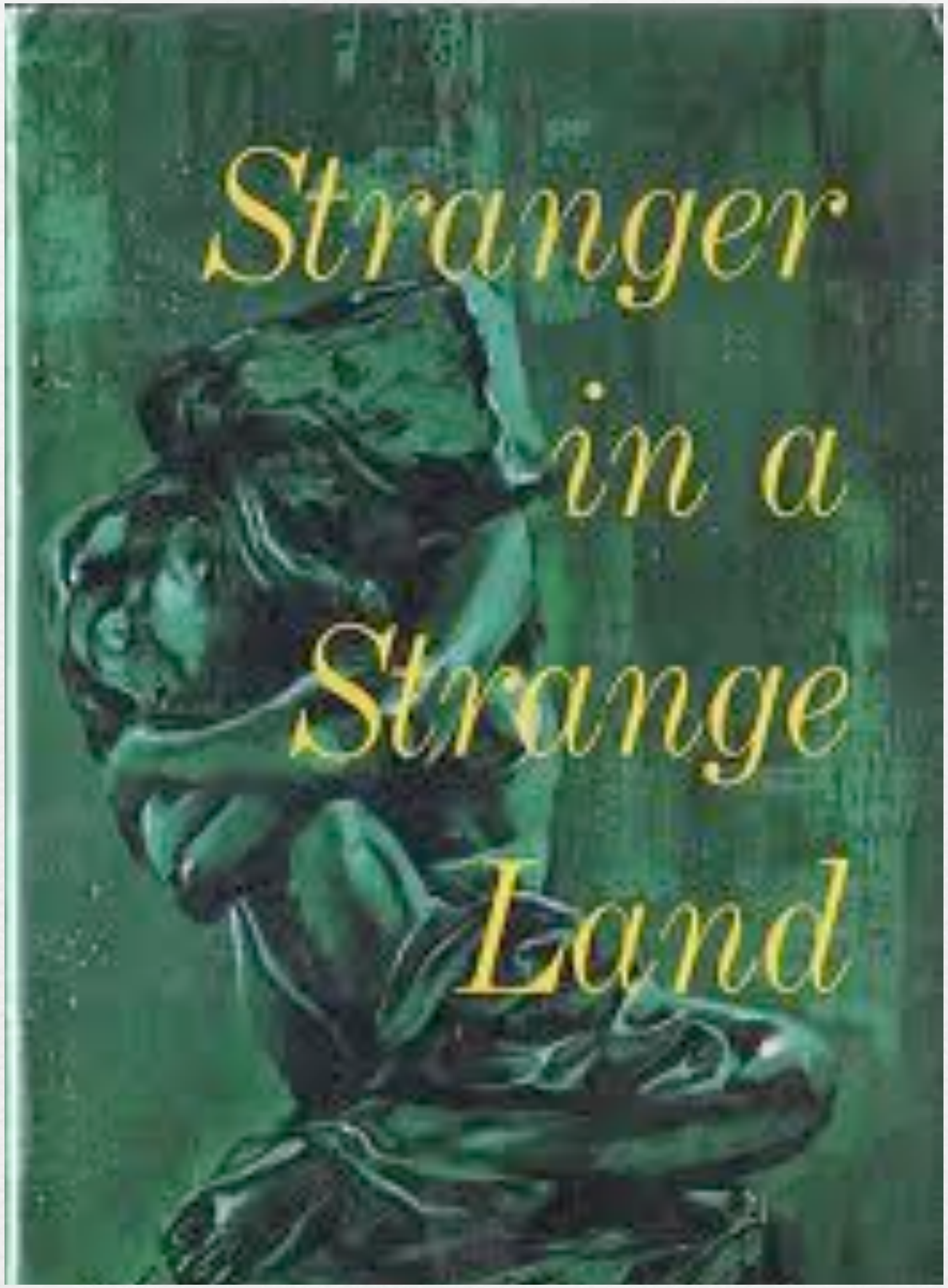
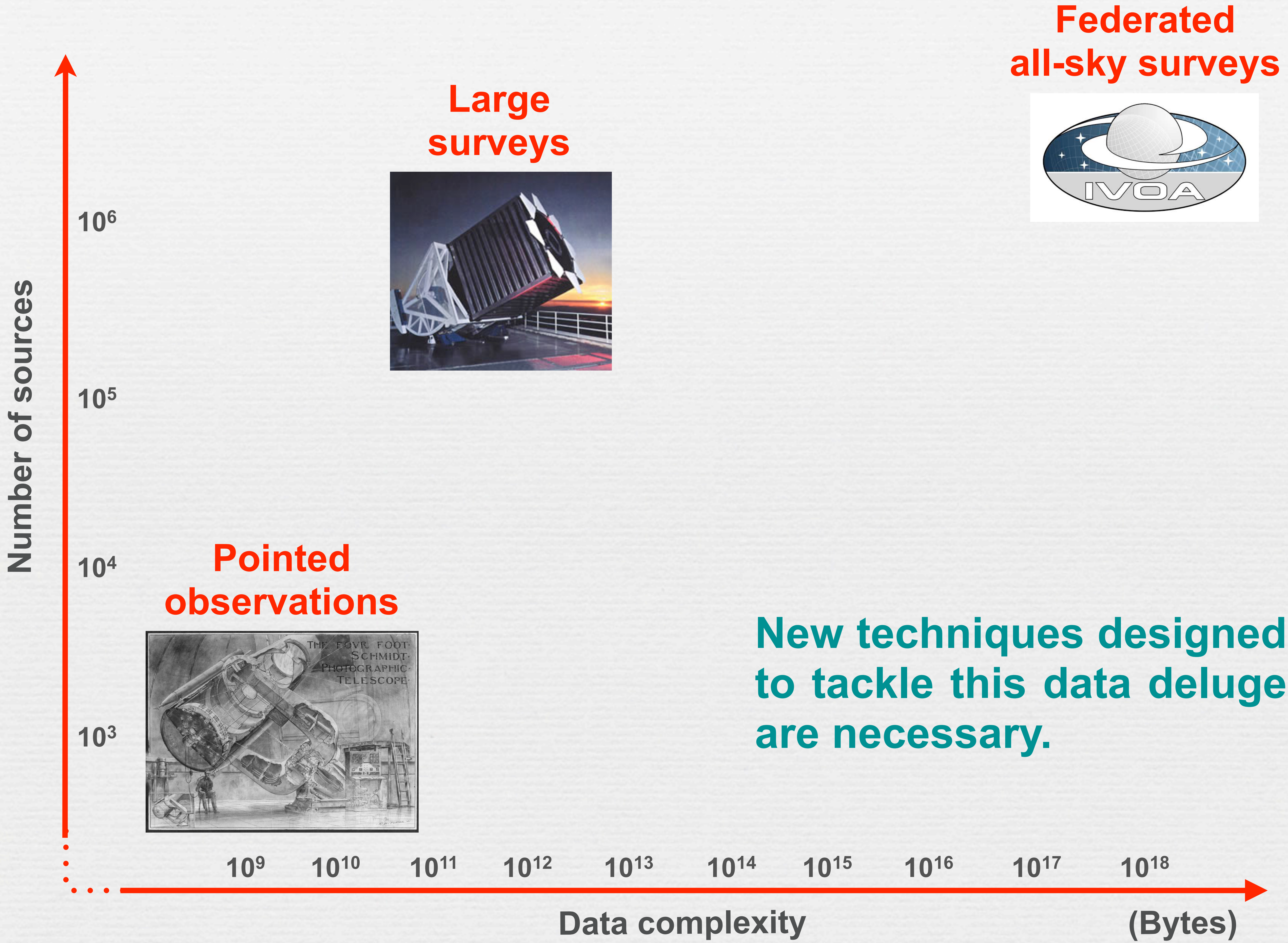


Between massive astronomical datasets and the Virtual Observatory

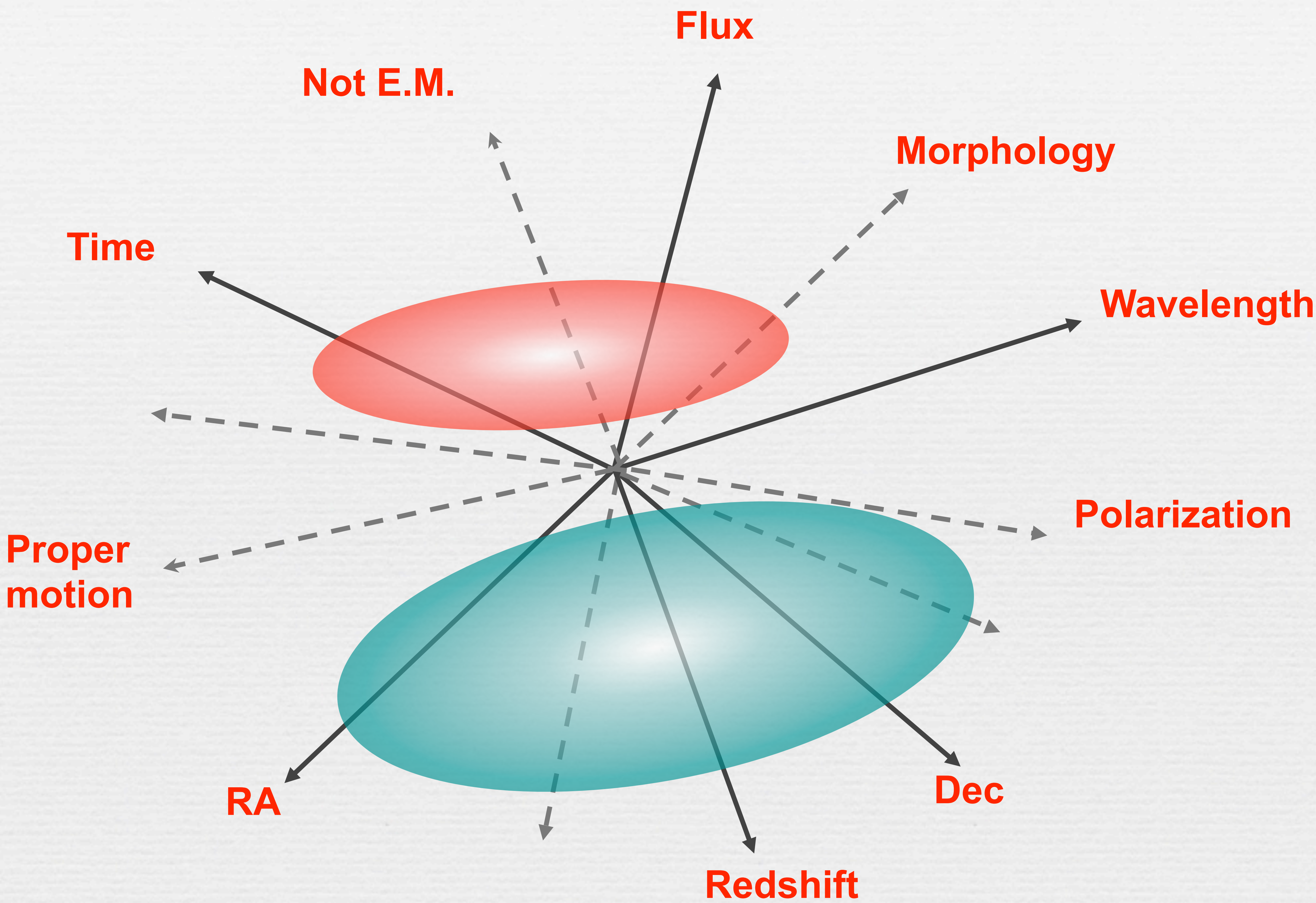
R. D'Abrusco

with O. Laurino, S. Cavuoti, G. Longo and the DAME gang

A paradigm shift



A growing parameter space



Most discoveries were made in small regions of subspaces or along some of these axes

Data Mining and Astronomy

'Data Mining (DM) is the process of extracting patterns from data.'

A science case for DM?

Machine learning can ease our access to the realm of 'candidates' (or probabilistic) astronomy. Many problems (cosmology, large scale structure, classification of sources) can be addressed with efficient selection methods and accurate measurement of statistical observables.

What Data Mining can do

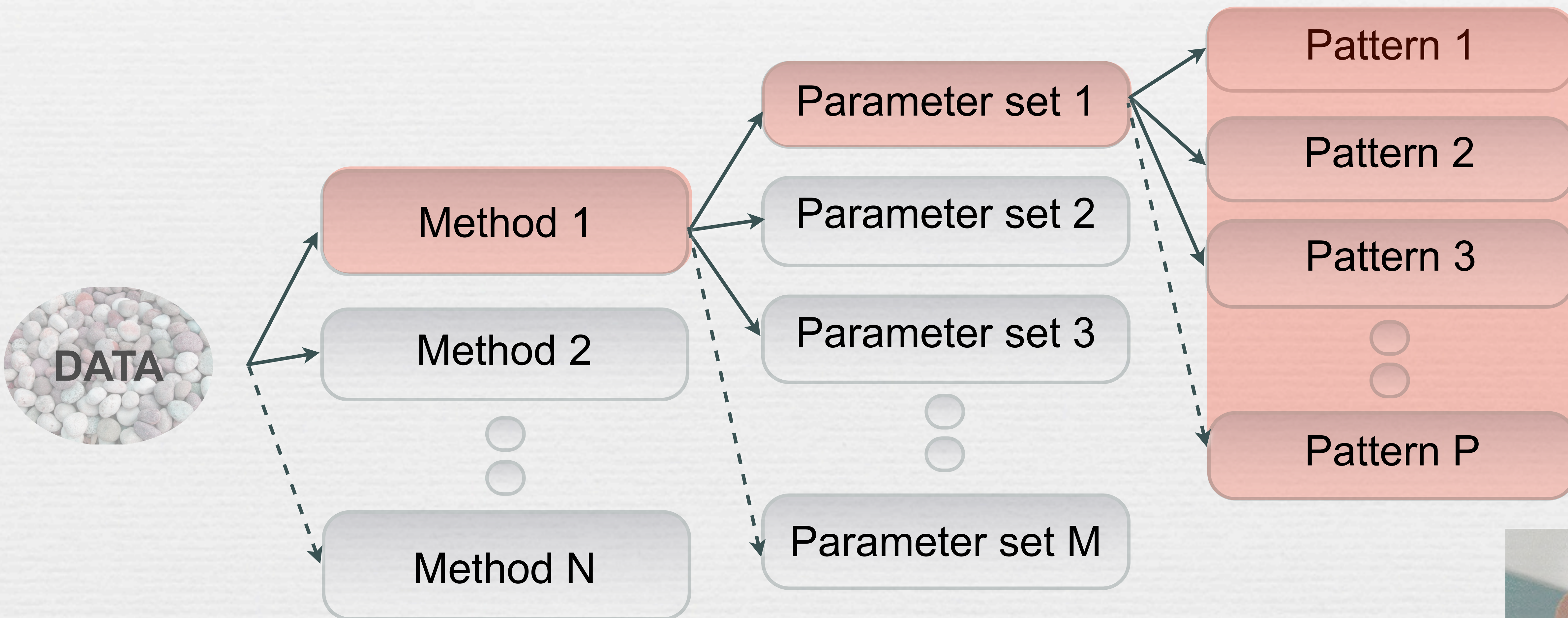
Data Mining (DM) and Machine Learning (ML) techniques can be used to perform multiple operations, common in astronomical research:

- Data exploration → Clustering, dimensionality reduction...
- Classification → Neural networks, k-means, Self Organizing Map, SVMs,...
- Regression → Neural Networks, Support Vector Machines...
- Data visualization → Dimensionality reduction, Principal Components, Principal Surfaces...

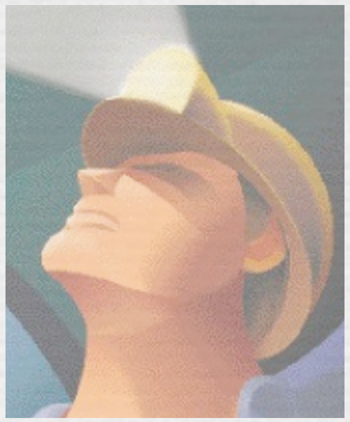
What Data Mining can't do

Provide a general recipe for all problems...

Criteria for the choice of the approach are the nature of the **specific astronomical problem**, the intricacies of the PS distribution, **computational performances,** implementation and **generalizability**.

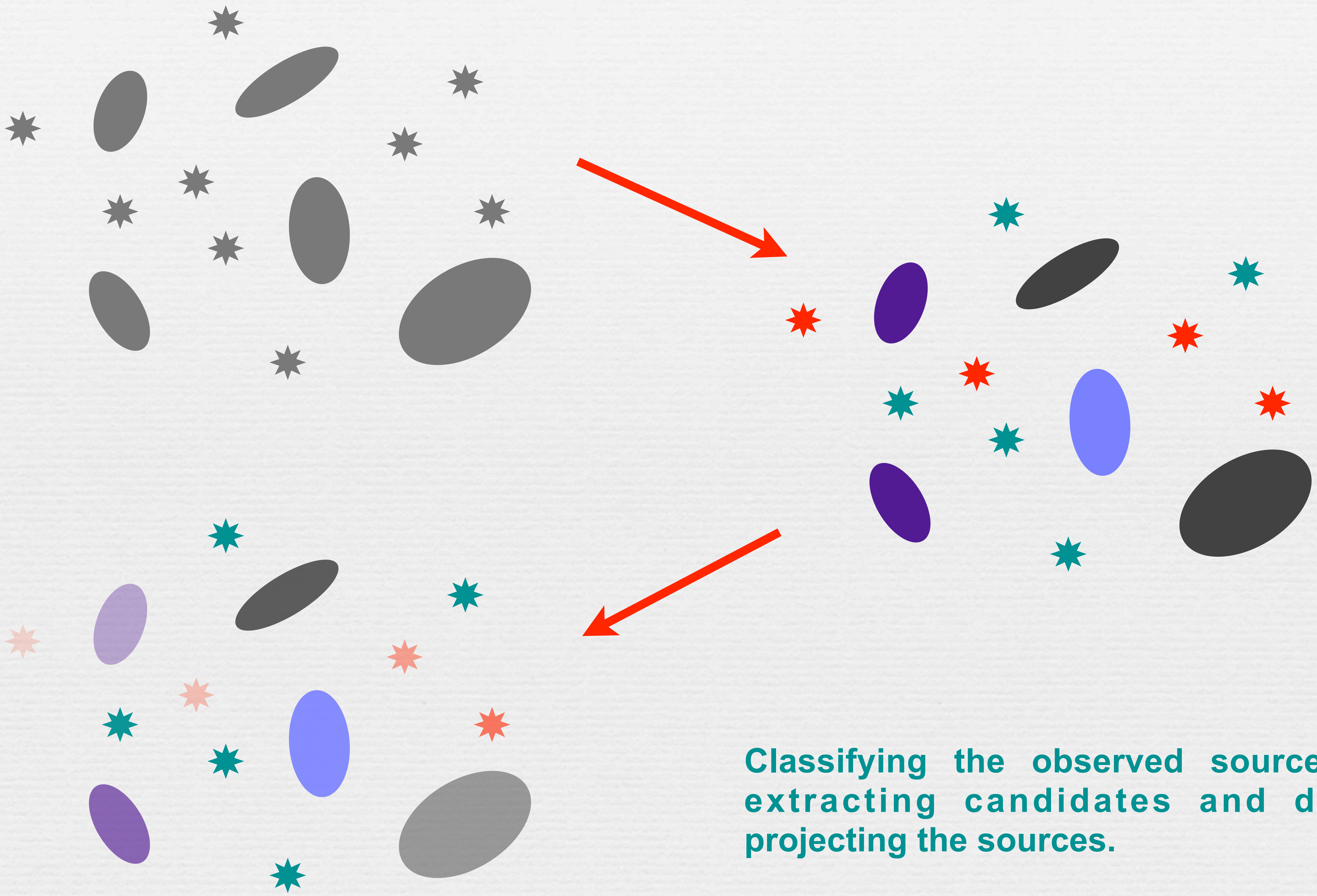


...and understand the results!



Human.
Still needed

Back to good old sky mapping



Classifying the observed sources, extracting candidates and de-projecting the sources.

The Basics

Raw materials:
'Data'

A belief system:
'Base of Knowledge'

Method:
'Statistical techniques'

Tools:
'Information Technology'

Candidate quasars extraction

Raw materials:
‘Data’

Dataset of photometric stellar sources

A belief system:
‘Base of Knowledge’

‘Optical spectroscopy is able to select quasars’

Method:
‘Statistical techniques’

Clustering algorithms

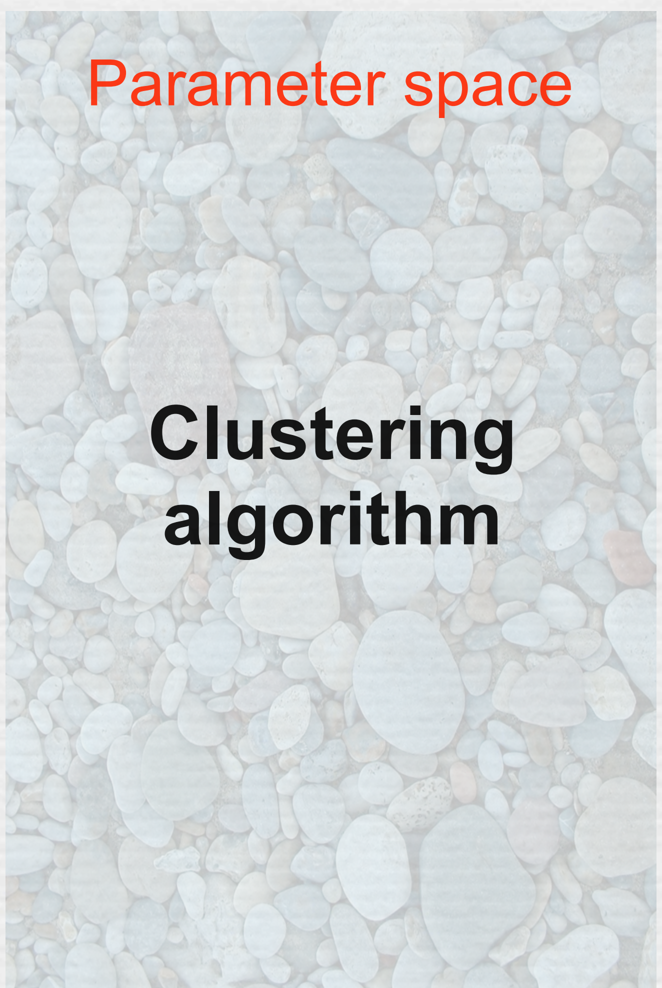
Tools:
‘Information Technology’

Virtual Observatory distributed computation

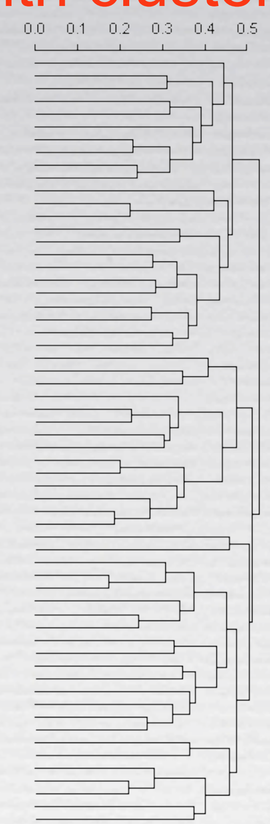
Quasars in the parameter space

Unsupervised clustering inside the colors space using spectroscopic classifications, available for the members of the BoK, as label.

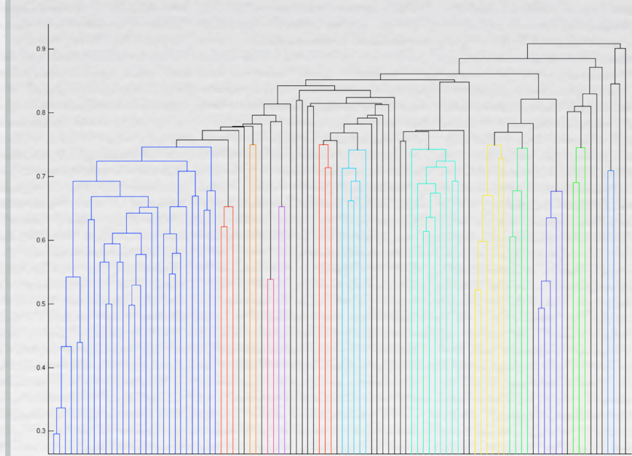
The statistical characterization of BoK clusters in the PS is exploited to select new candidates extracted from photometric samples (i.e. for which spectroscopy is unavailable).



Parameter space with clusters

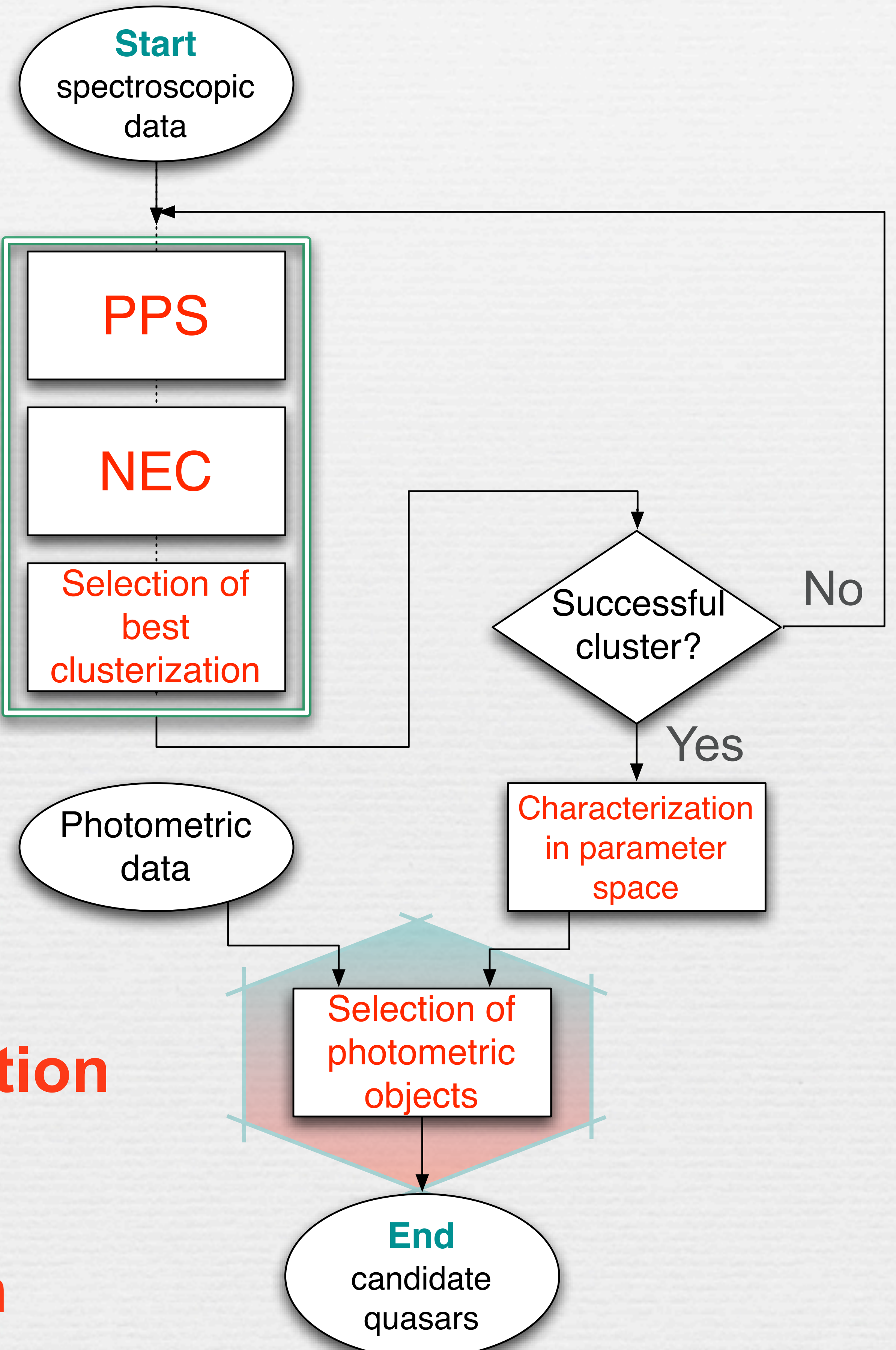


Parameter space with “labelled” clusters



Candidate quasars: the method

PS clustering



'Double clustering'

Candidates extraction

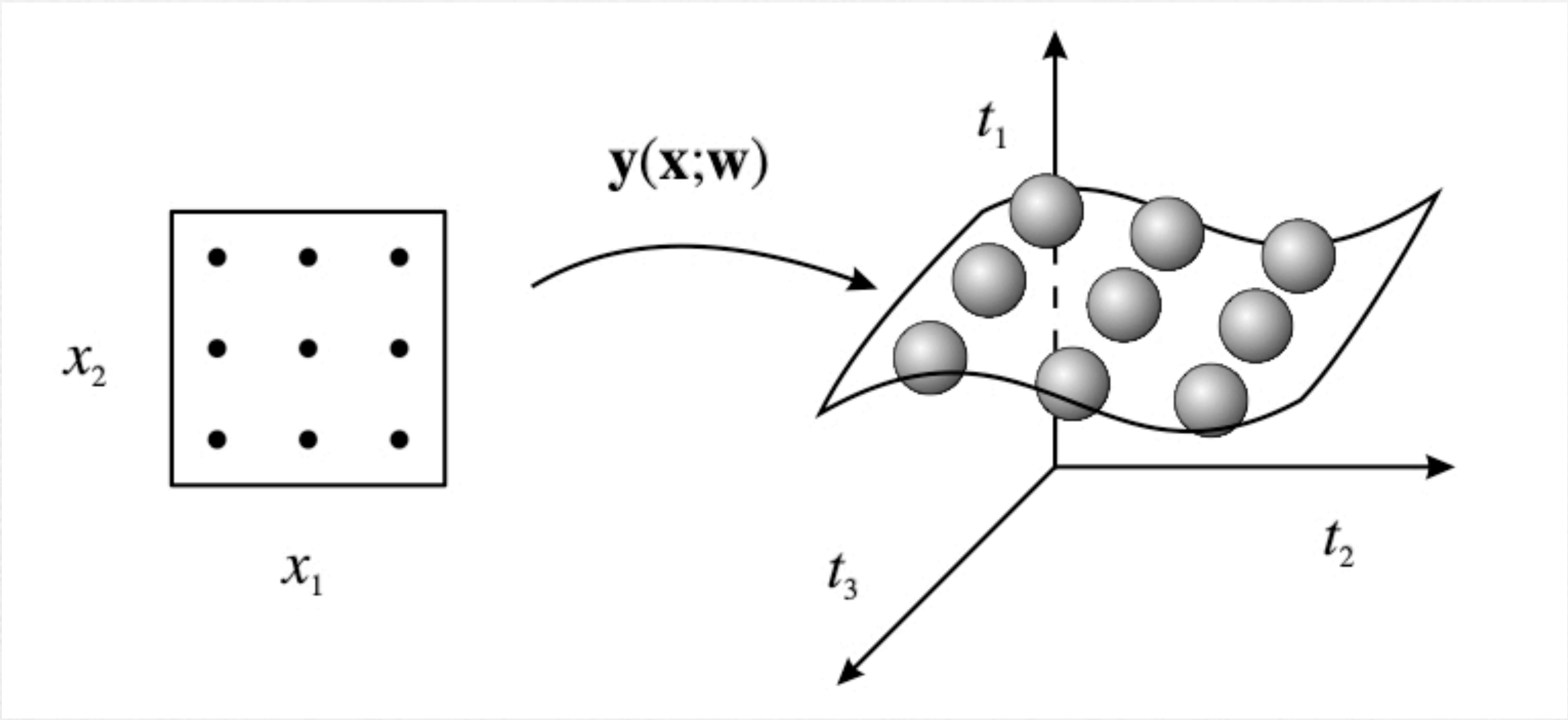
'Closest cluster'

z_{phot} estimation

'Weak Gated Expert'

Probabilistic Principal Surfaces

Generative Topographic Map + oriented covariance



PPS are a non linear extension of PCA which determine a parametric mapping from a Q-dimensional space a to D-dimensional space (Q << D), invert it and use it to connect points in the "real space" to points in the "latent space".

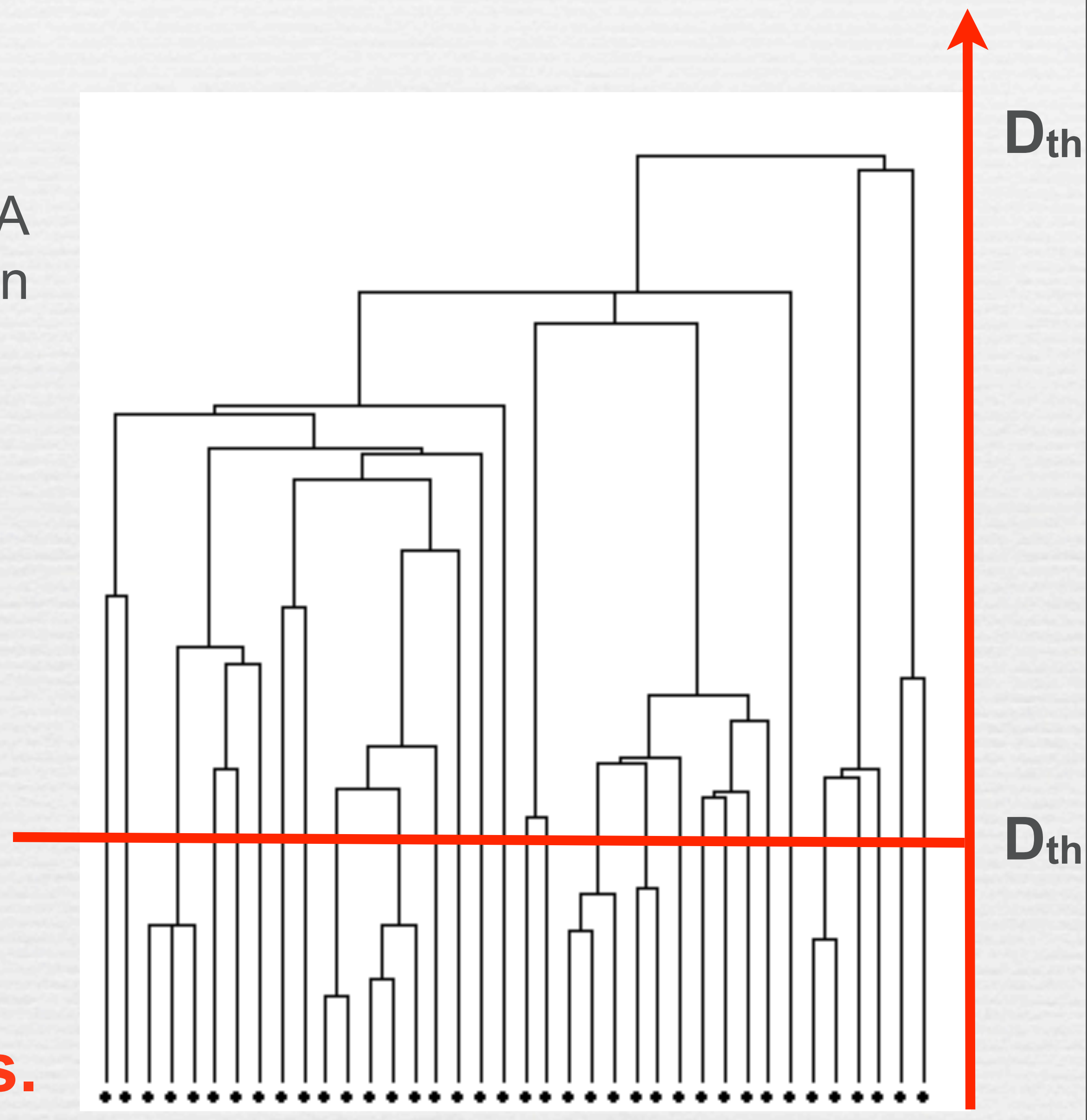
Close points in the original space are close in the latent space, where clustering is enhanced.

Negative Entropy Clustering

NEC is an agglomerative clustering based on "negative entropy", which express the 'non-gaussianity' of a multivariate distribution.

Given the clusters A and B, they are replaced by C = A ∪ B if and only if C resembles more strictly a gaussian than A and B respectively and the relation holds:

$$NegE(A \cup B) < D_{th}$$



Changes in metric reflect changes in topology of the PS distribution of sources.

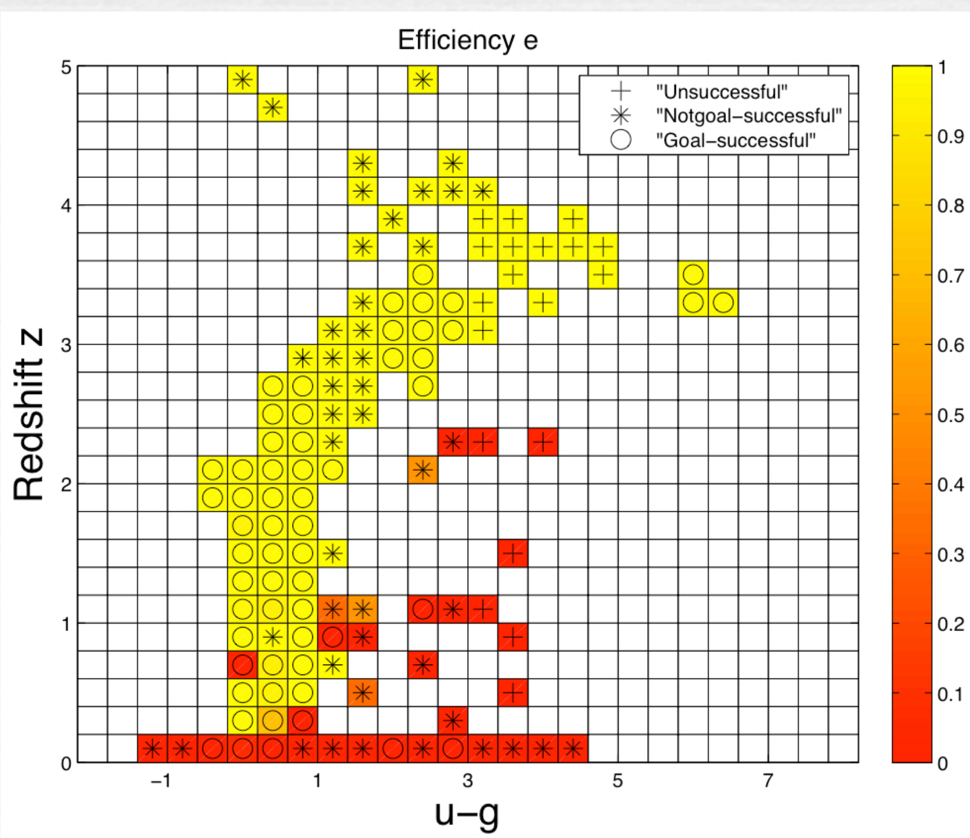
Candidate quasars: the results

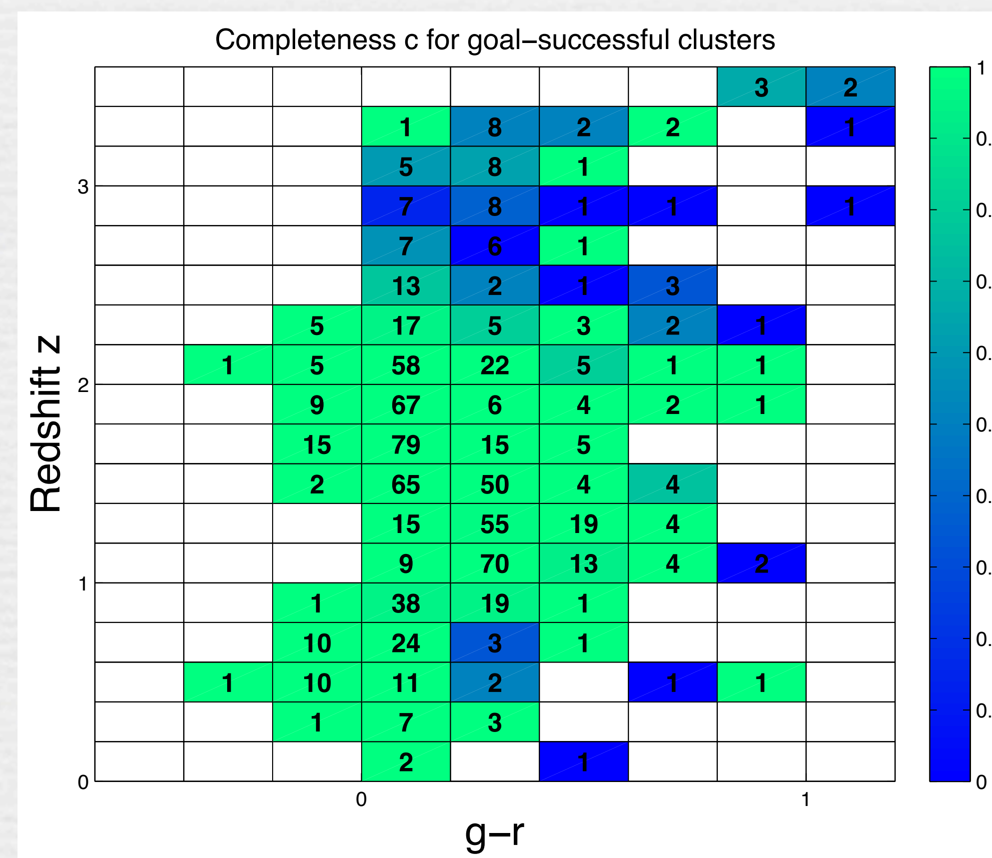
Global performance:

$e_{tot} = 85\%$
$c_{tot} = 91\%$
4 optical colors

$e_{tot} = 92\%$
$c_{tot} = 93\%$
7 optical + NIR colors

A map of how quasars are distributed into the SDSS optical color parameter space, with global and local information.





Largest catalog to date of optical candidate quasars
(D'Abrusco et al. 2009)

Optical AGNs

Raw materials:
'Data'

Sample of spectra of SDSS galaxies

Sample of photometry of SDSS galaxies

A belief system:
'Base of Knowledge'

'Line ratios can classify AG'

'Multi-λ photometry traces EL galaxies'

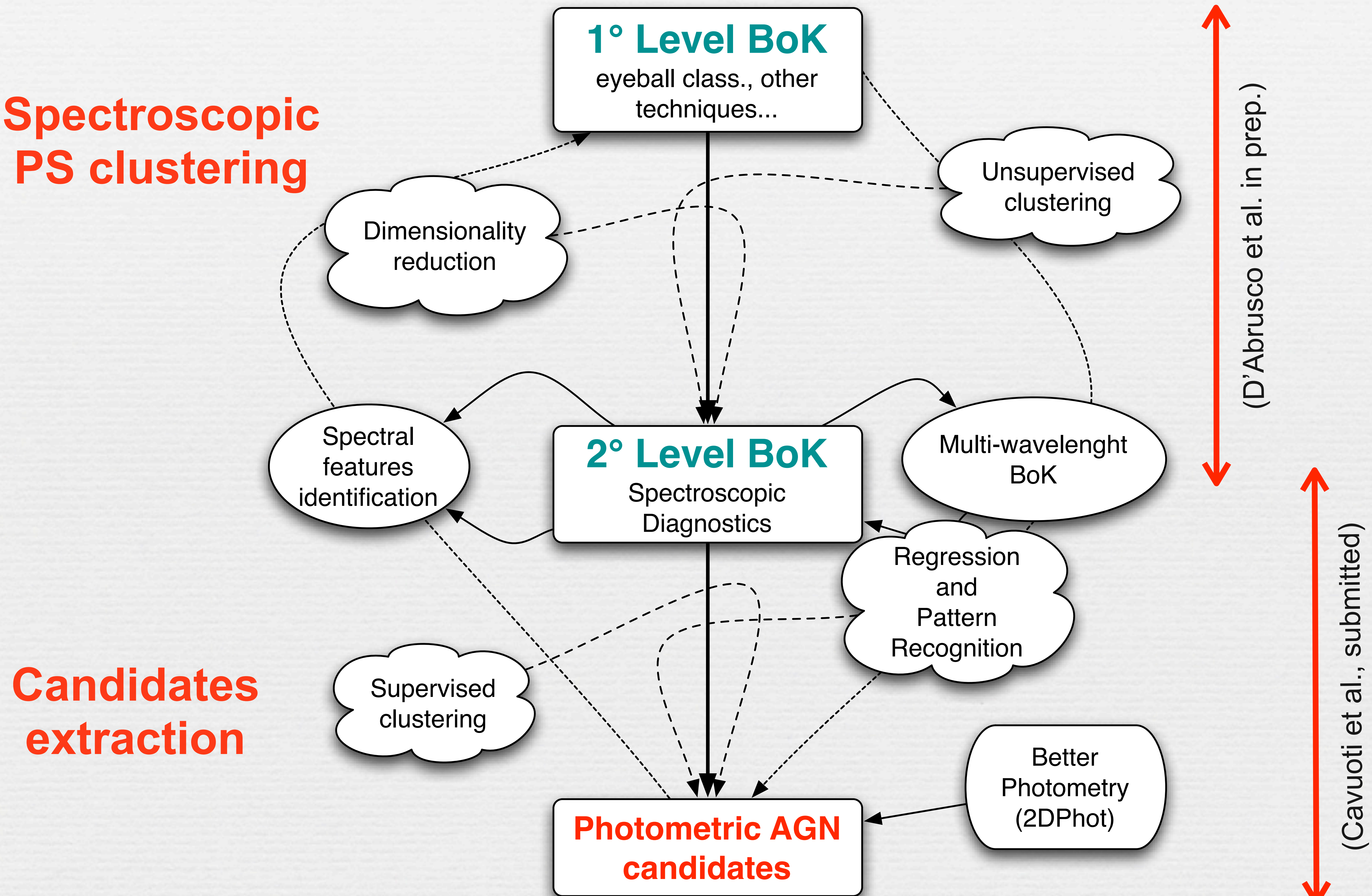
Method:
'Statistical techniques'

Support Vector Machines - NN

Tools:
'Information Technology'

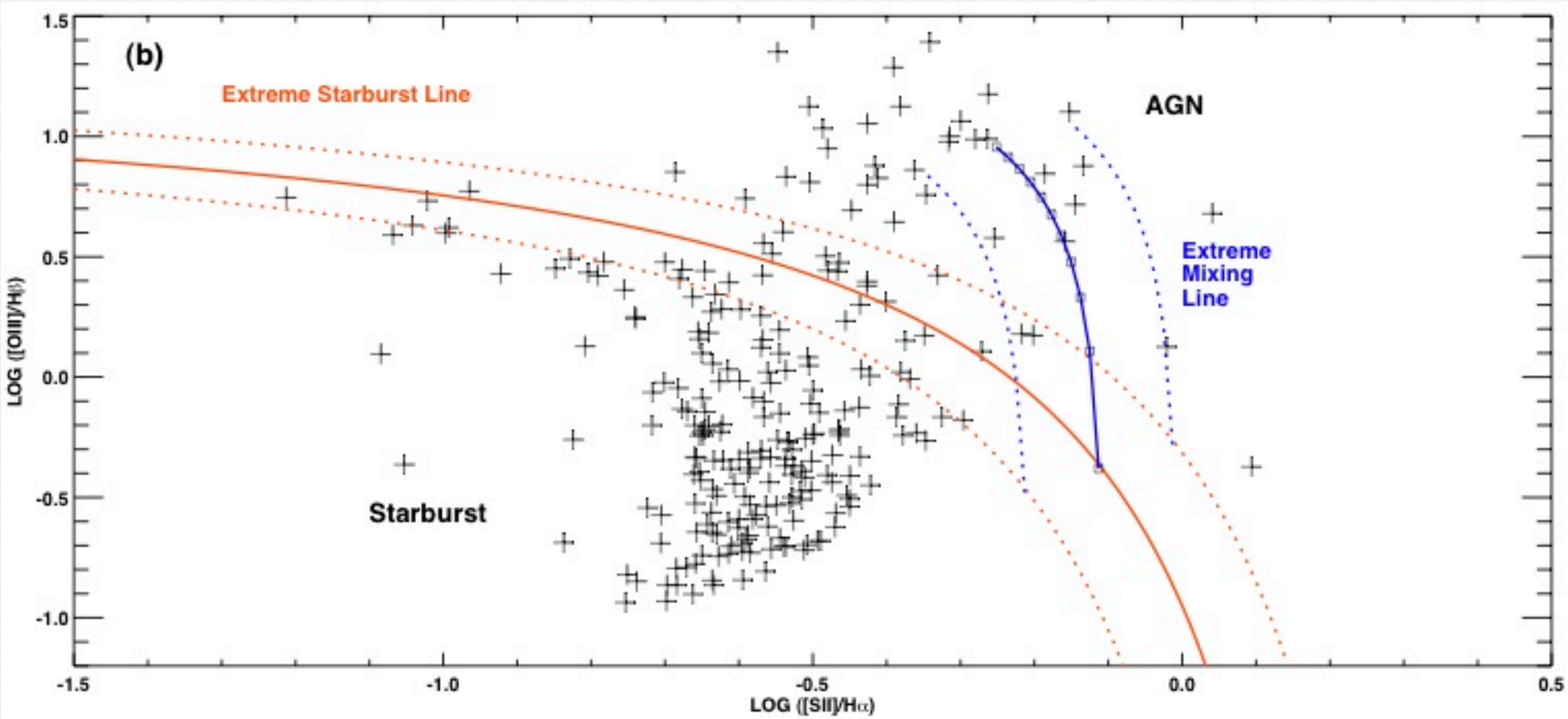
High perfomance computing

Optical Spectroscopic AGNs



Spectroscopic indicators

Spectroscopic diagnostics used to distinguish starburst galaxies from AGNs and classify AGNs in classes (Sey1, Sey2), based on line intensity ratios (BPT plots).



Kewley's line

$$\log\frac{[OIII]\lambda5007}{H_\beta} = \frac{0.61}{\log\frac{[NII]\lambda6583}{H_\alpha} - 0.47} + 1.19$$

Kauffman's line

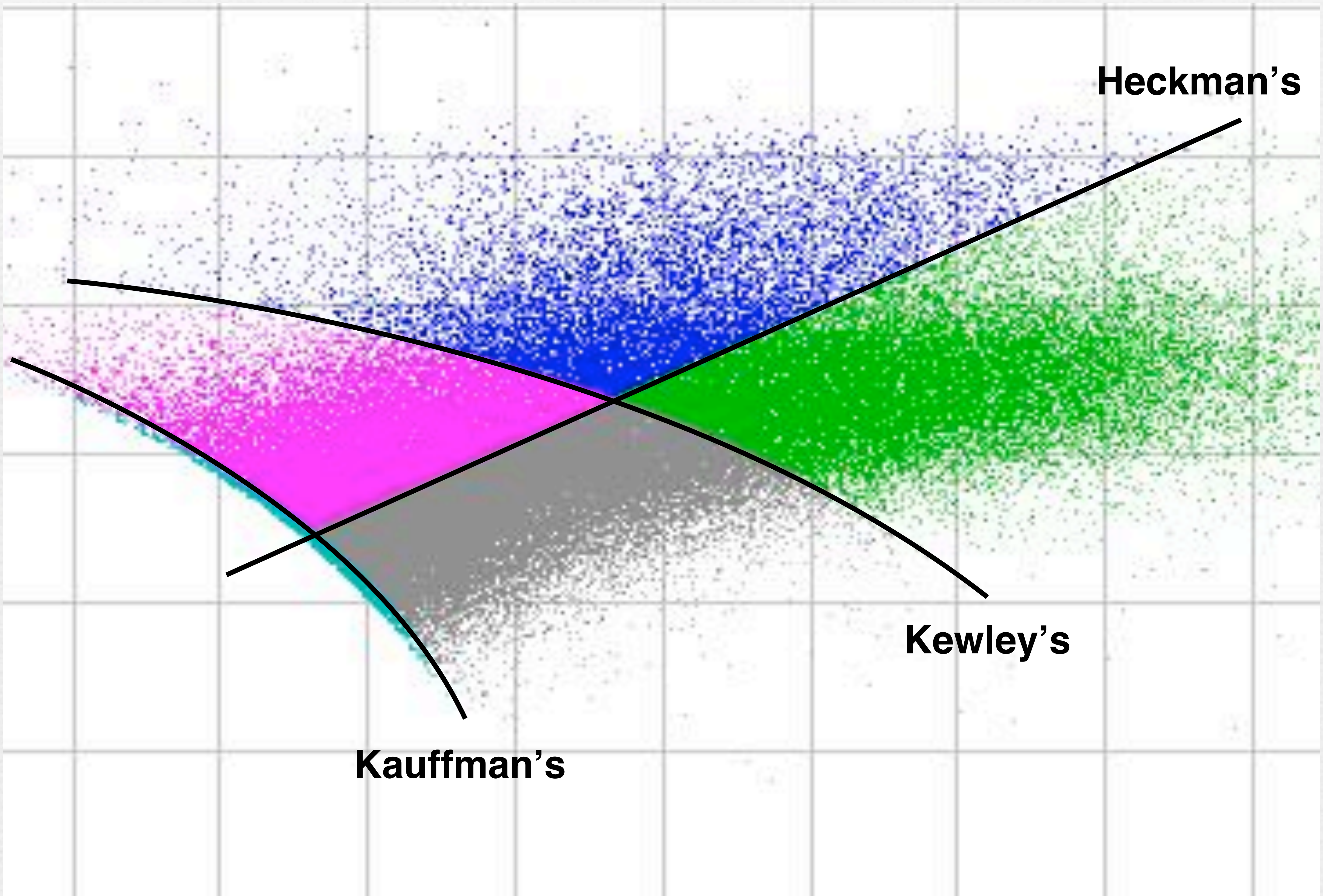
$$\log\frac{[OIII]\lambda5007}{H_\beta} = \frac{0.61}{\log\frac{[NII]\lambda6583}{H_\alpha} - 0.05} + 1.3$$

Heckman's line

$$\log\frac{[OIII]\lambda5007}{H_\beta} = \log\frac{[NII]\lambda6583}{H_\alpha} + 0.465$$

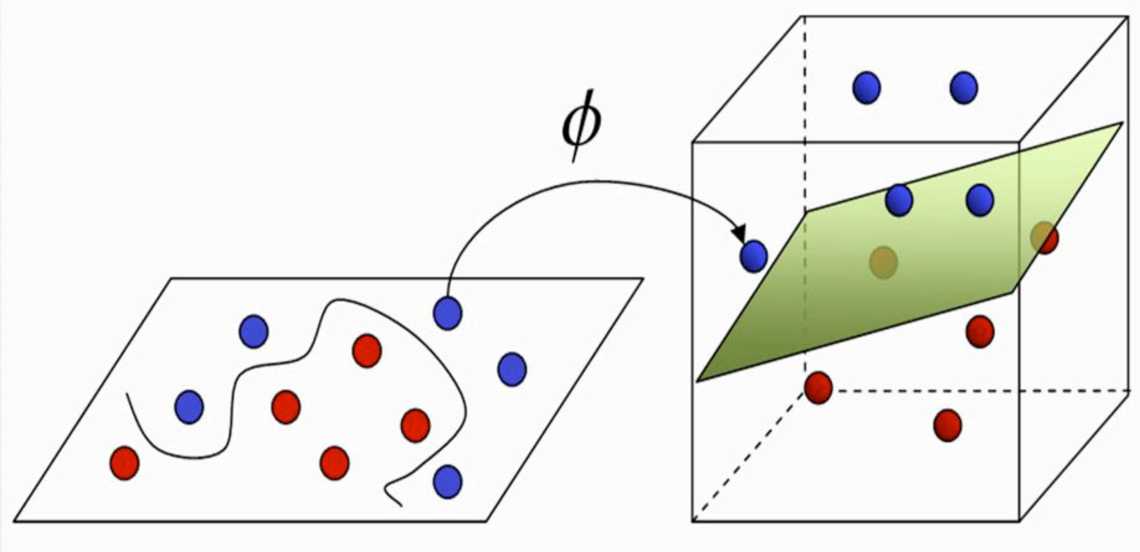
Spectroscopic indicators

Spectroscopic diagnostics used to distinguish starburst galaxies from AGNs and classify AGNs in classes (Sey1, Sey2), based on line intensity ratios (BPT plots).



Optical Spectroscopic AGNs

Support Vector Machines (SVMs) map input vectors to a higher dimensional space where a maximal separating hyperplane is constructed.



experiment	BoK	algorithm	efficiency	completeness
AGN vs Mix	BPT plot + Kewley line	MLP	76%	54%
	BPT plot + Kewley line	SVM	74%	55%
Type 1 vs 2	BPT plot + Kewley line	MLP	95%	~ 100%
	BPT plot + Kewley line	SVM	82%	98%
Seyfert vs LINER	BPT plot + Hecman & Kewley lines	MLP	80%	92%
	BPT plot + Kewley line	SVM	78%	89%

Another development will be the refinement of the BoK by using one 6D space of diagnostics instead of 3 2D spaces.

Photometric redshifts

$$\begin{matrix}\text{luminosity } L \\ \textbf{redshift } \boldsymbol{z} \\ \text{spectral type } T\end{matrix} \xrightarrow{f} \text{observed fluxes}$$

The inverse of this relation provides z_{phot}, statistical in nature but much simpler to measure than z_{spec}:

$$u,g,r,i,z,H,J,K,\ldots \xrightarrow{f^{-1}} \boldsymbol{z}, L, T$$

f^{-1} can be approximated by an empirical relation determined in the photometric parameter space, for a set of sources with z_{spec} available.

Photometric redshifts

Raw materials:
'Data'

Dataset of photometric stellar sources

A belief system:
'Base of Knowledge'

'Spectroscopic redshifts are accurate'

Method:
'Statistical techniques'

Clustering, Neural Networks

Tools:
'Information Technology'

Virtual Observatory distributed computation

Weak Gated Experts: a general DM framework

A composite approach probing different degeneracy regimes in different regions of the features space (PS).

- **PS exploration through unsupervised clustering** performed on the BoK to separate regions with qualitatively different relations between features and targets values;

- **A different 'Expert'** (a single regression machine) **is trained** in every distinct cluster extracted from the BoK distribution in the PS;

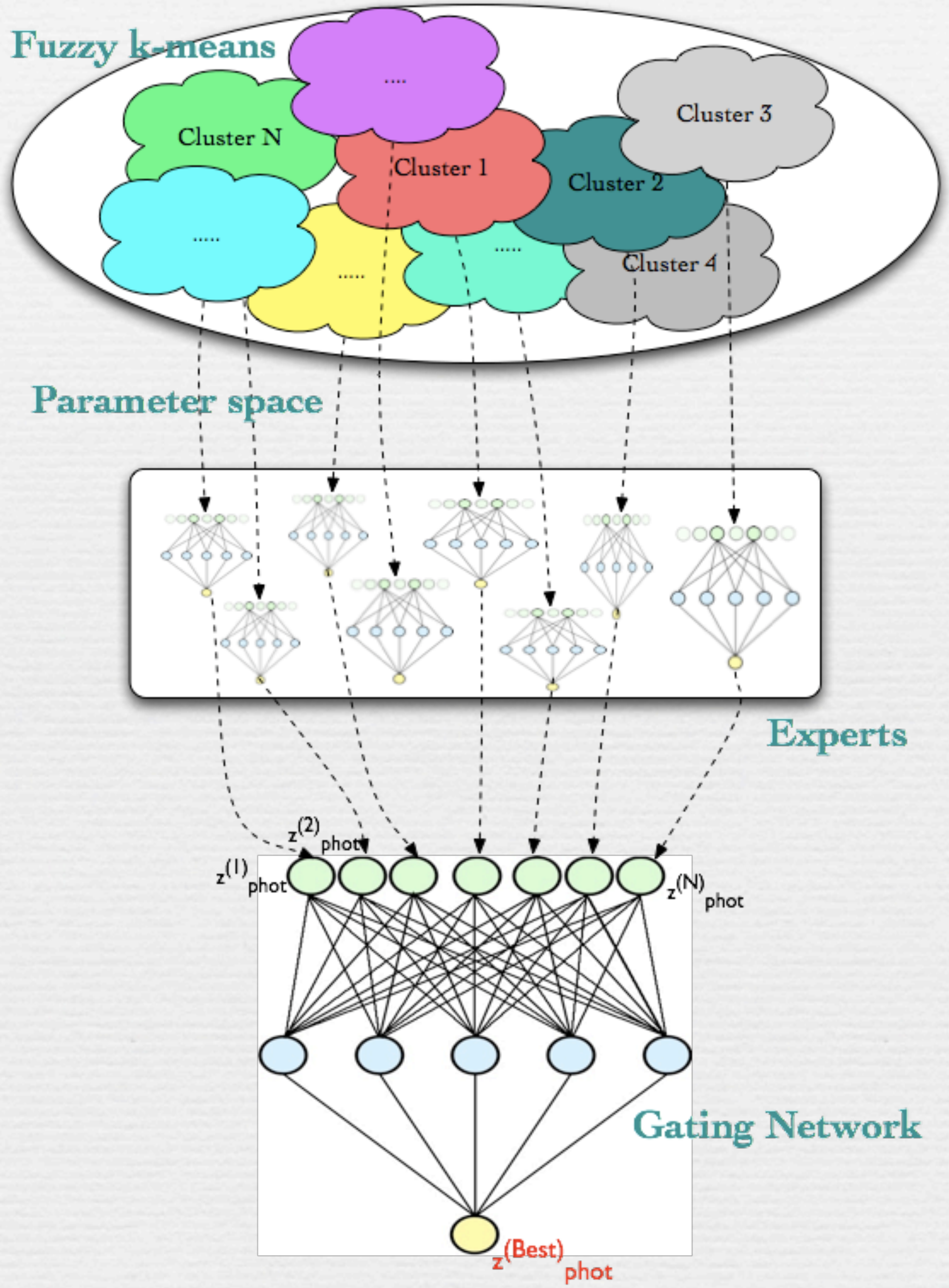
- **The 'Gating Expert' combines the outputs** of different experts and evaluates a more accurate 'merged' output value.

Photometric redshifts: the method

PS clustering

'Experts'

'Gating Expert'



Fuzzy K-means clustering

Neural Networks

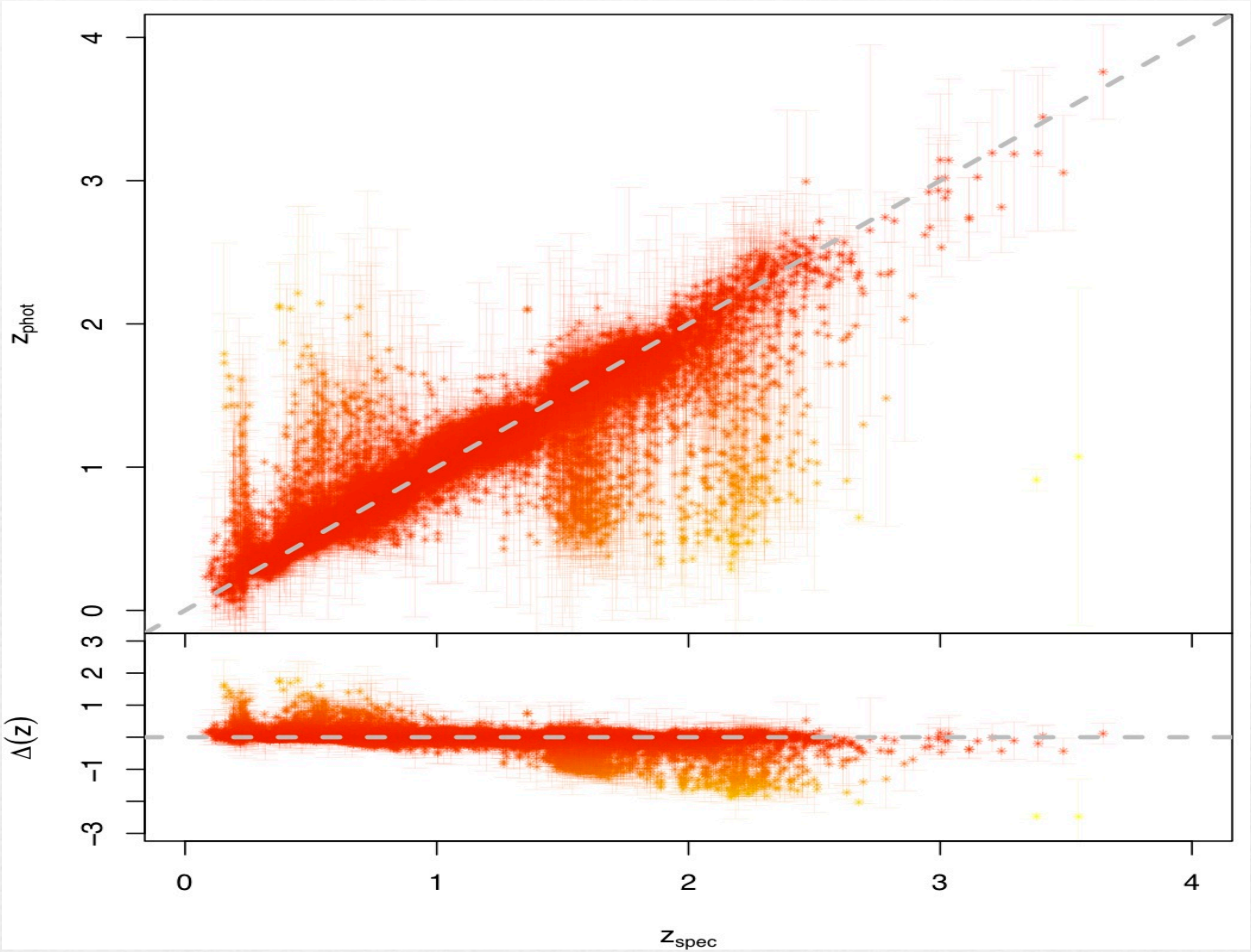
Neural Network
(different architecture)

Photometric redshifts: results (I)

$\langle \Delta z \rangle = 0.012$

$\sigma_{\Delta z} = 0.142$

$\%_{out} < 18\%$



Catalogs of photometric redshifts for optical candidate quasars, optical + UV candidate quasars and optical SDSS galaxies

Pro's of the WGE method

Method	Dataset	Variance	$\frac{\sigma^2}{1+z}$	$\mu\left(\frac{\Delta z}{1+z}\right)$	$\%\Delta_{0.1}$	$\%\Delta_{0.2}$	$\%\Delta_{0.3}$
kNN	S	**0.123**	**0.034**	0.095	54.9	73.3	80.7
kNNPDF	S	–	–	–	53.8	72.4	79.8
CZR	S	0.265	0.079	0.115	**63.9**	**80.2**	**85.7**
WGE	S	0.142	0.059	0.032	48.8	70.3	78.9
WGE+err	S	0.133	0.056	**0.025**	48.7	71.4	80.4
kNN	SG	**0.054**	**0.014**	0.060	70.8	85.8	90.8
kNNPDF	SG	–	–	–	71.8	86.4	90.8
CZR	SG	0.136	0.031	0.071	**74.9**	**86.9**	91.0
WGE	SG	0.058	0.030	0.022	67.9	85.2	91.1
WGE+err	SG	0.057	0.029	**0.012**	69.3	86.2	**91.3**

- **WGE provides errors and flag outliers**: it is trained to recognize distinct regimes in **both** z_{phot} and σ_{zphot};

- **Scalability**: WGE is able to crunch very **large datasets** with limited computational resources;

(Laurino et al., in prep.)

- **Fast training**: WGE readily improves to the data rate of very **large throughputs**;

- **WGE is versatile**: fits well with **different sources** and with general regression and classification problems;

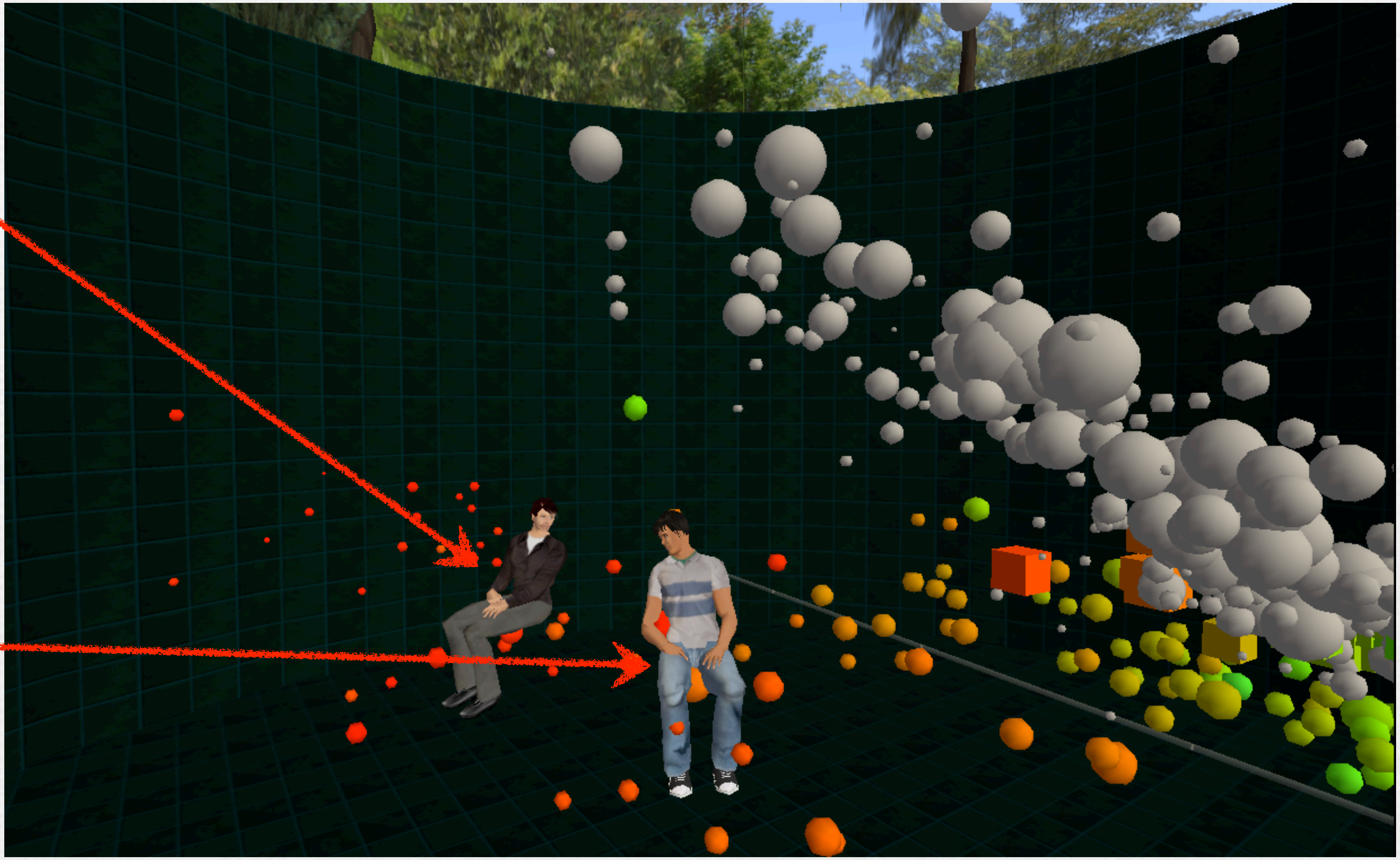
- **WGE is general**: can combine different methods (not based on data mining). Template fitting being included;

How we got there: immersive data exploration

Credit: MICA*

Omar Laurino

Raffaele D'Abrusco



SDSS stellar sources distribution recreated into Second Life (encoded: 4 optical colors, spectroscopic classification, redshift).

*www.mica-vw.org/wiki/index.php/Meta_Institute_for_Computational_Astrophysics

AGNs in the multi-wavelength PS

Characterization of the distribution of AGNs in a high dimensionality parameter space obtained by combining multi-wavelength data through clustering methods.

Multi-wavelength photometry → **proxy for the SEDs**

The primary purpose of this study is to obtain a possible census of AGN behavior in the 13-dimensional space of X-UV1-UV2-ugriz-YJHK-Radio photometry.

- **Classify AGNs according to their overall position in the PS**
- **Pick up outliers and determine their nature through correlations**

The data

X-ray selected AGNs from the Chandra Source Catalog (CSC) sources are used as basis of the parameter space distribution of sources.

Federation of archival photometric data from radio, IR, optical, ultraviolet and X-rays observations.

Crossmatching of catalogues is an issue, but others have already done most of the work at the level of single datasets using SDSS as reference dataset:

CSC-SDSS catalog (Evans et al. 2010)
UKIDSS-SDSS catalog (WFCAM Science Archive)
GALEX-SDSS catalog (Budavari et al. 2009)
VLA-First-SDSS catalog (Kimball & Izevic 2008)

Statistical challenges

- **'Curse of dimensionality'**
  - 11-d parameter space
  - ~10^4 X-ray detected sources in CSC

→ Very low density of the BoK into the PS

The choice of the clustering method(s) is crucial!

- **Censored analysis & Clustering**

≥ 15% of AGNs selected & detected in the other bands have no X-rays counterparts in CSC, but upper limits on their fluxes are available from limiting sensitivity.

No general approach to clustering with censored data

- **Outliers vs Clusters**

Most clustering methods more sensitive to homogeneous groups of sources (clusters) or to isolated sources (outliers).

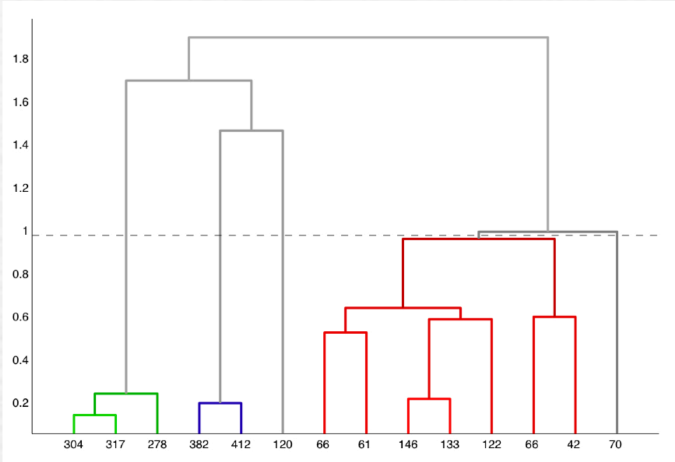
Trade-off between the two aspects

Clustering methods

The choice of the clustering methods depends on the features of the distribution of AGNs and the goal of the experiment. The performance are assessed through simulations. Three classes of interesting algorithms are being tested:

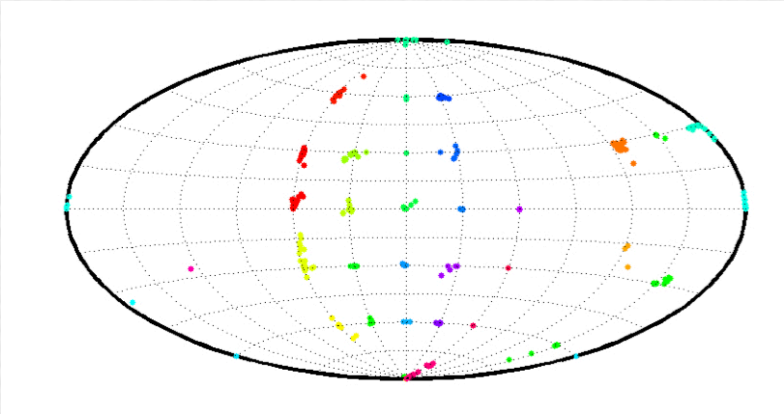
Hierarchical clustering

(no dimensionality reduction, different levels of complexity, intuitive visualization)



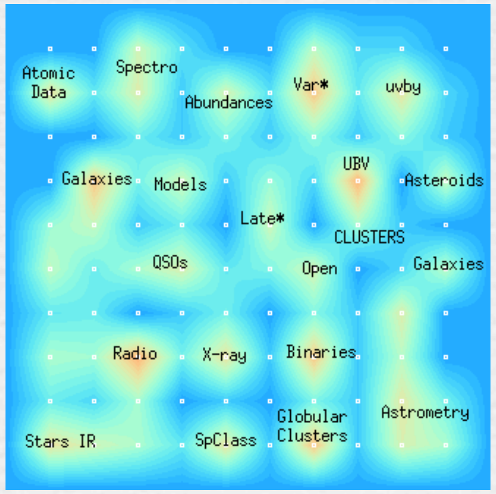
Principal Surfaces

(slow, dimensionality reduction, take care of outliers)



Self Organizing Mapping

(fast, NN based, nice but simplistic visualization, take care of outliers)



Conclusions

- **Astronomical techniques are evolving**: the application of new tools to the large databases that are becoming available, opens an exciting era of data-driven astronomy.

- **Extraction of optical candidate quasars** with unsupervised clustering leads to a substantial improvement of selection performances. Multi-wavelength BoK improves the efficiency and completeness.

- **Classification of optical AGNs** using supervised learning algorithms is promising in order to define future more reliable galaxy classifications without spectroscopic observations.

- **The WGE method** is for the estimation of the photometric redshifts classifications with different sparseness regimes is promising for accurate and fast determination of the 3D distribution of sources.

- **Clustering of AGNs in high dimensional parameter space** composed of multi-wavelength photometry could be interesting as a proxy for full SED characterization and the correlation between distinct classes defined in different spectral intervals.

Results are coming...
See you again in a couple of months.