SDSS DR12: Object classification & analysis

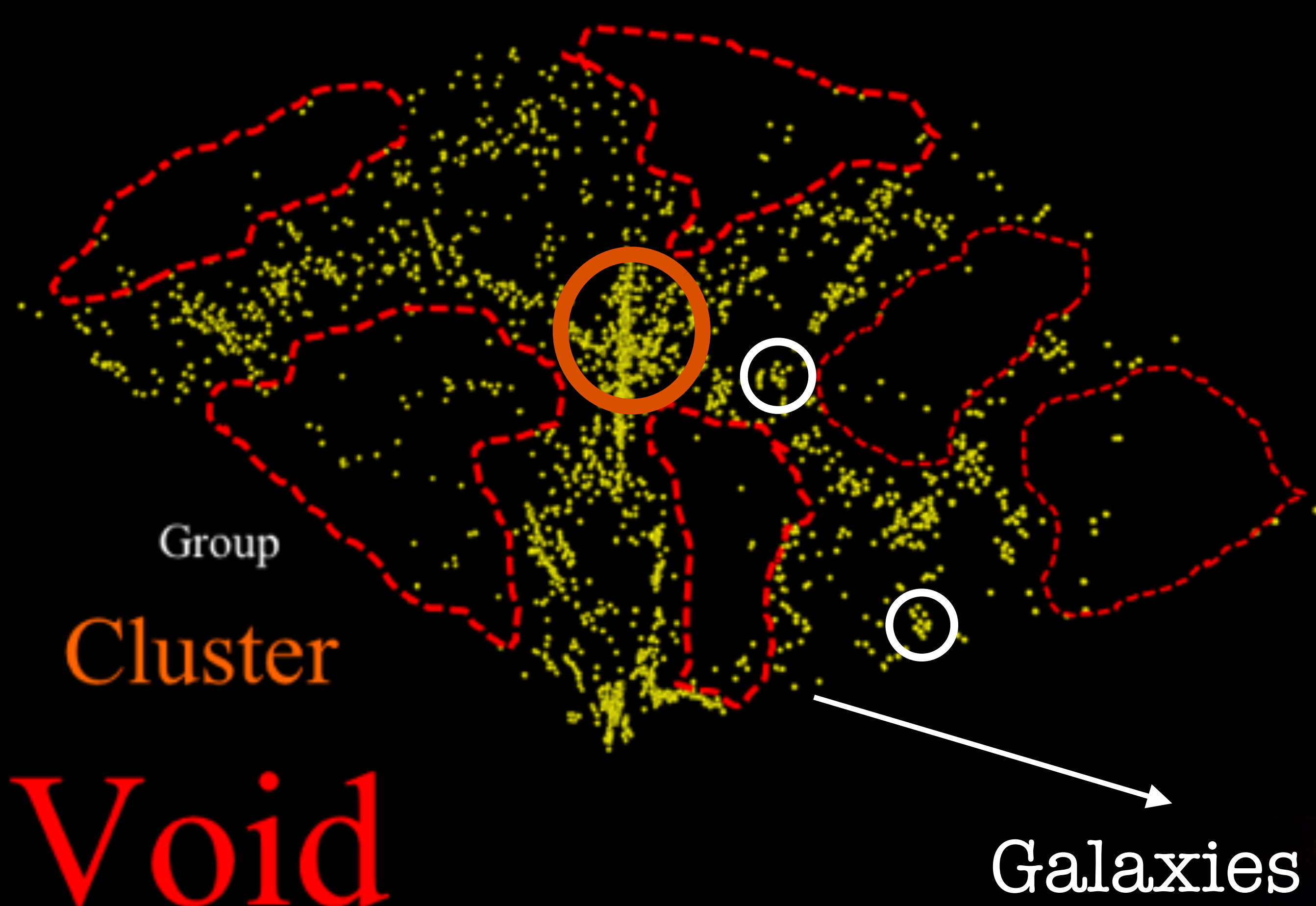
F. Habibi,
M. Moniez, R. Ansari & JE. Campagne

LSST-Webinaire
June 2017

Cosmological surveys

All sky surveys —> cosmic structures

Deep surveys —> structures formation & evolution



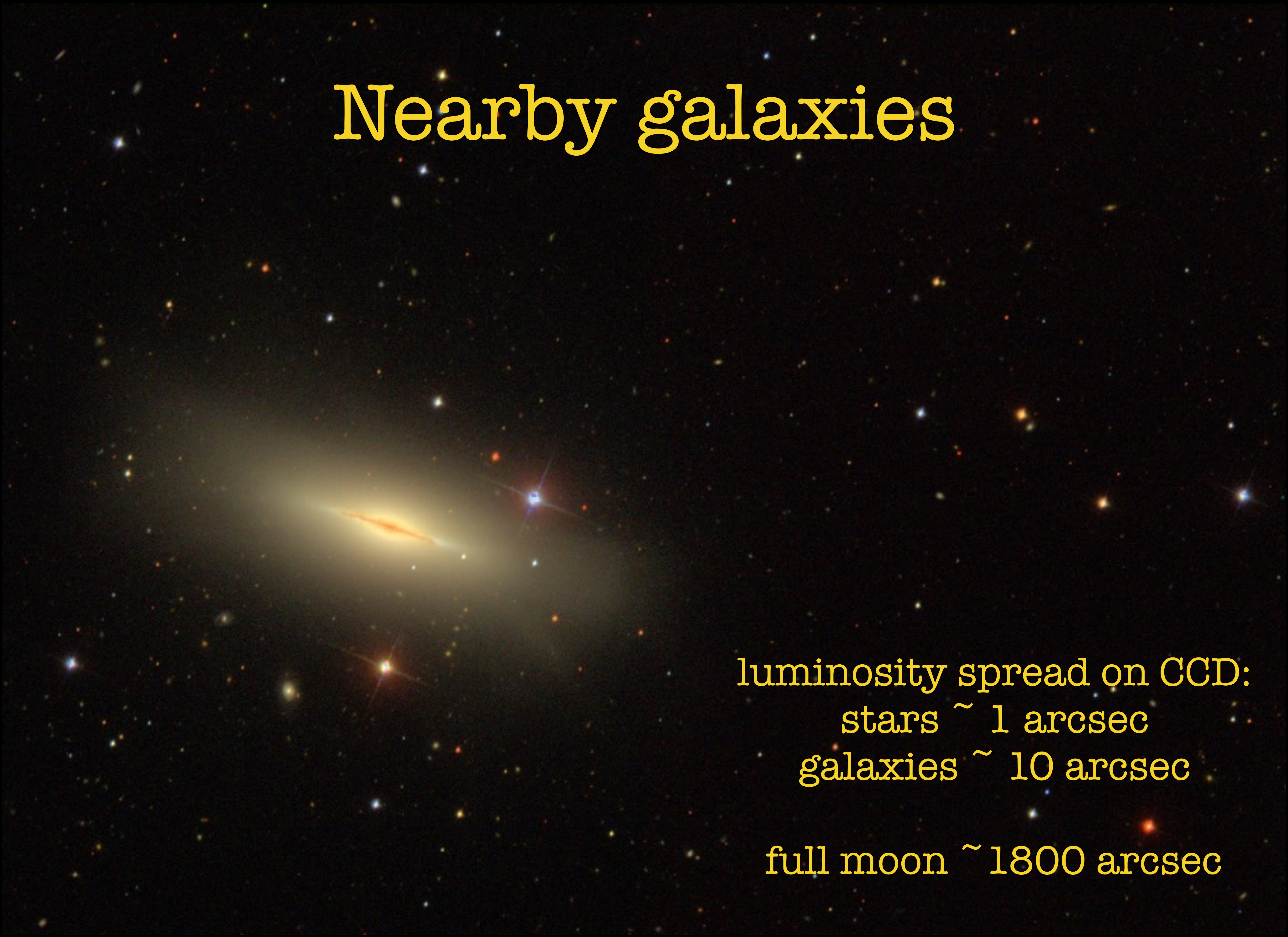
To know about the nature of Dark Matter & Dark Energy

Galaxies are units of cosmic structures

Object classification

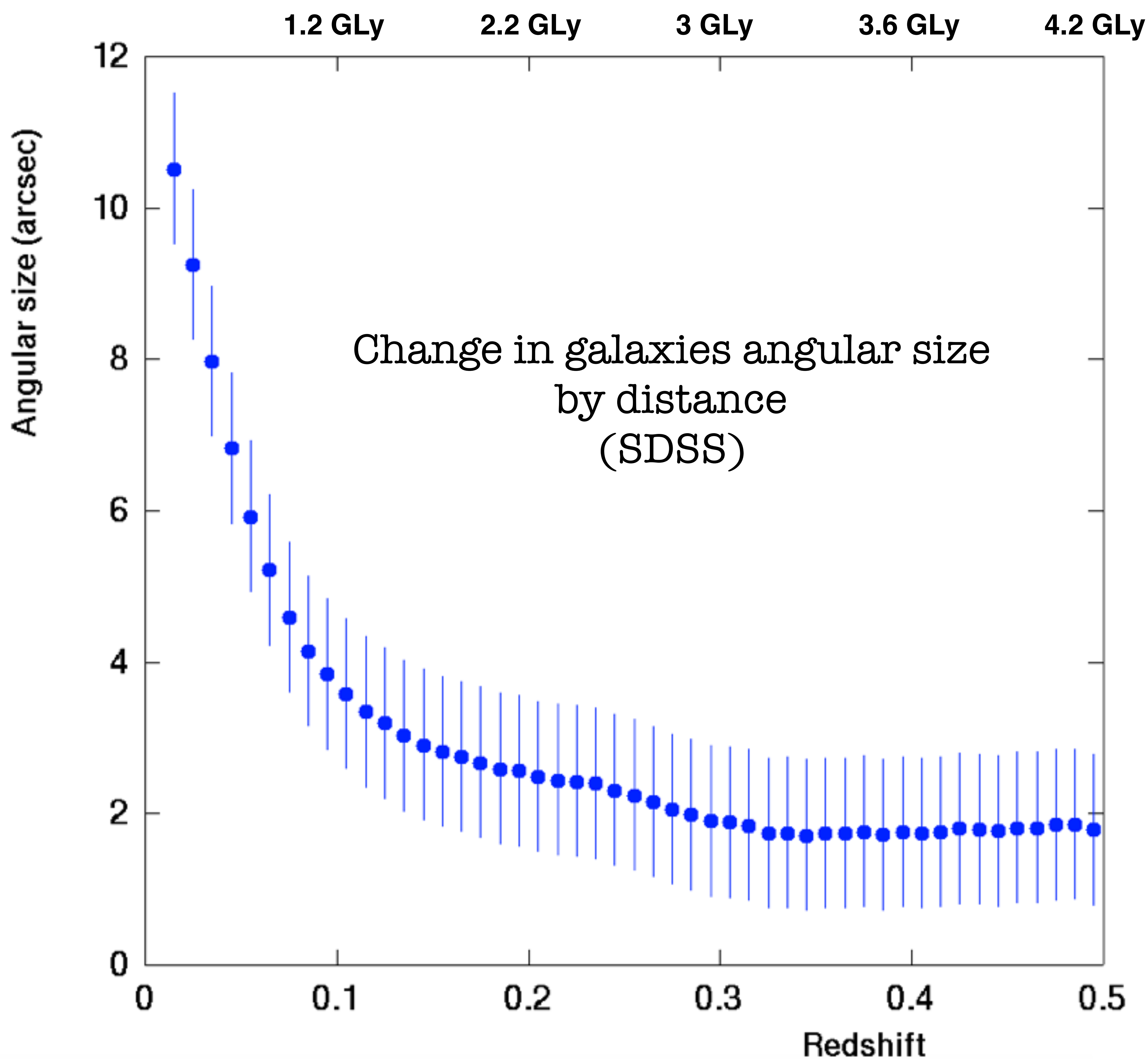
- **Cosmic structures contain galaxies.**
- **Images taken by surveys include QSOs and foreground stars in addition to galaxies.**
- **How to separate point-like sources from galaxies?**

Nearby galaxies



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

full moon ~1800 arcsec

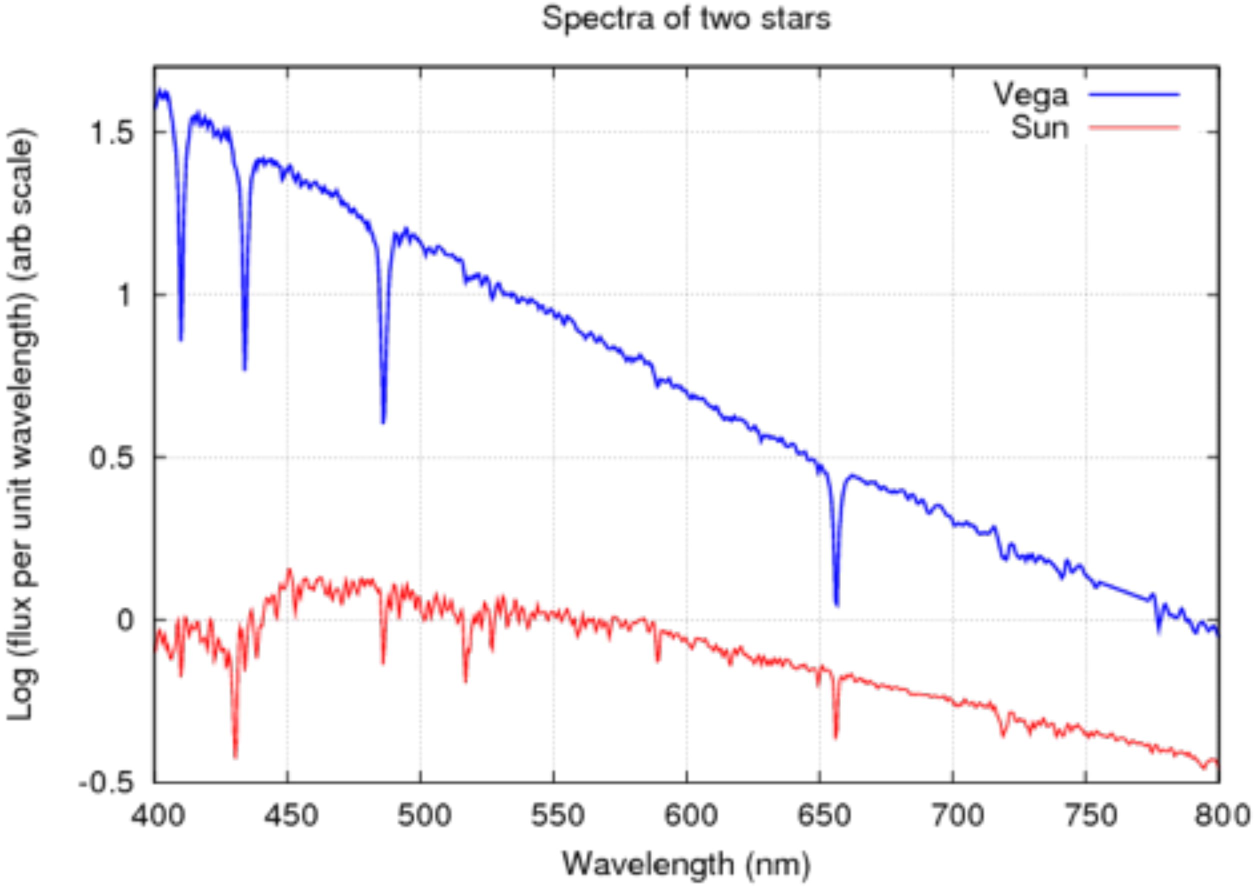

1.2 GLy
2.2 GLy
3 GLy
3.6 GLy
4.2 GLy
Change in galaxies angular size
by distance
(SDSS)
Angular size (arcsec)
Redshift
0
2
4
6
8
10
12
0
0.1
0.2
0.3
0.4
0.5

Far/faint galaxies

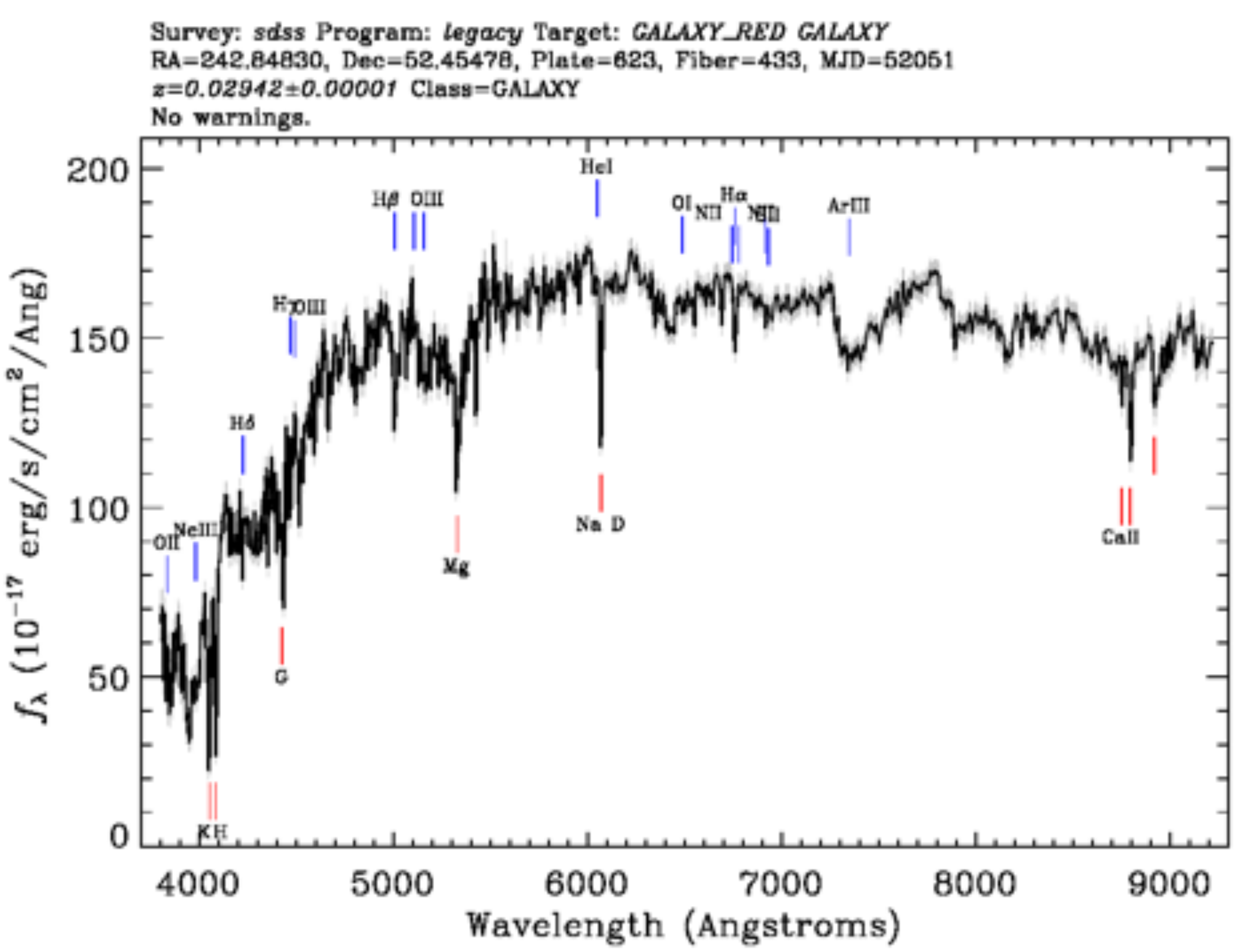
luminosity spread on CCD:

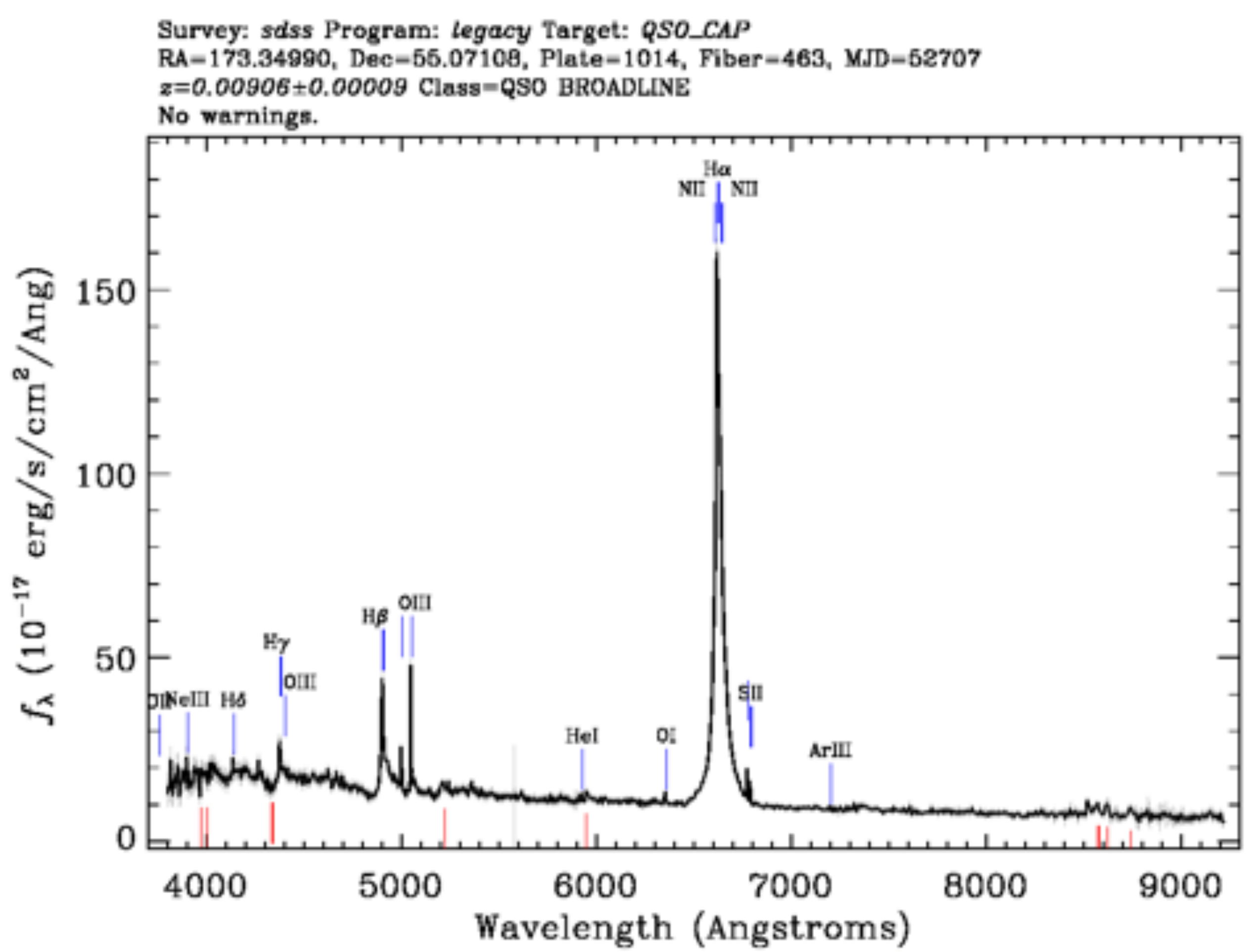
stars ~ 1 arcsec

galaxies ~ 1 arcsec



SEDs can separate the three objects





Aim

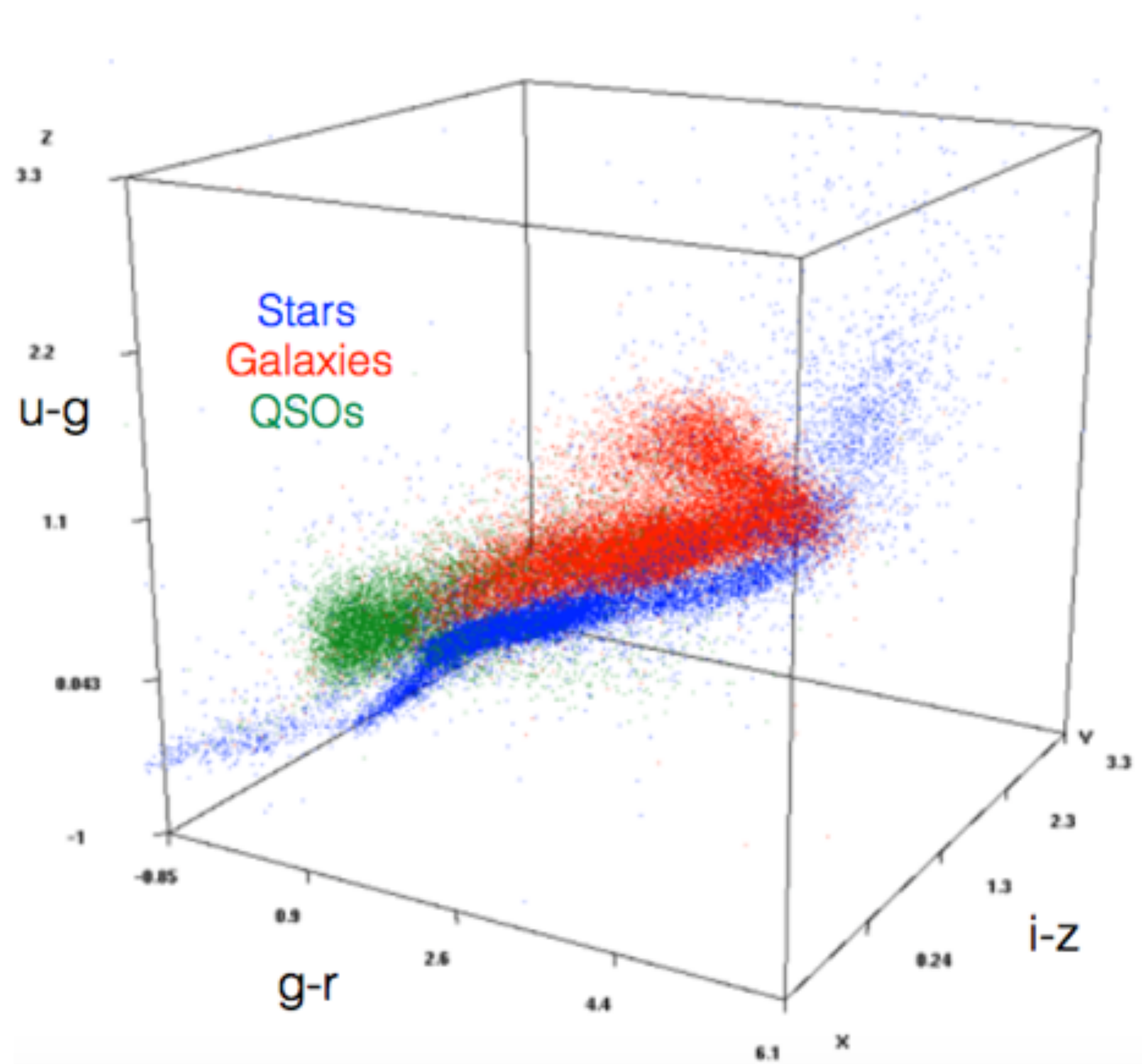
To separate galaxies from stars and QSOs, in the lack of spectroscopic data.

How?

Including all possible photometric information

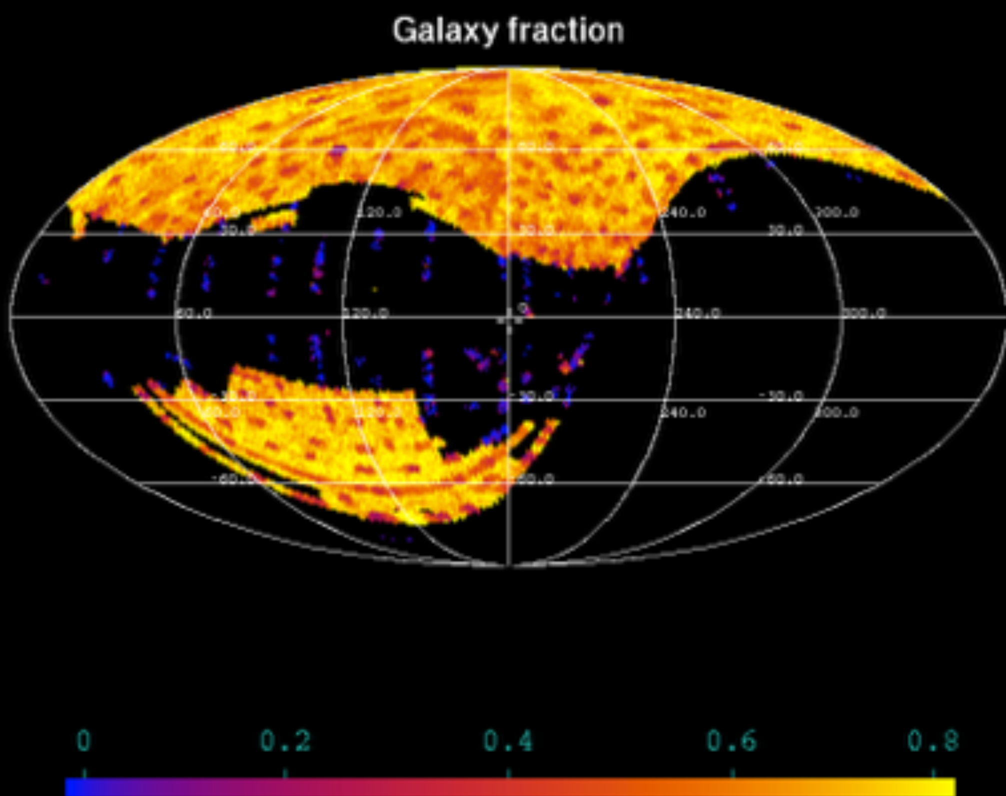
Colour indices:

proper “features” for supervised perceptrons



Using automatic classification for big number of data

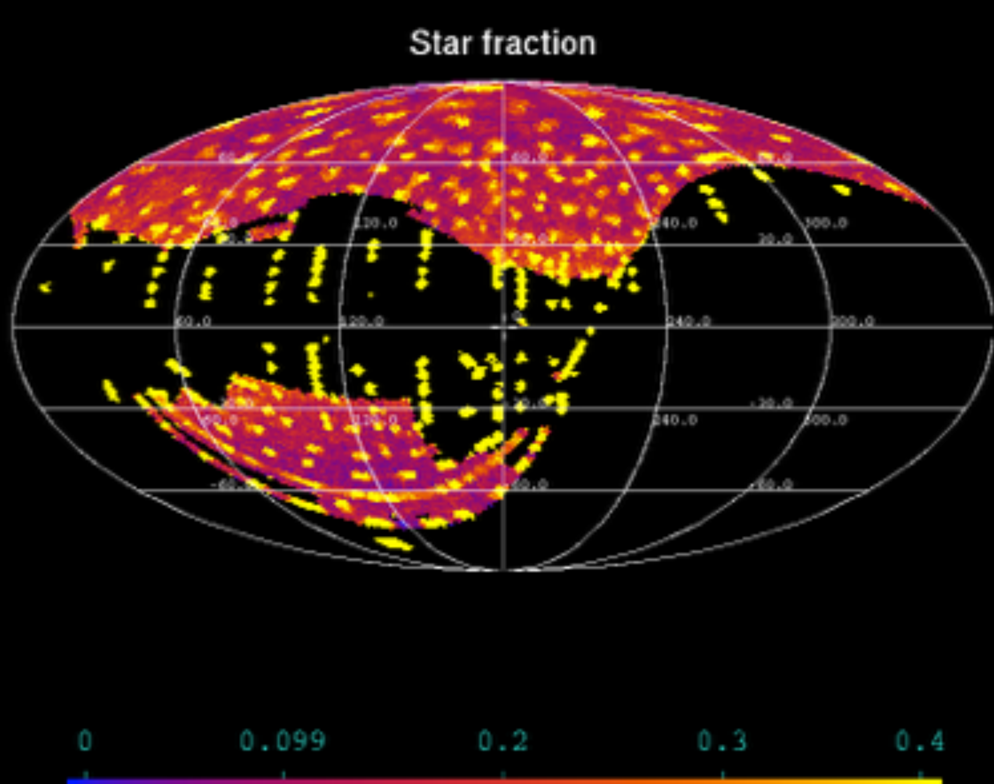
SDSS DR12 data

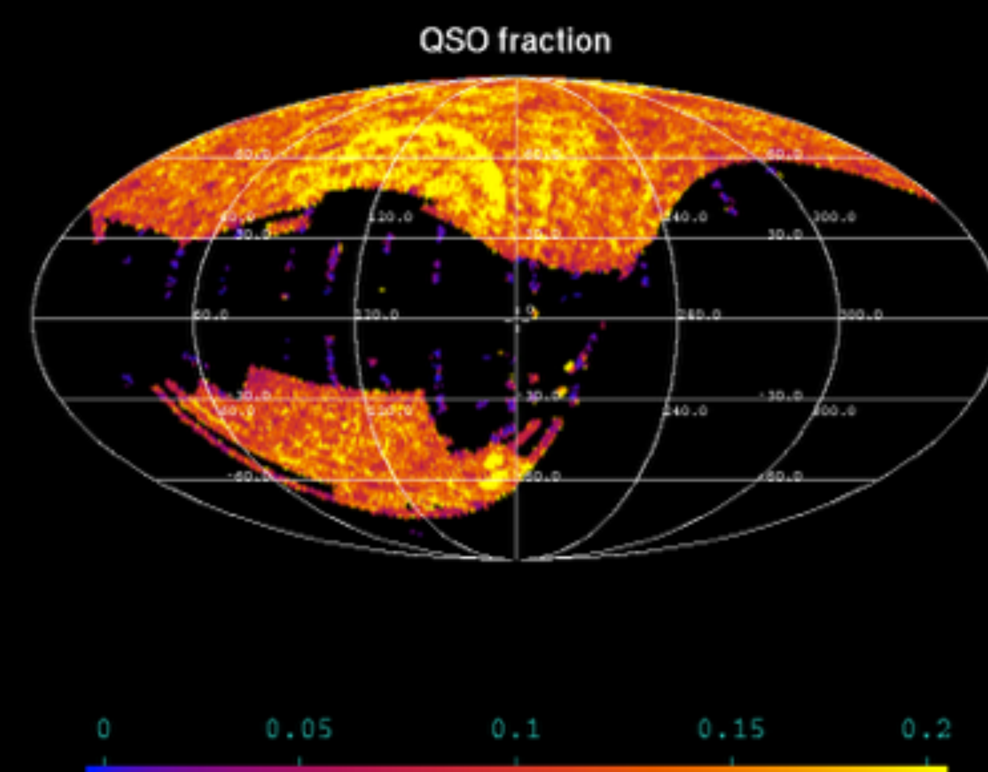


Different sub-surveys

Objects selected with both photometric and spectroscopic data available

Stars	Galaxies	QSOs	Total
928,464	2,484,161	566,475	3,979,100
23%	62%	15%	





SDSS DR12 photometry

- PSF magnitude
- Model magnitude: de Vaucouleurs / exponentiel profile

$$I(r) = I_0 \exp\{-7.67\ [(r/r_e)^{1/4}]\}$$

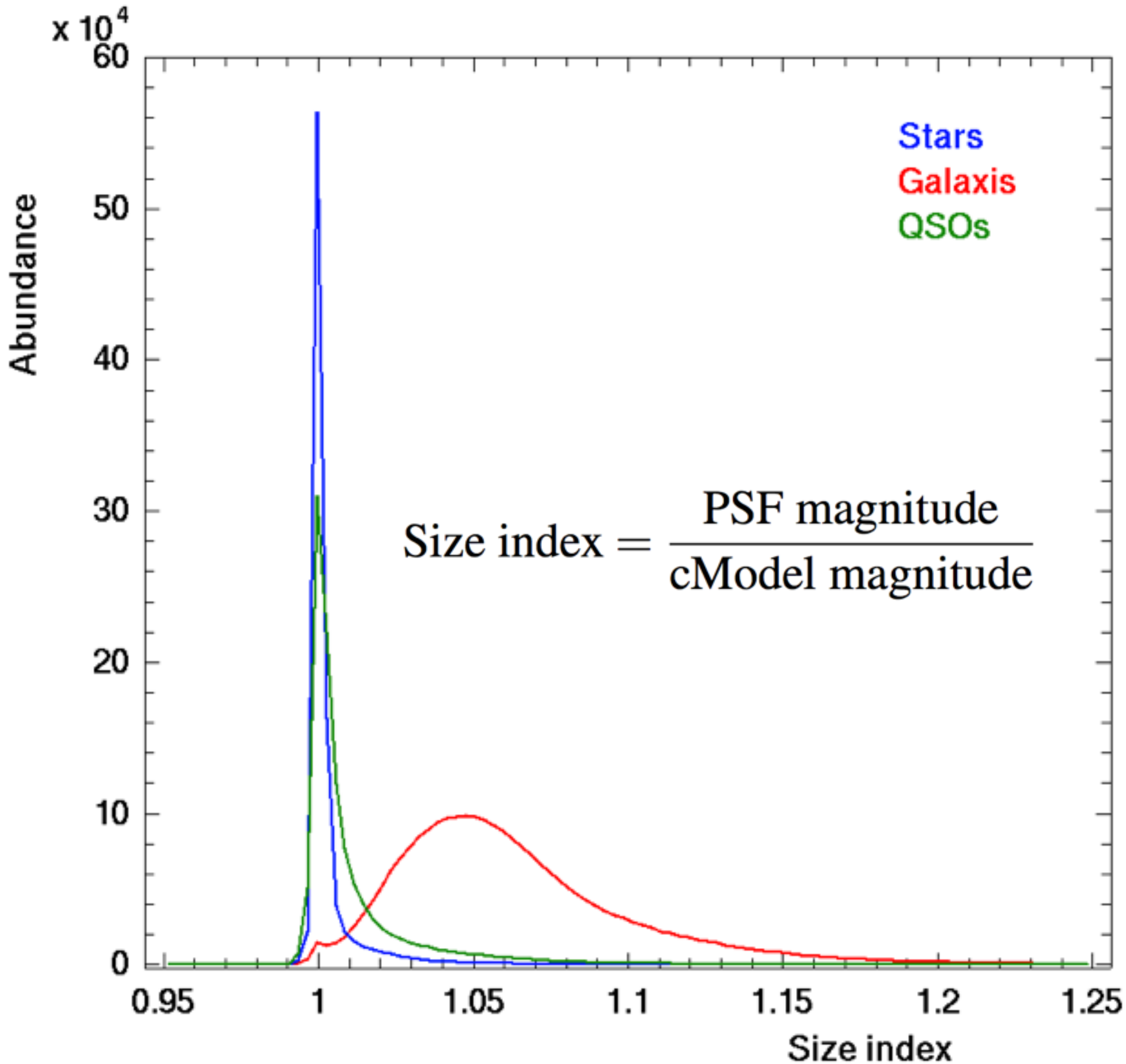
$$I(r) = I_0 \exp(-1.68 r/r_e)$$

- Composite model magnitude:

$$F_{\mathtt{composite}} = \mathtt{fracDeV}\ F_{\mathtt{deV}} + (1 - \mathtt{fracDeV})\ F_{\mathtt{exp}}$$

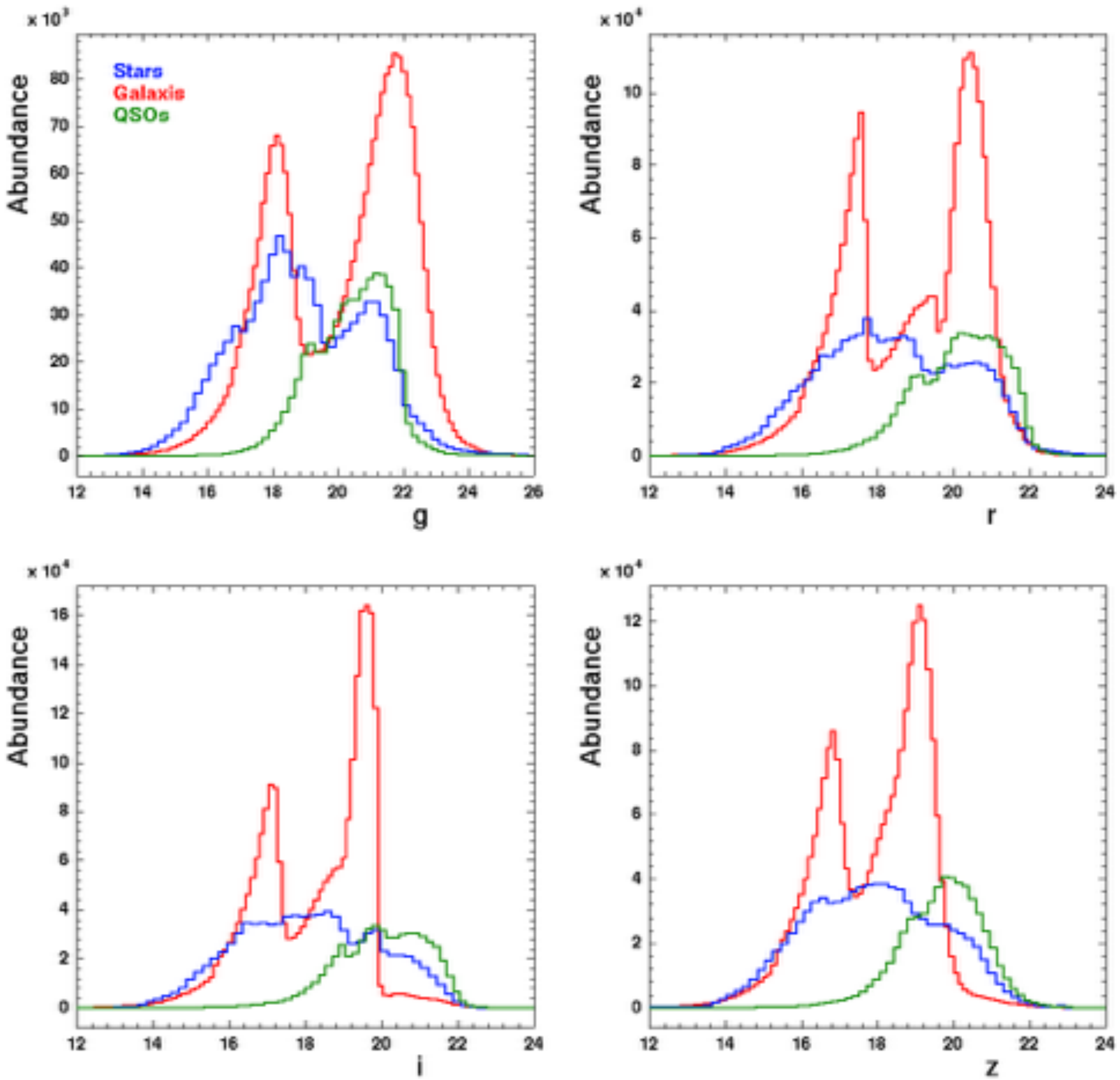
$$\text{Size index} = \frac{\text{PSF magnitude}}{\text{cModel magnitude}}$$

SDSS DR12 data



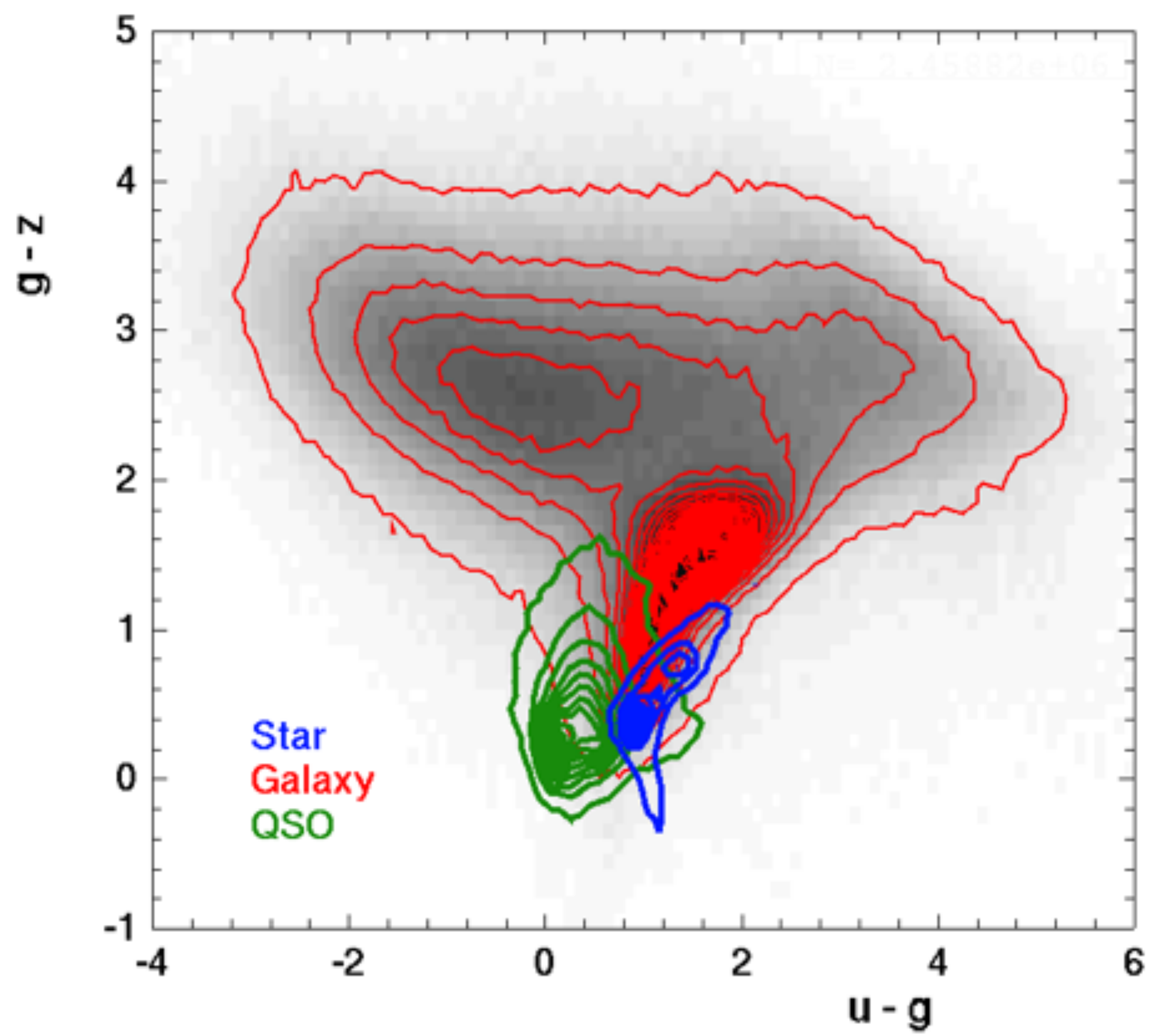
SDSS DR12 data

Magnitude distributions



Magnitudes:
insufficient to separate the objects

Colour-colour diagram



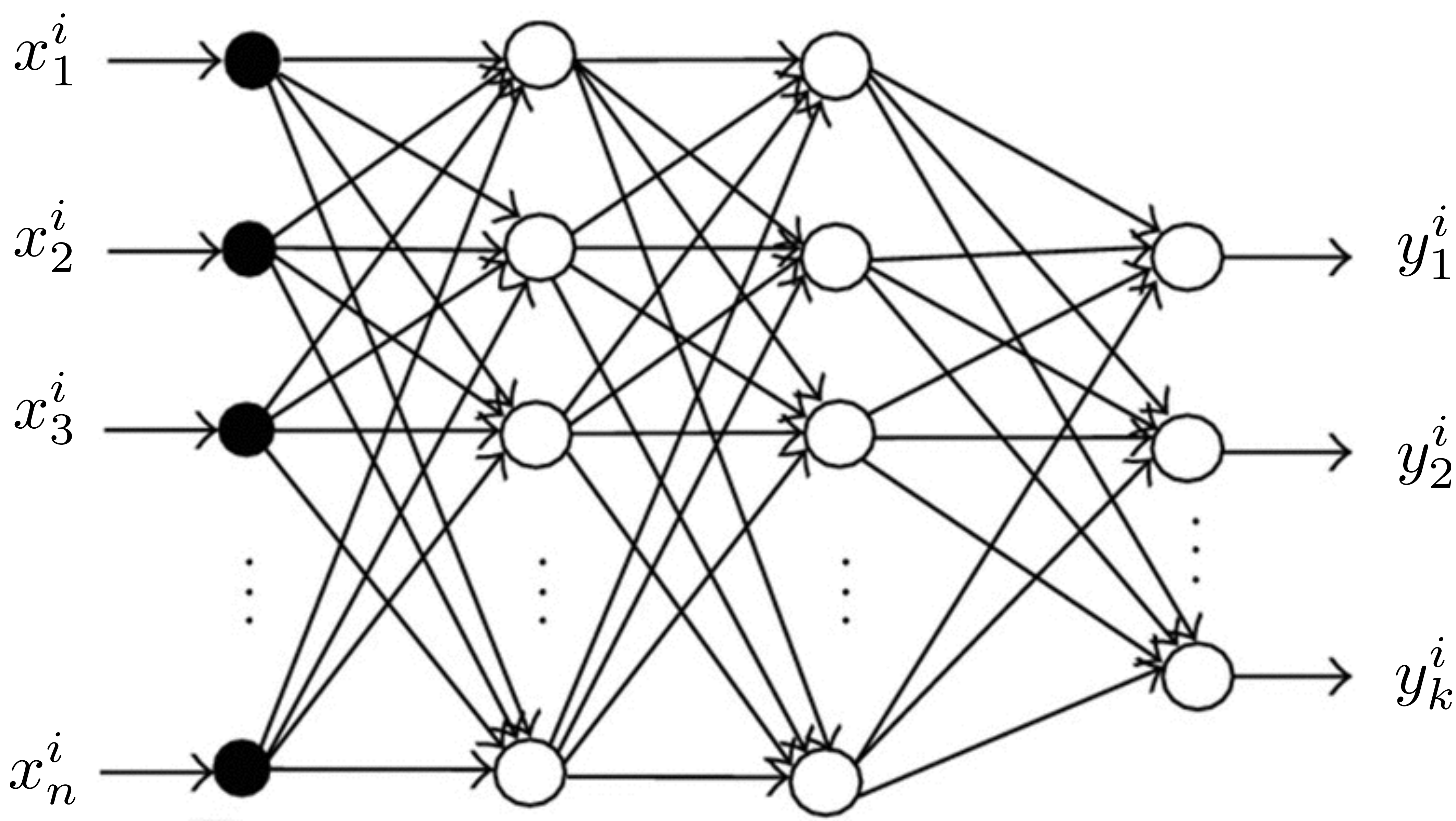
galaxies are redder
in average

Multi layer perceptron

m objects in the training set, $1 < i < m$

Each object contains n features

k number of classes labelled by vector y^i

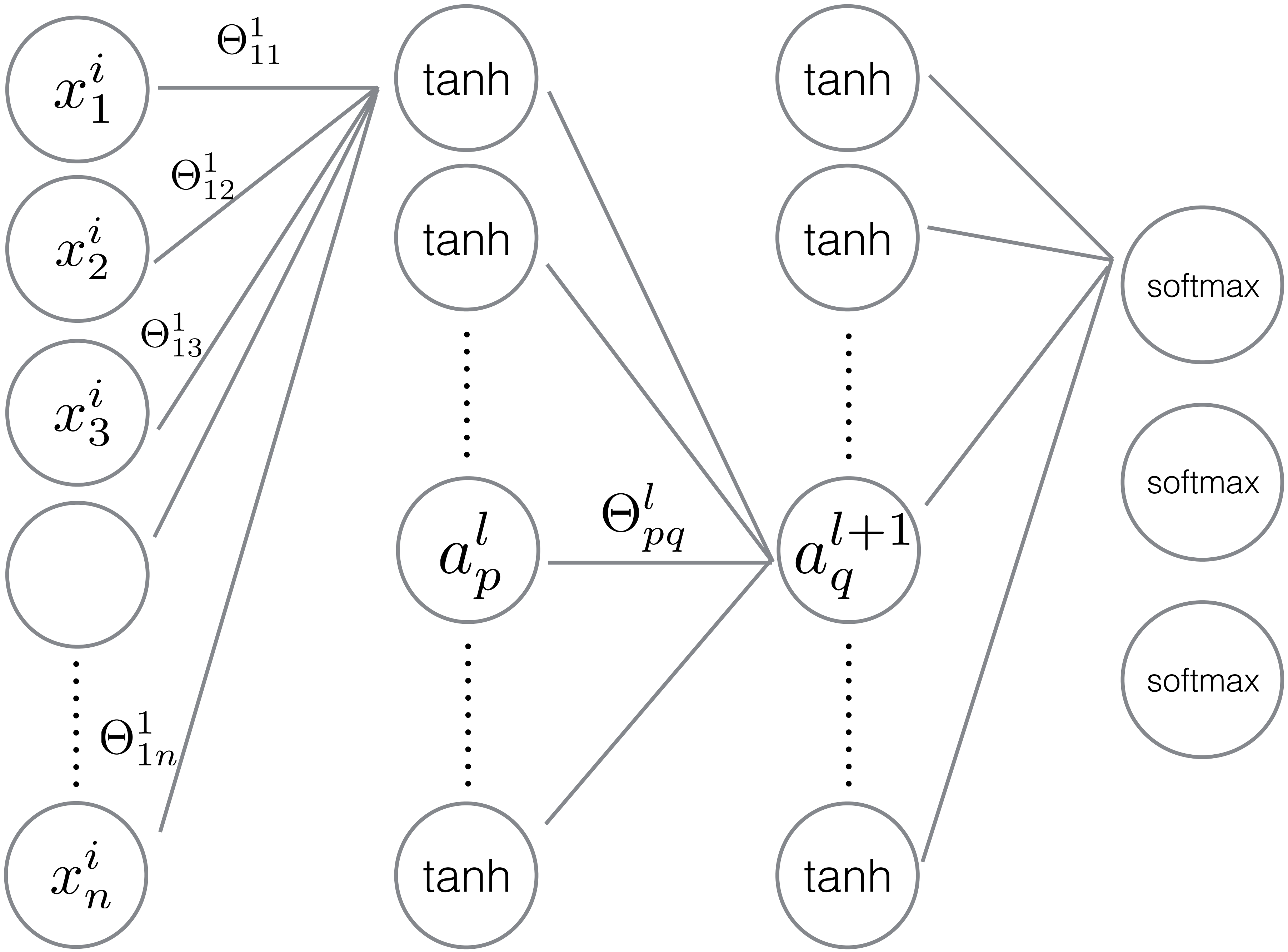


$$J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)}\, log((h_\Theta(x^{(i)}))_k) + (1 - y_k^{(i)})\, log(1 - (h_\Theta(x^{(i)}))_k)\right] + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_{j,i}^{(l)})^2$$

Multi layer perceptron

$$a_q^{l+1}(x^i) = g\,\Big(\sum_{p=0}^{s_l} \Theta_{pq}^l \, a_p^l(x^i)\Big)$$

$g(z)$: Activation function: Sigmoid, tanh, softmax and etc.



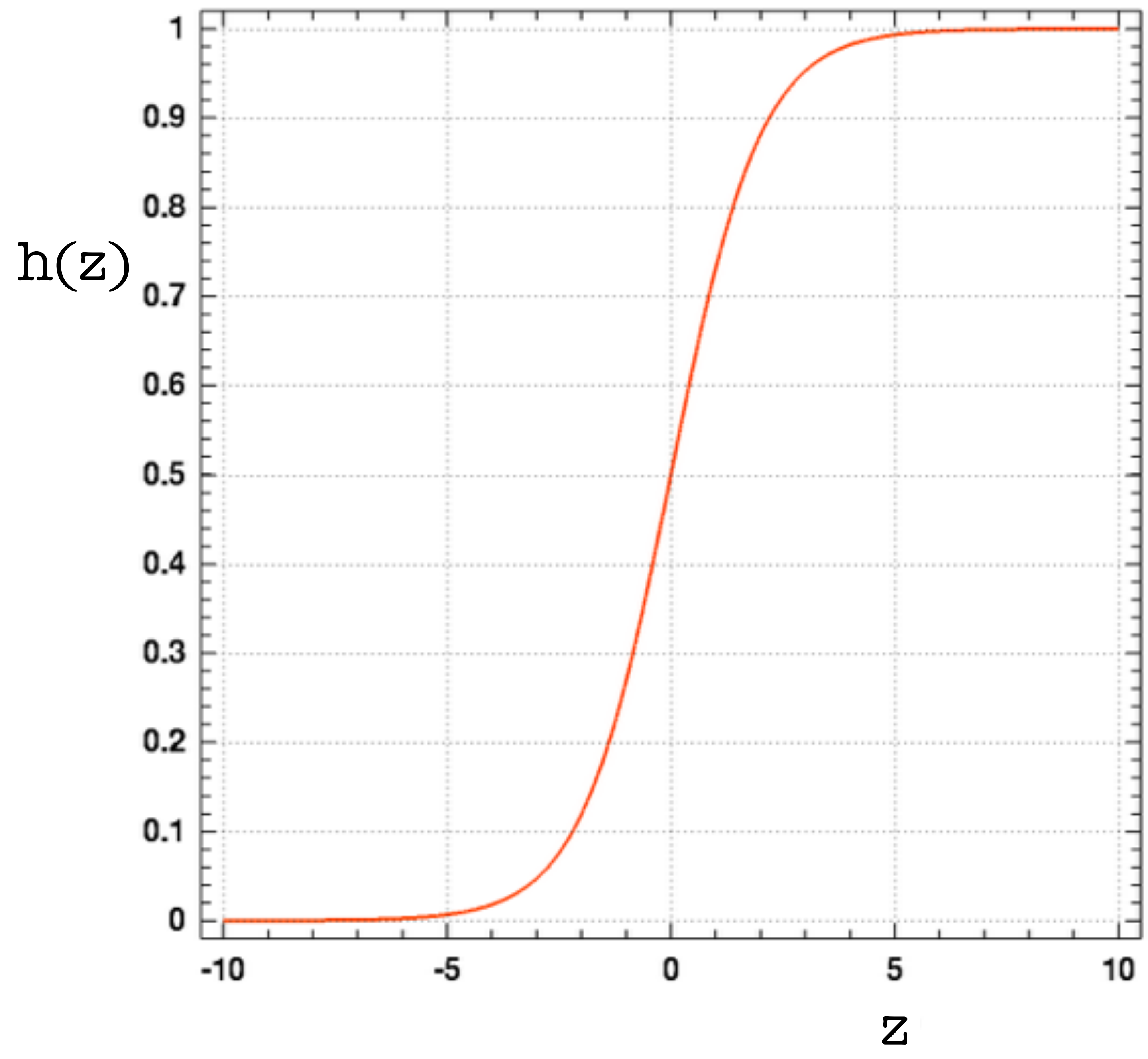
$$J(\Theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{K} y_k^{(i)}\, log((h_\Theta(x^{(i)}))_k) + (1 - y_k^{(i)})\, log(1 - (h_\Theta(x^{(i)}))_k)\right] + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_{j,i}^{(l)})^2$$

Activation

Sigmoid function

$$h(z) = \frac{1}{1+e^{-z}}$$

$$z = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n$$



Border hyper surface:

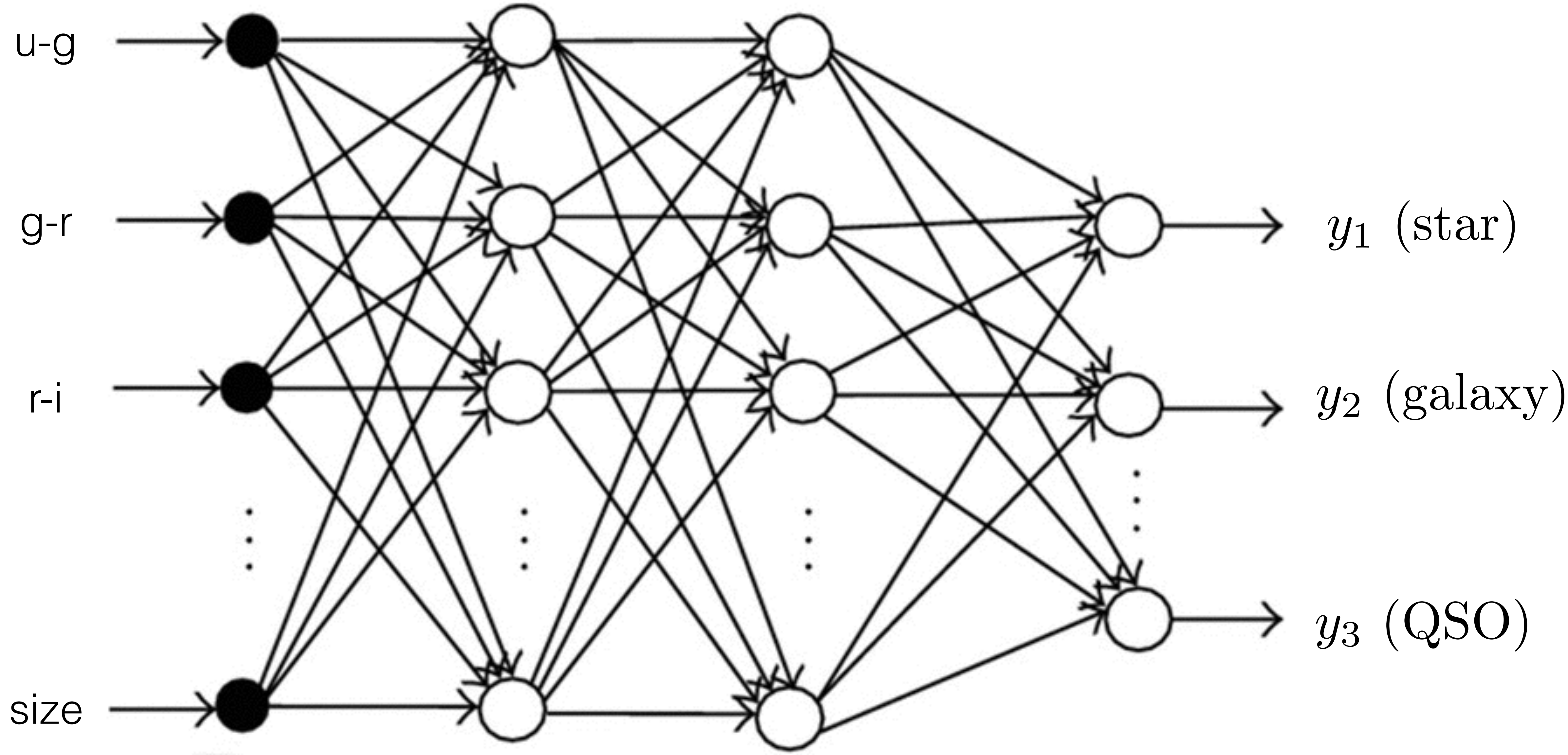
$$h(z=0) = 0.5$$

A single-minimum cost function:

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

Multi layer perceptron

Dense Neural Net with 2 hidden layers



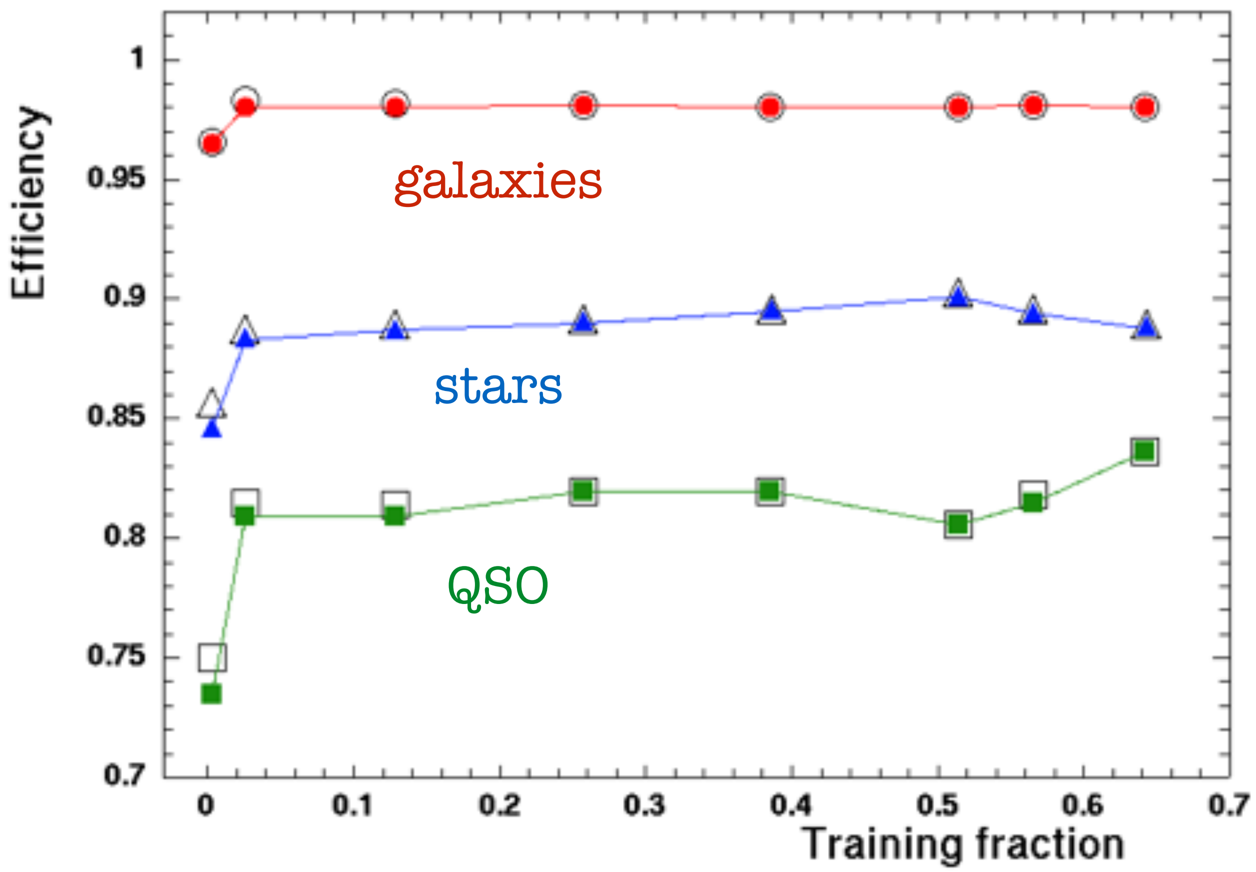
14
input features

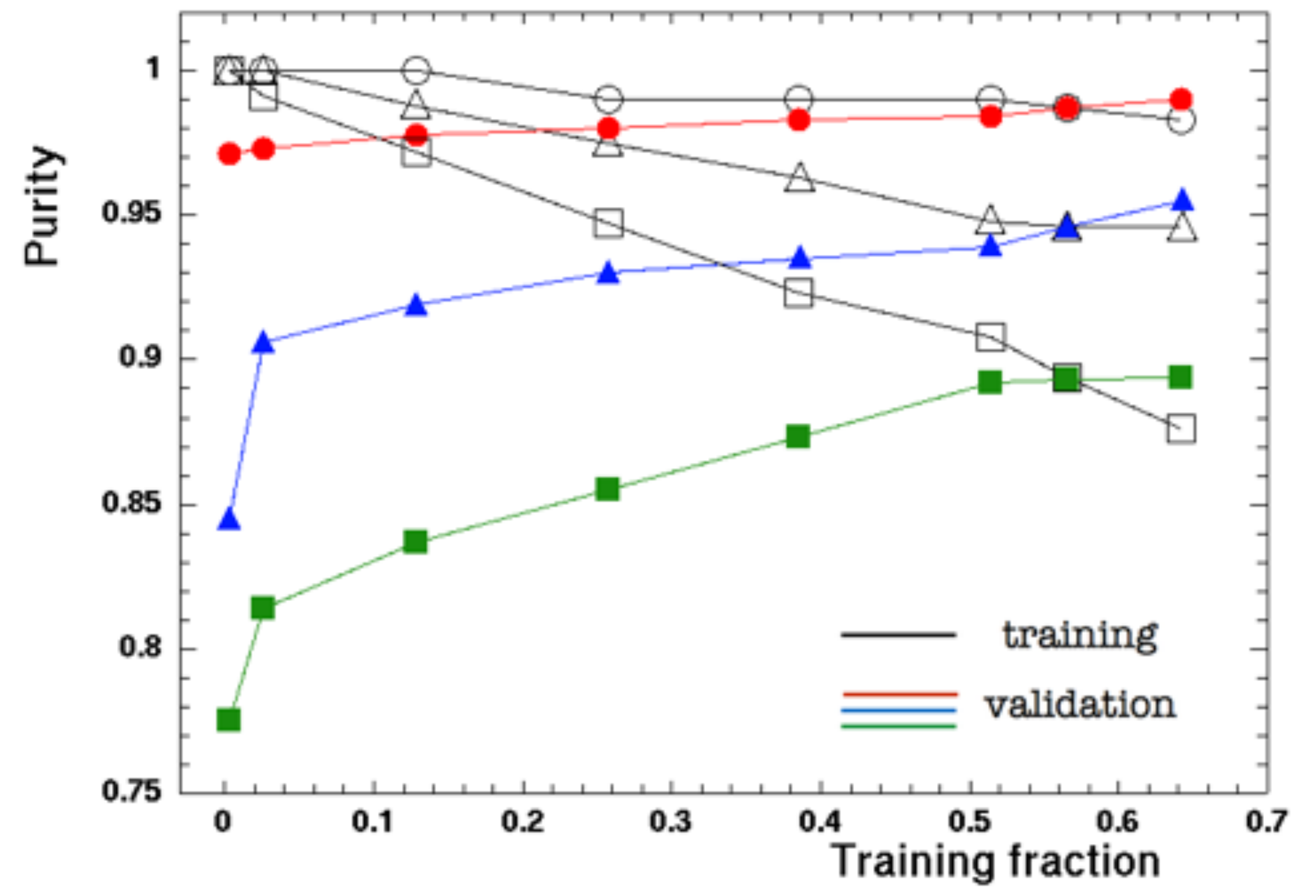
50
activations

20
activations

3
output classes

Training the MLP





$$\text{efficiency}_i = \frac{n_{i\to i}}{n_i}$$

$$\text{purity}_i = \frac{n_{i\to i}}{n_{i\to i} + \sum_{j\neq i} n_{j\to i}}$$

$$i : \text{galaxy, star or QSO}$$

Comparing
training and validation sets
to find optimum number
of objects for training

Training the MLP

Efficiency and purity of the classification

size index included

	Star	Galaxy	QSO	Total
Efficiency	89.4%	98.1%	81.5%	94%
Purity	94.6%	98.7%	89.3%	-

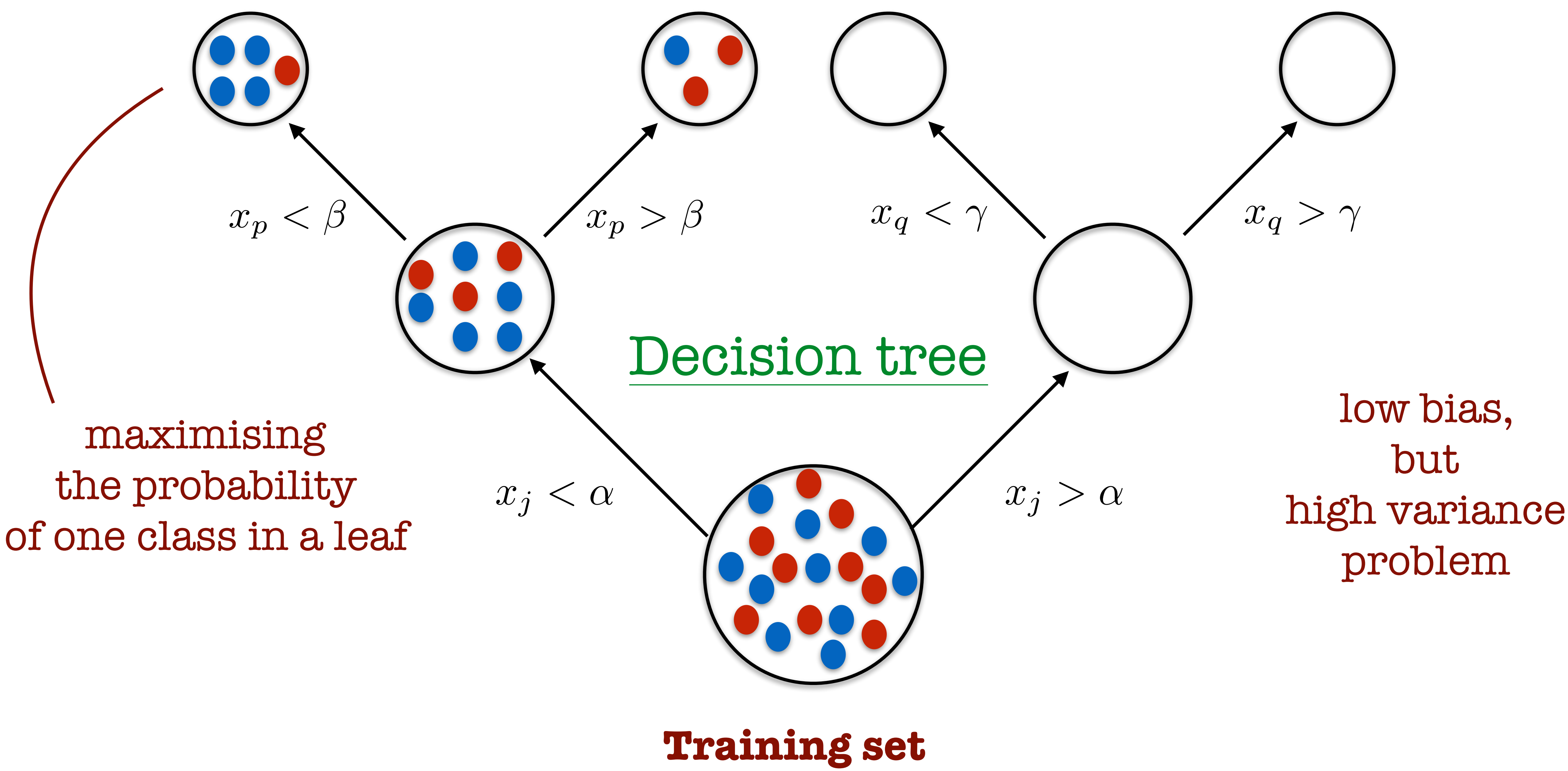
$$\text{Size index} = \frac{\text{PSF magnitude}}{\text{cModel magnitude}}$$

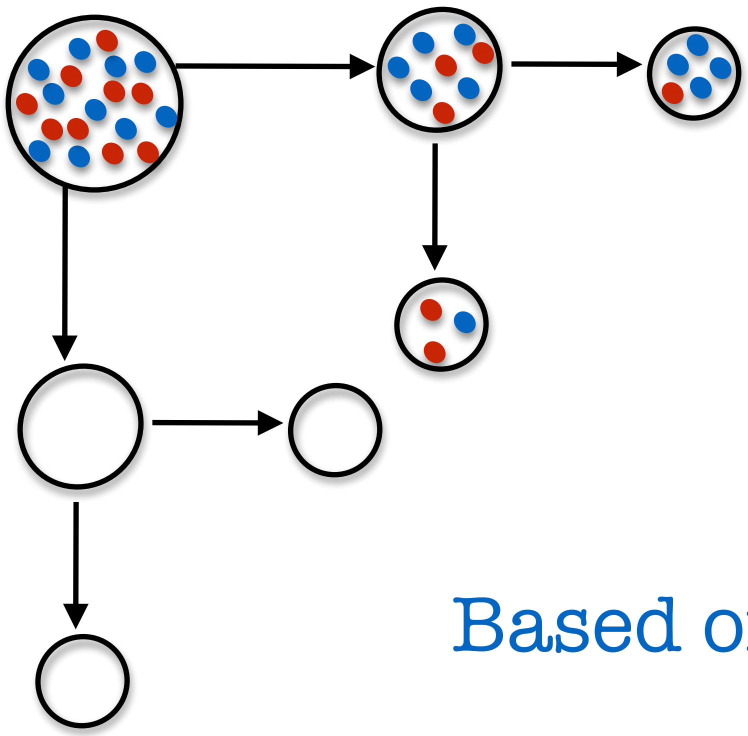
no size index

	Star	Galaxy	QSO	Total
Efficiency	86.6%	97.5%	78.5%	92%
Purity	93.0%	97.7%	88.1%	-

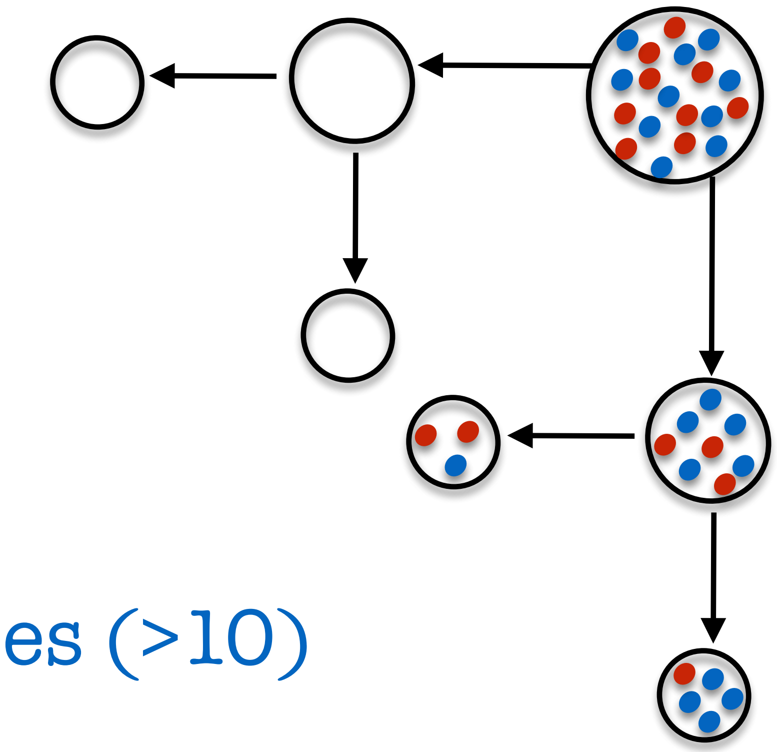
Random forest

constructing a classification model through feature filtering





Random forest

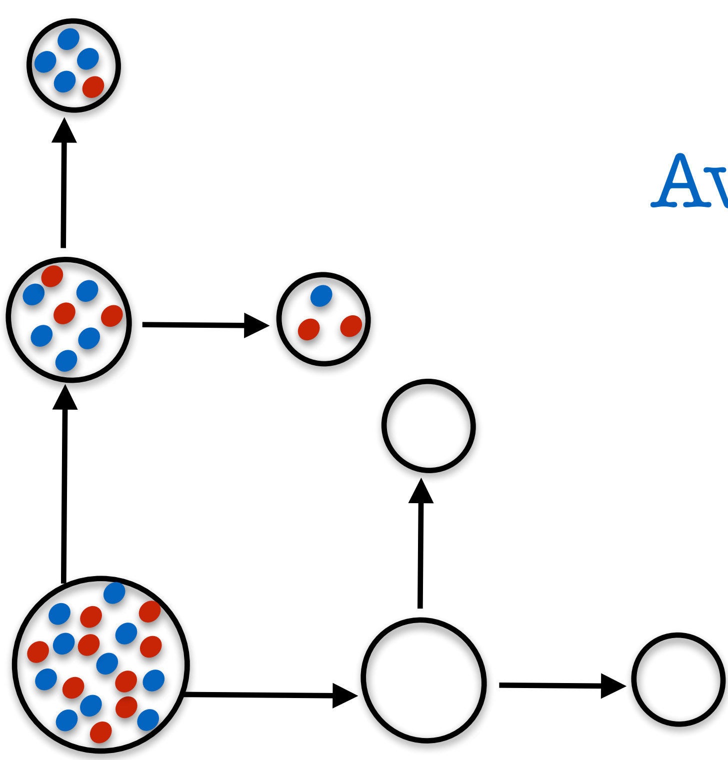


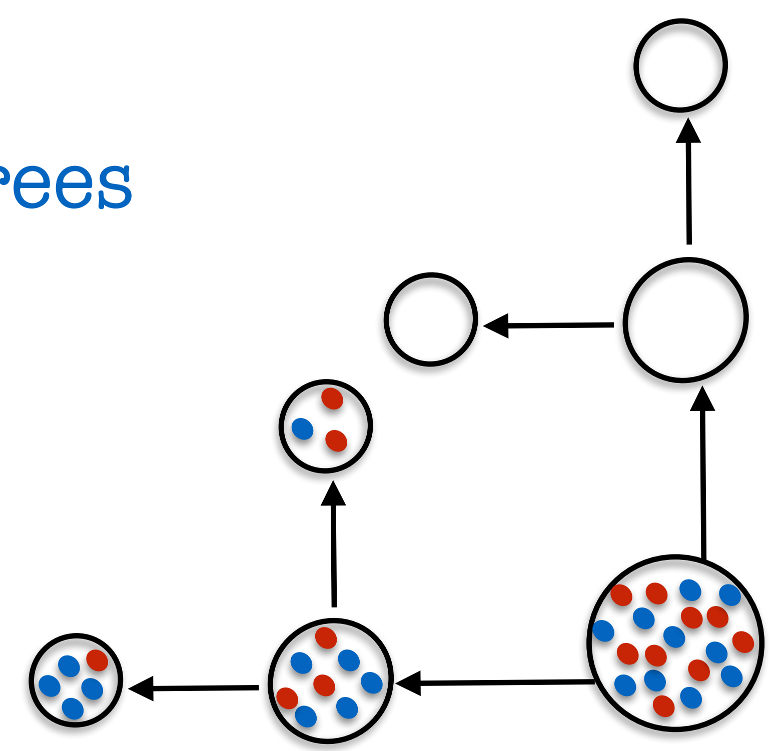
Based on large number of decision trees (>10)

Generating different samples of the training set through bootstrapping (sampling with replacement)

Feature (random) bagging at conjunctions

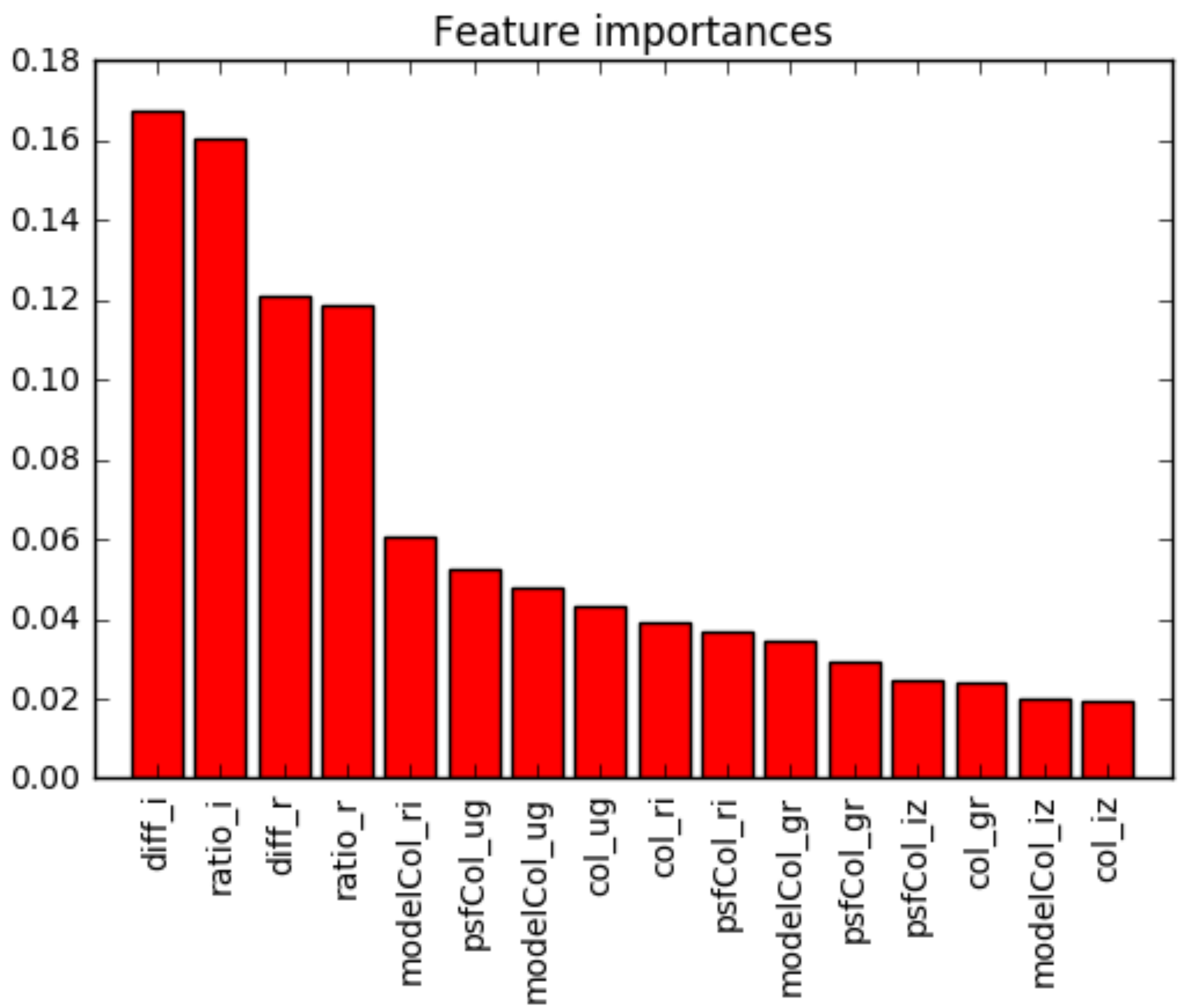
Average over predictions of all trees



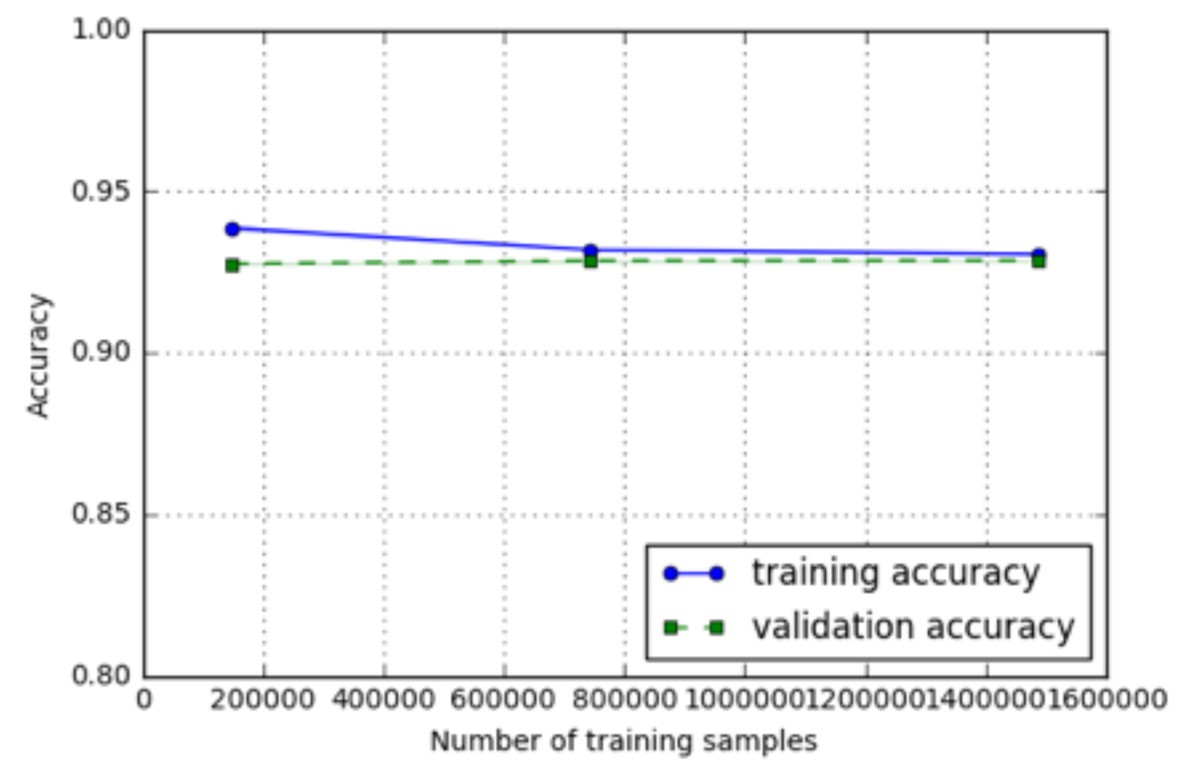


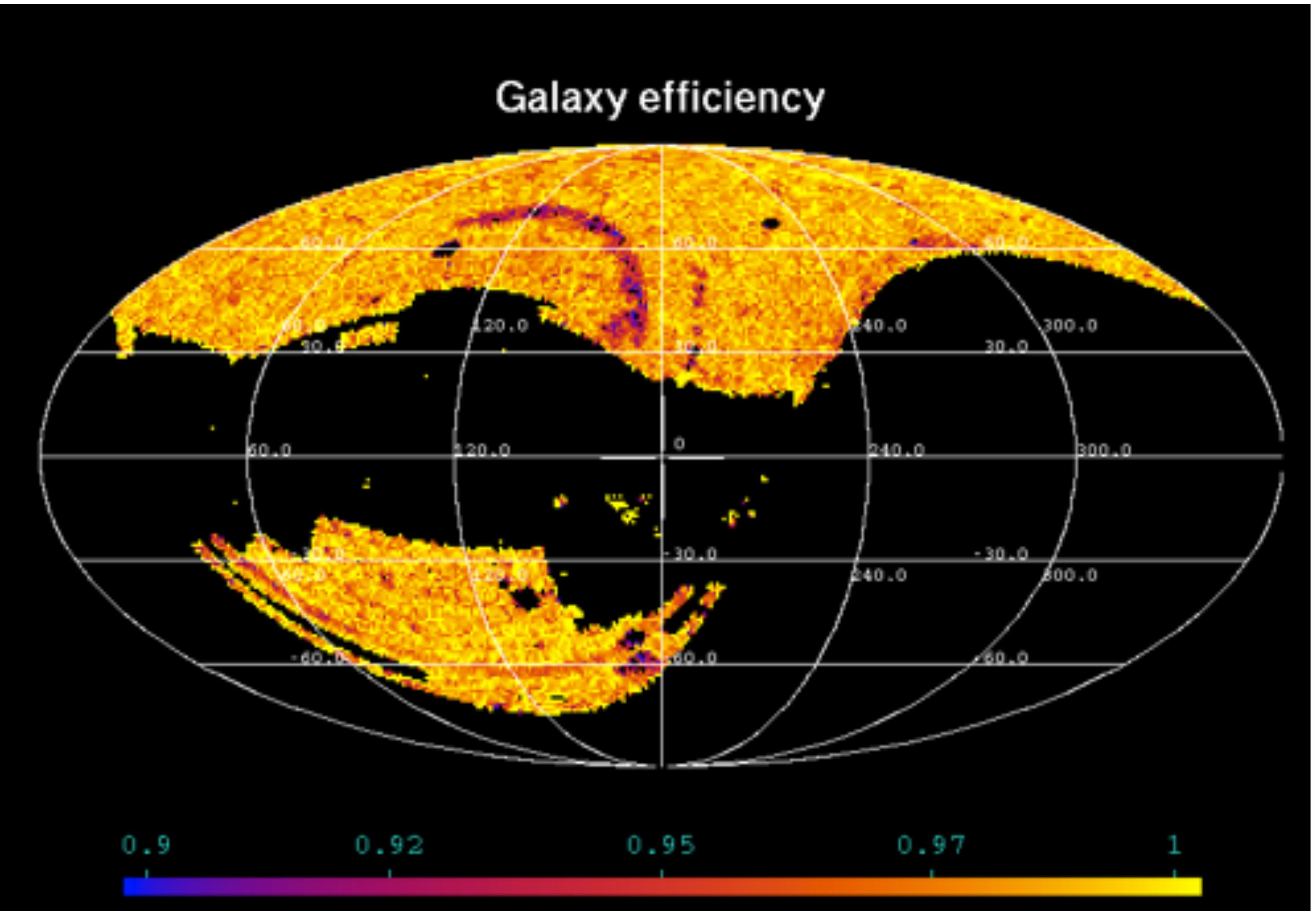
Random forest

Efficiency	Star	Galaxy	QSO	Total
RF	86.9%	98.0%	80.2%	93%
NN	89.4%	98.1%	81.5%	94%

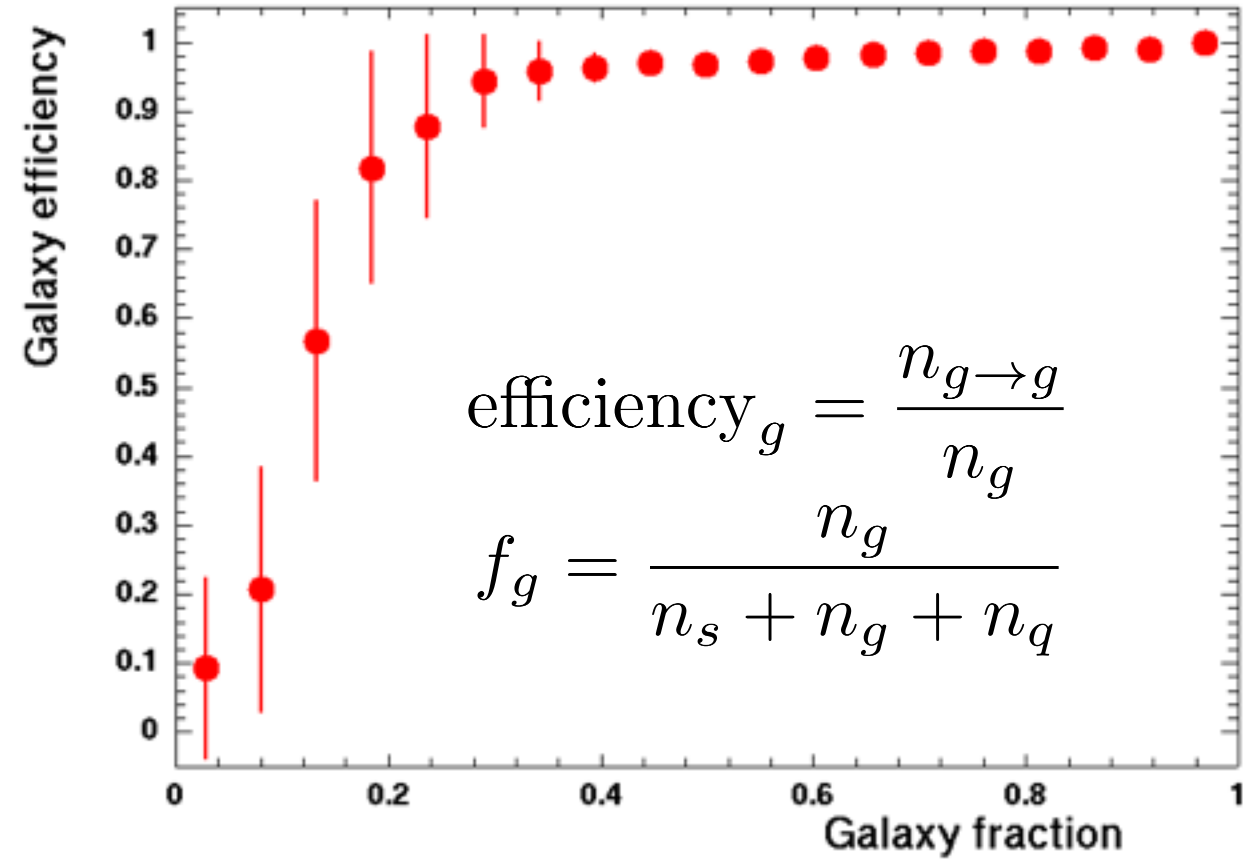


Size index is the most discriminative feature for RF

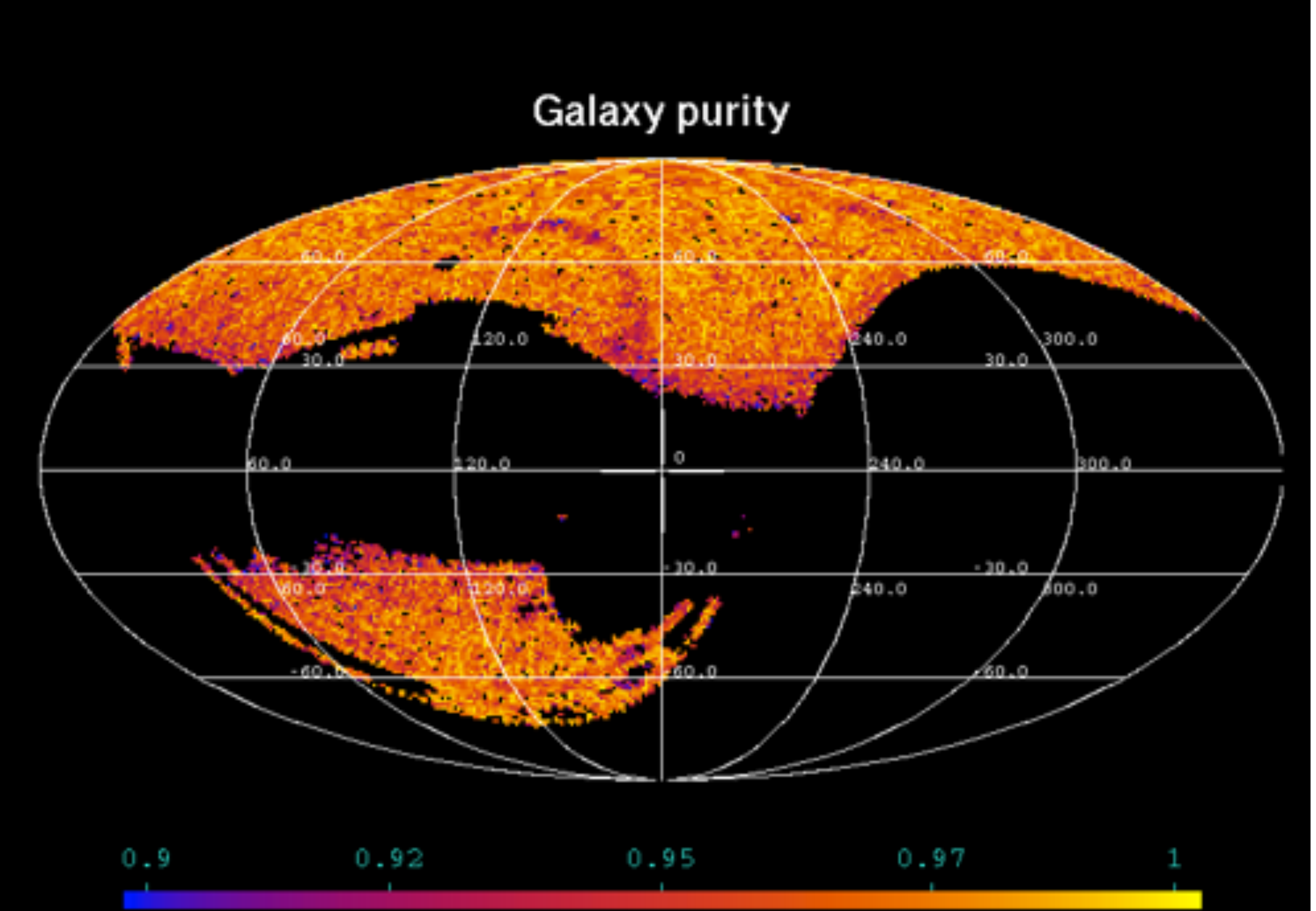




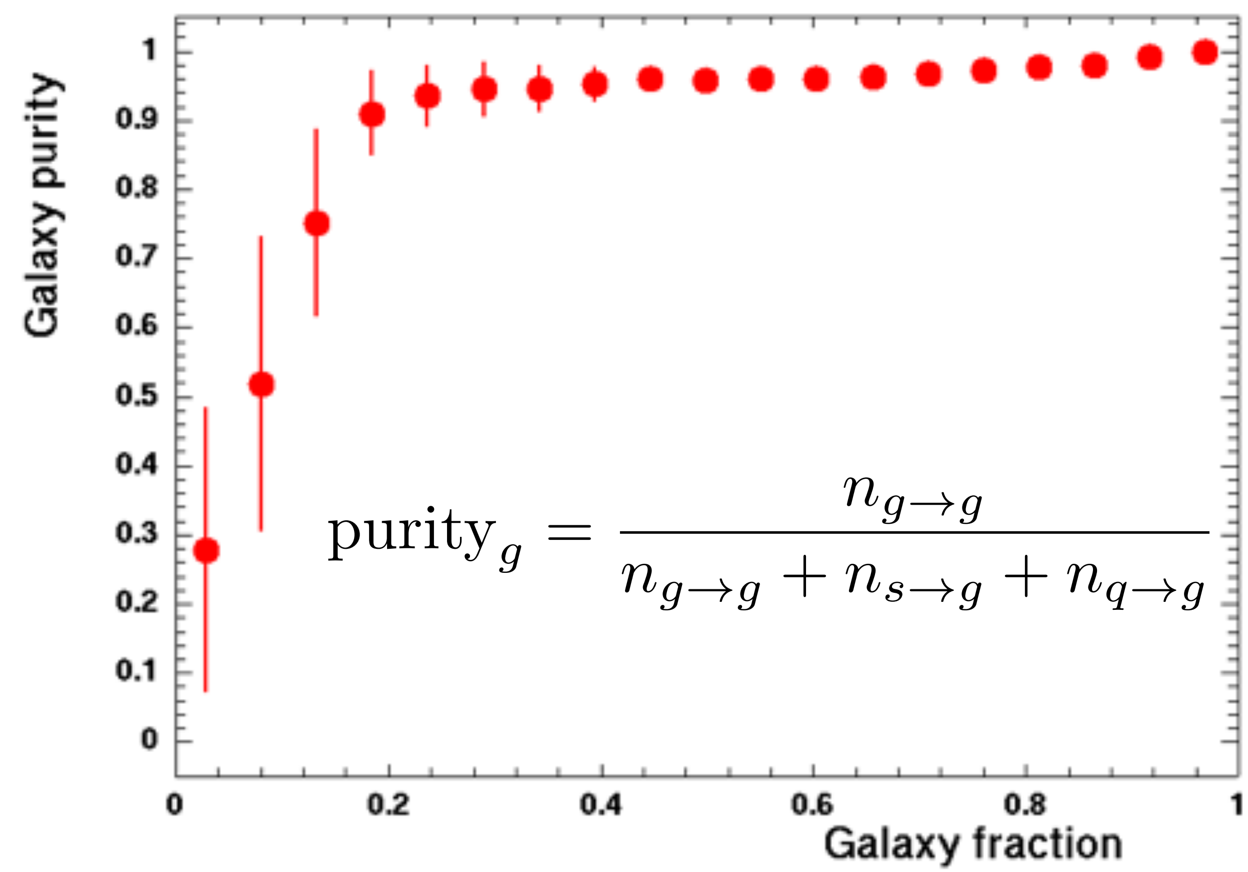
Galaxy efficiency
0.9
0.92
0.95
0.97
1



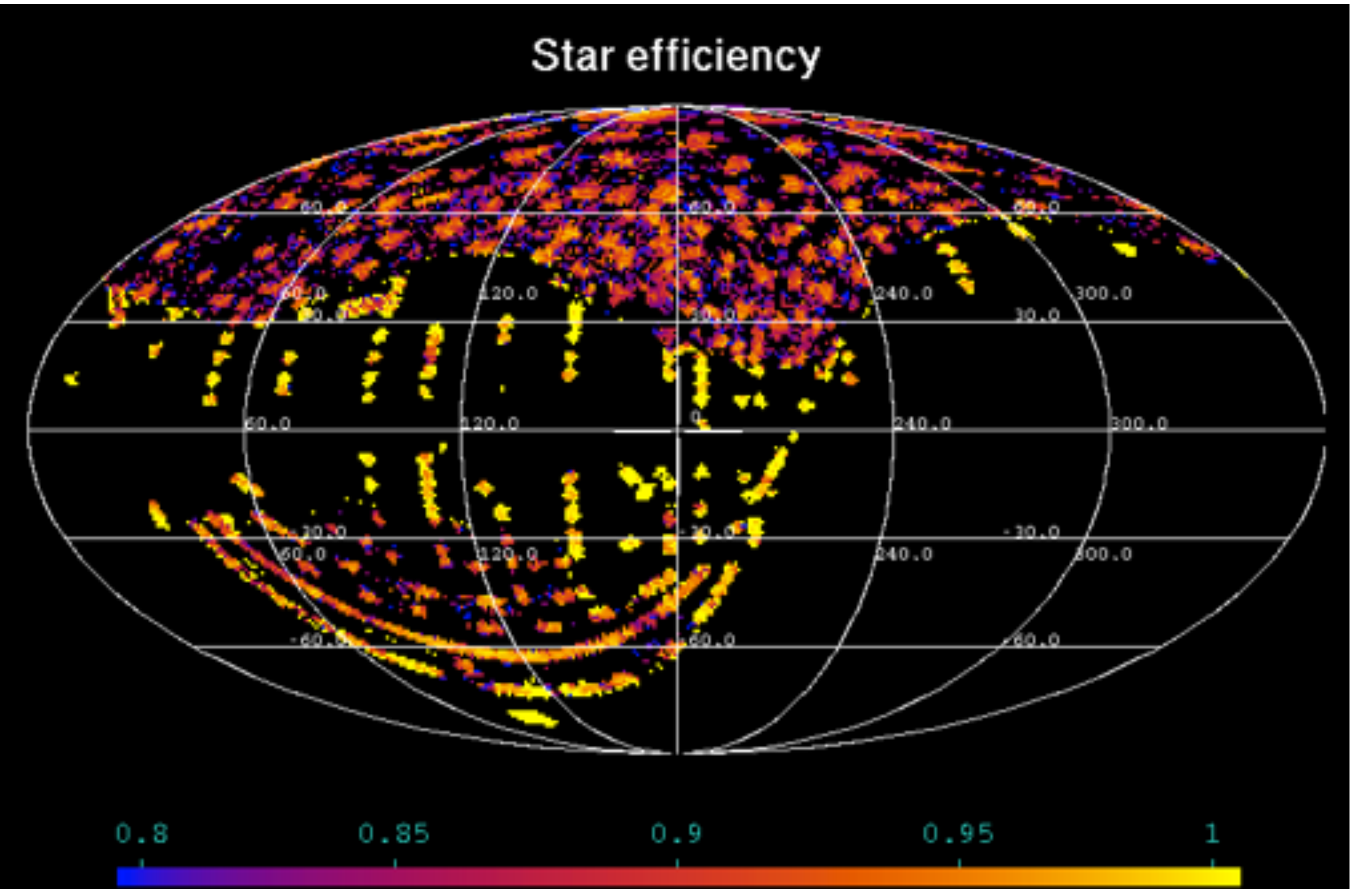
Galaxy efficiency
Galaxy fraction
$$\text{efficiency}_g = \frac{n_{g\to g}}{n_g}$$
$$f_g = \frac{n_g}{n_s + n_g + n_q}$$

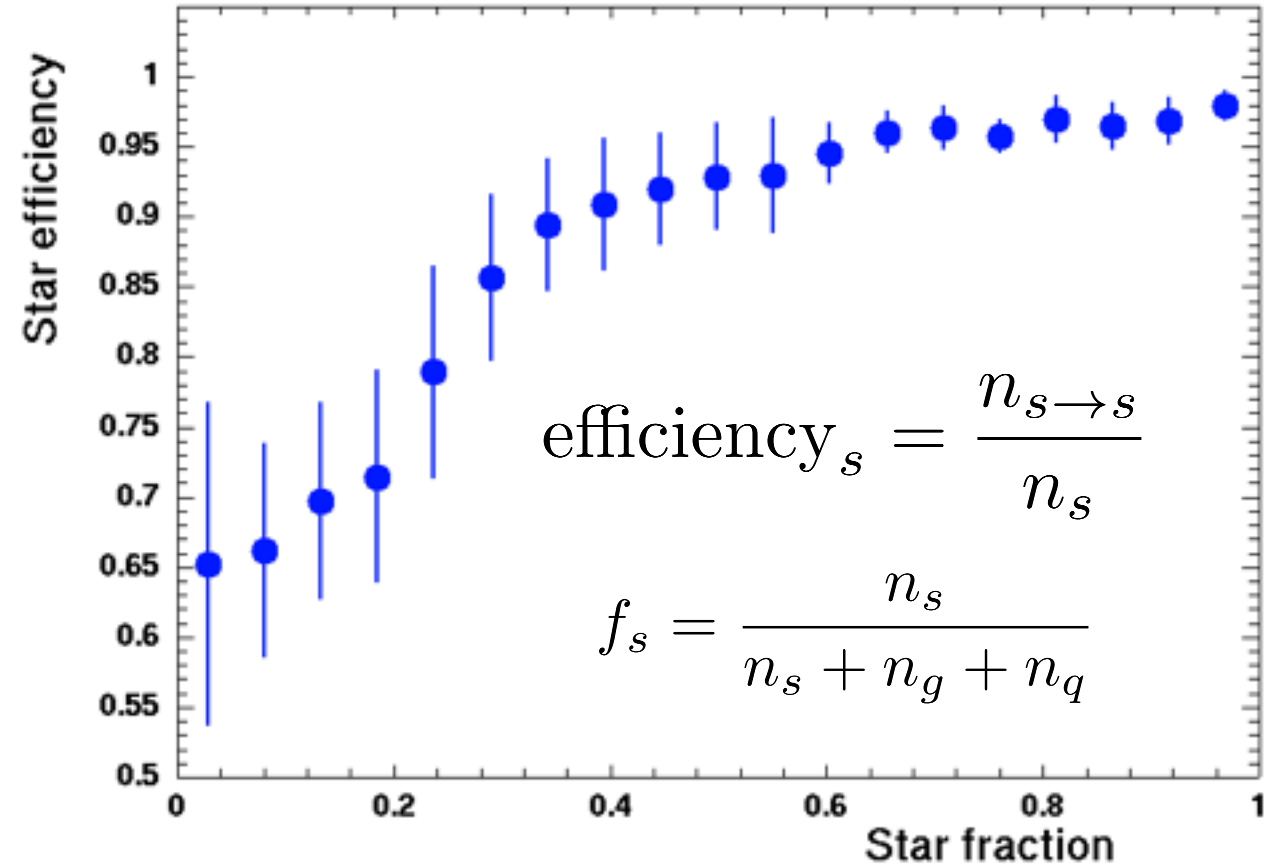


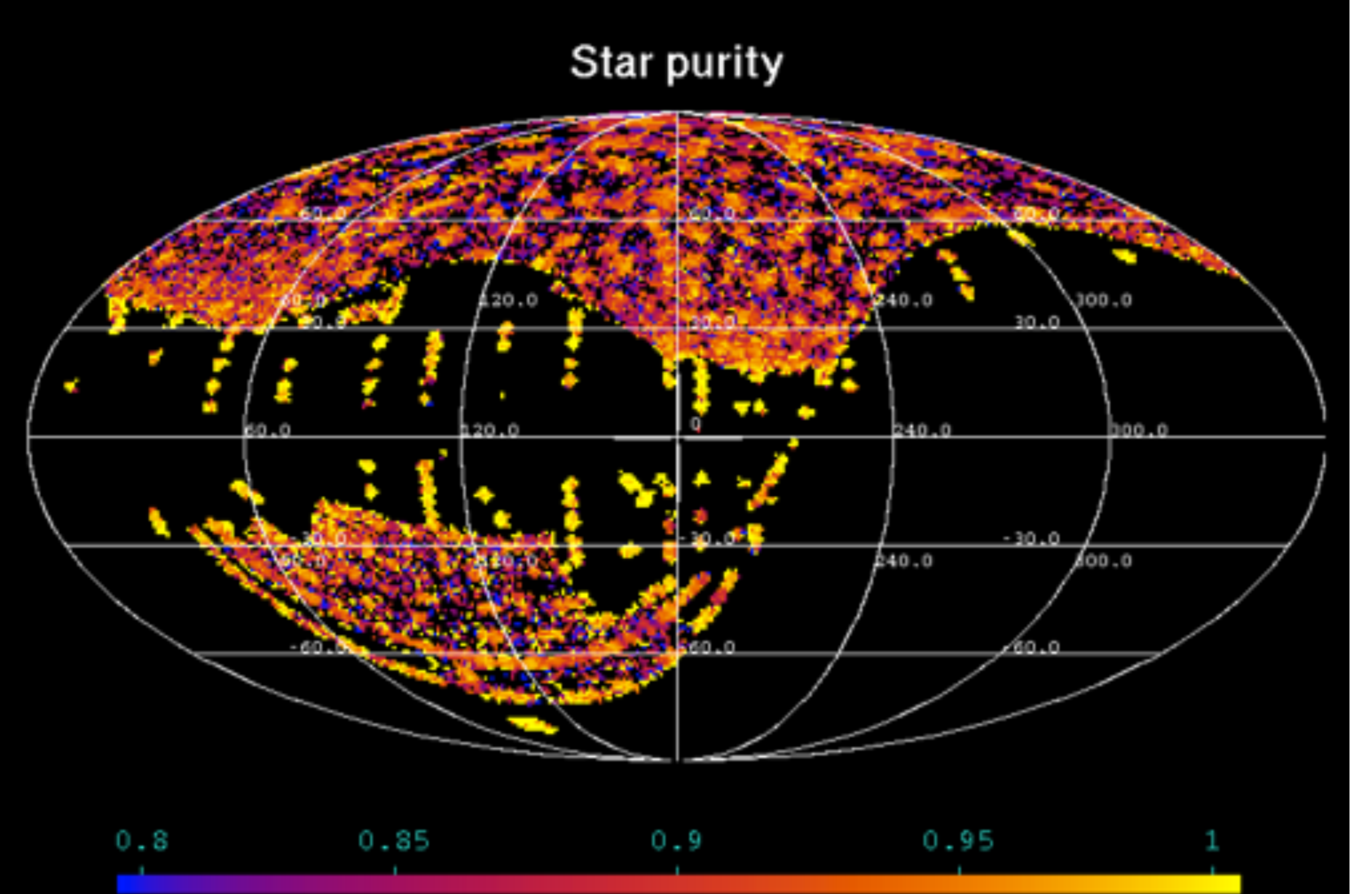
Galaxy purity
0.9
0.92
0.95
0.97
1

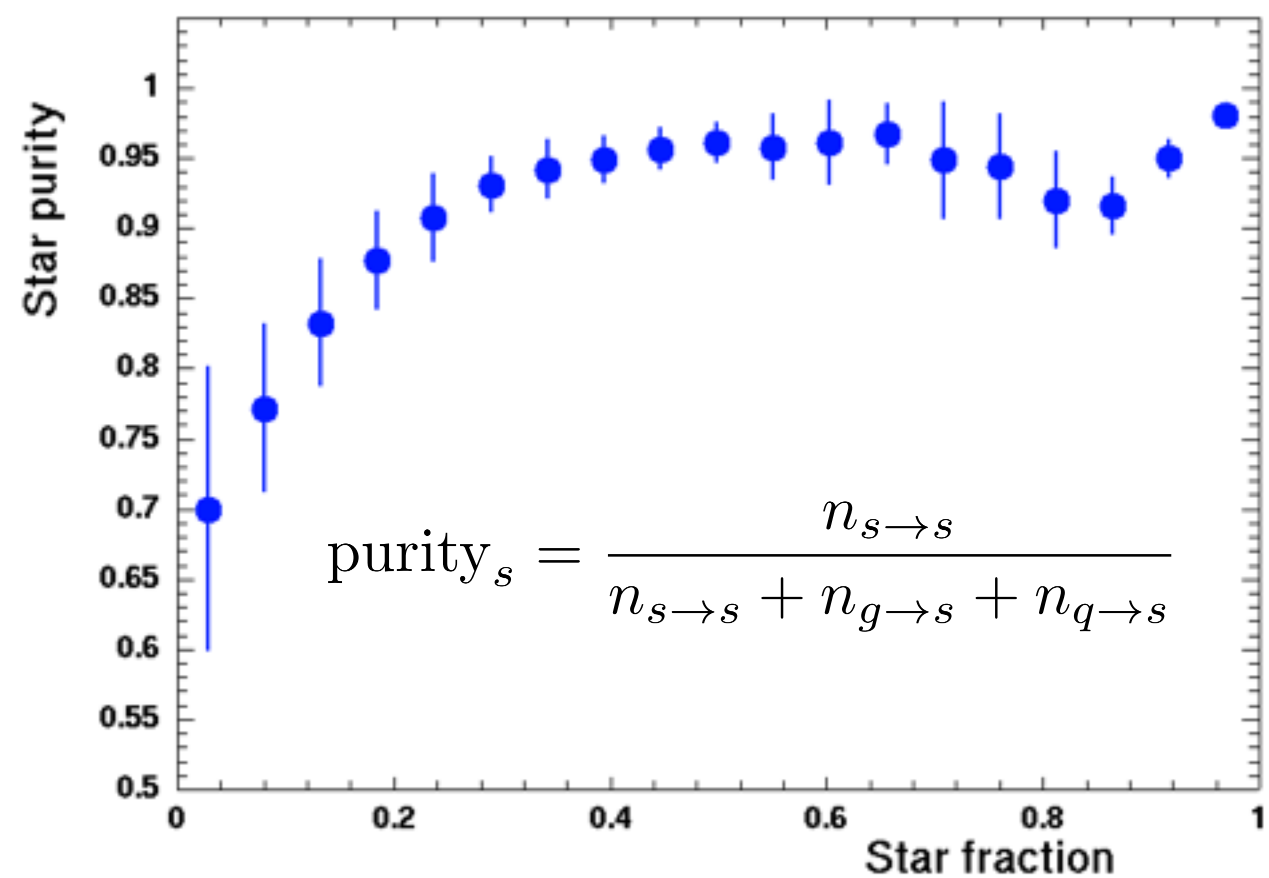


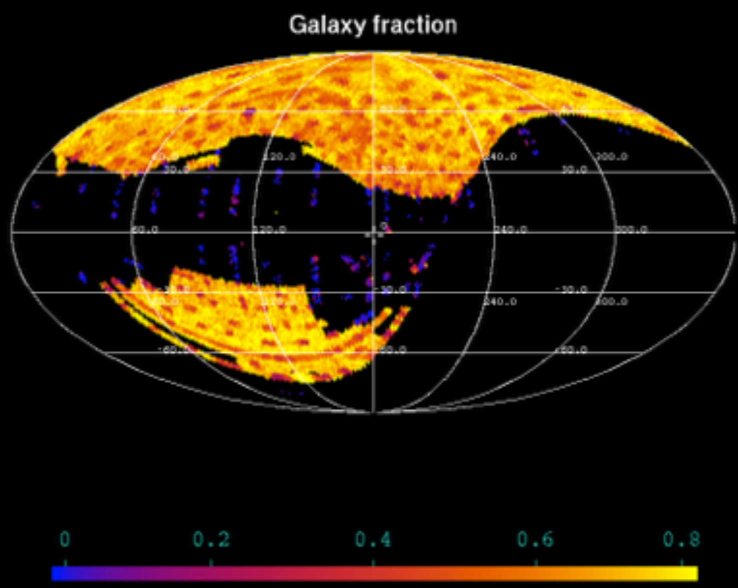
Galaxy purity
Galaxy fraction
$$\text{purity}_g = \frac{n_{g\to g}}{n_{g\to g} + n_{s\to g} + n_{q\to g}}$$

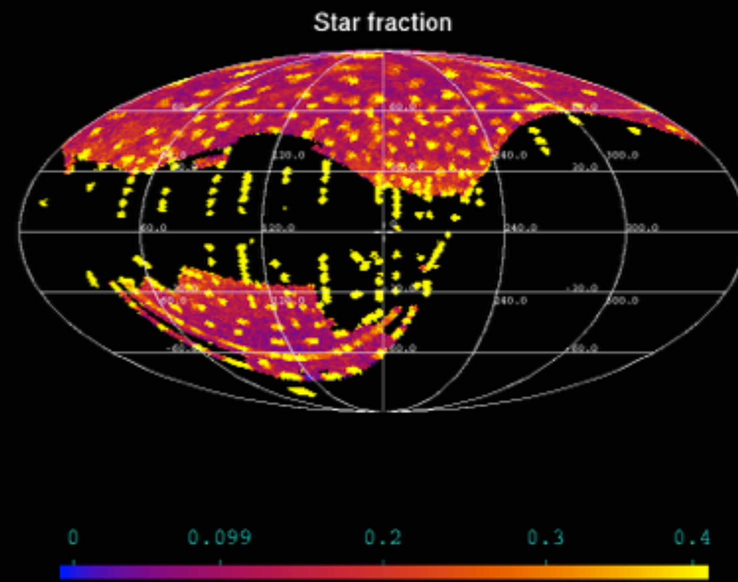

Star efficiency
0.8
0.85
0.9
0.95
1

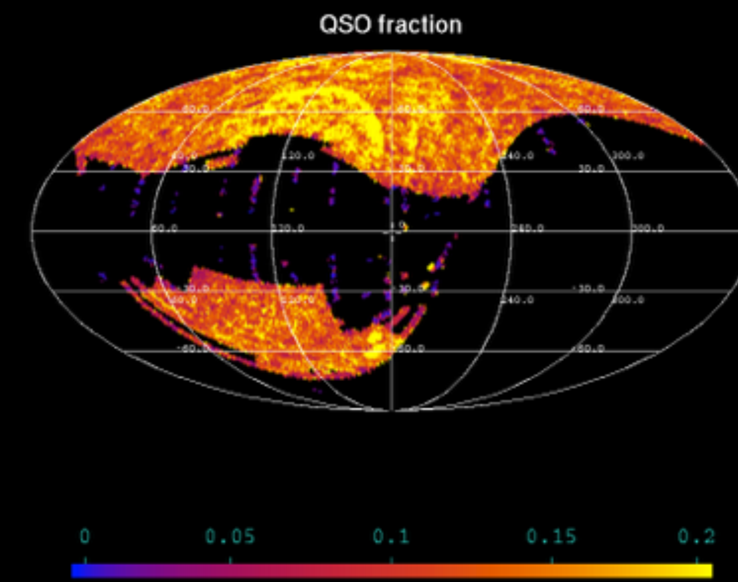

Star efficiency
Star fraction
$\mathrm{efficiency}_s = \frac{n_{s\to s}}{n_s}$
$f_s = \frac{n_s}{n_s + n_g + n_q}$


Star purity
0.8
0.85
0.9
0.95
1


Star purity
Star fraction
$\mathrm{purity}_s = \frac{n_{s\to s}}{n_{s\to s} + n_{g\to s} + n_{q\to s}}$



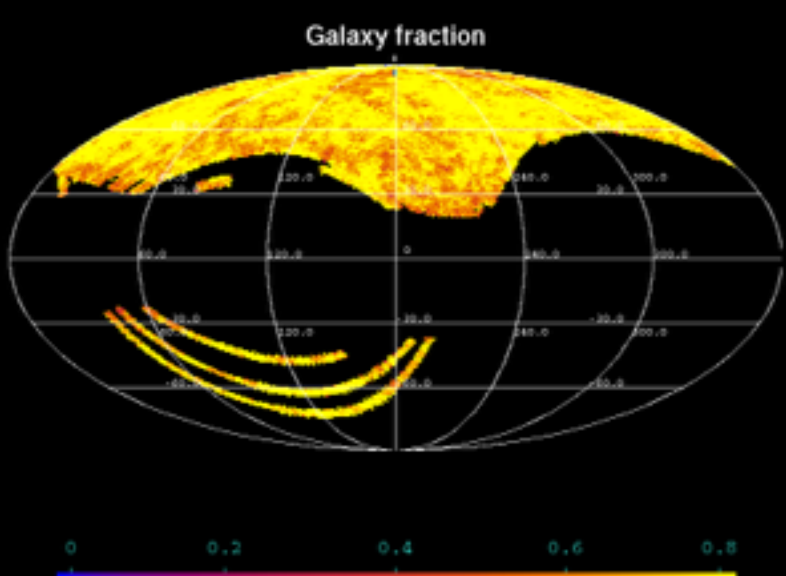


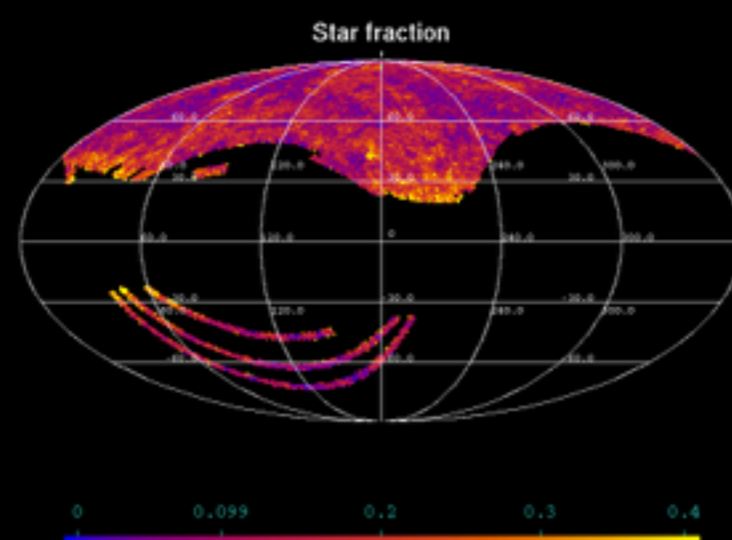


DR 12

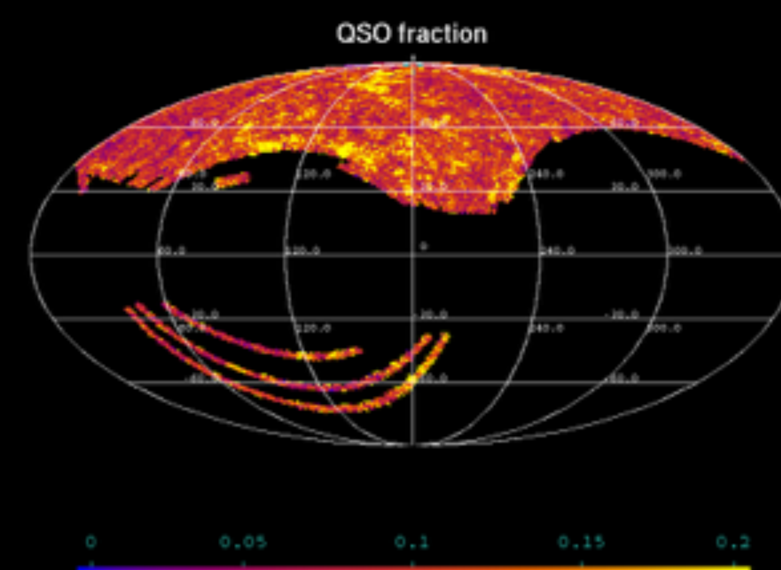
	Stars	Galaxies	QSOs	Total
Num	160,040	879,792	120,425	1,160,257
fraction	14%	76%	10%	
efficiency	95%	99%	90%	98%
purity	94%	99%	94%	

DR12
vs
Legacy

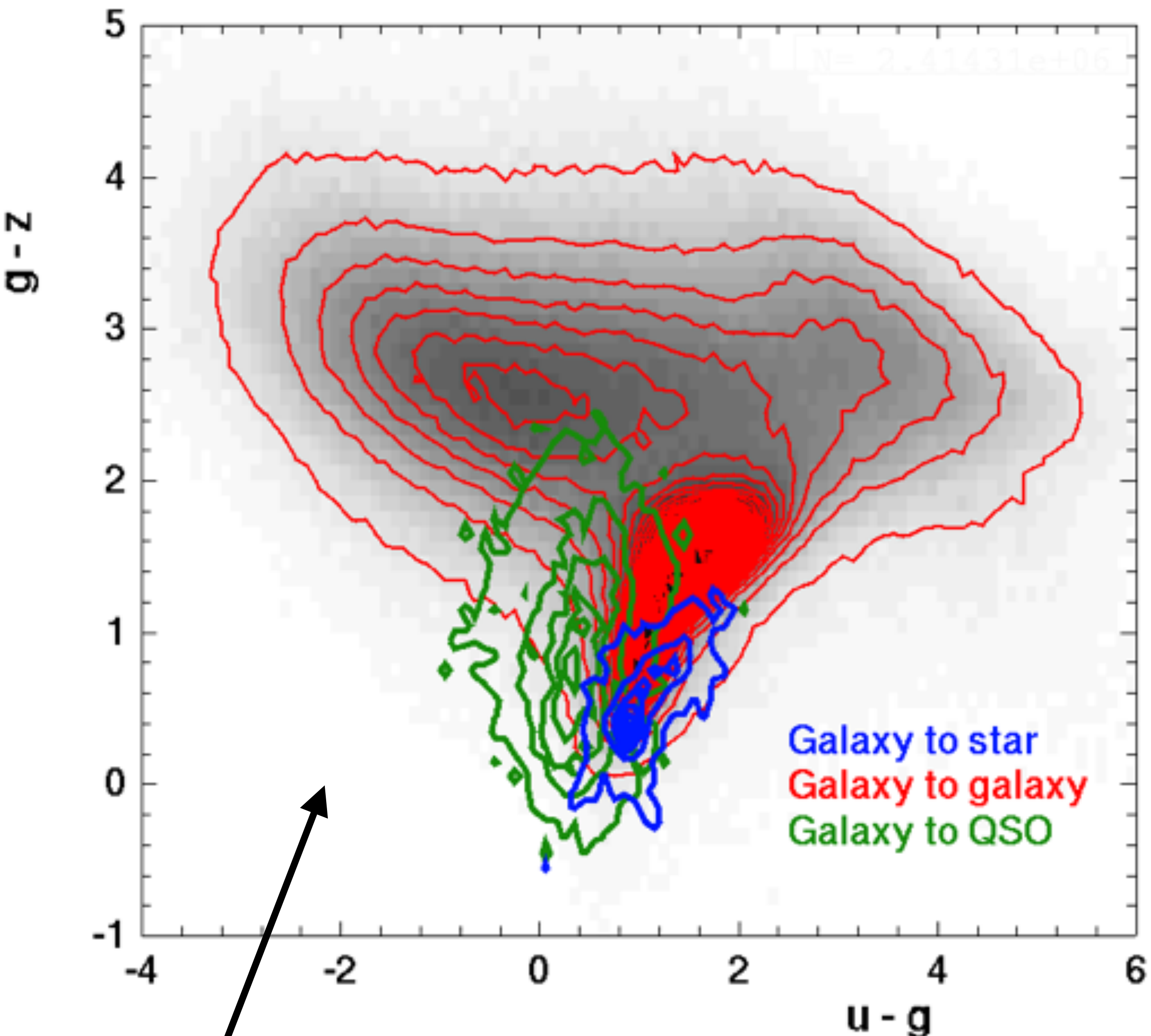




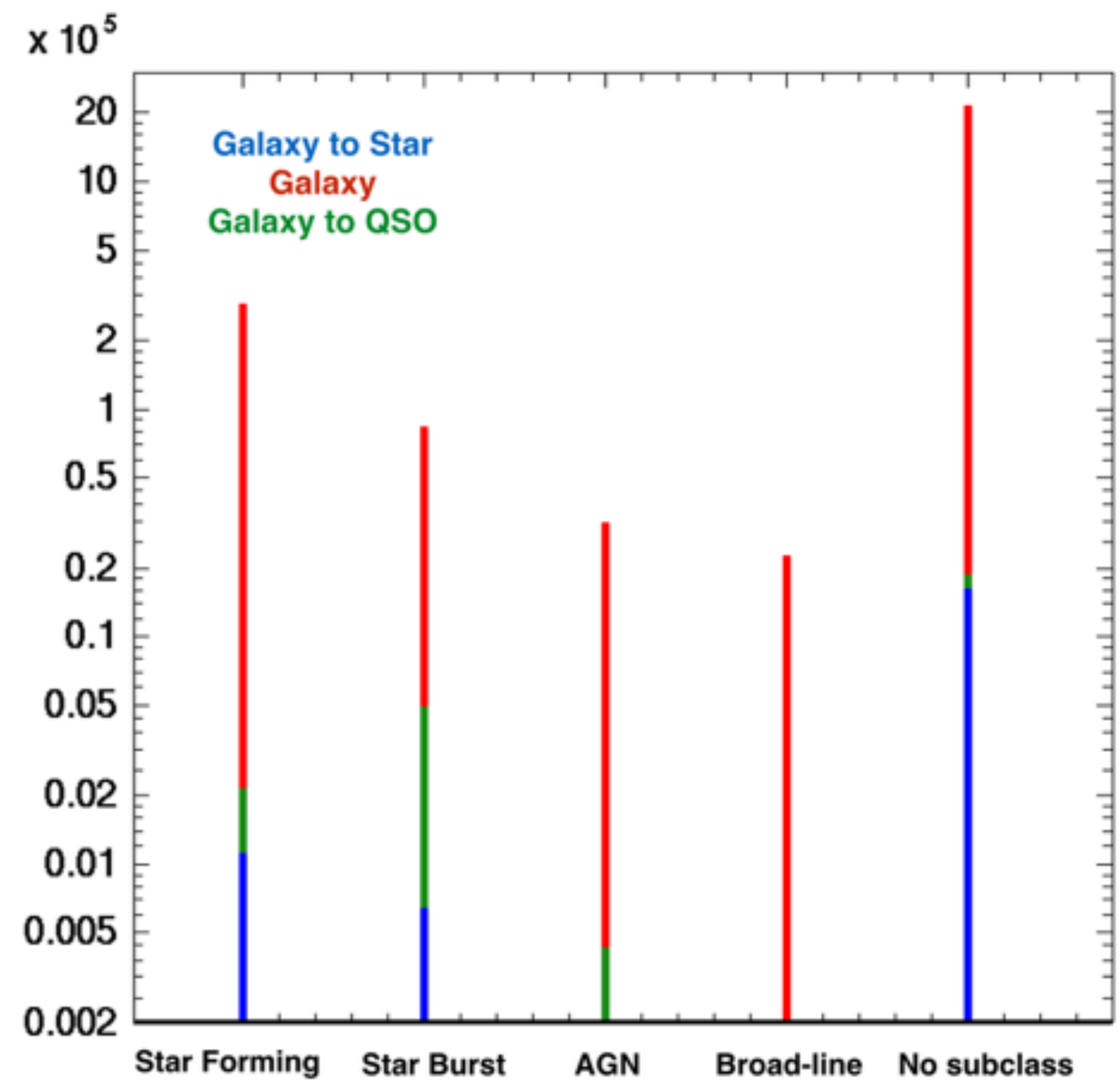
Legacy



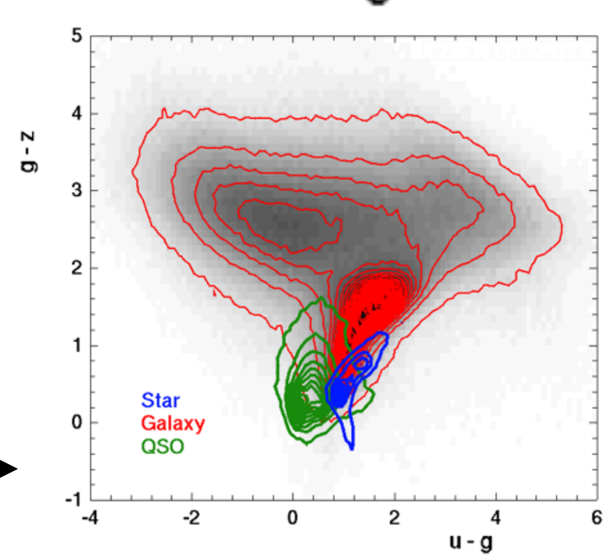
Galaxy misclassifications



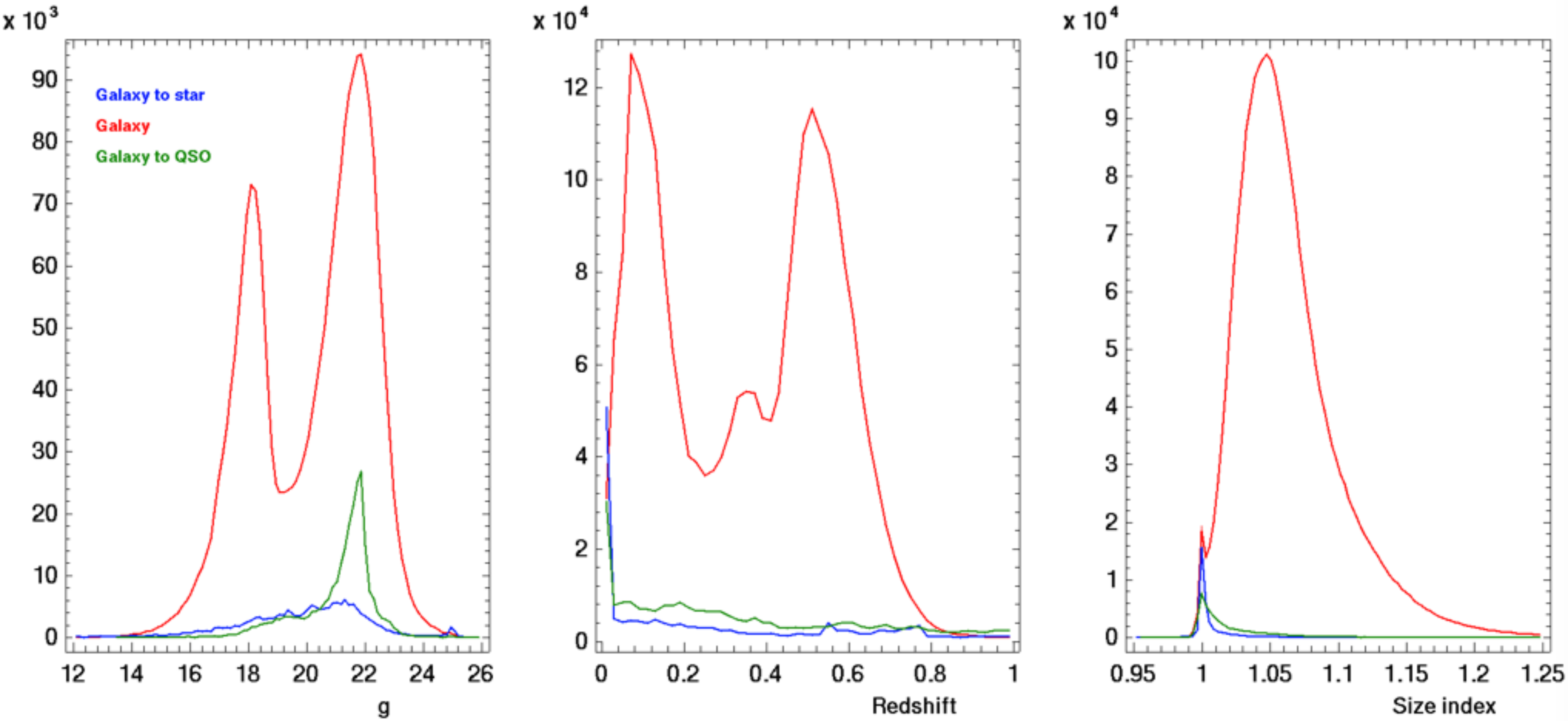
Misclassified sub-classes



Classified types
vs
spectroscopic types

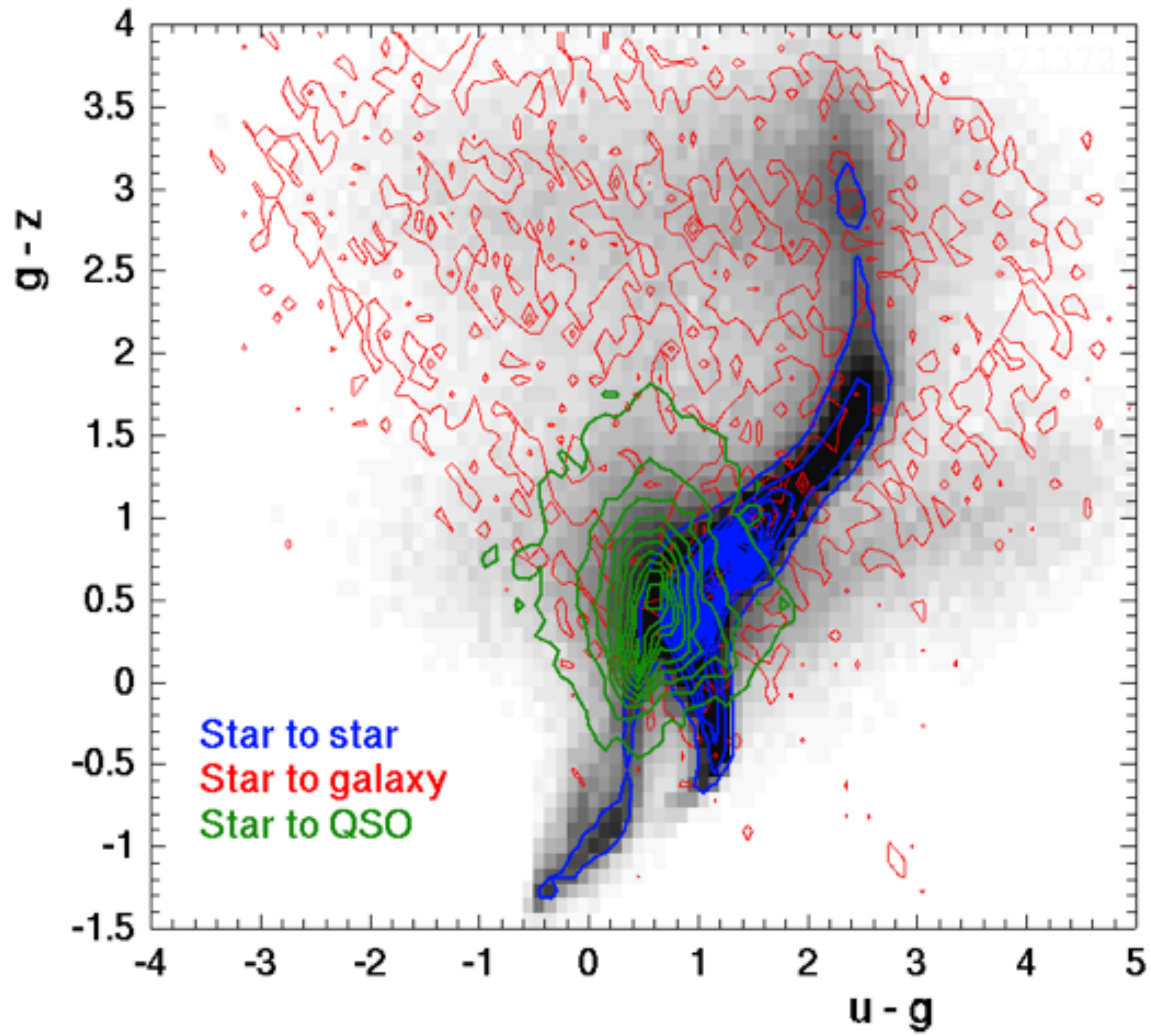


Galaxy misclassifications



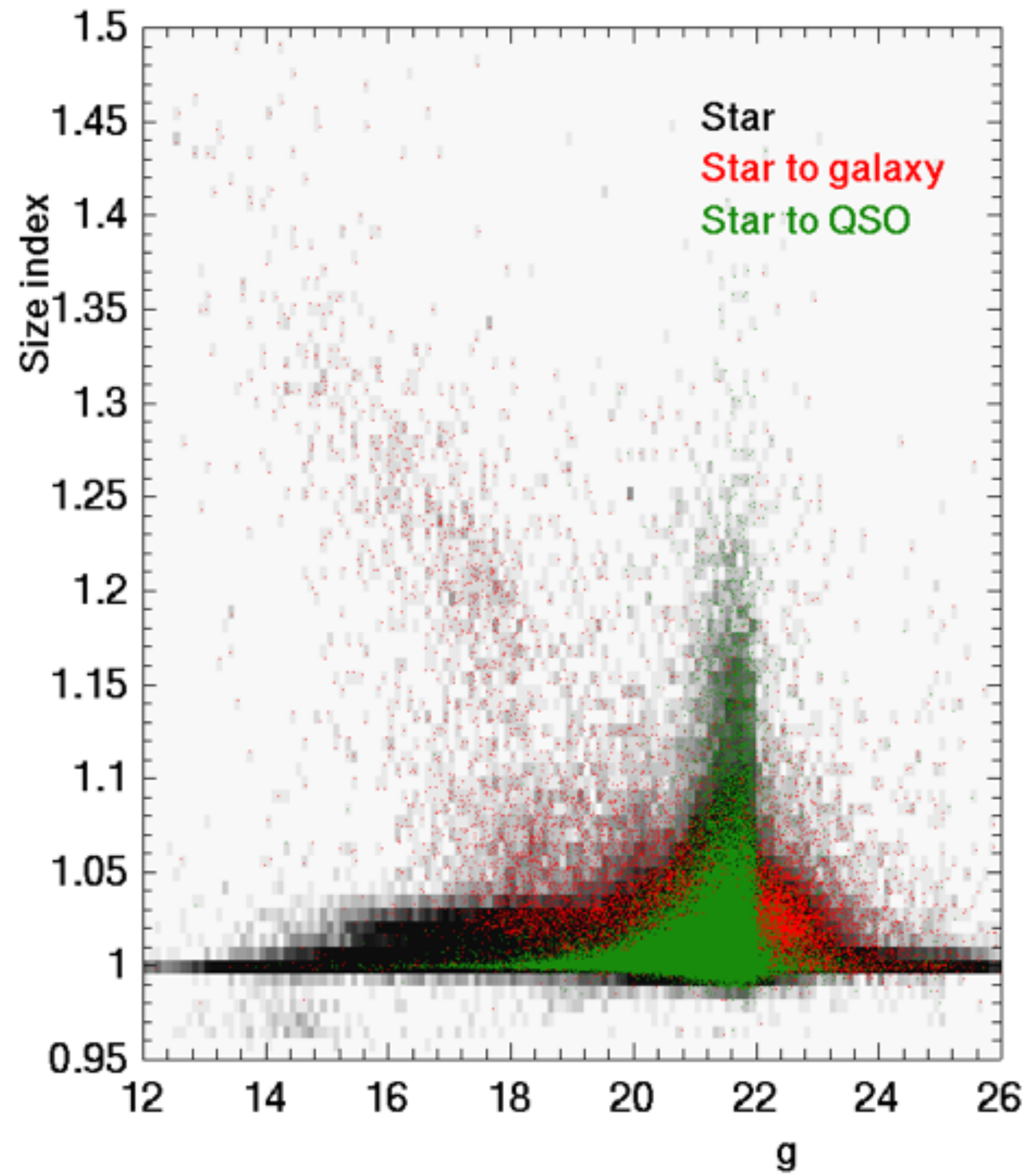
Faint nearby galaxies ==> point-like sources

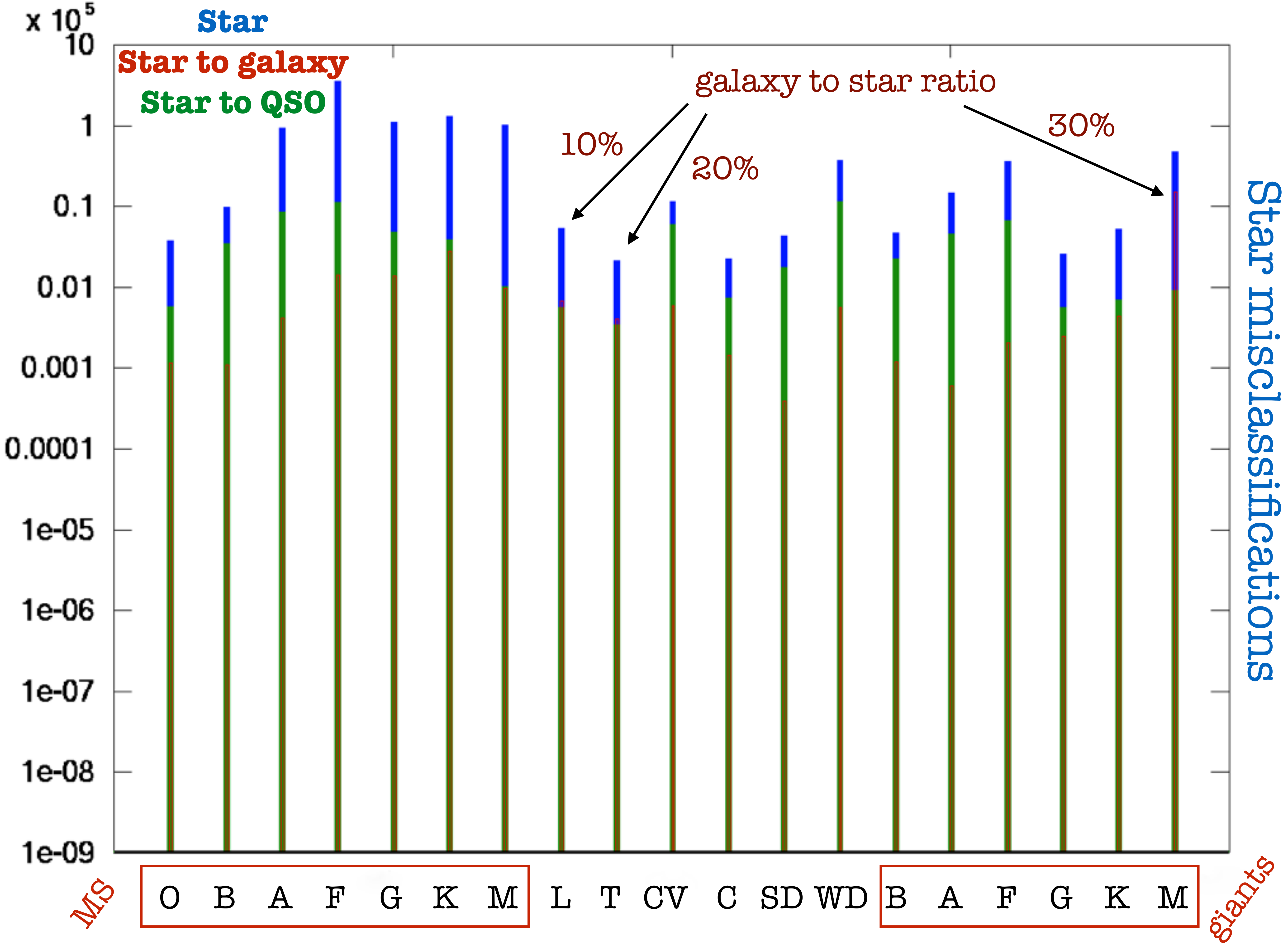
Star misclassifications



stars with scattered colours contaminates galaxy sample

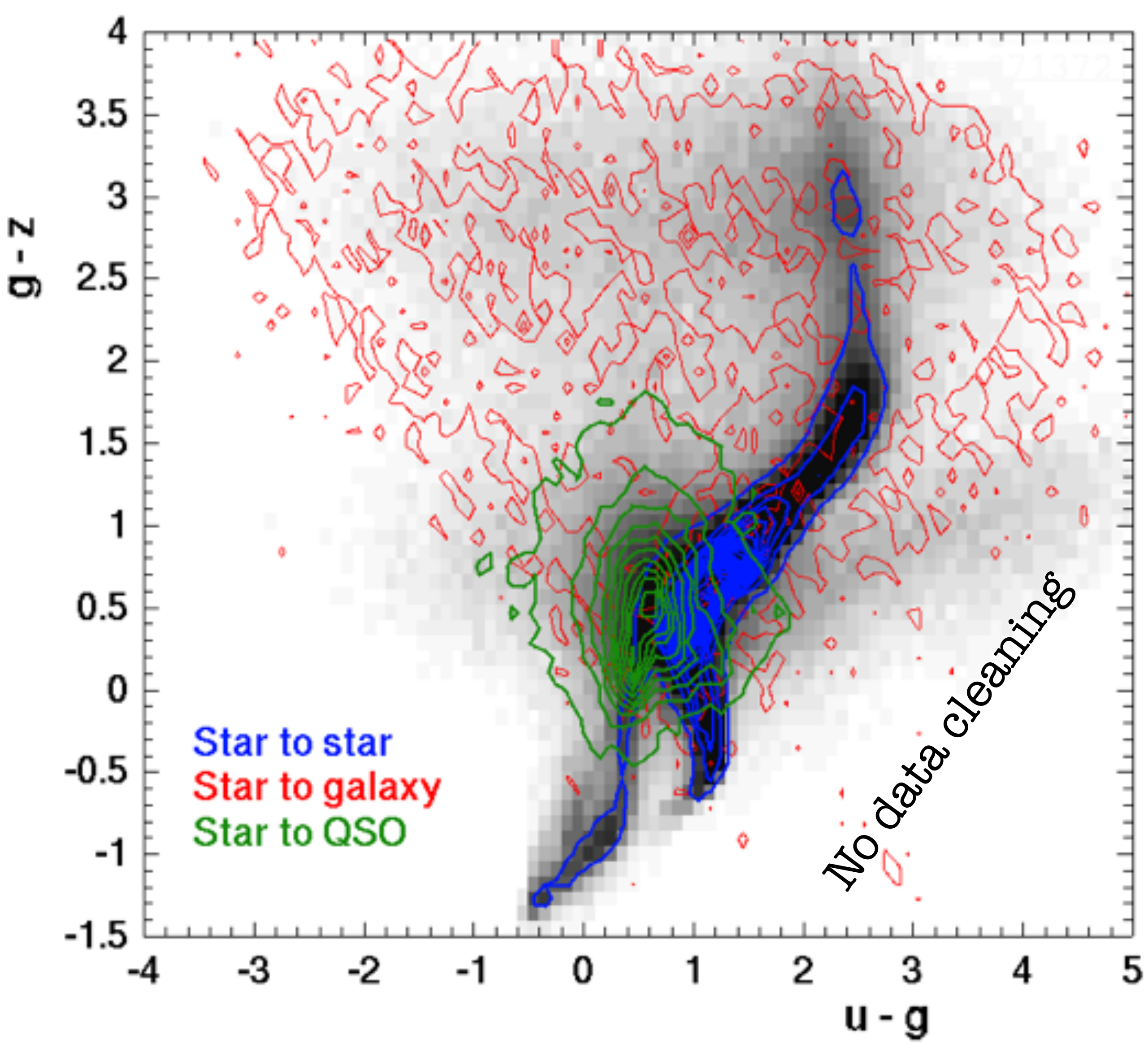
deviation from point-like source for faint stars

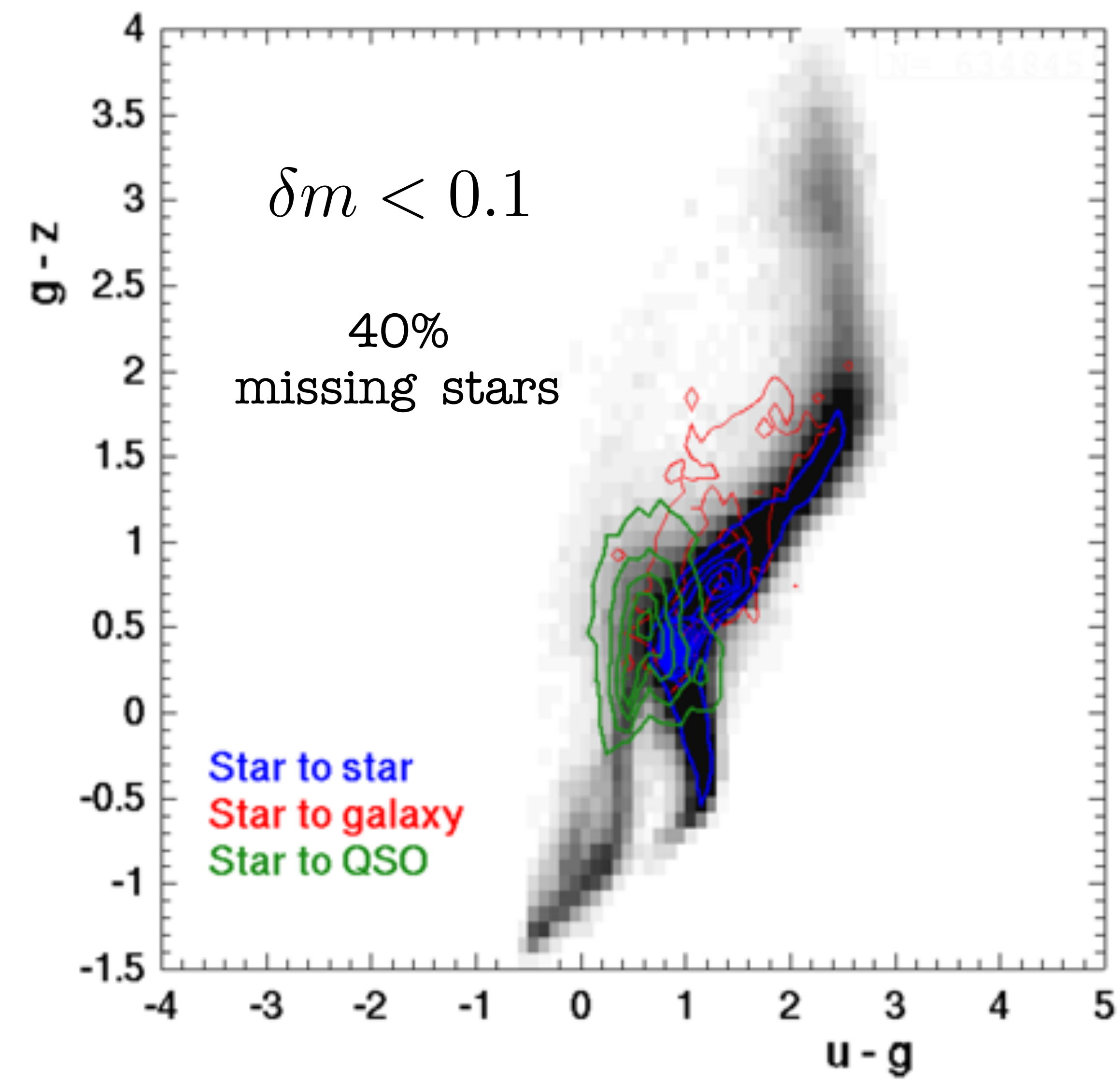



x 10^5
Star
Star to galaxy
Star to QSO
galaxy to star ratio
10%
20%
30%
Star misclassifications
10
1
0.1
0.01
0.001
0.0001
1e-05
1e-06
1e-07
1e-08
1e-09
MS
O B A F G K M
L T CV C SD WD
B A F G K M
giants

Star misclassifications

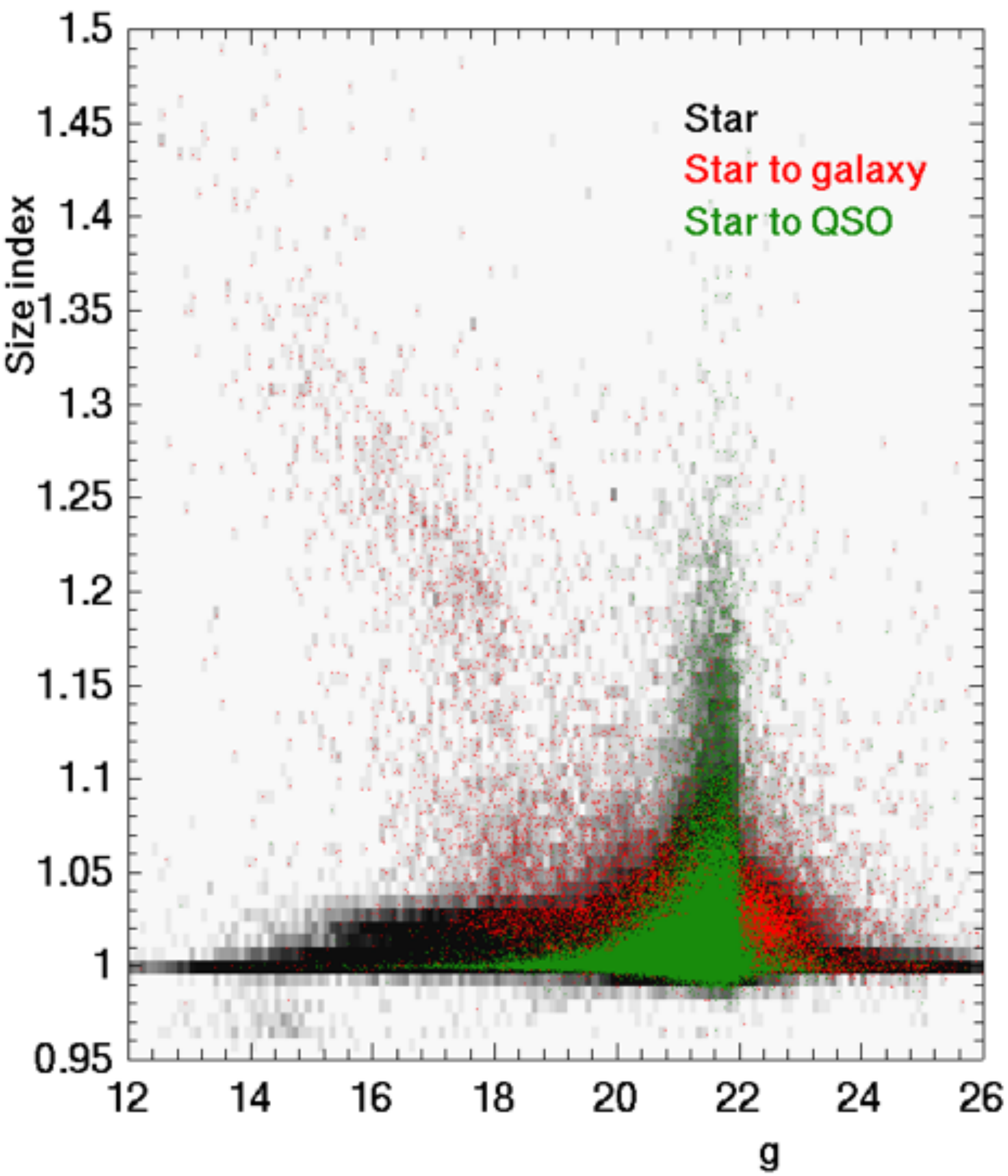
Photometric quality affects the colour measurement accuracy

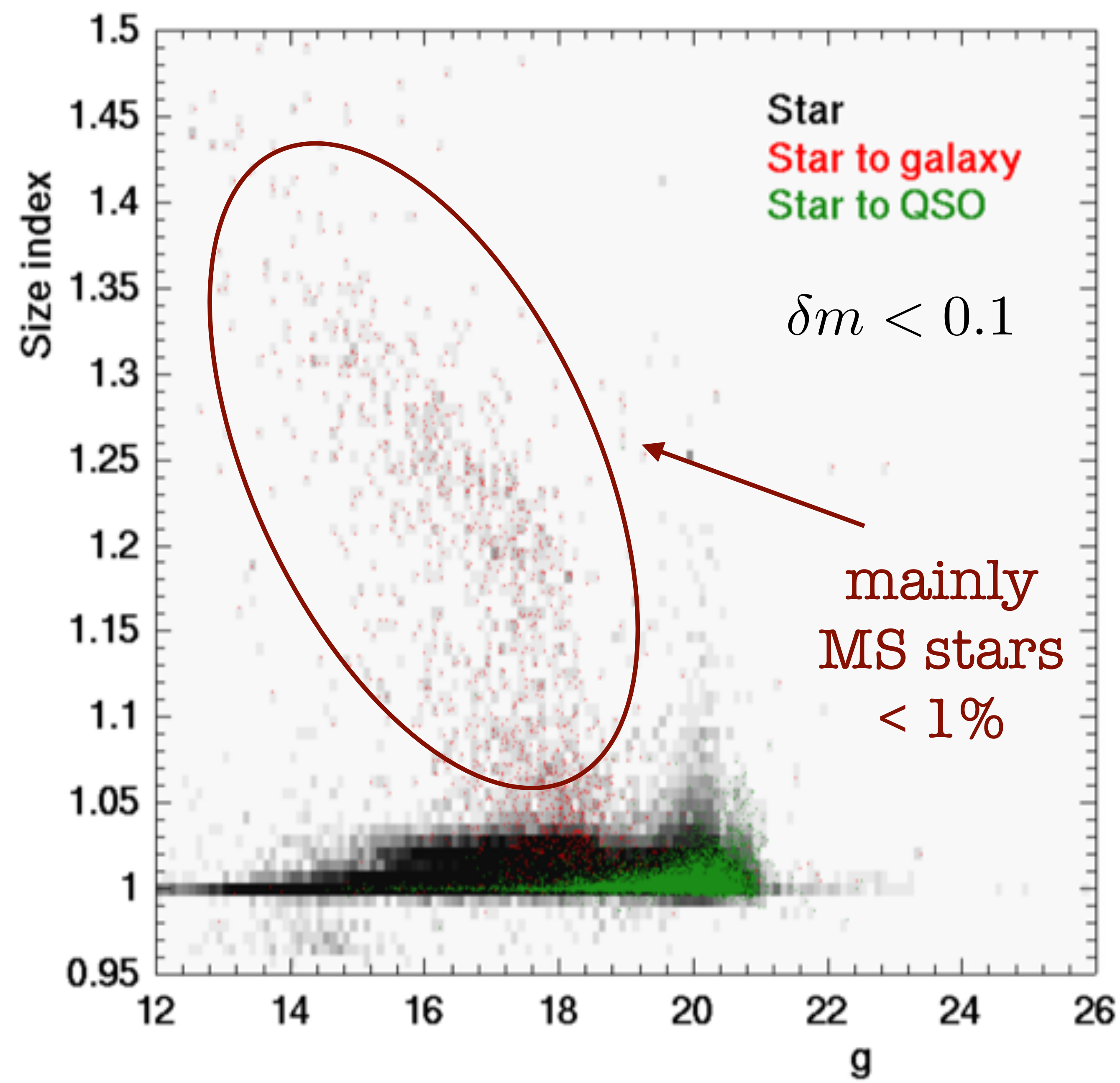




	Star	Galaxy	QSO	Total
Efficiency	97.1%	97.6%	93.8%	97%
Purity	97.5%	98.9%	94.9%	-

Star misclassifications





Summary

- Colour indices and apparent angular size can separates galaxies from stars and QSOs
- For SDSS DR12 a 4-layer MLP separates galaxies from the point-like sources by precision better than 98%
- Observational strategy with uniform sky coverage improves the classification efficiency
- Faint nearby galaxies can be misclassified as point-like sources while redshifted galaxies tend less to be misclassified
- M-giant stars, faint red L and T stars mainly contaminate the classified galaxy sample

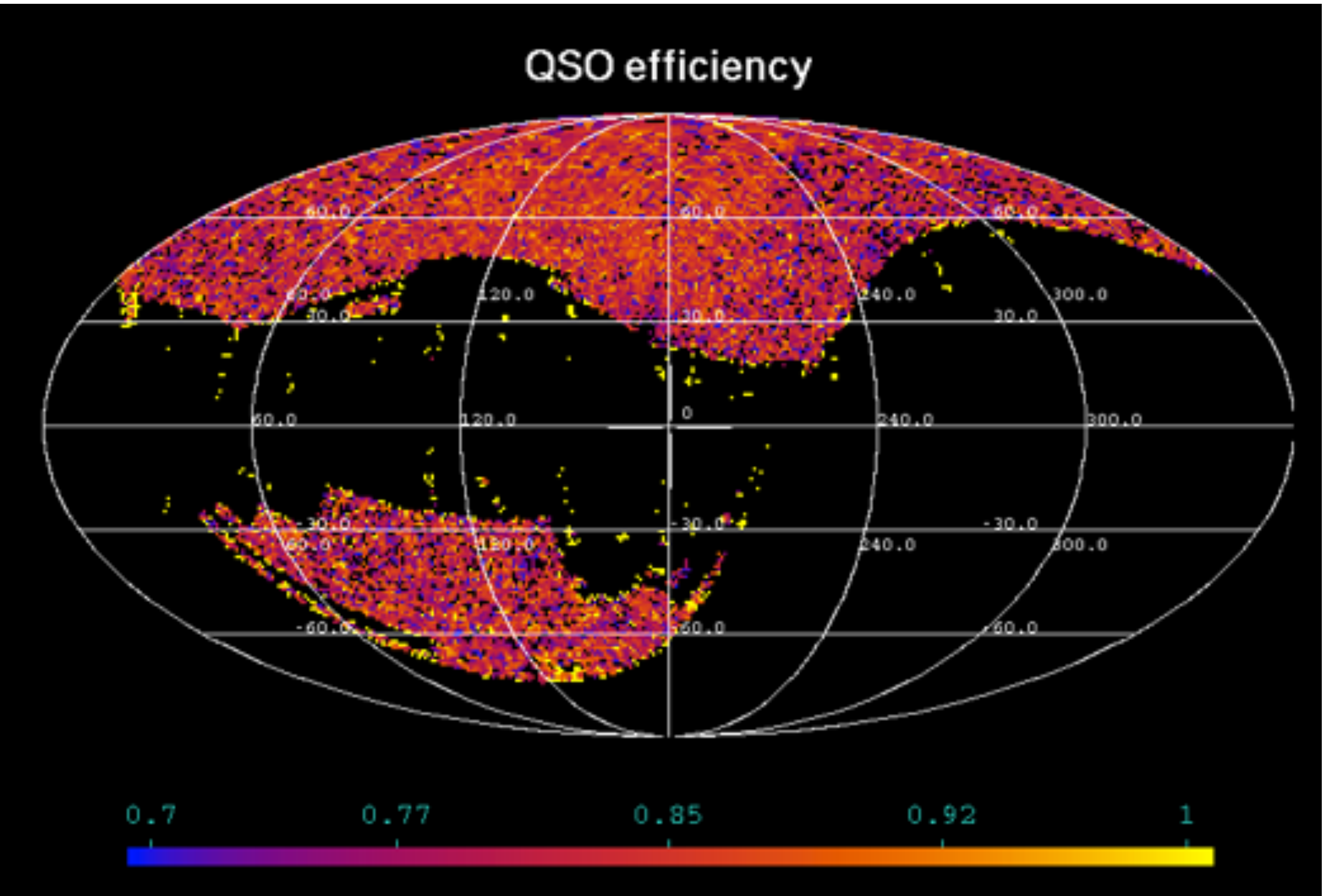
Classification for the LSST

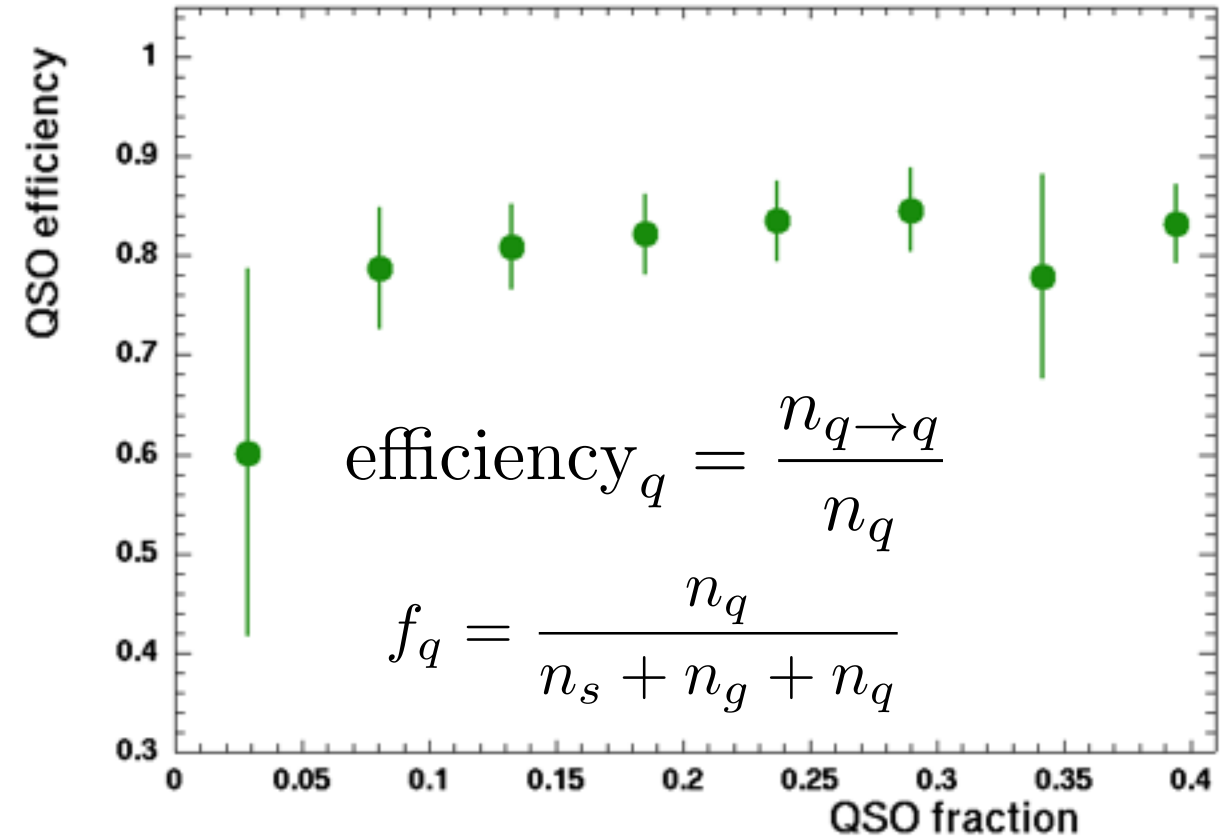
- Generating different galaxy types according to their luminosity function and the LSST apparent magnitude limits

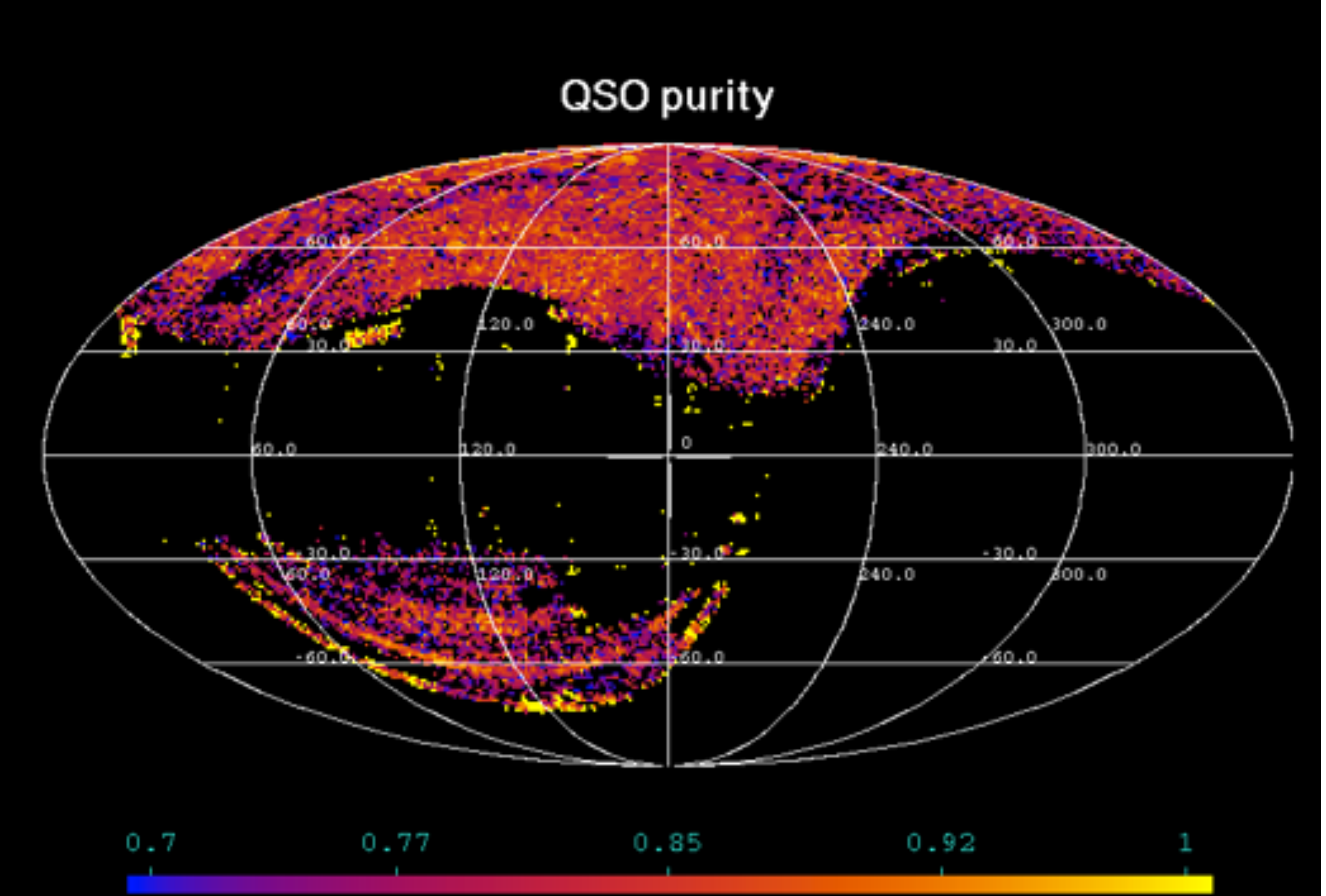
- Simulating the colour indices according to galaxy redshifted SEDs and LSST pass-band filters

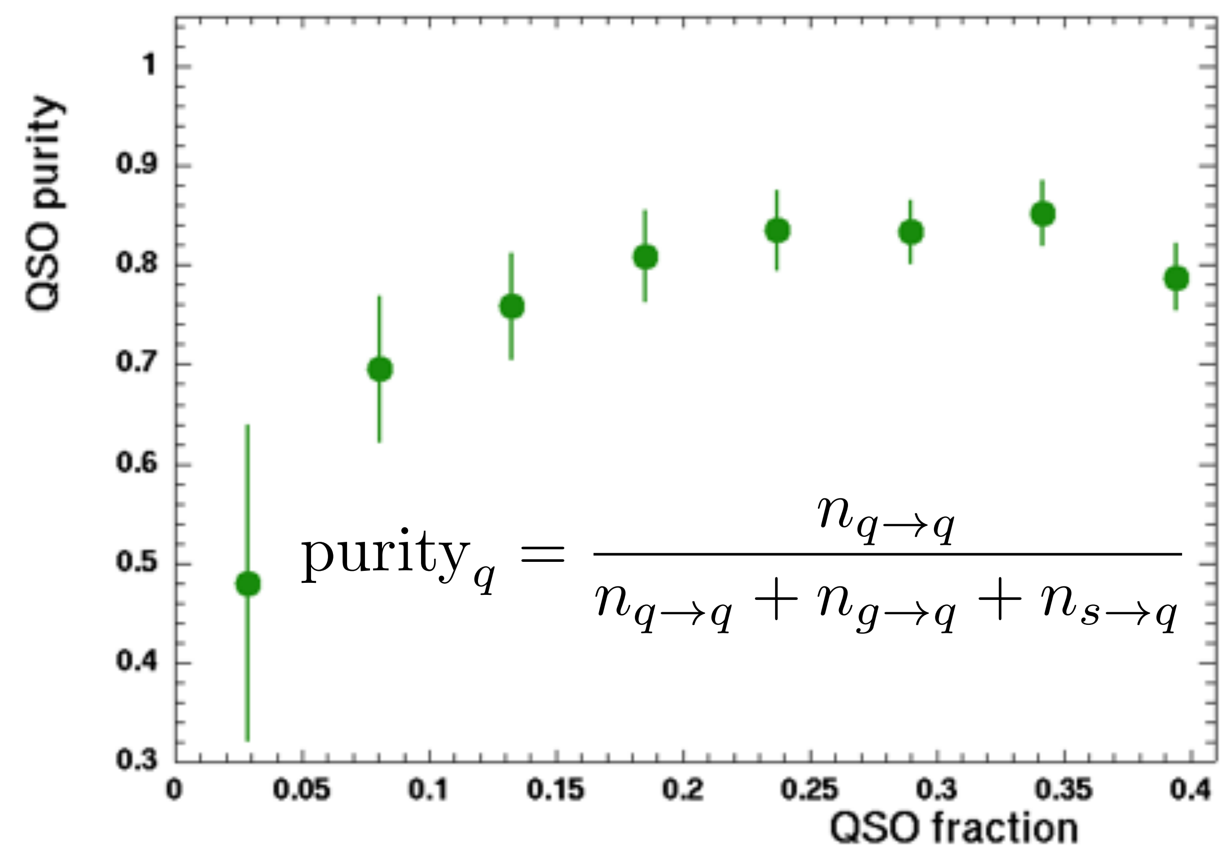
- Including the stars

Backups

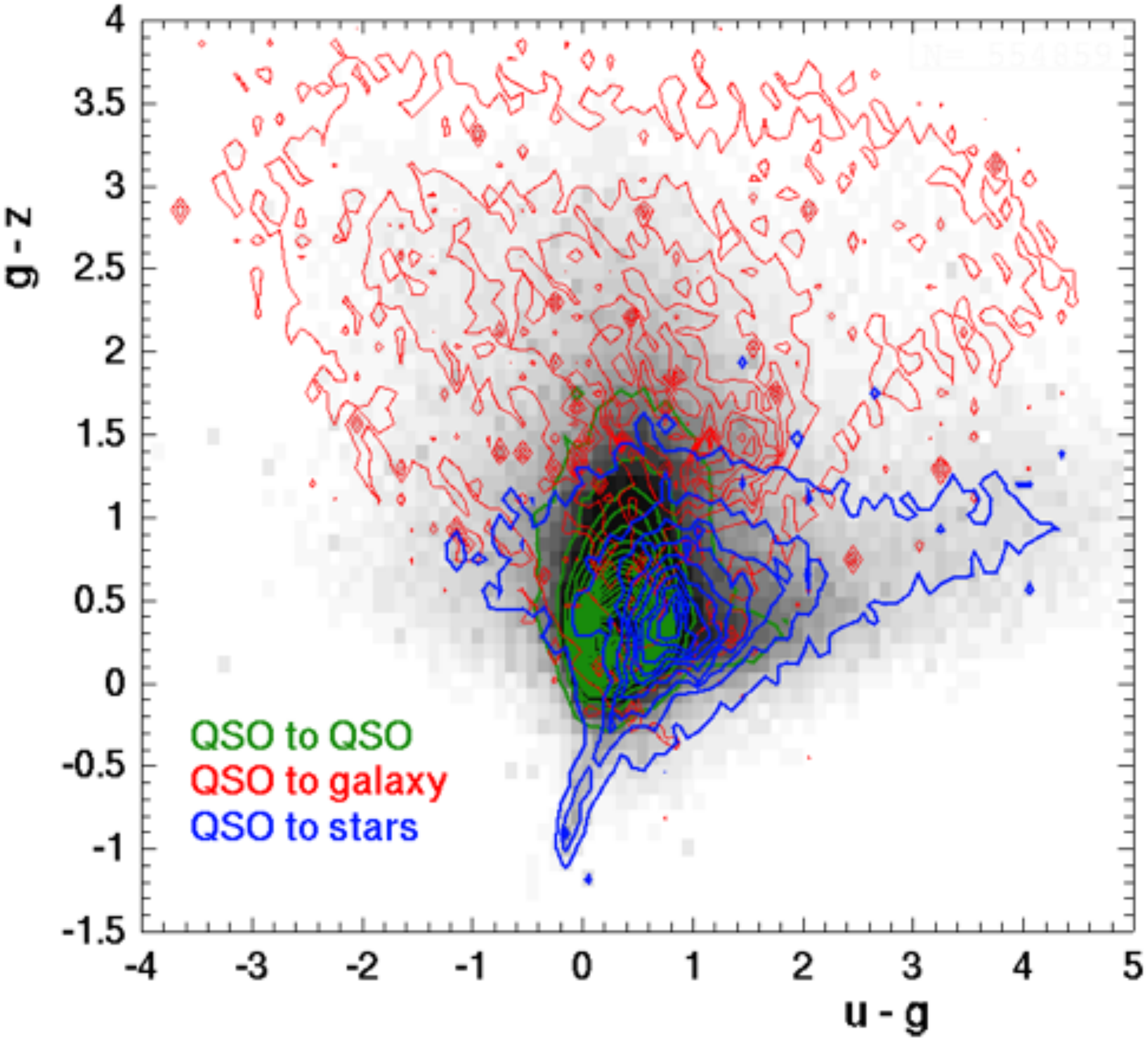

QSO efficiency
0.7
0.77
0.85
0.92
1

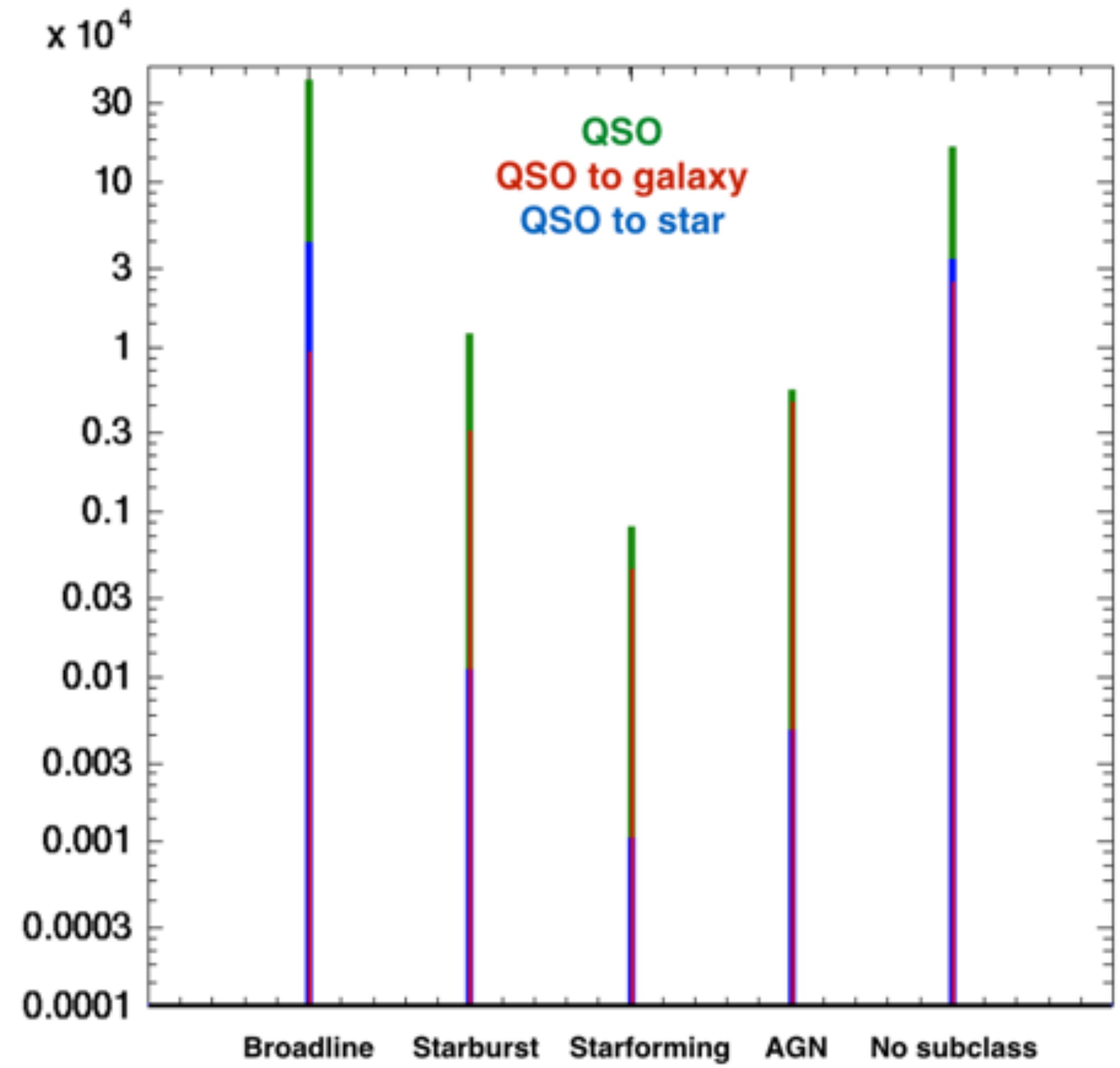

QSO efficiency
QSO fraction
$\mathrm{efficiency}_q = \frac{n_{q\to q}}{n_q}$
$f_q = \frac{n_q}{n_s + n_g + n_q}$


QSO purity
0.7
0.77
0.85
0.92
1

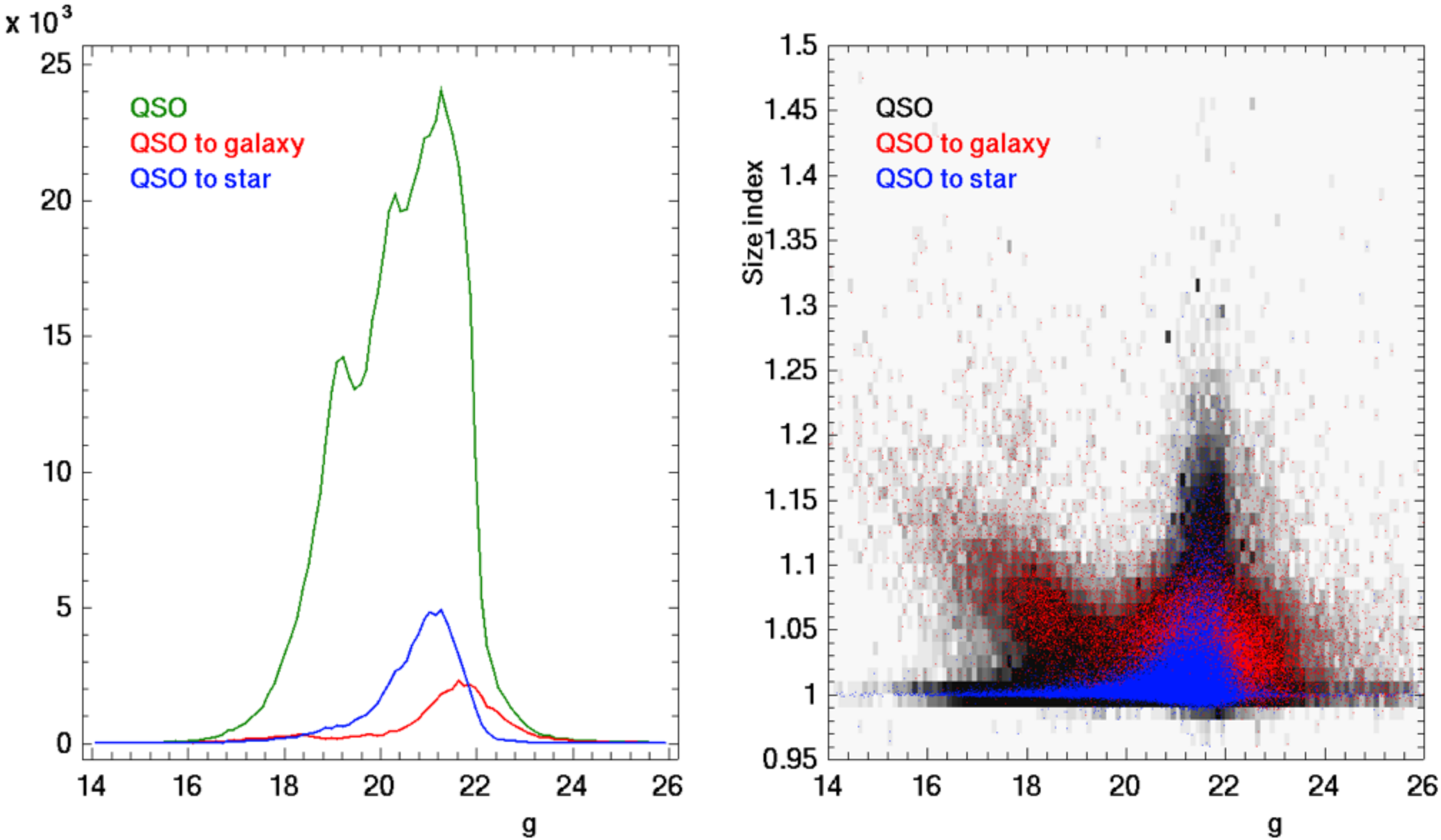

QSO purity
QSO fraction
$\mathrm{purity}_q = \frac{n_{q\to q}}{n_{q\to q} + n_{g\to q} + n_{s\to q}}$

QSOs misclassifications

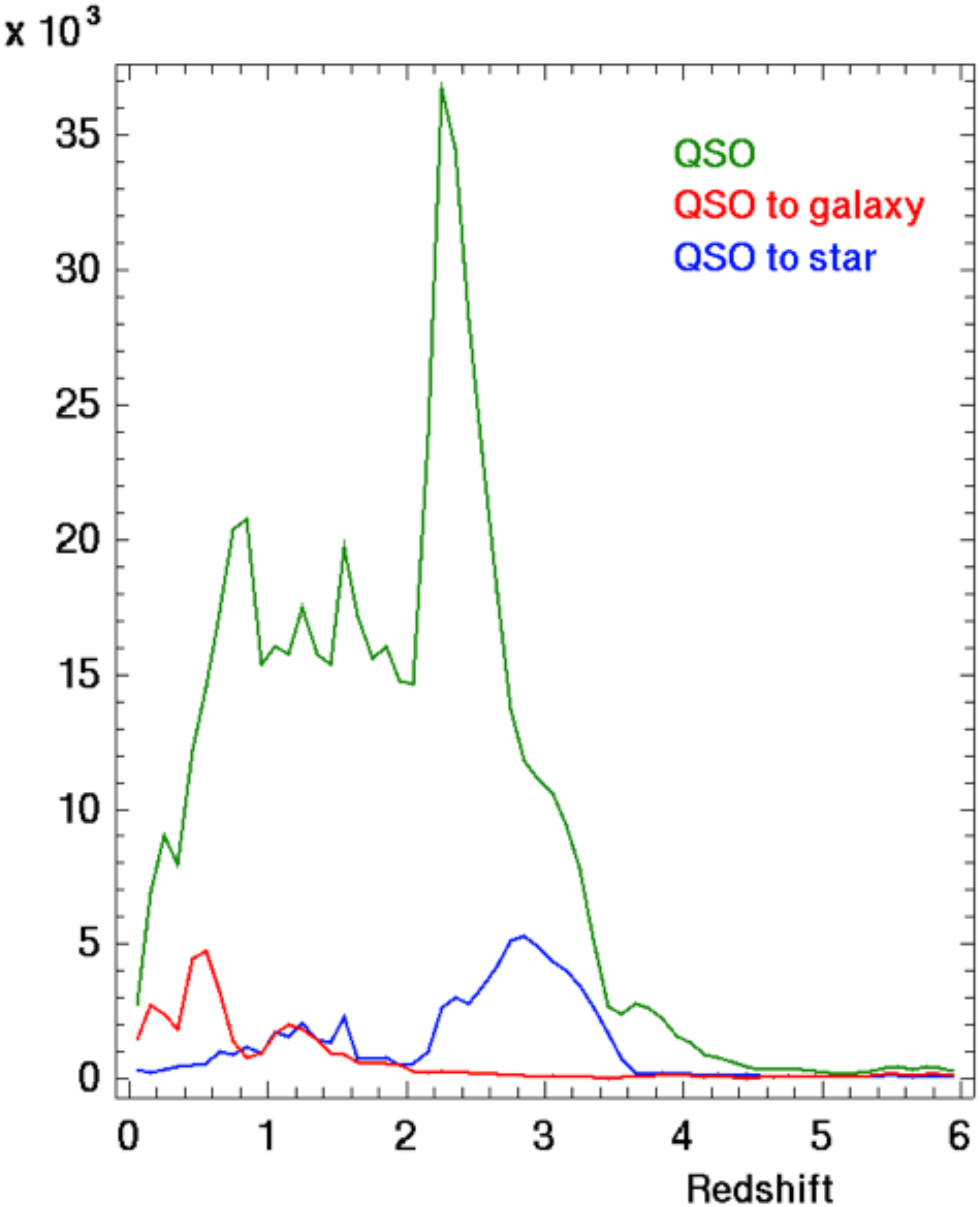


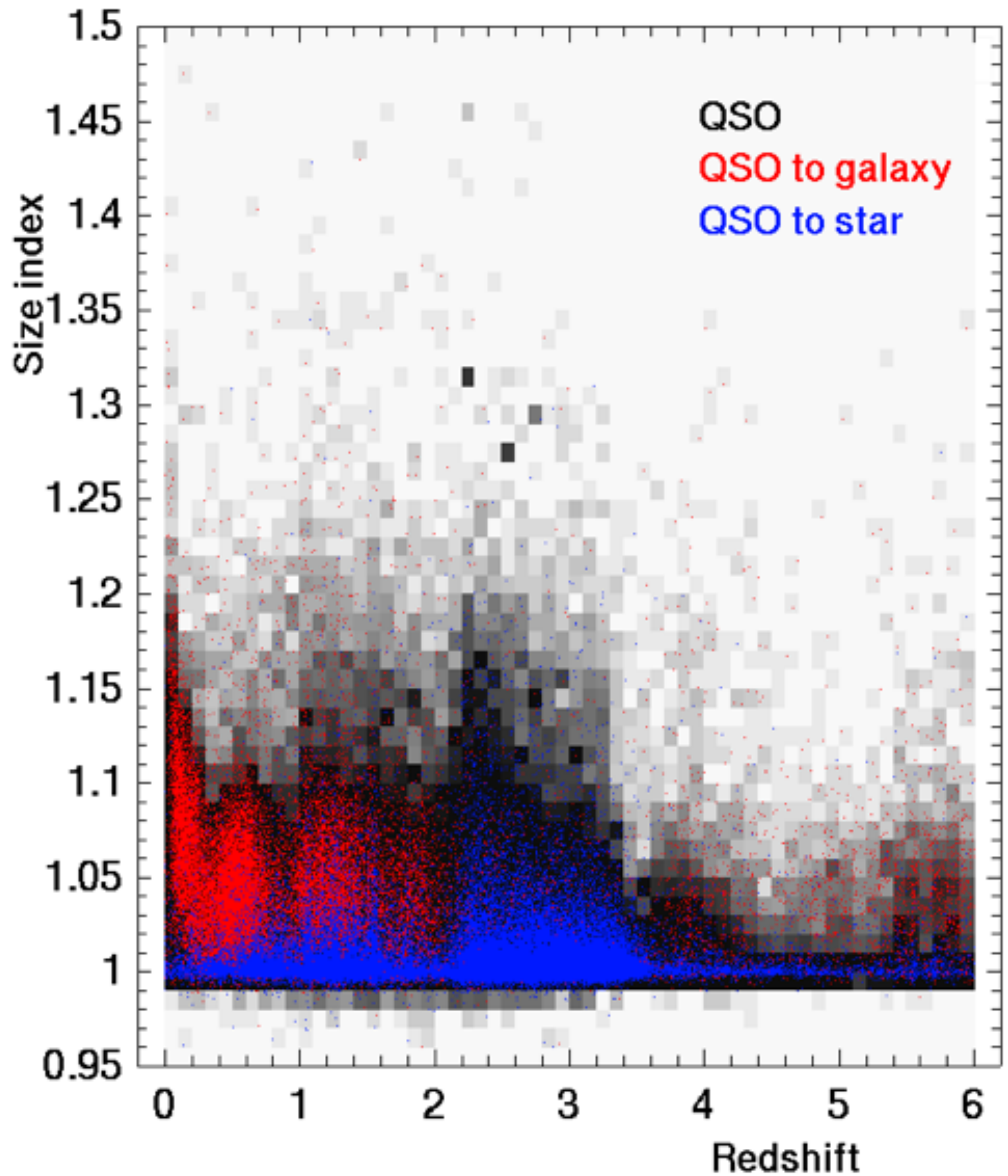


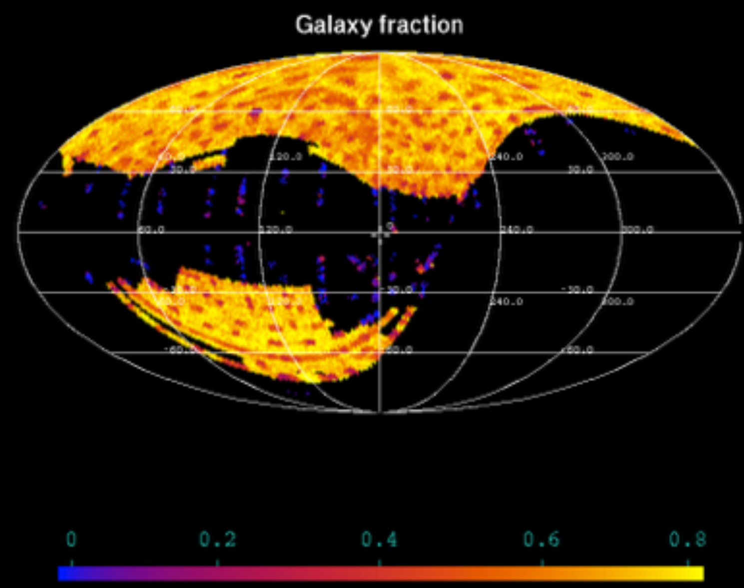
QSOs misclassifications



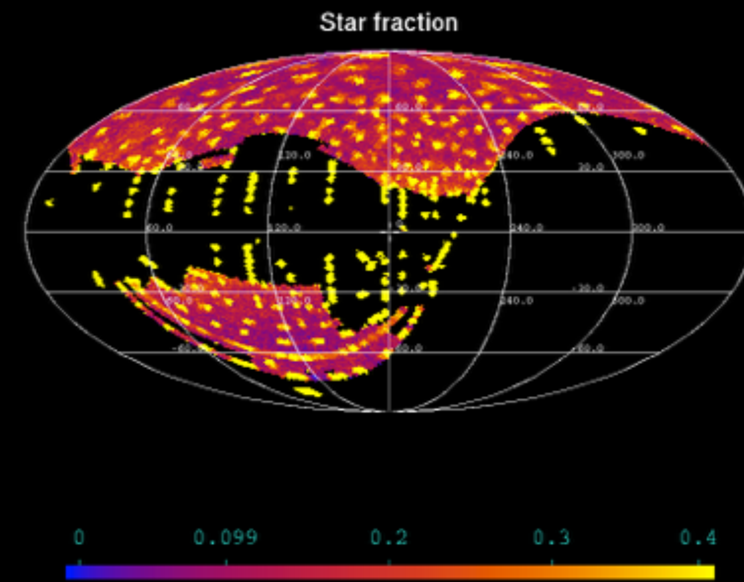
QSOs misclassifications

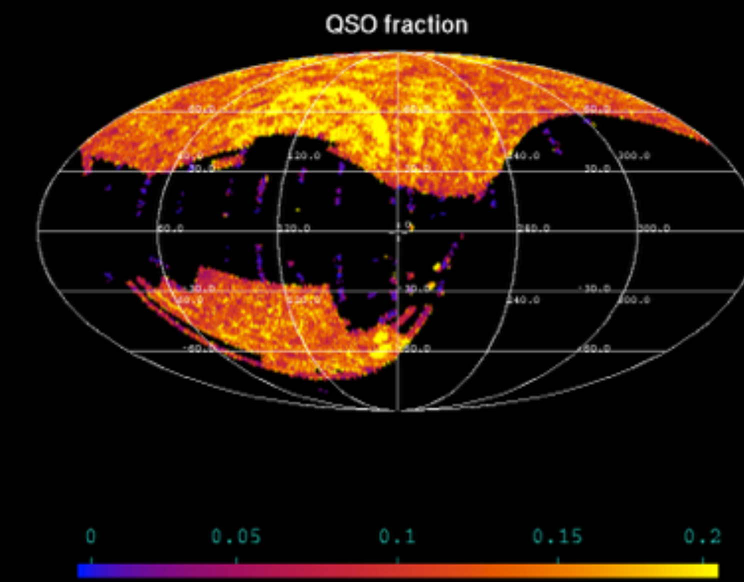






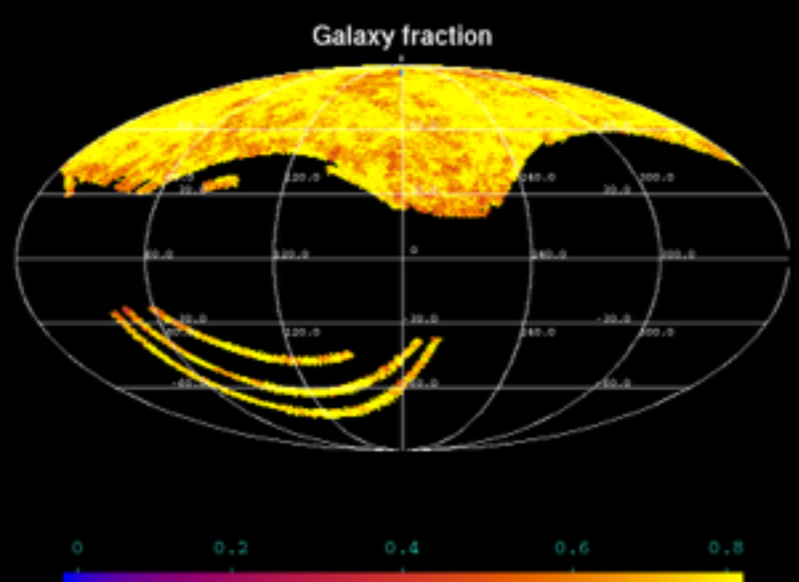
DR 12

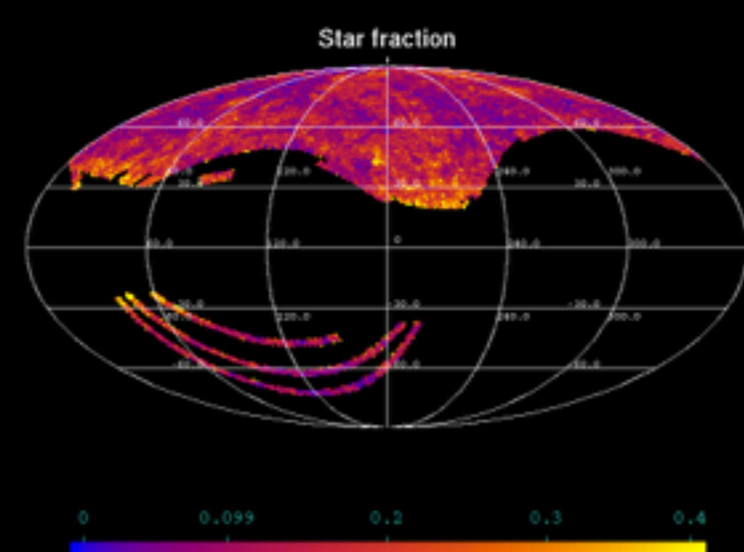




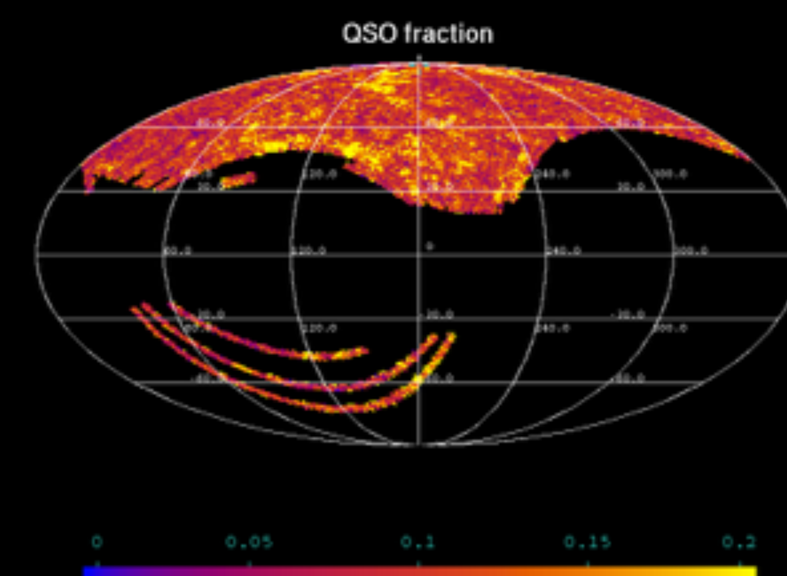
	Stars	Galaxies	QSOs	Total
efficiency	95%	99%	90%	98%
purity	94%	99%	94%	
efficiency	90%	99%	90%	97%
purity	96%	98%	96%	

DR12
vs
Legacy

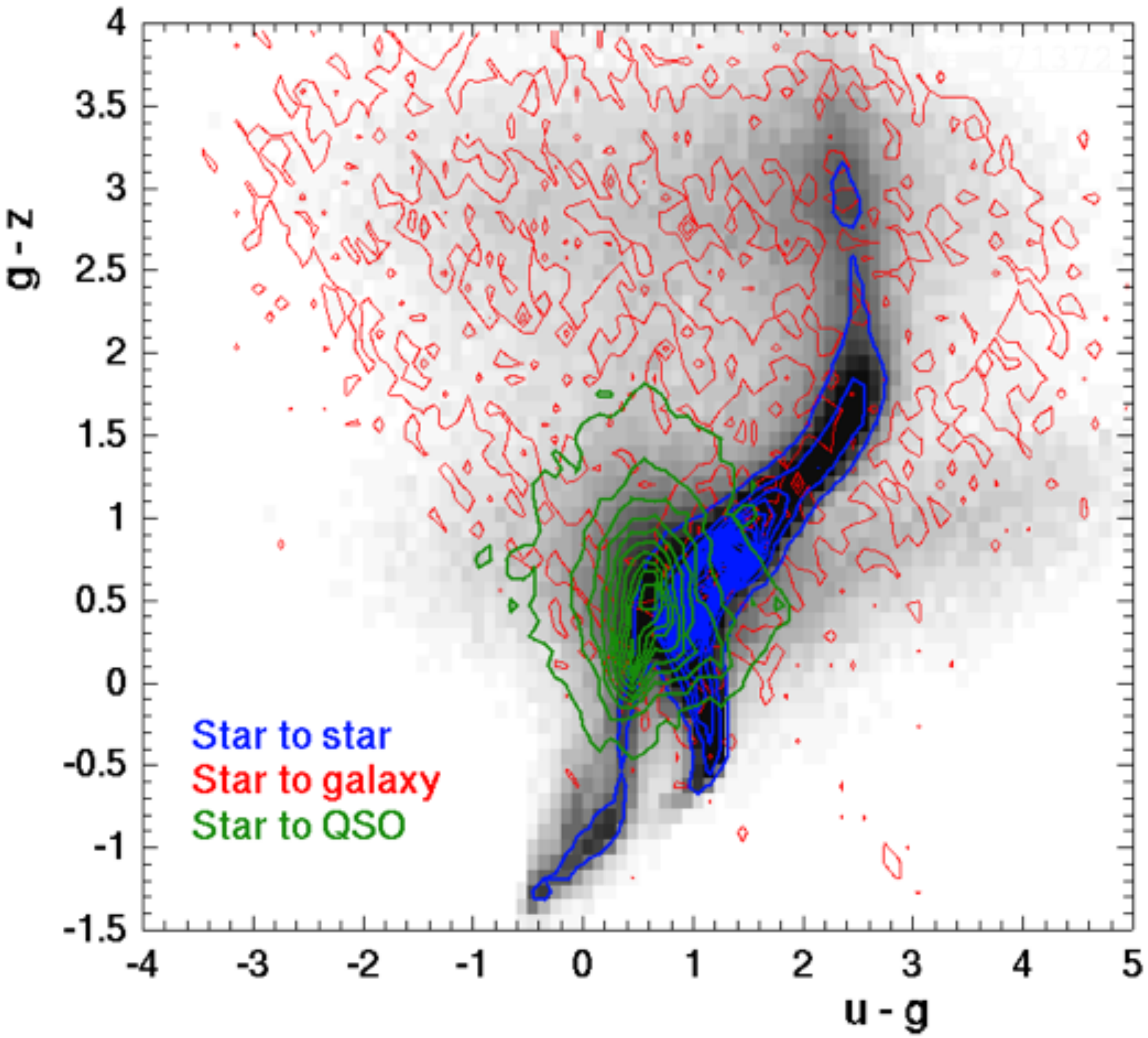




Legacy



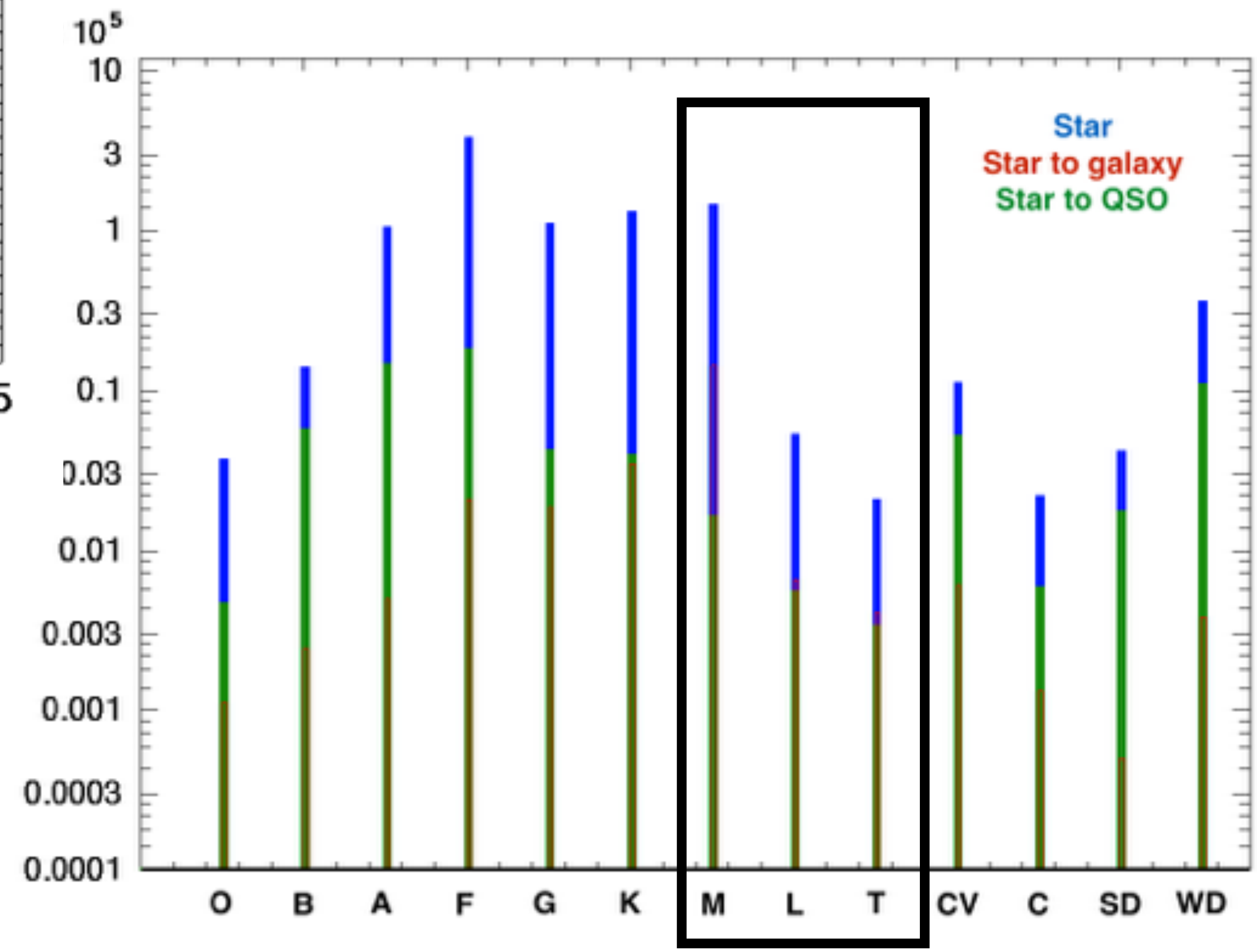
Star misclassifications



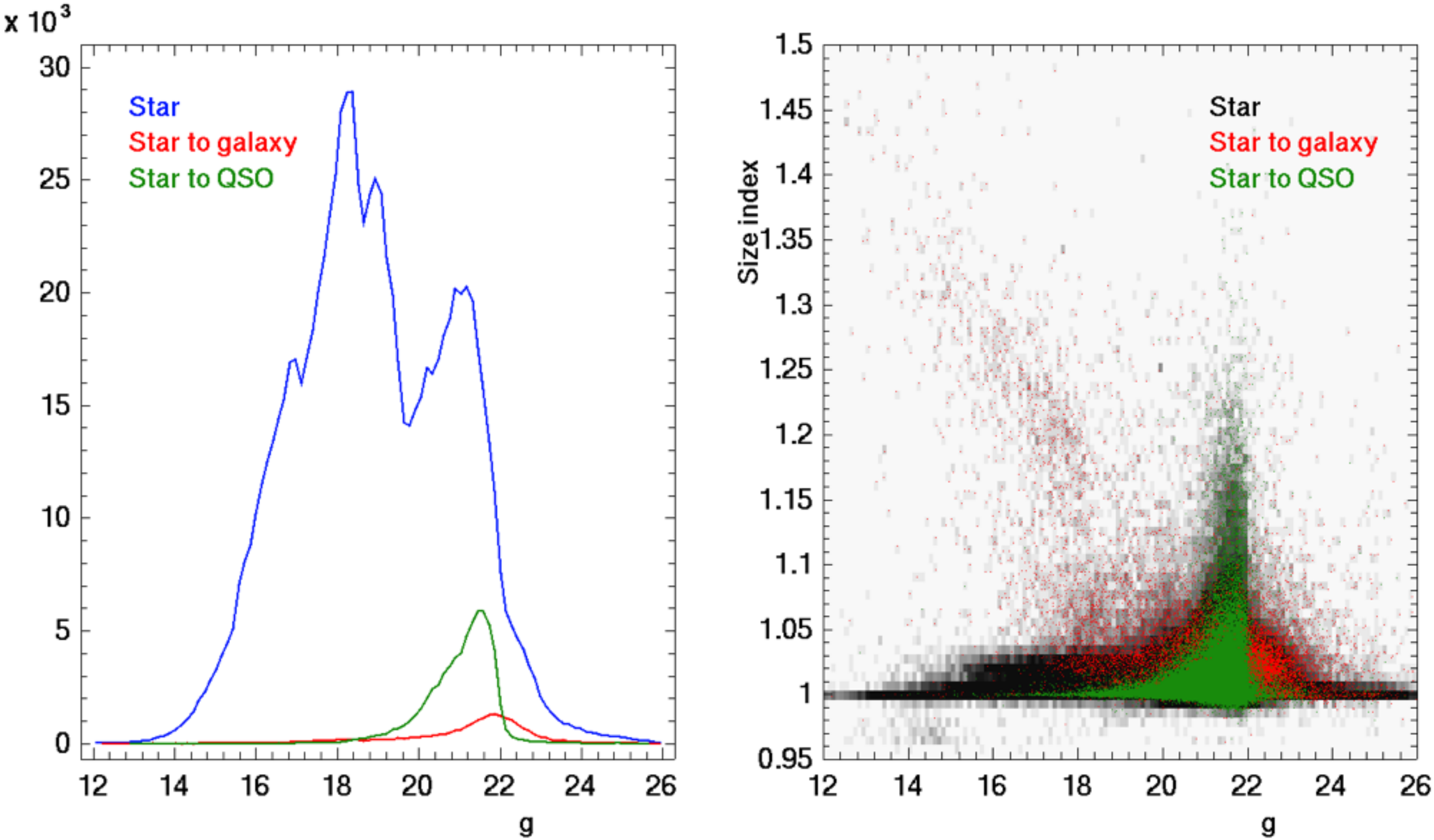
stars with
scattered colours
contaminates galaxy sample

Misclassified sub-classes

> 10% of M, L, T stars ==> galaxy



Star misclassifications



deviation from point-like source for faint stars