Object classification in SDSS DR12

Farhang Habibi
LAL - CNRS, France

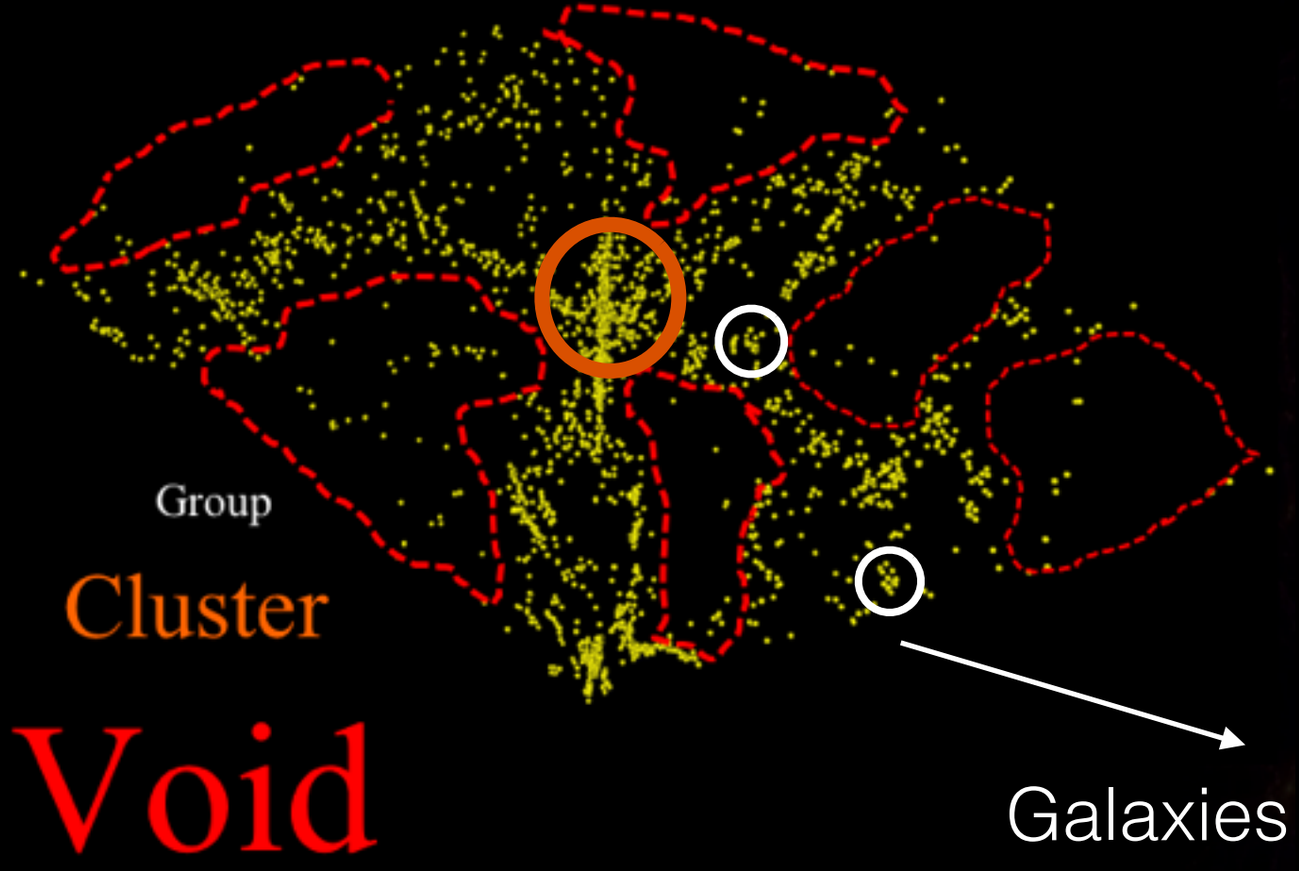
Aim

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

Cosmological surveys

All sky surveys —> cosmic structures

Deep surveys —> structures formation & evolution



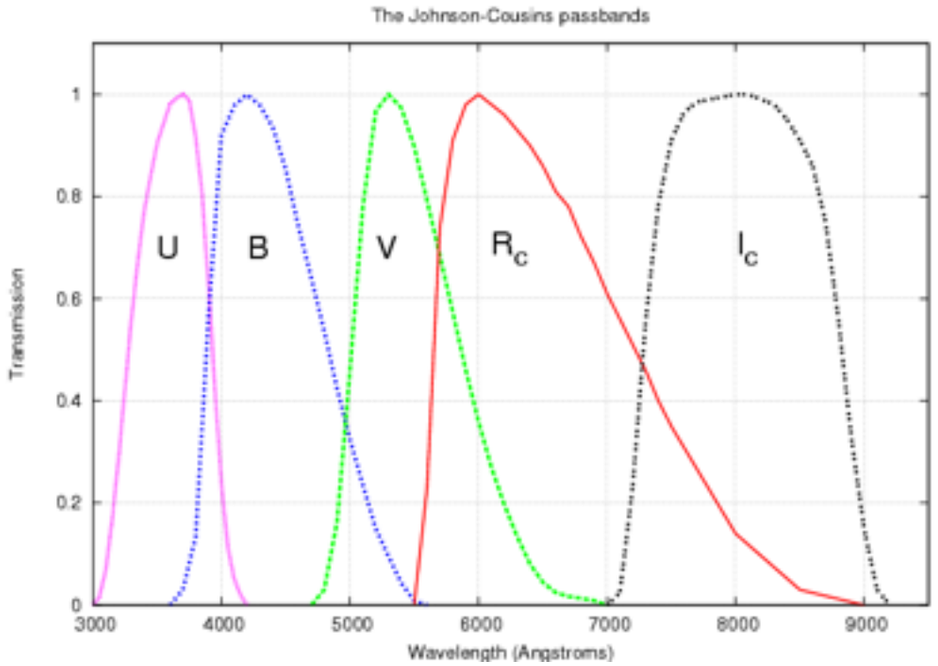
To know about the nature of
Dark Matter & Dark Energy

Galaxies are units of cosmic structures

Object classification

- **Cosmic structures contain galaxies.**
- **Images taken by surveys include galaxies, QSOs and foreground stars.**
- **How to separate these three objects?**

Features from photometric data

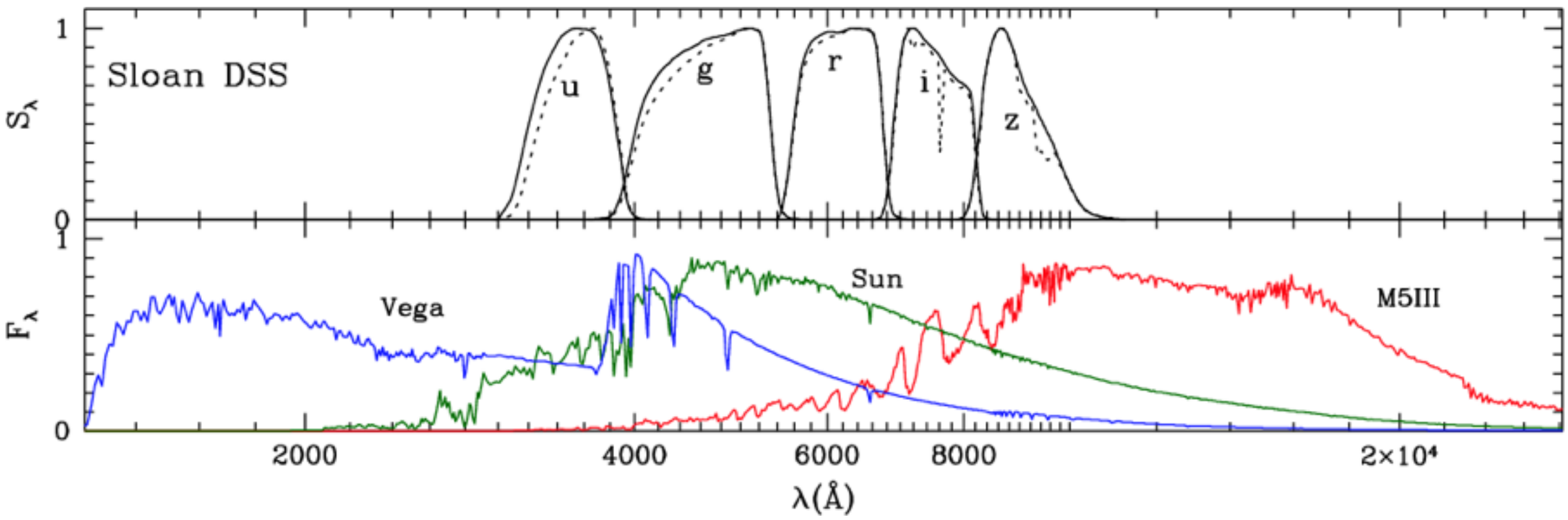


Magnitude in a filter:

$$m(f_\lambda) - m_0(f_\lambda) = -2.5 \log \frac{F(f_\lambda)}{F_0(f_\lambda)}$$

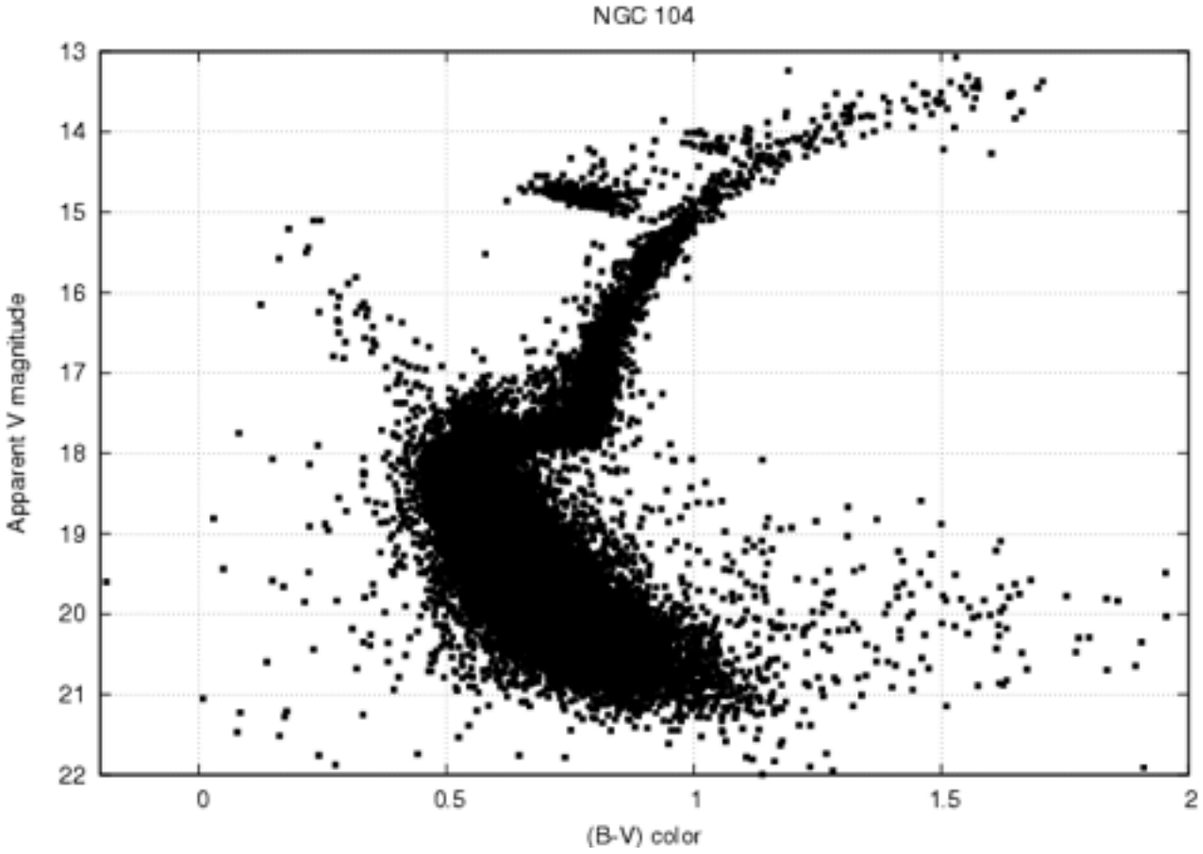
Colour index:

$$m(f_{\lambda_1}) - m(f_{\lambda_2})$$

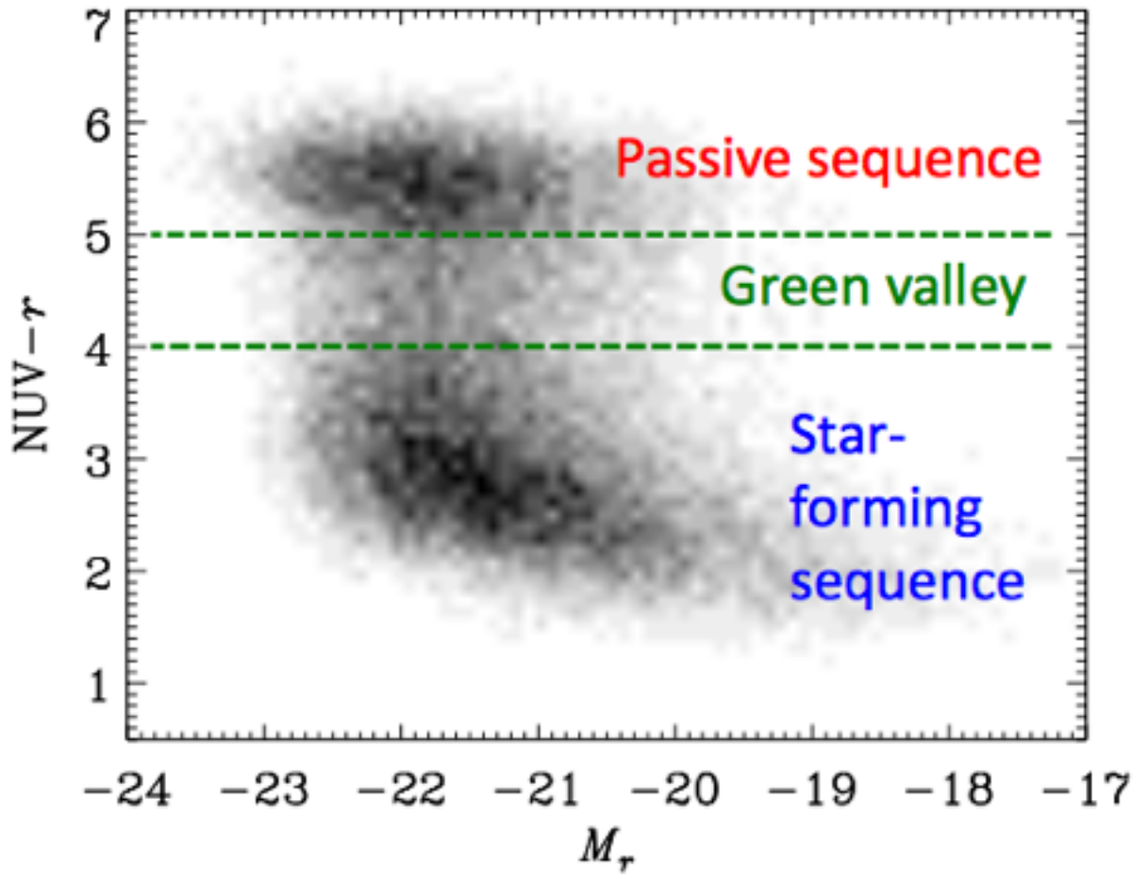


Features from photometric data

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

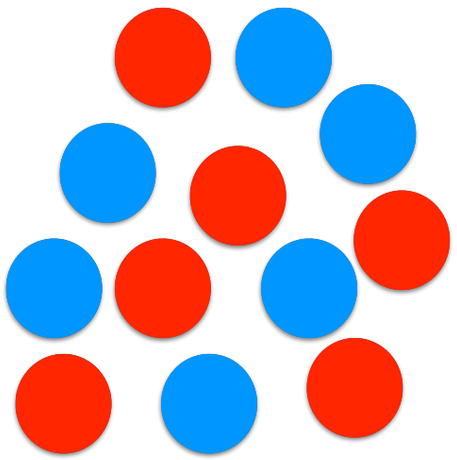


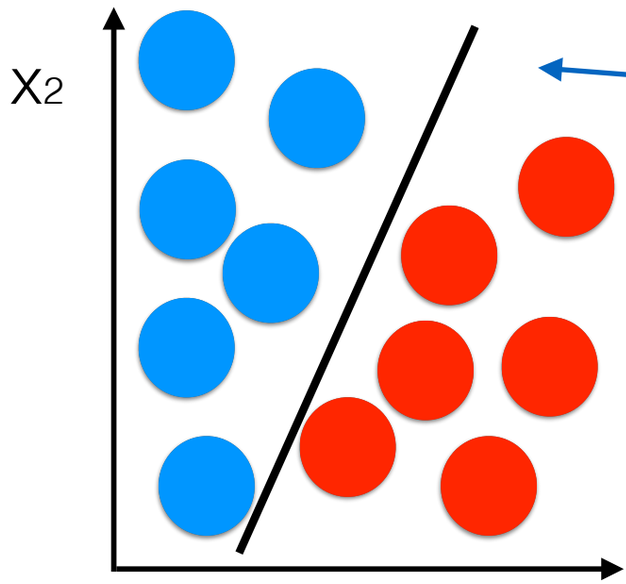
Stars



Galaxies

Supervised classification





Border curve

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

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

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

Label vector

$$\vec{y} = \begin{pmatrix} y^{(1)} = 0 \\ y^{(2)} = 1 \\ y^{(3)} = 1 \\ \vdots \\ y^{(m)} = 0 \end{pmatrix}$$

Feature matrix

$$X = \begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{pmatrix}$$

Logistic Regression

Supervised classification

How to classify a mixture of objects with different classes?

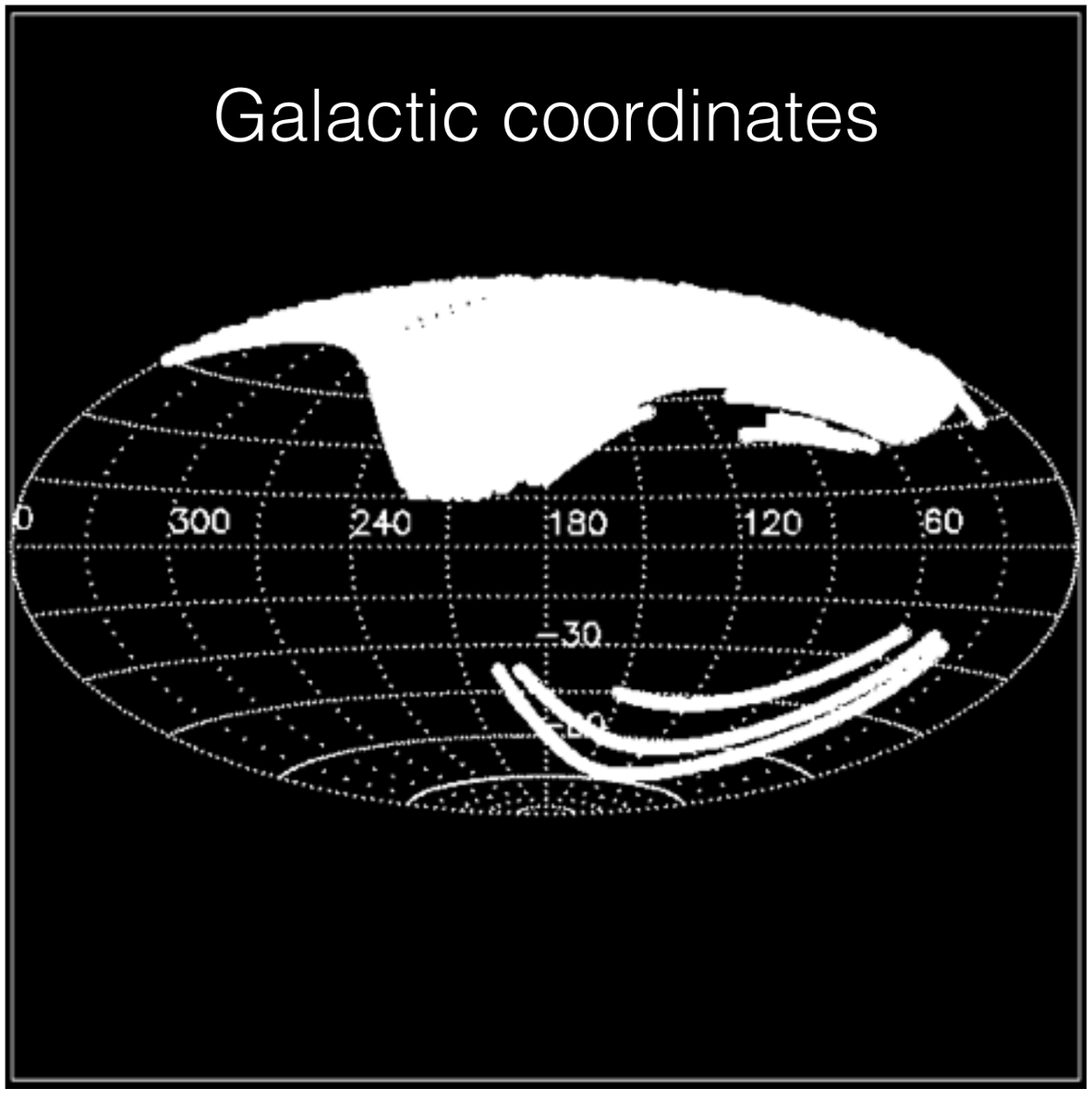
For objects with known classes:

Make a training set: Features + Labels (0 or 1)
Determine the parameters of the border curve

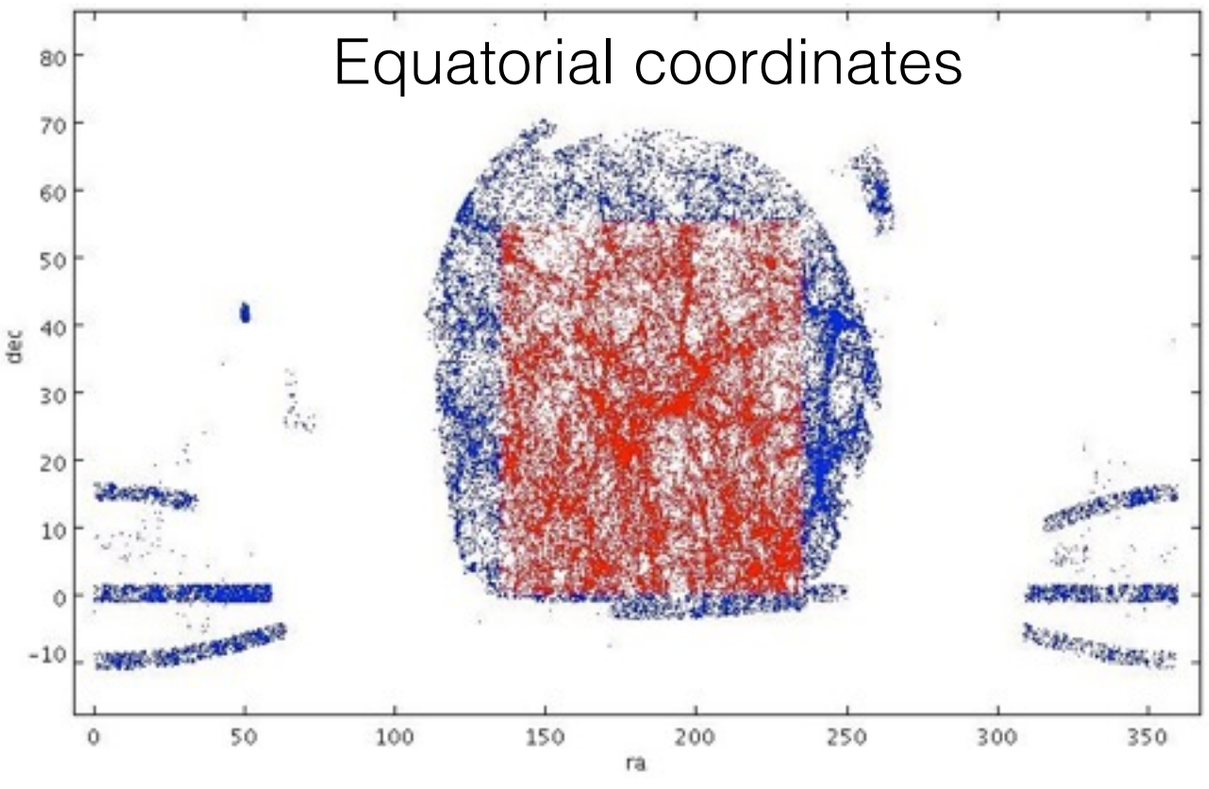
For objects with unknown classes:

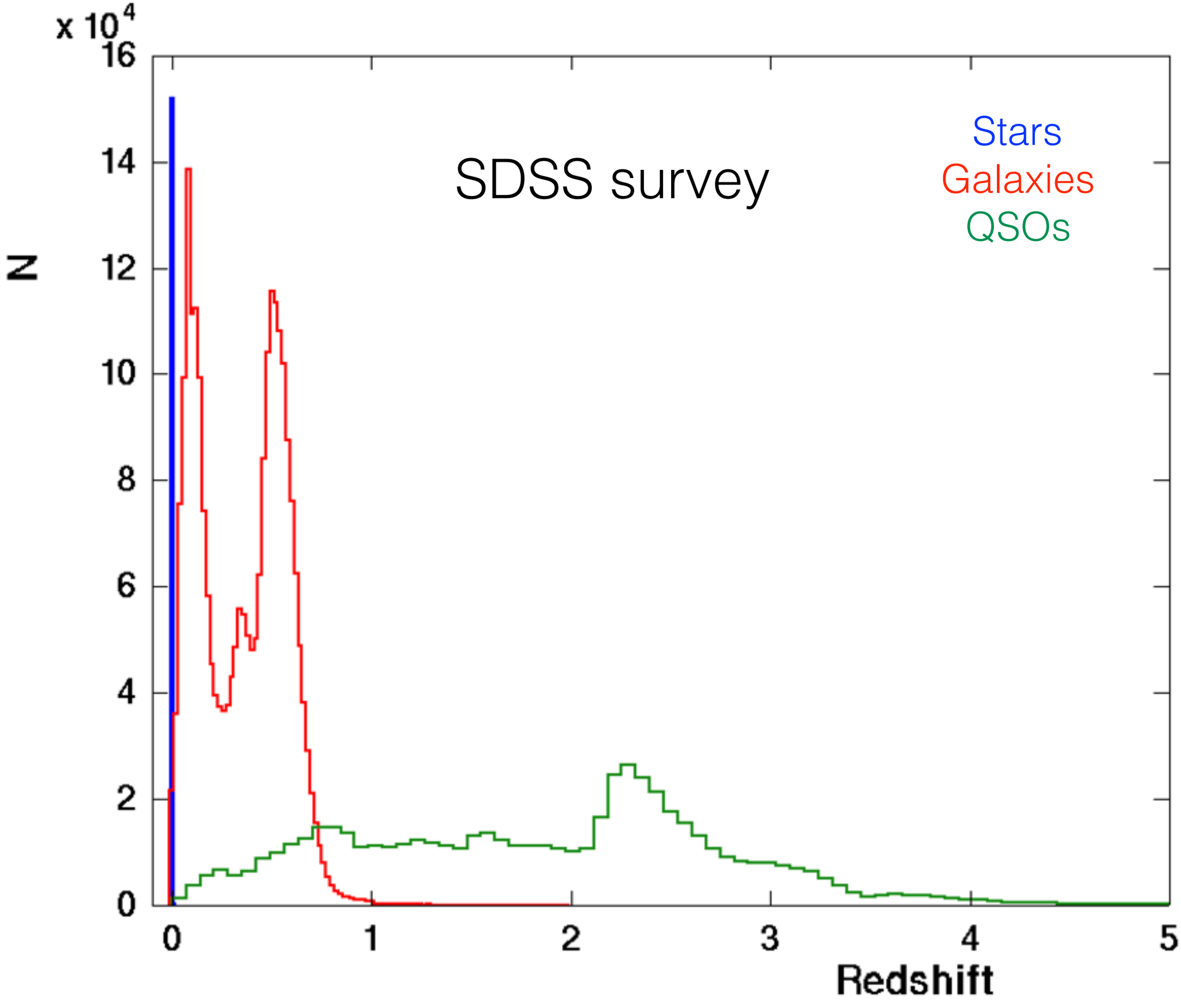
Label them using their features + the determined parameters

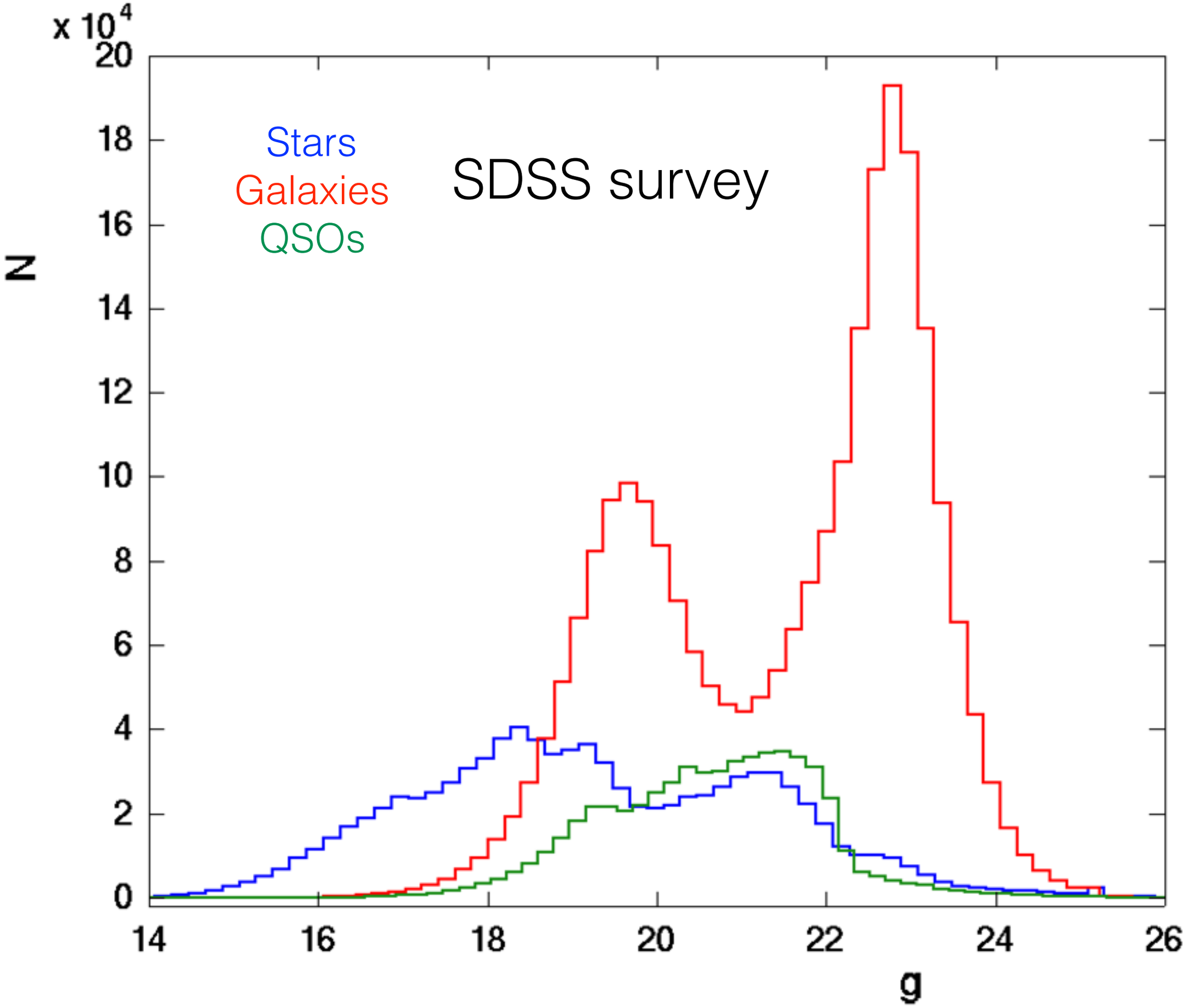
SDSS survey



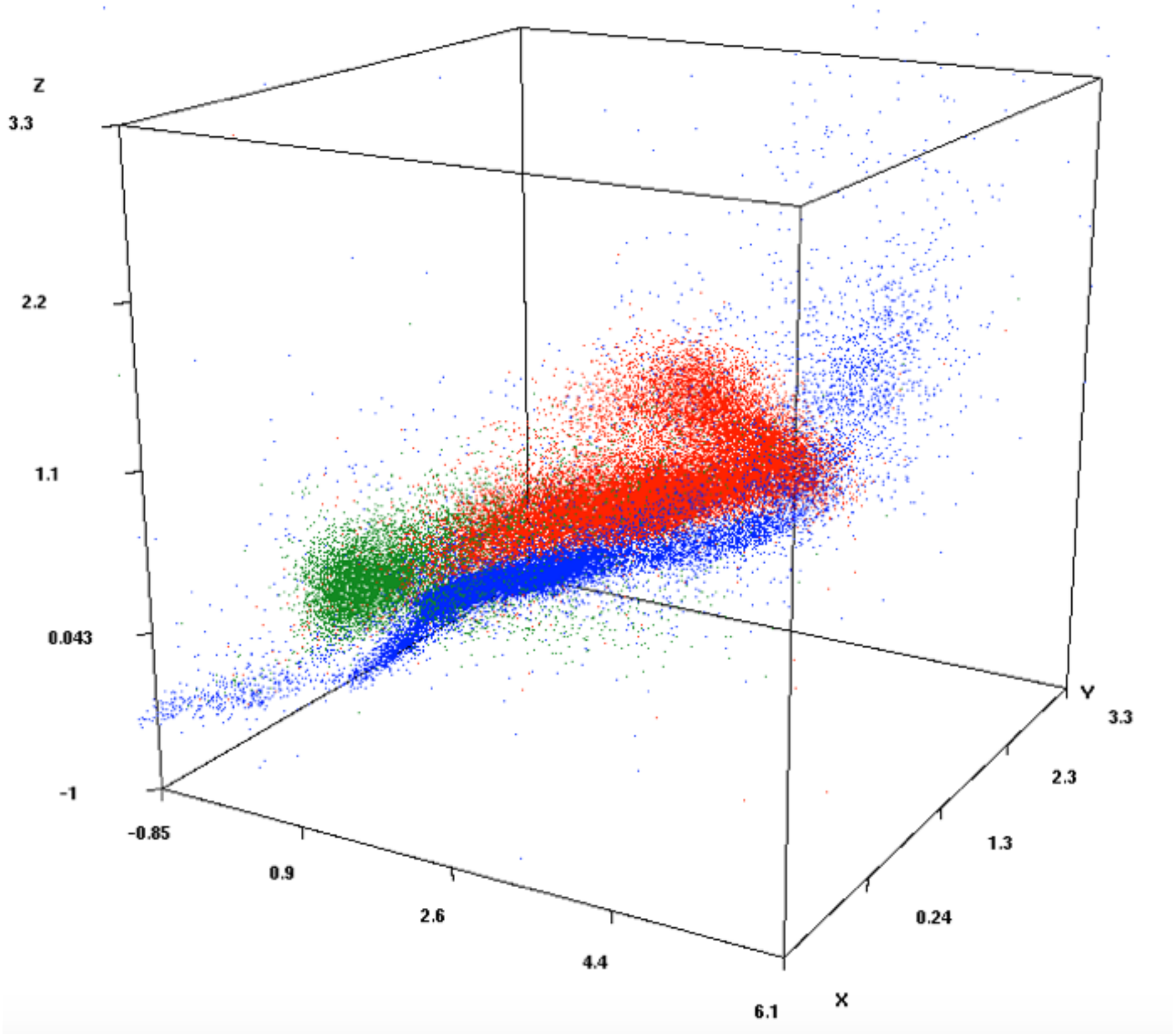
2 m class telescope
complete up to ~ 2.6 GLy
~ 4 million spectroscopically classified objects



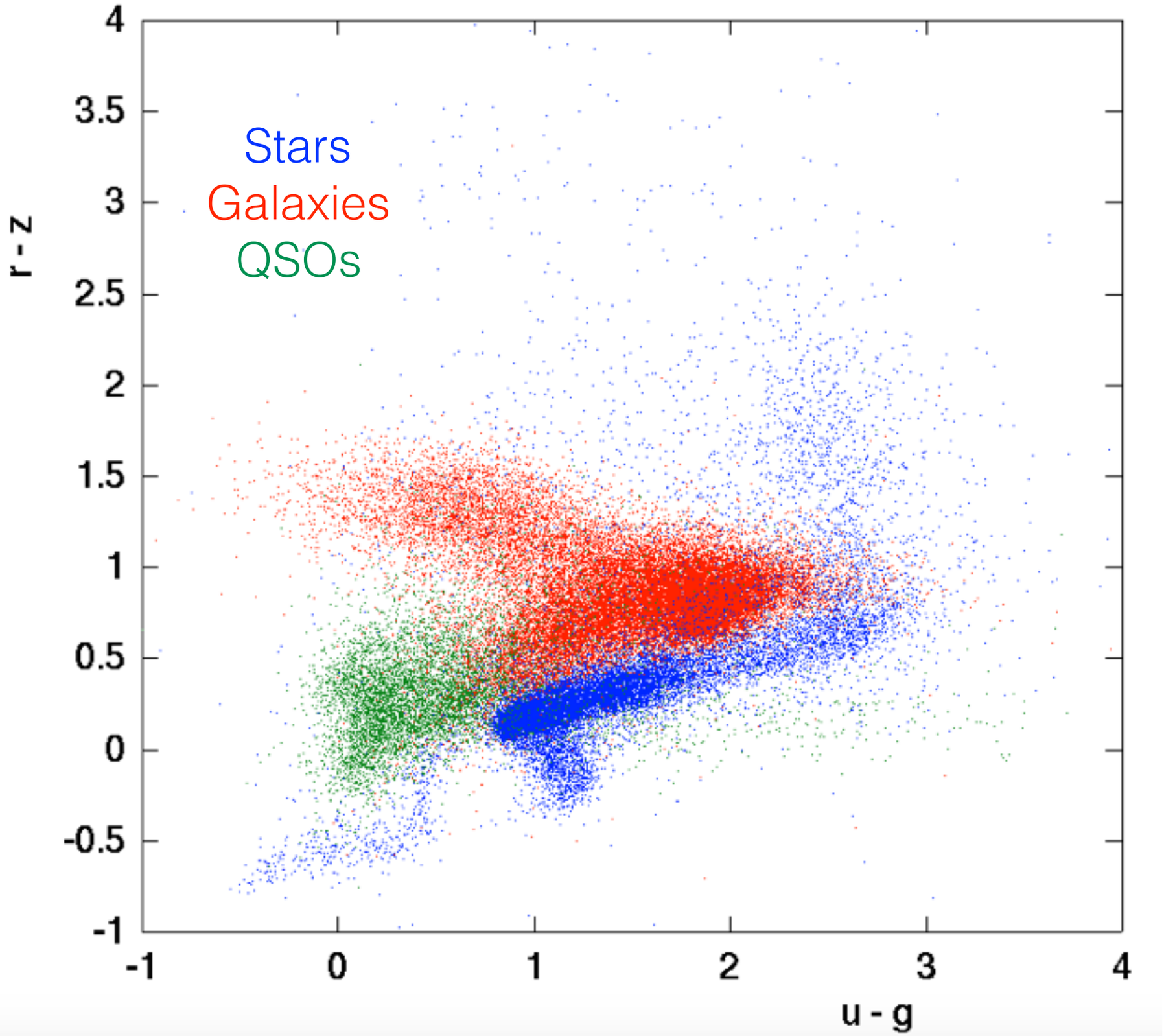

x 10^4
16
14
12
10
8
6
4
2
0
N
SDSS survey
Stars
Galaxies
QSOs
0
1
2
3
4
5
Redshift


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

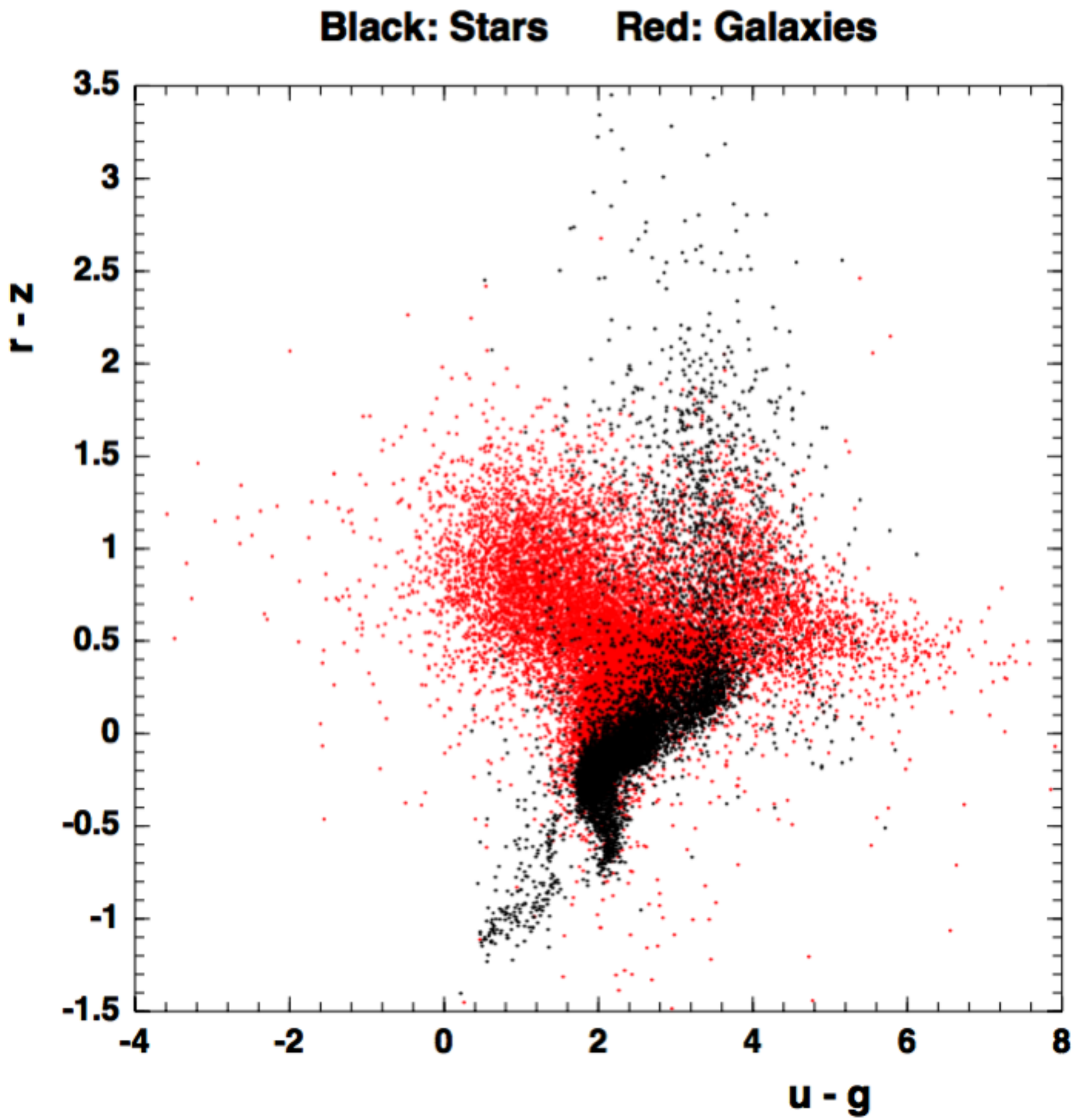


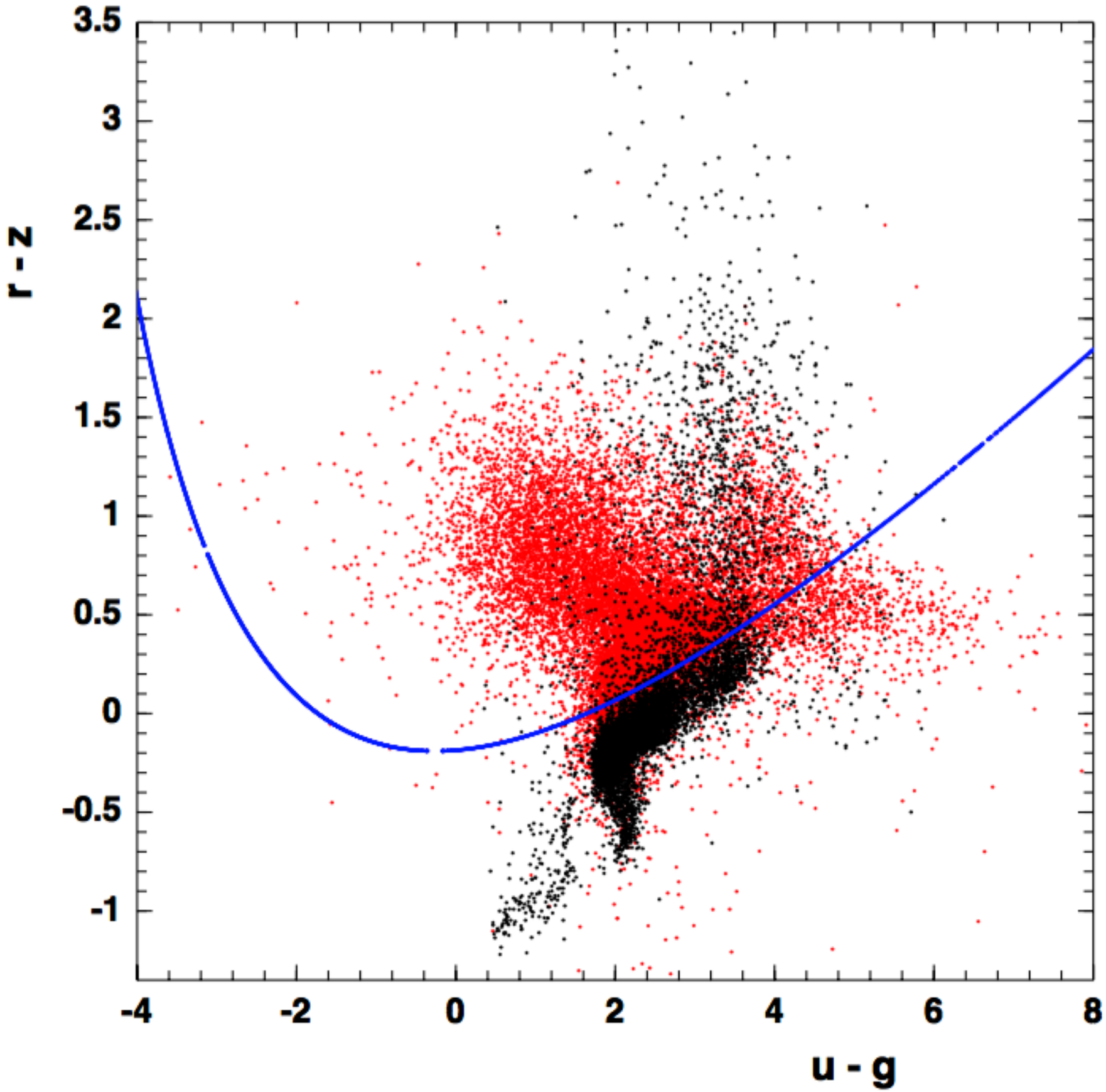
Colour indices as “features” for classification



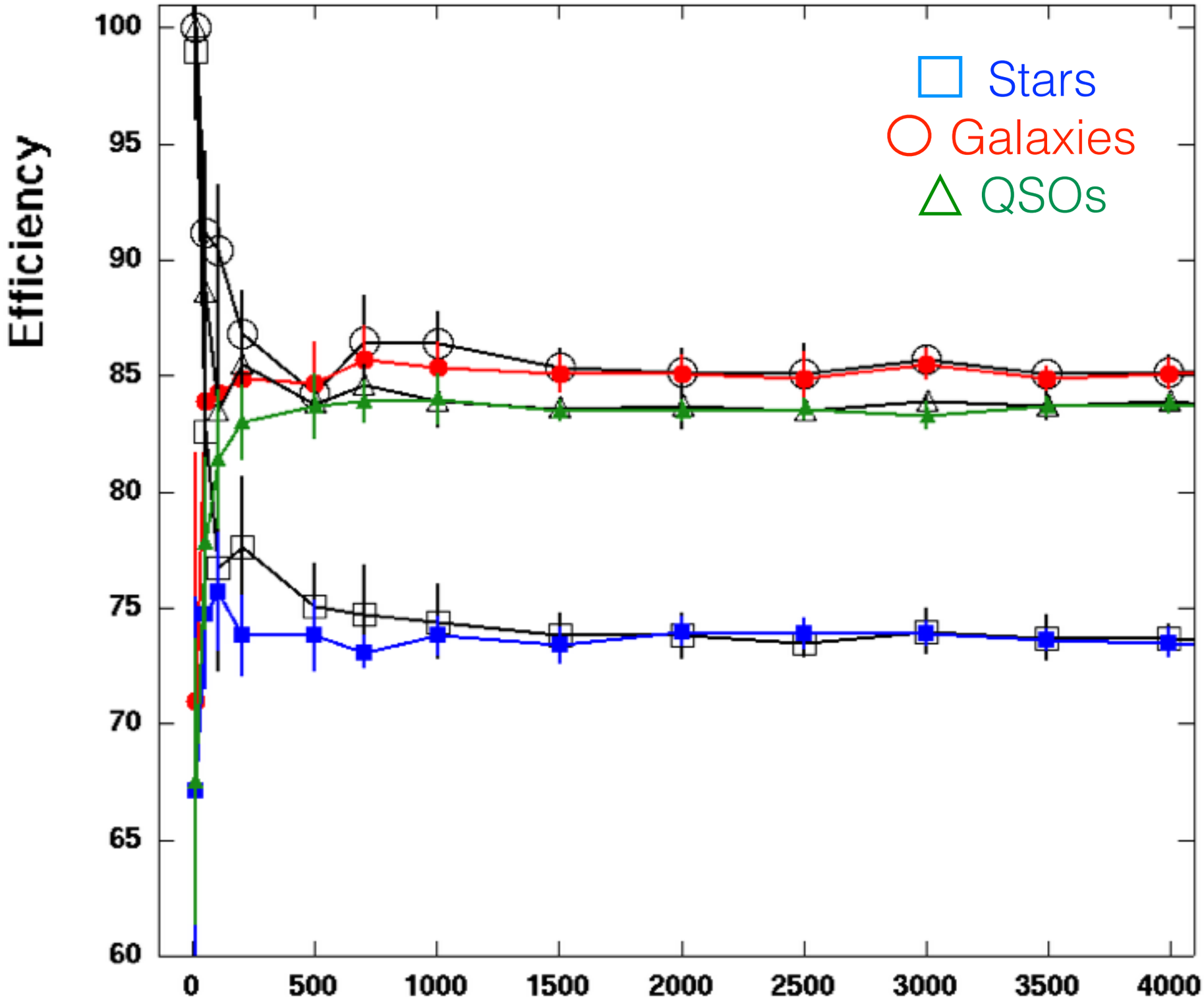
Classification features

- 4 independent colours: u-g, u-r, u-i, u-z
- Consider 4 colours, the g-magnitude and all multiplications, they construct 20 features defining a hyper parabola in a 5 dimension colour-magnitude space.

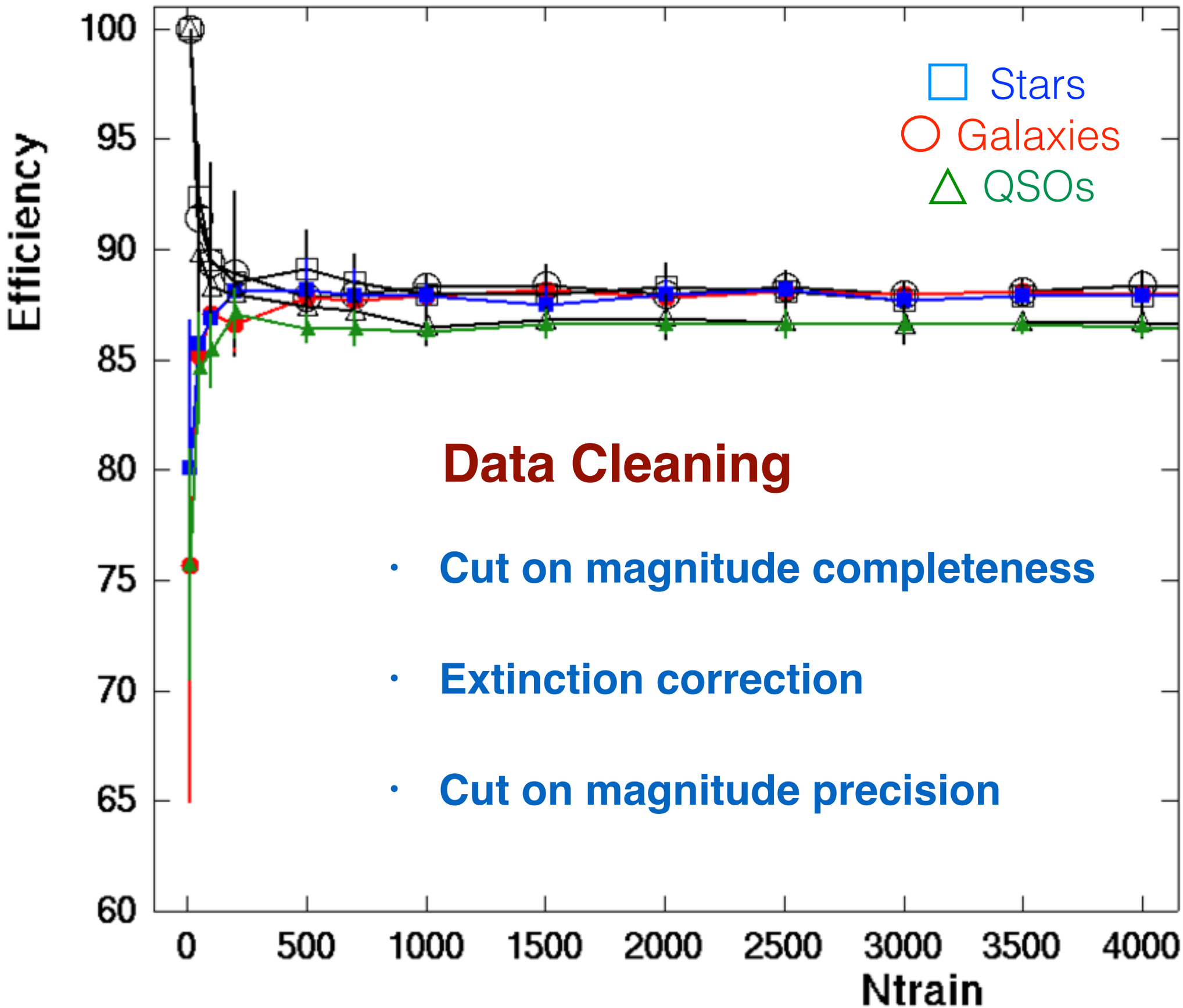

Black: Stars
Red: Galaxies
r - z
u - g
3.5
3
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-4
-2
0
2
4
6
8


Black: Stars
Red: Galaxies
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

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



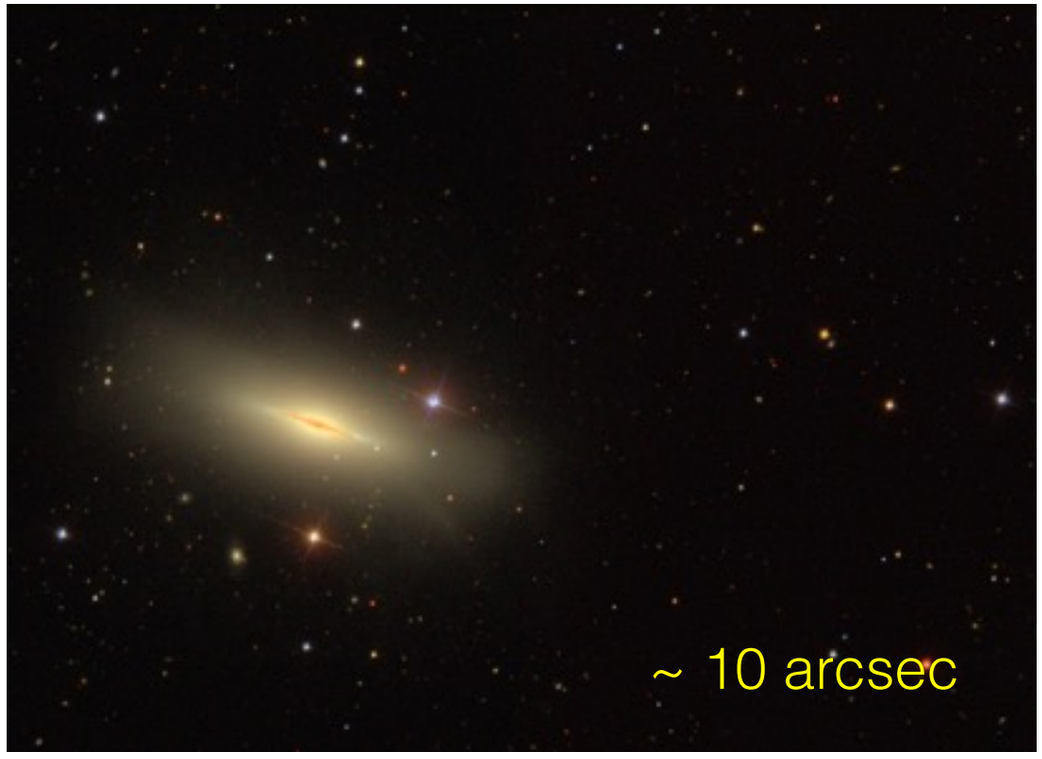
number of training objects per class ⟵ Ntrain

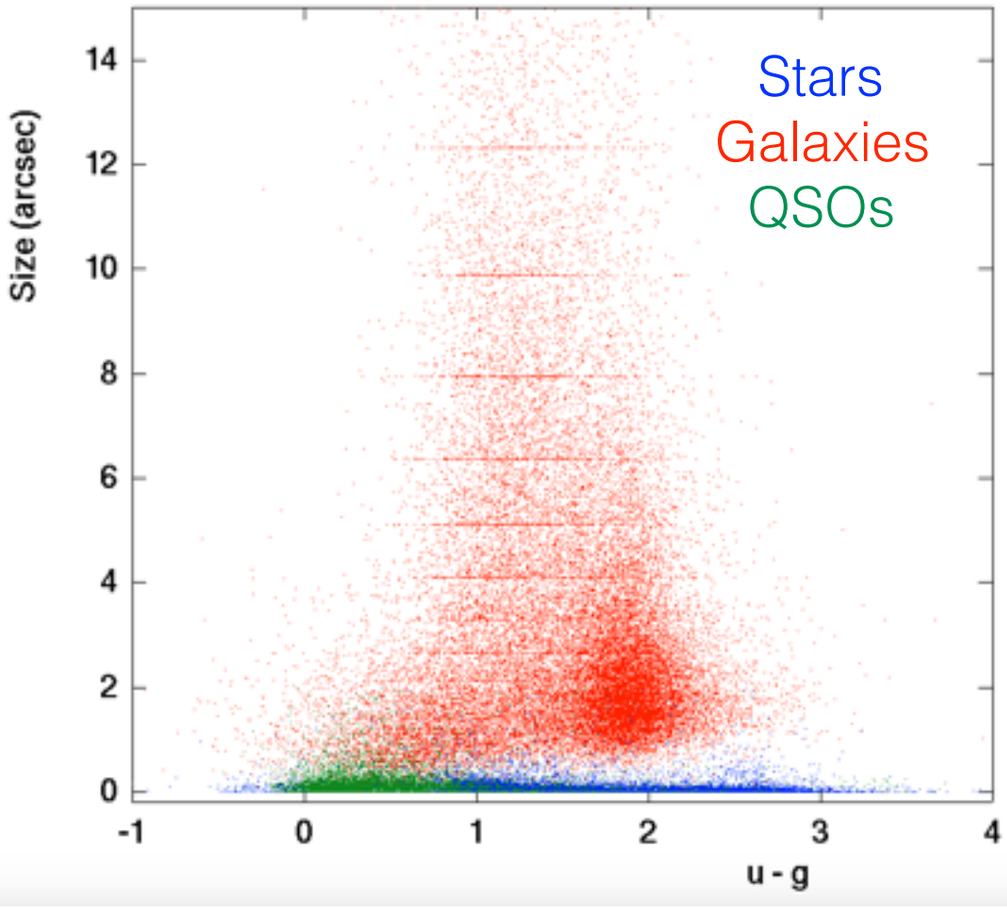


Data Cleaning

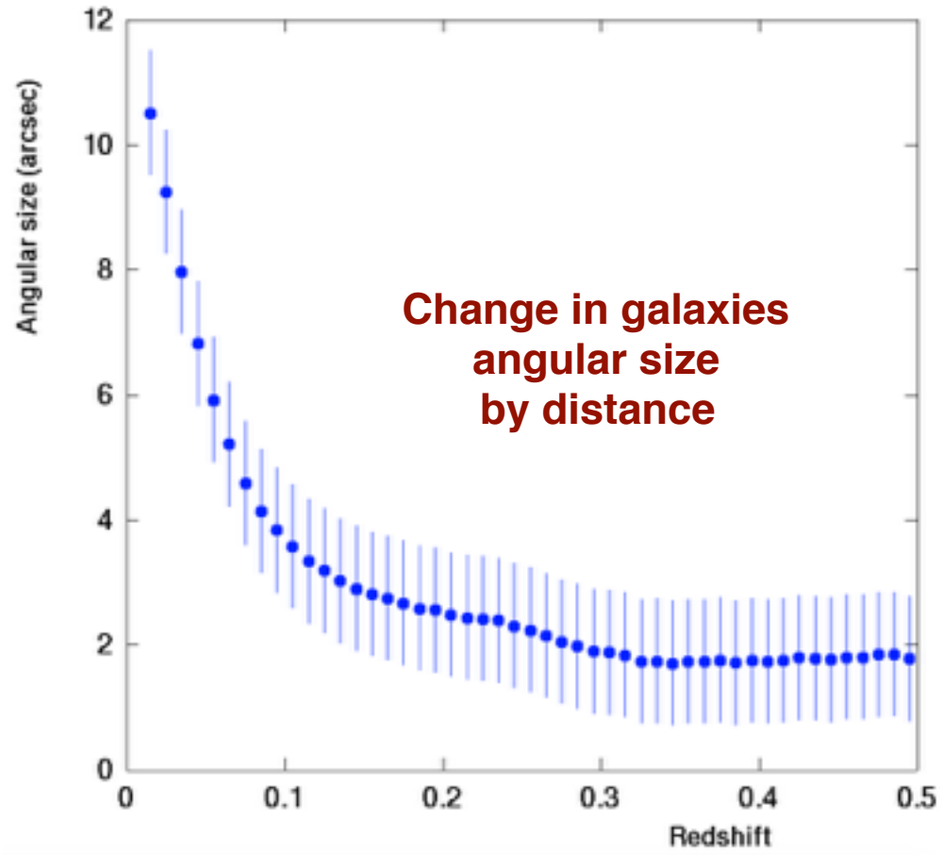
- Cut on magnitude completeness
- Extinction correction
- Cut on magnitude precision

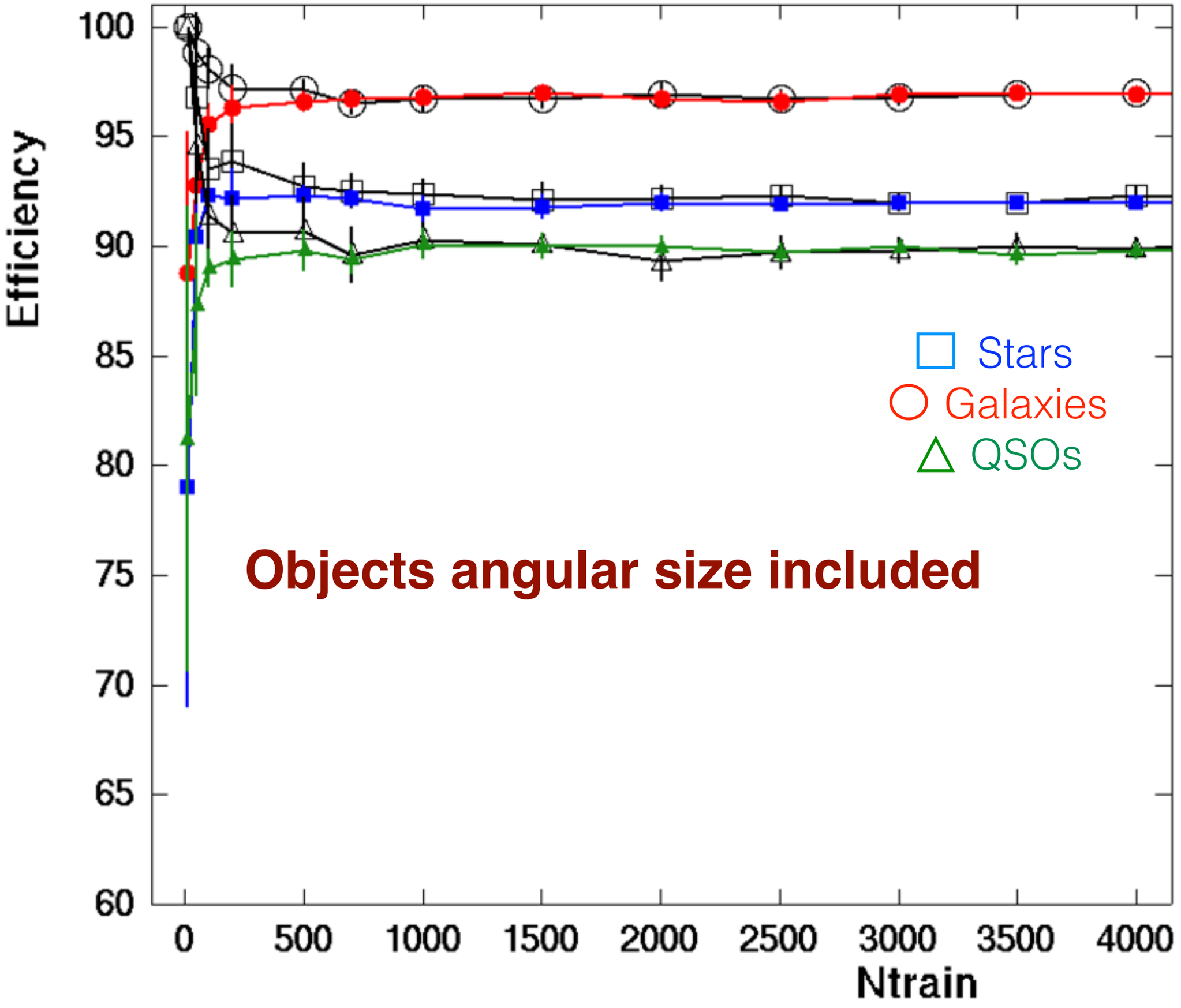
Is it possible to still improve the classification quality?


~ 10 arcsec

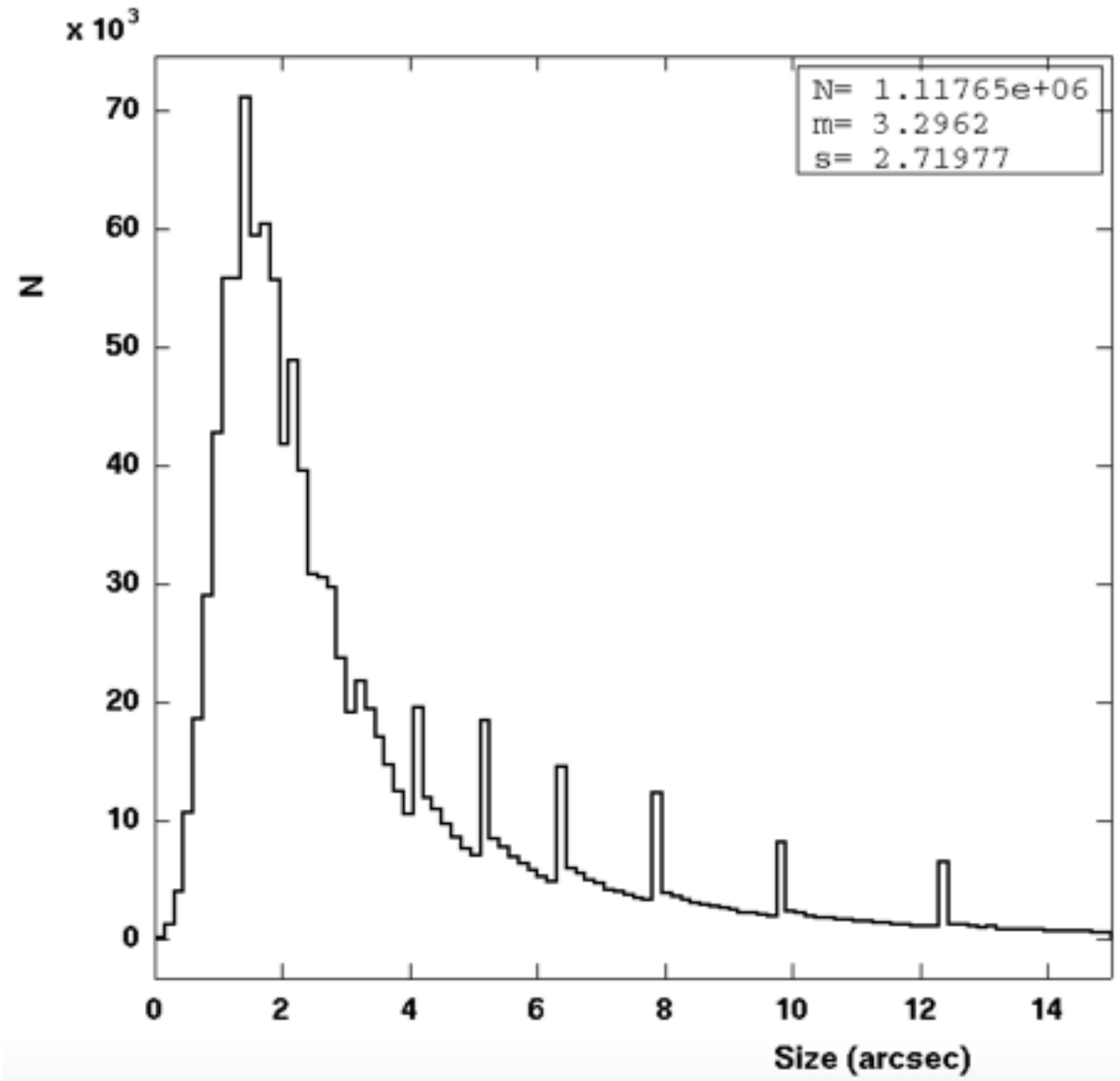

Stars
Galaxies
QSOs
Size (arcsec)
14
12
10
8
6
4
2
0
-1
0
1
2
3
4
u - g

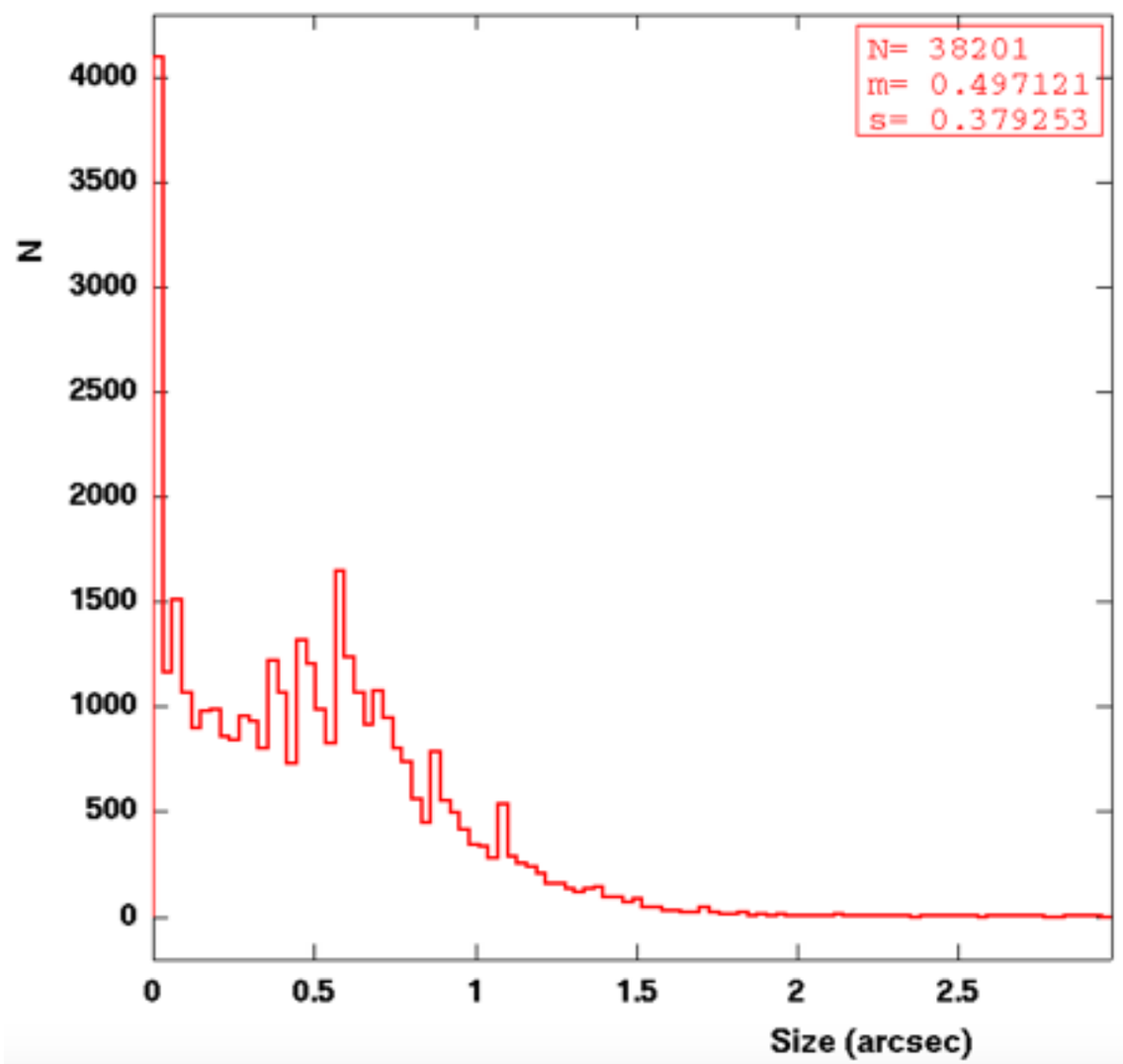

~ 1 arcsec


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

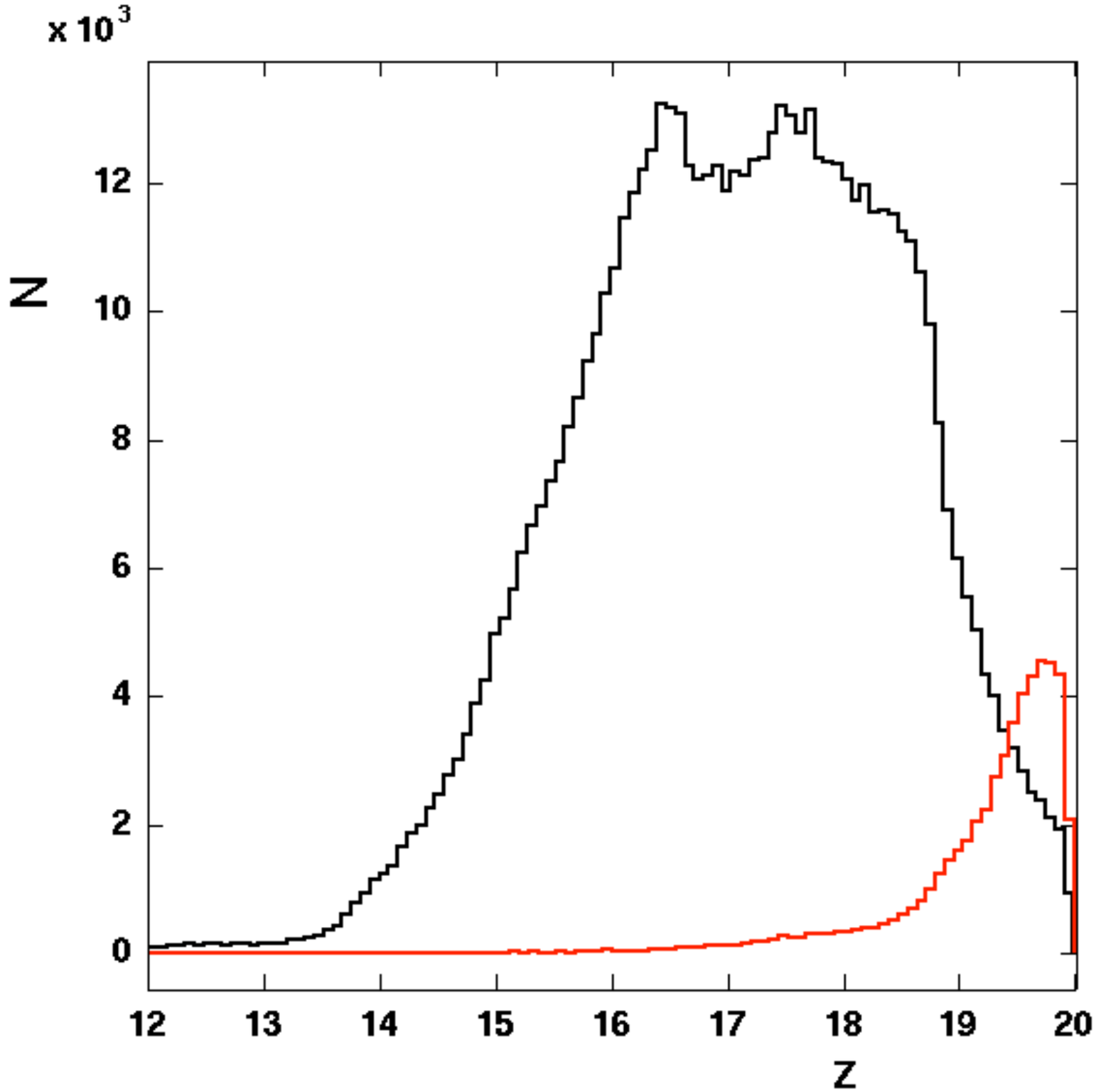

Efficiency
100
95
90
85
80
75
70
65
60
0
500
1000
1500
2000
2500
3000
3500
4000
Ntrain
Stars
Galaxies
QSOs
Objects angular size included

Wrongly and **correctly** classified galaxies

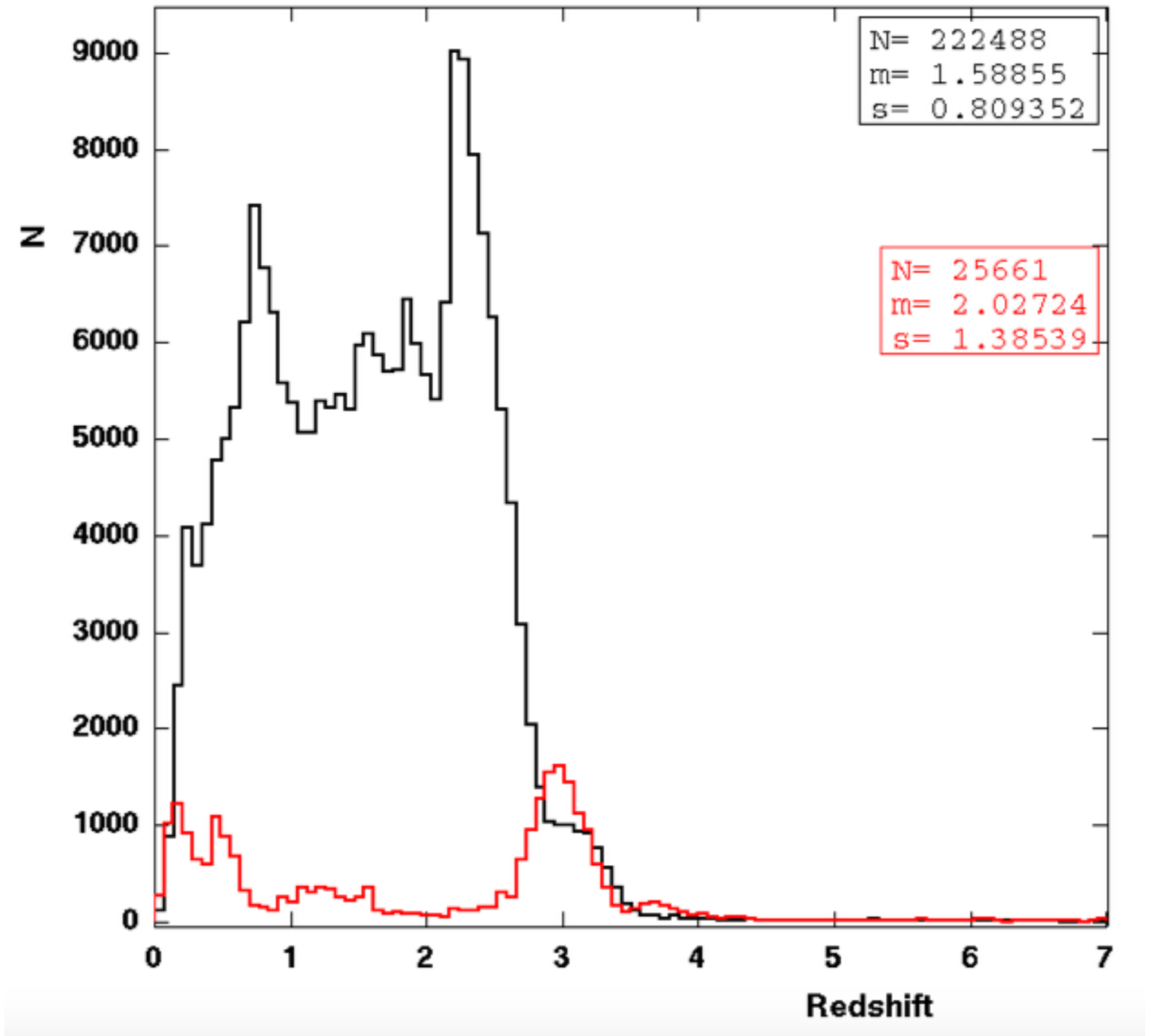




Wrongly and **correctly** classified stars



Wrongly and **correctly** classified QSOs



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

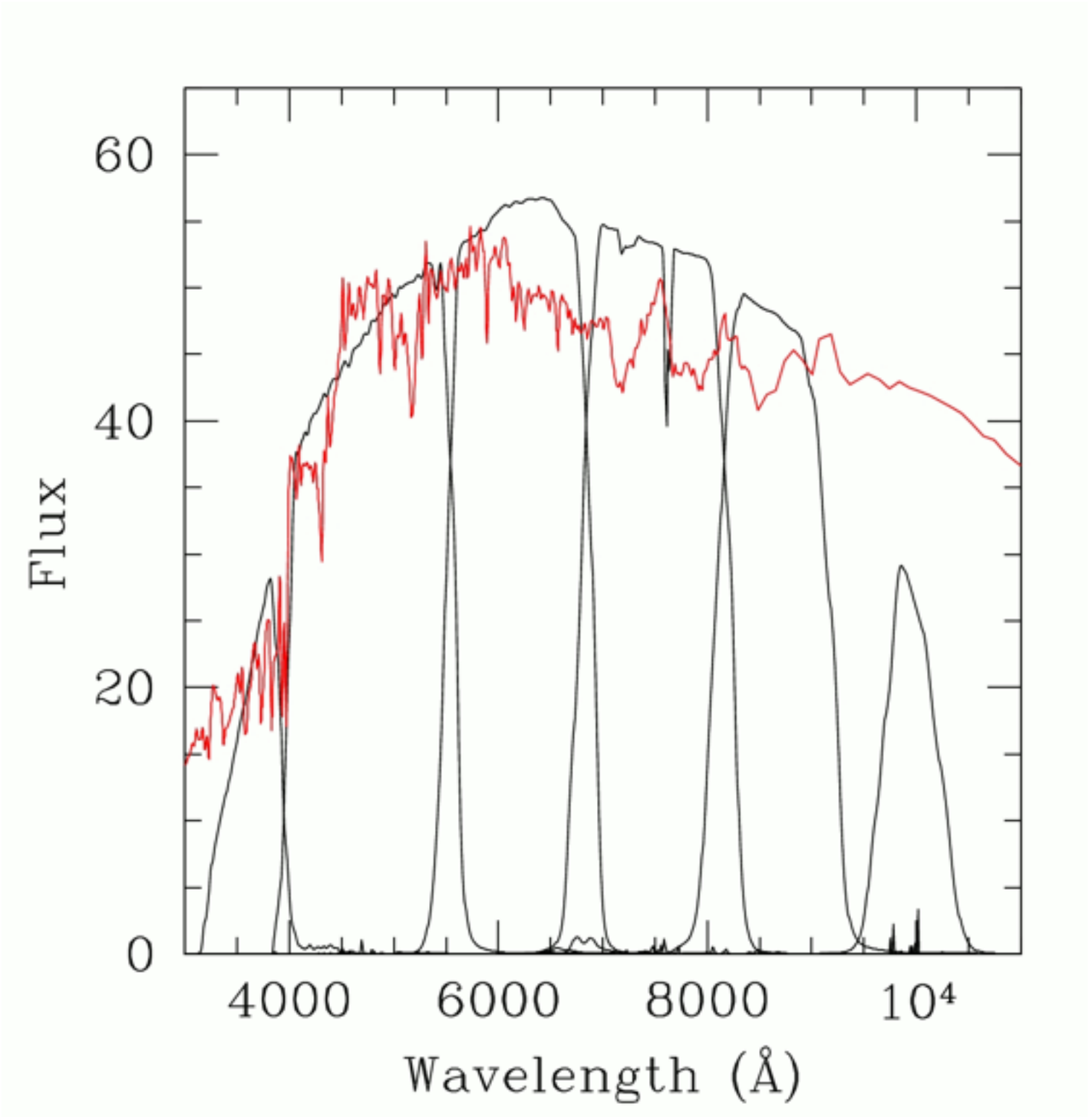
- Mean redshift of the QSOs classified
  wrongly: redshift = 2 (further QSOs)
  correctly: redshift = 1.5

Comparison with Random Forest classifier

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

A basic classifier works nicely so far!

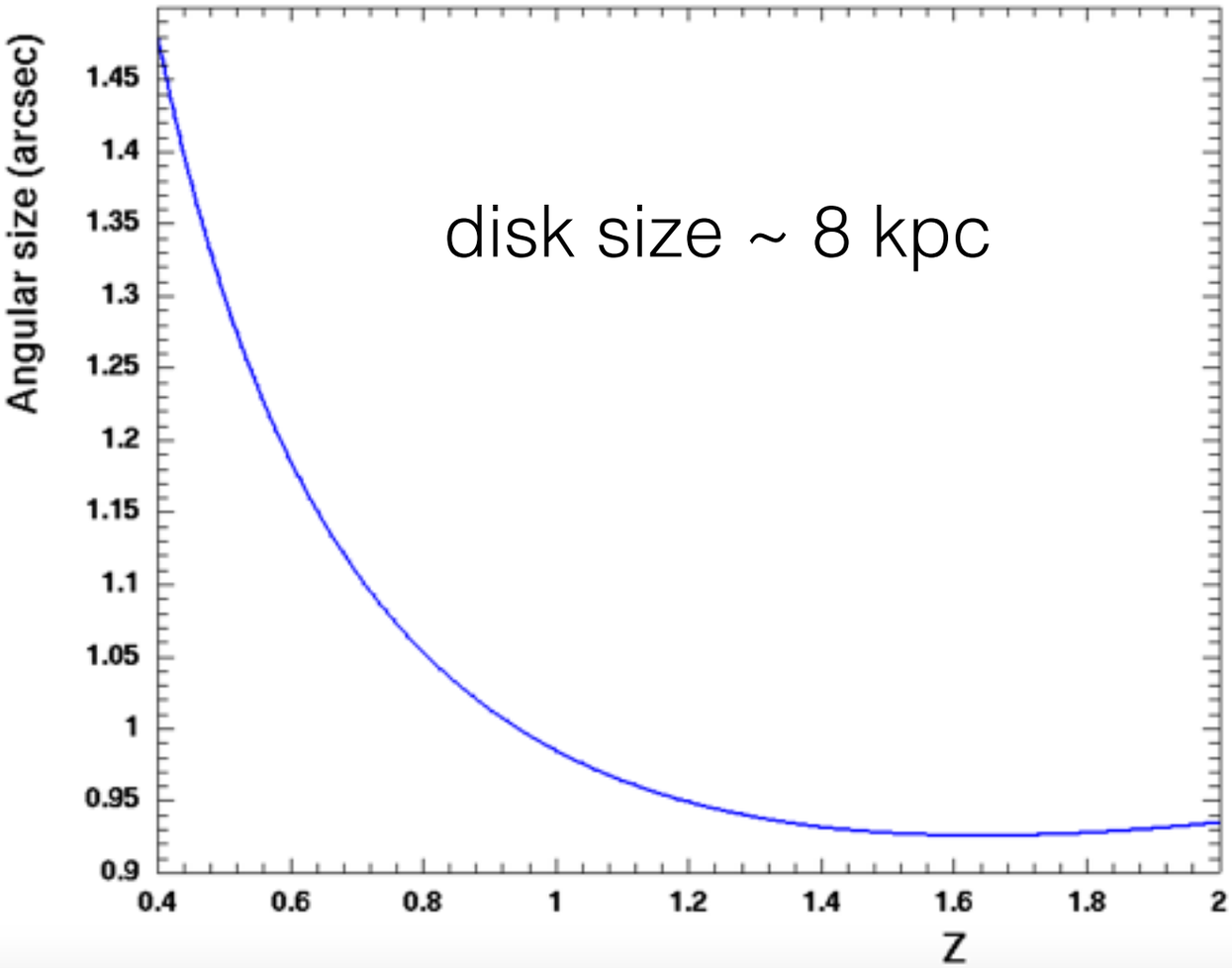
Classification for LSST objects



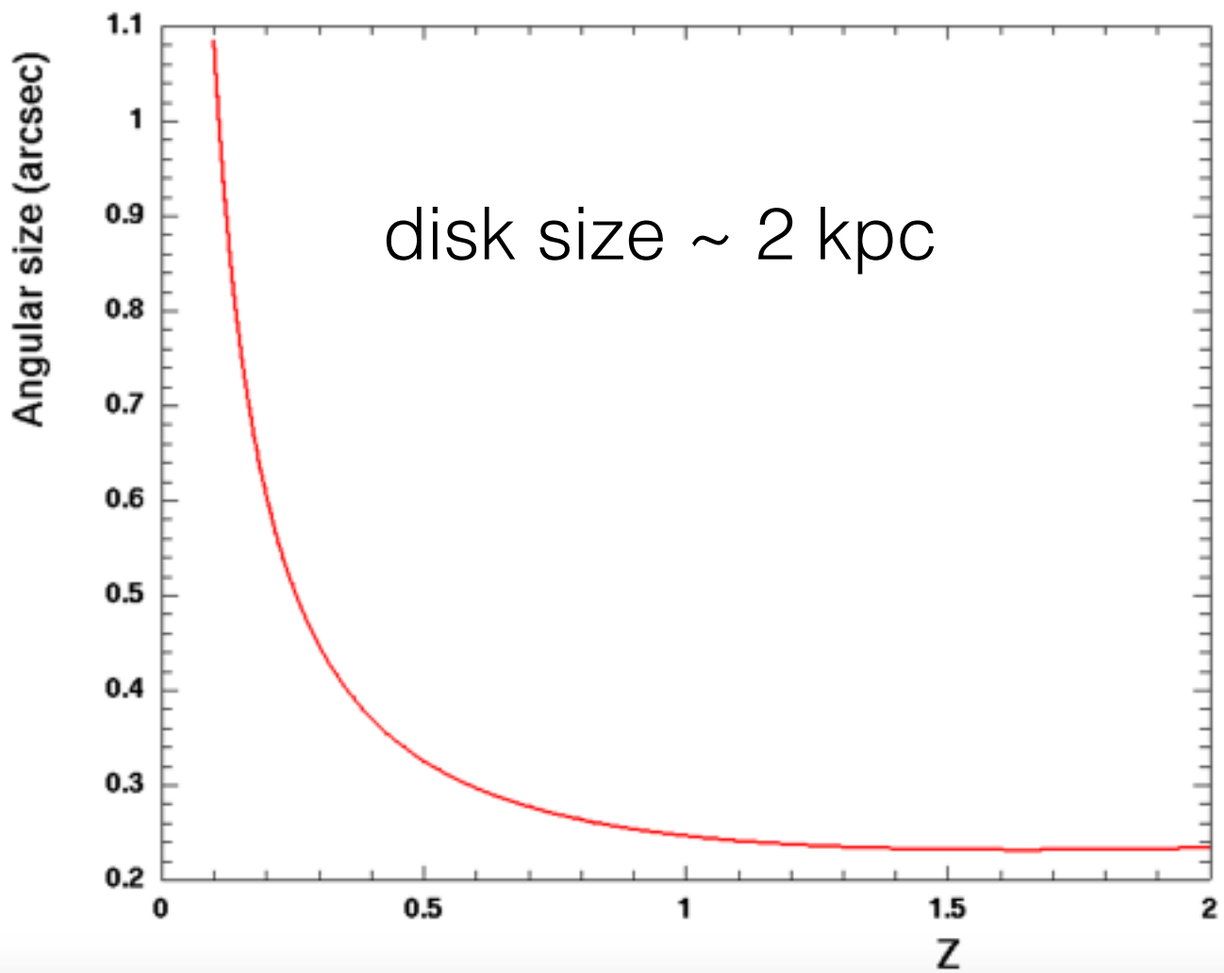
- Generating galaxies at different redshifts
- Assigning SED to galaxies
- Computing magnitudes and colours
- Assigning angular size to galaxies

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

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?