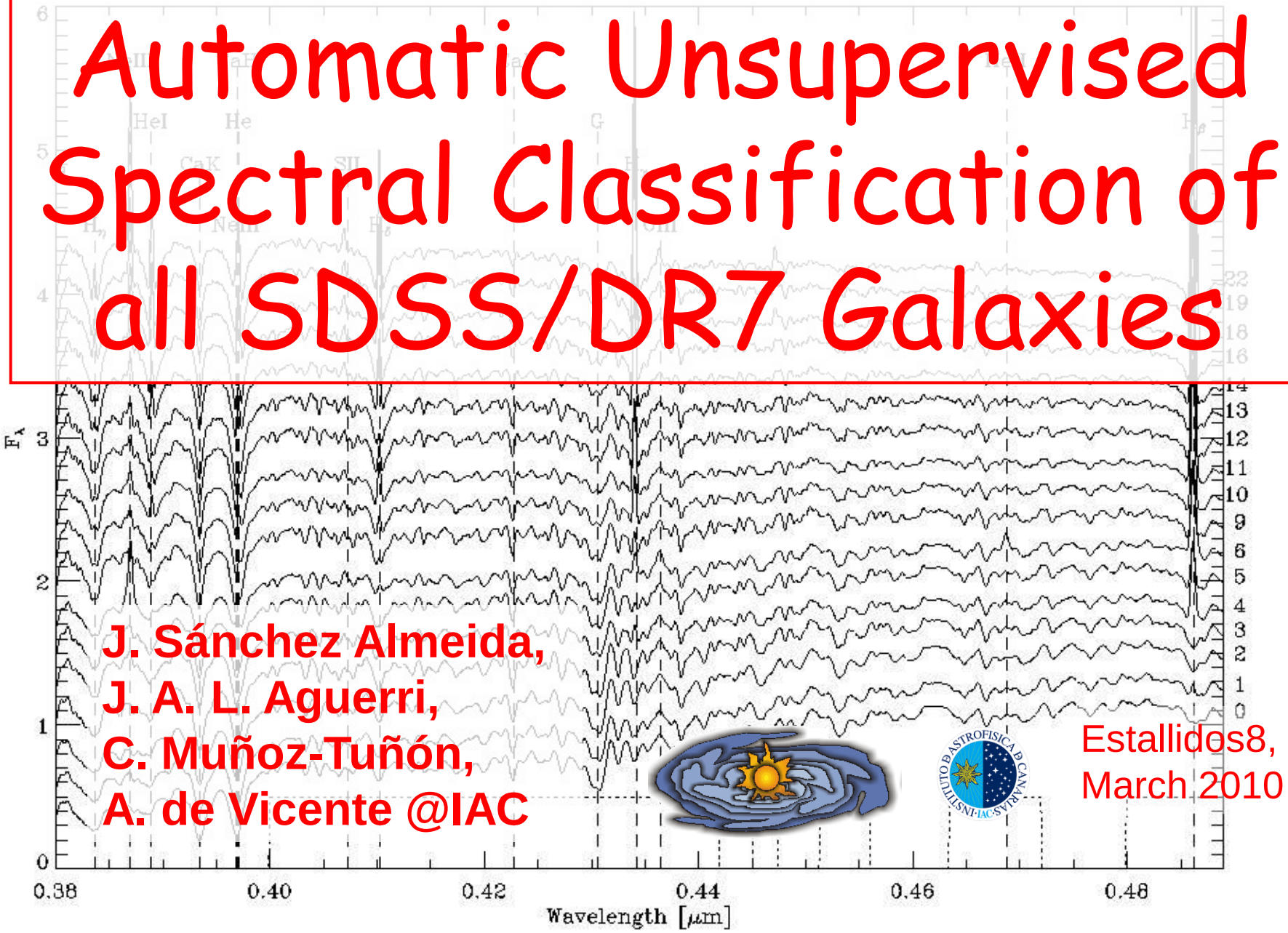
Automatic Unsupervised Spectral Classification of all SDSS/DR7 Galaxies



J. Sánchez Almeida,
J. A. L. Aguerri,
C. Muñoz-Tuñón,
A. de Vicente @IAC

Estallidos8,
March 2010

Summary

- Motivation
- The classification method: k-means clustering algorithm
- ASK classification of the full SDSS/DR7
- Properties of the classes
- Uses
- Conclusions

Motivation

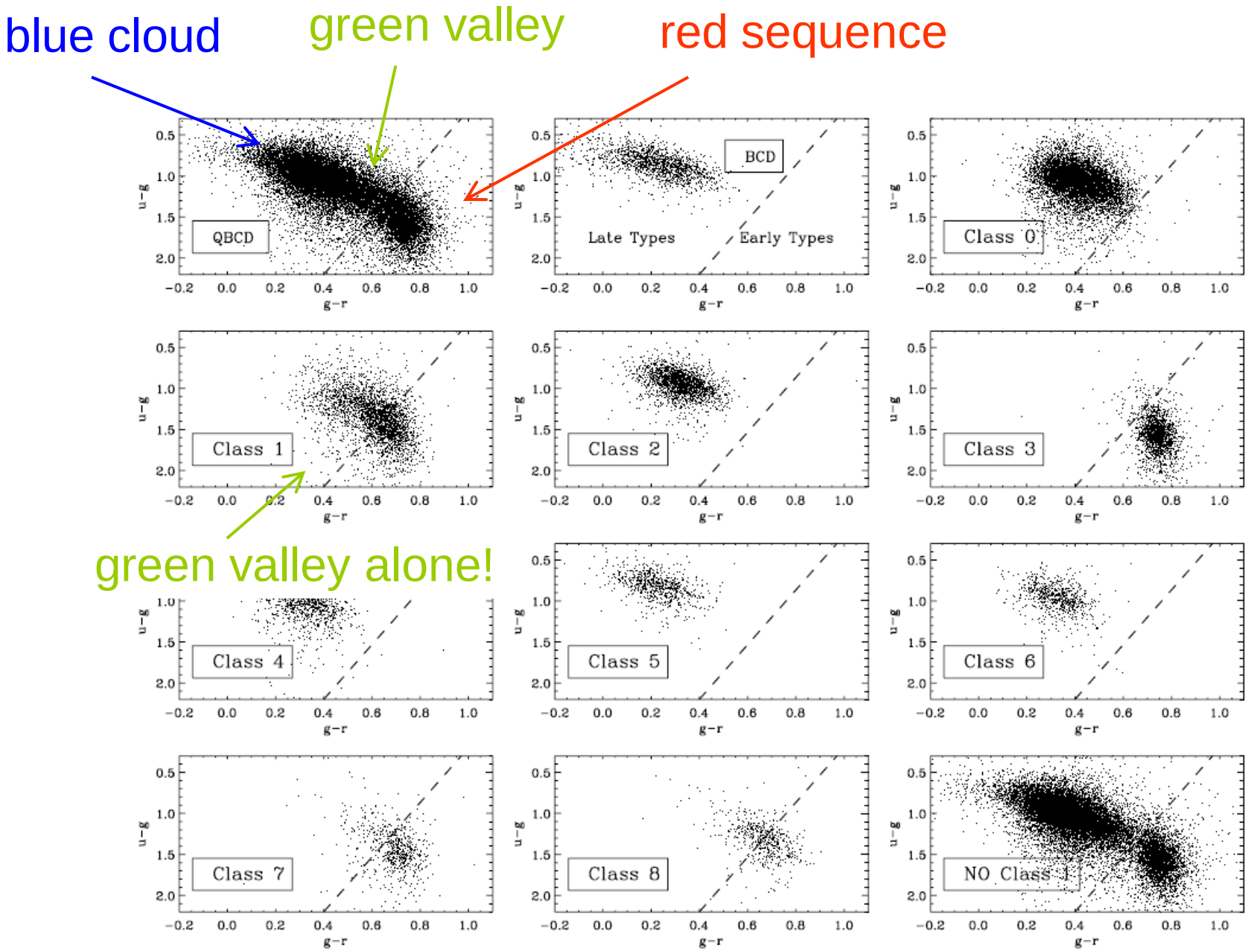
- Need to simplify

  The nebulae are so numerous that they cannot be studied individually. Therefore, it is necessary to know whether a fair sample can be assembled from the most conspicuous objects and, if so, the size of the sample required. (Hubble, 1936)

- spectral catalogs far more complete than ever now freely available (SDSS/DR7)

- ***k-means*** separates galaxies in the green valley

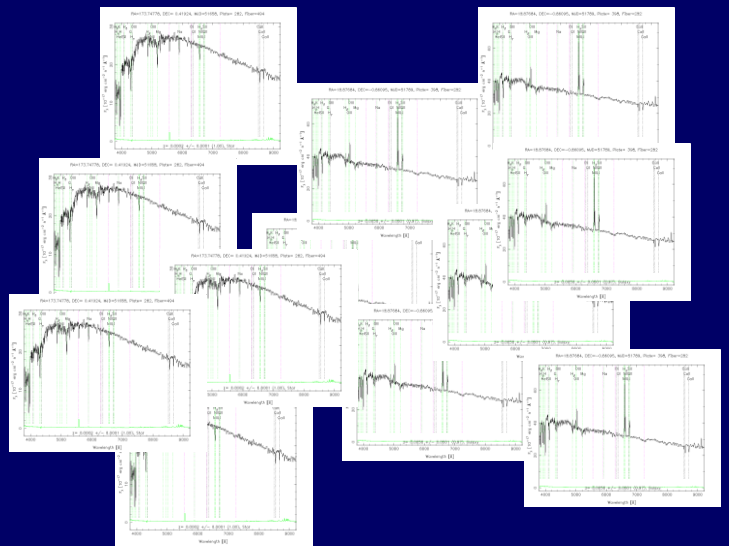
  In the local universe, galaxies come in two colors: red and blue (e.g., Balogh et al. 2004). They are loosely connected with Hubble types (E ≡ red, S ≡ blue)

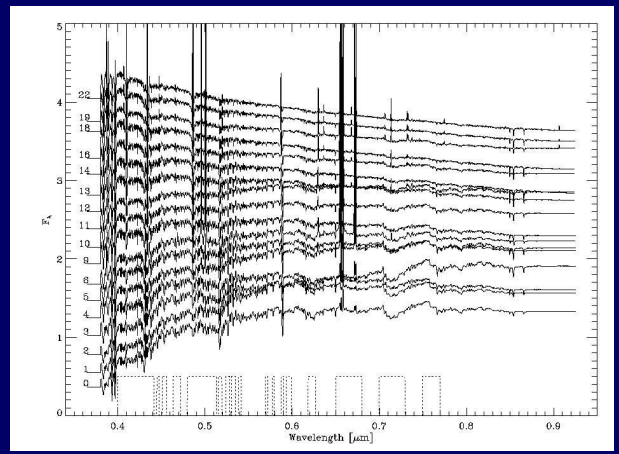

blue cloud
green valley
red sequence
green valley alone!
QBCD
BCD
Late Types
Early Types
Class 0
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8
NO Class 1
u−g
g−r

SA et al. 2009

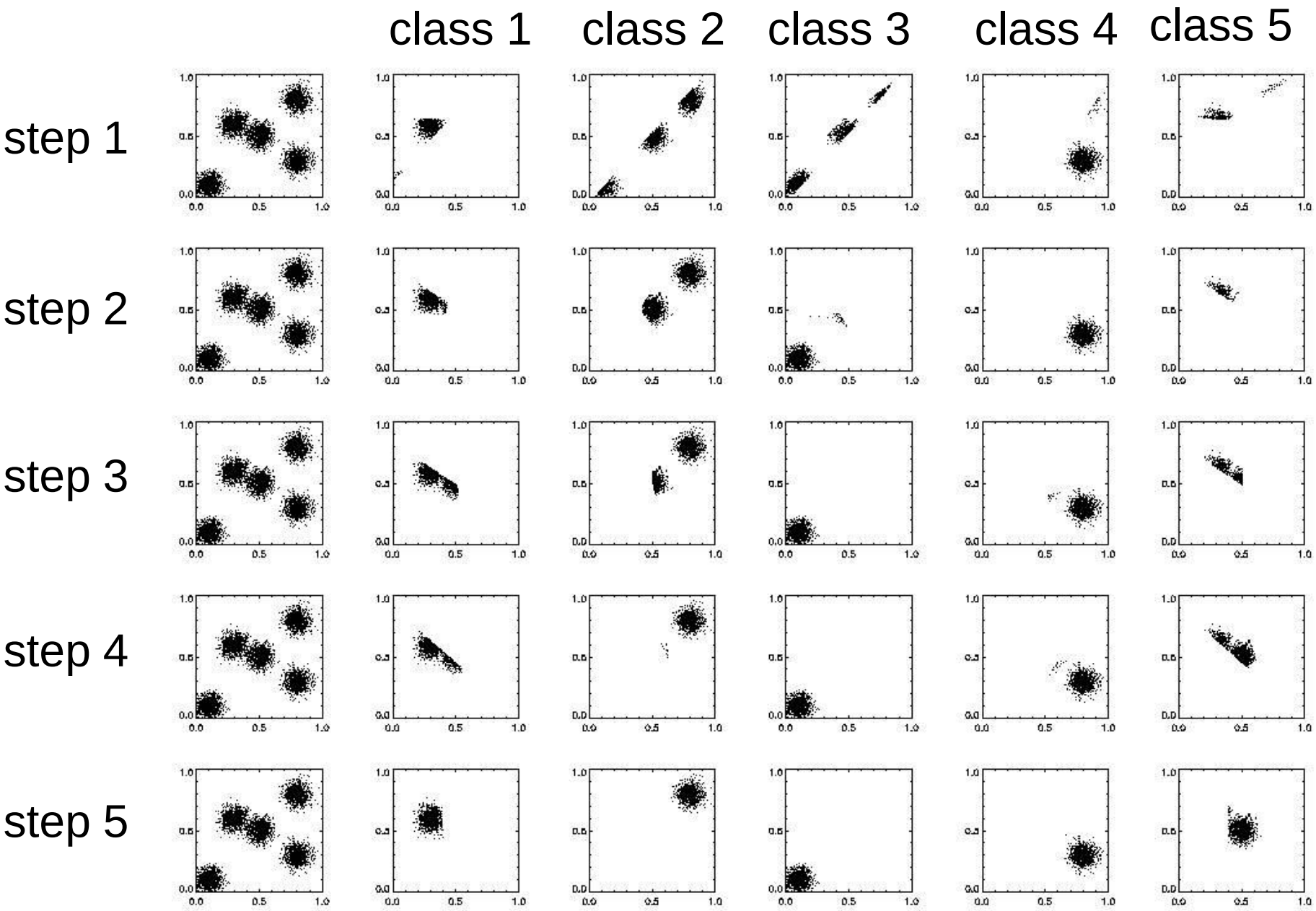
The classification method: k-means clustering algorithm

pixels properties
cluster around 10 ⟶
RGB classes





How does k-means work?



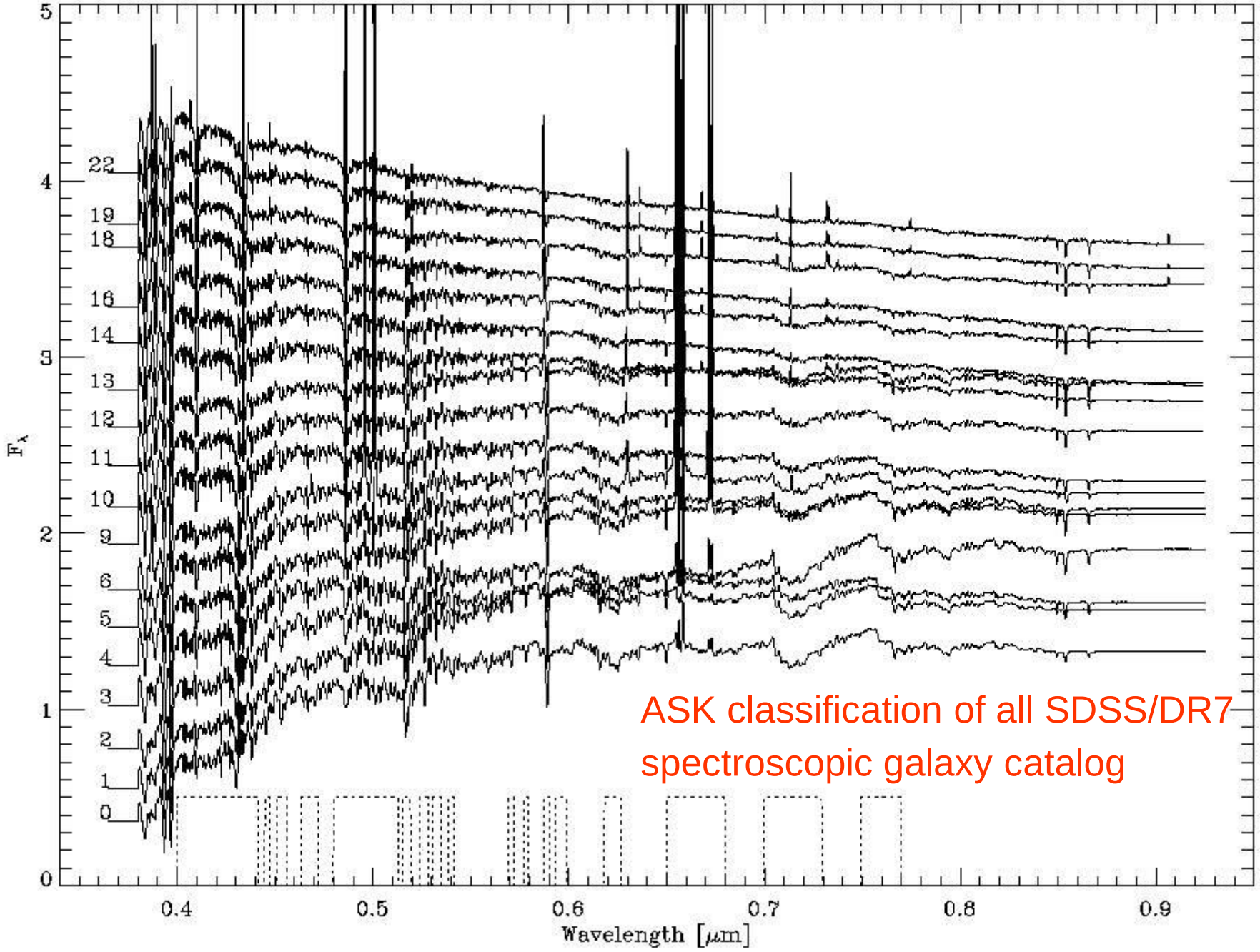
Automatic spectral K-means (ASK) classification of the full SDSS/DR7

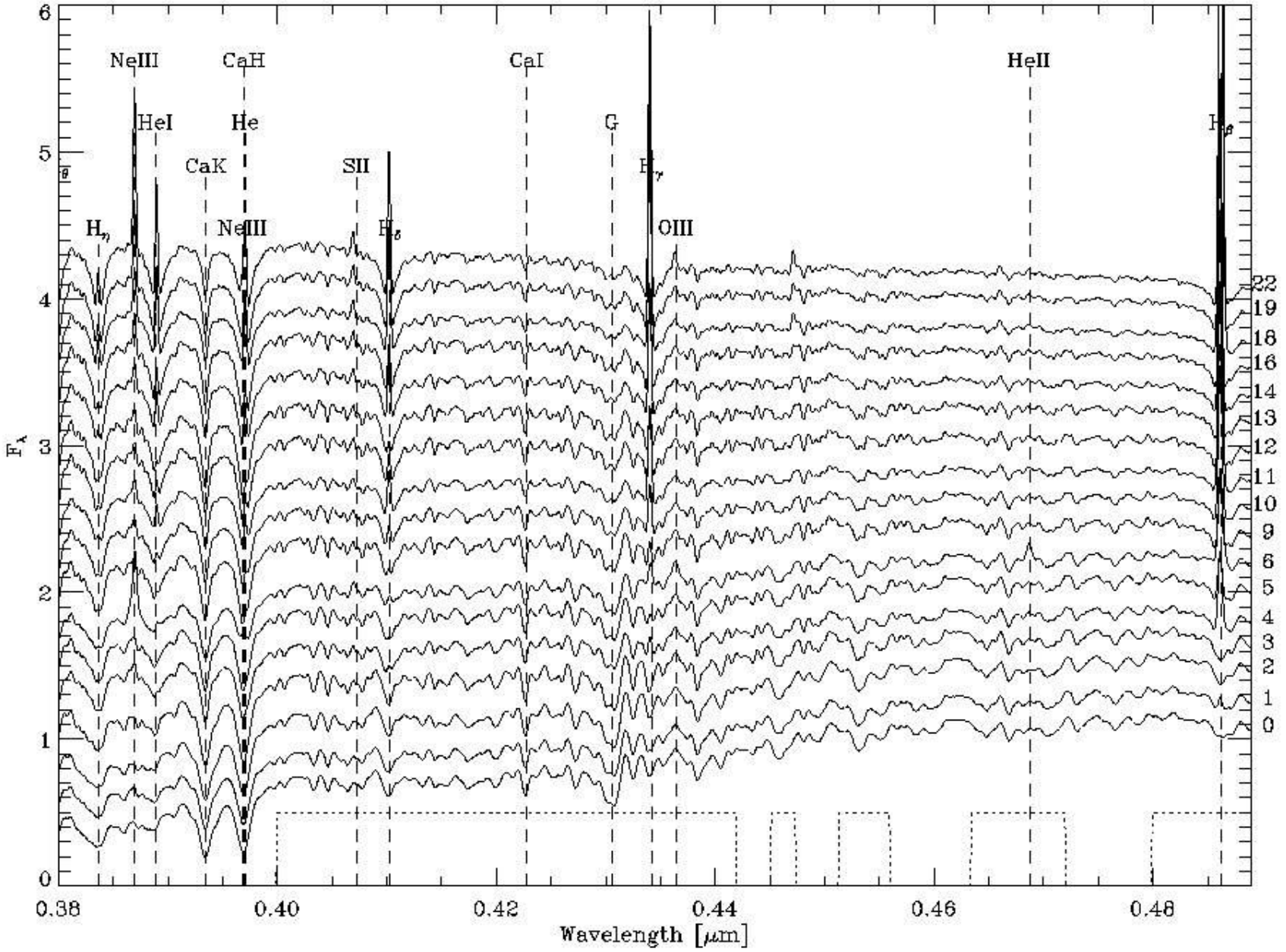
- It works for SDSS/DR7 spectra. 3800 – 9300 Å, ≈1.5 Å pixels, selected spectral regions, normalized to the mean flux in the *g*-band.
- Computationally intensive: 788677 spectra x 1637 pixels (≈11.6 Gb). 50 iterations. 150 initiallizations.

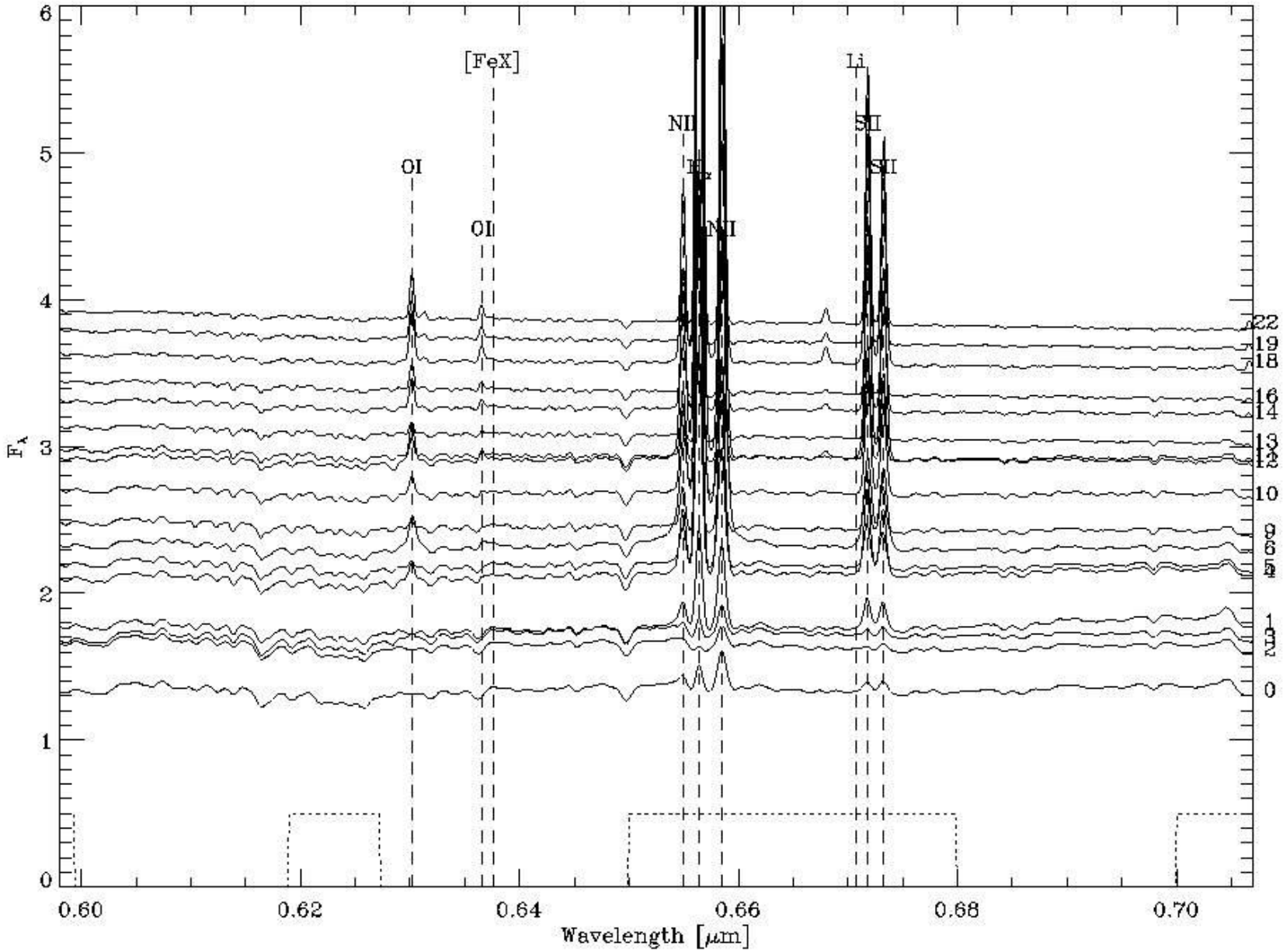
  IDL 300 min/ classification (31 days for 150) using a fast 8-core Intel Xenon 2.66GHz 32Bb RAM.

  Fortunately the algorithm can be parallelized. Fortran MPI 1 hour per 150 initializations using the cluster of 48 Intel Xenon CPUs (2.4 GHz) at IAC (de Vicente).

- 99% of the 788677 galaxies can be assigned to only 17 major classes. We order them by *u-g* color.


ASK classification of all SDSS/DR7
spectroscopic galaxy catalog
Wavelength [μm]
F_λ
0.4
0.5
0.6
0.7
0.8
0.9
0
1
2
3
4
5
22
19
18
16
14
13
12
11
10
9
6
5
4
3
2
1
0


NeIII
CaH
CaI
HeII
HeI
He
G
CaK
SII
OIII
NeIII
0
1
2
3
4
5
6
22
19
18
16
14
13
12
11
10
9
6
5
4
3
2
1
0
0.38
0.40
0.42
0.44
0.46
0.48
Wavelength [μm]
F_λ

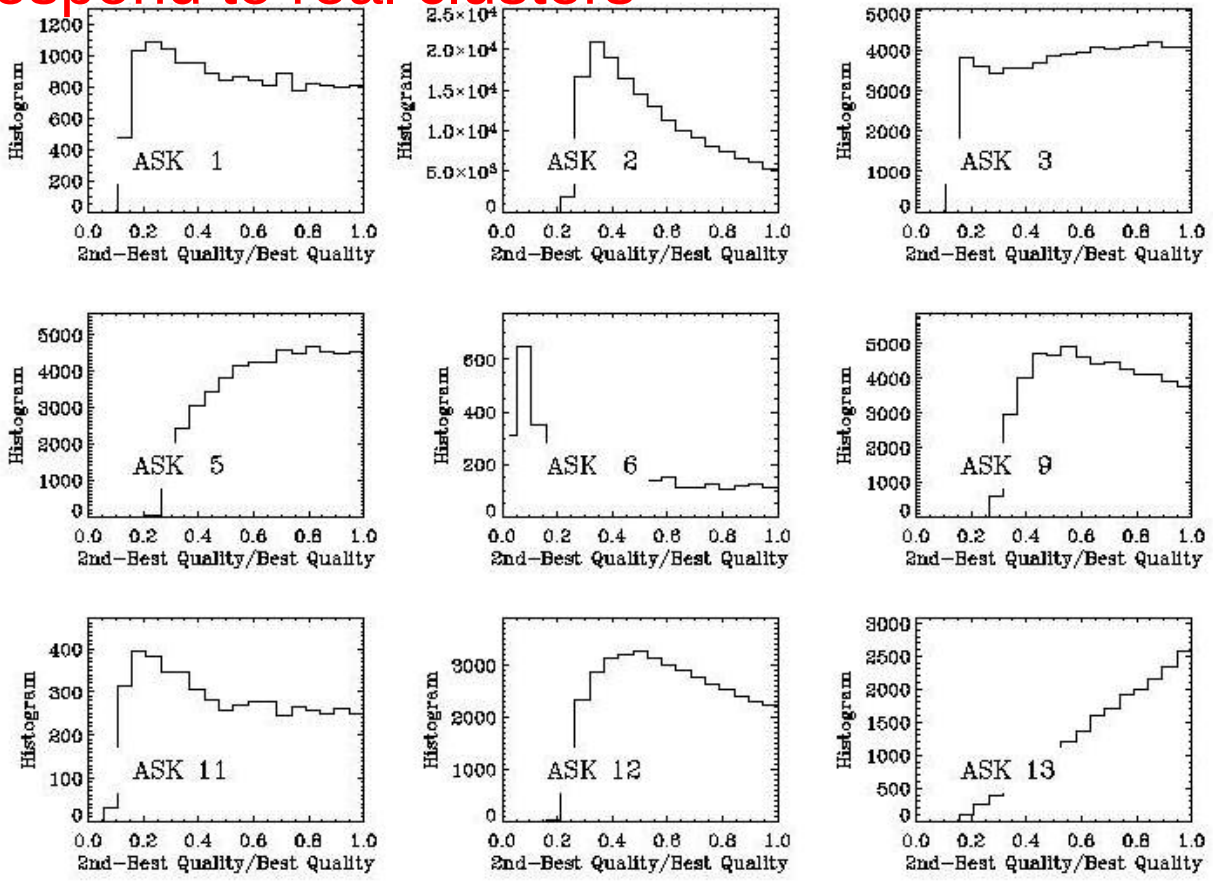


[FeX]
OI
OI
NII
Hα
NII
Li
SII
SII
22
19
18
16
14
13
12
10
9
6
5
4
1
3
2
0
F_λ
Wavelength [μm]
0.60
0.62
0.64
0.66
0.68
0.70
0
1
2
3
4
5
6

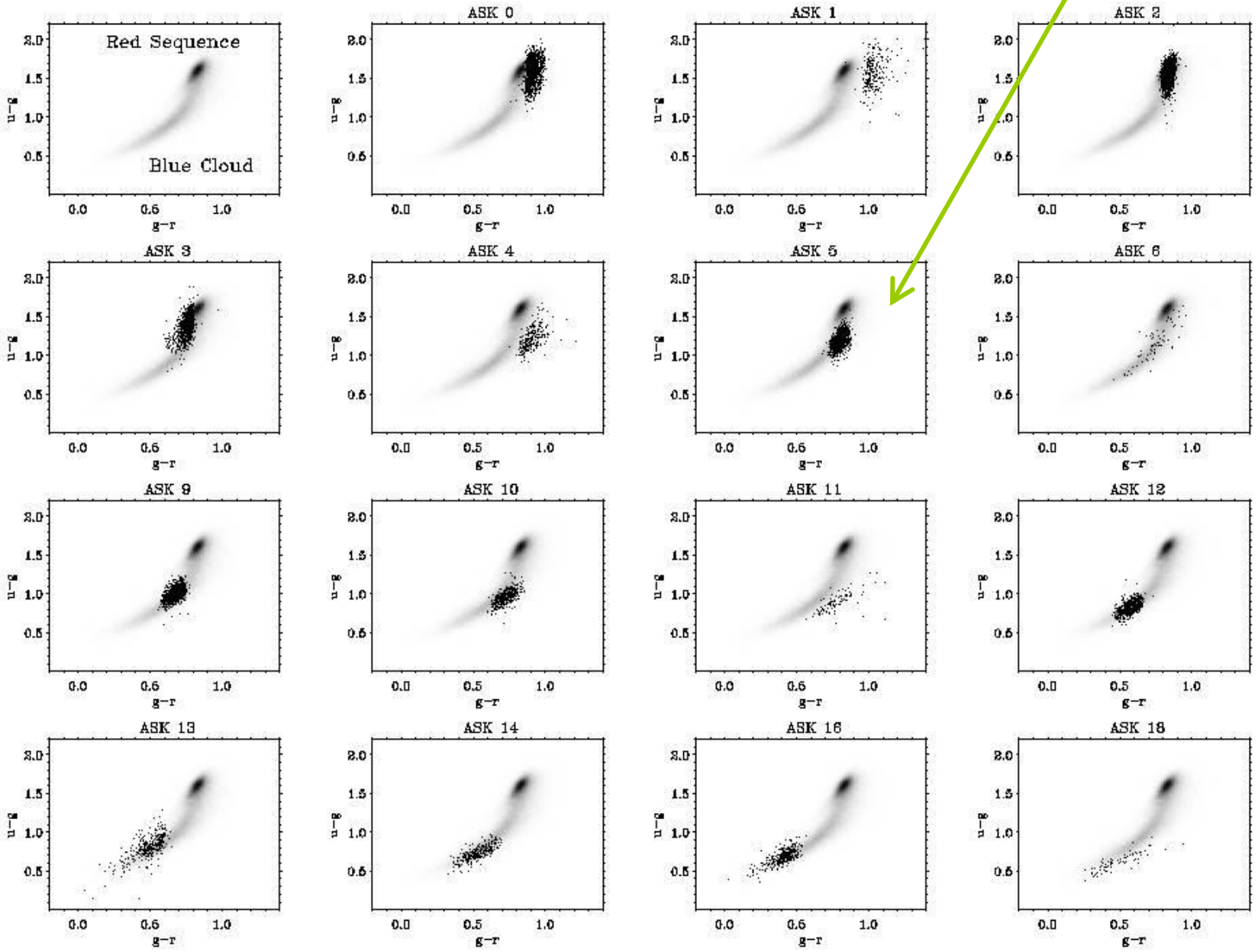
Properties of the ASK classes

Are there true clusters in the classification space?

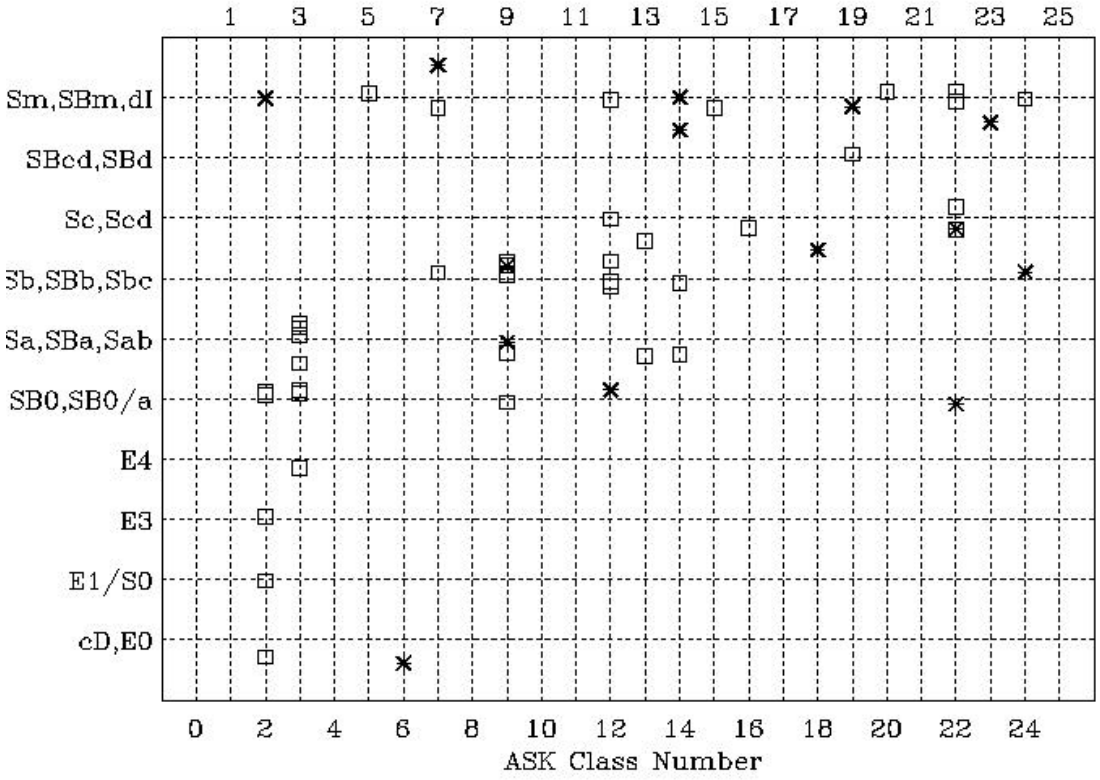
The spectra form a continuum, as judged from the existence of borderline galaxies. However, some of the classes do correspond to real clusters

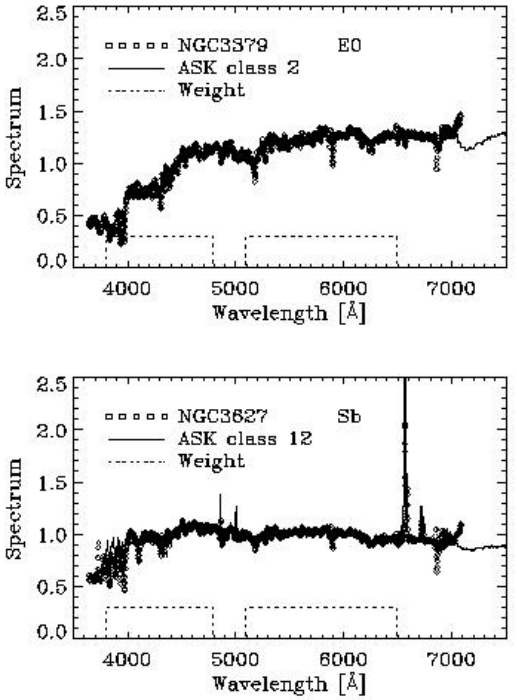


ASK classes distinguish galaxies in the green valley



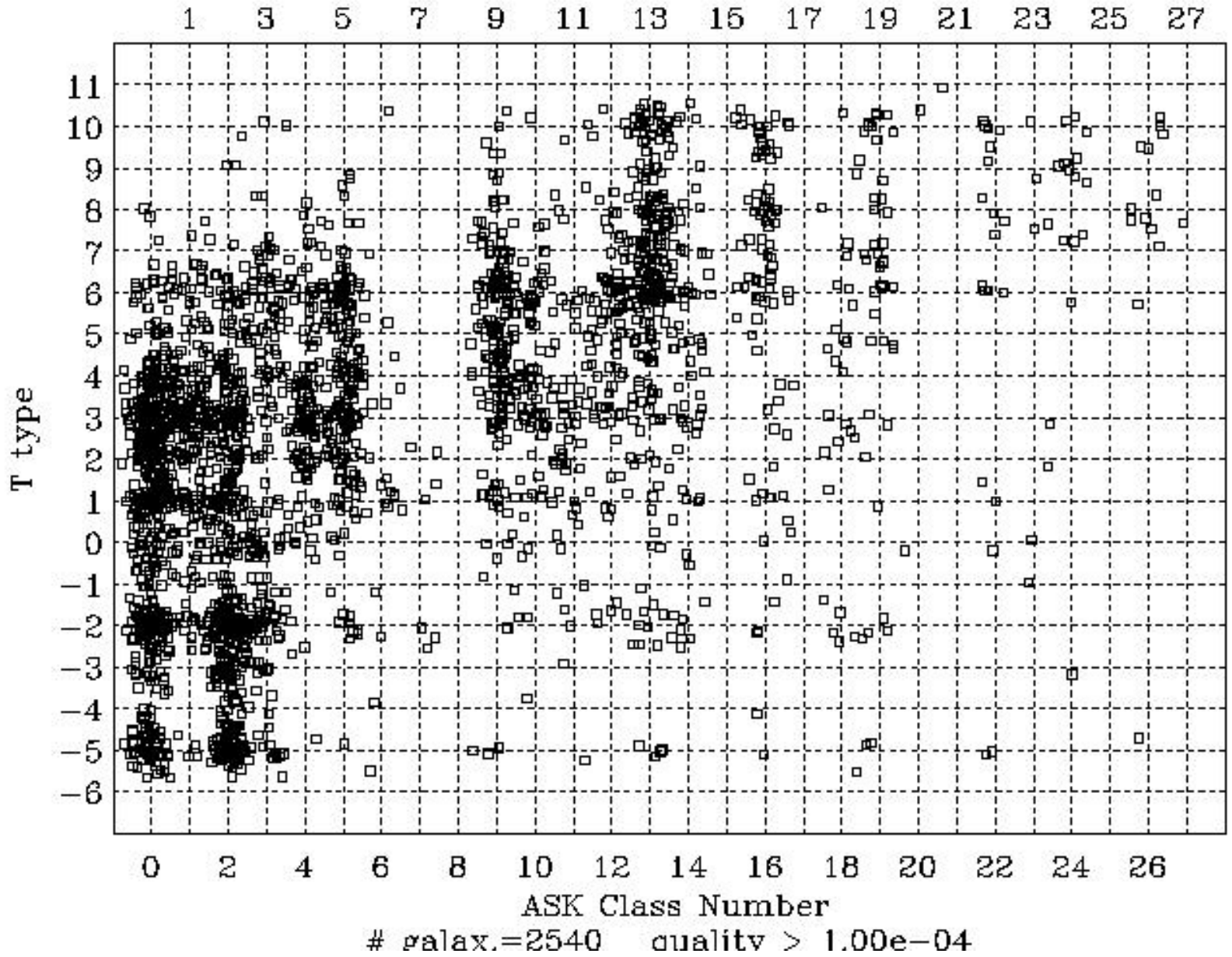
ASK class vs morphological classification





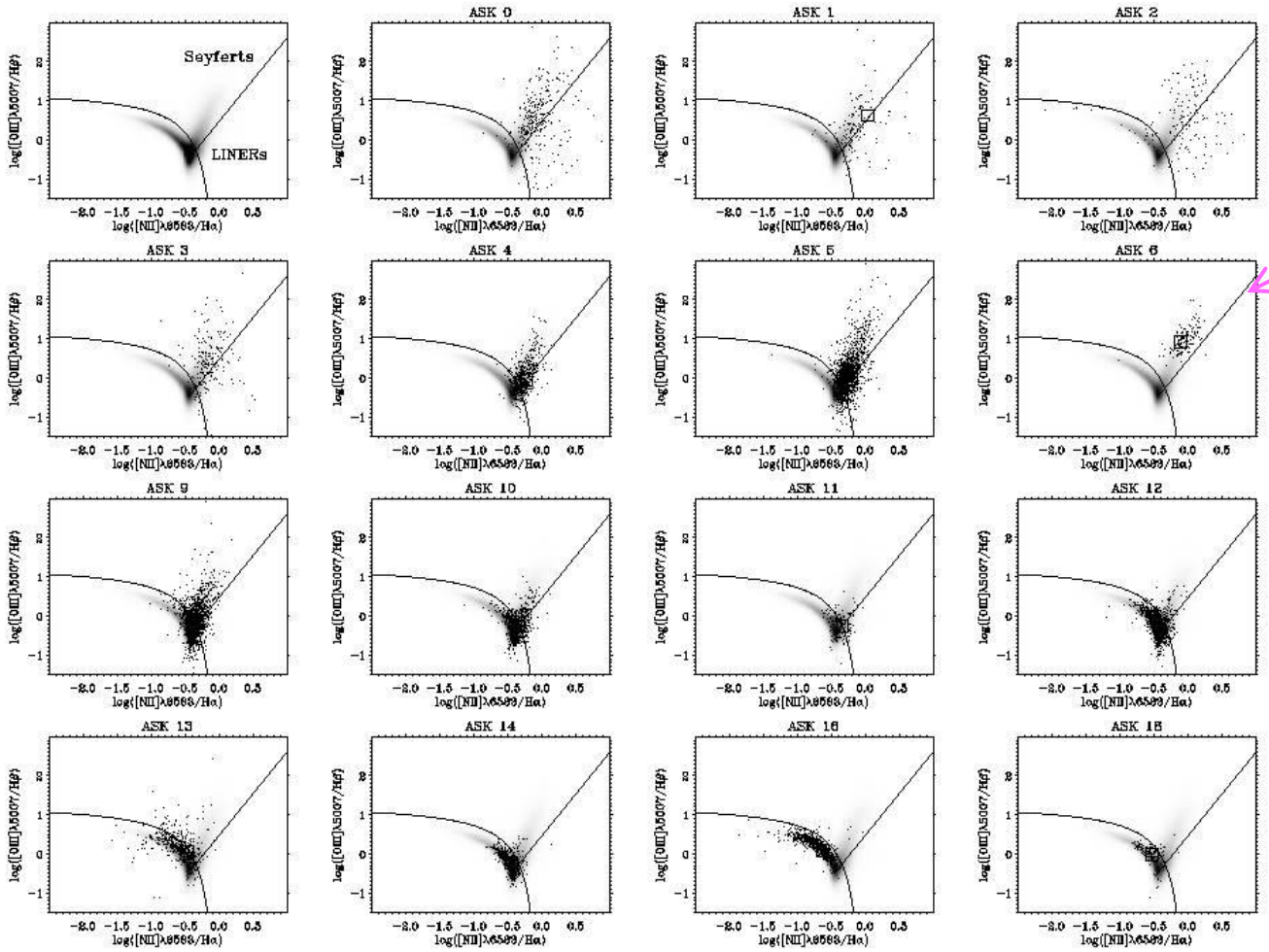
Kennicutt 02

There is a clear trend for the small ASK numbers (red galaxies) to be associated with the early-types, and vice versa. However the relationship presents a large intrinsic scatter.



galaxies in the RC3 catalog (de Vaucouleurs et al. 91)

ASK class vs AGN activity



ASK 6, pure Seyfert galax

red galaxies present AGN activity ASK 0,1,2

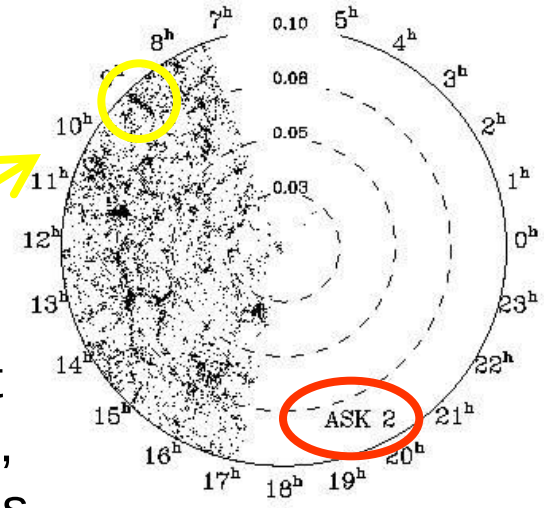
green galaxies also resent AGN activity ASK 3,4,5,6

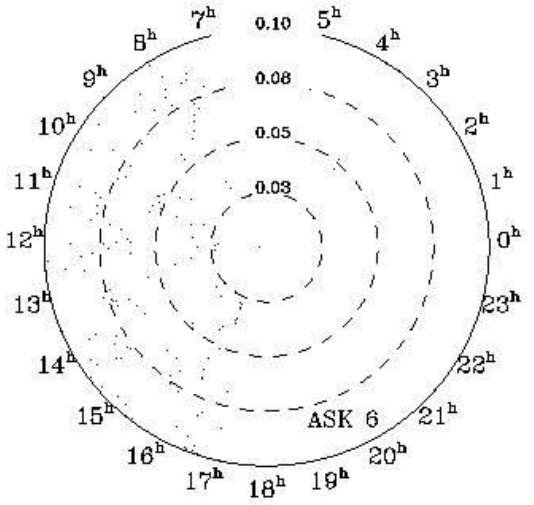
blue galaxies resent starburst activity ASK ≥ 7

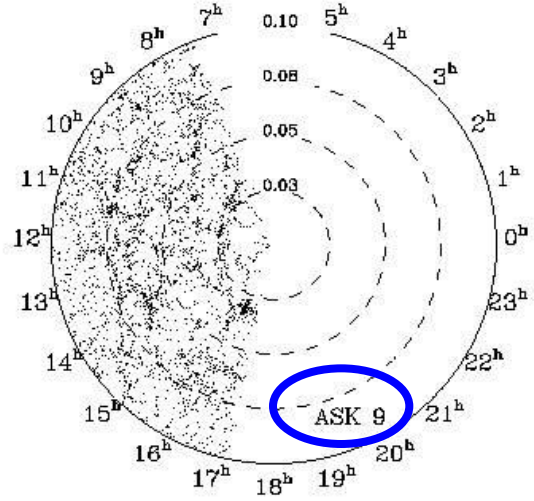
Cone diagram, redshift < 0.1

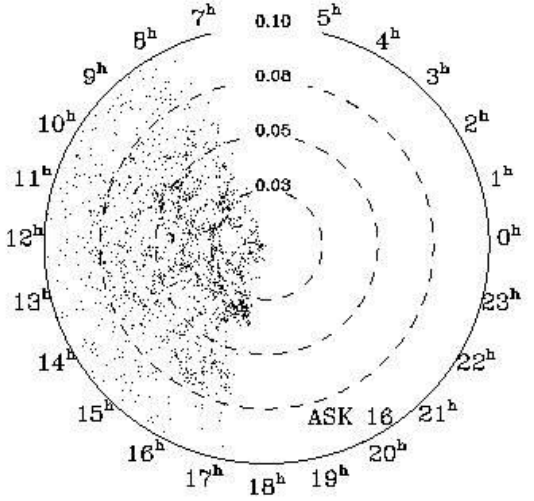
$35^o < DEC < 45^o$

Clear *finger of god* effect present only in red types, meaning that red galaxies tend to be in clusters, whereas blue types are more spread out.



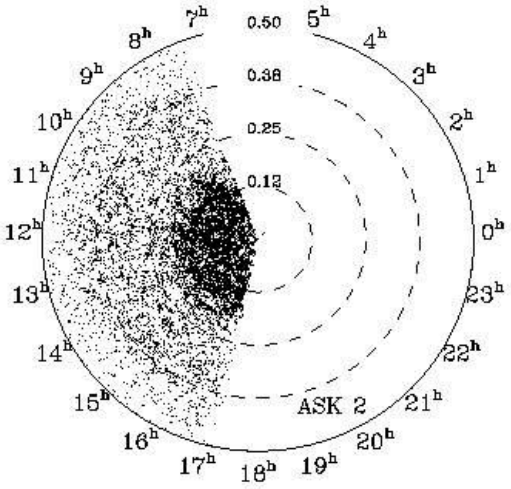


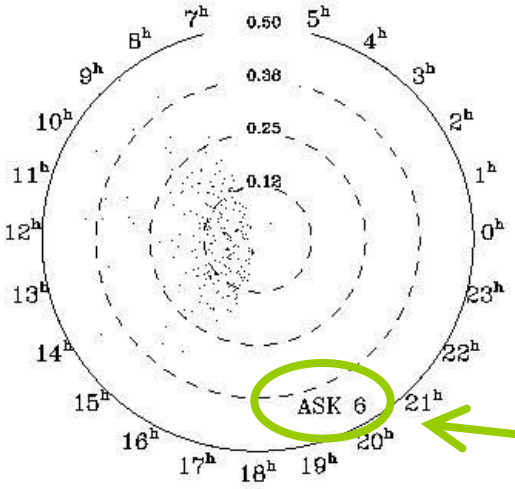


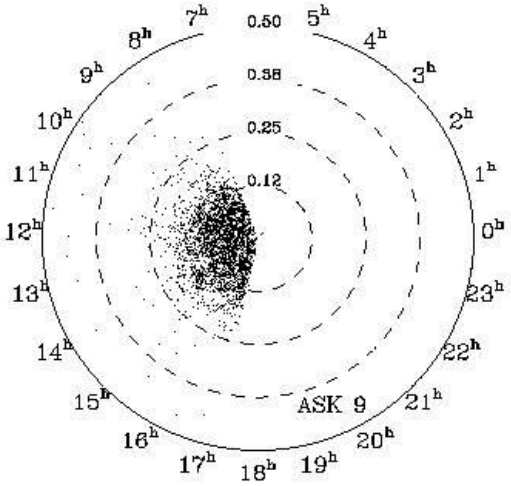


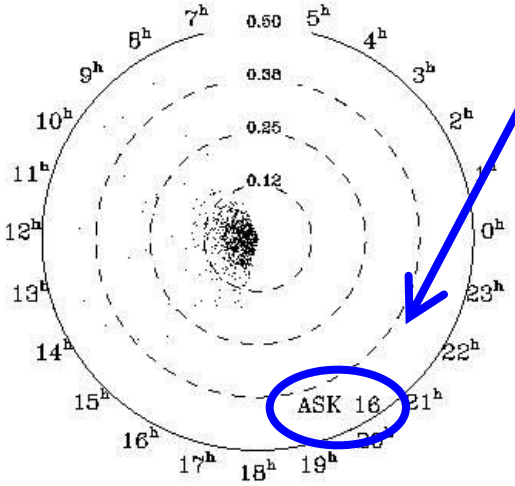
Cone diagram, redshift < 0.5

$35^o < DEC < 45^o$









- Seyferts (ASK 6) are spread out.
- Blue types are nearby.

Uses

- The classification is freely available to anyone (Spanish VO)
- Classification, meaning, that each class come with a number of physical properties that can be assigned to your target galaxy once its class is known.
- Complete template set for redshift determination, and galaxy classification. Drawback: limited wavelength range … but trivial extension down to 2500 Å.
- Target Selection. Green valley galaxies, Seyfert Galaxies,…
- Trivial extension to stellar spectra (work in under way)
- (New) specific classifications focused on a particular spectral features (e.g., low metal targets, double AGN, A.B. Morales Luis will talk about one of these applications)

Conclusions

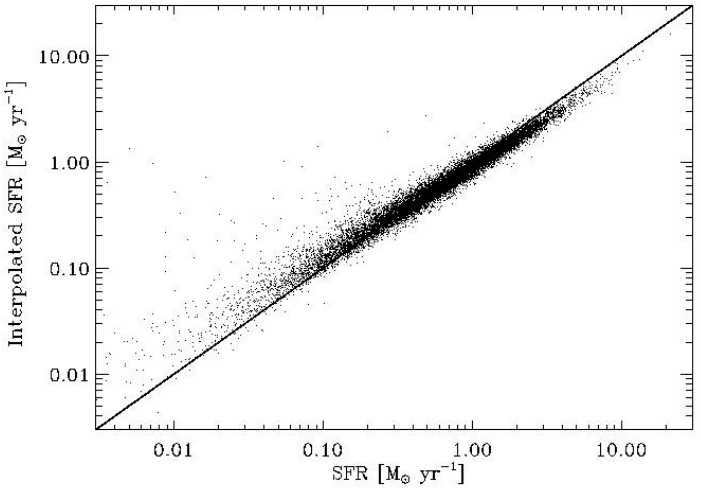
- Developed an unsupervised classification method for galaxy spectra (ASK)

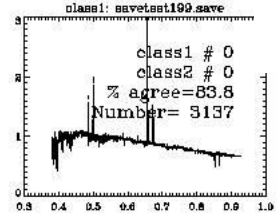
- Classify the some 930000 galaxies in the final data release of Sloan into only 17 major classes (SDSS/DR7) .

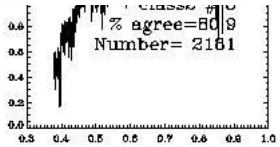
- Publically available.

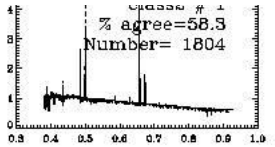
- With many potential applications, from templates for redshift determinations to target selection; see AB's talk!

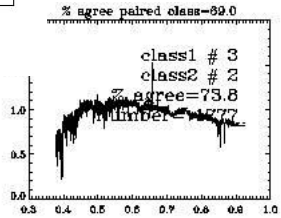
Flammarion woodcut

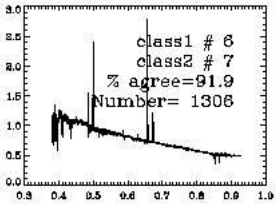

Interpolated SFR [M$_\odot$ yr^{-1}]
SFR [M$_\odot$ yr^{-1}]
0.01
0.10
1.00
10.00

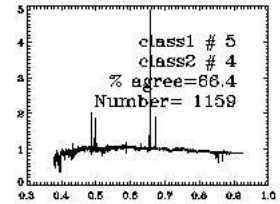

class1: nsvetsat199.save
class1 # 0
class2 # 0
% agree=83.8
Number= 3137

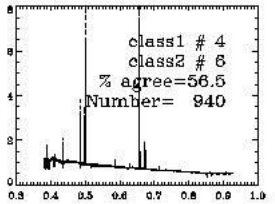

% agree=80.9
Number= 2161

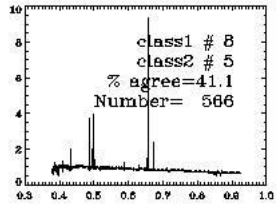

% agree=58.3
Number= 1804

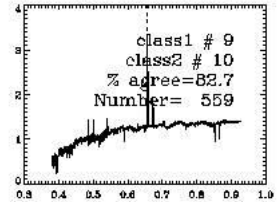

% agree paired class=82.0
class1 # 3
class2 # 2
% agree=73.8

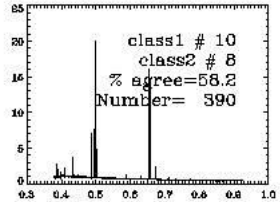

class1 # 6
class2 # 7
% agree=91.9
Number= 1308

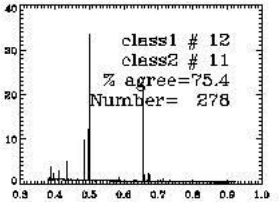

class1 # 5
class2 # 4
% agree=66.4
Number= 1159

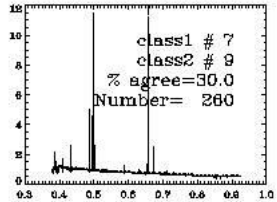

class1 # 4
class2 # 6
% agree=56.5
Number= 940

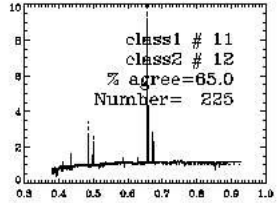

class1 # 8
class2 # 5
% agree=41.1
Number= 566

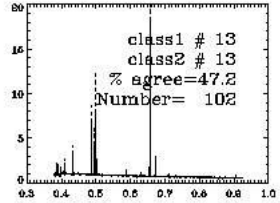

class1 # 9
class2 # 10
% agree=82.7
Number= 559

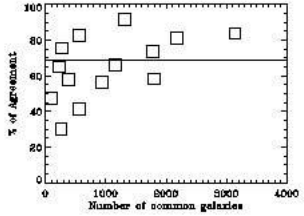
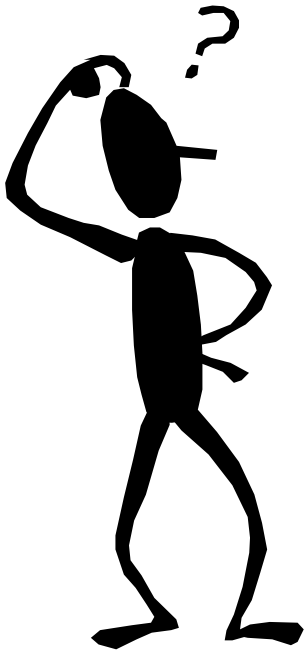
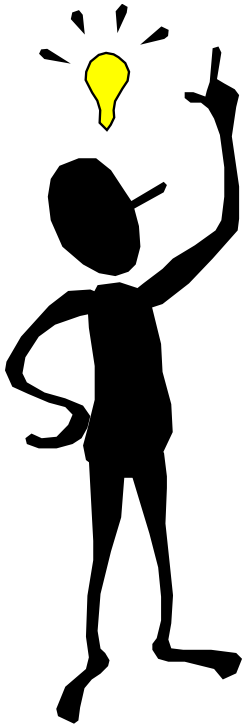

class1 # 10
class2 # 8
% agree=58.2
Number= 390


class1 # 12
class2 # 11
% agree=75.4
Number= 278


class1 # 7
class2 # 9
% agree=30.0
Number= 260


class1 # 11
class2 # 12
% agree=65.0
Number= 225


class1 # 13
class2 # 13
% agree=47.2
Number= 102


% of Agreement
Number of common galaxies