Object classification in SDSS DR12

Farhang Habibi
Reza Ansari & Marc Moniez

LAL

Aim

To automatically separate stars, galaxies and Quasars by using the colour indices in the absence of spectroscopic data.

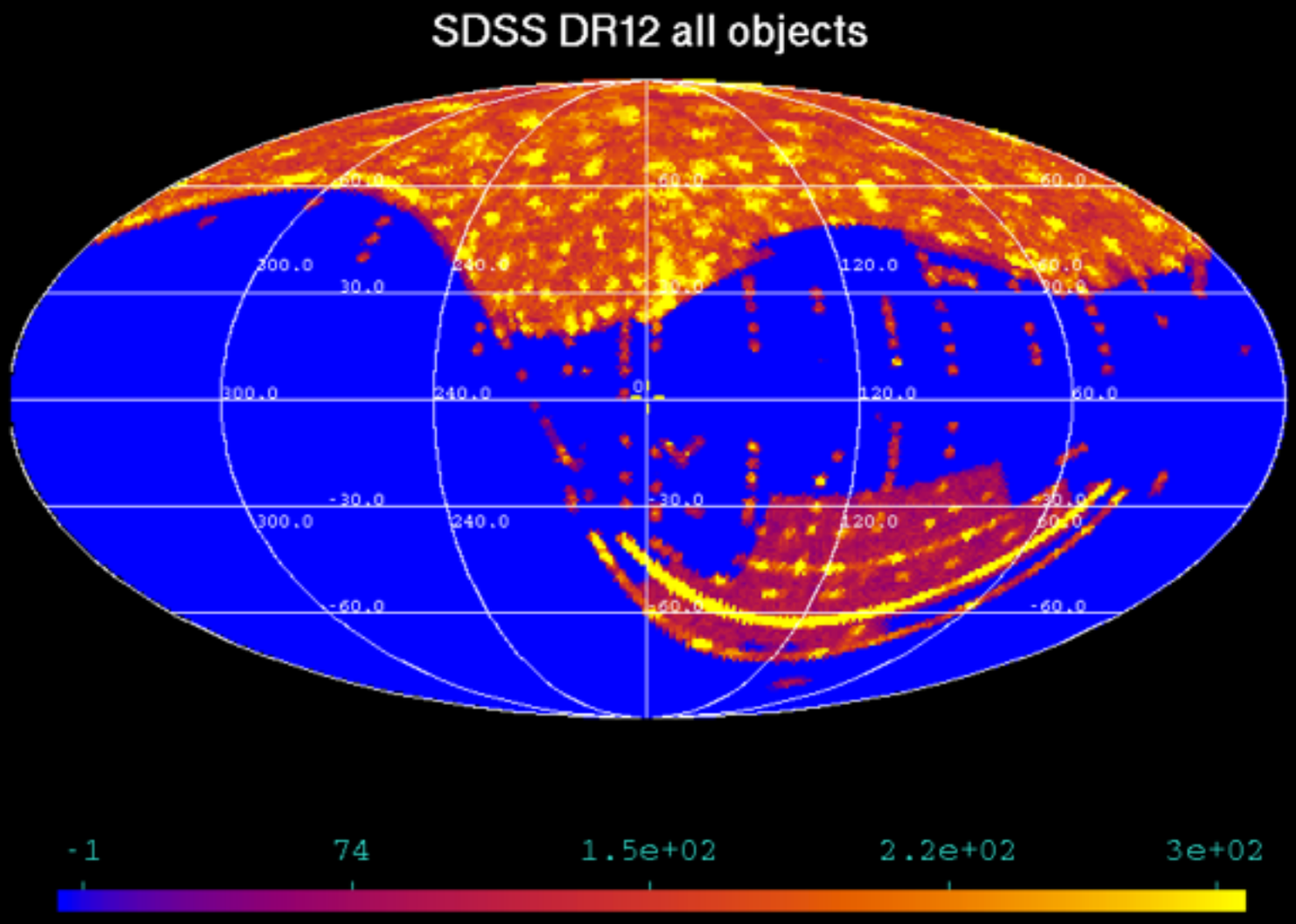
2 m class telescope

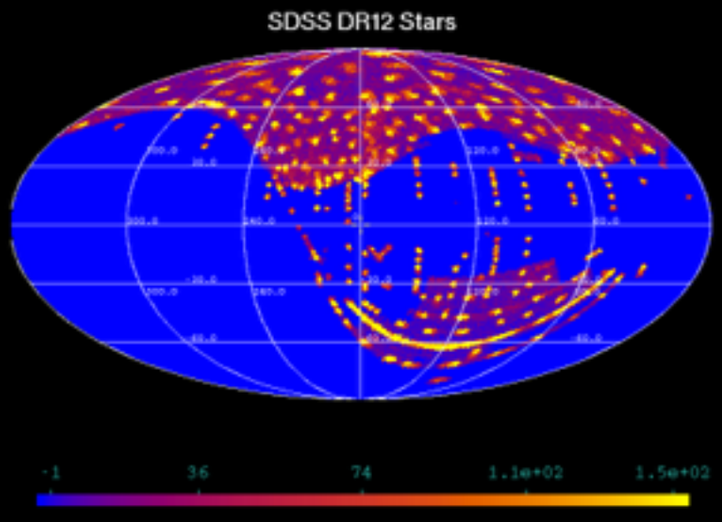
~ 4 million spectra

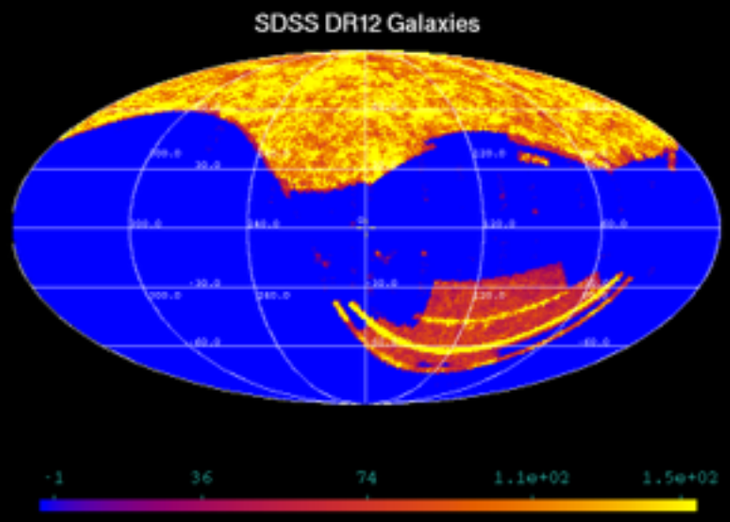
~ 60% galaxies

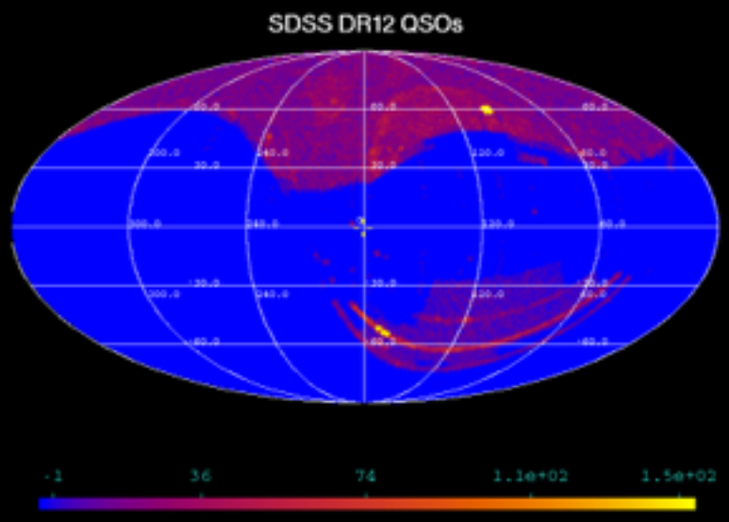
~ 30% stars

~ 10% QSOs

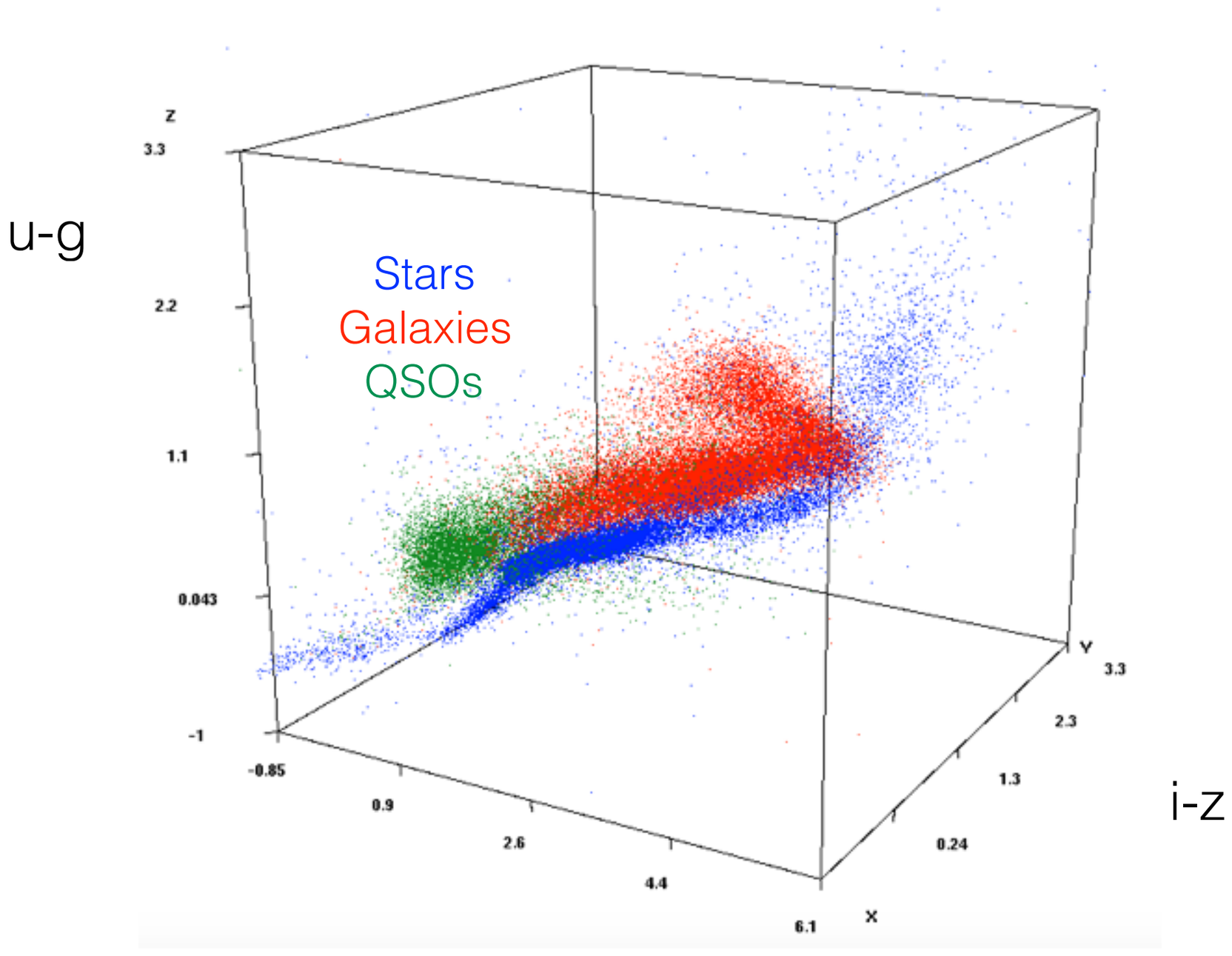


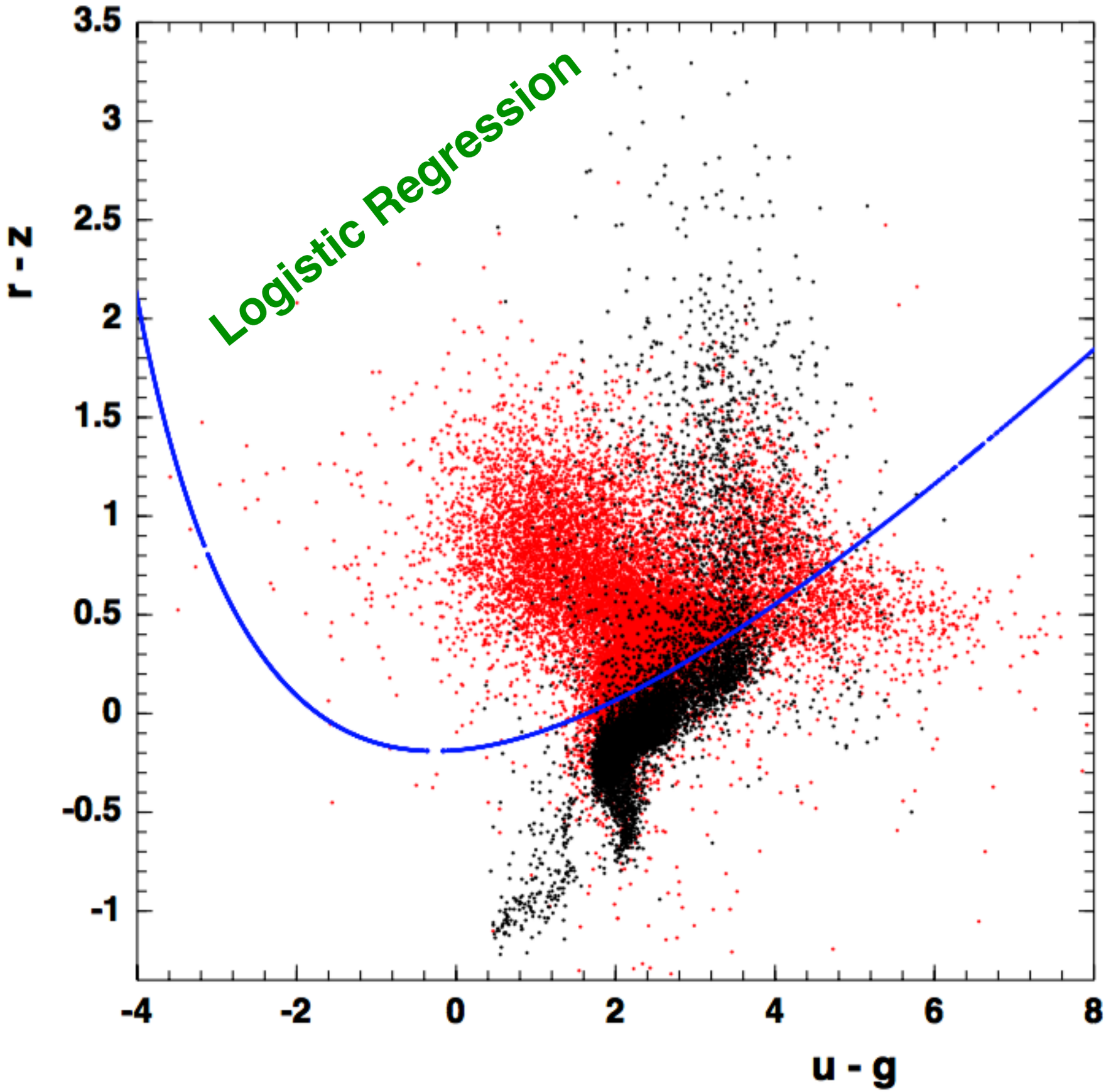




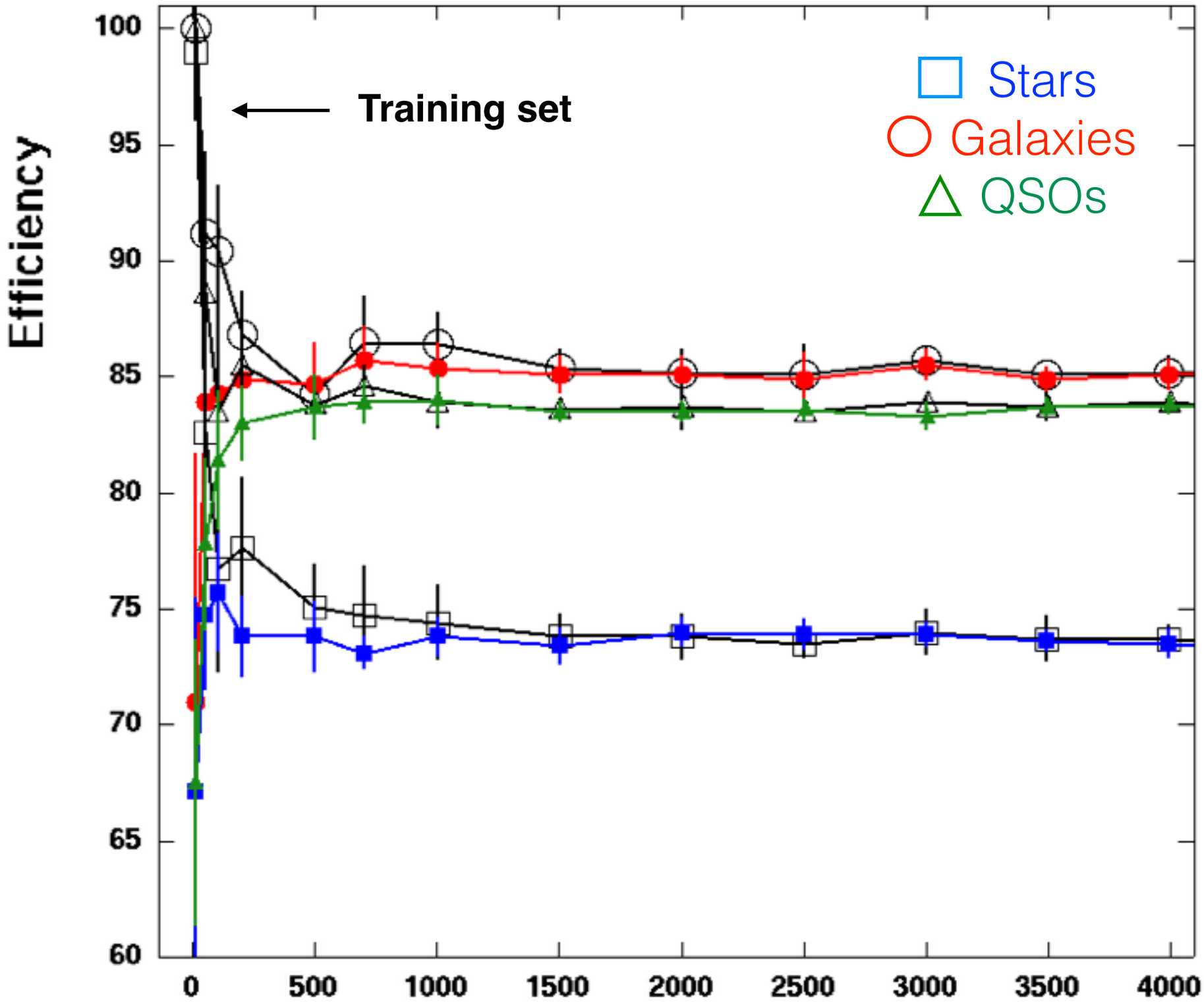


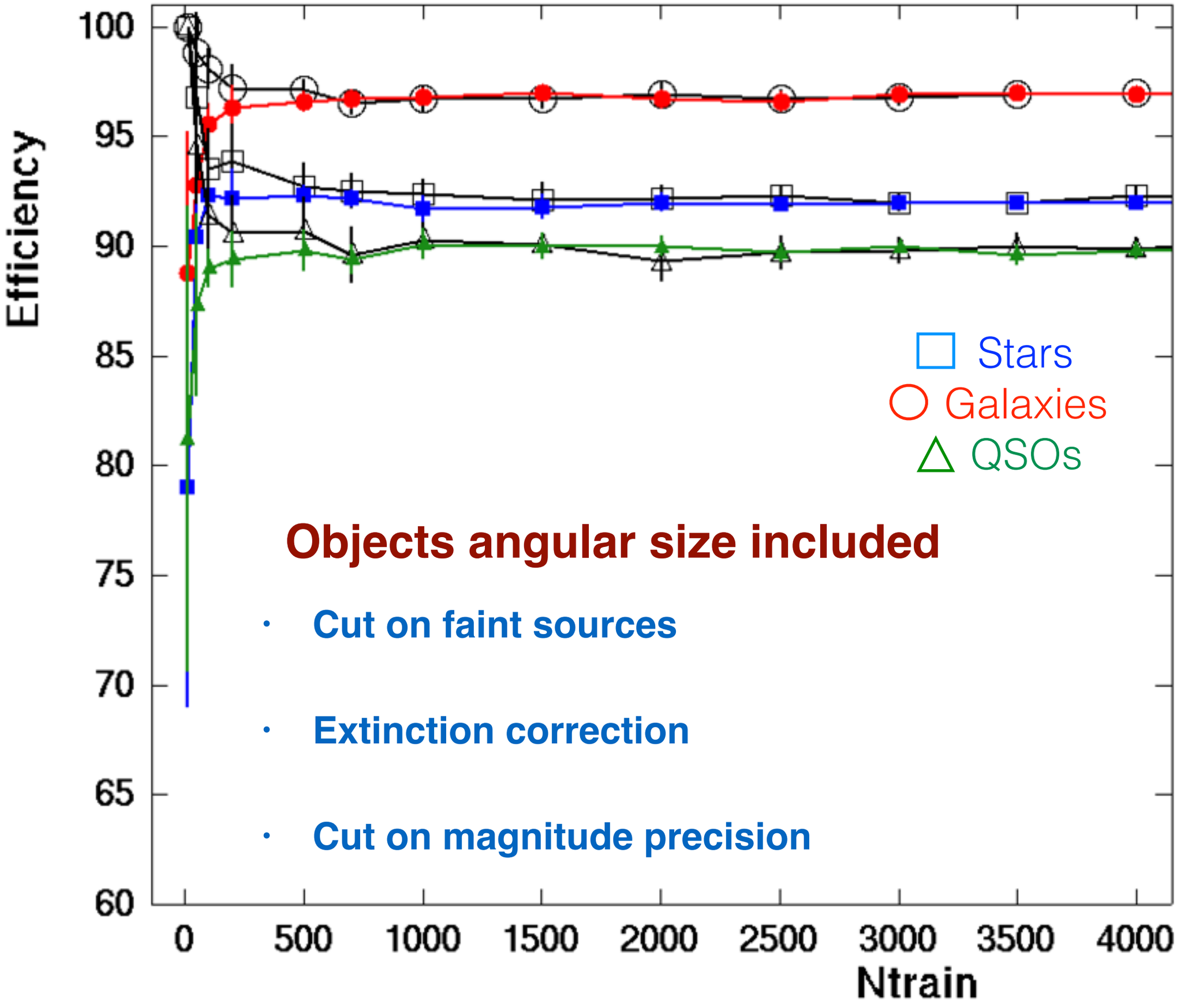
4 independent **colours** **+** **g-magnitudes** **+** their multiplications
construct **20 features**
defining a 5-dimensional hyper parabola
in colour-magnitude space.




Black: Stars
Red: Galaxies
Logistic Regression
r - z
u - g
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-4
-2
0
2
4
6
8

$$\text{efficiency} = \frac{\text{number of correctly classified objects}}{\text{number of objects}}$$




Efficiency
100
95
90
85
80
75
70
65
60
0
500
1000
1500
2000
2500
3000
3500
4000
Ntrain
Stars
Galaxies
QSOs
Objects angular size included
Cut on faint sources
Extinction correction
Cut on magnitude precision

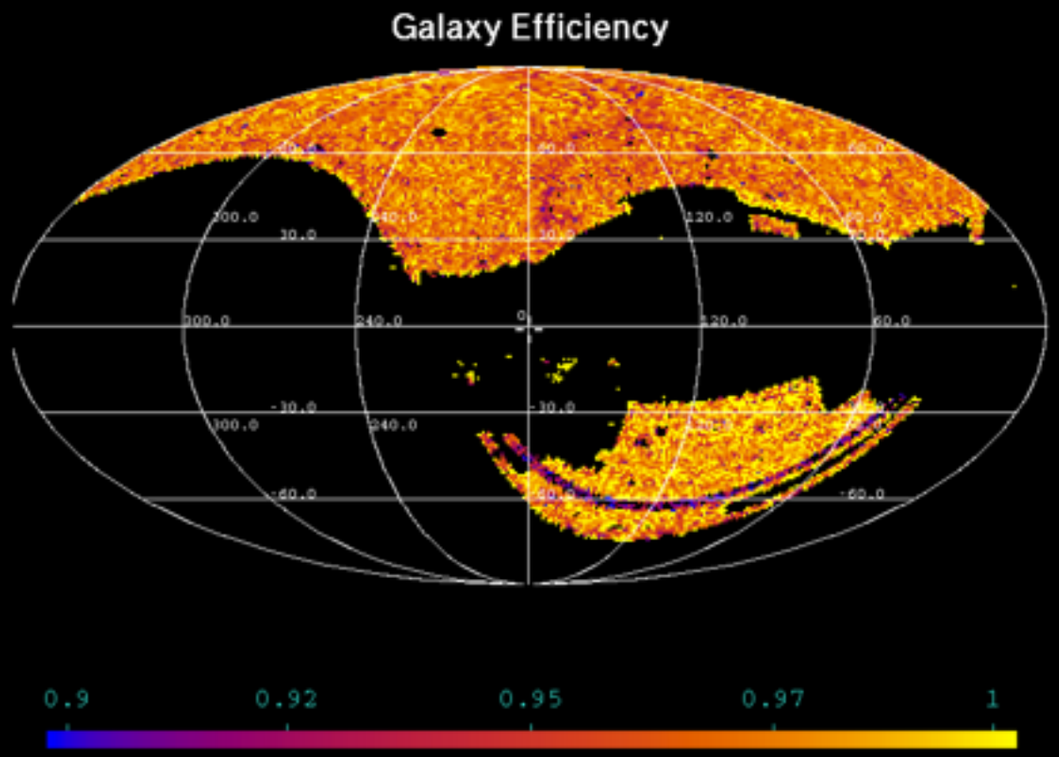
Results from the classification

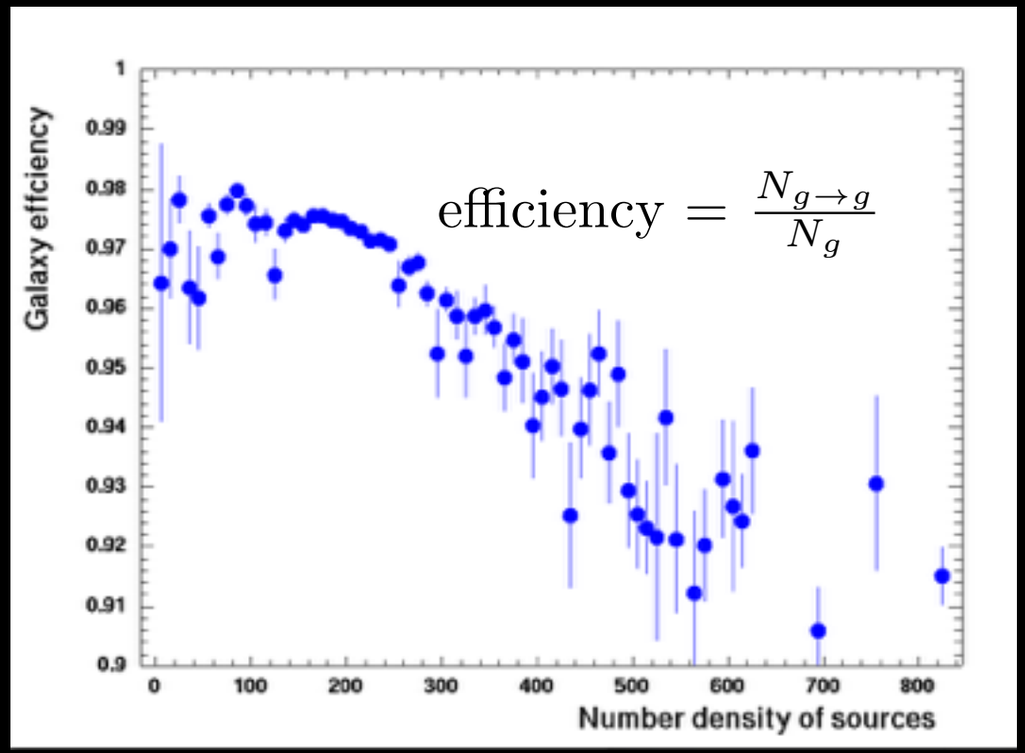
- Classification efficiency for the
  whole sample: 94%
  galaxies: 97%
  stars: 92%
  QSOs: 90%

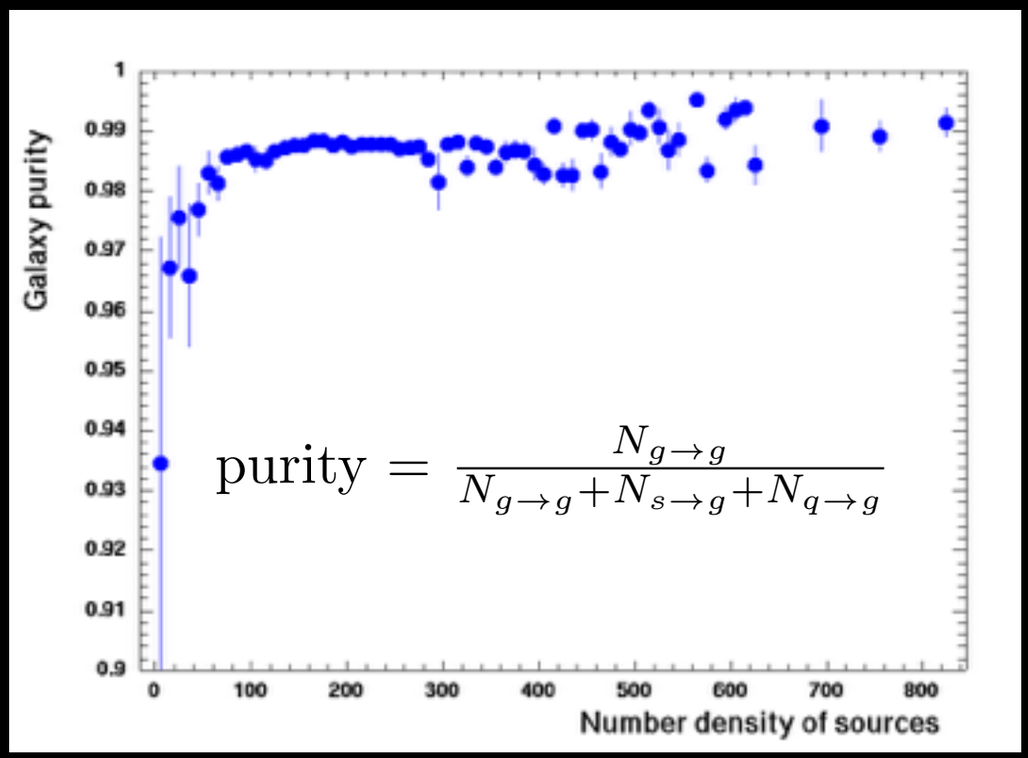
- Mean size of the galaxies classified
  wrongly: 0.5 arcsec
  correctly: 3 arcsec

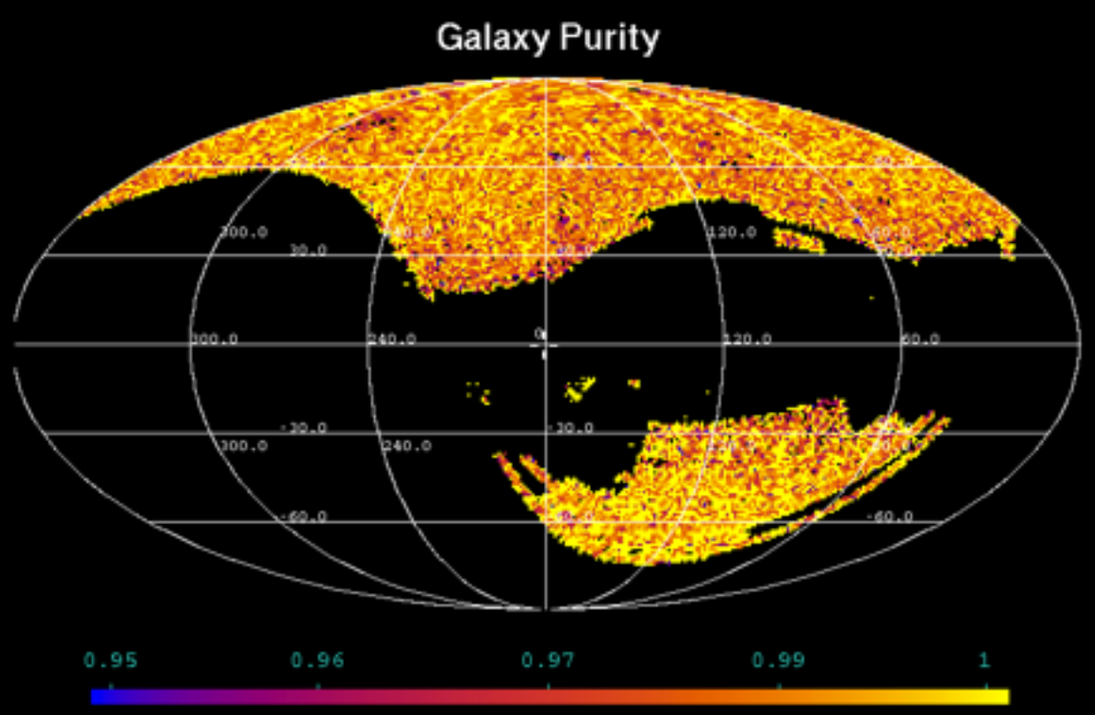
- Mean magnitude (extinction corrected) of the stars classified
  wrongly: z = 19 (fainter stars)
  correctly: z = 17

- Mean redshift of the QSOs classified
  wrongly: redshift = 2 (further QSOs)
  correctly: redshift = 1.5


Galaxy Efficiency
0.9
0.92
0.95
0.97
1

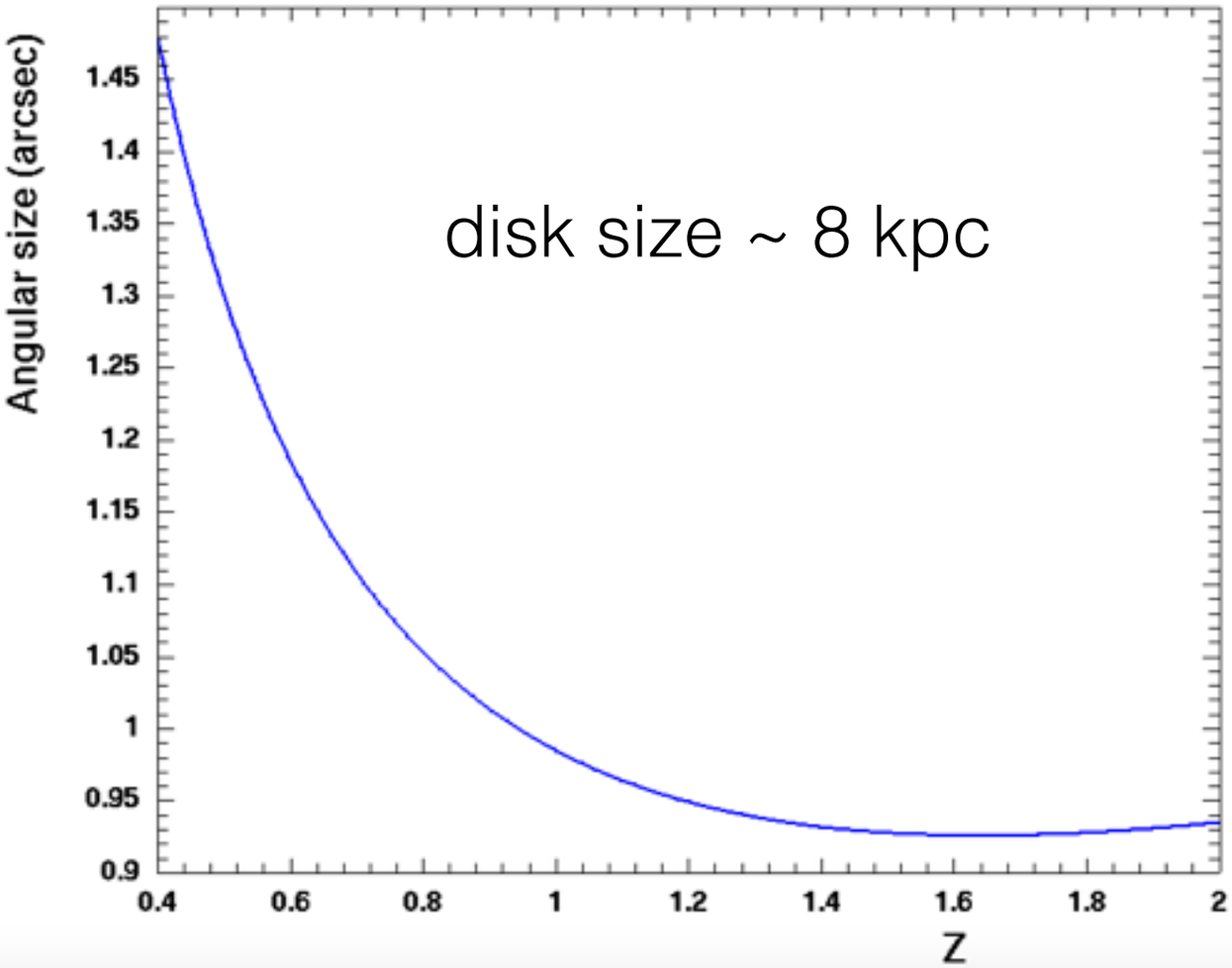

Galaxy effciency
Number density of sources
efficiency = $\frac{N_{g\to g}}{N_g}$


Galaxy purity
Number density of sources
purity = $\frac{N_{g\to g}}{N_{g\to g}+N_{s\to g}+N_{q\to g}}$

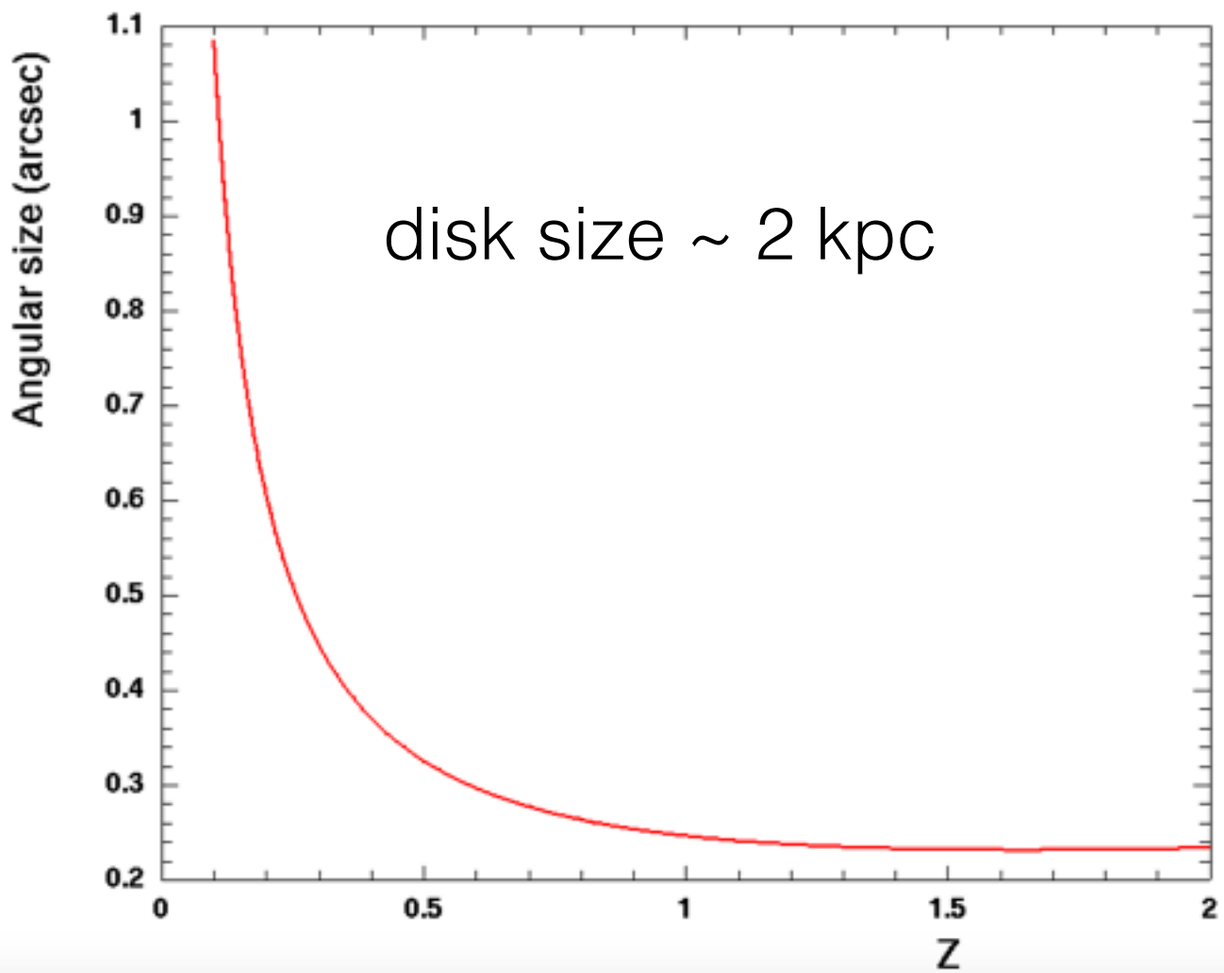

Galaxy Purity
0.95
0.96
0.97
0.99
1

Classification for LSST objects

MW-like galaxies can be resolved by morphology
but not for faint galaxies (dwarfs)

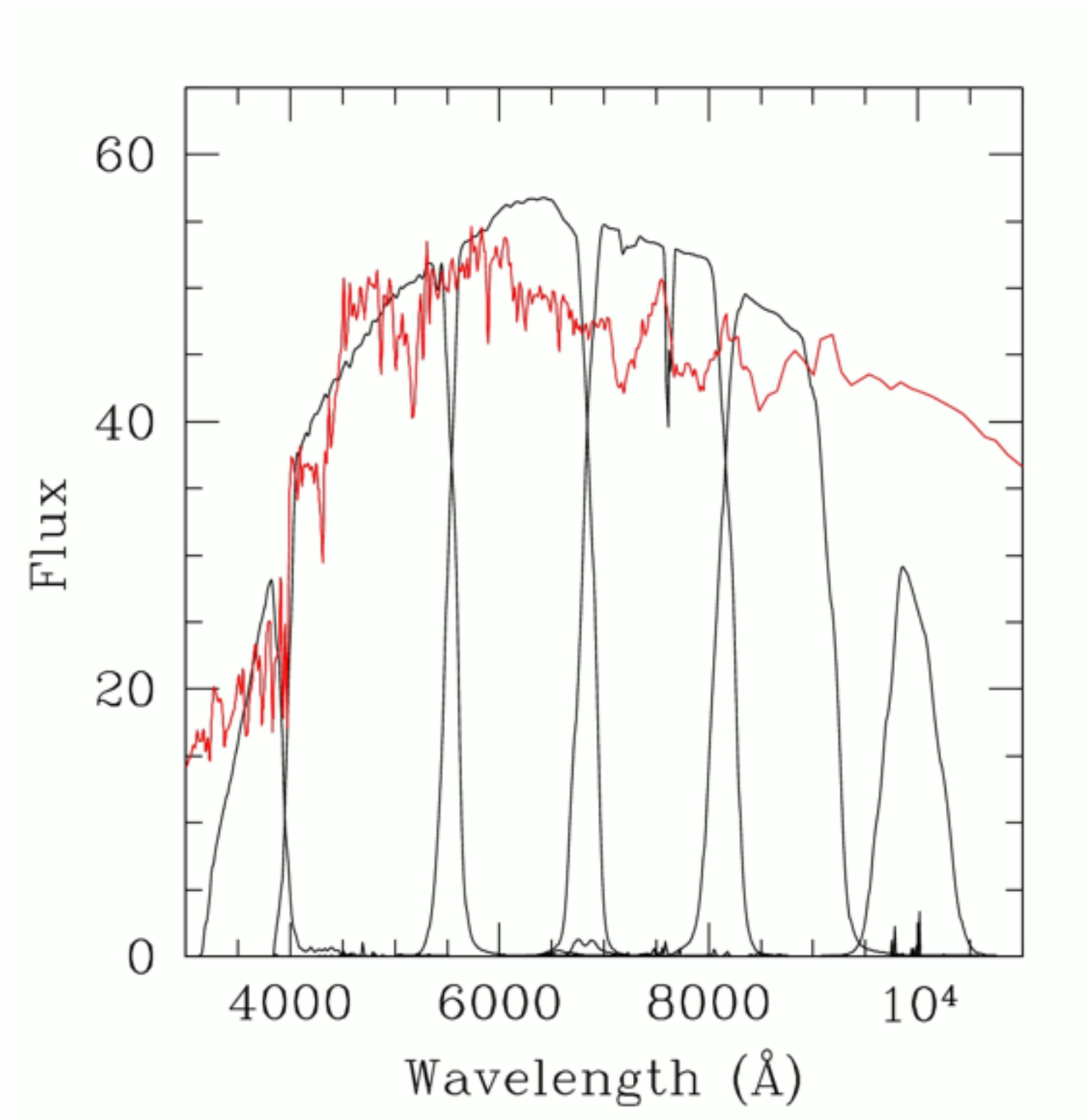


disk size ~ 8 kpc



disk size ~ 2 kpc

Classification for LSST objects

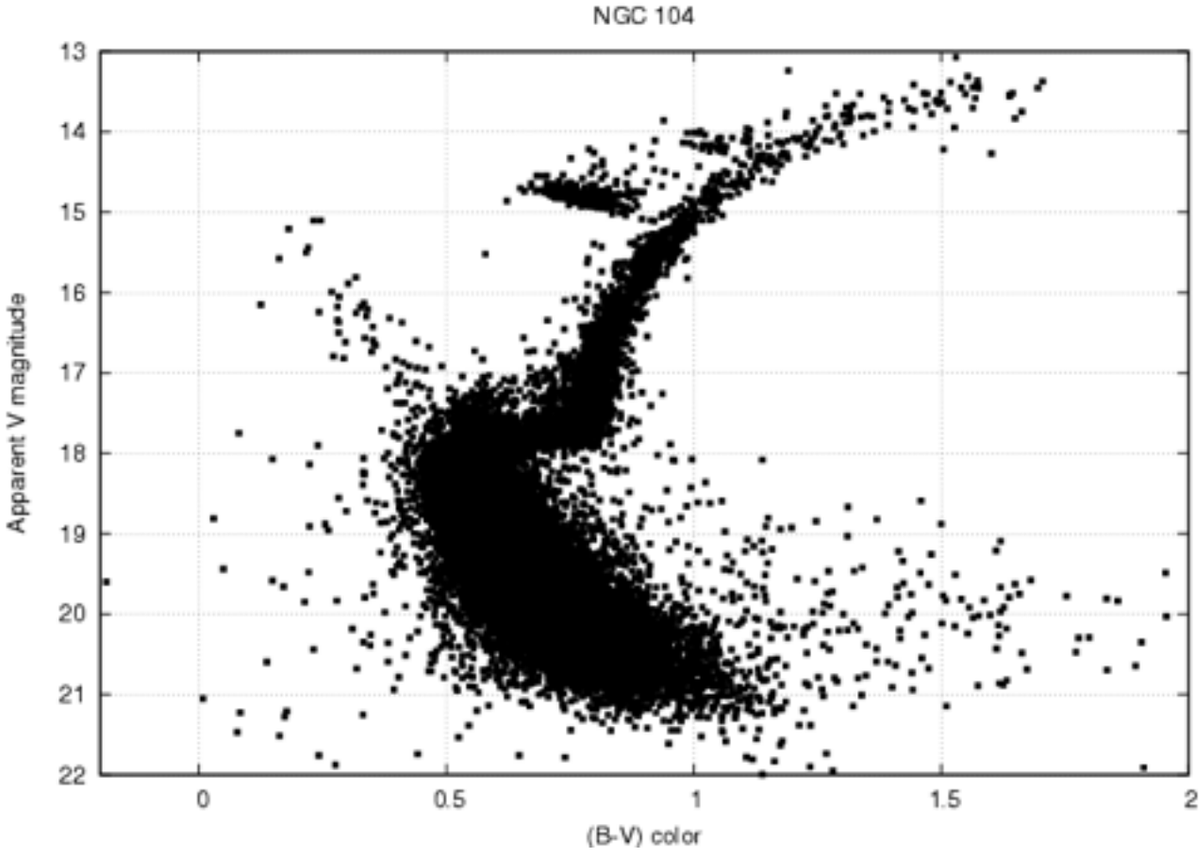


- Generating galaxies at different redshifts from their luminosity function
- Assigning SED to galaxies
- Computing magnitudes and colours
- Assigning angular size to galaxies

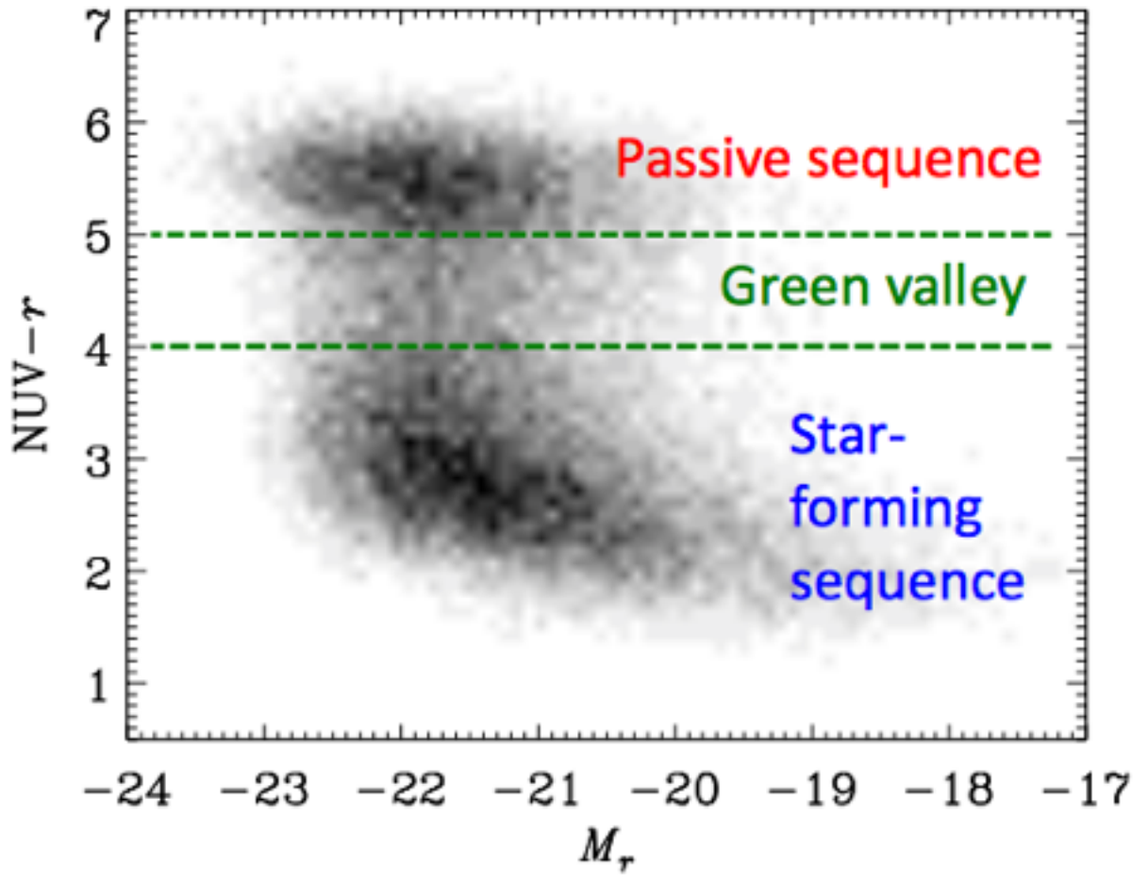
Backups

Features from photometric data

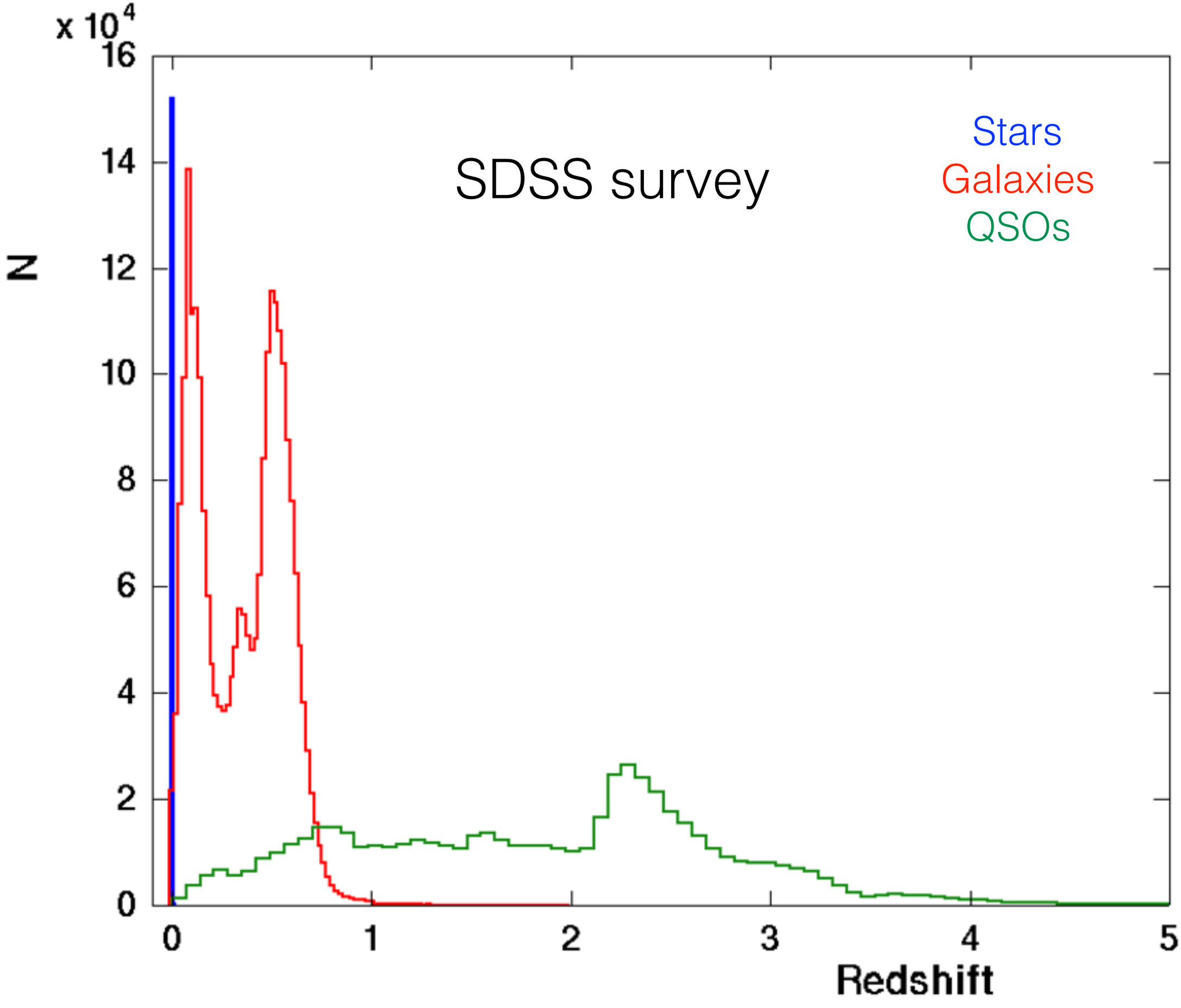
Colour indices and magnitudes can be used to classify the celestial objects

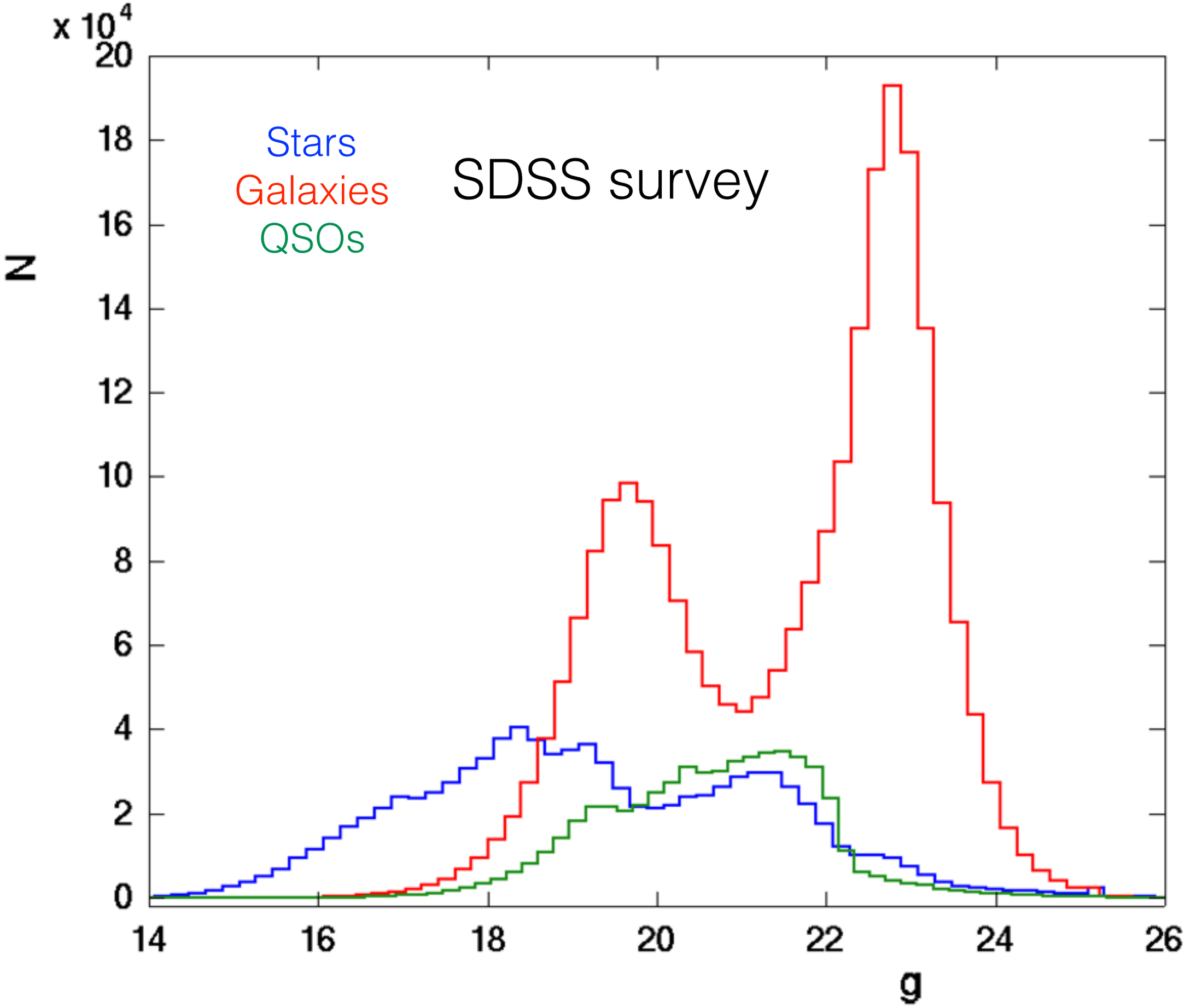


Stars

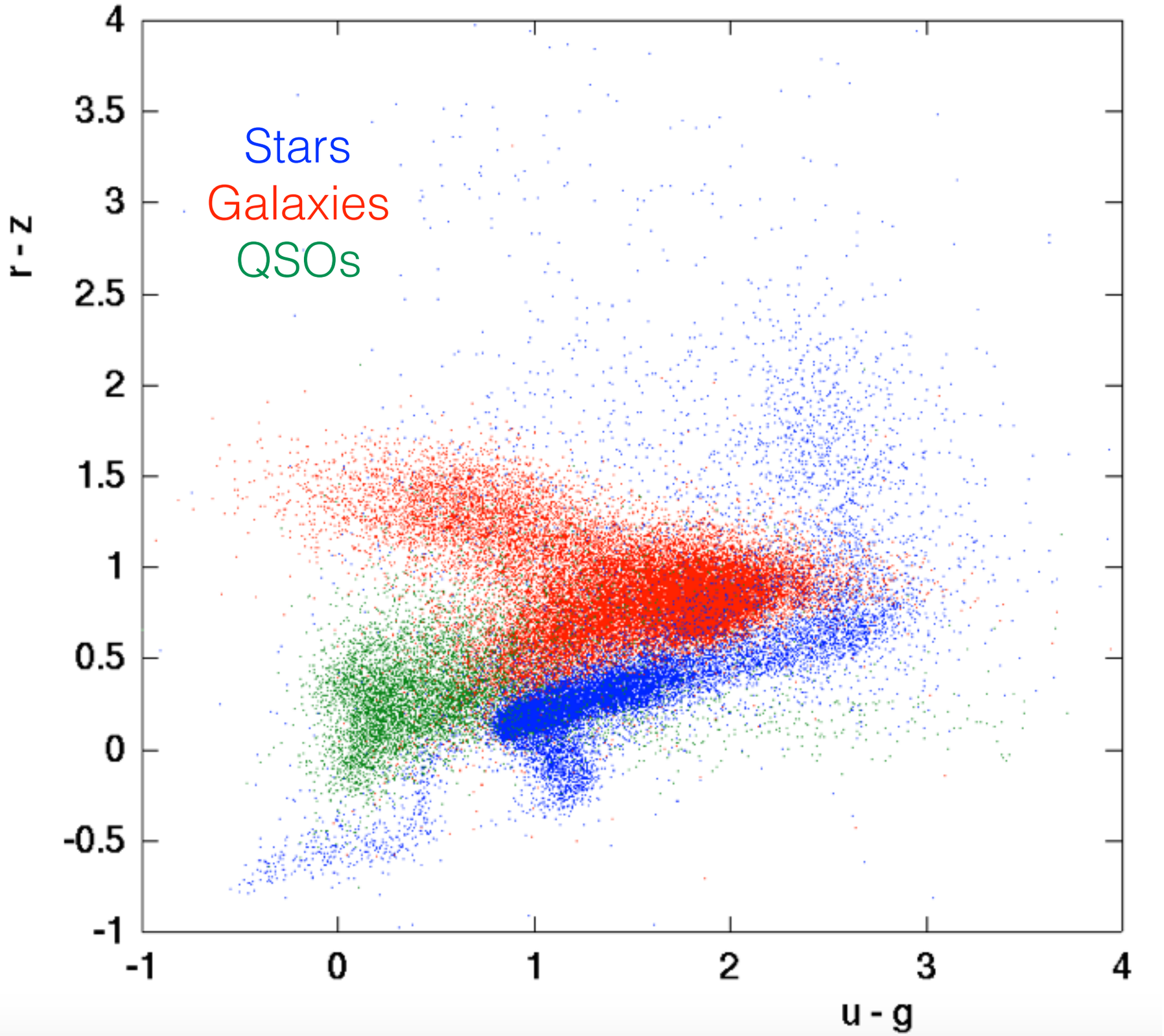


Galaxies

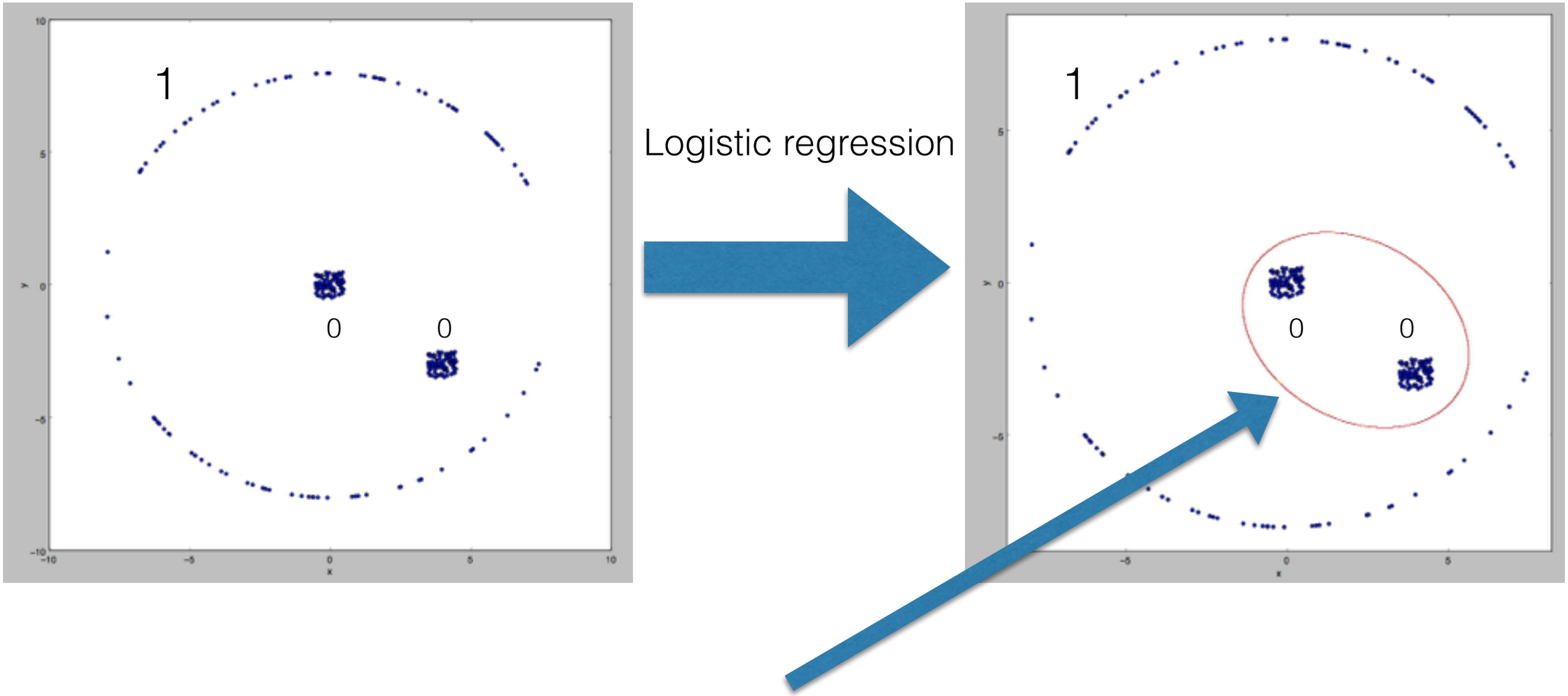

x 10^4
16
14
12
10
8
6
4
2
0
N
SDSS survey
Stars
Galaxies
QSOs
0
1
2
3
4
5
Redshift


x 10^4
20
18
16
14
12
10
8
6
4
2
0
N
Stars
Galaxies
QSOs
SDSS survey
14
16
18
20
22
24
26
g

Colour indices as “features” for classification



Supervised Classification



Parameters of the separating curve are derived by the logistic regression method.

Logistic regression

(thanks to Andrew Ng)

Cost function to be minimised

$$J(\theta) = -\frac{1}{m}[\Sigma y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(x^(i))) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))]$$

Sigmoid (logistic) function

$$h_\theta = \frac{1}{1 - e^{-\theta^T x}}$$

m : total number of objects in the training set

i : object's index

x_i : vector of features of an object

y_i : object's label, 0 for stars, 1 for galaxies

θ : vector of parameters to be fitted

Logistic regression

(thanks to Andrew Ng)

$$h_\theta(\vec{x}) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n)}}$$

Hyper border surface:

$$\theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n = 0$$

The cost function:

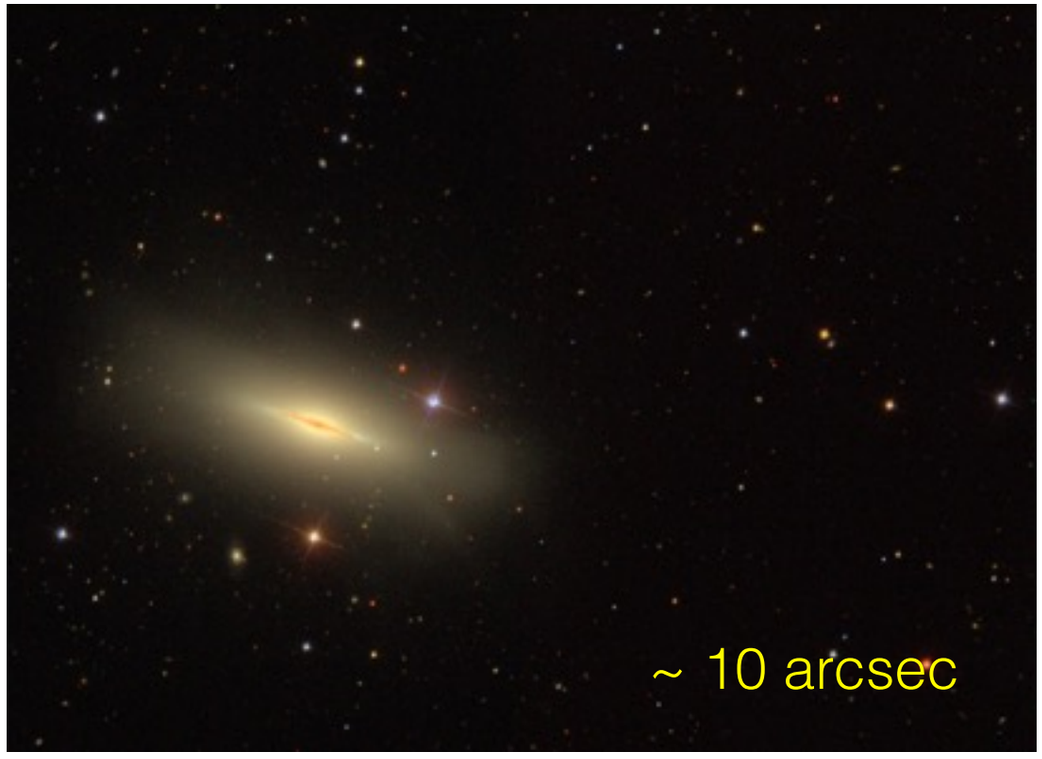
$$J = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log(h_\theta(\vec{x}^{(i)})) + (1 - y^{(i)}) \log(1 - h_\theta(\vec{x}^{(i)}))]$$

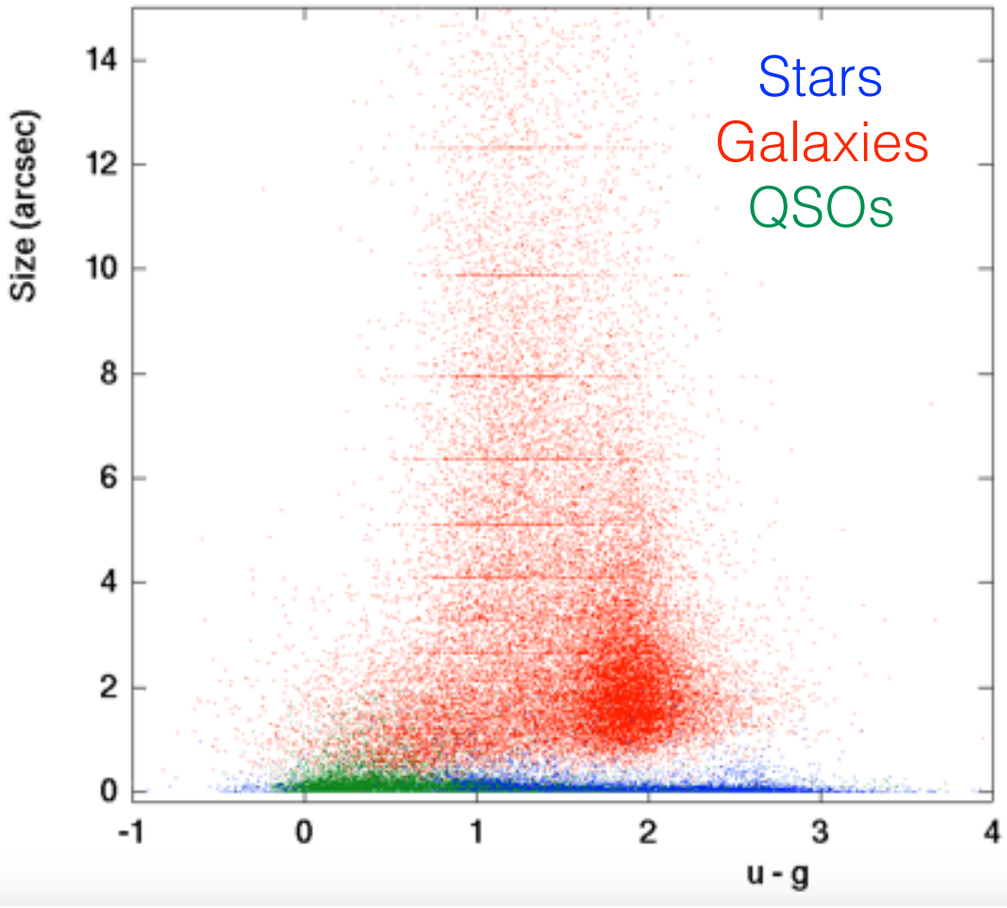
Logistic regression

- We take into account size of objects and10 colours (c1=u-g, c2=u-r, …) plus one magnitude (u) and their quadratic function (ci.cj) to have 77 features.

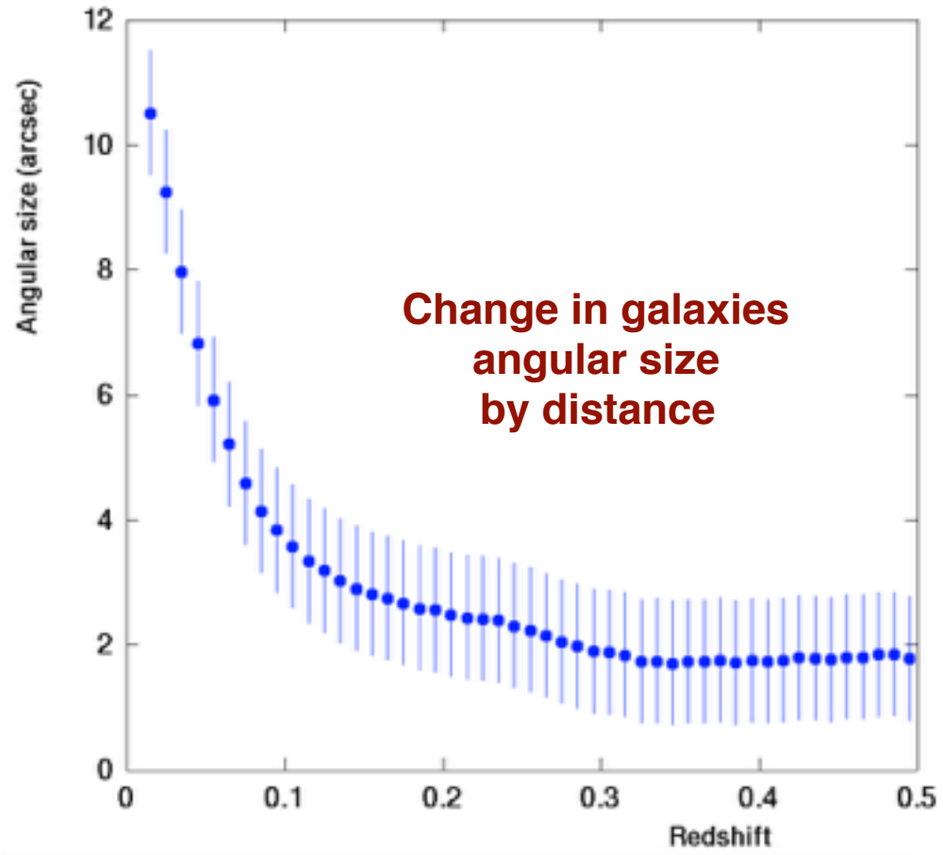
- The separation region is constrained by a 12 dimension hyper parabola defined by 79 parameters.

- From ~670,000 stars, ~1,100,000 galaxies and 250,000 QSOs we randomly put 20000 form each object into the training sample.

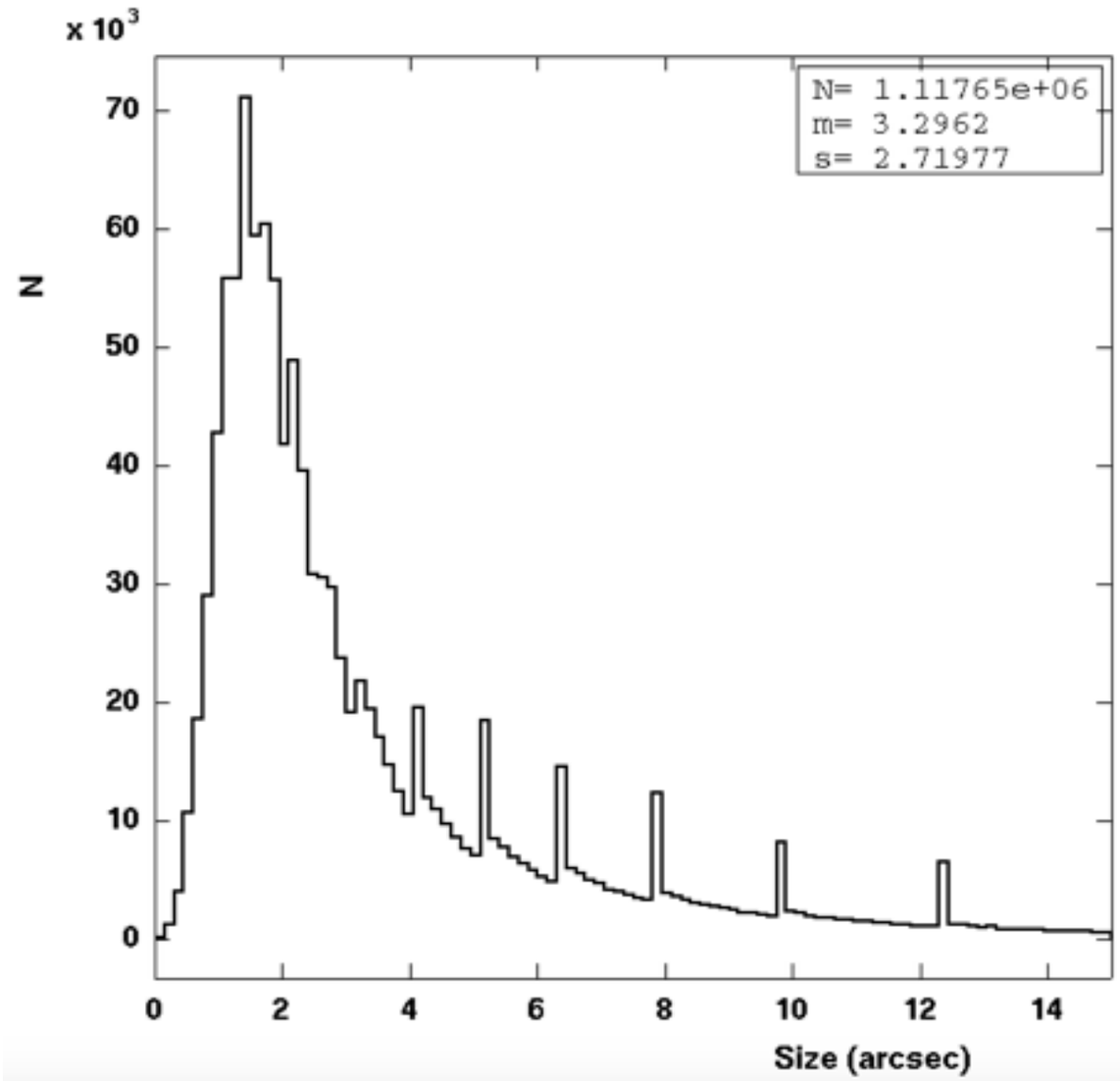

~ 10 arcsec

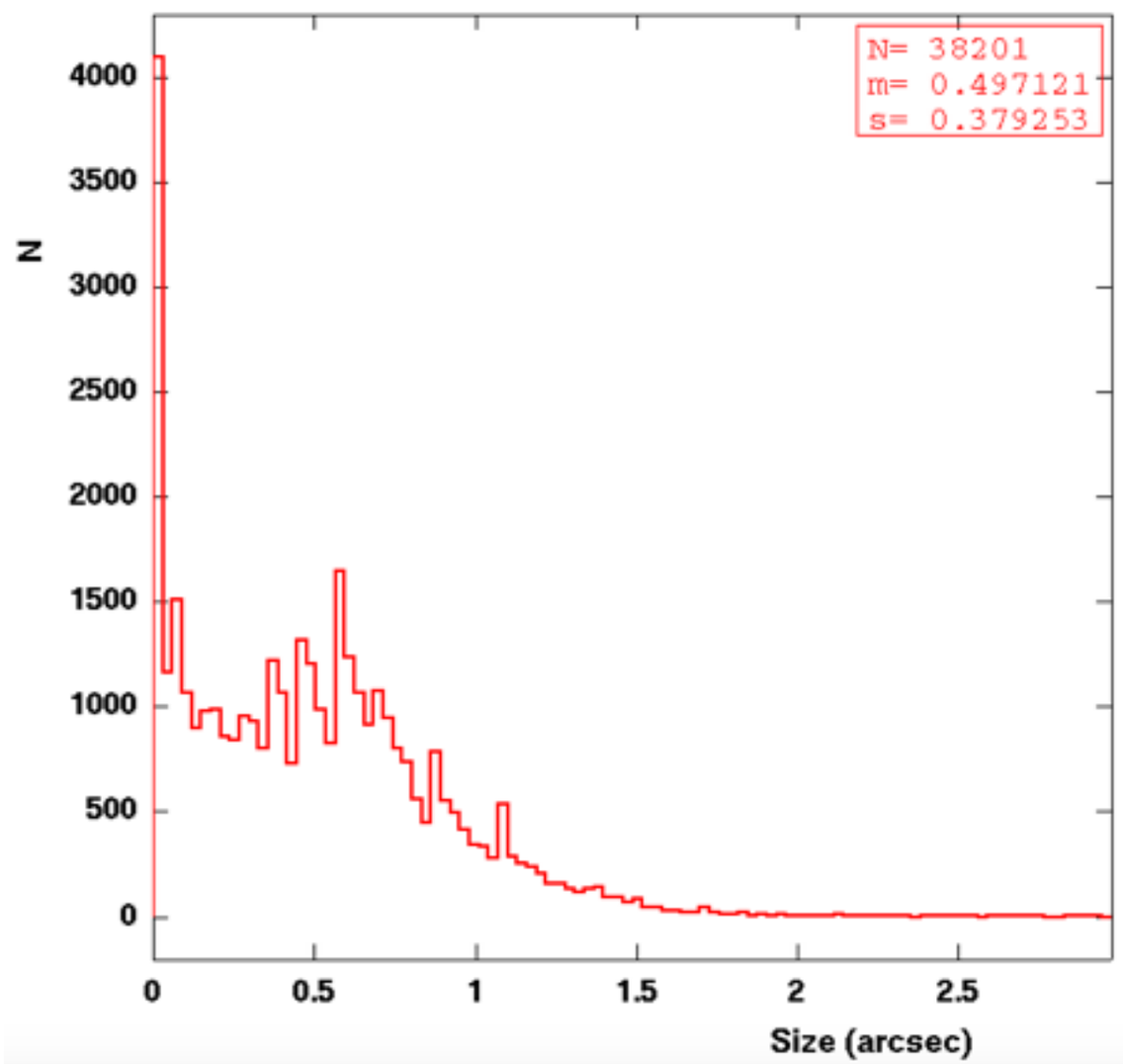

Stars
Galaxies
QSOs
Size (arcsec)
14
12
10
8
6
4
2
0
-1
0
1
2
3
4
u - g


~ 1 arcsec

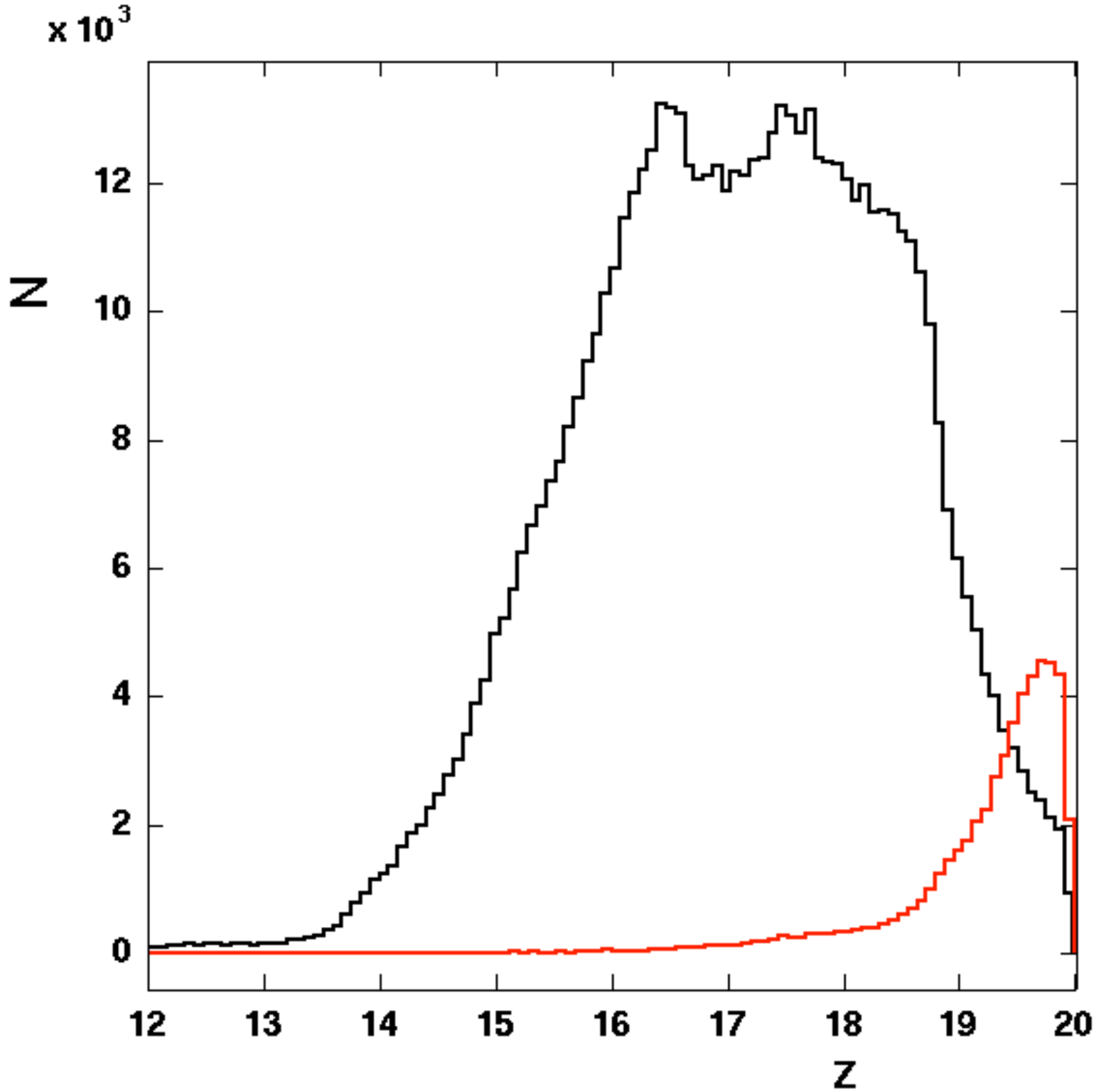

Angular size (arcsec)
12
10
8
6
4
2
0
0
0.1
0.2
0.3
0.4
0.5
Redshift
Change in galaxies
angular size
by distance

Wrongly and **correctly** classified galaxies

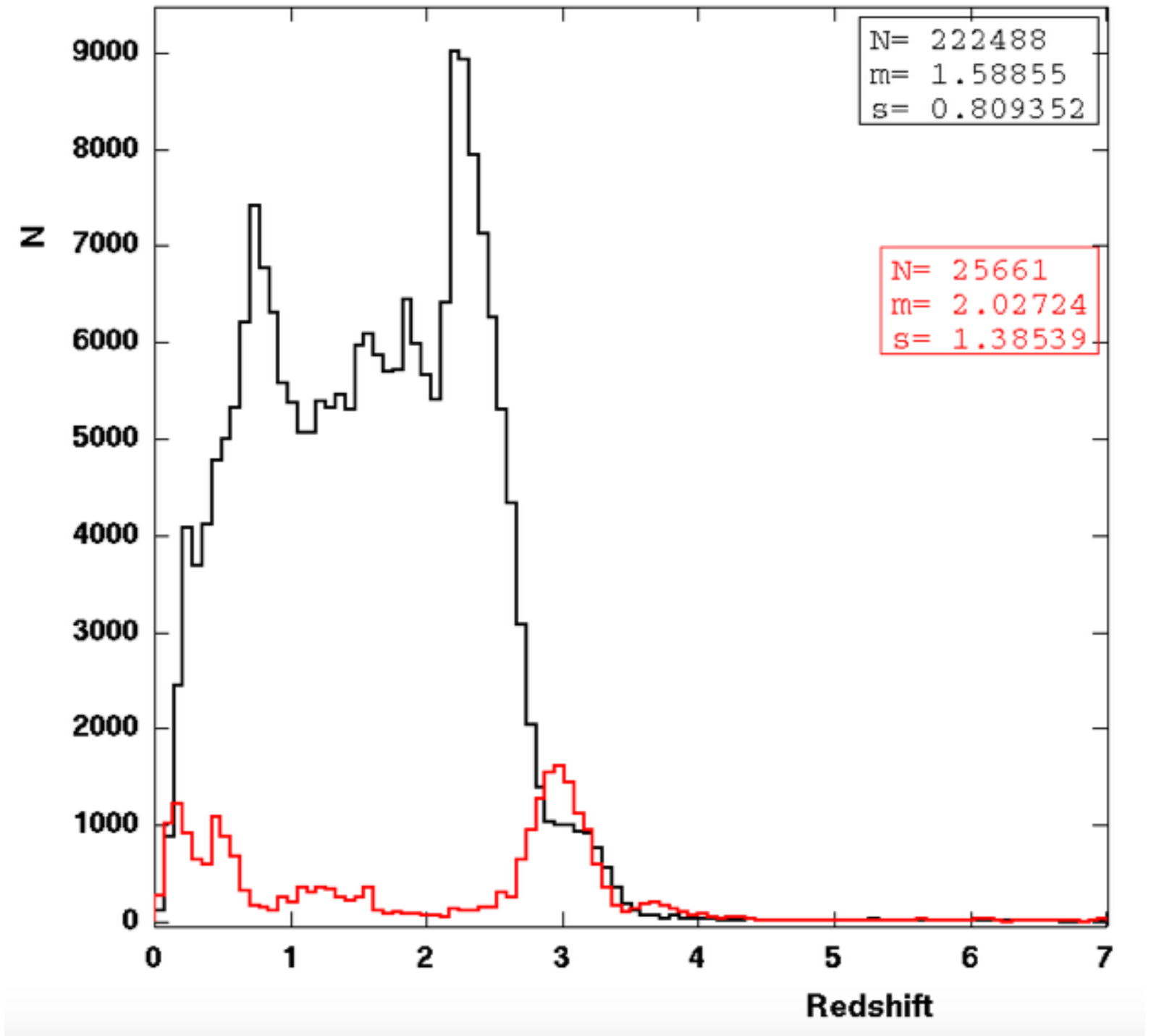




Wrongly and **correctly** classified stars



Wrongly and **correctly** classified QSOs



Comparison with Random Forest classifier

- Classification efficiency:
  whole sample: 96%
  galaxies: 97%
  stars: 94%
  QSOs: 91%

A basic classifier works nicely so far!

Classification for LSST objects

Including fainter stars to the sample

Computing the contamination of the photo-z sample

Conclusions & Perspectives

- in SDSS DR12, ~ 94% of galaxies, stars and QSOs can be correctly separated using their colours and size by implementing Logistic Regression.

- 3% of galaxies (small angular size) can be mis-classified as point-like sources.

- 8% of (faint) stars can be mis-classified as galaxy-QSO.

- 10% of (further) QSOs can be mis-classified as galaxy-star.

- Classifying the simulated objects according to the LSST observation ability (higher redshifts and fainter objects).

- What is the effect of misclassified objects on photo-z determination of galaxies?