Three data analysis problems

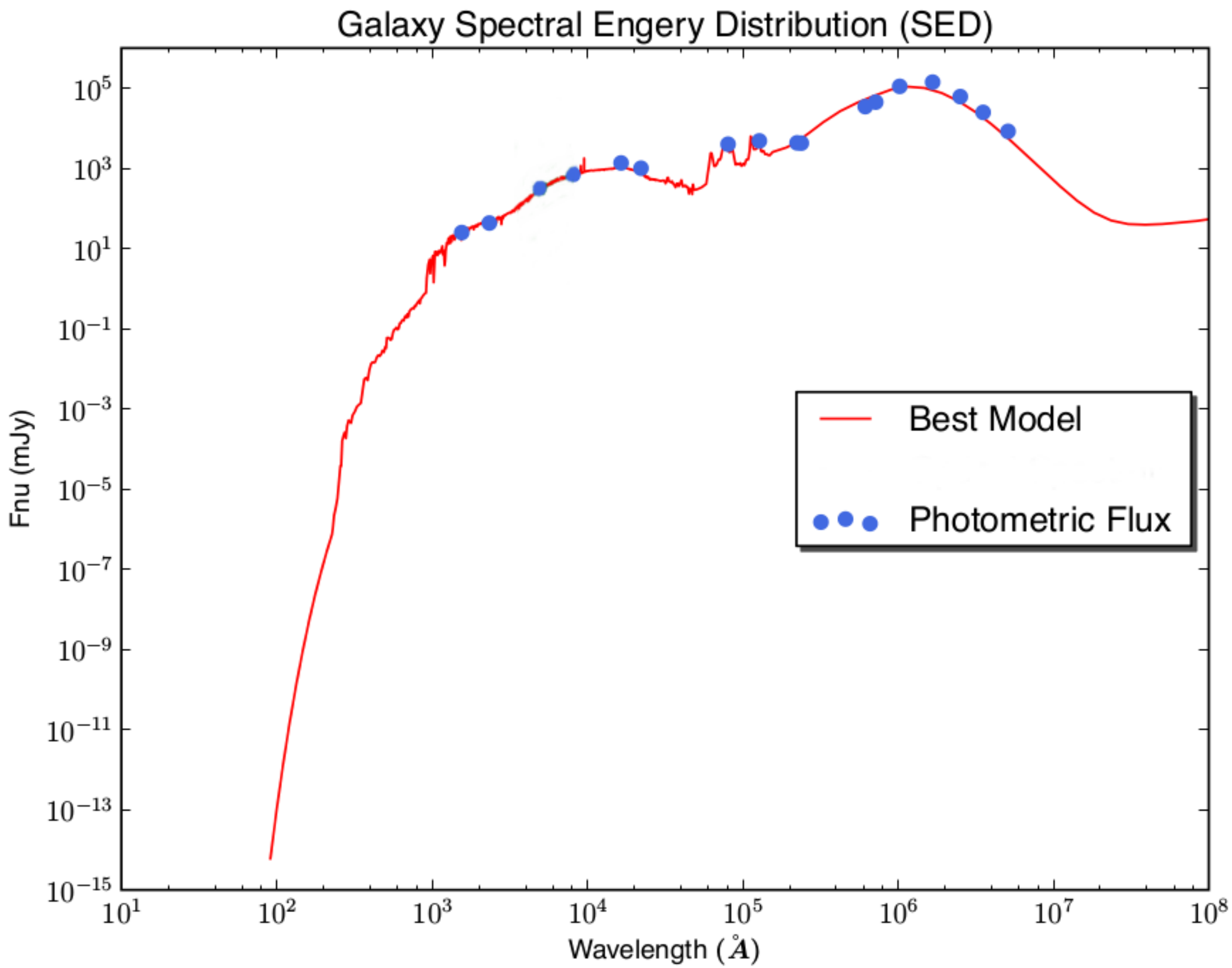
Andreas Zezas

University of Crete
CfA

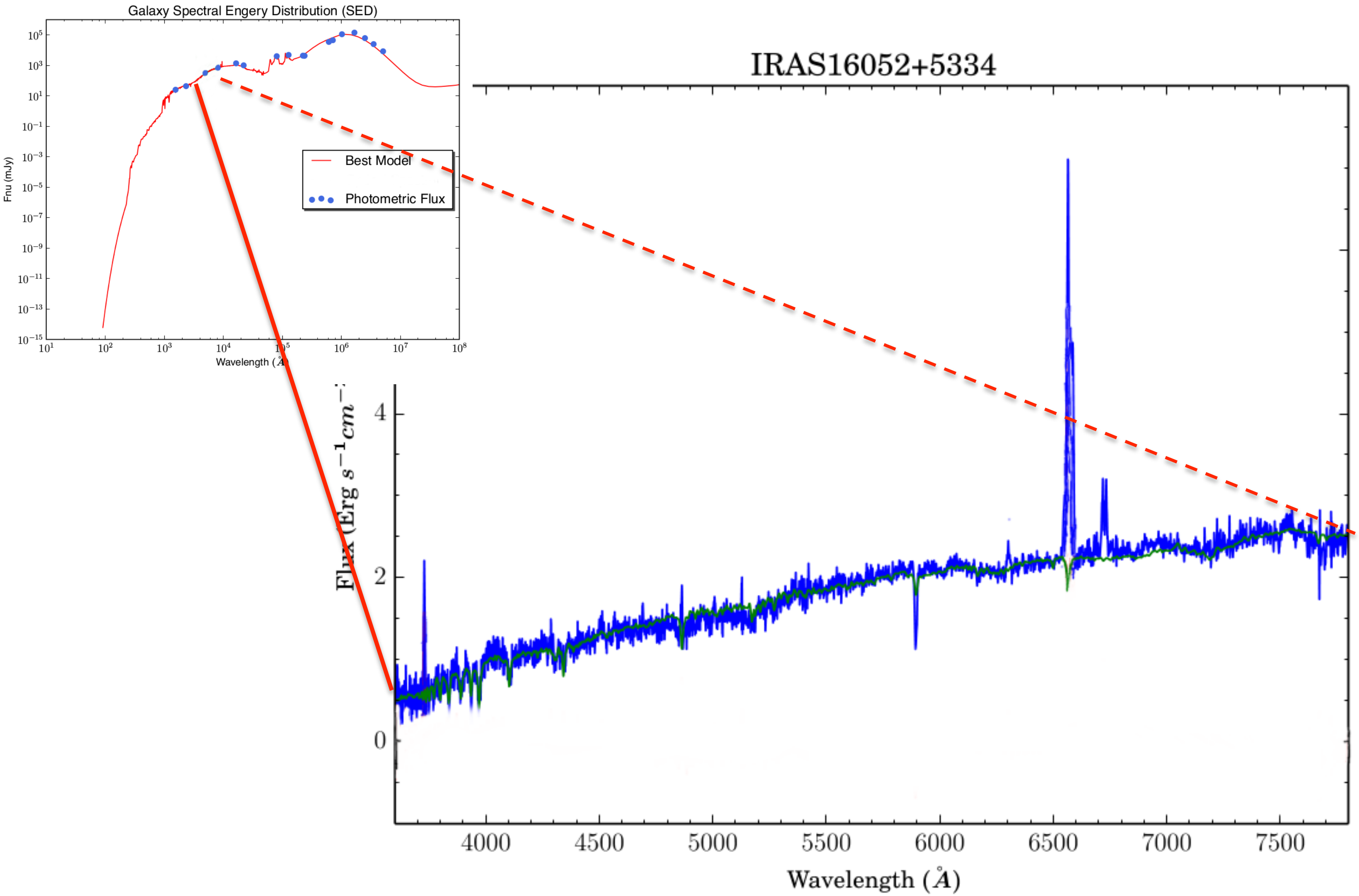
Two types of problems:

- Fitting
- Source Classification

Fitting: complex datasets

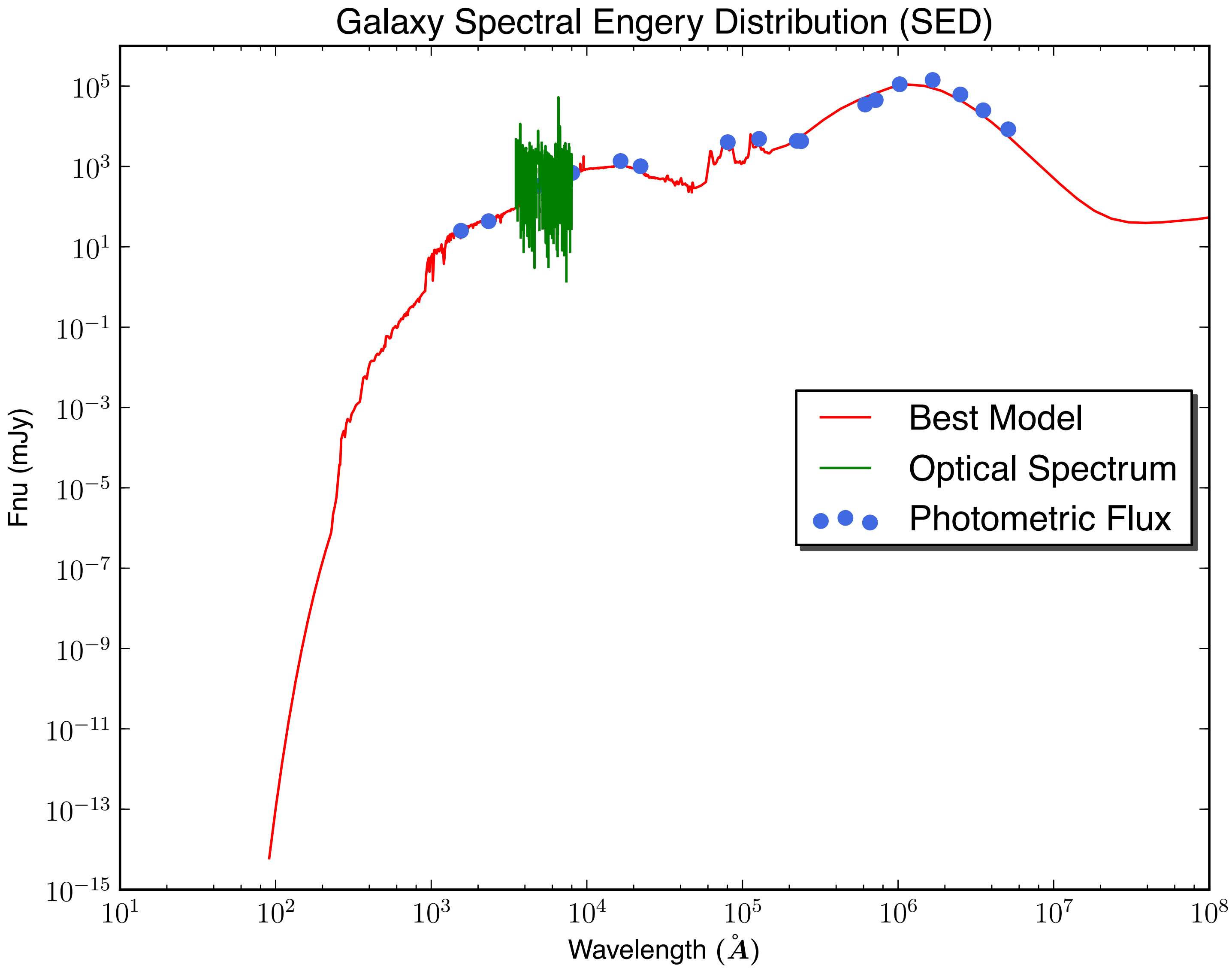


Fitting: complex datasets

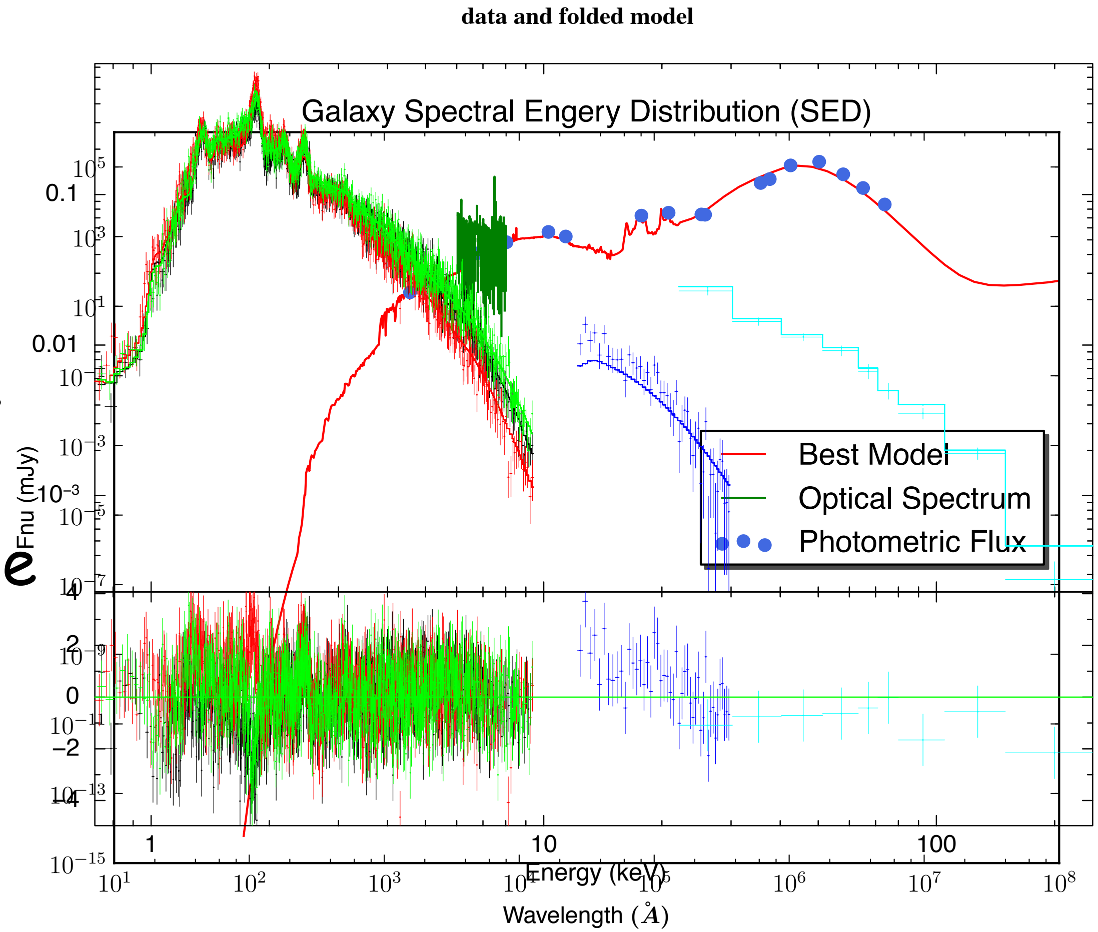


Maragoudakis et al. in prep.

Fitting: complex datasets



Fitting: complex datasets



Iterative fitting may work, but it is inefficient and confidence intervals on parameters not reliable

How do we fit jointly the two datasets ?

VERY common problem !

Problem 2

Model selection in 2D fits of images

A primer on galaxy morphology

Three components:

spheroidal

$$I(R) = I_e \exp\left[-7.67\left[\left(\frac{R}{R_e}\right)^{1/4} - 1\right]\right]$$

exponential disk

$$I(R) = I_0 \exp\left(\frac{r}{r_h}\right)$$

and nuclear point source (PSF)

Fitting: The method

Use a generalized model

$$I(R) = I_e \exp\left[-k\left[\left(\frac{R}{R_e}\right)^{1/n} - 1\right]\right]$$

n=4 : spheroidal
n=1 : disk

Add other (or alternative) models as needed

Add blurring by PSF

Do χ^2 fit (e.g. Peng et al., 2002)

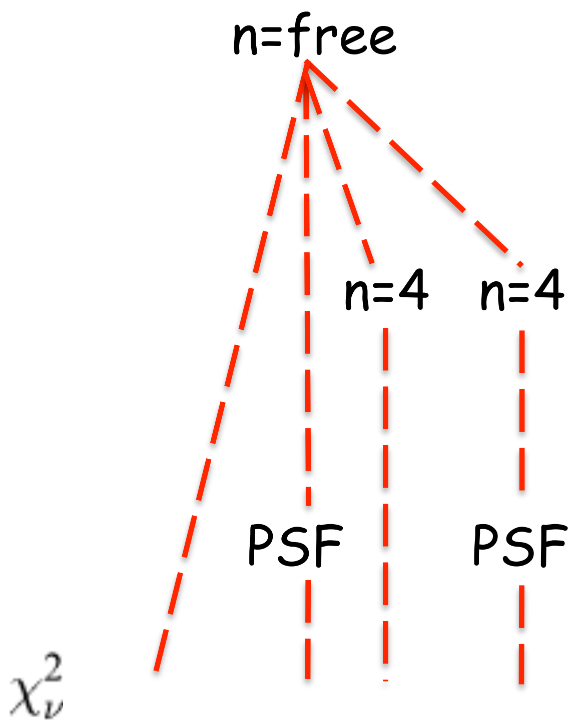
$$\chi^2_\nu = \frac{1}{N_{\mathrm{dof}}} \sum_{x=1}^{nx} \sum_{y=1}^{ny} \frac{\left(\mathrm{flux}_{x,y} - \mathrm{model}_{x,y}\right)^2}{\sigma^2_{x,y}}$$

$$\mathrm{model}_{x,y} = \sum_{\nu=1}^{nf} f_{\nu,x,y}(\alpha_1 \ldots \alpha_n)$$

Fitting: The method

Typical model tree

$$I(R) = I_e \exp\left[-k\left[\left(\frac{R}{R_e}\right)^{1/n} - 1\right]\right]$$



Fitting: Discriminating between models

Generally χ^2 works

BUT:

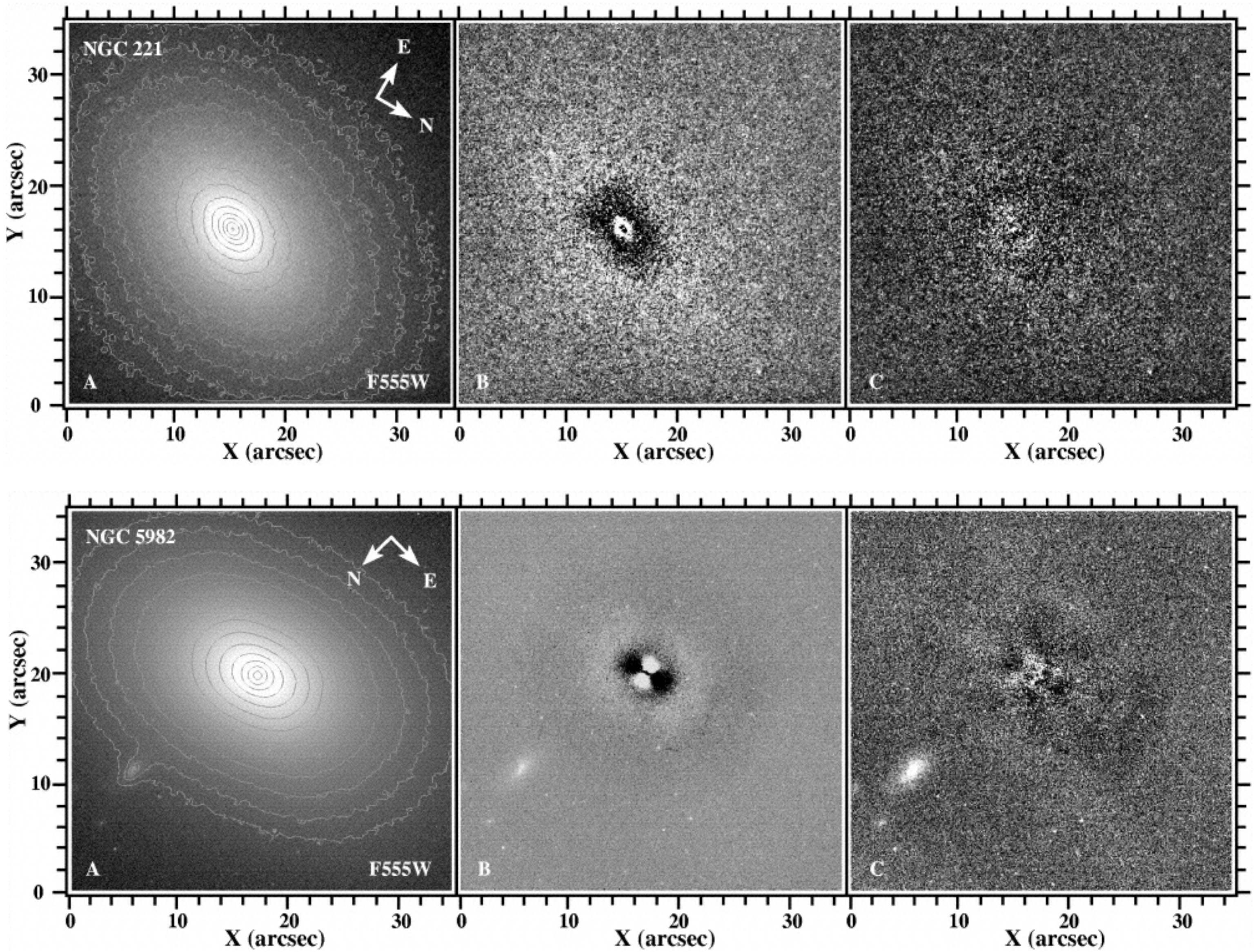
Combinations of different models may give similar χ^2

How to select the best model ?

Models not nested: cannot use standard methods

Look at the residuals

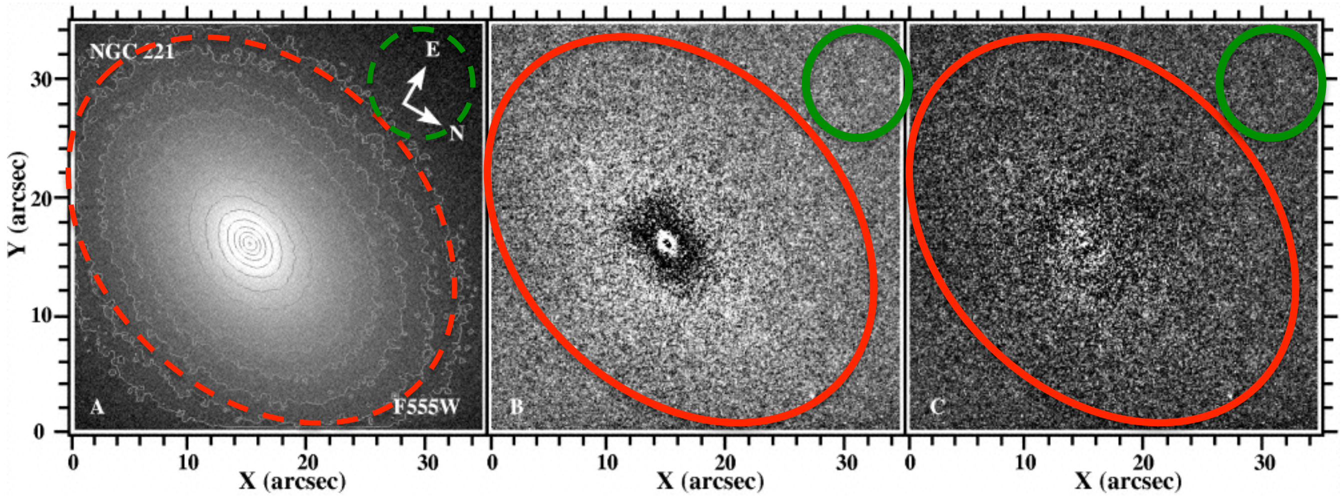
Fitting: Discriminating between models



Fitting: Discriminating between models

Excess variance

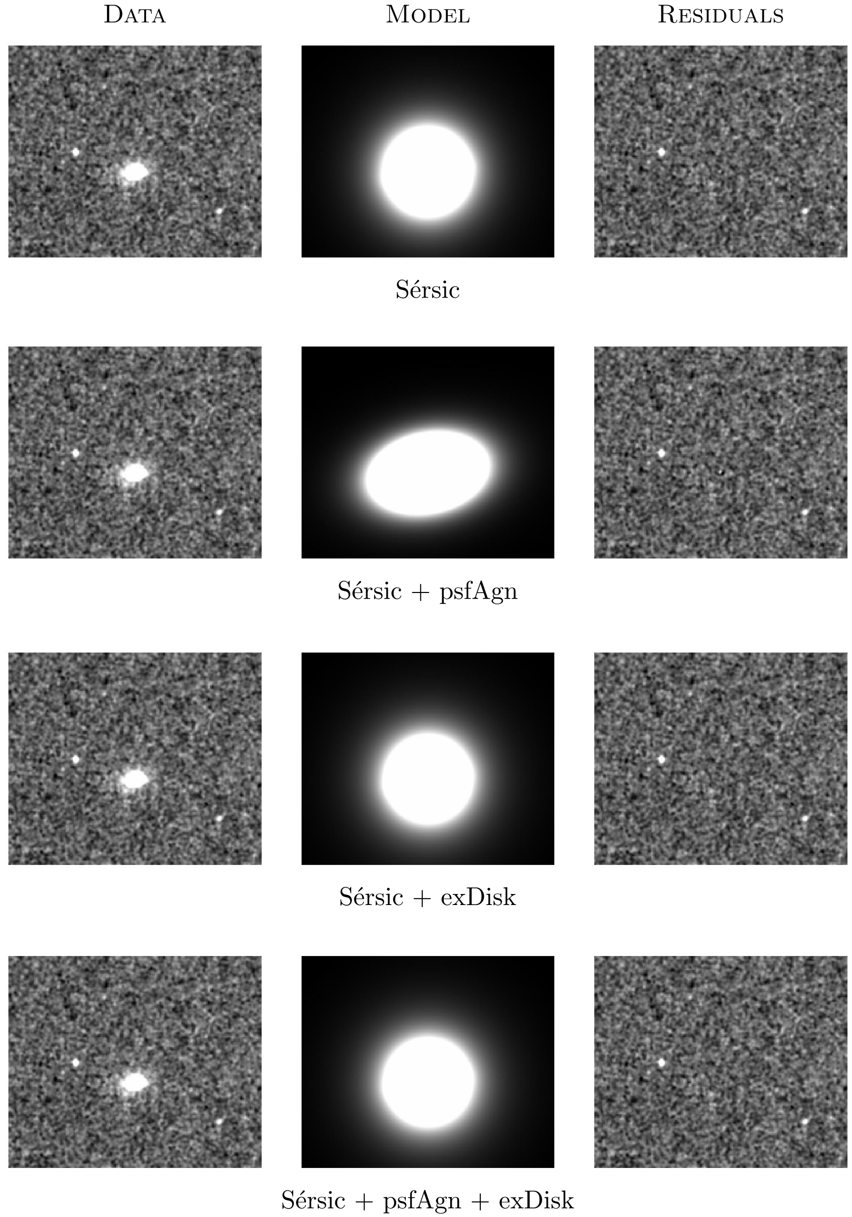
$$\sigma_{XS}^2 = \sigma_{obj}^2 - \sigma_{sky}^2$$



Best fitting model among least χ^2 models
the one that has the lowest exc. variance

Fitting: Examples

Model	χ^2_ν	σ^2_{XS}
(1)	(2)	(3)
Sérsic	1.107	1.722(0.120)
Sérsic + psfAgn	1.107	1.657(0.118)
Sérsic + exDisk	1.107	1.770(0.121)
Sérsic + psfAgn + exDisk	1.106	1.472(0.113)



Bonfini et al. in prep.

Fitting: Problems

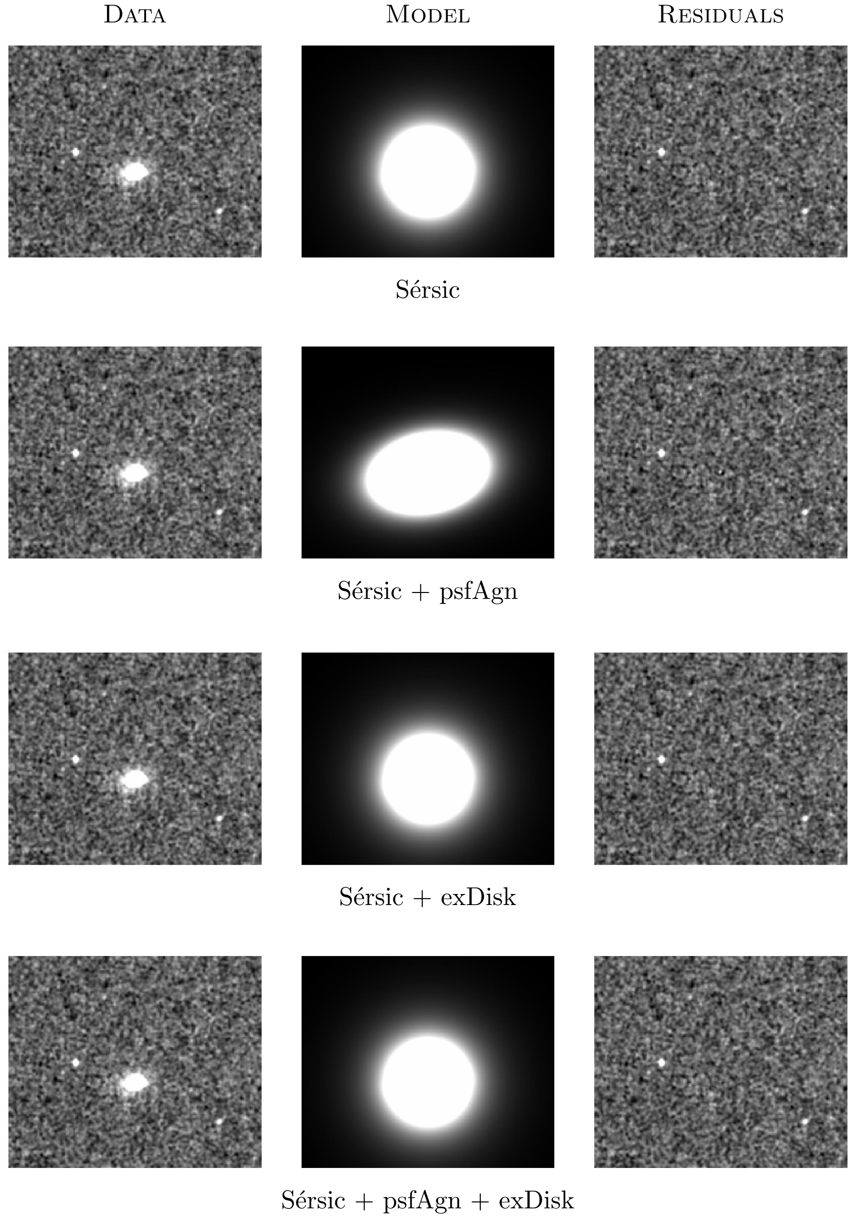
However, method not ideal:

It is not calibrated

Cannot give significance

Fitting process computationally intensive

Require an alternative, robust, fast, method



Problem 3

Source Classification

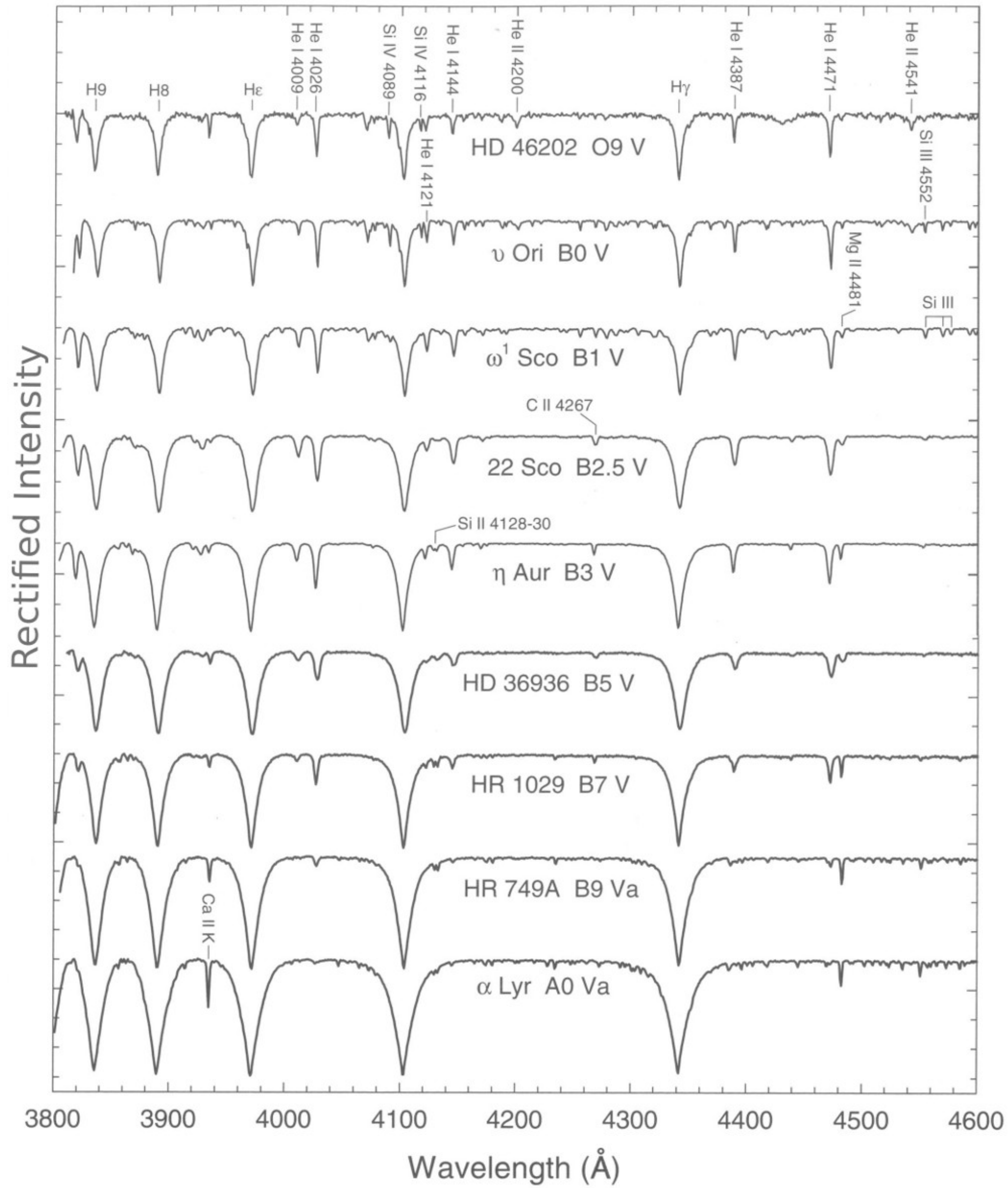
(a) Stars

Classifying stars

Relative strength of lines discriminates between different types of stars

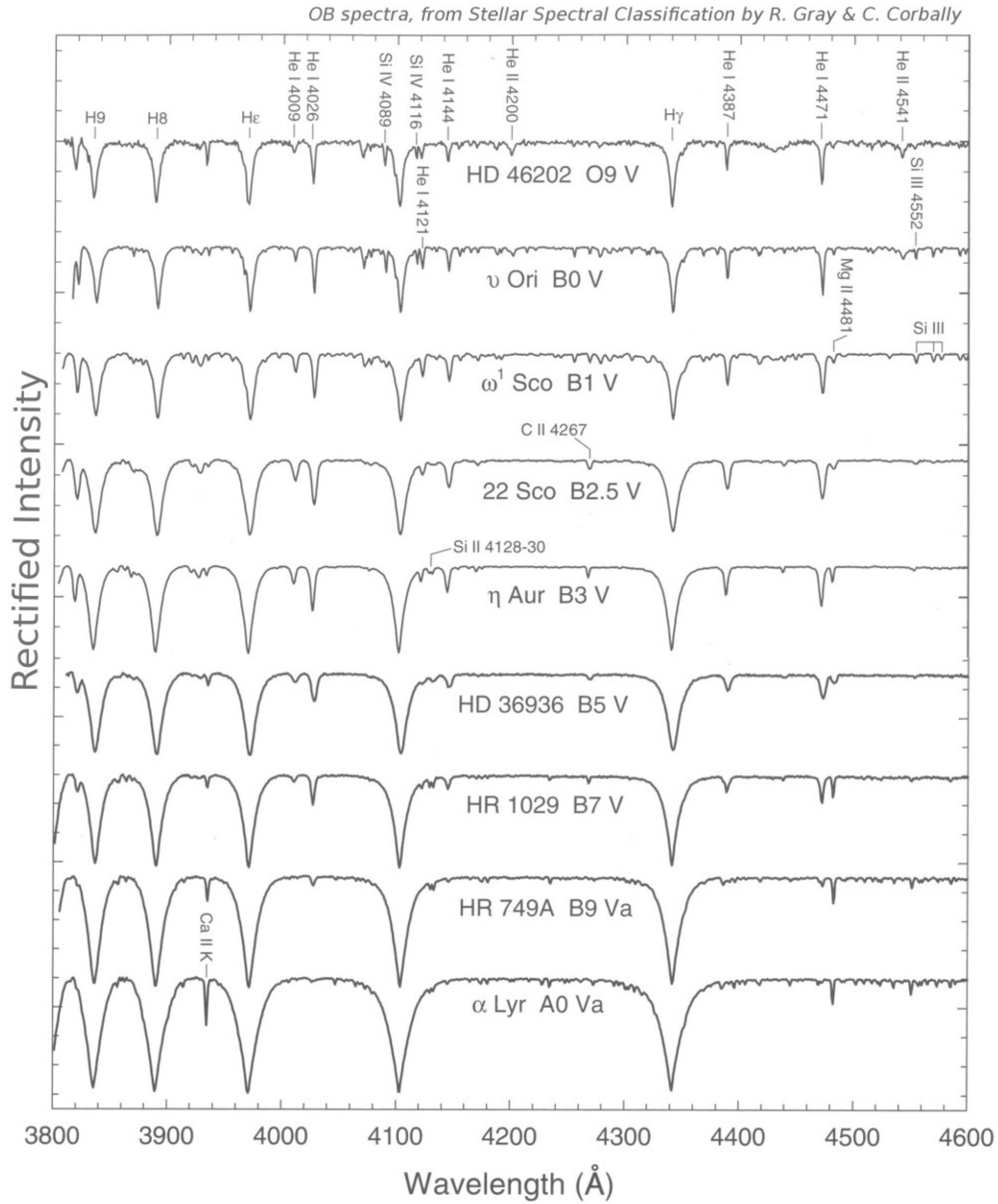
Currently done "by eye"
or
by cross-correlation analysis

OB spectra, from Stellar Spectral Classification by R. Gray & C. Corbally

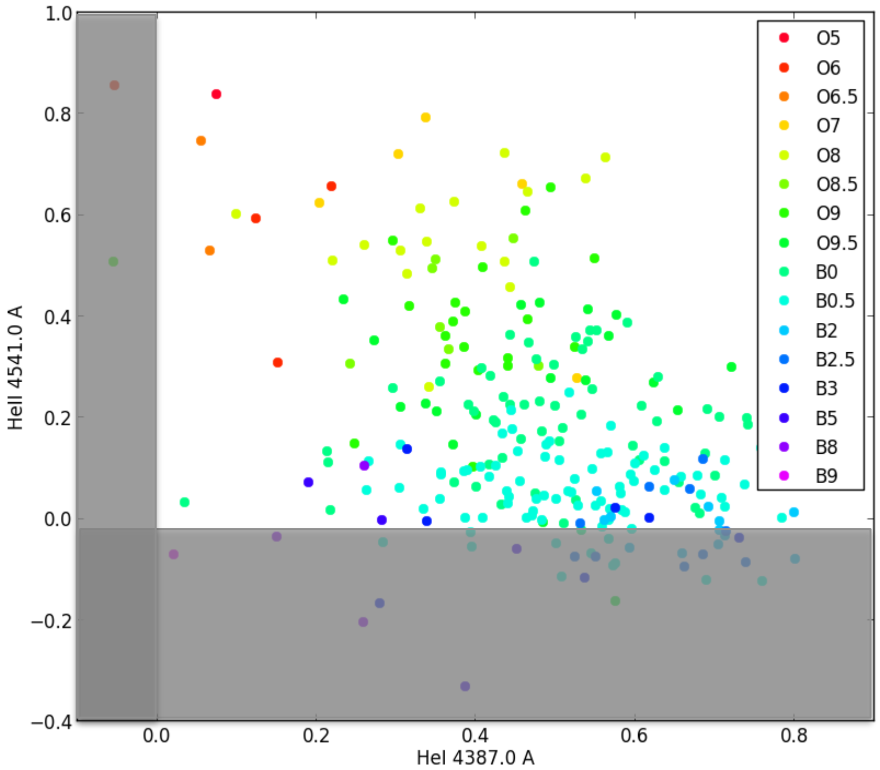


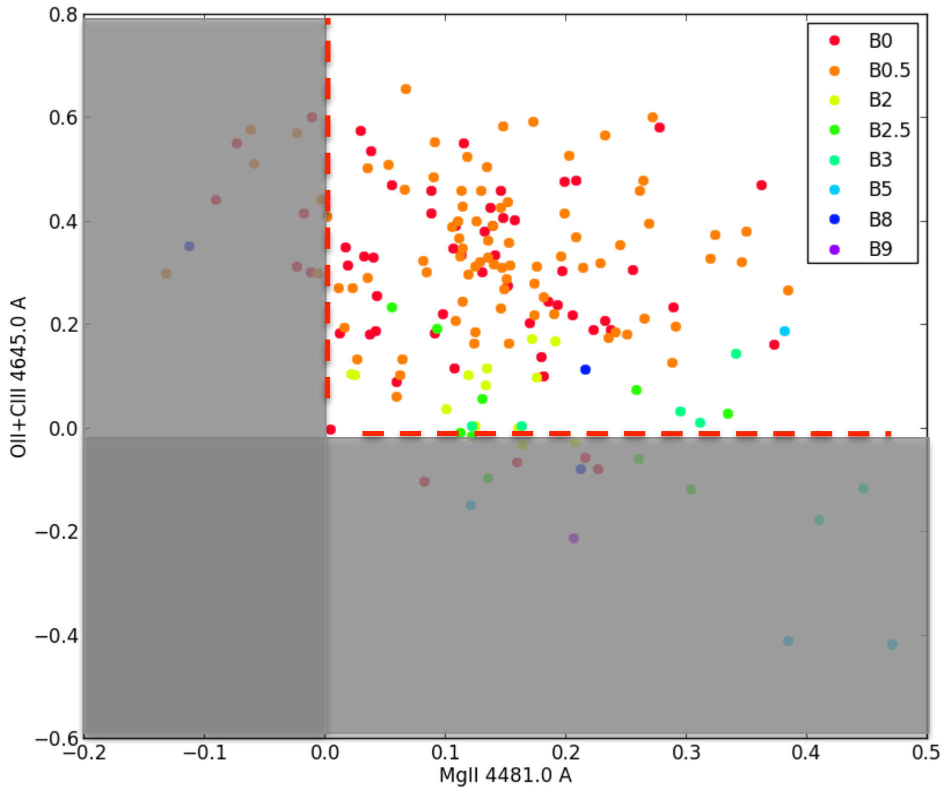
Classifying stars

Would like to define a quantitative scheme based on strength of different lines.



Classifying stars



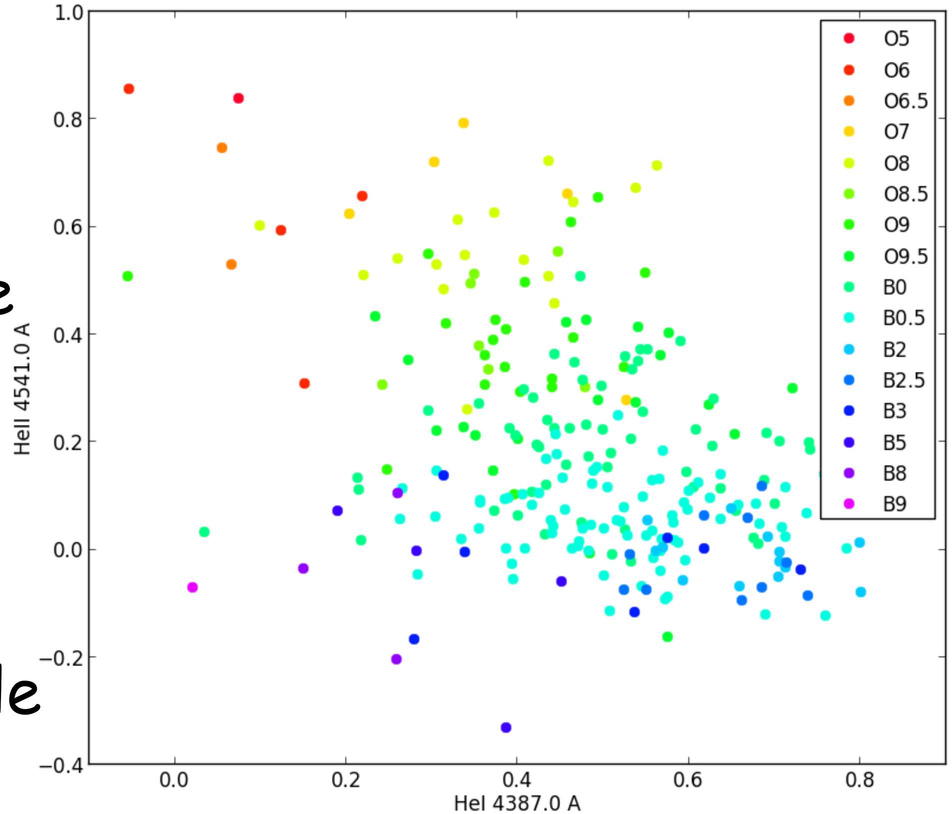


Maravelias et al. in prep.

Classifying stars

Not simple....

- Multi-parameter space
- Degeneracies in parts of the parameter space
- Sparse sampling
- Continuous distribution of parameters in training sample (cannot use clustering)
- Uncertainties and intrinsic variance in training sample

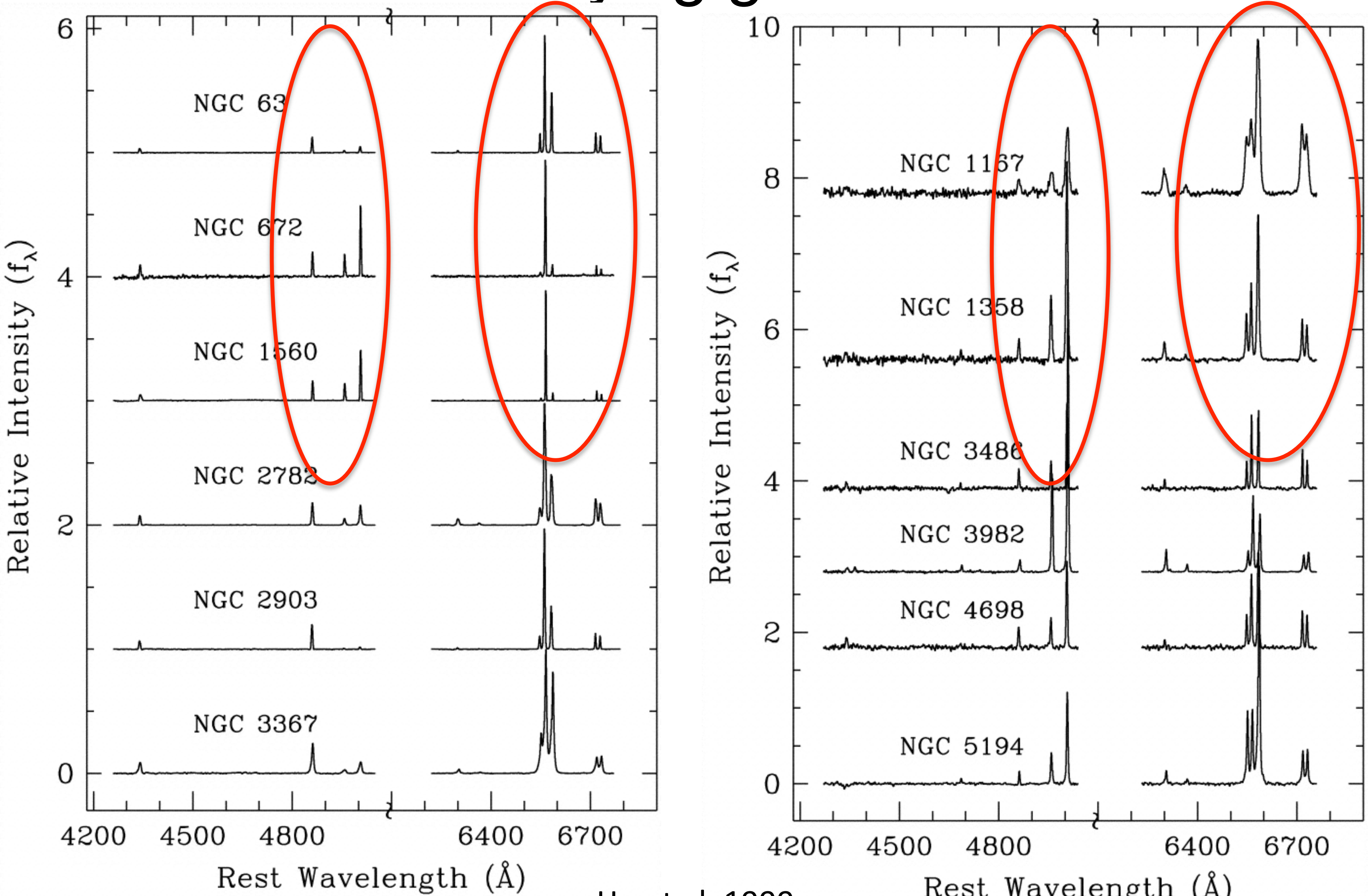


Problem 3

Source Classification

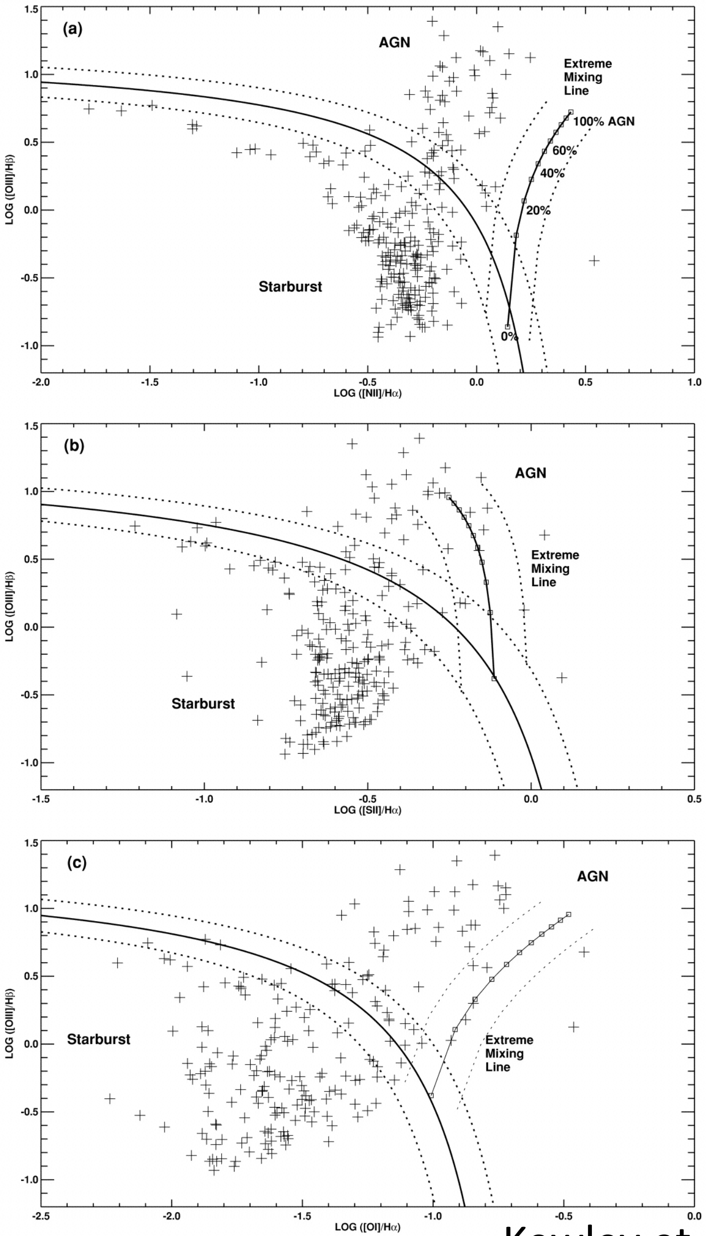
(b) Galaxies

Classifying galaxies

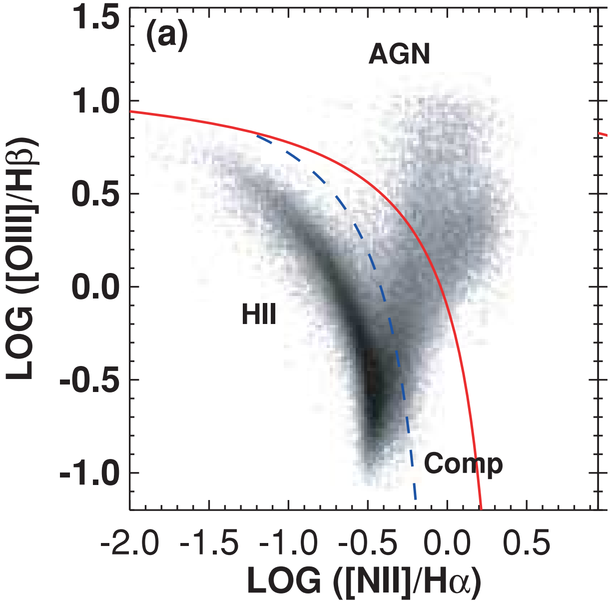


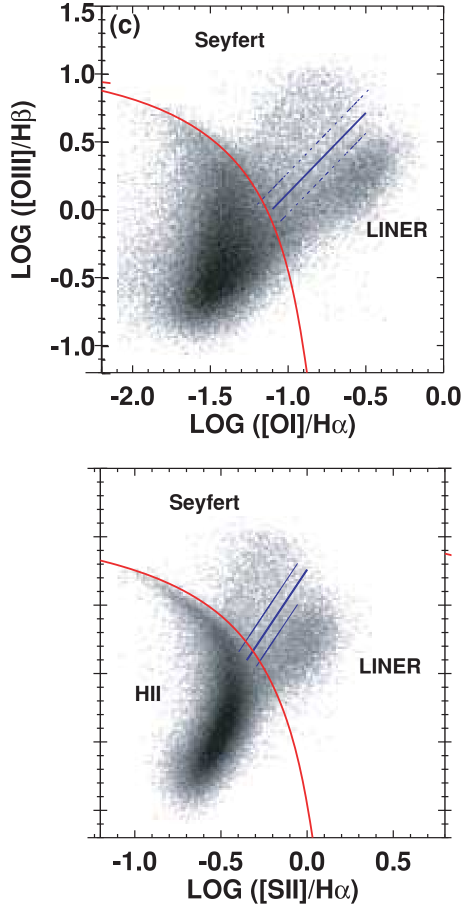
Ho et al. 1999

Classifying galaxies



Kewley et al. 2001



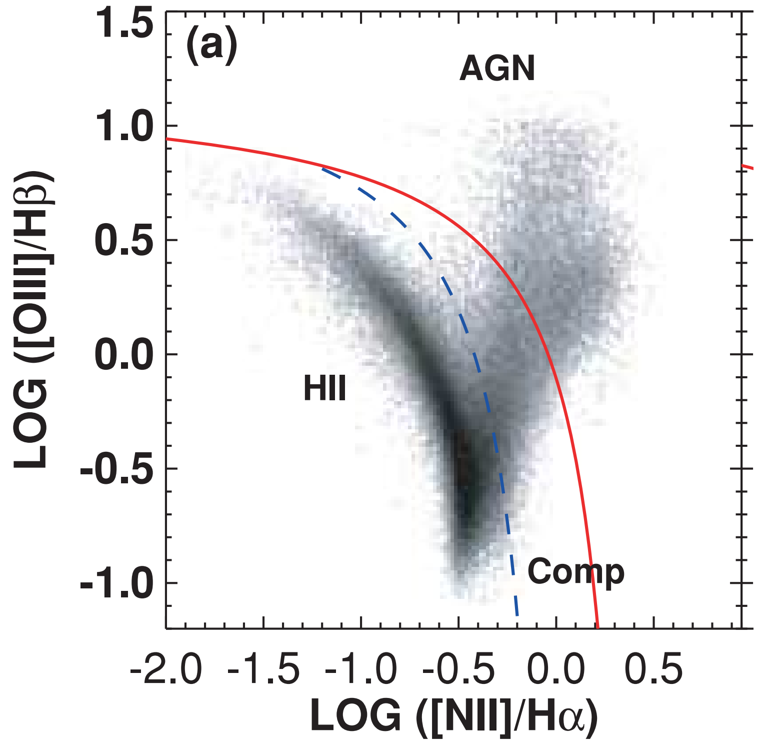


Kewley et al. 2006

Classifying galaxies

Basically an empirical scheme

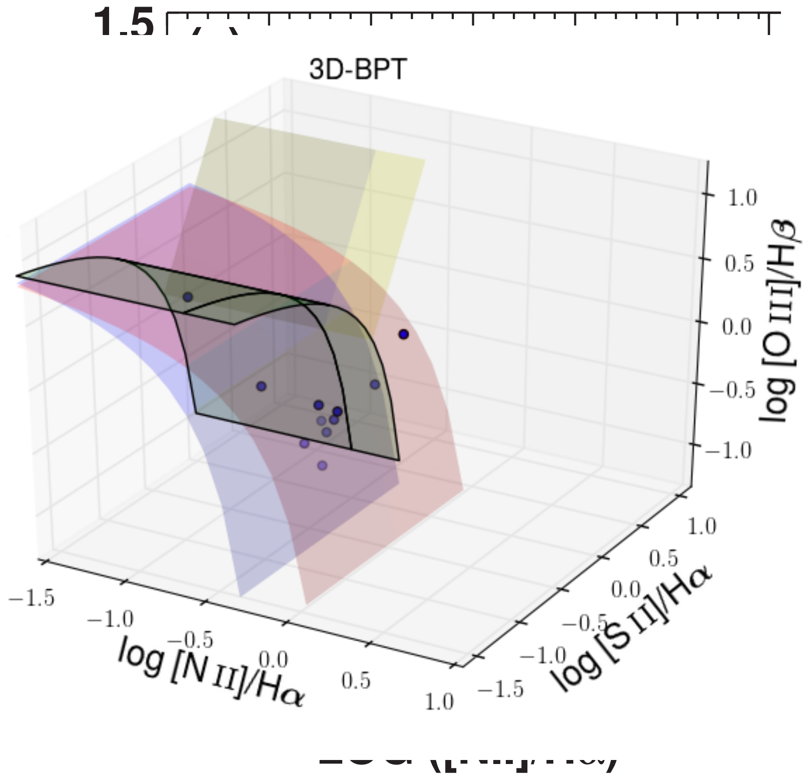
- Multi-dimensional parameter space
- Sparse sampling - but now large training sample available
- Uncertainties and intrinsic variance in training sample



➔ Use observations to define locus of different classes

Classifying galaxies

- Uncertainties in classification due to
  - measurement errors
  - uncertainties in diagnostic scheme
- Not always consistent results from different diagnostics



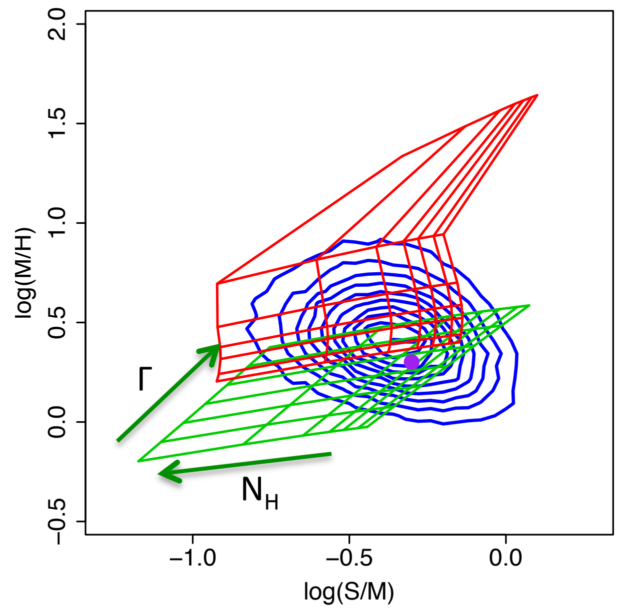
Maragoudakis et al in prep.

→Use ALL diagnostics together

→ Obtain classification with a confidence interval

Classification

- Problem similar to inverting Hardness ratios to spectral parameters
- But more difficult
  - We do not have well defined grid
  - Grid is not continuous



		N_H					
		0.250–0.500	0.125–0.250	0.075–0.125	0.050–0.075	0.025–0.050	0.010–0.025
	1.75–2.00	**11.36%**	**13.93%**	**3.35%**	**1.00%**	0.53%	0.24%
	1.50–1.75	**5.56%**	**13.70%**	**5.99%**	**2.34%**	**1.70%**	0.67%
Γ	1.25–1.50	**1.80%**	**7.76%**	**5.61%**	**3.11%**	**2.82%**	**1.56%**
	1.00–1.25	0.38%	**2.71%**	**2.87%**	**2.26%**	**2.33%**	**1.58%**
	0.75–1.00	0.07%	0.54%	**0.82%**	0.75%	**1.00%**	0.81%
	0.50–0.75	0.01%	0.09%	0.15%	0.18%	0.23%	0.17%

Taeyoung Park's thesis

Summary

- Model selection in multi-component 2D image fits
- Joint fits of datasets of different sizes
- Classification in multi-parameter space
  - Definition of the locus of different source types based on sparse data with uncertainties
  - Characterization of objects given uncertainties in classification scheme and measurement errors

All are challenging problems related to very common data analysis tasks.

Any volunteers ?