Object classification in SDSS DR12

Farhang Habibi
LAL - Université Paris SUD-11

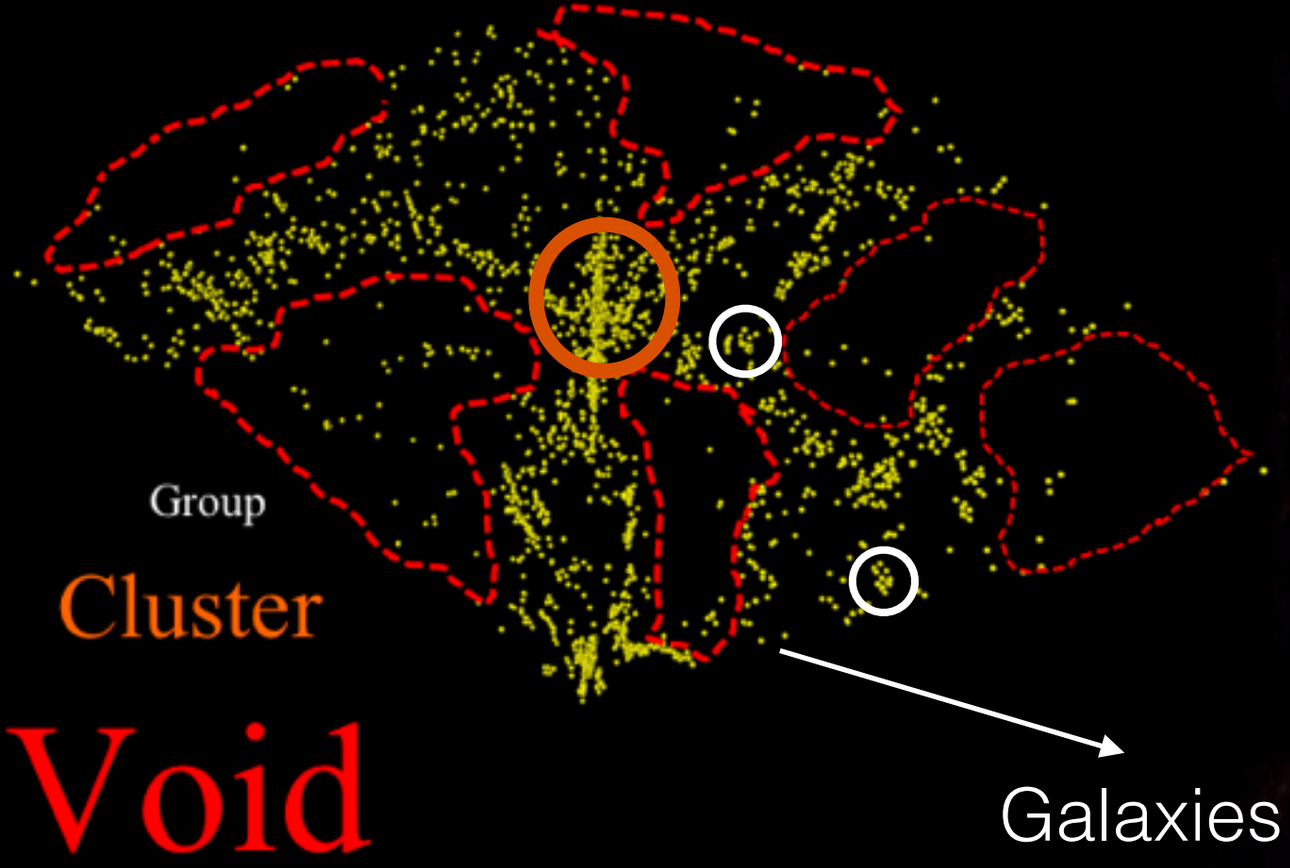
Aim

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

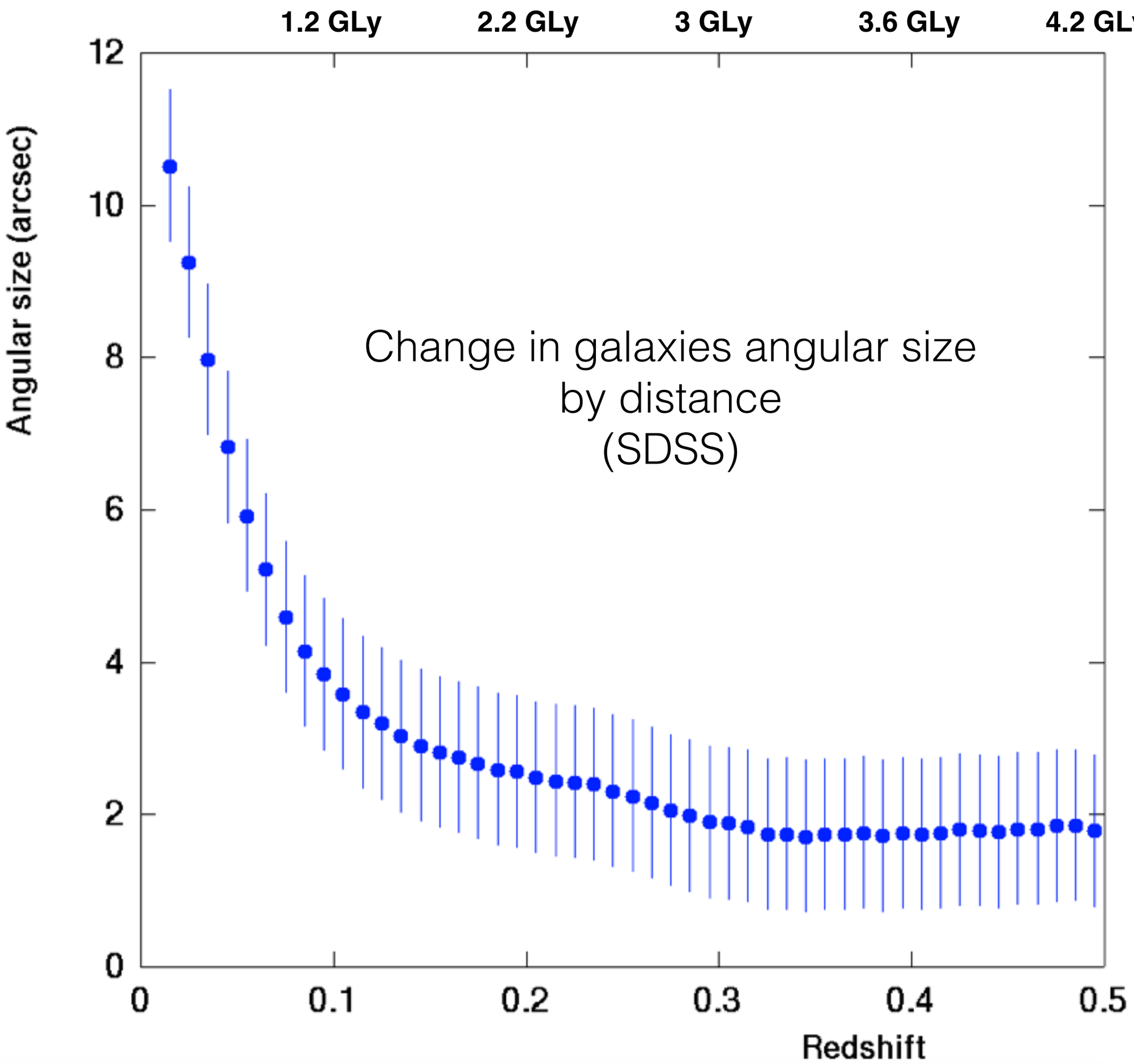
Object classification

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

Nearby galaxies

luminosity spread on CCD:
stars ~ 1 arcsec
galaxies ~ 10 arcsec

full moon ~1800 arcsec


1.2 GLy
2.2 GLy
3 GLy
3.6 GLy
4.2 GLy
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
(SDSS)

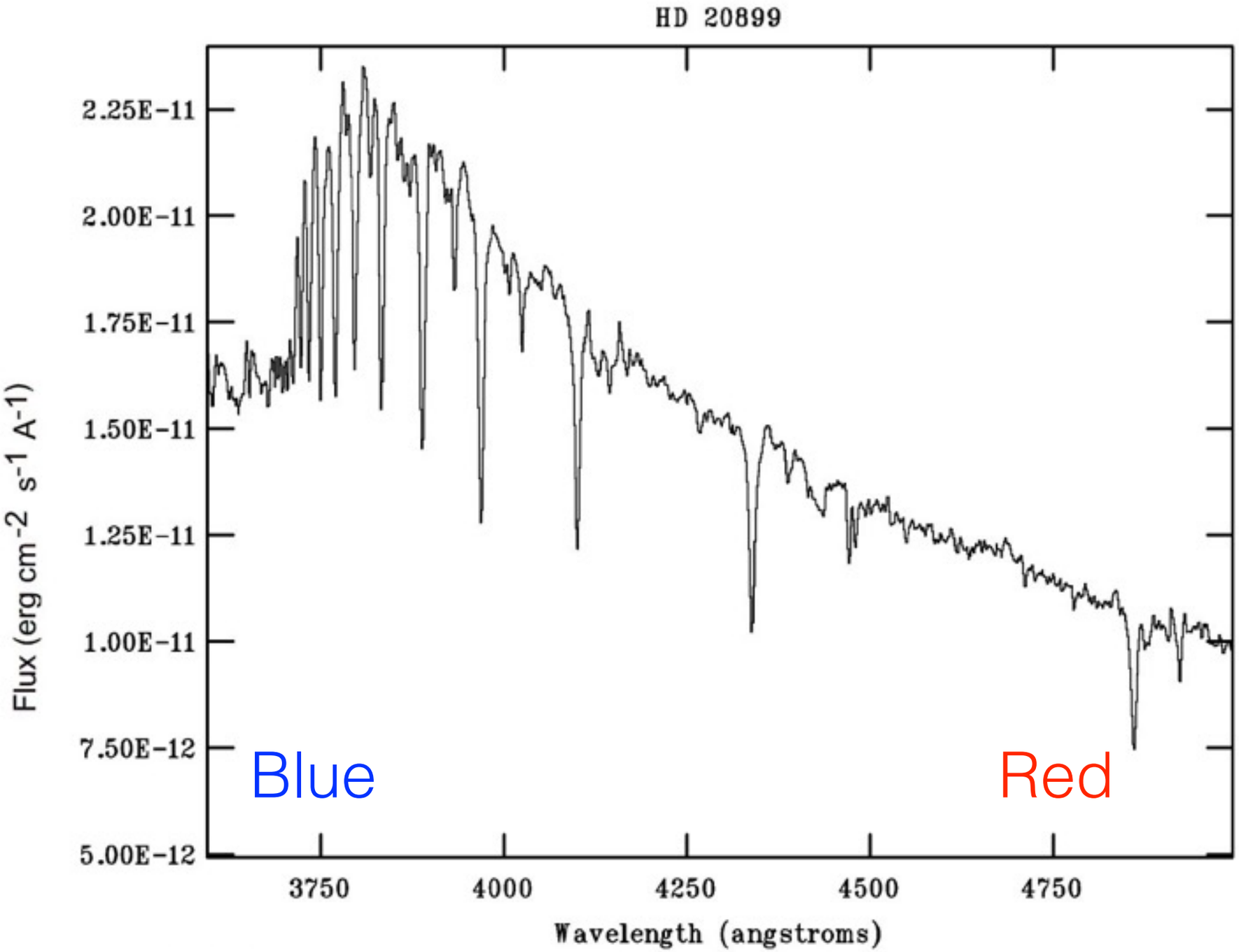
Far galaxies

luminosity spread on CCD:

stars, QSOs ~ 1 arcsec

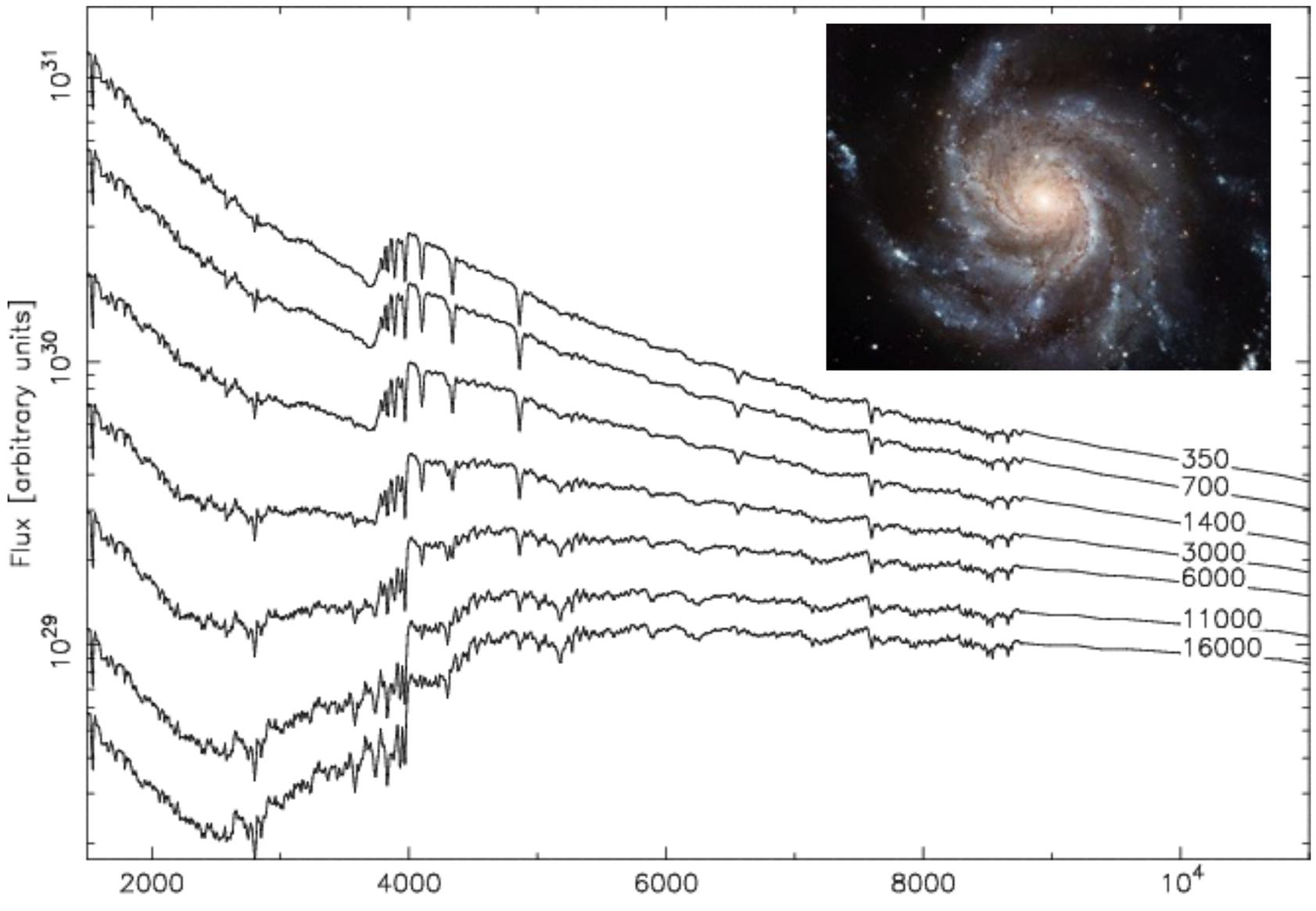
galaxies ~ 1 arcsec

Stellar energy spectrum



Jacoby, G. H., Hunter, D. A., and Christian, C. A. 1984, Astrophys. J. Supp. Ser., 56, 257

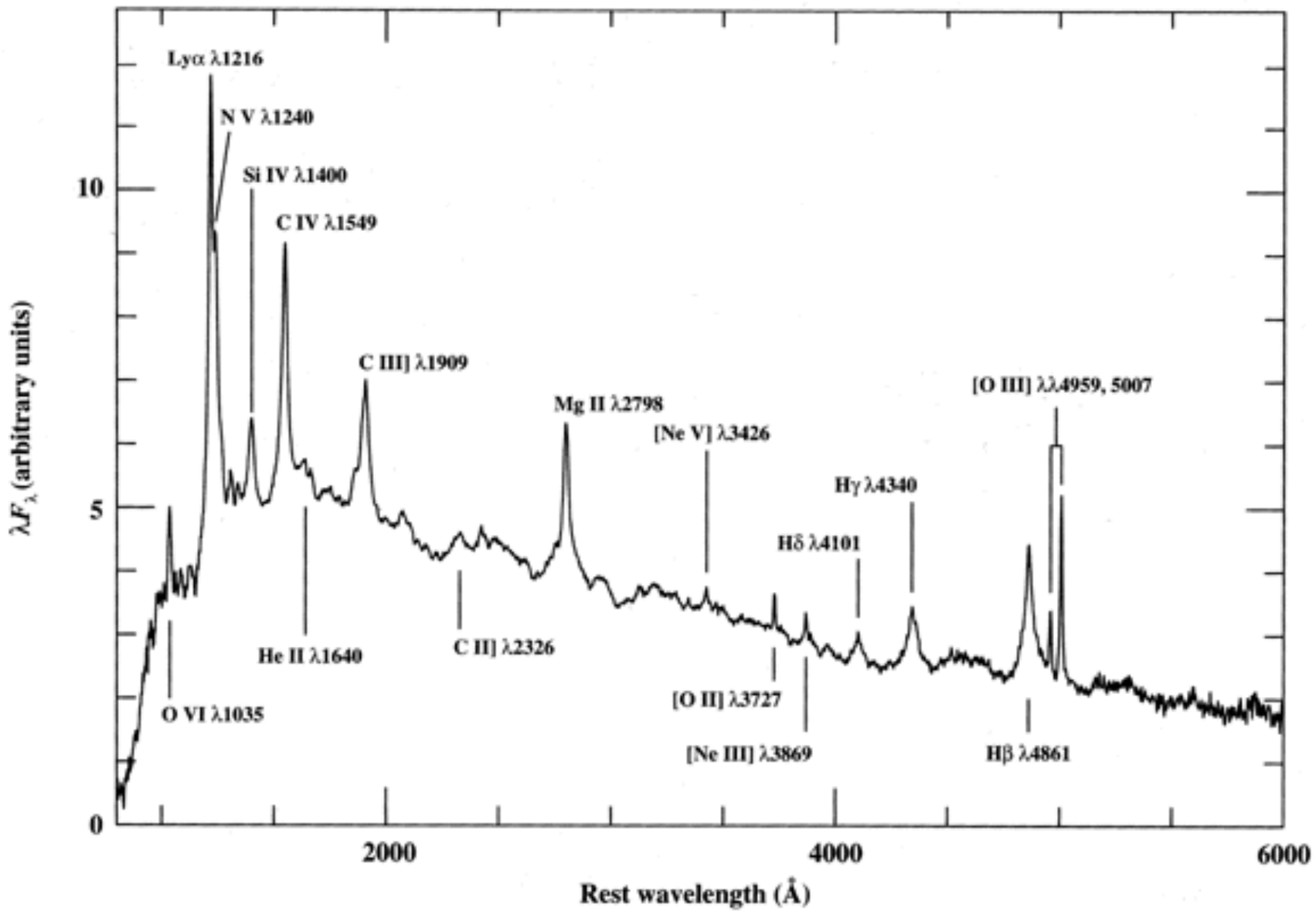
Galactic energy spectrum



Blue

Red

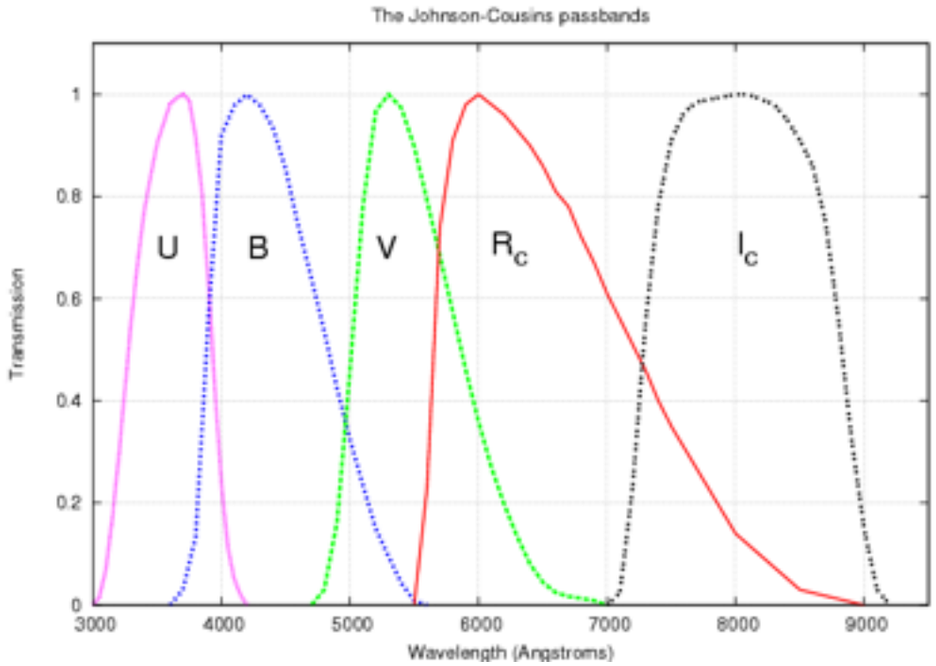
QSO energy spectrum



Blue

Red

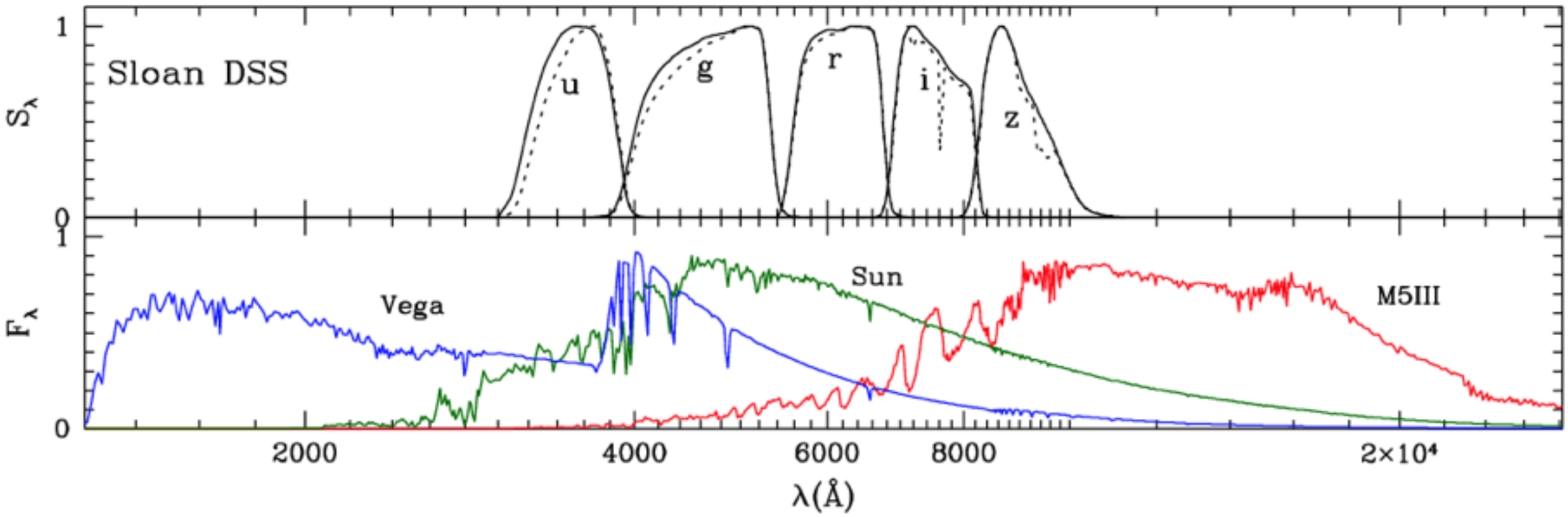
Photometric systems



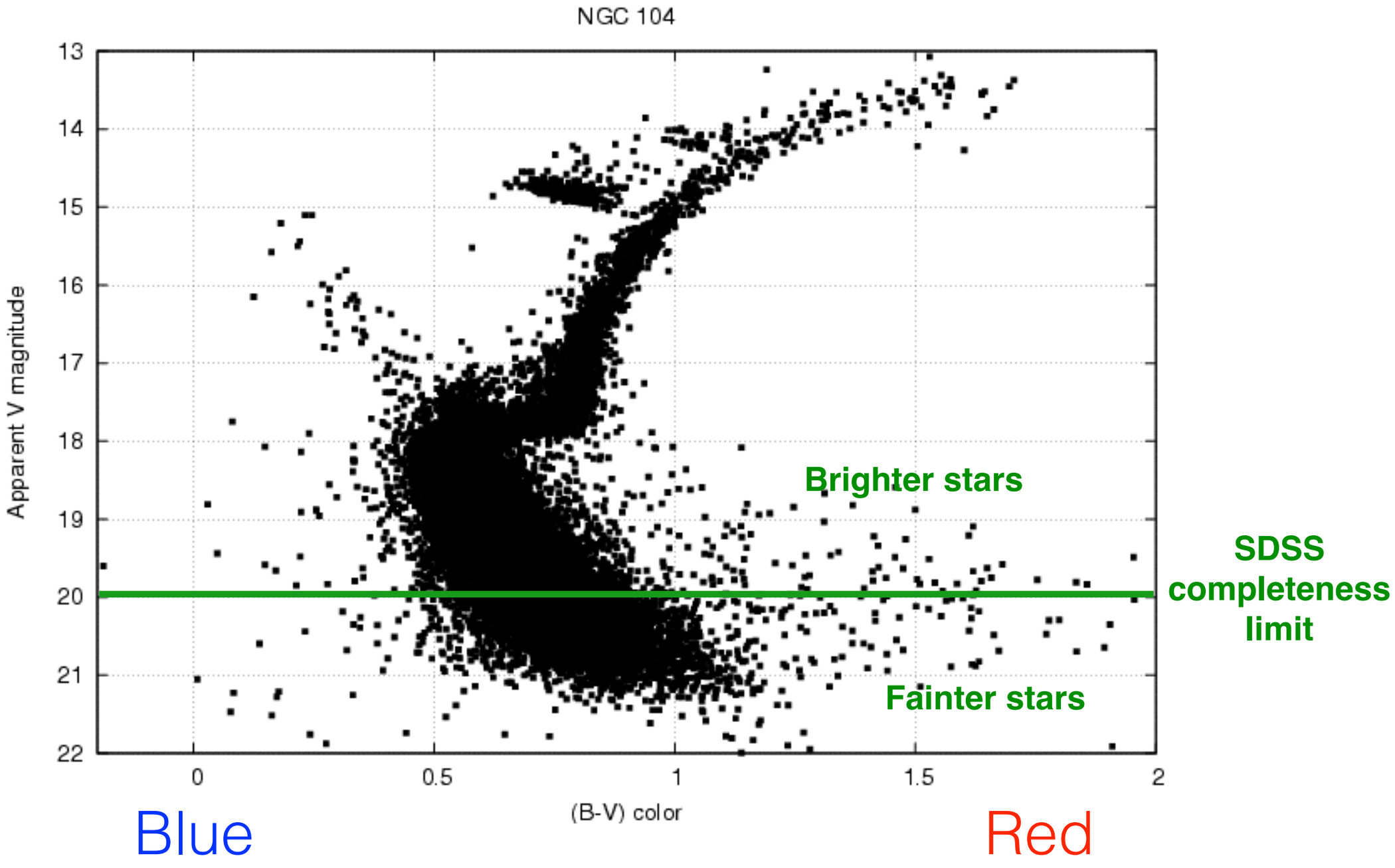
<u>Magnitude</u> in a filter
~ -Log (<u>Flux</u> in the same filter)

<u>Colour</u> = <u>Mag</u>(filter 1) - <u>Mag</u>(filter 2)

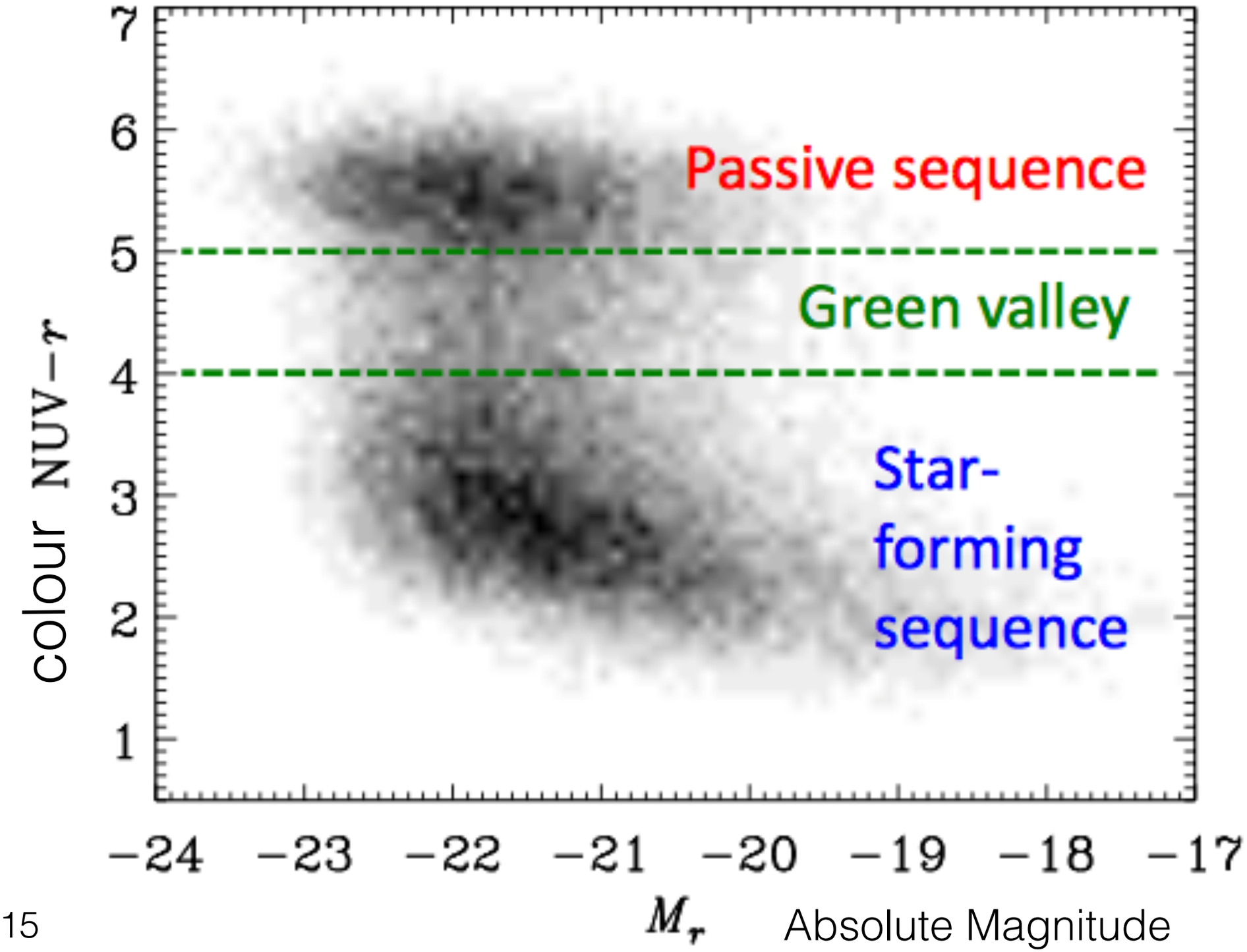
higher colour index: Redder object



Stellar colour-magnitude diagram

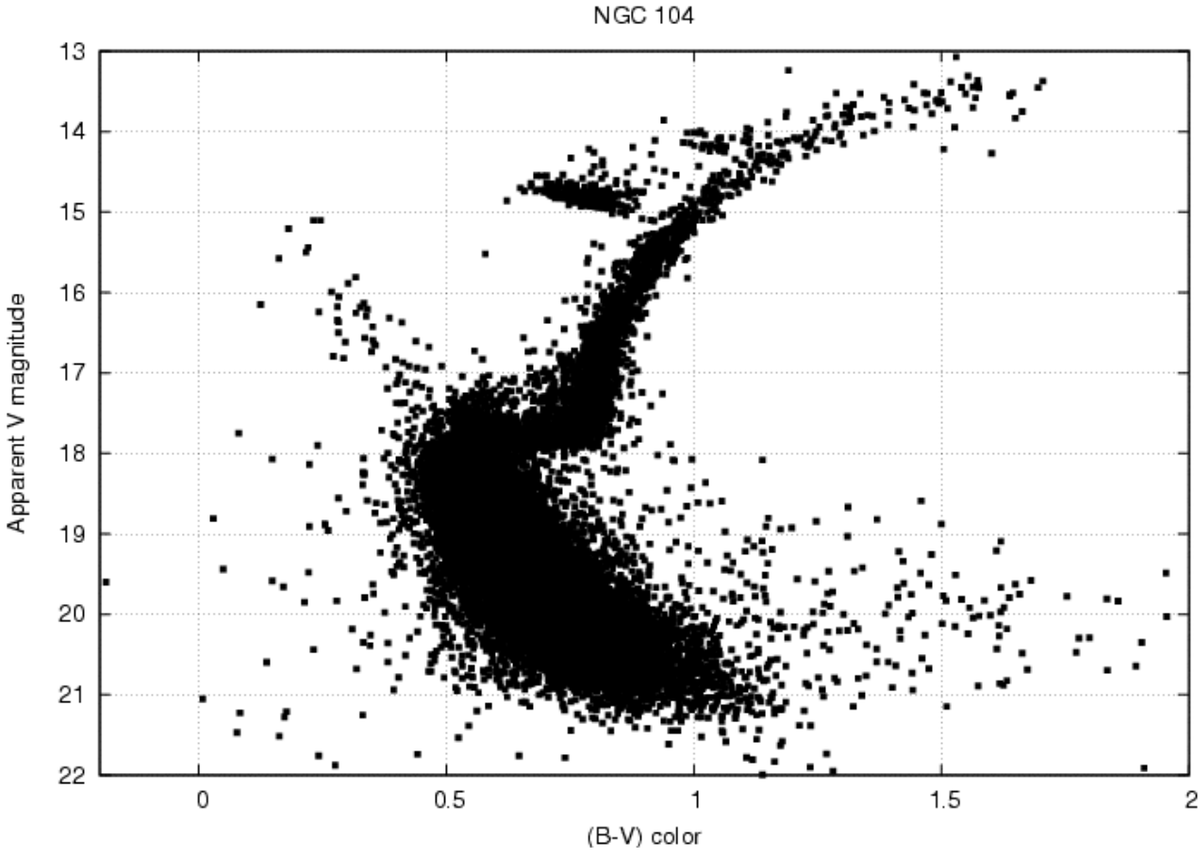


galactic colour-magnitude diagram

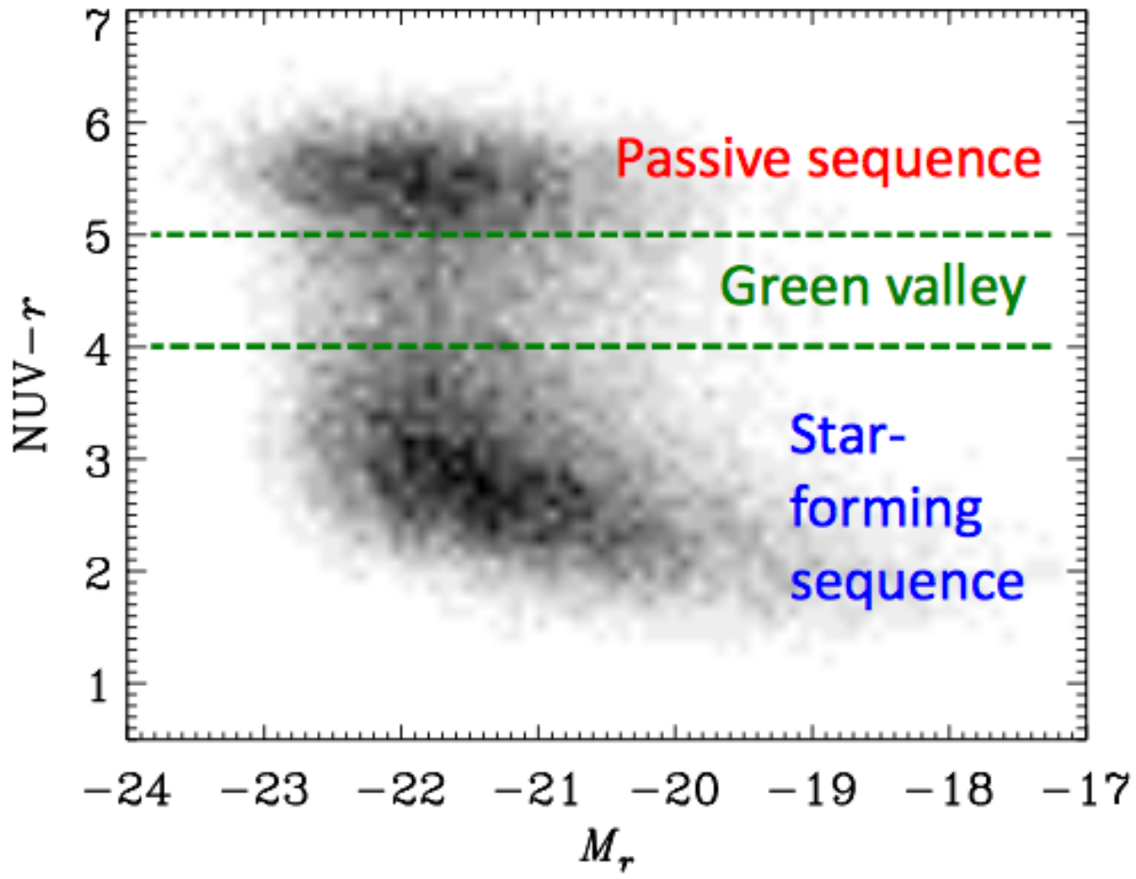


S. Salim 2015

Colour indices can be used to classify the celestial objects

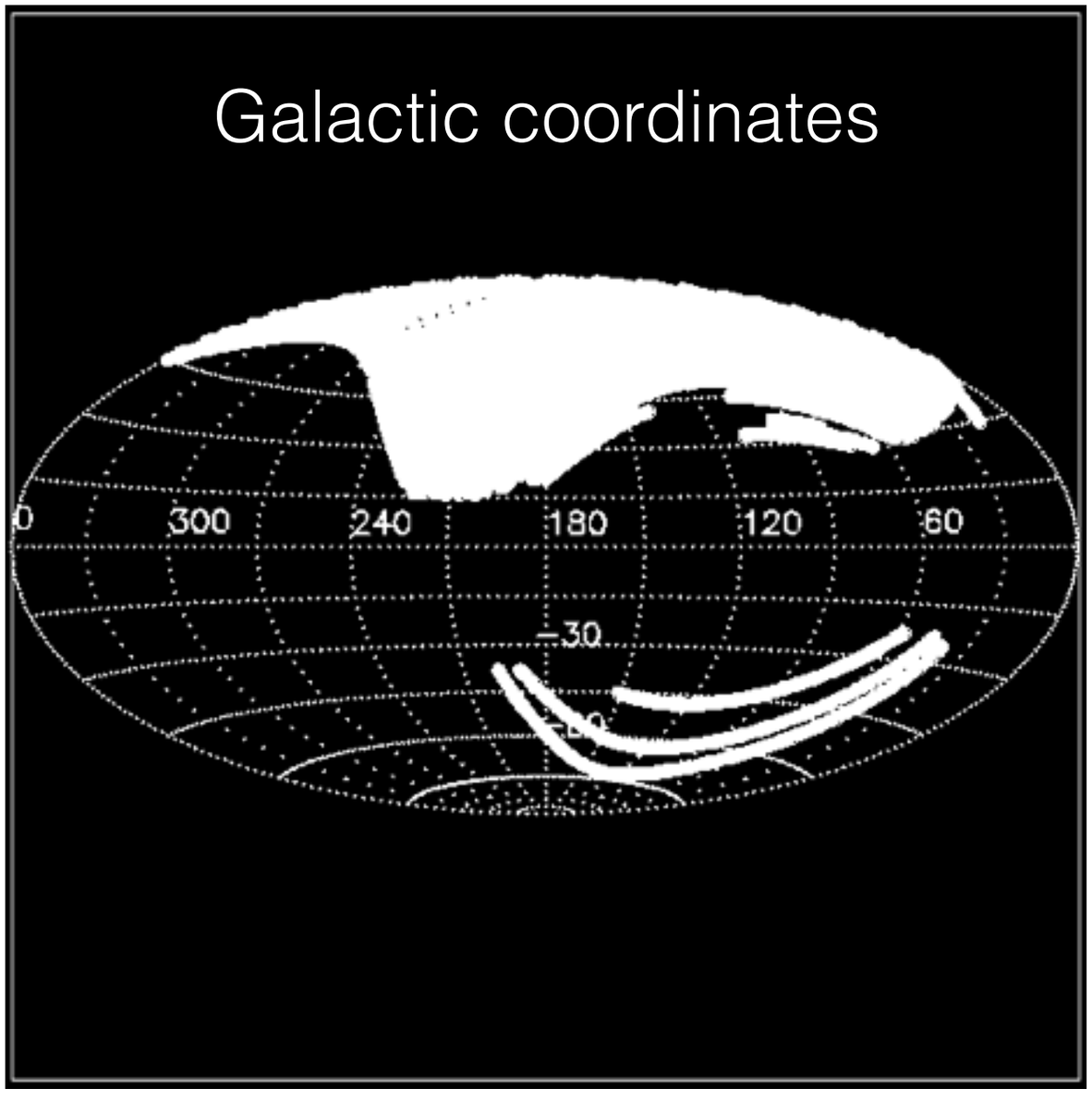


Stars

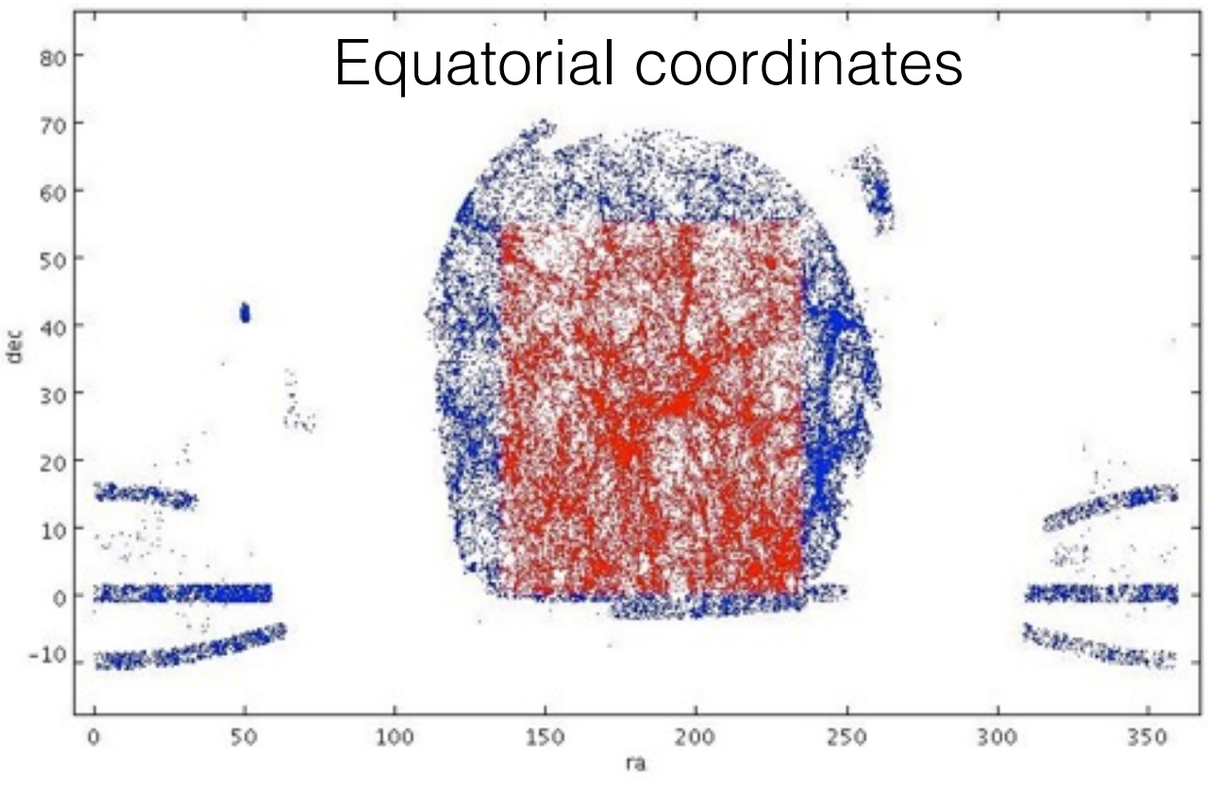


Galaxies

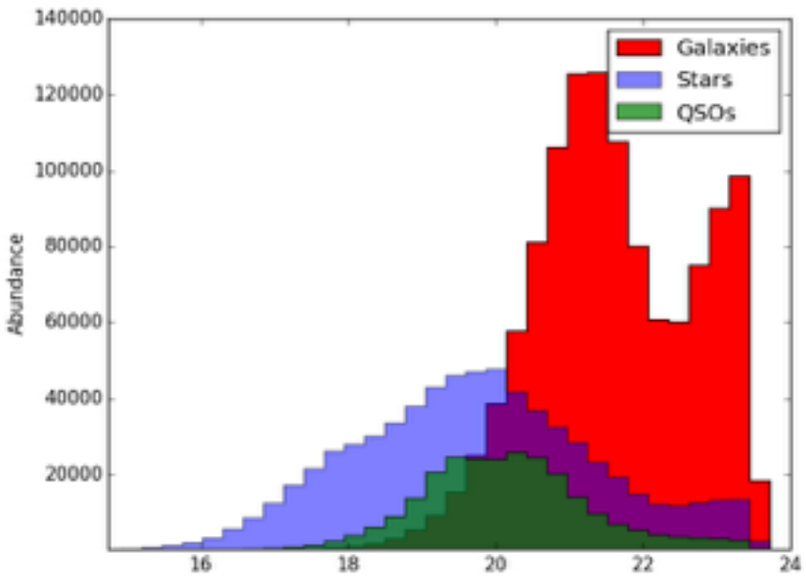
SDSS survey
(Sloan digital sky survey)

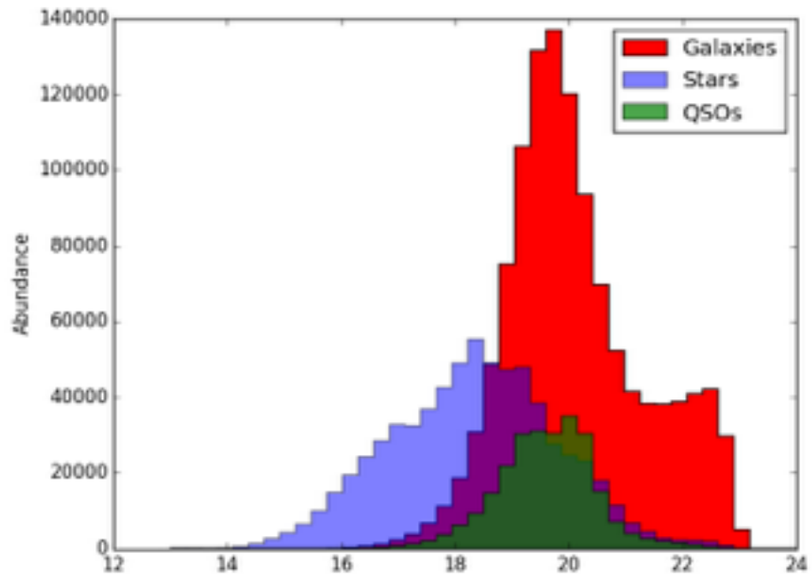


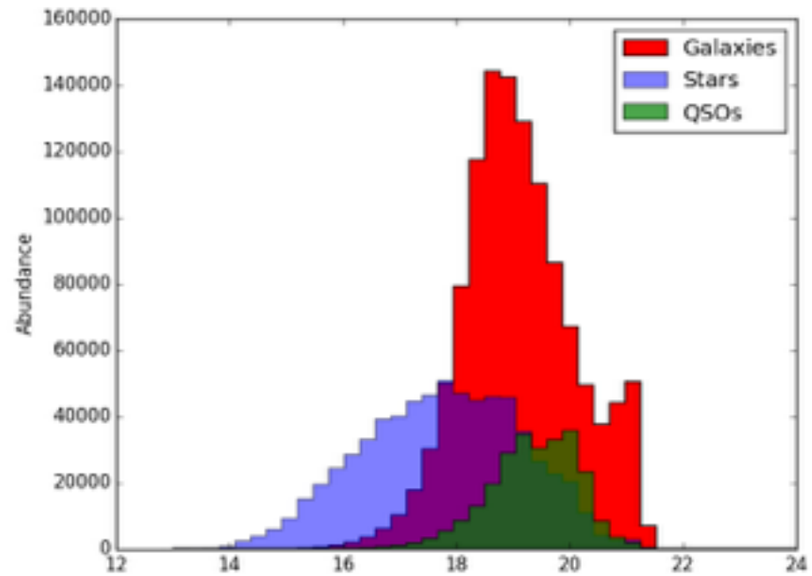
.2 m class telescope
.complete up to ~ 2.6 GLy
~ 4 million spectroscopically classified objects

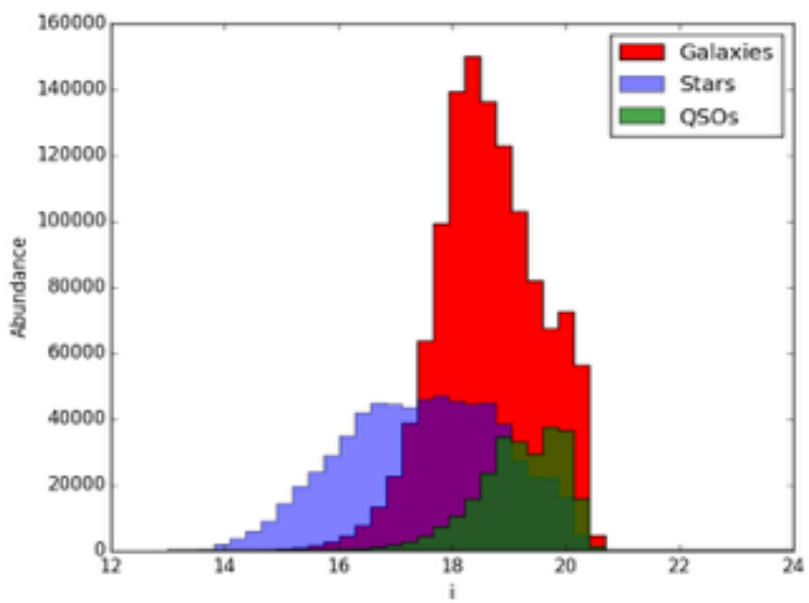


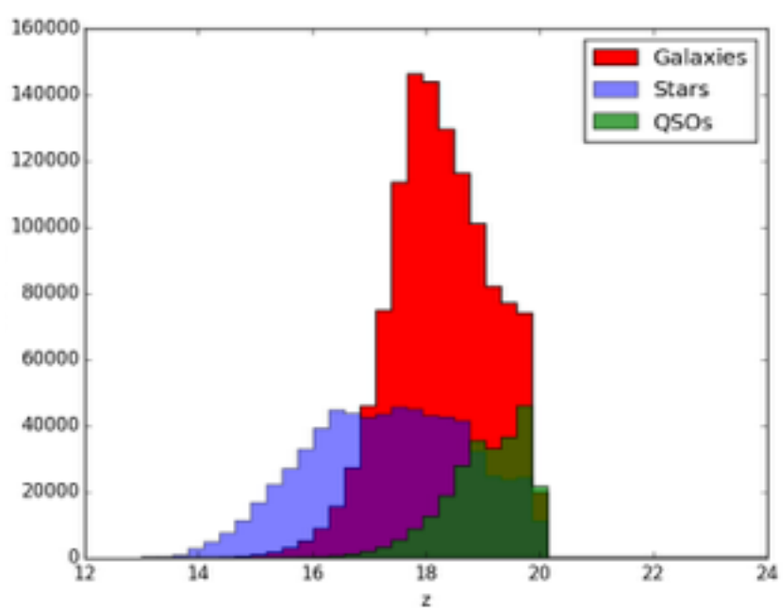
SDSS DR12 photo-spec sample
~ 2,100,000 objects (after data cleaning)



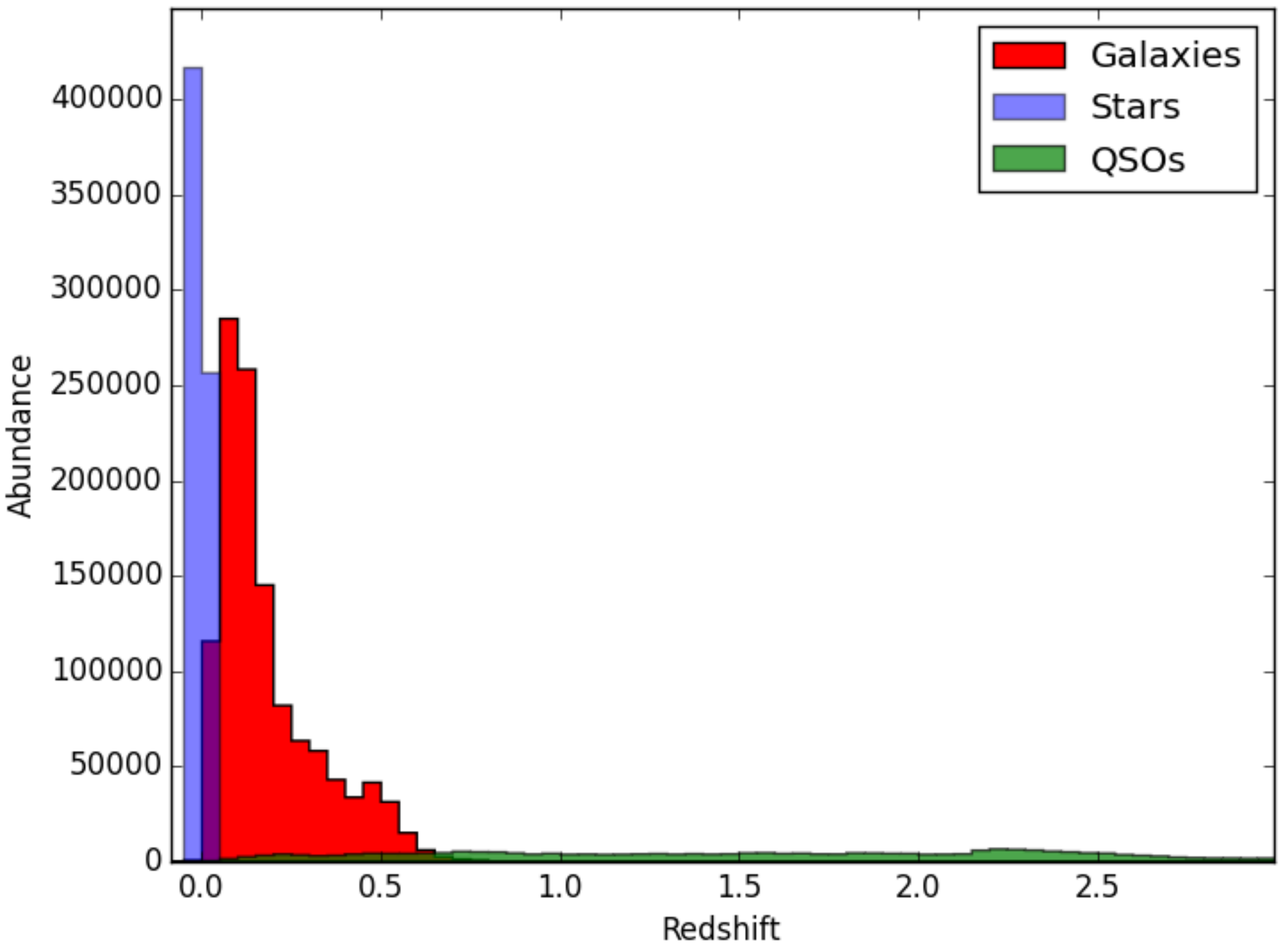




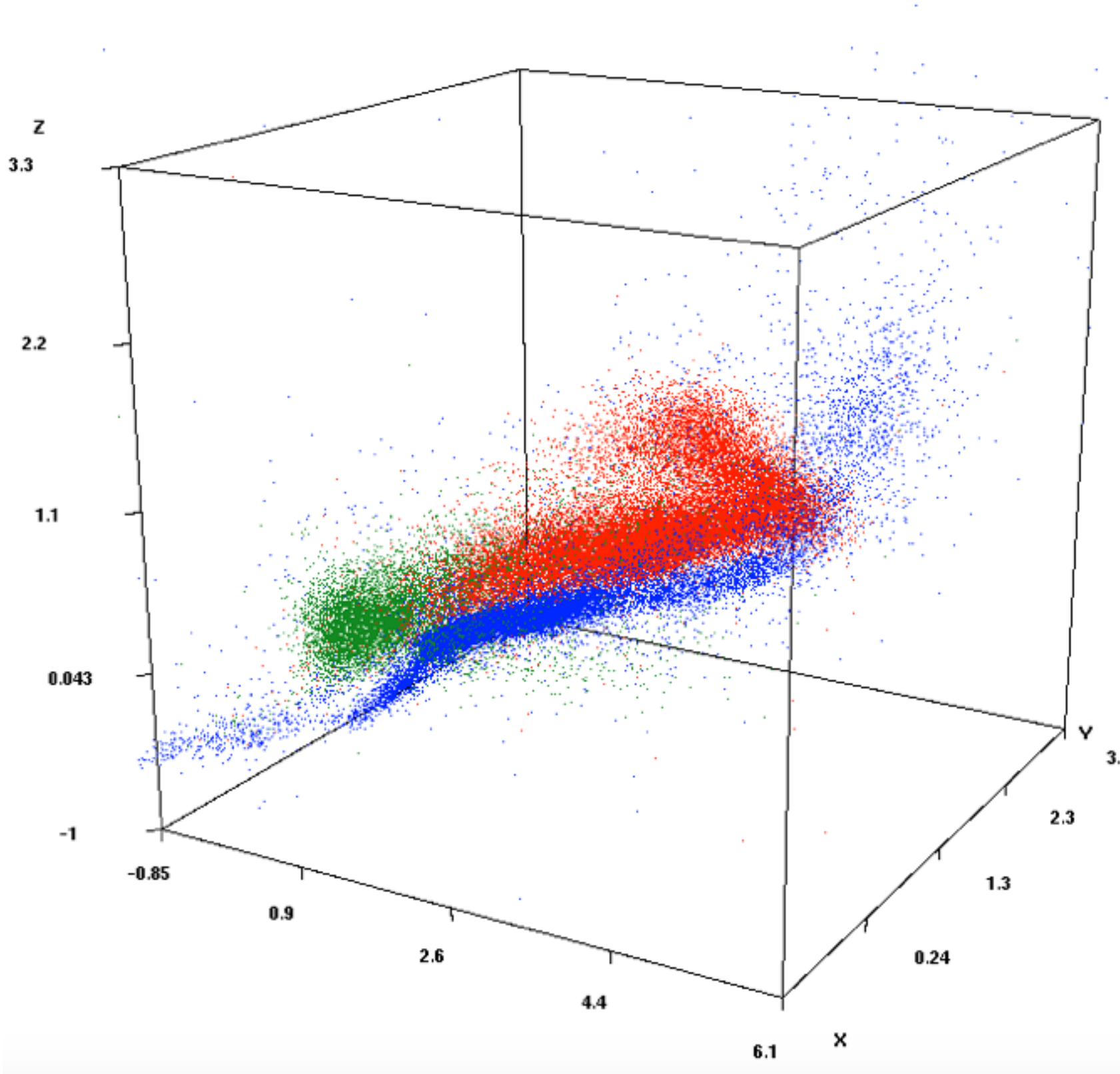




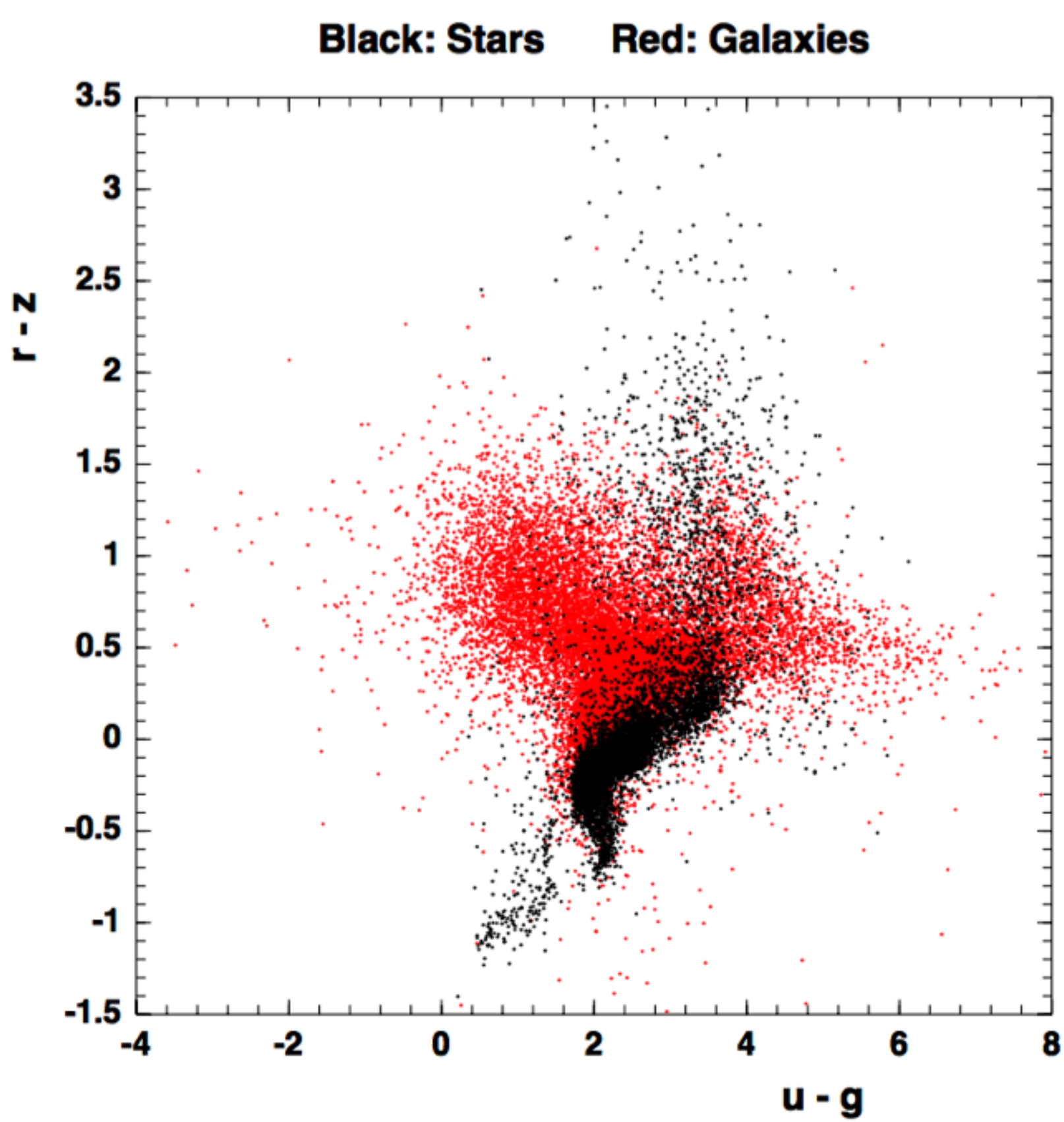
SDSS DR12 photo-spec sample
~ 2,100,000 objects (after data cleaning)



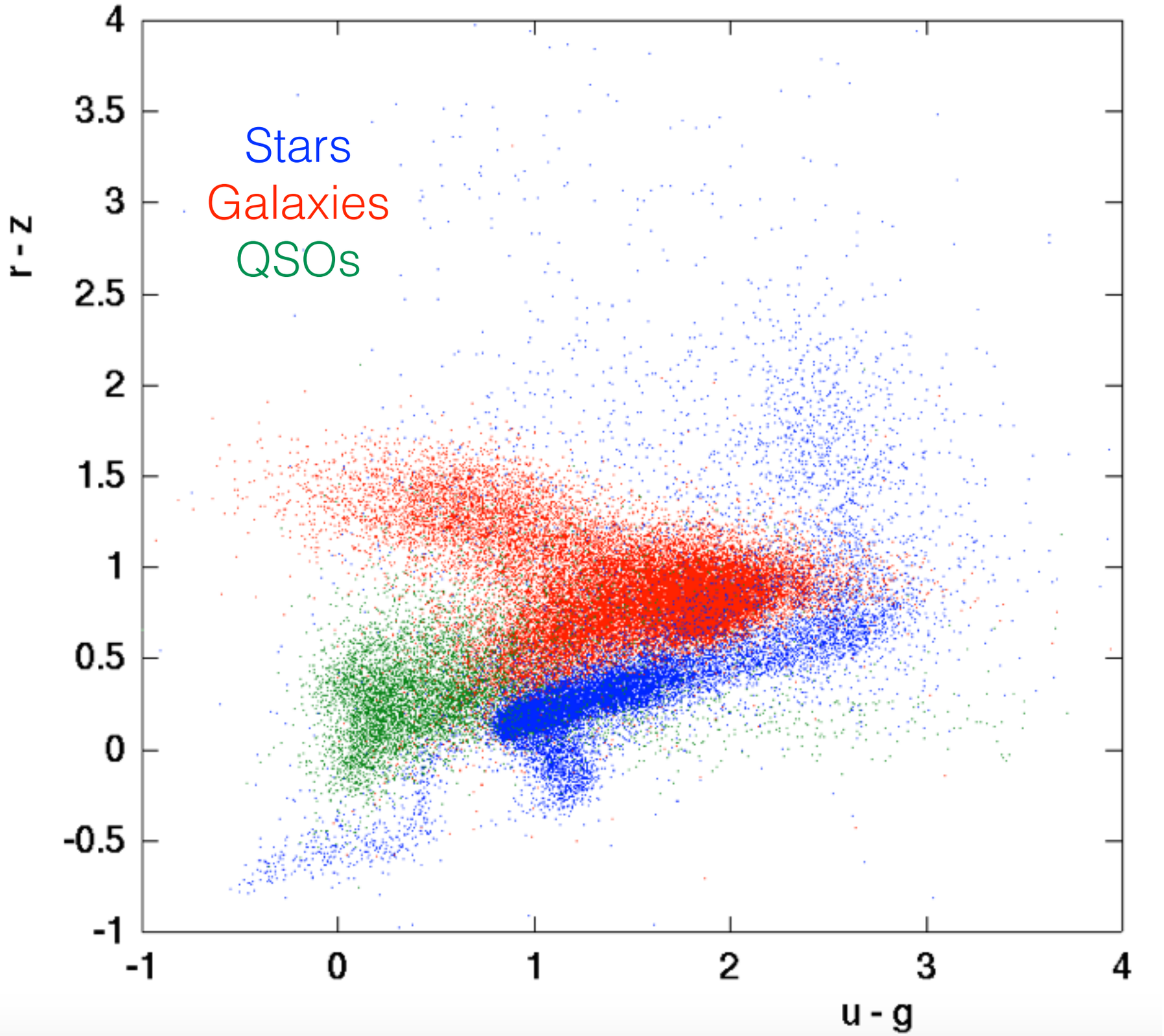
Colour indices as “features” for classification



Colour indices as “features” for classification

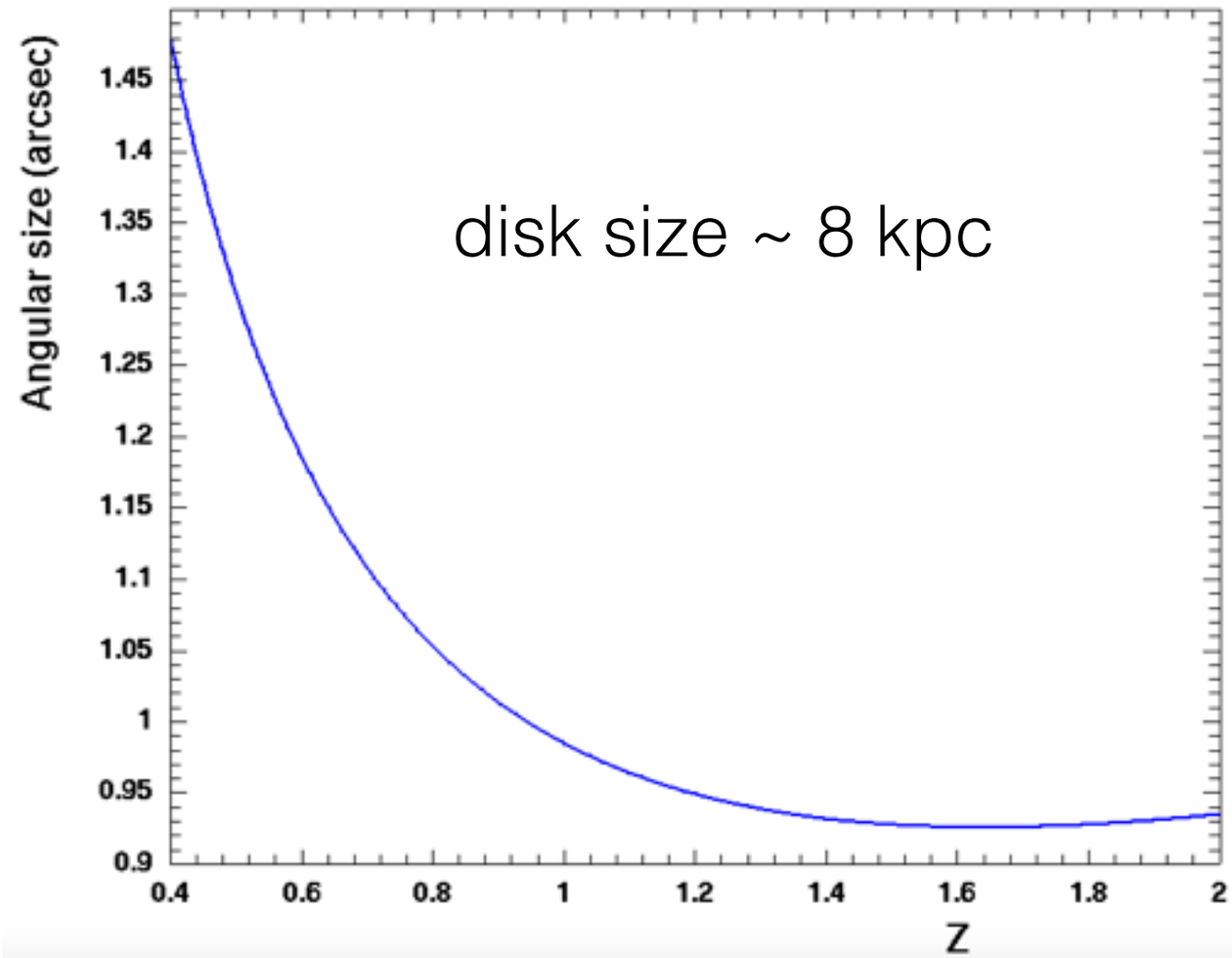


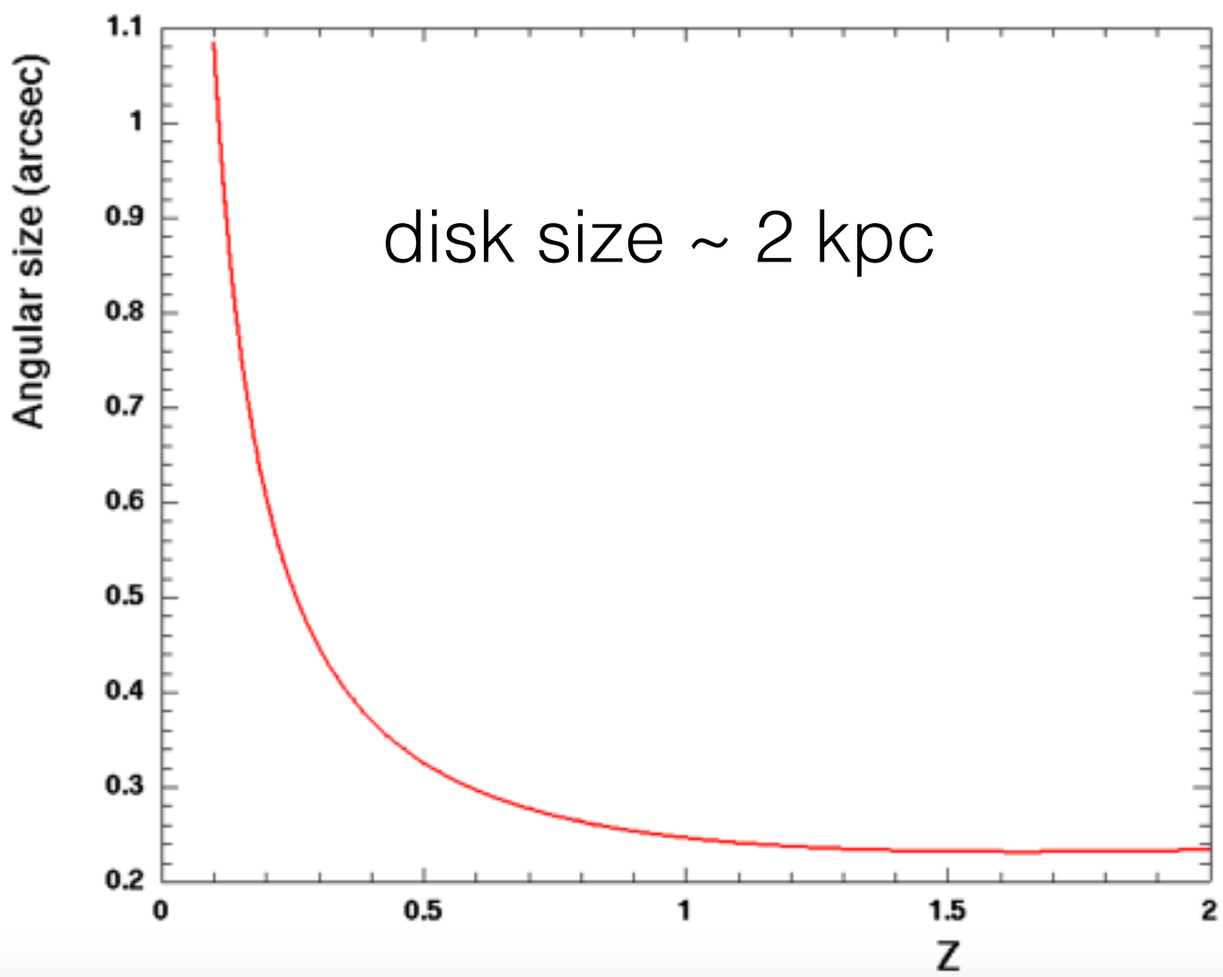
Colour indices as “features” for classification



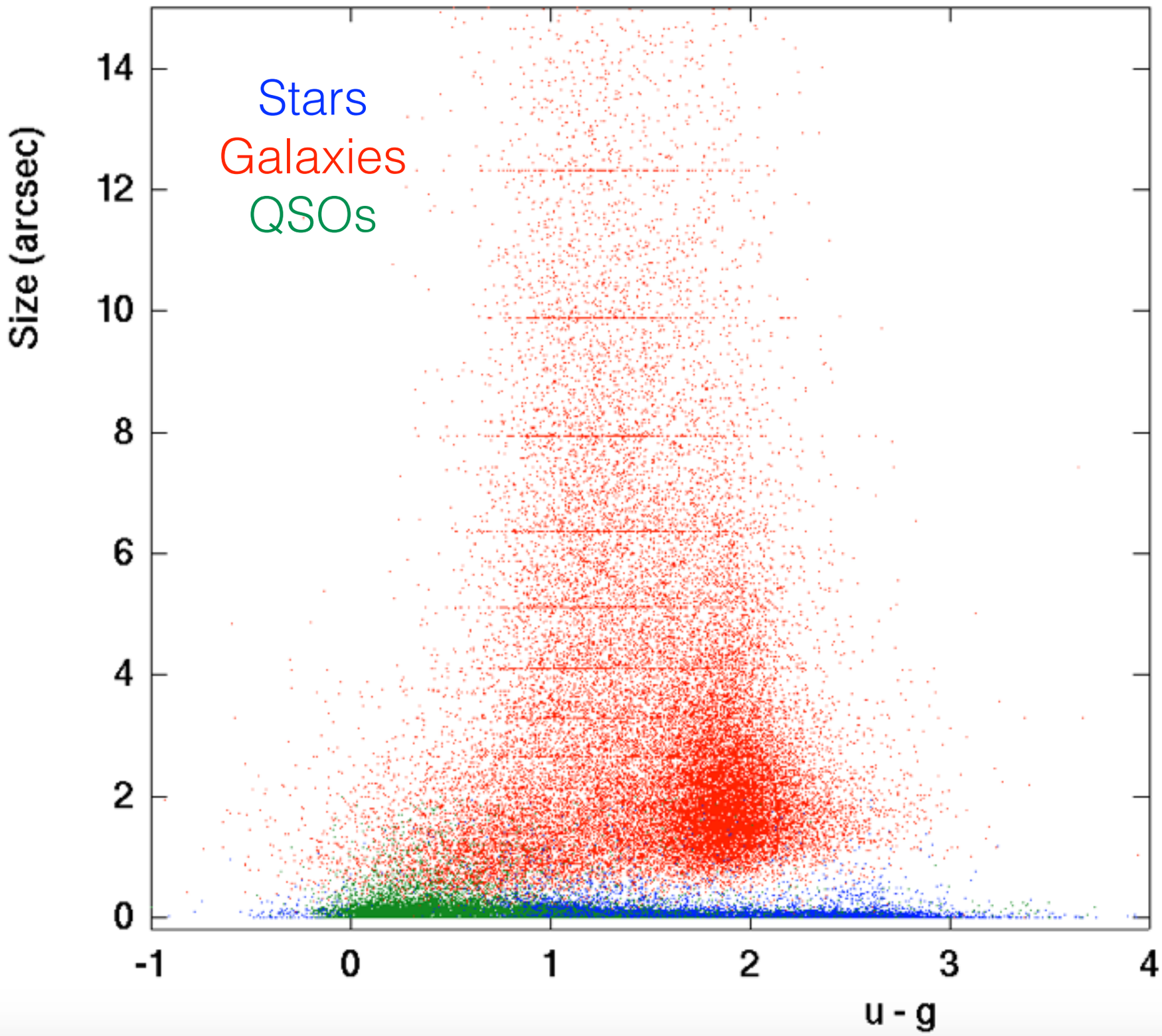
Object's size as "feature" for classification

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

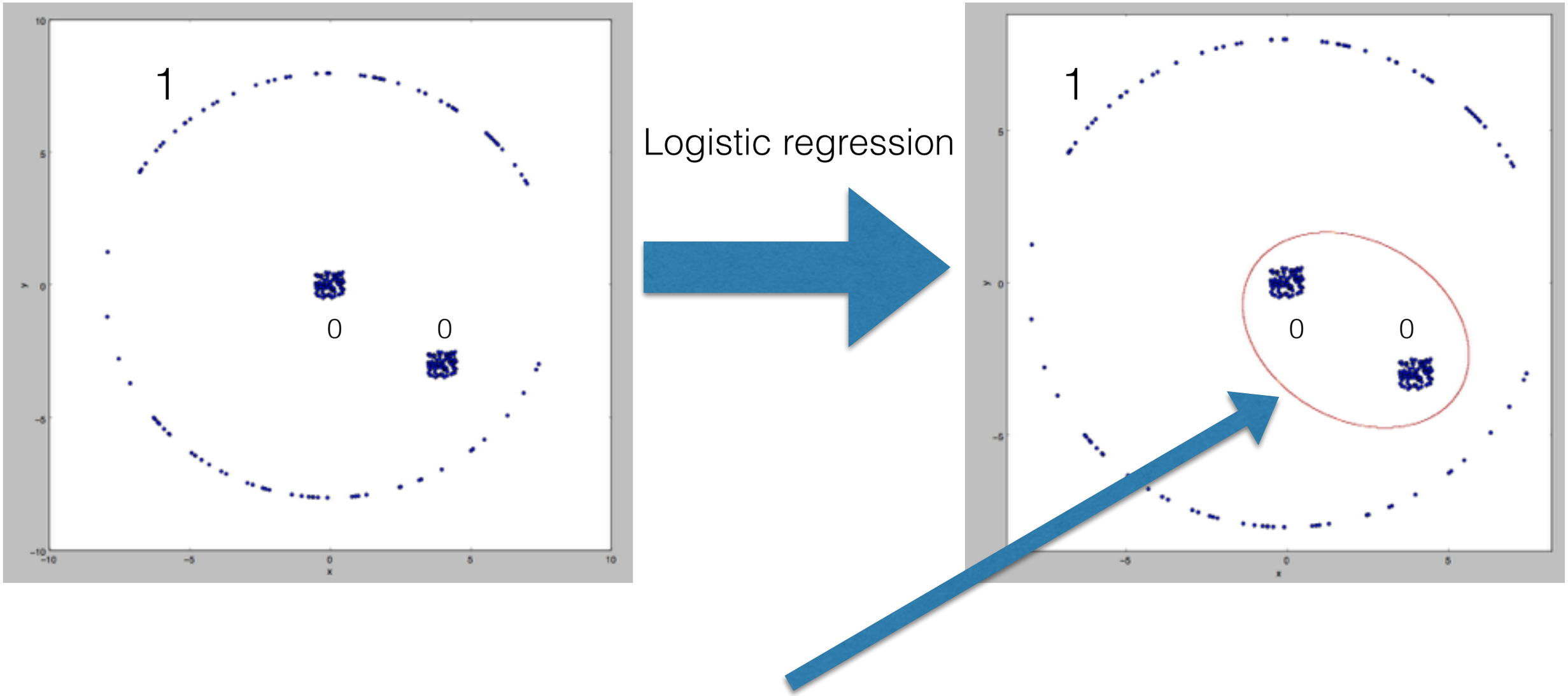




Object's size as "feature" for classification



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

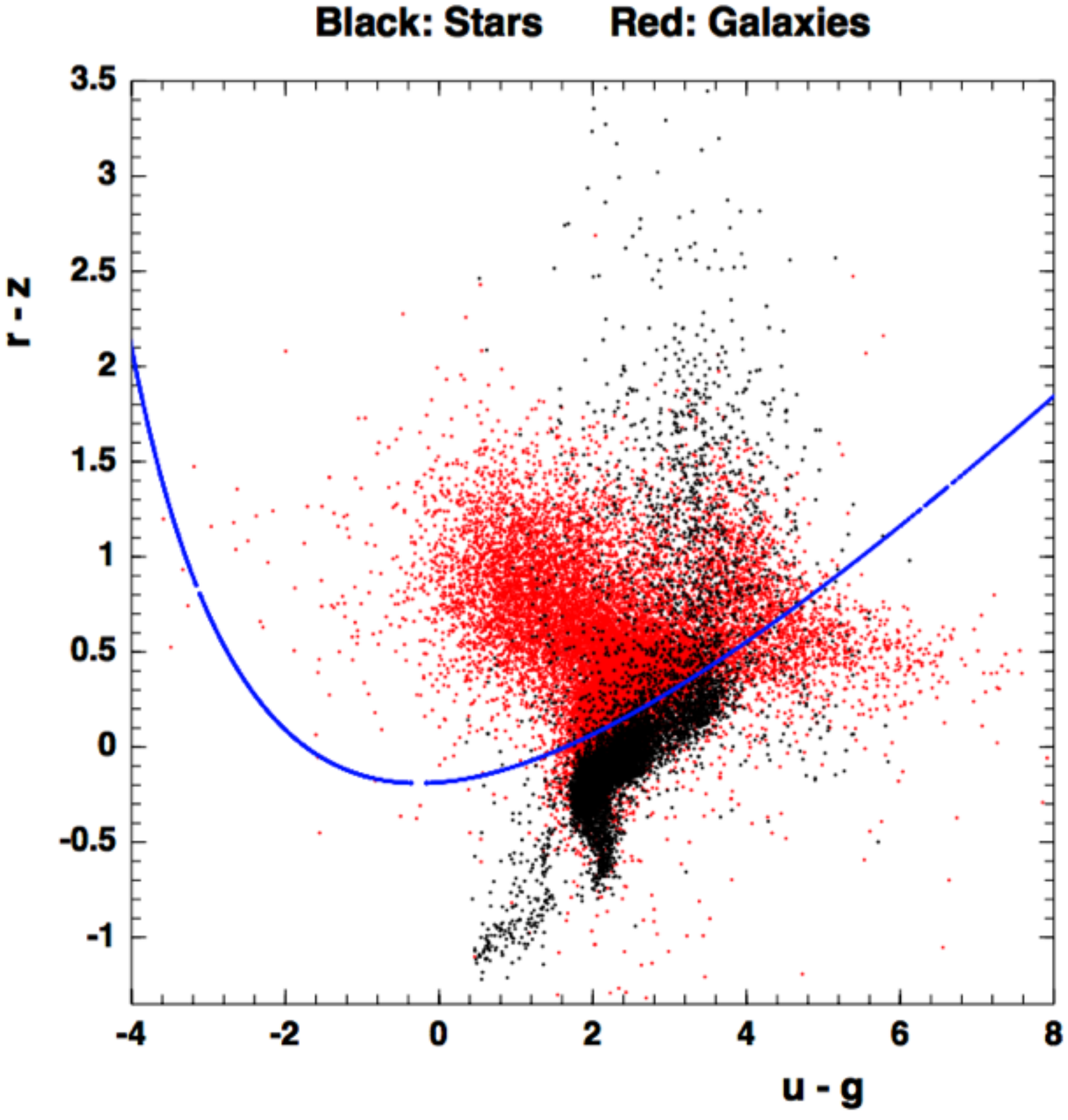
Logistic regression

- 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.

Results from the L.R. fit



Results from the classification

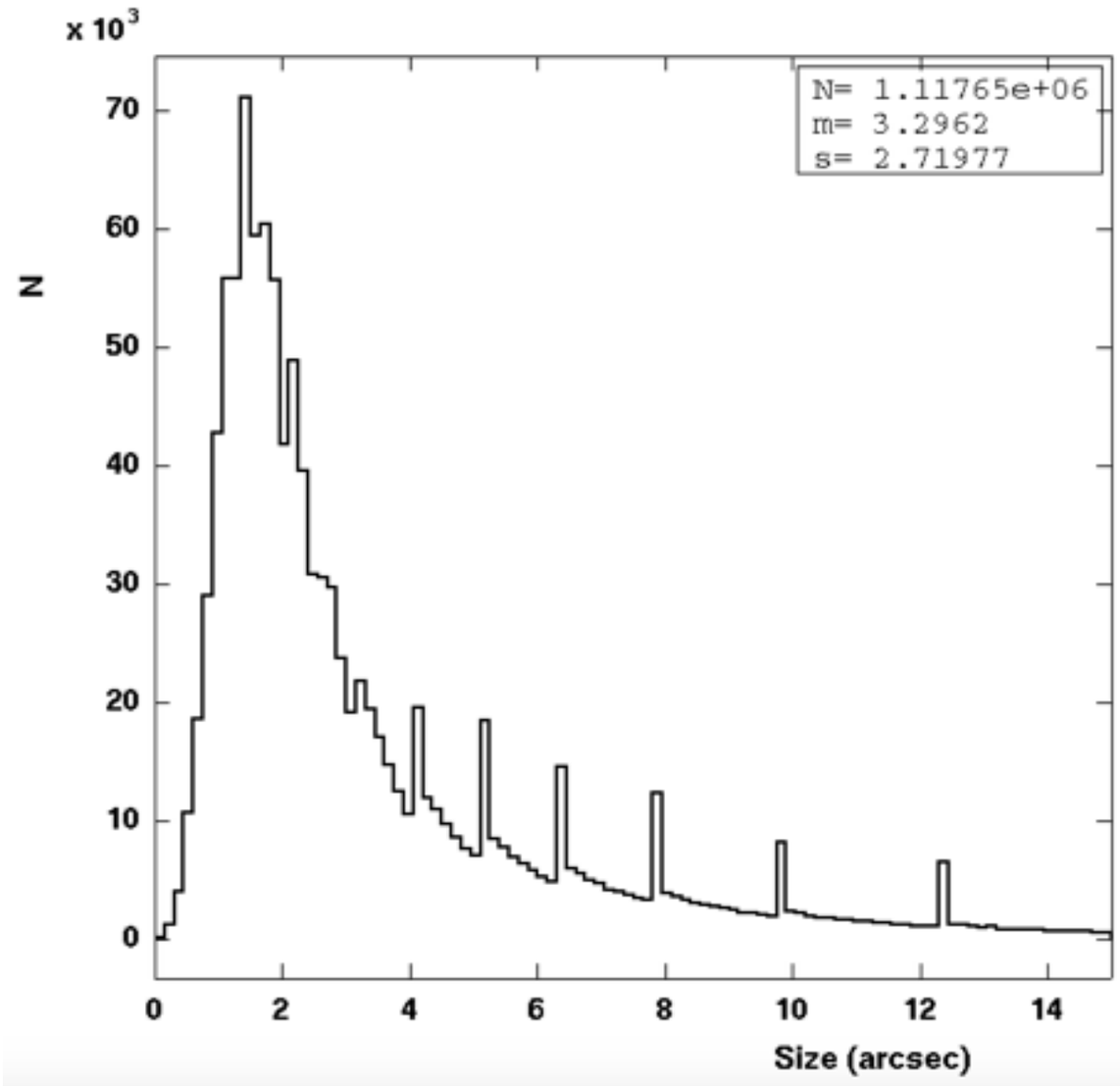
- Classification efficiency for the
  whole sample: 94%
  galaxies: 96%
  stars: 92%
  QSOs: 88%

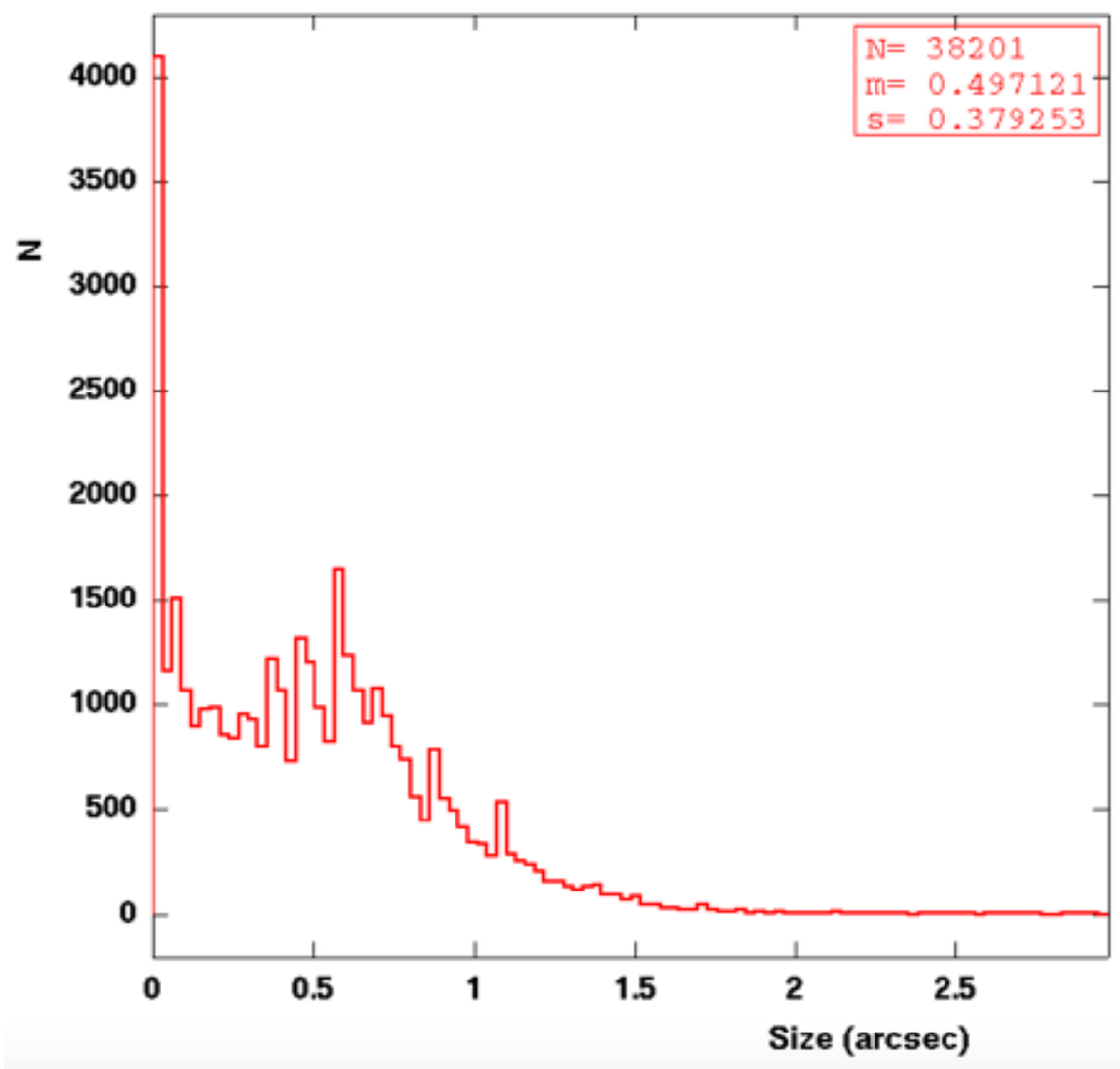
- Mean size of the galaxies classified
  wrongly: 0.5 arcsec
  correctly: 3 arcsec

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

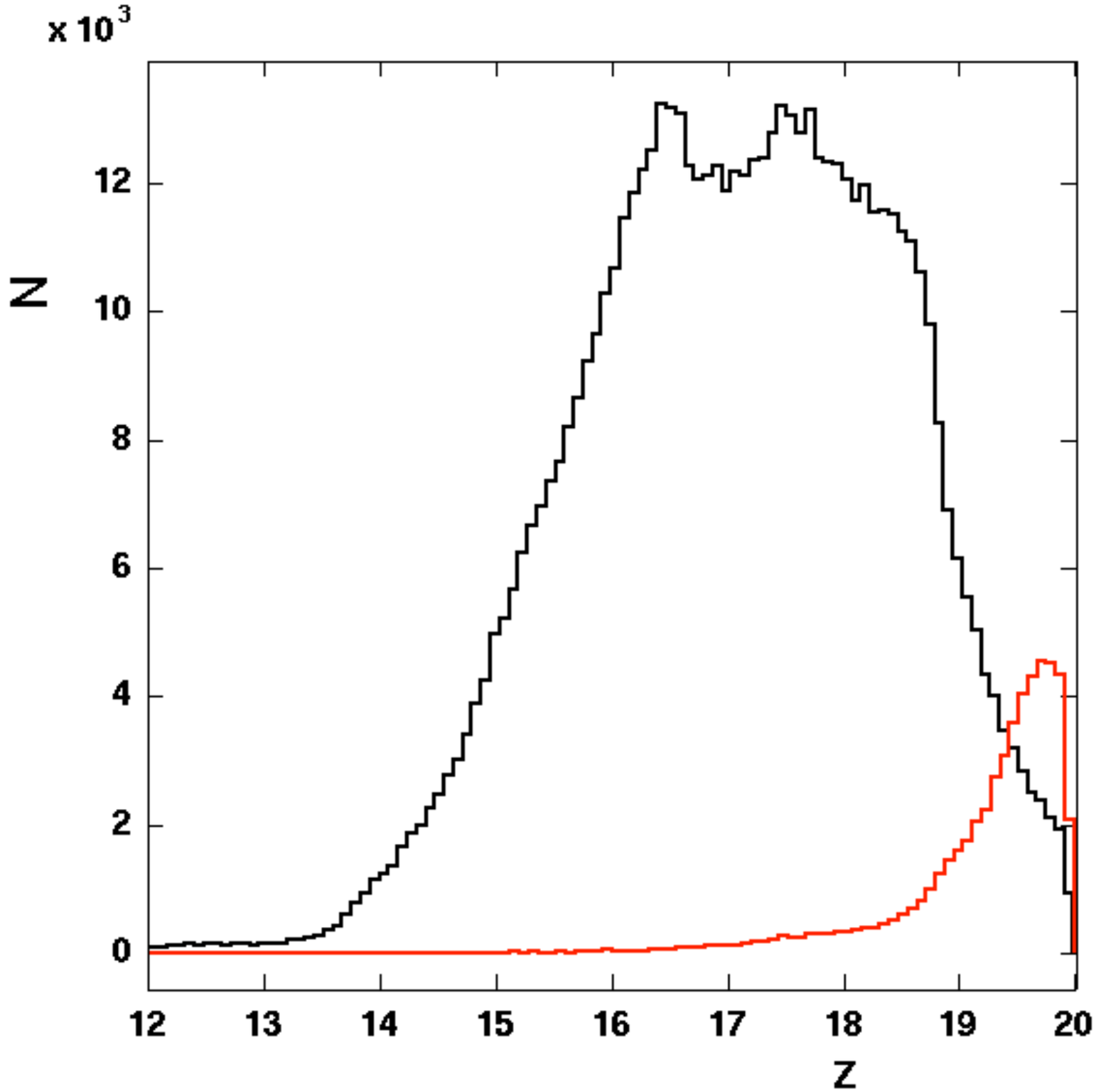
- Mean redshift of the QSOs classified
  wrongly: redshift = 2 (further QSOs)
  correctly: redshift = 1.5

Wrongly and **correctly** classified galaxies

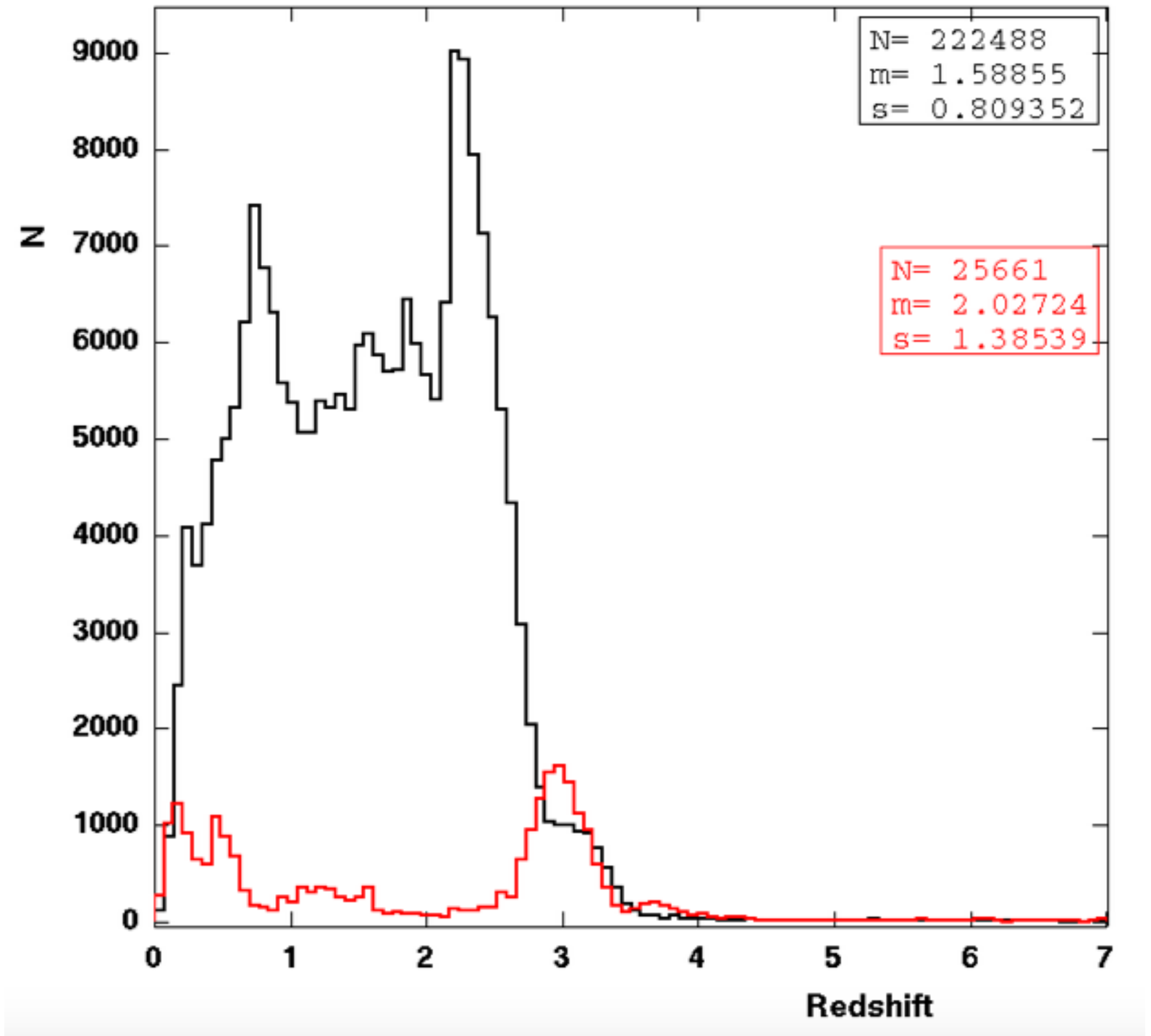




Wrongly and **correctly** classified stars



Wrongly and **correctly** classified QSOs



Ongoing work

Comparison with other classifiers

1. Random forest technique:
Whole sample efficiency ~ 95%
efficiency per object to be investigated
(special thanks to Mehdi Cherti)

A basic classifier works nicely so far!

2. Go deeper in finding the sources
of the misclassifications.

Conclusions & Perspectives

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

- 4% of galaxies (small angular size) can be mis-classified as point-like sources.

- 9% of (faint) stars can be mis-classified as galaxy-QSO.

- 12% 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 and cosmological parameters?