A Knowledge Discovery Workflow for Blazars

Raffaele D'Abrusco

Harvard-Smithsonian Center for Astrophysics

Knowledge Discovery workflow

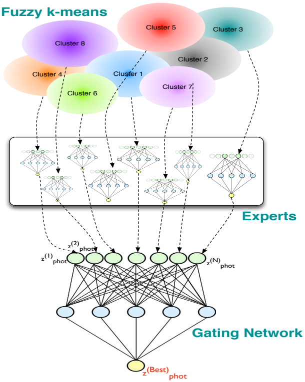
A Knowledge Discovery - *KD* - workflow is a sequence of analysis steps accomplished through distinct *KD* techniques to extract the most knowledge out of (usually) large amount of complex data.

Goals

- **Discovery**
  - Find new complex correlations;
  - Expand known correlations to more dimensions;
  - Find new simple correlations, so far overlooked;
- **Using the discovery**
  - Insight into astrophysics;
    - Classification, regression, new ways to look at things...
    - Optimized use of astronomical archival data;

An example: clustering & Quasars

A priori **knowledge** (spectroscopic quasars) and **Unsupervised Clustering can be used** to determine efficient ways **to extract candidate quasars from optical datasets and to optimize the training of regressors** (like neural networks) **for the determination of photometric redshifts** of extragalactic sources (Weak Gated Expert - *WGE*).



- The UC method splits the color space distribution of the sources into homogeneous aggregations;
- Multiple distinct *experts* (neural networks) are trained on different regions of the *features* space;
- The *gate* combines the outputs of the *experts* to maximize the accuracy of the redshift reconstruction and minimize the bias.

A question

What if the goal is not the improvement of the accuracy of a quantity obtained by regression (z_{phot}) or the classifications of sources (star *vs* quasars)?

What if the goal is to find out whether any pattern happens to occur in a generics *feature* space using unsupervised clustering techniques?

The *tenet*

The clusters in the *feature* space reflect similarities shared by cluster members.

Anisotropies in the distribution of clusters populations

relative to other observables reflect the existence of significant patterns.

The **CLaSPS** method

Clustering-Labels-Scores Patterns Spotter (CLaSPS)

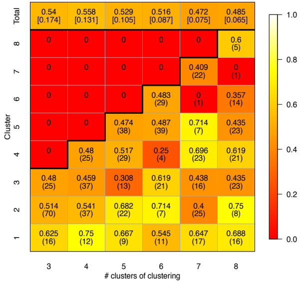
1. A UC algorithm is used to produce clusterings in the *parameter* space generated by any subset of the observables (the ***features***);
2. Other observables not employed for the clustering (the ***labels***), are used as *tags* to identify interesting set of clusters using the ***score***;
3. The patterns in the selected set of clusters are selected and studied.

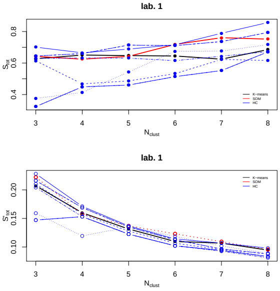
The choice of the clustering(s)

The degree of correlation between the distribution of cluster members in the *feature* space and their distribution in the *labels* space can be quantified using the *score*:

$$S_{tot} = \frac{1}{N_{\mathrm{clust}}} \cdot \sum_{i=1}^{N_{\mathrm{clust}}} S_i = \frac{1}{N_{\mathrm{clust}}} \sum_{i=1}^{N_{\mathrm{clust}}} \left(\sum_{j=1}^{M^{(j)}-1} \| f_{ij} - f_{i(j+1)} \| \right)$$

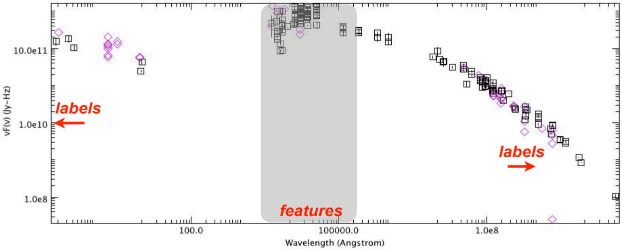
where f_{ij} is the fraction of members of the i-th cluster with values of the *label* in the j-th class.





An interesting finding

CLaSPS has been applied on a sample of AGNs with multi-wavelength observations spanning from radio to γ-rays to **characterize their SEDs in the colors space**.

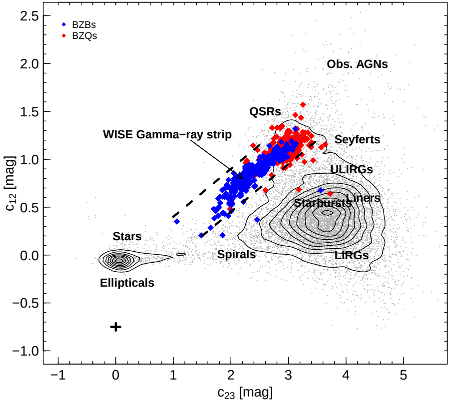


Dataset	$\rightarrow$	AGNs catalog
Features	$\rightarrow$	UV(*Galex*) + Optical(*SDSS*)+ NIR(*UKIDSS*) + IR(*WISE*)
Labels	$\rightarrow$	AGNs class., blazars spectral class. γ-ray emission

Three clusters mostly populated by blazars had large values of the *scores* using AGNs classification, the γ-ray detection and FSRQs-BL Lacs spectral classification for blazars as *labels*. **Such pattern depends on the peculiar *WISE* colors of blazars.**

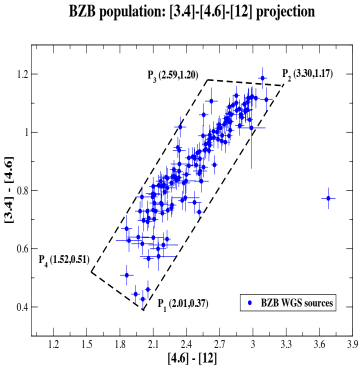
The *WISE* blazars *locus*

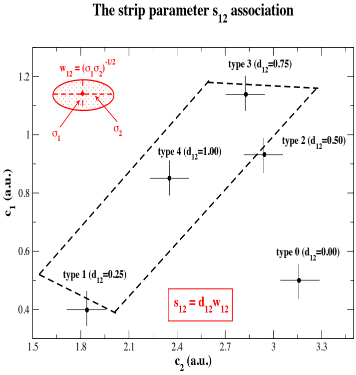
The blazars occupy a peculiar region of the mid-Infrared color space generated by *WISE* magnitudes. ***This pattern had been overlooked so far***.



A first *WISE* blazars *locus* modelization

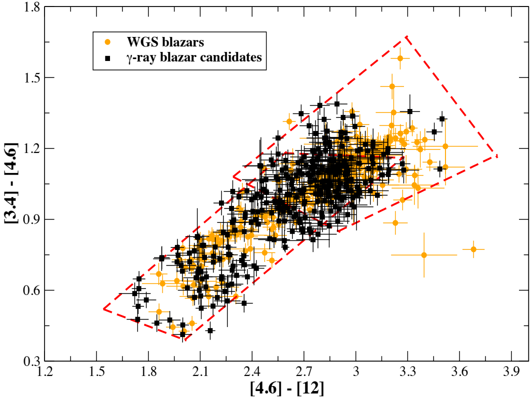
The first modelization of the wse blazars *locus* was created by determining the boundaries for each color-color plane projection to contain a minimum fraction of total sources (95%). Regions mostly occupied by BL Lacs and FSRQs have been modeled separately.





Unidentified *Fermi* γ-ray sources

Out of **313** "clean" 2FGL Unidentified γ-ray sources within the *WISE* Preliminary Release footprint ($\sim$55% of the sky), we have associated **156** to *WISE* candidate blazars.



A better modelization

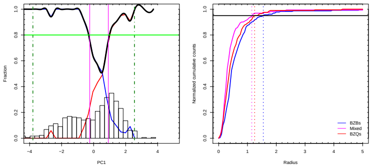
The *WISE* blazars *locus* can be also thought as a *classifier* whose parameters can be optimized by supervised learning, just like a **neural network**.

This systematic approach provides:

- A quantitative criterion to pick the “best” model of the *locus* in terms of:
  - Accuracy of the reconstruction of the *locus*;
  - Completeness vs efficiency of the classification;
  - Complexity of the geometrical model;
- Extensibility (adding new constraints from other wavelengths/*features*);
- Easily updatability (new version of blazars catalog, *WISE* photometric dataset, release, etc.);

The new model

Best model: cylinder(s) in the Principal Component space.



Regions are separated according to the distribution of the spectral classification of the WGS sources (the *training set*) coaxial with the PC1.

The radii of the three regions (BL Lacs, FSRQ-dominated and mixed) **are determined based on the radial distribution of the WGS sources (the *training set*).**

Discrete protoscore

$$ps_{\mathrm{disc}} = 1/n_{\mathrm{extr}}$$

where n_{extr} is the number of *extremal* points inside the region (for each region of the *locus*).

Normalized continuos protoscore

$$ps_{\mathrm{cont}} = \frac{1}{6^n \cdot ps^n_{\mathrm{disc}}}$$

where n is an index used to tweak the efficiency and completeness of the association process.

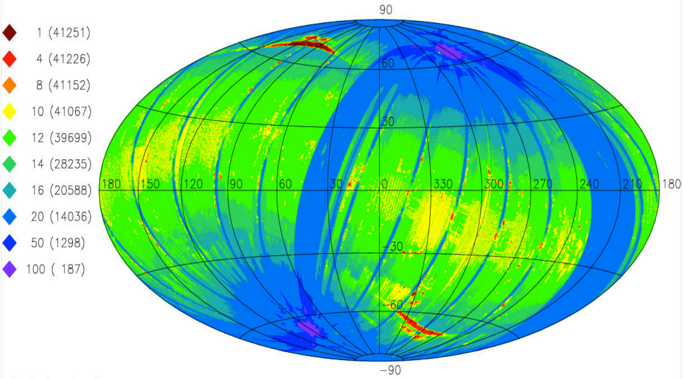
Final score

$$s = ps_{\mathrm{cont}} \cdot w_V$$

where $w_V = ||V_{\mathrm{err.ellips.}} - V_{\mathrm{reg}}||/V_{\mathrm{reg}}$ weights according to the volume of the error ellipsoid of the source.

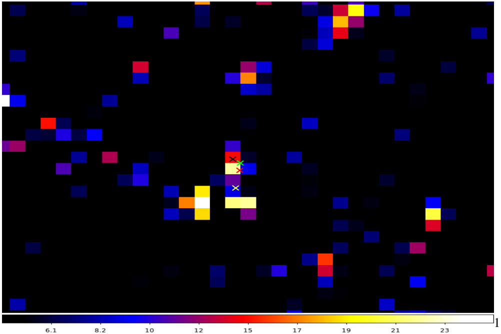
Spatially unconstrained search

The new model can be also used to perform **"not spatially constrained" extraction of *WISE* candidate blazars from the *WISE* photometric catalog**.



Can we find new blazars?

γ-ray sources in the 2FGL catalog have ($TS \geq 25$). Due to their extreme variability, many blazars might not have made into the catalog (where data were integrated over 2 years timespan).



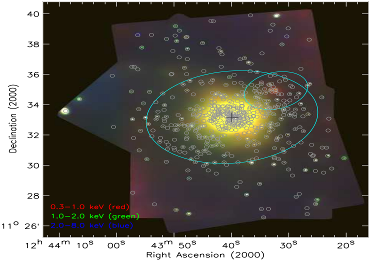
The extraction of *WISE* candidate blazars can be used to **select *Fermi* blazars below the *TS* threshold** and to search for **blazars in the positions where γ-ray transients were observed.**

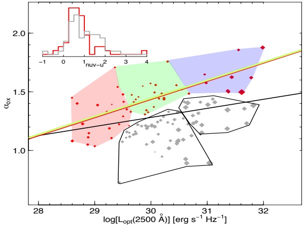
[1](Courtesy of G. Migliori)

Other projects

1. Characterization of the globular clusters-LMXBs connection in the **optical/X-rays/spatial *feature* space** (NGC4649 and other galaxies);

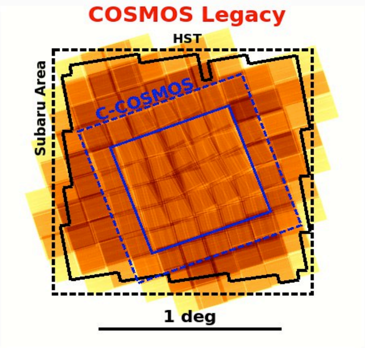
2. **Application to a sample of X-ray selected AGNs with wide-band multi-λ photometry**, with already known correlations found by CLaSPS.





CLaSPS and COSMOS Legacy Survey

2.8 Ms exposure time on Chandra were just awarded (P.I. F. Civano) to observe 2 deg^2 containing the original *Chandra*-COSMOS field. Expected to detect 4500 X-ray sources to $F_{\rm lim} \sim 2{\cdot}10^{-16}$ cgs in $[0.5, 2]$ keV energy band.



- Unparalleled multi-wavelength coverage: 47 wide and narrow bands from X-ray to radio.
- Suited to characterize the SEDs of AGNs and constrain the dependence of SMBHs on their environment as function of the host galaxies properties.
- **A rich, complex and large dataset**!

CLaSPS development

Handling upper-limits and NaN's (regardless of their origins) becomes crucial with observationally rich complex samples.

- Observations or upper-limits in a band can be translated into a binary *labels* and used to characterize the clustering in the *feature* space...
- ...but still, discarding sources of the sample with not-measured *features* can drastically reduce the size and richness of the dataset and, potentially, throw away valuable information.
- Comparison with the results on similar datasets *features*-wise to check robustness, assess variance, validate outliers, etc.

Exploring the application of **Feature-Distributed Clustering (FDC)** and **Object-Distributed Clustering (ODC)** methods, borrowed from *consensus clustering*.

Summary

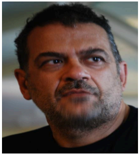
The discovery of the *WISE* blazars *locus* with CLaSPS and its application as a tool for the classification and extraction of candidate blazars is an example of astronomical *KD* workflow involving unsupervised and supervised methods.

- **Archival data** can be re-used and interpreted from a fresh point of view;
- **Variability**: how does variability fit into this scenario? Do mid-infrared and γ-ray/X-ray variability affect the blazars *locus*? If so, how and why?

- A few examples of *KD* workflows giving interest results raise awareness of these new "integrated methods in the astronomical community;
- Like already happened in many other fields, *KD* will become (is becoming) the only chance to make sense out of the overwhelming amount of data from observations.

Acknowledgements






G. Fabbiano (*CfA*)
O. Laurino (*CfA*)
G. Longo (*Univ. of Naples*)
F. Massaro (*SLAC*)

- UC & Classification/Regression → [D'Abrusco, R. et al. 2009, MNRAS, 396, 223], [Laurino, O., D'Abrusco, R. et al. 2011, MNRAS, 418, 4]
- *CLaSPS* → [D'Abrusco, R. et al. 2012, ApJ, 755, 2, 92]
- *WISE* Blazars → [D'Abrusco, R. et al. 2012, ApJ, 748, 68D], [Massaro, F., D'Abrusco, R. et al. 2012, ApJ, 752, 61M]