Unsupervised learning in a Nutshell

A probabilistic approach for galaxy emission-line classification

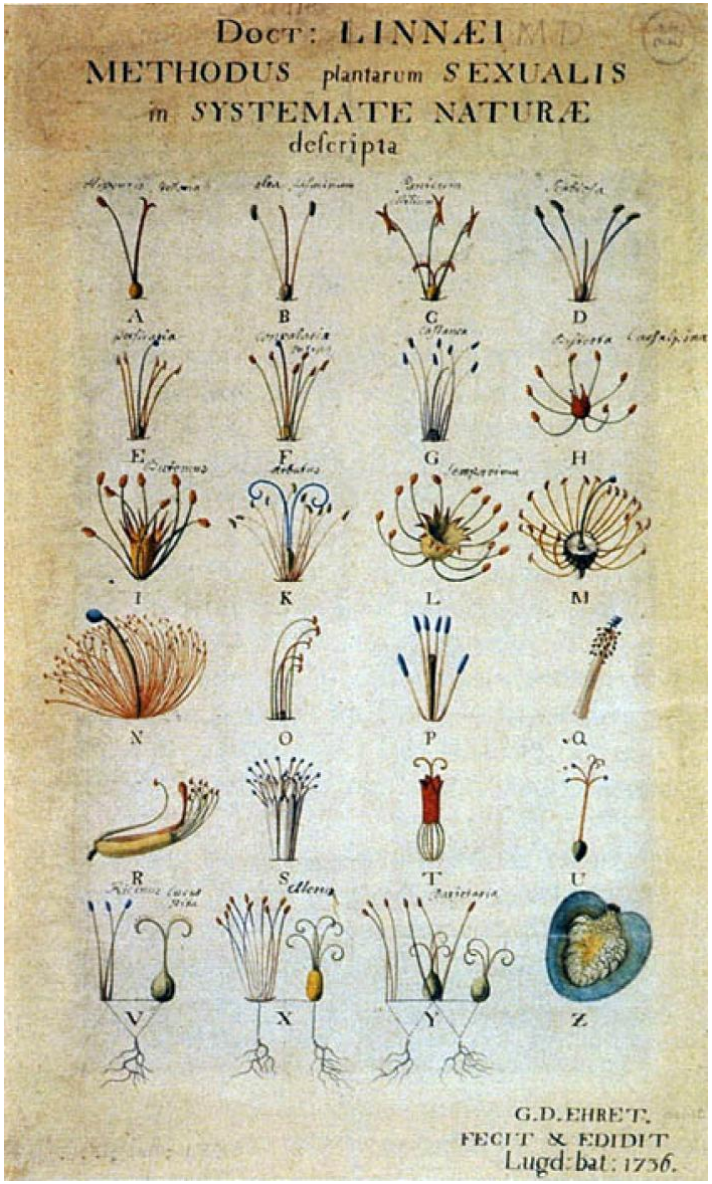
Rafael S. de Souza

EWASS-Prague, 2017
Astroinformatics

The first step in wisdom is to know the things themselves, Carl Linnaeus

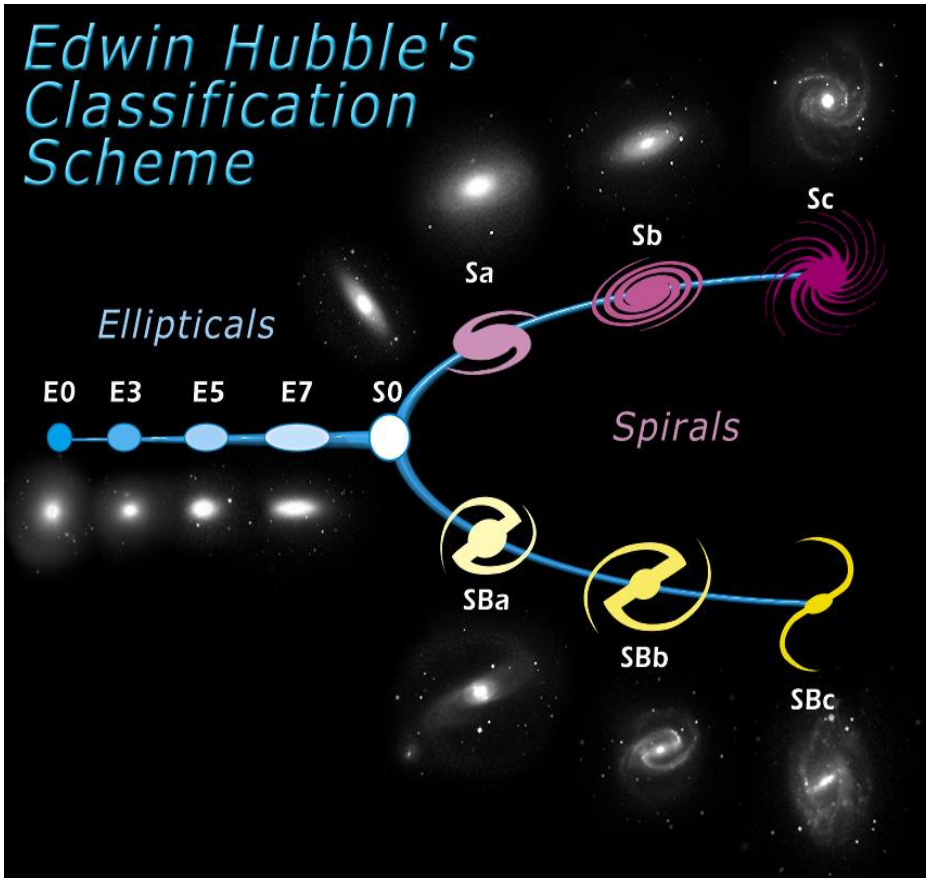
Classification and name-giving will be the foundation of our science.

Hierarchical classification of organisms based on similarities and differences



Classification of objects is a major driver in natural sciences

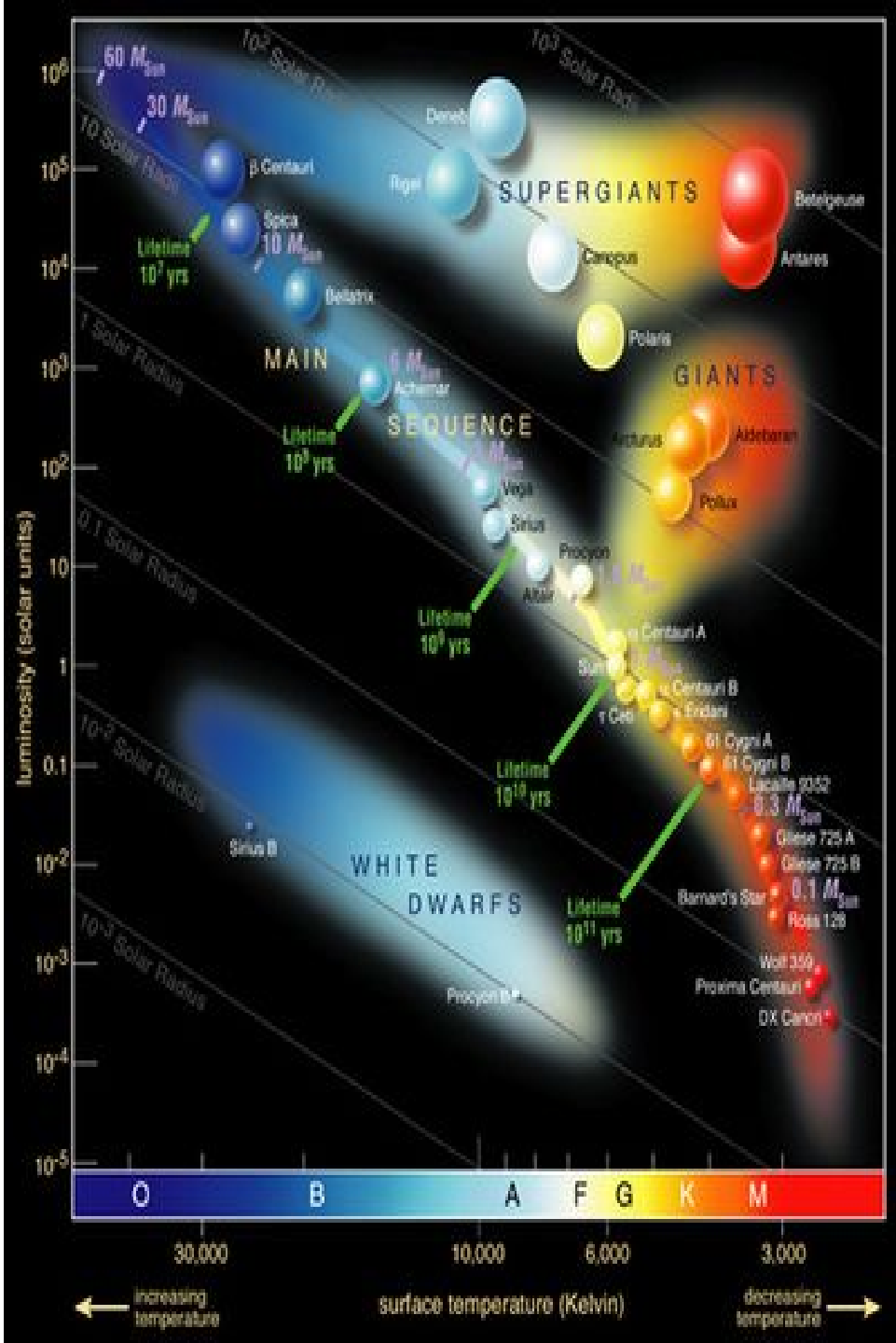
Morphological feature space



HR diagram: 2D-Stellar classification

Subspace of features

Ad hoc dimensionality reduction

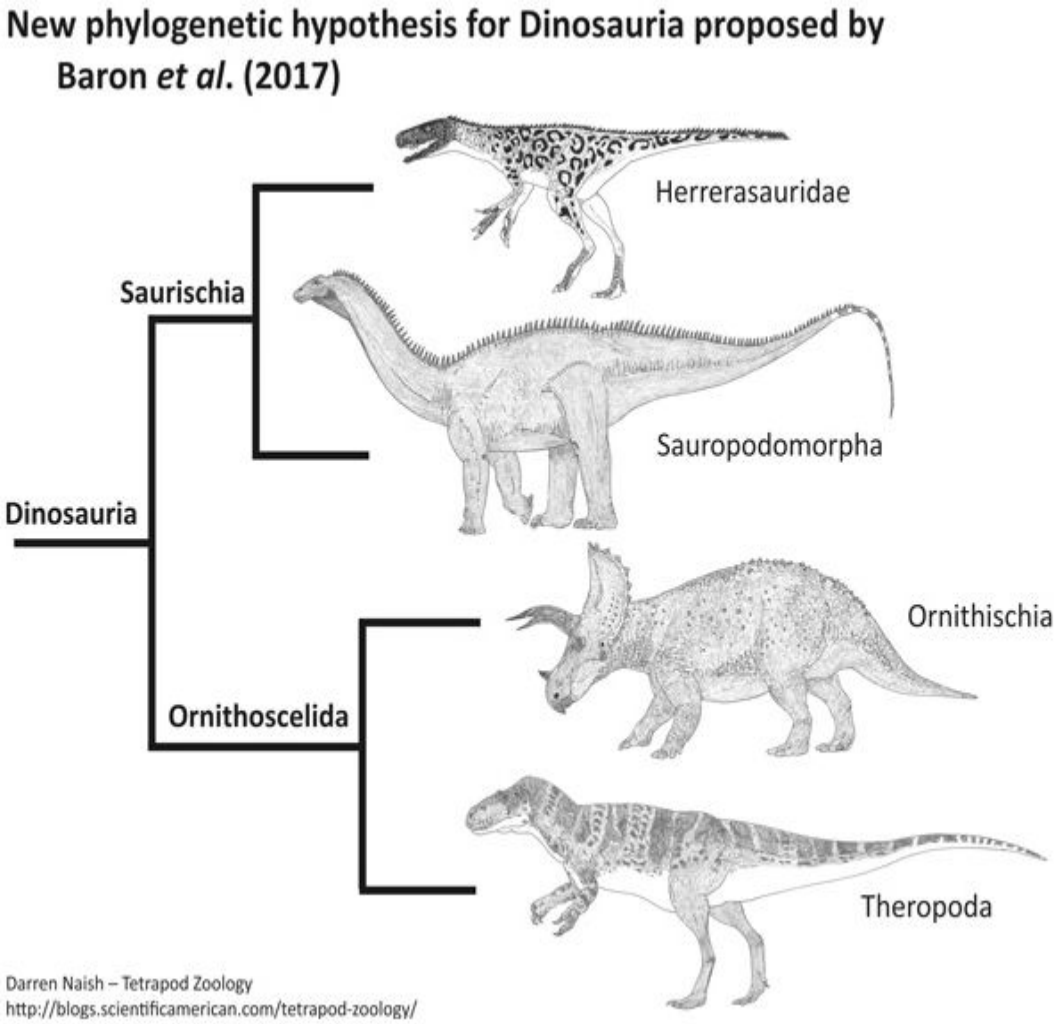


Taxonomical classification

Feature space strongly affects clustering

- Morphological
- Genetic
- Eating habits

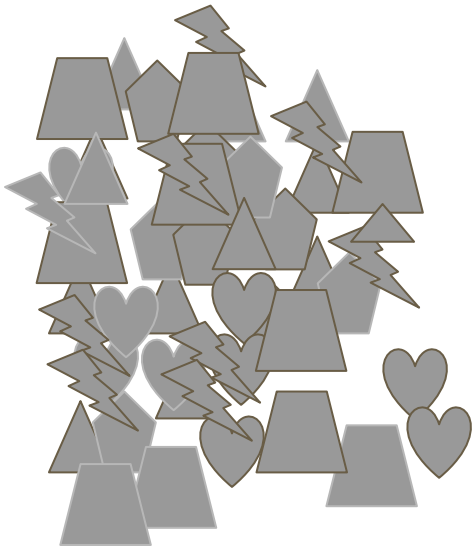
⋮



Unsupervised Learning Pipeline

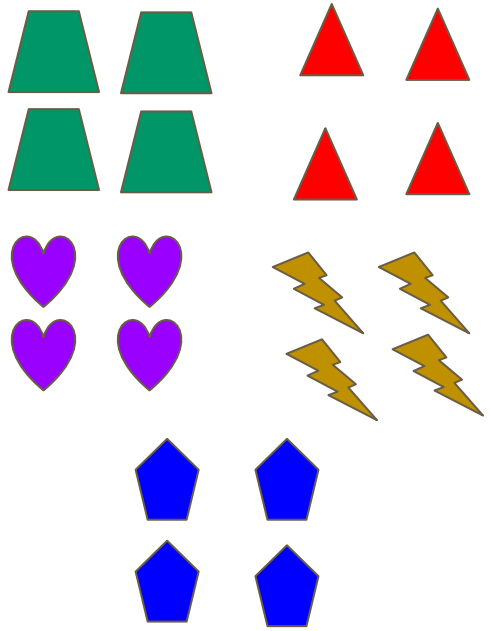
Raw data

Clustering



Feature

space



External

validation

Validation

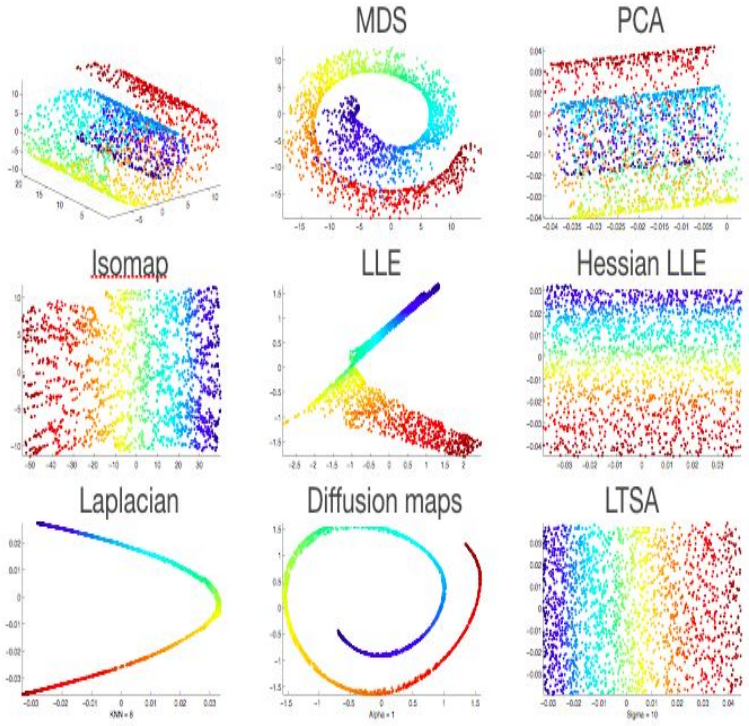
Internal

Literature --Name given

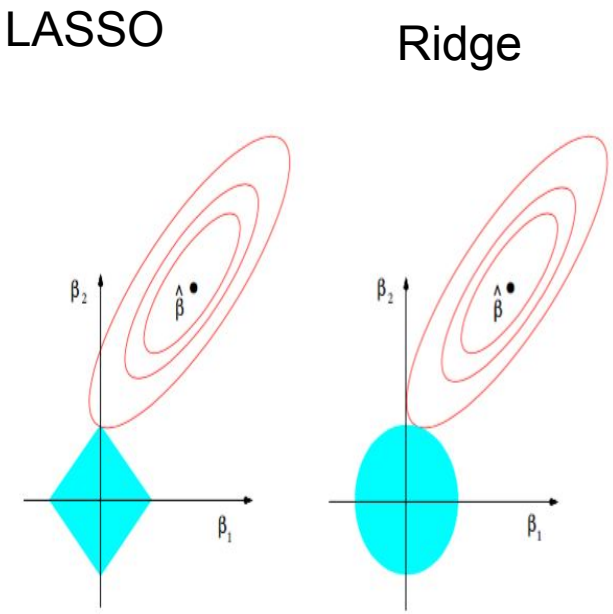
Data-driven groups

Pre-Processing

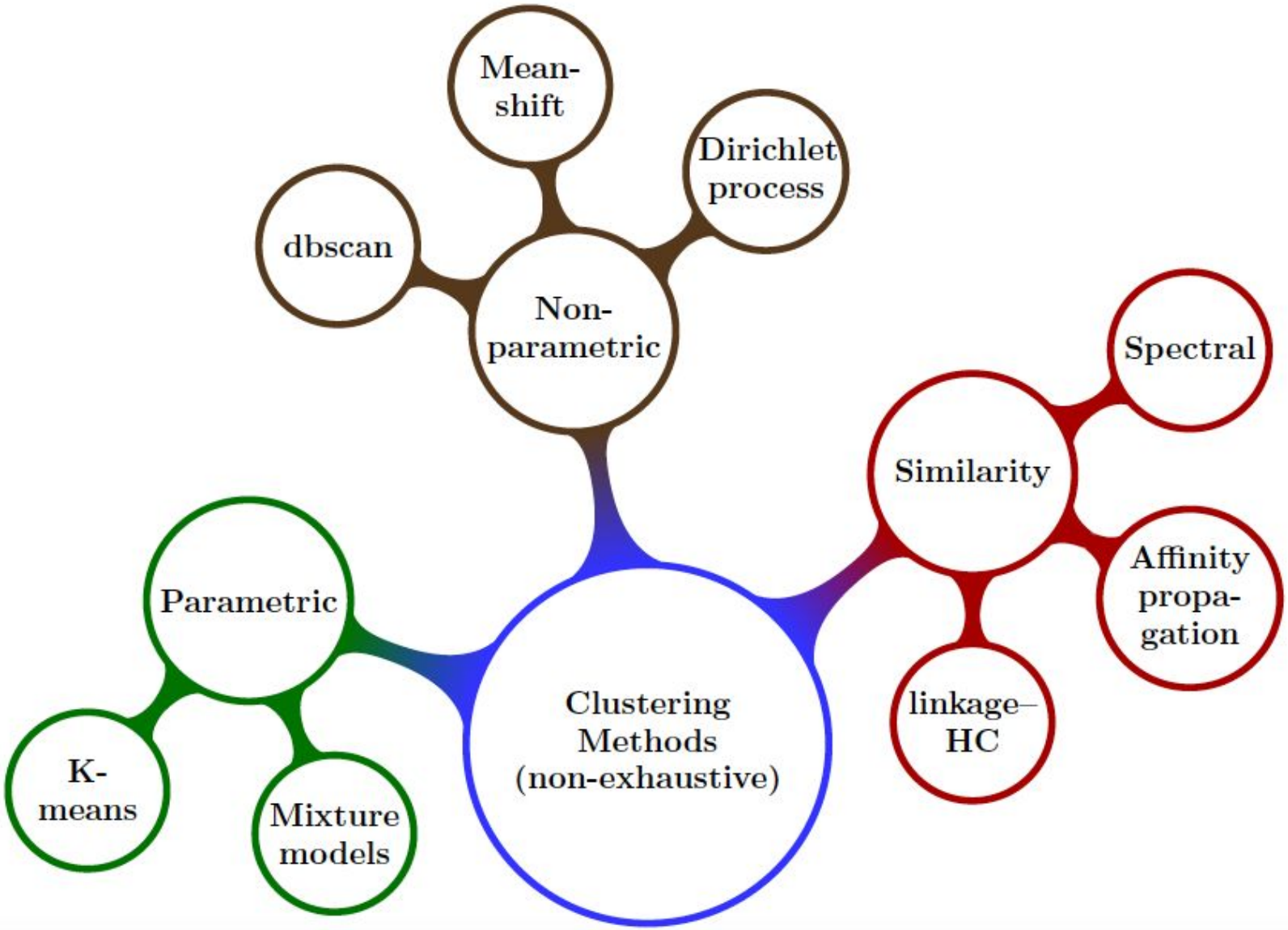
Feature extraction/dimensionality



Feature Selection

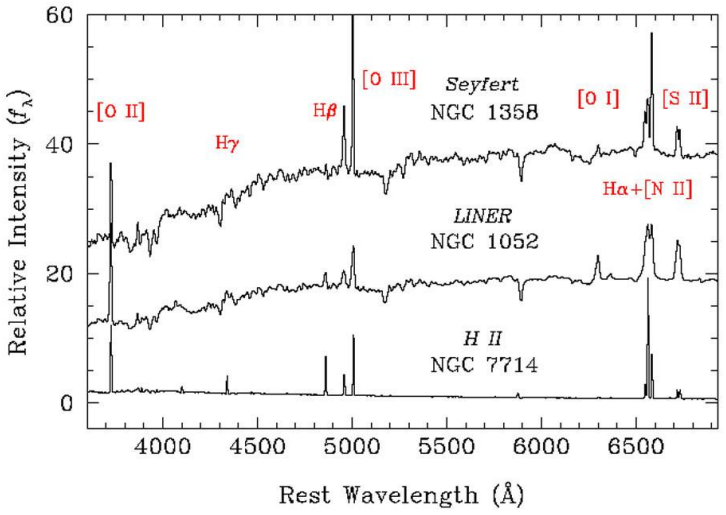


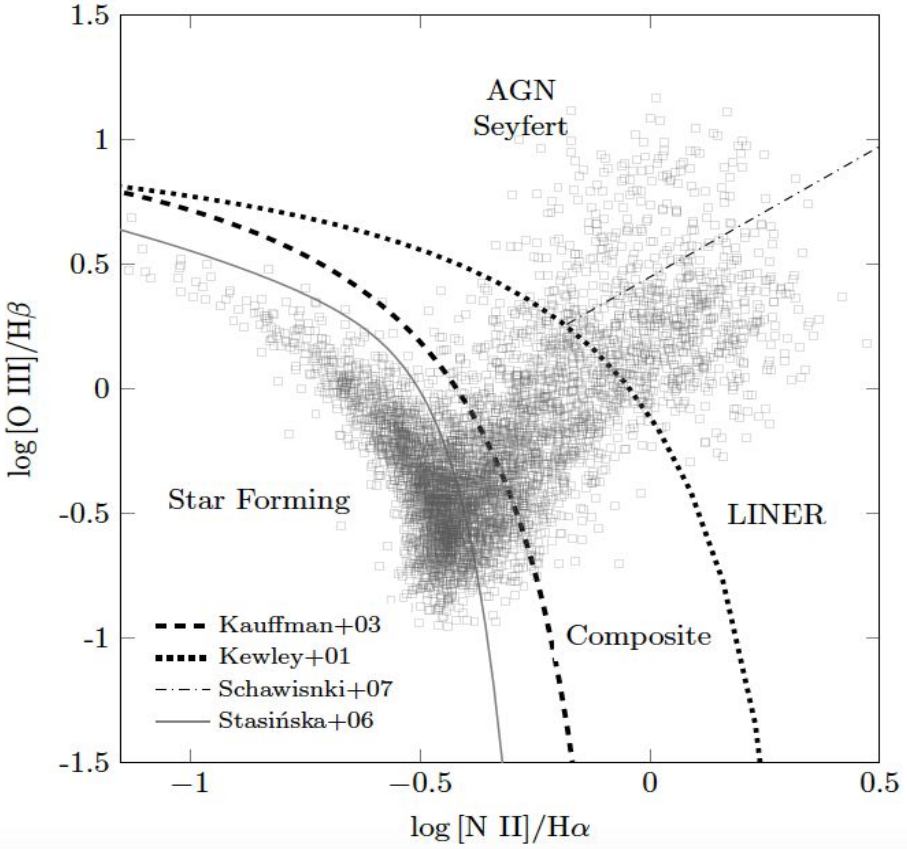
Clustering methods

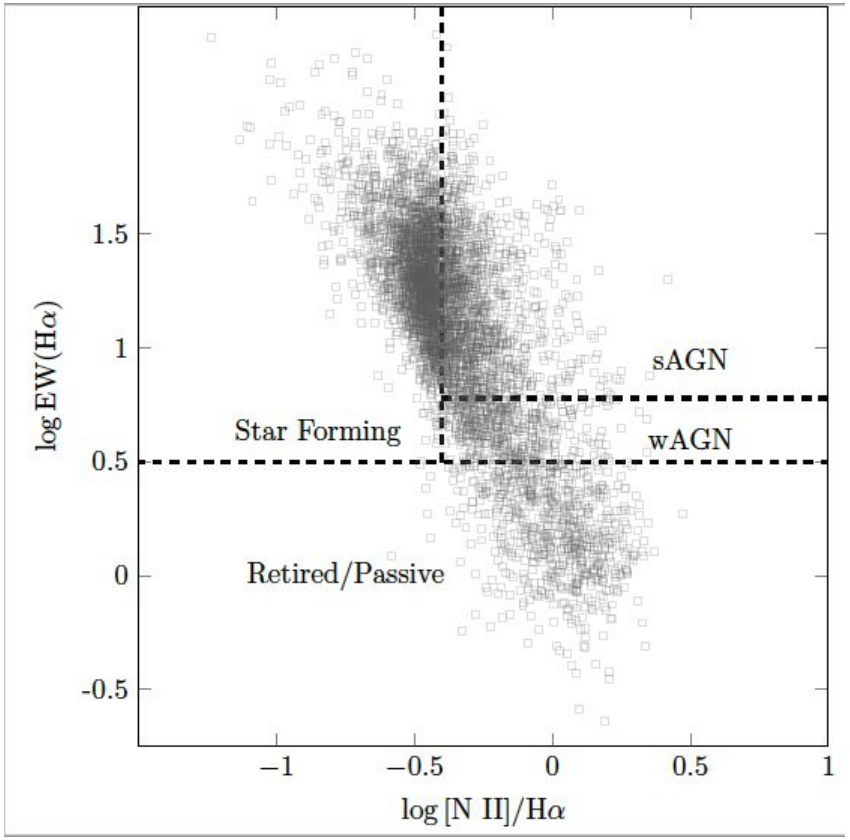


Scientific Case

Test a GMM to discriminate galaxies based on their ionization sources

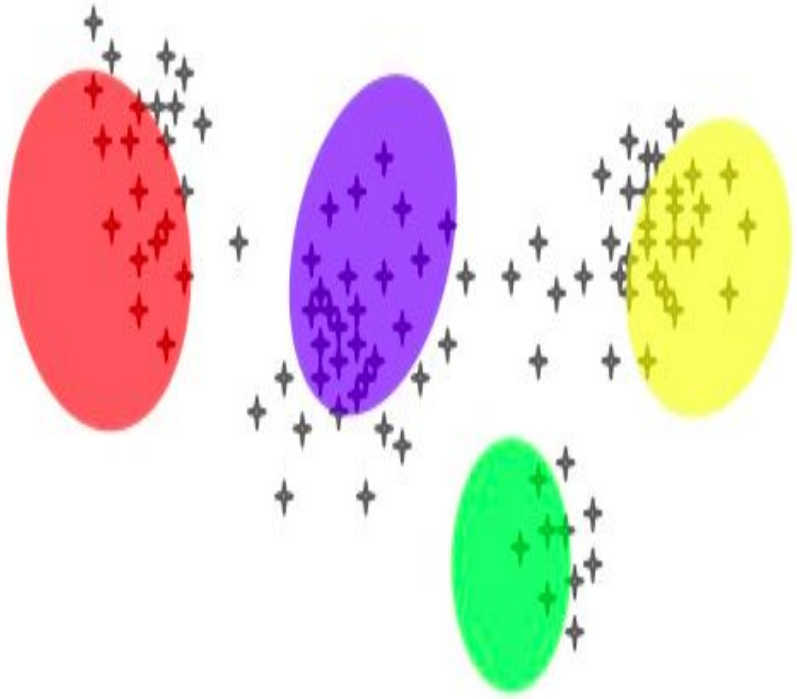






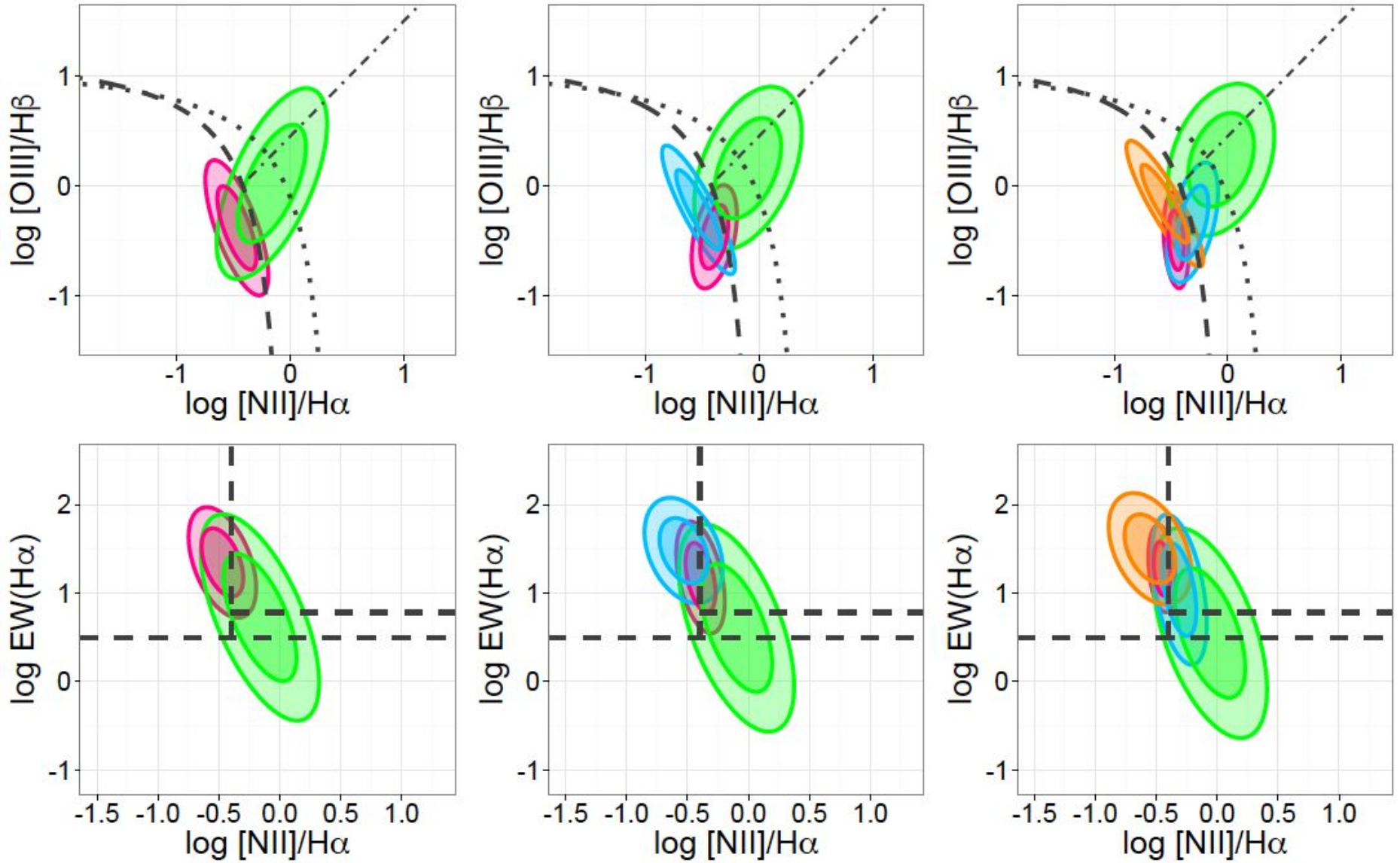
Make it simple, but solid: Exploit model-based GMM

- Soft classification
- Generative model
- Stability
- Other Mixture models are possible: Gamma mixture, Poisson mixture, ...

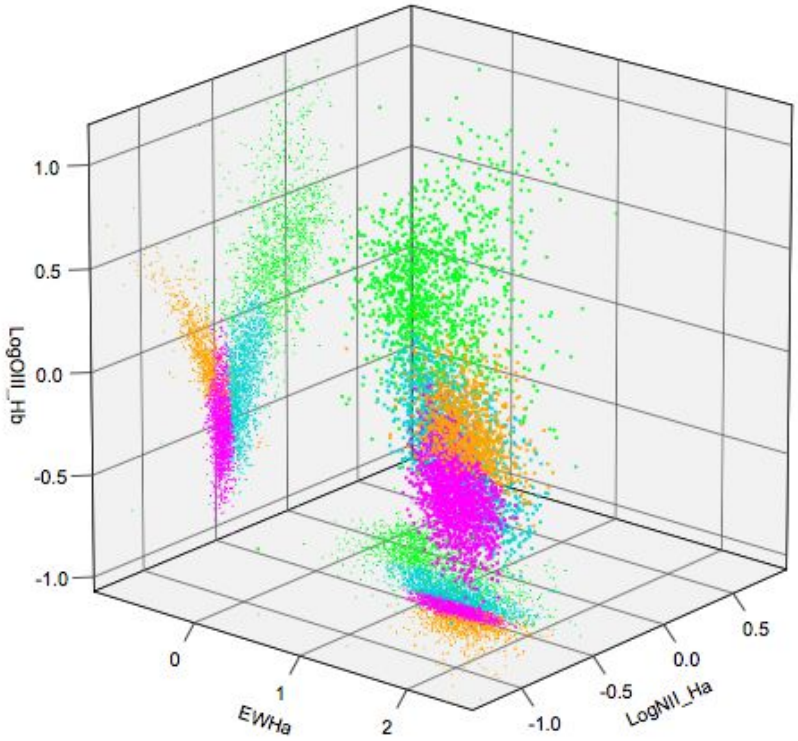


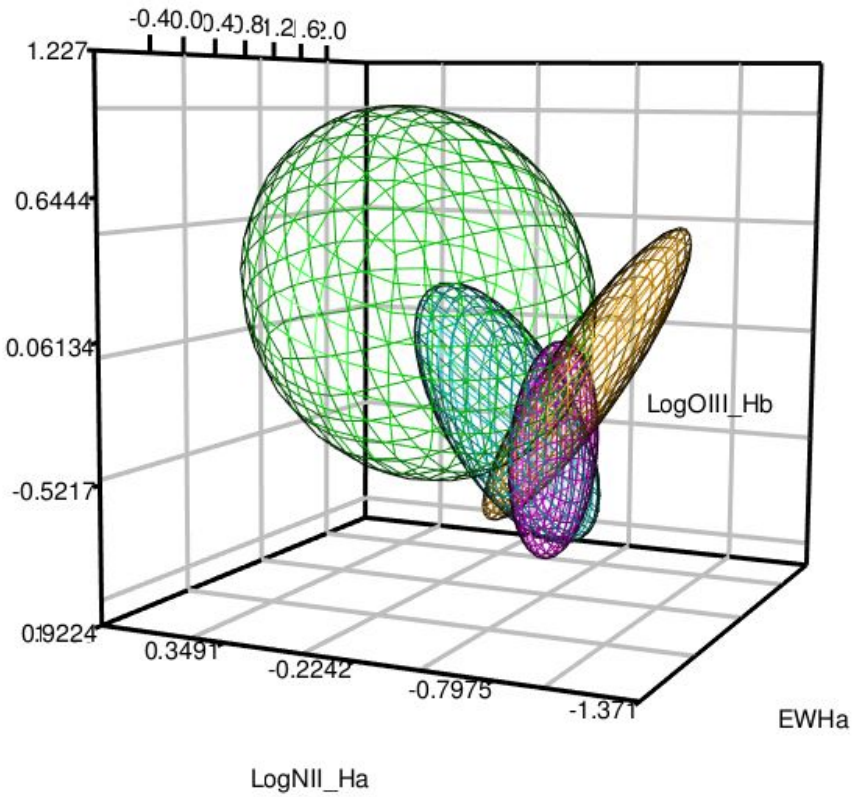
A probabilistic approach to emission-line galaxy classification

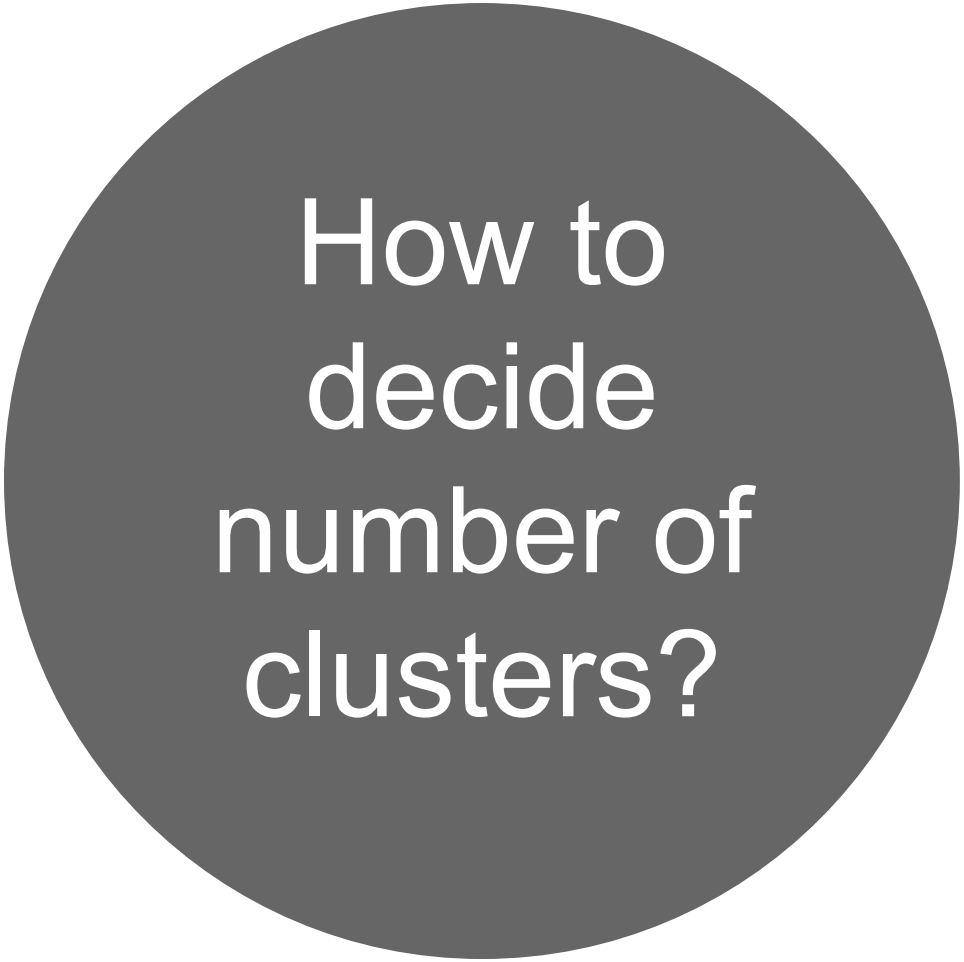
R. S. de Souza, M. L. L. Dantas, M. V. Costa-Duarte, E. D. Feigelson, M. Killedar, P.-Y. Lablanche, R. Vilalta, A. Krone-Martins, R. Beck, F. Gieseke



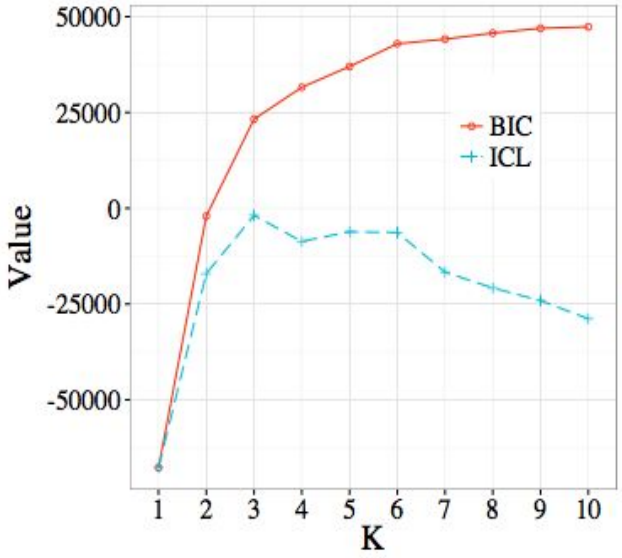
Full solution

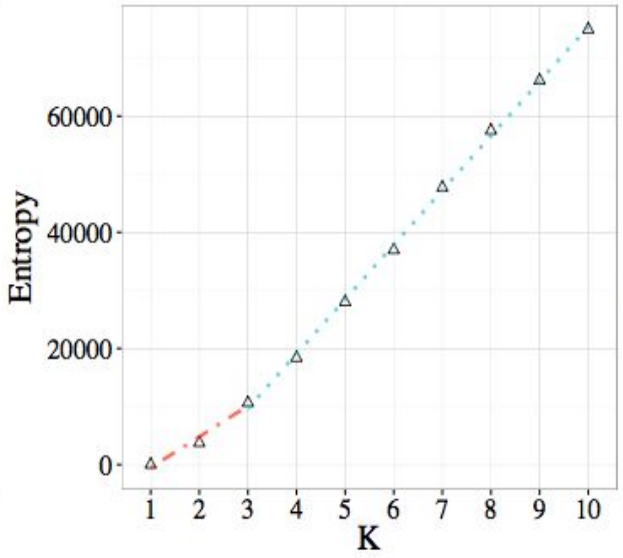


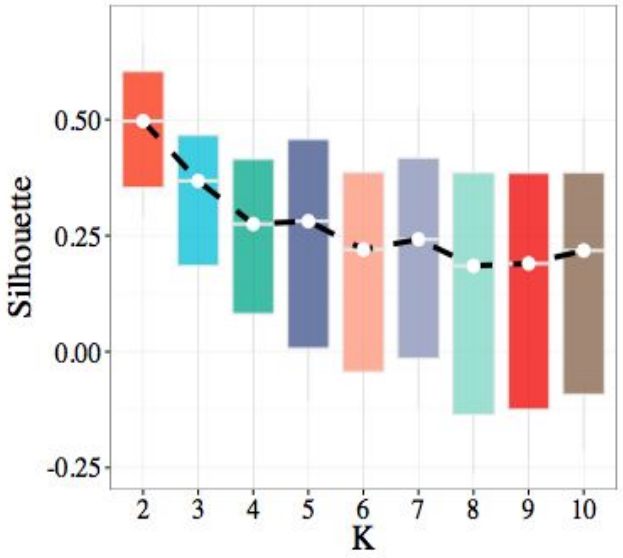



How to
decide
number of
clusters?

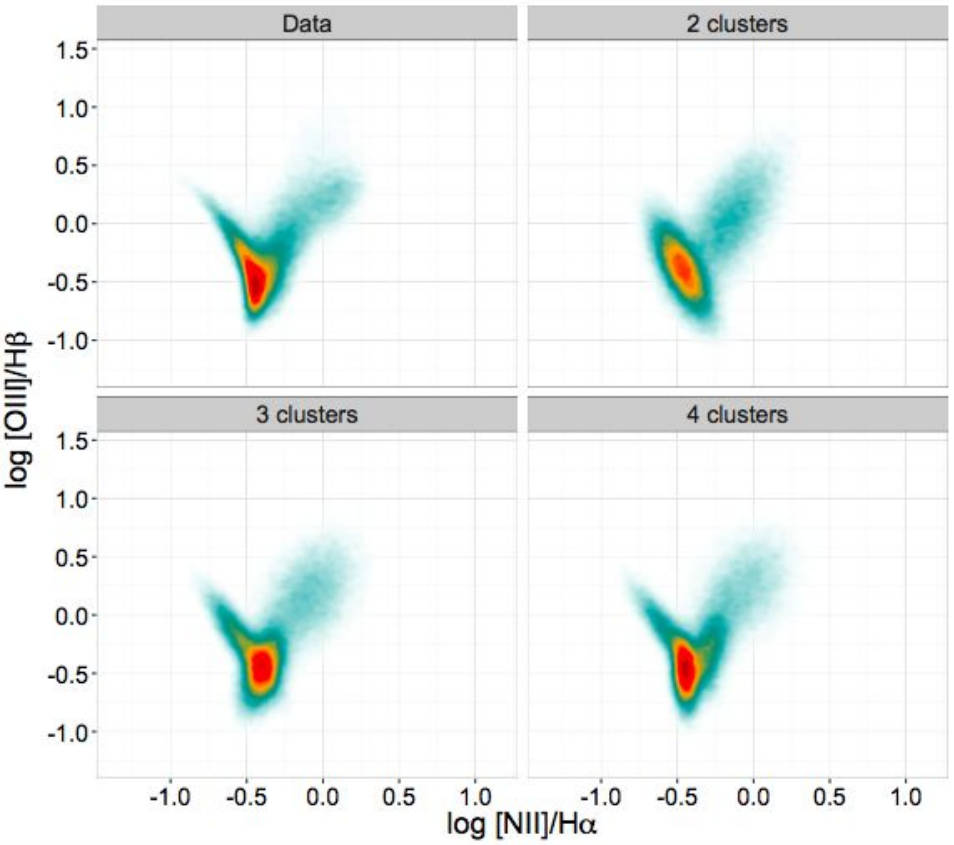
Few internal validation methods

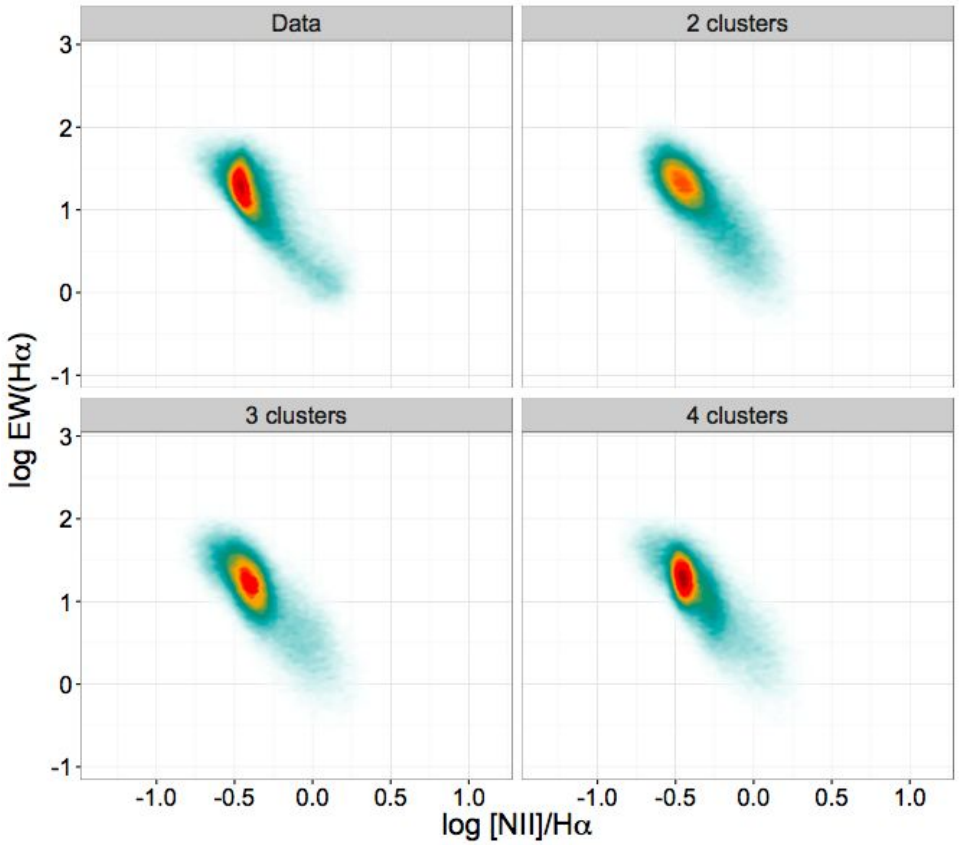




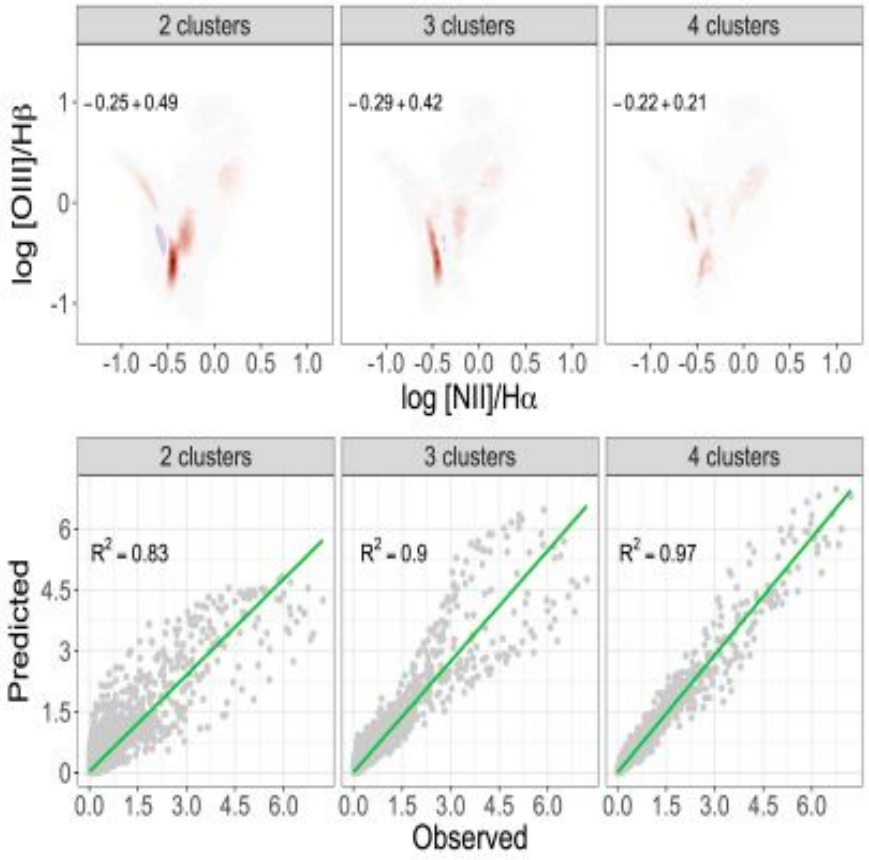


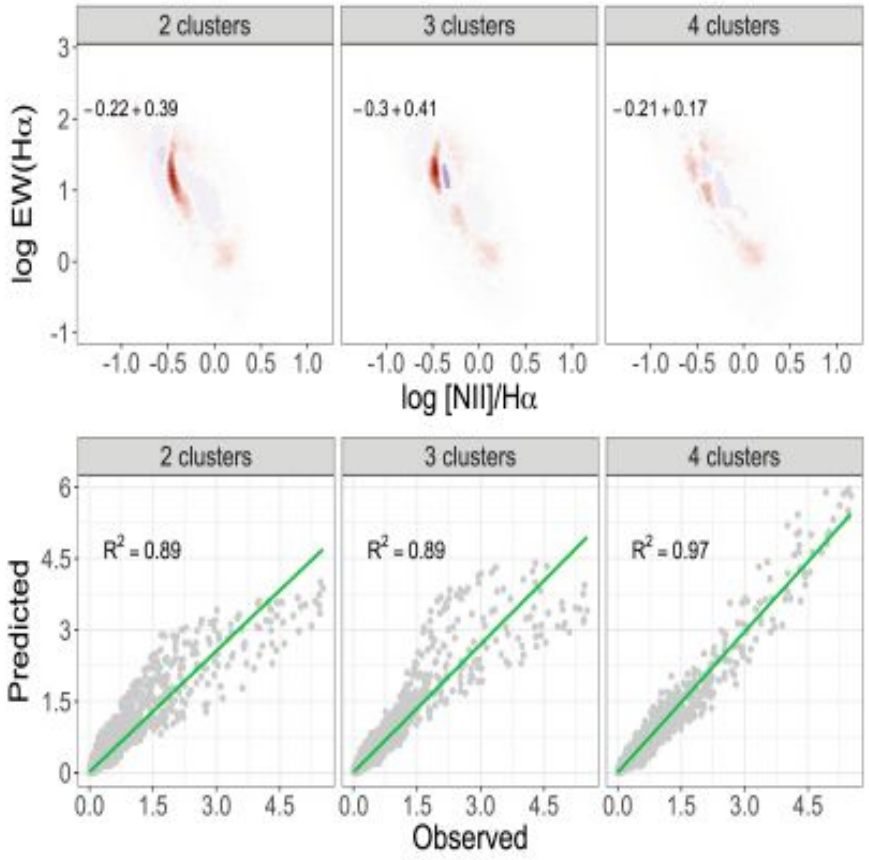
Explore sampling properties of GMMs to simulate synthetic data



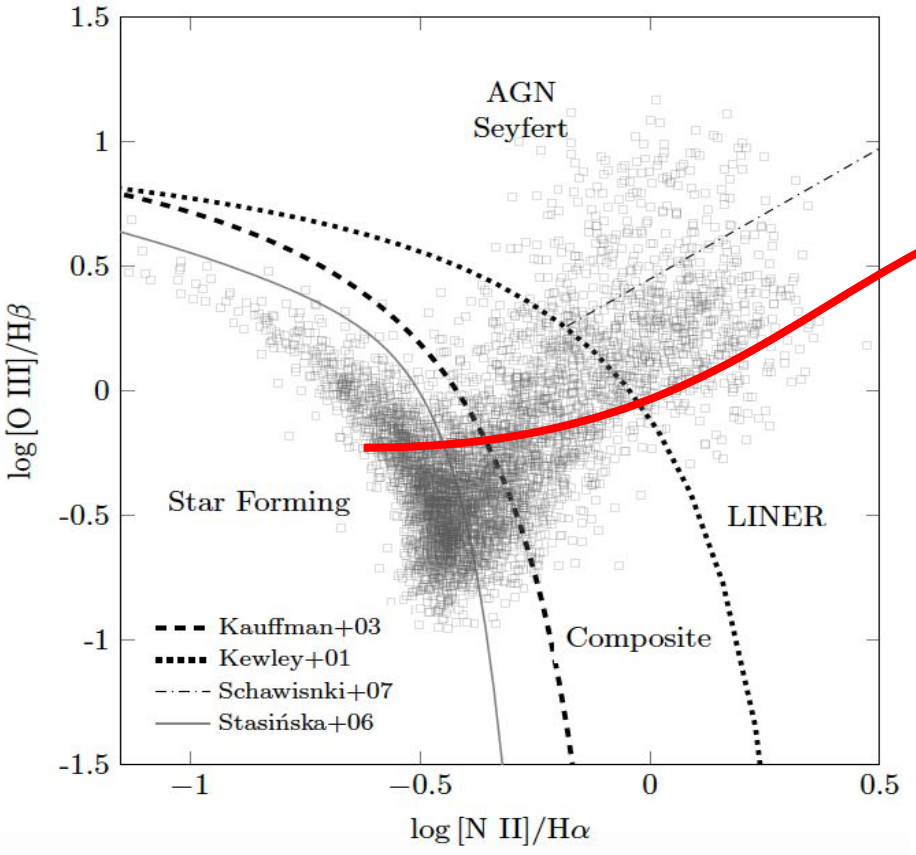


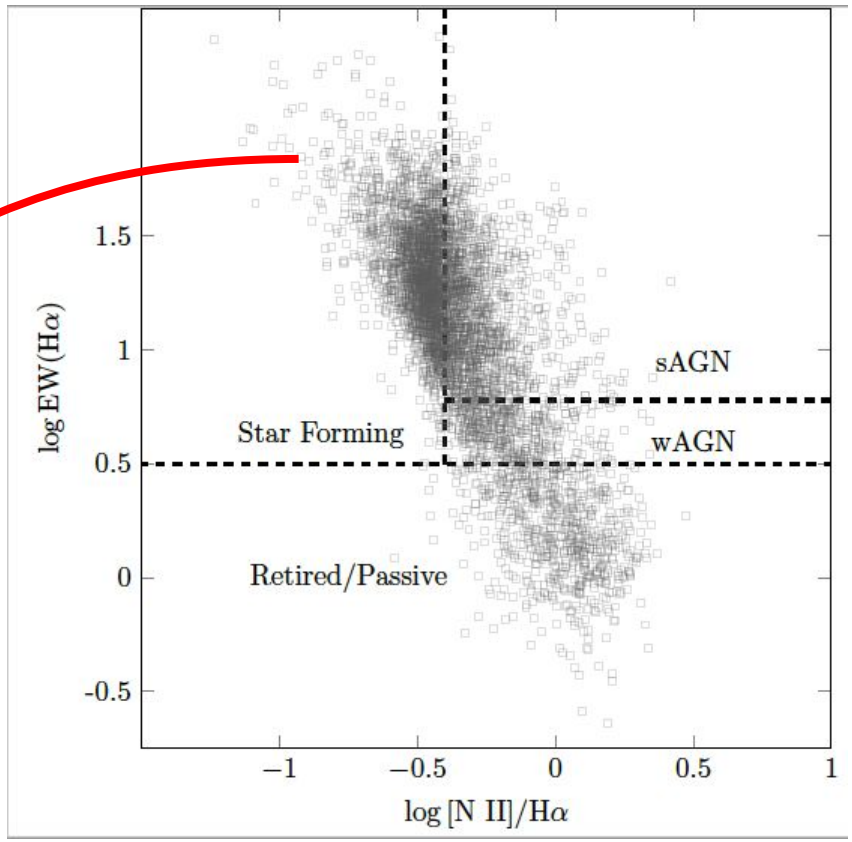
Model-based clustering allows residual analysis

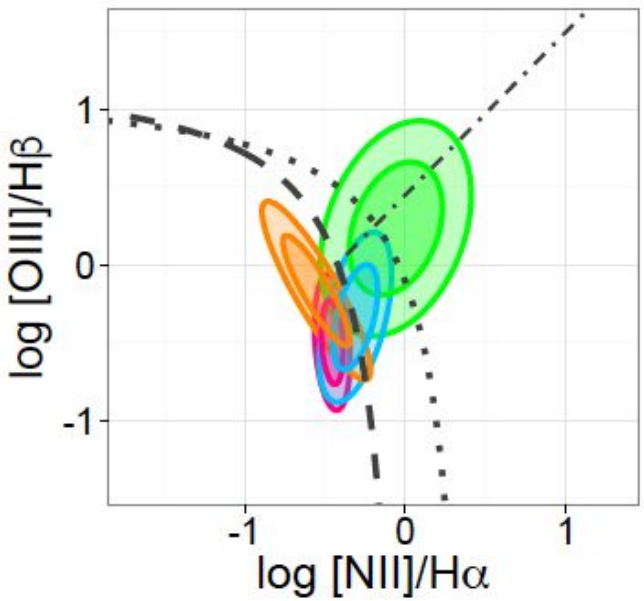


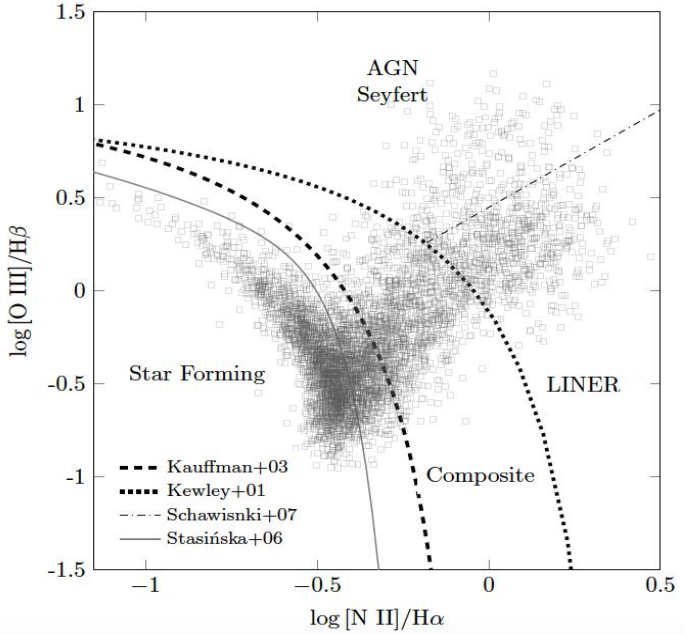


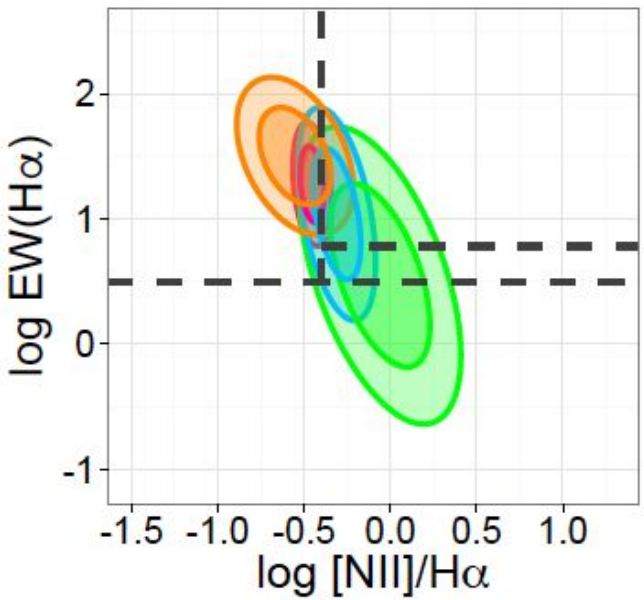
How to measure similarity among groups?

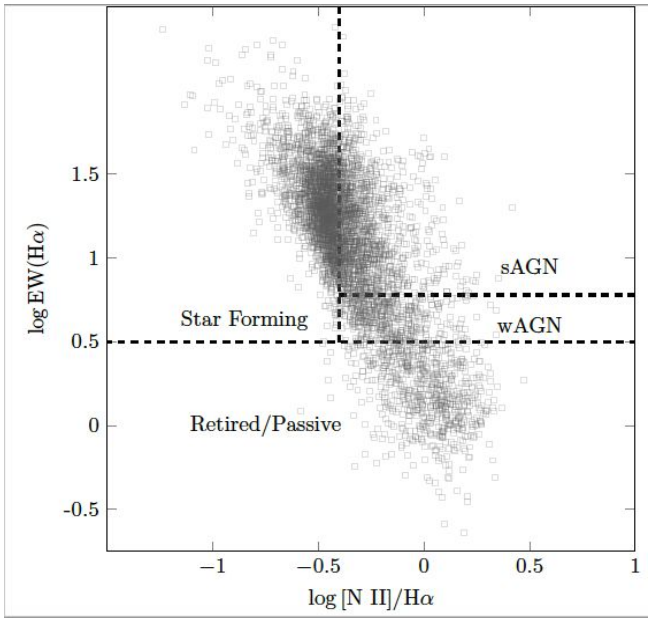





log [OIII]/Hβ
log [NII]/Hα
1
0
-1
-1
0
1

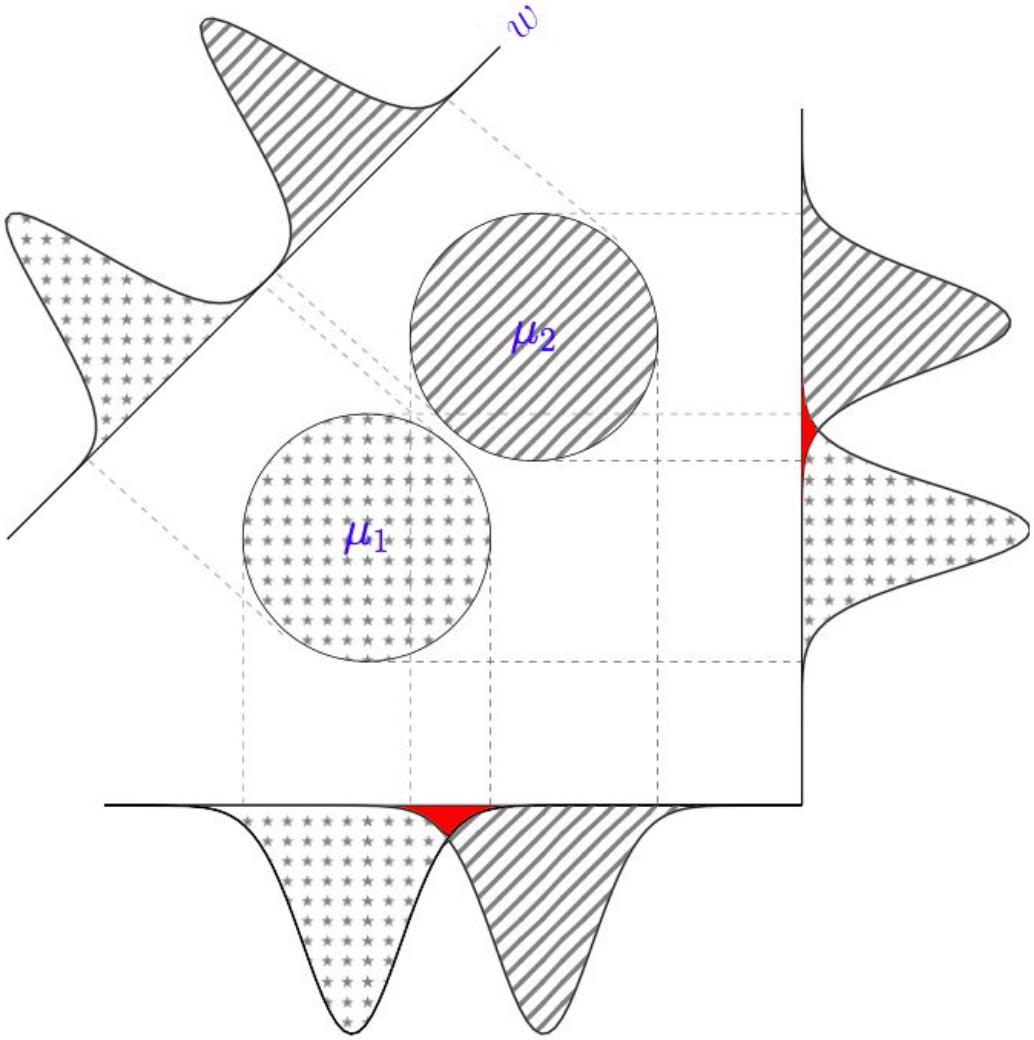

AGN
Seyfert
Star Forming
LINER
Composite
Kauffman+03
Kewley+01
Schawisnki+07
Stasińska+06
log [O III]/Hβ
log [N II]/Hα
1.5
1
0.5
0
-0.5
-1
-1.5
−1
−0.5
0
0.5


log EW(Hα)
log [NII]/Hα
2
1
0
-1
-1.5 -1.0 -0.5 0.0 0.5 1.0


sAGN
Star Forming
wAGN
Retired/Passive
log EW(Hα)
log [N II]/Hα
1.5
1
0.5
0
-0.5
−1
−0.5
0
0.5
1

Linear Discriminant Analysis

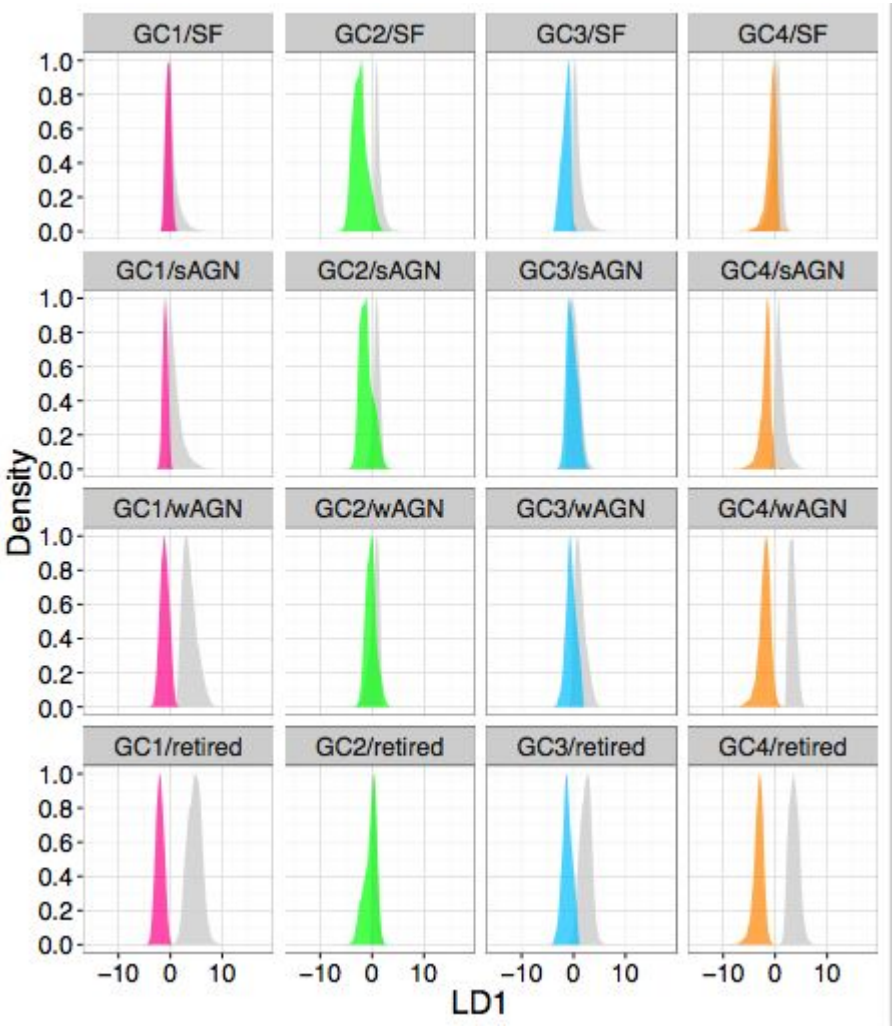
Project groups into a lower dimensional subspace, while preserving their distances



Kullback-Leibler divergence-relative entropy

$$D(f_i||f_j) = \int_{\mathbf{x}} f_i(\mathbf{x}) \ln \frac{f_i(\mathbf{x})}{f_j(\mathbf{x})} d\mathbf{x}.$$

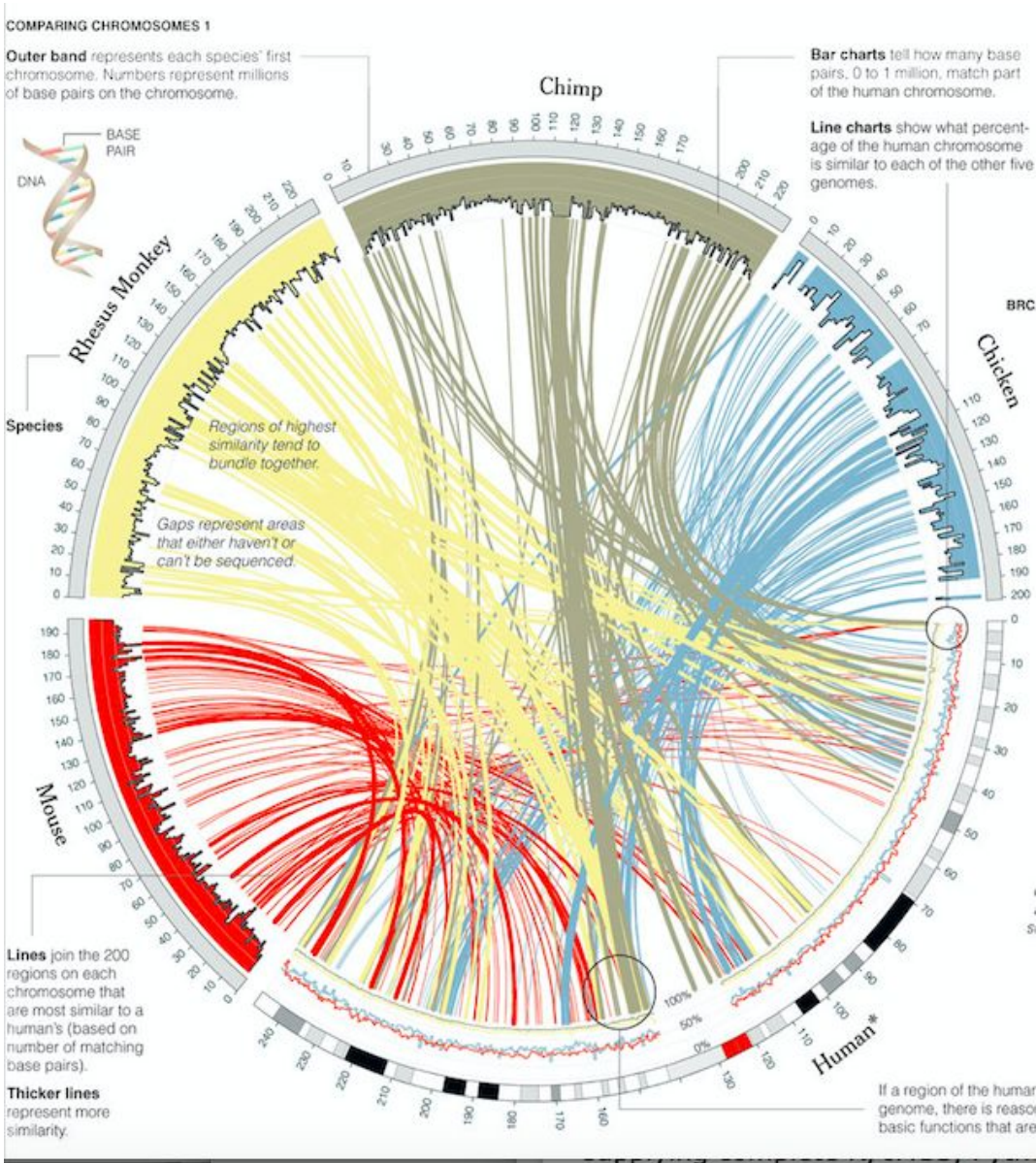
Measures how one probability distribution diverges from a second expected probability distribution



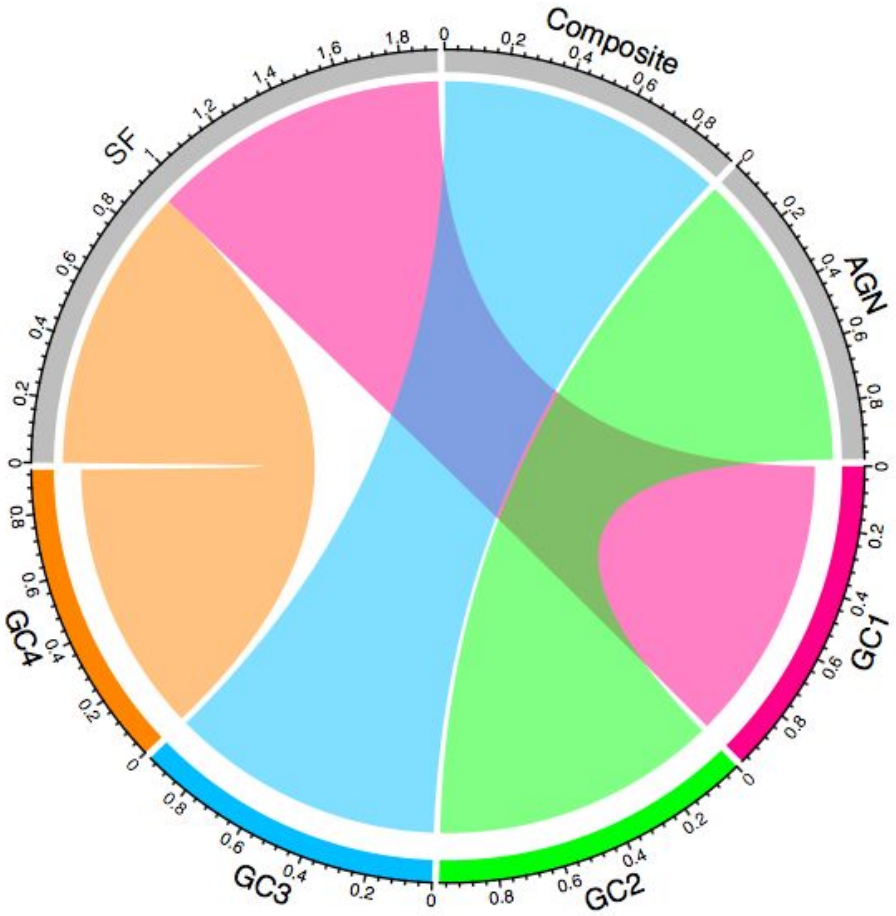
Genomics

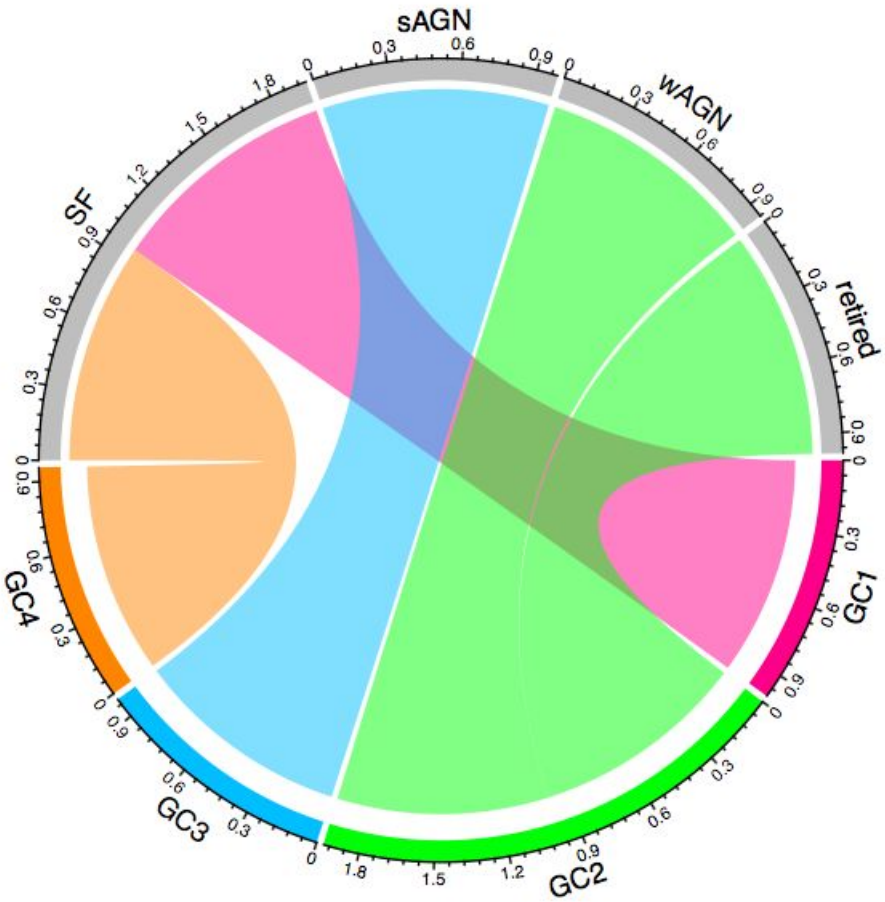
Visualize multiple associations

Chord diagrams--borrowed from Bioinformatics

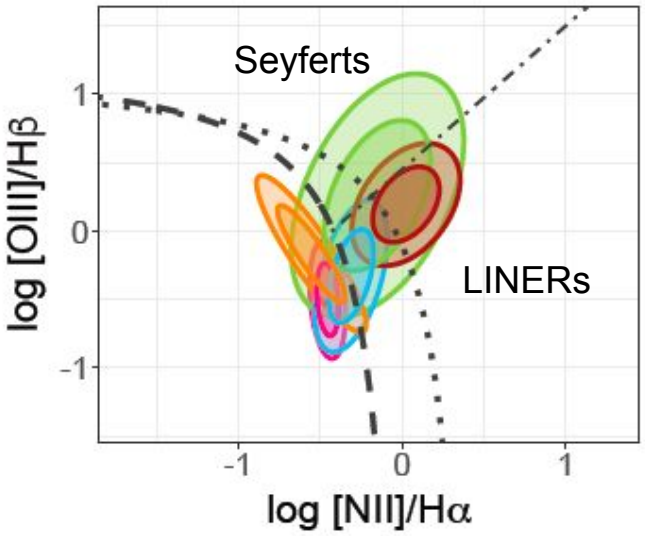


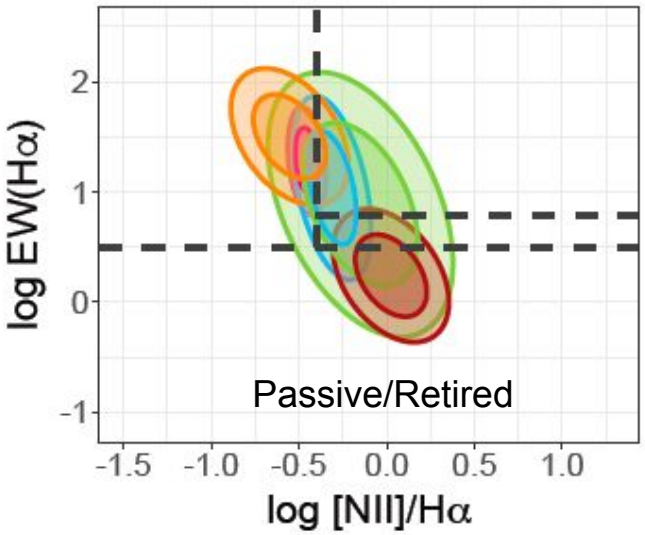
Statistical vs Astrophysical classes

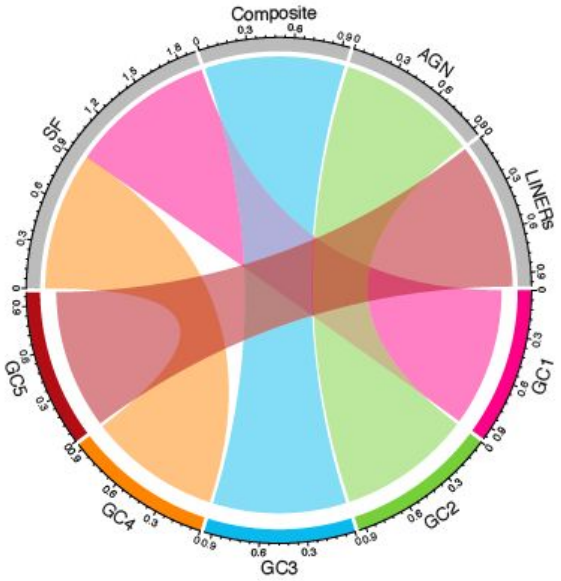


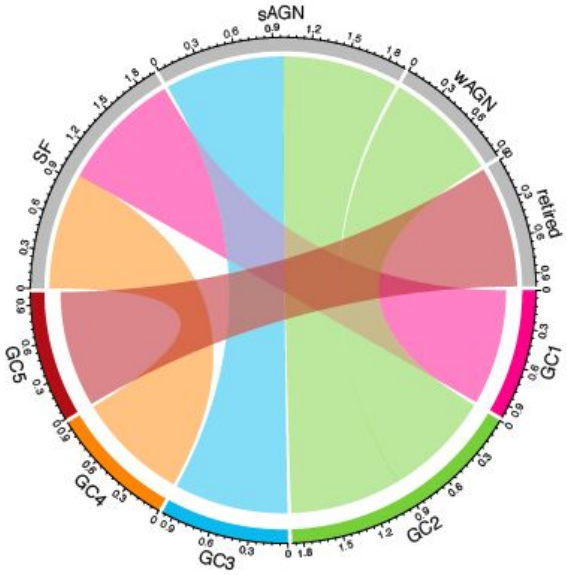


An extra group spots the dichotomy Seyfert/LINER









Remarks

- Soft classification provides more realistic predictions, account for boundaries uncertainties;
- Generative models combined with spatial analysis provides a solid recipe to define groups;
- External clustering validation adds semantics to groups via previous domain knowledge.

GMM_Catalogue

Catalogue with probabilistic classification of galaxies based on their ionization source

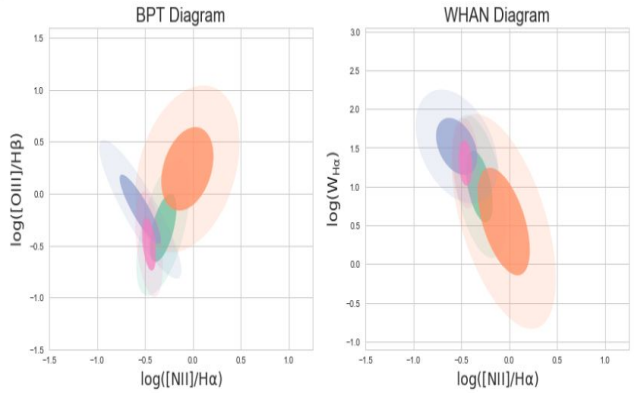
View the Project on GitHub

```
In [17]: plt.figure(figsize=(14,6))

         ax1 = plt.subplot(121)
         for i in enumerate(['C0','C1','C2','C3']):
             plot_BPT_ell(gmm4.covariances_[i[0]], gmm4.means_[i[0]], ax=ax1, col=i[1])

         ax2 = plt.subplot(122)
         for i in enumerate(['C0','C1','C2','C3']):
             plot_WHAN_ell(gmm4.covariances_[i[0]], gmm4.means_[i[0]], ax=ax2, col=i[1])

         set_plt_style()
```



This project is maintained by COINtoolbox

Hosted on GitHub Pages — Theme by orderedlist

Galaxy Classification via Gaussian Mixture Models

Catalogue with probabilistic classification of galaxies based on their ionization source using Gaussian Mixture Models

This is one of the products of the third edition of the COIN Residence Program, which took place in August/2016 in Budapest (Hungary).

The catalogue is given bellow. Check the individual folders for detailed information on the files presented here.

GMM Catalogue

This catalogue was designed to provide soft/probabilistic classifications for galaxies based on their ionization source. The data consists of a combination of the BPT and WHAN diagrams.

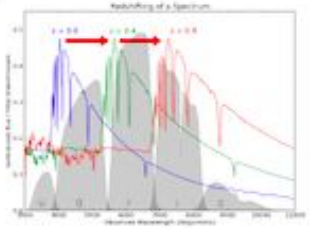
The probabilistic classification of galaxies via GMM is constructed using data from the SDSS DR7 and the SEAGal/STARLIGHT catalogues and is available at Catalogue.

Short tutorial for the use of GMM in Python

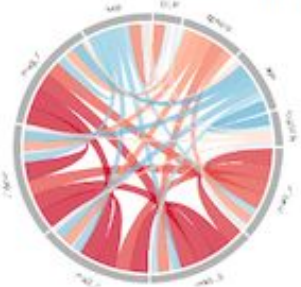
Created by Pierre-Yves and Madura Killedar and available at GMM Python tutorial

COINtoolbox

CosmoPhotoZ
Fast photo-z estimation via GLMs.

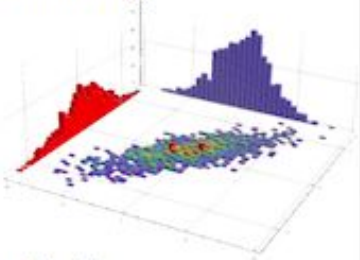


AMADA
Analysis of Muldimensional Astronomical DAtasets



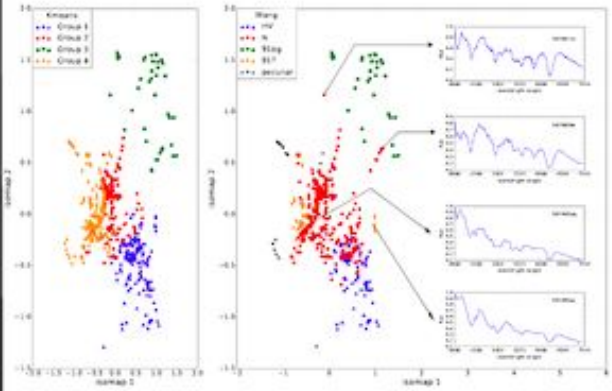
Cosmoabc

Likelihood free inference for cosmology

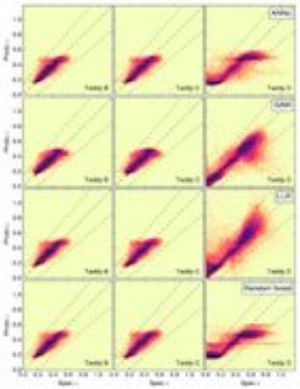


DRACULA

Dimensionality Reduction And Clustering for Unsupervised Learning in Astronomy



Happy and Teddy
Catalogues for realistic photo-z validation



COIN
Cosmostatistics Initiative