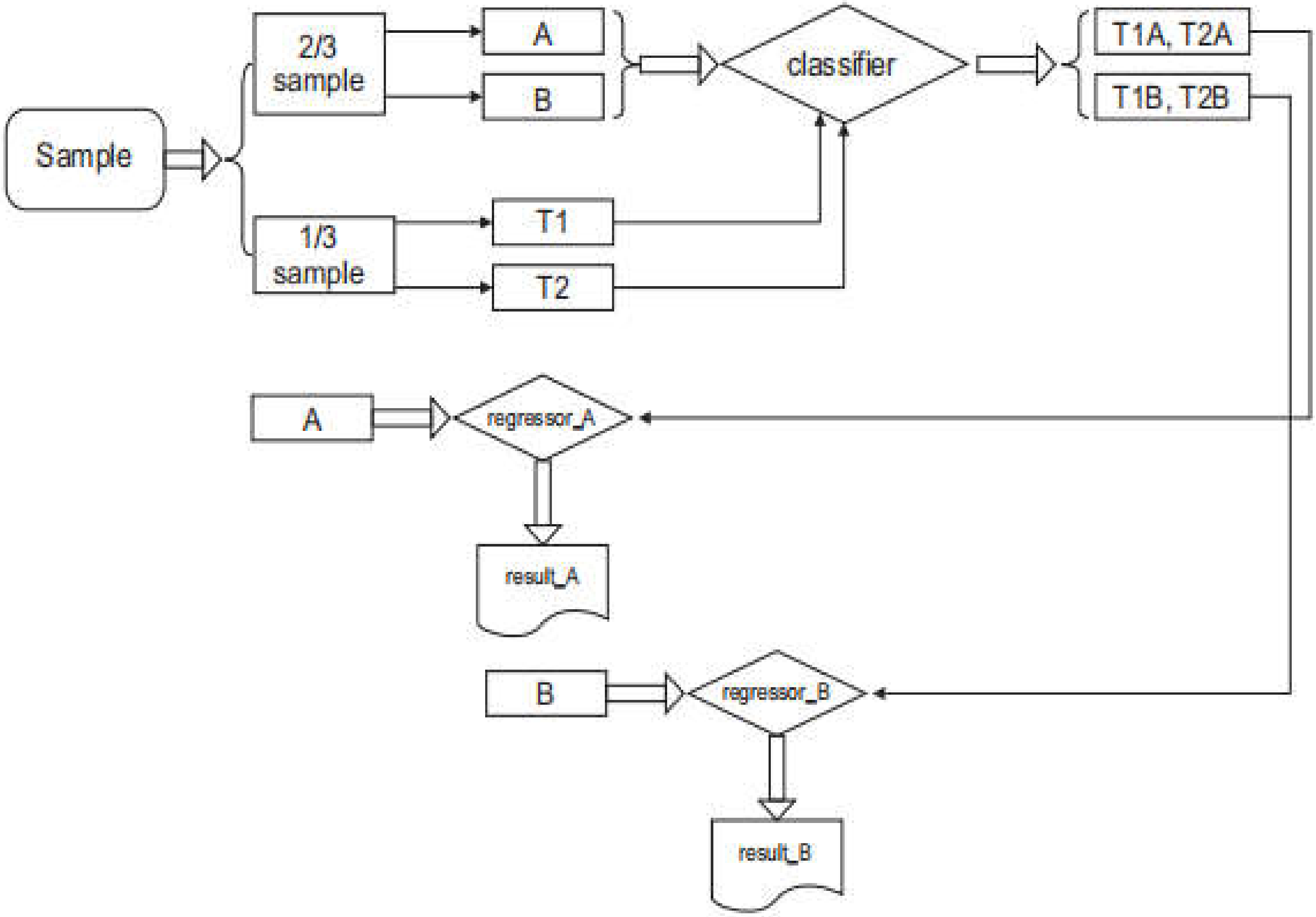


Fig. 3 Flow chart of photometric redshift estimation based on two subsamples.

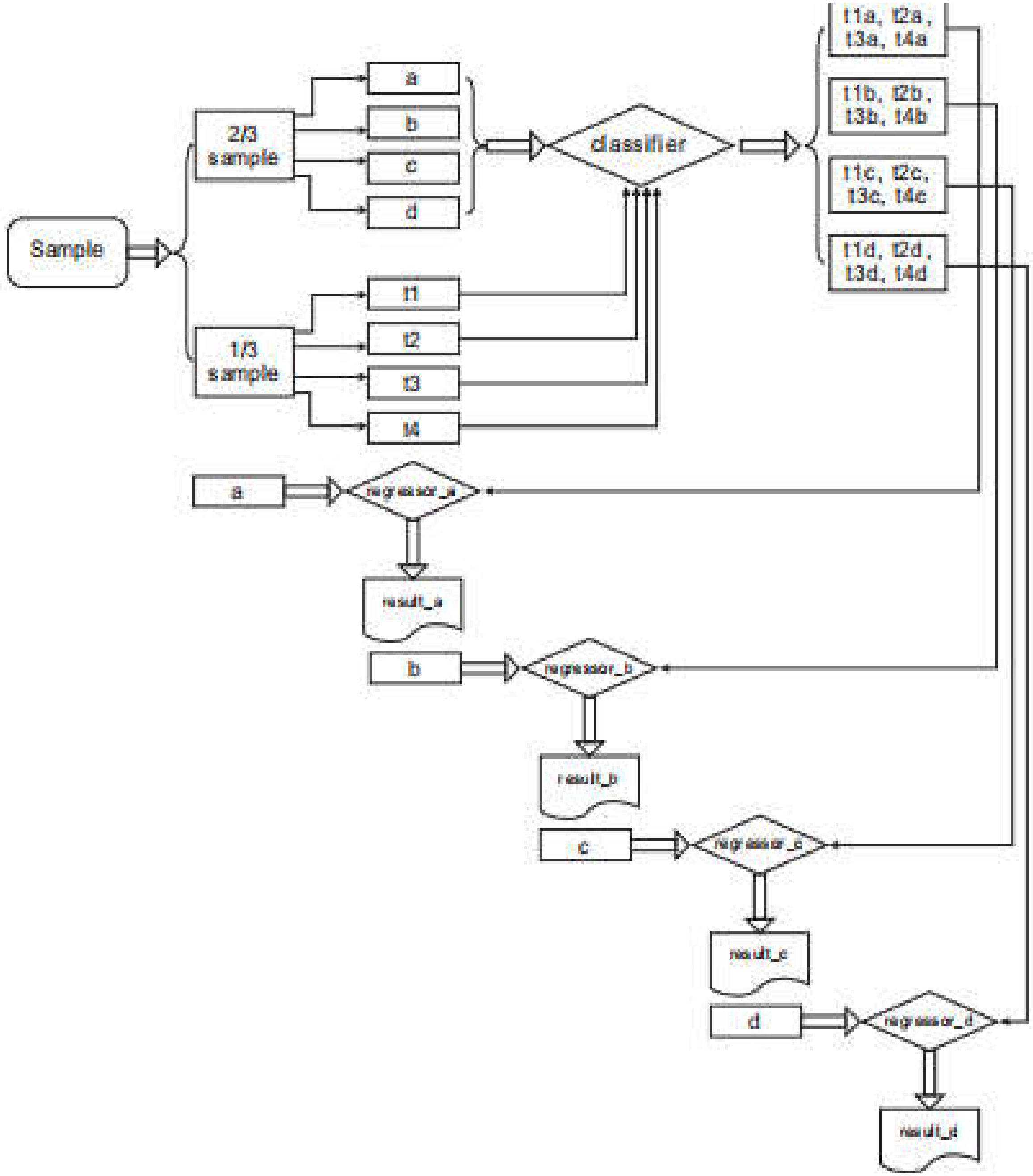


Fig. 5 Flow chart of photometric redshift estimation based on four subsamples.

Table 3 Performance of photometric redshift estimation for different datasets with random forest after classifying one sample into two subsamples by random forest.

Data Set (Test set)	Algorithm	Model Parameters	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ
SDSS (T1)	RF_RF	*n_estimators*=300 *max_depth*=15	55.08	72.07	84.36	0.3550
SDSS (T2)	RF_RF	*n_estimators*=300 *max_depth*=15	84.77	89.55	90.31	0.2810
SDSS (T1+T2)			67.74	79.52	86.90	0.3235
SDSS-WISE (T1)	RF_RF	*n_estimators*=300 *max_depth*=15	75.77	89.55	94.52	0.2022
SDSS-WISE (T2)	RF_RF	*n_estimators*=300 *max_depth*=20	93.01	96.40	96.97	0.1660
SDSS-WISE (T1+T2)			81.60	91.87	95.35	0.1900

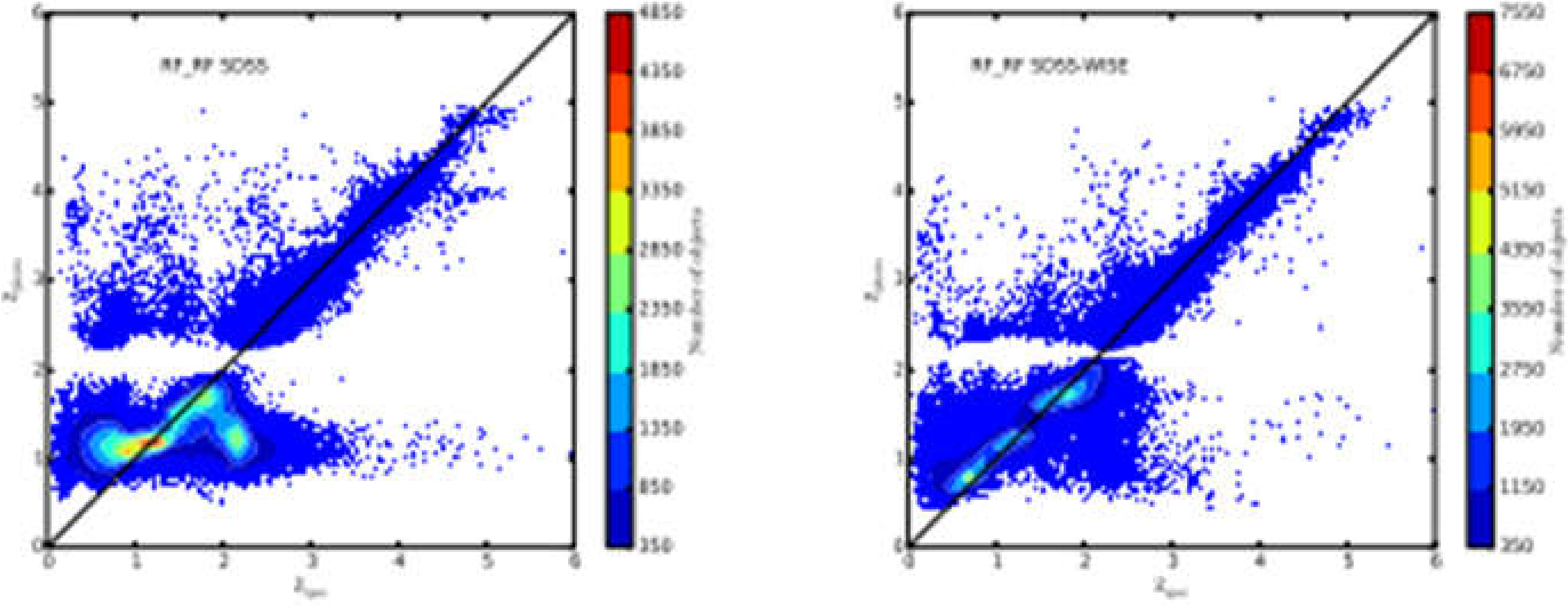

RF_RF SDSS
RF_RF SDSS-WISE
Number of objects

Table 5 Performance of photometric redshift estimation for different datasets with random forest after classifying one sample into four subsamples by random forest.

Data Set (Test set)	Algorithm	Model Parameters	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ
SDSS (t1)	RF_RF	*n_estimators*=200 *max_depth*=15	65.65	75.88	80.27	0.3760
SDSS (t2)	RF_RF	*n_estimators*=300 *max_depth*=15	73.81	85.05	87.07	0.2854
SDSS (t3)	RF_RF	*n_estimators*=300 *max_depth*=15	82.04	87.16	88.16	0.3131
SDSS (t4)	RF_RF	*n_estimators*=50 *max_depth*=15	95.35	96.49	96.68	0.1935
SDSS (t1+t2+t3+t4)			75.56	83.62	85.82	0.3213
SDSS-WISE (t1)	RF_RF	*n_estimators*=300 *max_depth*=15	80.43	90.16	93.35	0.1916
SDSS-WISE (t2)	RF_RF	*n_estimators*=300 *max_depth*=15	83.35	93.69	95.45	0.1860
SDSS-WISE (t3)	RF_RF	*n_estimators*=300 *max_depth*=15	91.95	95.48	96.26	0.1770
SDSS-WISE (t4)	RF_RF	*n_estimators*=200 *max_depth*=20	97.31	98.30	98.52	0.1420
SDSS-WISE (t1+t2+t3+t4)			85.33	93.01	94.97	0.1843

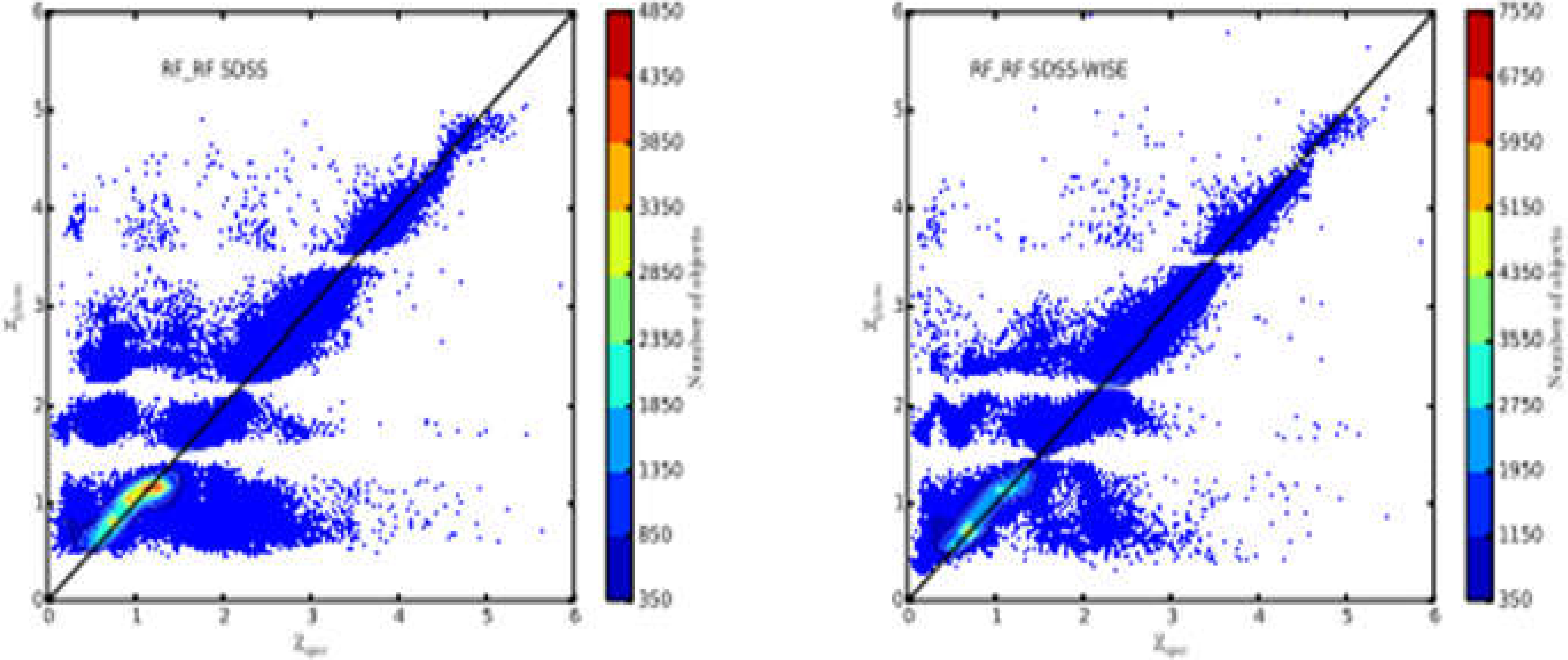

RF_RF SDSS
RF_RF SDSS-WISE
0
1
2
3
4
5
6
4850
4350
3850
3350
2850
2350
1850
1350
850
350
7550
6750
5950
5150
4350
3550
2750
1950
1150
350

Table 4. Performance comparison of photometric redshift estimation with the SDSS-WISE sample for different schemes.

Scheme	Algorithm	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	$\sigma_{\rm rms}$
one sample	RF	79.87	91.37	95.23	0.1907
two subsamples	RF_RF	81.60	91.87	95.35	0.1900
four subsamples	RF_RF	85.33	93.01	94.97	0.1843
four subsamples	RF_RF by correction	85.76	93.28	95.19	0.1699

RF_RF means that random forest is used to build the classifier and the regressor. RF_RF by correction represents that they adopted the estimated redshift value from the regressor with four subsamples but kept the estimated value from one sample by random forest near the three cutoff points (± 0.3) during photometric redshift estimation period. The performance metrics of RF_RF by correction all increase compared to those with one sample, two subsamples, four subsamples, especially $\sigma_{\rm rms}$ reduces to 0.1699. It is evident that this strategy is effective and applicable when the accuracy of photometric redshift estimation is improved.

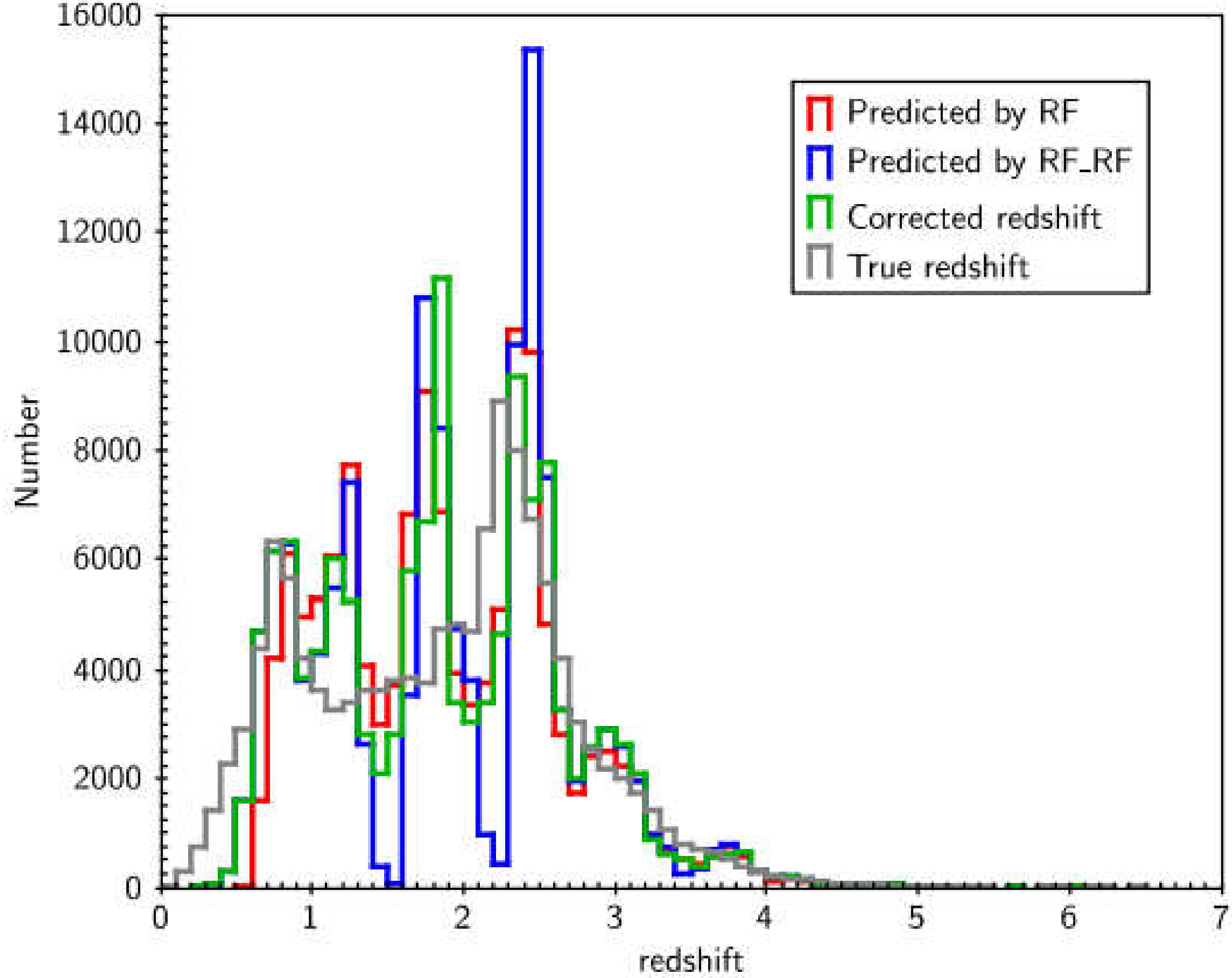


Fig. 7 Redshift distribution. Grey line represents true redshift; red line for estimated redshift from one sample by random forest; blue line for estimated line from four samples by RF_RF; green line for estimated redshift from four samples by RF_RF and corrected near the cutoff.

Improvements on photo-z

- More photometric data in other bands

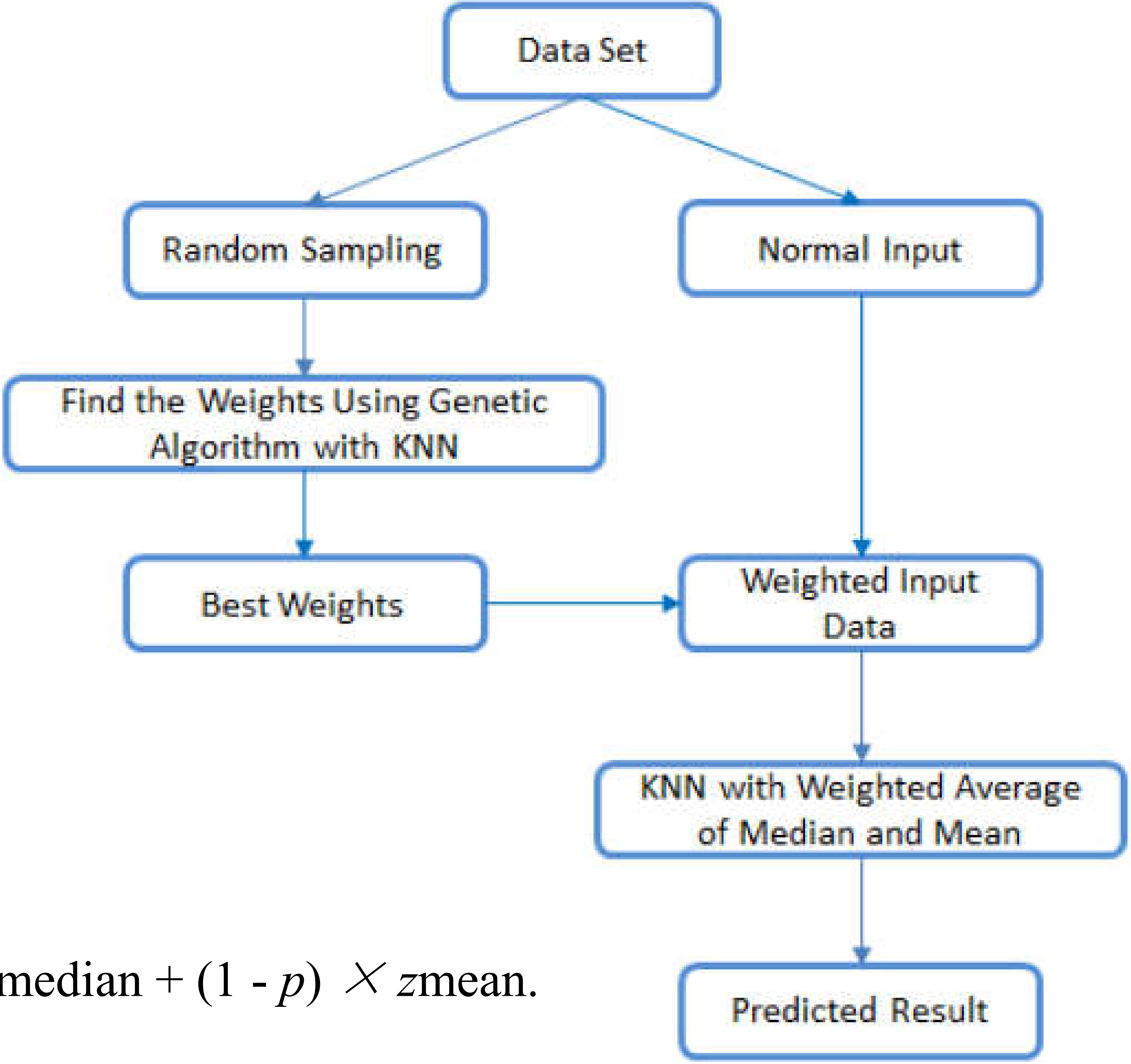
  -- UV(GALEX), near-IR(JHK),radio …
- Much more accurate data
- Feature selection/extraction/weighting/reconstruction
- Better methods/hybrid methods/ensemble methods
- Algorithm selection and optimization
- How to deal with training sample?
- How to get a complete and respective sample?
- How to balance between accuracy and efficiency?

Summary

- Scintific issue
- Choice of ML method. (The more complex method is not always a better solution in big data era.)
- Data preparation is important.
- Each step of ML need be careful.
- High performance computing and parallel computing
- Team work

Thank you for your attention!!

We adopt the quasar sample from the data release 14 Quasar catalog (DR14Q) of the SDSS-IV/eBOSS (Pâris et al. 2018). The DR14Q catalog contains 526,356 unique quasars. Discarding the records which contain default SDSS magnitudes, $zWarning = -1$ and full magnitude errors large than 5, the number of the SDSS quasar sample is 445,958. When further getting rid of the records with default $W1$ and $W2$, the number of the SDSS-WSIE quasar sample is 324,333. Here all magnitudes are adopted AB magnitudes. The AB magnitude conversion and extinction correction process is referred to Schindler et al. (2017). The better quality data are obtained by the limitation with $sciencePrimary = 1$, $Mode = 1$, $zWarrning = 0$, excluding the records using flags such as BRIGHT, SATURATED, EDGE and BLENDED, and removing objects whose magnitude errors are larger than 0.2 in five optical bands and larger than 0.3 in two infrared bands. At this situation, the number of the SDSS-WISE quasar sample adds up to 261,705.



$$p \times z\text{median} + (1 - p) \times z\text{mean}.$$

Table 2 Performance of photometric redshift estimation of different models for the SDSS sample with 5m_4c

Algorithm	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ	MSE	R^2	Time(s)
LASSO	32.41	73.22	82.05	0.4977	0.3777	-0.6983	115
SVR	60.81	79.76	84.33	0.3709	0.2933	0.0732	1403
NN	59.90	79.11	86.70	0.3475	0.2411	0.3286	3834
XGBoost	62.30	80.27	87.41	0.3303	0.2281	0.3908	2819
KNN	62.18	80.00	86.99	0.3344	0.2353	0.3512	137
RF	63.29	80.54	87.42	0.3263	0.2277	0.3887	16574
GK	**66.48**	**81.80**	**87.53**	**0.3169**	0.2340	**0.4016**	115

Table 4 Performance of photometric redshift estimation of different models for the SDSS-WISE sample with 7m_6c

Algorithm	$\delta_{0.1}(\%)$	$\delta_{0.2}(\%)$	$\delta_{0.3}(\%)$	σ	MSE	R^2	Time(s)
LASSO	50.54	78.87	89.58	0.3479	0.2085	0.4882	98
SVR	70.85	88.39	93.76	0.2365	0.1262	0.7422	1336
NN	77.11	90.83	95.24	0.2075	0.1064	0.7935	3749
XGBoost	78.83	91.27	95.44	0.1950	0.0989	0.8085	3129
KNN	78.57	91.10	95.20	0.1983	0.1036	0.7956	282
RF	79.76	91.53	95.37	0.1908	0.0998	0.8036	12944
GK	**83.25**	**92.85**	**95.61**	**0.1777**	**0.0982**	**0.8179**	319