Knowledge Discovery Workflows in the Exploration of Complex Astronomical Datasets

Raffaele D'Abrusco

Harvard-Smithsonian Center for Astrophysics

Galilean experimental method

Hunts Needle in a Haystack

Setting the stage

Knowledge Discovery - *KD* - is the "automatic processing of large amount of data to extract patterns that can represent knowledge about the data".

HOW LONG does it take to find a needle in a haystack? Jim Moran, Washington, D. C., publicity man, recently dropped a needle into a convenient pile of hay, hopped in after it, and began an intensive search for (a) some publicity and (b) the needle. Having found the former, Moran abandoned the needle hunt.

Hunts Needle in a Haystack

KD in the real world

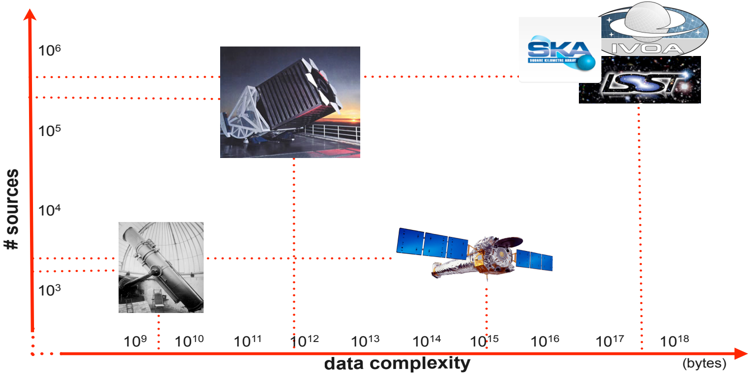
Outside our Real and Virtual Domes, *KD* methodology has already shaped how Data are processed and Knowledge is extracted, in several (expected and unexpected) fields:

- Social sciences: advertisement placement, social networks...
- Finance: market analysis tool, derivatives trading...
- Life science: genetics, epidemiology, drug testing....
- Security: face recognition, behavior tracking...
- **Google and the like**...

And for most of these fields, KD is the only possibility to make sense out of the overwhelming amount of data gathered.

The opportunity in Astronomy

The advancement of astronomical technology (**hardware and software**) allows to go larger, deeper and with higher resolution, both spatially and spectrally, changing the nature of astronomical data.



Facilities like LSST, SKA, ALMA, *Euclid*, etc... and the access and federation to archival data provided by the VO's will boost this change by **making large multivariate datasets** (spanning also the time axis) **easily available**.

Not just a needle in the haystack

A *KD workflow* is a sequence of analysis steps accomplished through *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...

While high-dimensional regions of the observable parameters space are still completely unexplored, **not all low-dimensionality *feature* spaces have been investigated yet**, as in principle we look into places where they expect to find something. **A systematic way to search for "something" is necessary as it does not depend on our biases/prioritization/limited availability of time and resources**.

Not just a needle in the haystack

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

Goals:

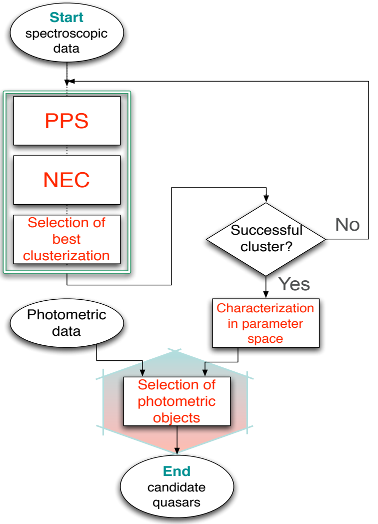
- **Discovery**
  - Find new complex correlations;
  - Expand known correlations to more dimensions;
  - Find new simple correlations, so far overlooked;
- **Using the discovery**
  - Insight into astrophysics;
    - Classification, regression, new ways to look at things...

While high-dimensional regions of the observable parameters space are still completely unexplored, **not all low-dimensionality *feature* spaces have been investigated yet**, as in principle we look into places where they expect to find something. **A systematic way to search for "something" is necessary as it does not depend on our biases/prioritization/limited availability of time and resources**.

A first try

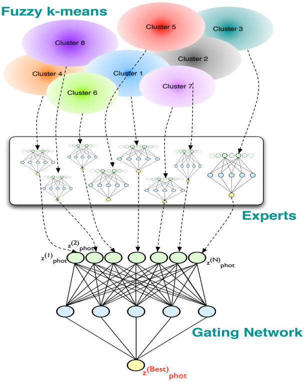
Extraction of optical candidate quasars from the SDSS photometric dataset using spectroscopic base of knowledge.

A combination of two unsupervised clustering (UC) techniques and the use of *a priori* knowledge available for a subset of confirmed SDSS quasars was used to **extract optical candidate quasars from photometric data**.



The Weak Gated Expert

The Weak Gated Expert (*WGE*) is a KD procedure for the determination of $z_{\rm phot}$ for galaxies and quasars, based on clustering in the color space and the training of an *ensemble* of neural networks for regression.



- The UC algorithm split the *feature space* into more homogeneous chunks to prevent under or over-fitting of the *experts*;
- Multiple distinct *experts* (neural networks) are trained on different regions of the *features* space;
- The *gate* combines the outputs of the single *experts* in order to maximize the accuracy of the reconstruction and minimize biases.

A more general question

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

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

The *tenet*

Spontaneous aggregations of sources in their observable space, the clusters, reflect similarities common traits shared by these sources. **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)

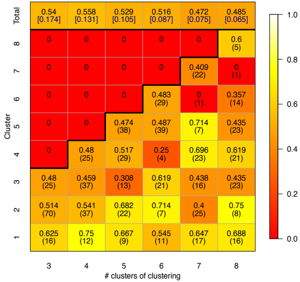
1. A UC algorithm is used to produce clusterings in the *parameter* space generated by any subset of the observables (the ***features***);
2. Other observables not employed for the clustering (the ***labels***), are used as *tags* to identify interesting set of clusters using the ***score***;
3. The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

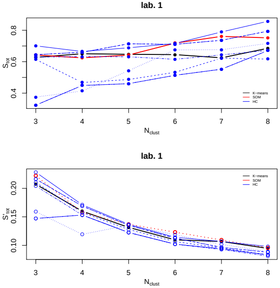
The choice of the clustering(s)

Set of clusters (or single clusters) are picked according to the degree of correlation between the distribution of cluster members in the *feature* space and their distribution in the *labels* space.

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

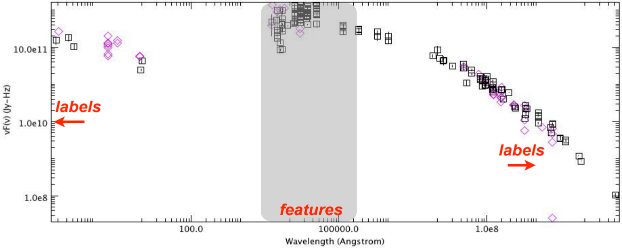
where f_{ij} is the fraction of members of the i-th cluster with values of the *label* in the j-th class.





An interesting finding

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

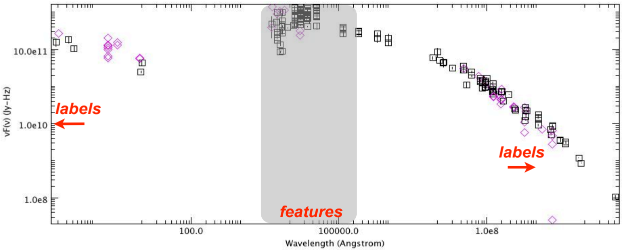


Dataset	→	AGNs catalog
Features	→	UV(*Galex*) + Optical(*SDSS*)+ NIR(*UKIDSS*) + IR(*WISE*)
Labels	→	AGNs class., Blazars spectral class. γ-ray emission

Three clusters composed of Blazars stood out with large values of the *scores* spectral classification as *label*. Further experiments using as *labels* the γ-ray detection and FSRQs-BL Lacs classifications **showed that such patterns of Blazars depend on *WISE* mid-Infrared colors.**

An interesting finding

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

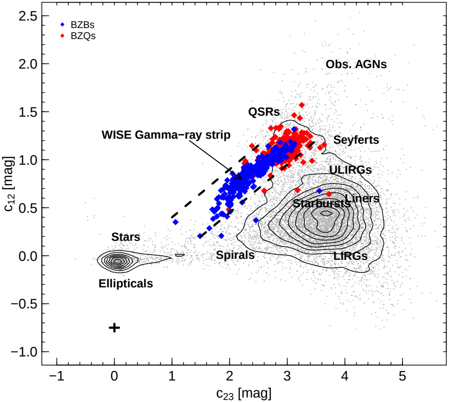


Dataset	$\rightarrow$	AGNs catalog
Features	$\rightarrow$	UV(*Galex*) + Optical(*SDSS*)+ NIR(*UKIDSS*) + IR(*WISE*)
Labels	$\rightarrow$	AGNs class., Blazars spectral class. γ-ray emission

Three clusters composed of Blazars stood out with large values of the *scores* spectral classification as *label*. Further experiments using as *labels* the γ-ray detection and FSRQs-BL Lacs classifications **showed that such patterns of Blazars depend on *WISE* mid-Infrared colors.**

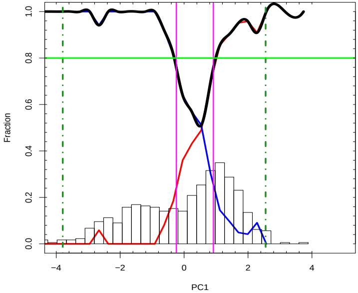
The *WISE* Blazars strip

This pattern in the IR *WISE* color space of Blazars would have been apparent even in this low dimensional projection of the multi-λ *feature* space that we studied with *CLaSPS*, but it had been overlooked so far.



Another step in the workflow

The *WISE* Blazars *locus* can be used as a **supervised classifier**.



The *WISE* Blazars *locus* is modeled in the Principal Component space generated by *WISE* colors space as three distinct subregions dominated by different spectral subclasses of sources (BL Lacs, FSRQ-dominated and mixed).

Discrete protoscore

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

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

Normalized continuos protoscore

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

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

Final score

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

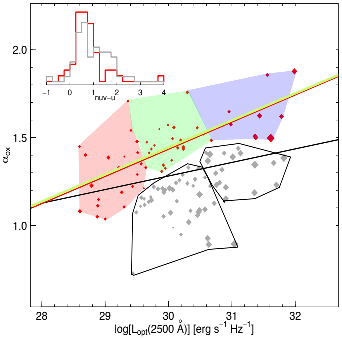
where $w_V = ||V_{\text{err.ellips.}} - V_{\text{reg}}||/V_{\text{reg}}$ weights according to the volume of the error ellipsoid of the source.

Some more applications

- Mixing mid-IR and high-energy variability;
- Classification for Unassociated *Fermi* sources;
- Extraction of new *WISE* candidate blazars with validations using archival multi-wavelength data

More science with **CLaSPS**:

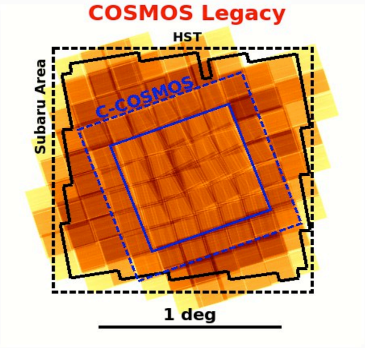
1. The characterization of the globular clusters-LMXBs connection in different galaxies;
2. **Application to a sample of X-ray selected AGNs with wide-band multi-λ photometry**, with already known correlations found by CLaSPS.



CLaSPS and Legacy COSMOS

Here comes the Super *Chandra*-COSMOS!

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.



- COSMOS multi-wavelength coverage is unparalleled: 47 wide and narrow bands spanning the whole spectrum.
- Perfect to characterize the SEDs of AGNs and constrain the dependence of SMBHs on their environment, as a function of the host galaxies properties.
- **A treasure for astronomical data miners!**

Improvements

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

1. Observations or upper-limits in a band can be translated into a binary *labels* and used to characterize the clustering in the *feature* space...
2. ...but still, discarding sources of the sample with not-measured *features* can drastically reduce the size and richness of the dataset.
3. Significant comparison with results on similar datasets *features*-wise to check robustness, assess variance, etc.

Feature-Distributed Clustering (FDC) methods can be used to address points 1 and 2, while simulations and **Object-Distributed Clustering (ODC)** techniques are useful for point 3.

Improvements

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

1. Observations or upper-limits in a band can be translated into a binary *labels* and used to characterize the clustering in the *feature* space...
2. ...but still, discarding sources of the sample with not-measured *features* can drastically reduce the size and richness of the dataset.
3. Significant comparison with results on similar datasets *features*-wise to check robustness, assess variance, etc.

Feature-Distributed Clustering (FDC) methods can be used to address points 1 and 2, while simulations and **Object-Distributed Clustering (ODC)** techniques are useful for point 3.

Stuff that helps

- The core *CLaSPS* functionalities (*KD* algorithms, statistics and visualization) originally implemented in R
- The *connective tissue* of the workflow (retrieval of archival data, pre-processing, post-processing) is Python
- Specific data-related tasks are carried out by the *passepartout* for the *realm of tables*: STILTS.
- All experiments run on my laptop or desktop in my office (OK for small datasets).

Handy stuff that would help

What's missing?

- A **high-level description of *KD* workflows** in astronomy (to compare and improve methods with different applications/use cases/domain);
- A **repository** for code, workflows and template datasets;
- A scalable platform for *KD* workflows to tackle massive and complex datasets! (My computers won't cope with data anymore very soon...);
- Widespread adoption of versatile data access protocols (**TAP** interface, *casJobs*-like access points, etc.) from data centers
- Astronomers should learn **SQL, SQL, SQL**, machine learning, statistics,...

Handy stuff that would help

What's missing?

- A **high-level description of *KD* workflows** in astronomy (to compare and improve methods with different applications/use cases/domain);
- A **repository** for code, workflows and template datasets;
- A scalable platform for *KD* workflows to tackle massive and complex datasets! (My computers won't cope with data anymore very soon...);
- Widespread adoption of versatile data access protocols (**TAP** interface, *casJobs*-like access points, etc.) from data centers
- Astronomers should learn **SQL, SQL, SQL**, machine learning, statistics,...

The future

The future of astronomy will give me (us?) something to cheer about:

Astronomy is becoming a data-intensive discipline

- Exciting science ahead for the *brave and lucky* ones
- *KD* experts acquire *transferable skills* and expertise valued outside the academia
- (Average) astronomers' awareness of KD usefulness (somewhat) growing
- *KD* know-how starting to percolate into the astronomical community

Interesting scientific results will boost *KD* adoption!

The future

The future of astronomy will give me (us?) something to cheer about:

Astronomy is becoming a data-intensive discipline

- Exciting science ahead for the *brave and lucky* ones
- *KD* experts acquire *transferable skills* and expertise valued outside the academia
- (Average) astronomers' awareness of KD usefulness (somewhat) growing
- *KD* know-how starting to percolate into the astronomical community

Interesting scientific results will boost *KD* adoption!

The future

The future of astronomy will give me (us?) something to cheer about:

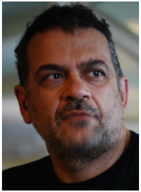
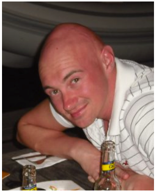
Astronomy is becoming a data-intensive discipline

- Exciting science ahead for the *brave and lucky* ones
- *KD* experts acquire *transferable skills* and expertise valued outside the academia
- (Average) astronomers' awareness of KD usefulness (somewhat) growing
- *KD* know-how starting to percolate into the astronomical community

Interesting scientific results will boost *KD* adoption!

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]

Thank you!