Department of Physical Sciences - University “Federico II”, Naples

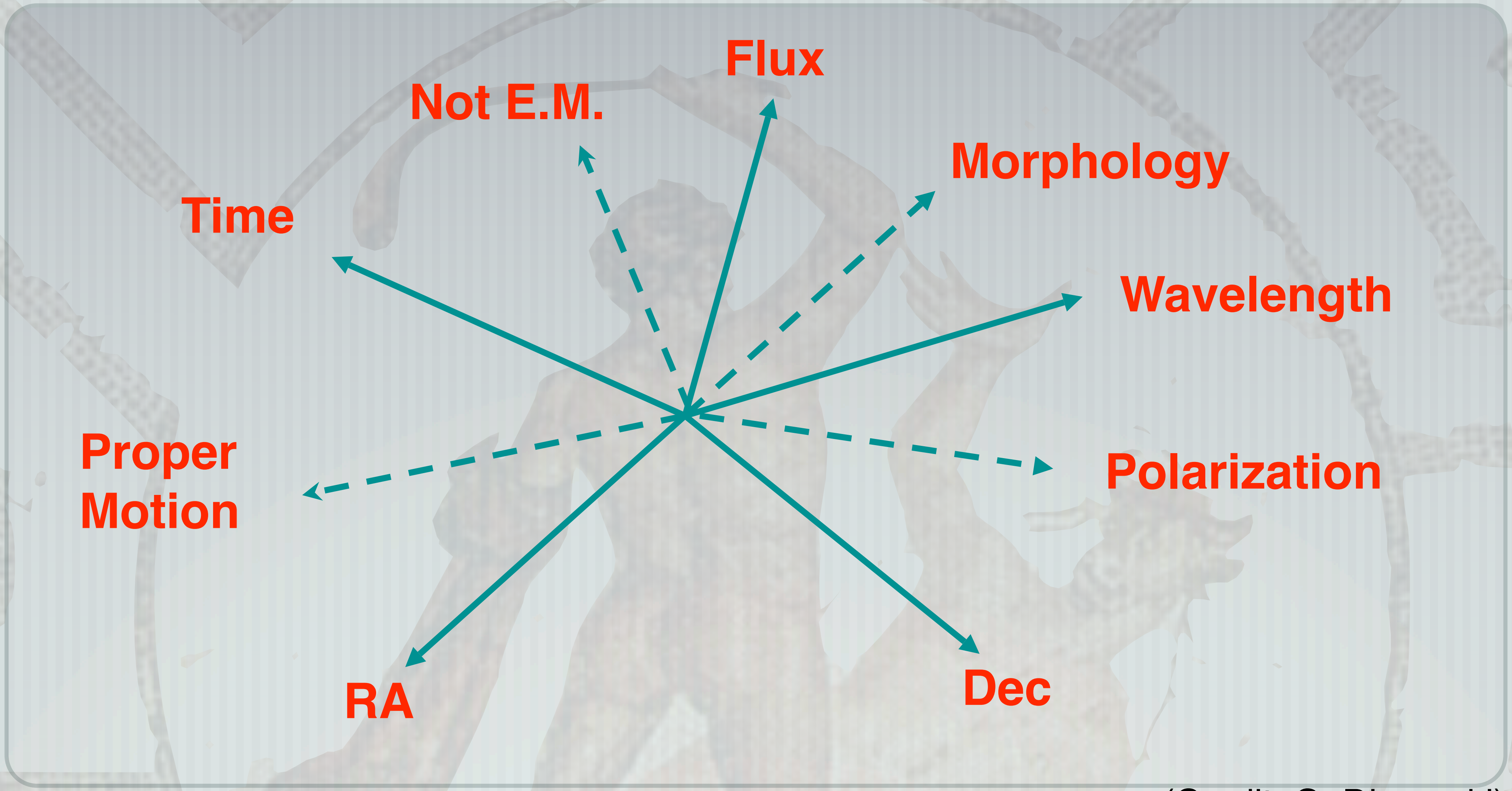
Mining the SDSS for AGNs: the problem of the partition of the parameter space

R. D’Abrusco

in coll. with G. Longo, **S. Cavuoti** and the VONeural team

Astronomical Data Analysis V - Heraklion 9/5/08

Astronomical observational parameter space



(Credit: G. Djorgvski)

Along each axis the measurements are characterized by the position, extent, sampling and resolution. From a mathematical point of view any observation is just a point or a manifold in R^N (with metrics that might be not euclidean.)

All astronomical measurements span some volume in this parameter space. Until now, discoveries were made along some of these axes or in little projected regions of the space, but now new tools and techniques are available.

Tools for astronomy

Clustering & pattern recognition (e.g. structure selection in both real and parameter space)

Classification (e.g. star/galaxy classification, AGNs selection)

Regression (e.g. photometric redshift)

Potential for discovery ⟷ [N_{obj} (data volume) / $N^2_{surveys}$ (connections)] ⟷ [Big surveys / Data federation]

Data Mining algorithms scale very badly:

Clustering ~ $N \log(N) \times N^2$, $\sim D^2$

Correlations ~ $N \log(N) \times N^2$, $\sim D^k$ $(k \geq 1)$

Bayesian Likelihood ~ N^m $(m \geq 3)$, $\sim D^k$ $(k \geq 1)$

Computing power is a finite resource:

Dimensionality reduction and\or algorithm optimization are needed.

Tools for astronomy

Clustering & pattern recognition (e.g. structure selection in both real and parameter space)

Classification (e.g. star/galaxy classification, AGNs selection)

Regression (e.g. photometric redshift)

Potential for discovery ⟷ [N_{obj} (data volume) $N^2_{surveys}$ (connections) ⟷ **VO**

Data Mining algorithms scale very badly:

Clustering ~ $N \log(N) \times N^2$, ~ D^2
Correlations ~ $N \log(N) \times N^2$, ~ D^k ($k \geq 1$)
Bayesian Likelihood ~ N^m ($m \geq 3$), ~ D^k ($k \geq 1$)

Computing power is a finite resource:

Dimensionality reduction and\or algorithm optimization are needed.

The strategy for QSOs and AGNs selection

The scientific start-up: only a small fraction of the QSOs predicted by models and X-ray observations are found in optical and infrared surveys. Data mining techniques can be used to exploit both the abundance of optical\infrared data and the accuracy and windows opened by other-than-optical observations.

In addition, QSOs identification and AGNs classification are complex topics which offer a challenge for the effectiveness of data mining algorithms in astronomy.

- **QSOs identification**: to avoid the risk of loosing objects due to misleading or incomplete classification schemes, unsupervised approaches are to be preferred (by-product: serendipitous discovery of outliers and rare objects).

- **AGNs classification**: a more classical selection algorithm learning how to classify AGNs "by example" can be applied to this kind of problem. The efficiency of selection depends on the parameters and the BoK chosen for the training.

Photometric selection of QSOs

Several algorithms for "general purpose" photometric identification of candidate QSOs select sources according to different techniques exist.

- Optical surveys: looking for counterparts of strong radio sources (but only ~ 10% of QSO are radio-loud).
- Ultraviolet and optical surveys: looking for star-like sources bluer than stars.
- Multi-colour surveys: looking for star-like objects in colour parameter space lying outside compact regions ("star locus") occupied by stars.

Overall performances of a generic targeting algorithm are expressed by two parameters:

Completeness $$c = \frac{\text{candidate quasars identified by the algorithm}}{\text{a priori known quasars}}$$

Efficiency $$e = \frac{\text{confirmed quasars identified by the algorithm}}{\text{candidate quasars selected by the algorithm}}$$

Photometric selection of QSOs

Several algorithms for "general purpose" photometric identification of candidate QSOs select sources according to different techniques exist.

- Optical surveys: looking for counterparts of strong radio sources (but only ~ 10% of QSO are radio-loud).

Not used anymore (in general)

- Ultraviolet and optical surveys: looking for star-like sources bluer than stars.
- Multi-colour surveys: looking for star-like objects in colour parameter space lying outside compact regions ("star locus") occupied by stars.

Overall performances of a generic targeting algorithm are expressed by two parameters:

Completeness $c = \dfrac{\text{candidate quasars identified by the algorithm}}{\text{a priori known quasars}}$

Efficiency $e = \dfrac{\text{confirmed quasars identified by the algorithm}}{\text{candidate quasars selected by the algorithm}}$

Photometric selection of QSOs

Several algorithms for "general purpose" photometric identification of candidate QSOs select sources according to different techniques exist.

- Optical surveys: looking for counterparts of strong radio sources (but only ~ 10% of QSO are radio-loud).

Not used anymore (in general)

- Ultraviolet and optical surveys: looking for star-like sources bluer than stars.

Good

- Multi-colour surveys: looking for star-like objects in colour parameter space lying outside compact regions ("star locus") occupied by stars.

Overall performances of a generic targeting algorithm are expressed by two parameters:

Completeness $$c = \frac{\text{candidate quasars identified by the algorithm}}{\text{a priori known quasars}}$$

Efficiency $$e = \frac{\text{confirmed quasars identified by the algorithm}}{\text{candidate quasars selected by the algorithm}}$$

Photometric selection of QSOs

Several algorithms for "general purpose" photometric identification of candidate QSOs select sources according to different techniques exist.

- Optical surveys: looking for counterparts of strong radio sources (but only ~ 10% of QSO are radio-loud).

Not used anymore (in general)

- Ultraviolet and optical surveys: looking for star-like sources bluer than stars.

Good

- Multi-colour surveys: looking for star-like objects in colour parameter space lying outside compact regions ("star locus") occupied by stars.

Better

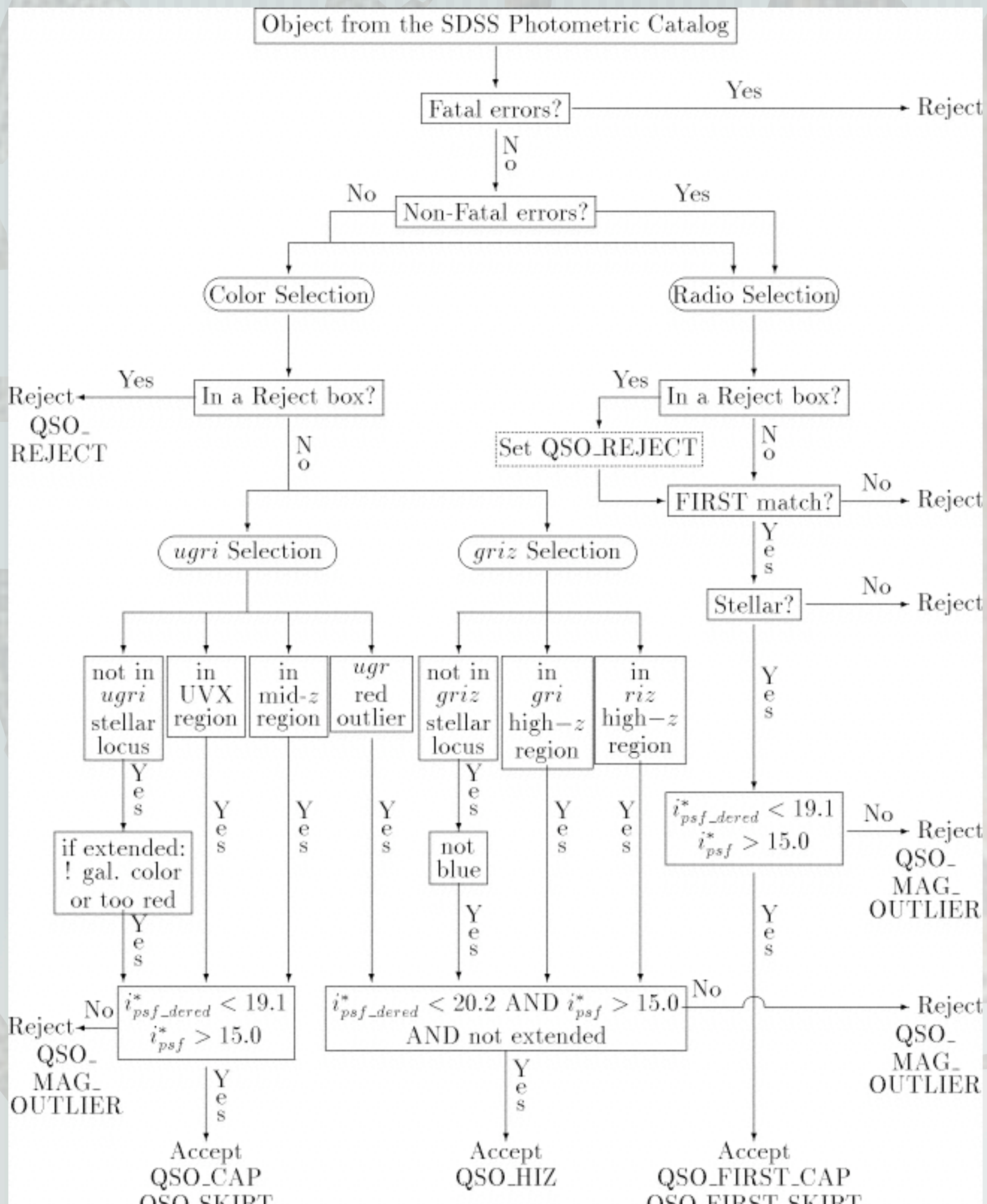
Overall performances of a generic targeting algorithm are expressed by two parameters:

Completeness $$c = \frac{\text{candidate quasars identified by the algorithm}}{\text{a priori known quasars}}$$

Efficiency $$e = \frac{\text{confirmed quasars identified by the algorithm}}{\text{candidate quasars selected by the algorithm}}$$

SDSS QSOs targeting algorithm (I)

SDSS QSO candidate selection algorithm [*Richards et al. 2002*] targets star-like objects as QSO candidate according to their position in the SDSS colours space (u-g,g-r,r-i,i-z), if one of these requirements is satisfied:



- **QSOs are supposed to be placed >4σ far from a cylindrical region containing the "stellar locus"** (S.L.), where σ depends on photometric errors.

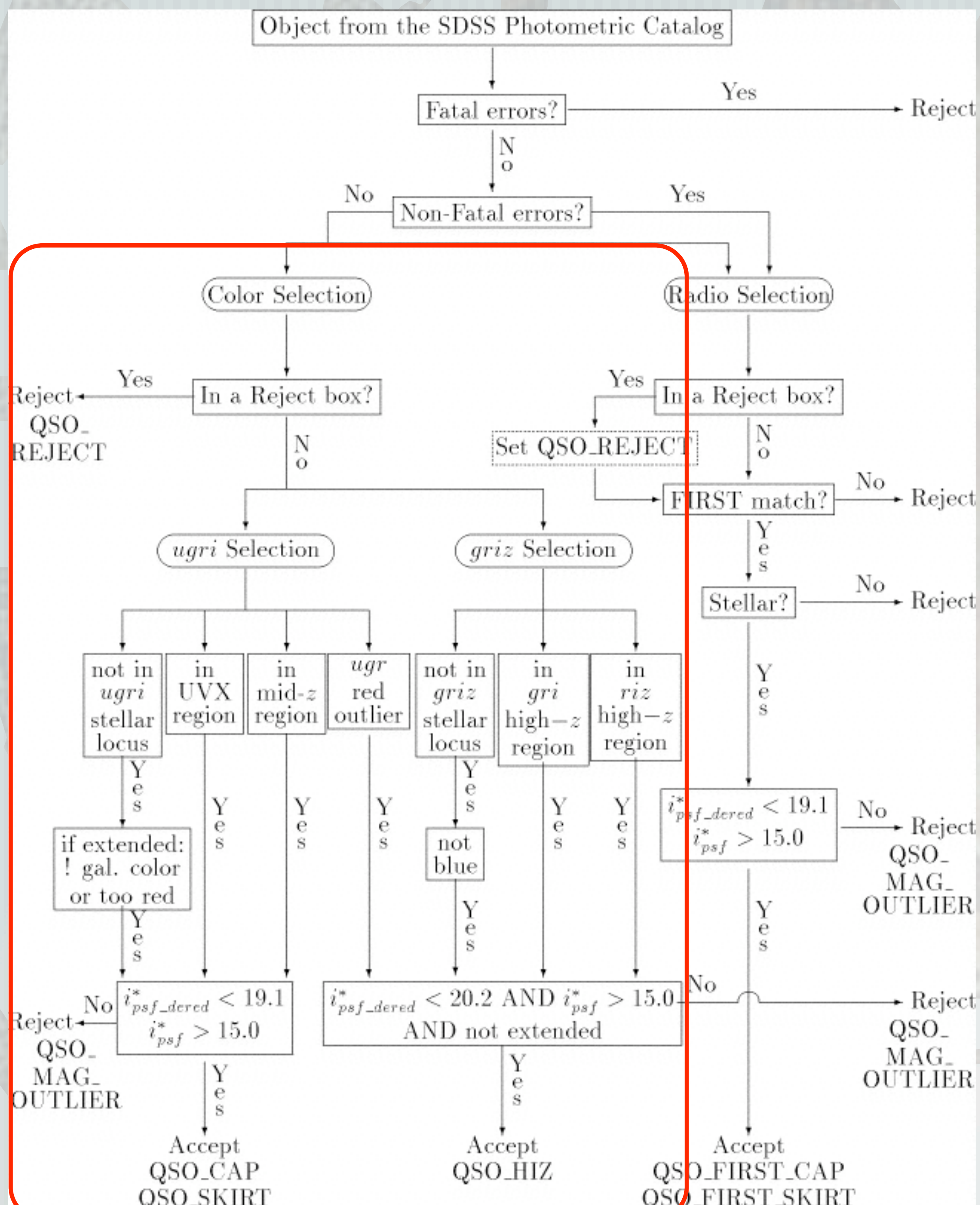
OR

- **QSOs are supposed to be placed inside the inclusion regions**, even if not meeting the previous requirement.

c = 95%, e = 65%
(integrated)

SDSS QSOs targeting algorithm (I)

SDSS QSO candidate selection algorithm [*Richards et al. 2002*] targets star-like objects as QSO candidate according to their position in the SDSS colours space (u-g,g-r,r-i,i-z), if one of these requirements is satisfied:



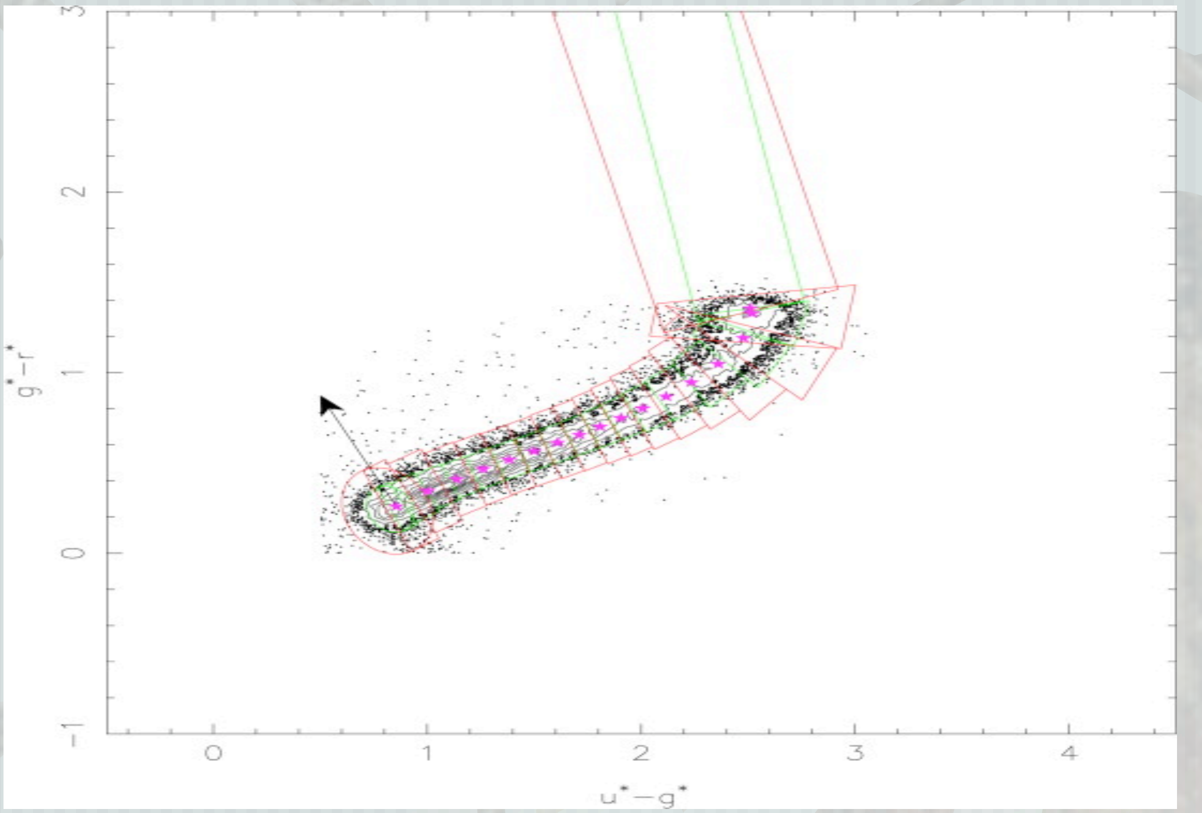
- **QSOs are supposed to be placed >4σ far from a cylindrical region containing the "stellar locus"** (S.L.), where σ depends on photometric errors.

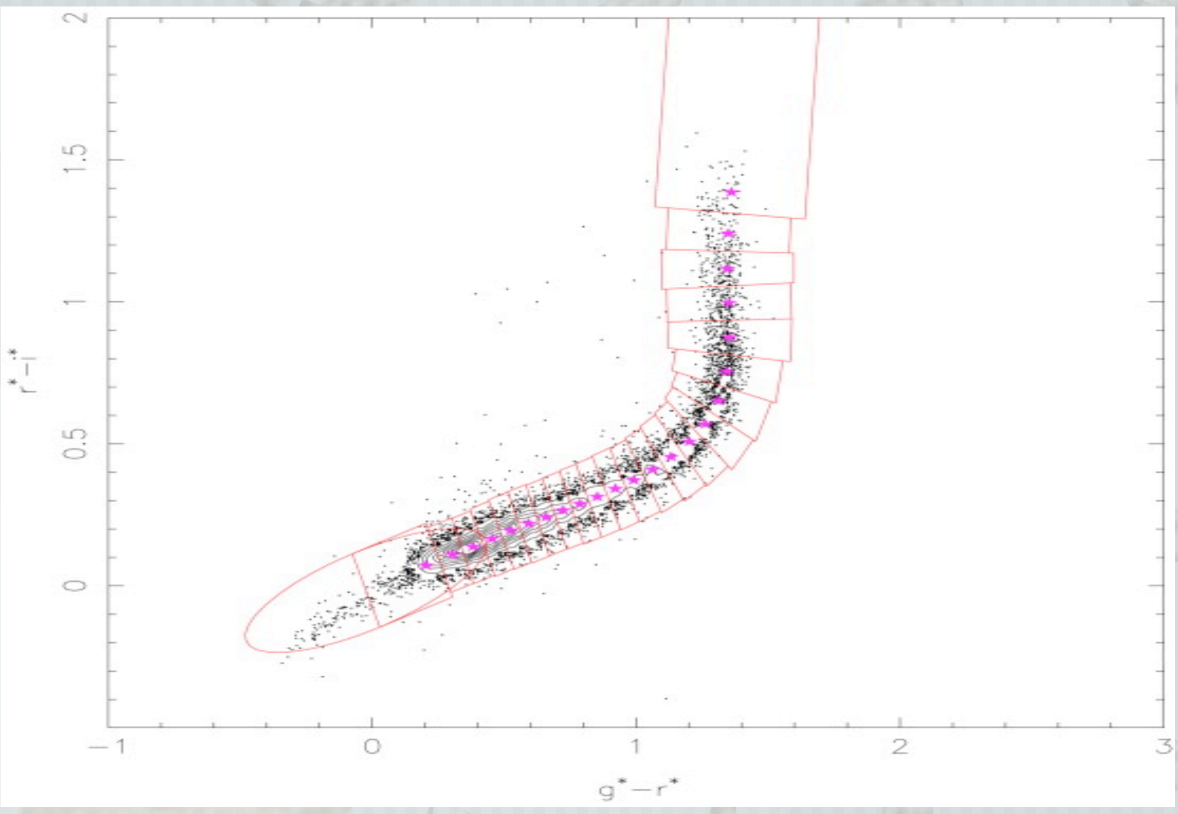
OR

- **QSOs are supposed to be placed inside the inclusion regions**, even if not meeting the previous requirement.

c = 95%, e = 65%
(integrated)

SDSS QSOs targeting algorithm (II)





1. Defined as **inclusion regions** are regions where S.L. meets QSO's area (due to absorption from Lyα forest entering the SDSS filters, which change continuum power spectrum power law spectral index). All objects in these areas are selected so to sample the [2.2, 3.0] redshift range (where QSO density is also declining), but at the cost of a worse efficiency [*Richards et al. 2001*].

2. Defined as **exclusion regions** are those regions outside the main "stellar locus" clearly populated by stars only (usually WDs). All objects in these regions are discarded.

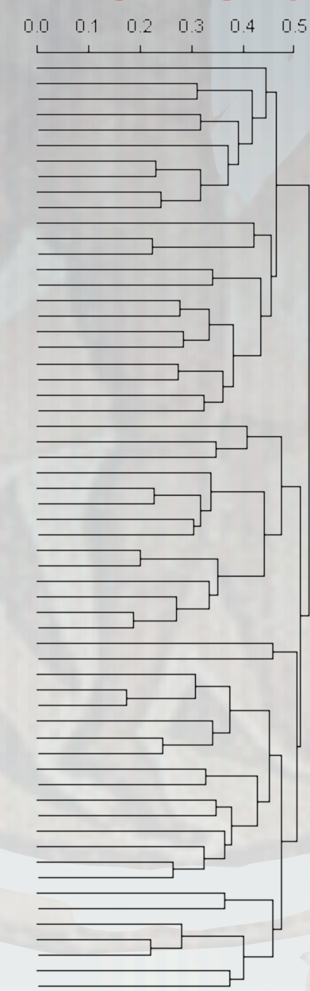
Unsupervised clustering for QSOs

Our candidate QSO selection algorithm is based on unsupervised clustering inside colours space and exploits mixed (spectroscopic+photometric) datasets. Once clusters have been somehow detected, knowledge-base (**spectroscopic types**) is used (i.e., “labels” associated to objects within each cluster) to understand the mixture of objects contained in each cluster and to perform a statistical analysis.

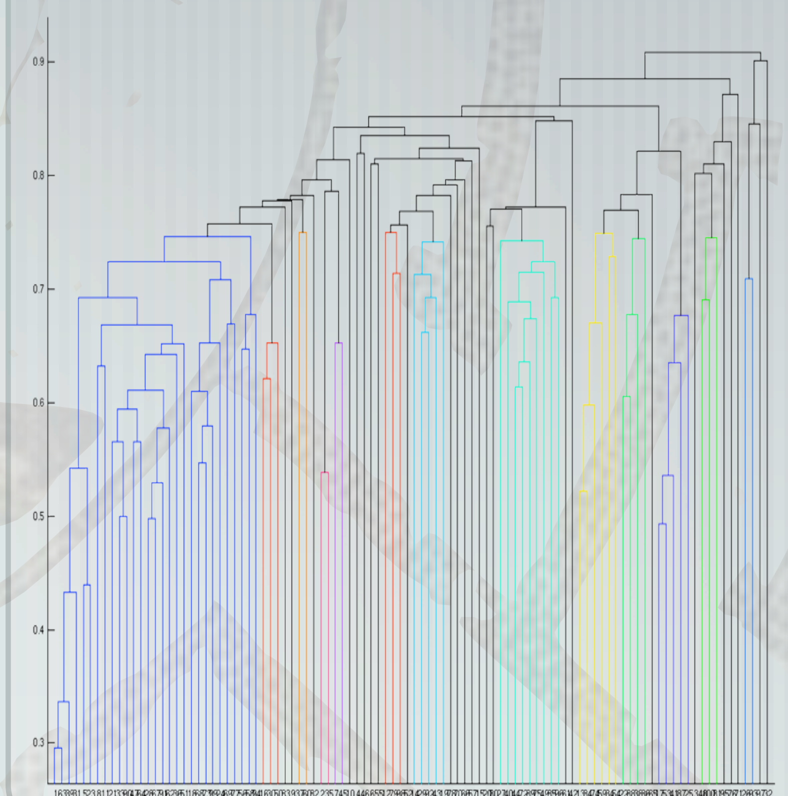
Parameter space

Clustering algorithms

Parameter space with clusters



Parameter space with “labelled” clusters



The algorithm as a whole

Clustering is usually performed on single objects, but this approach may be too sensitive to single outliers to be extensively used in highly non linear parameter space as astronomical ones. We perform a **pre-clustering** on the real distribution of points inside the parameter space, and then used a **clustering algorithm** to aggregate the pre-clusters produced.

1. **Pre-clustering algorithm:** this phase can be accomplished performing a reduction of dimension of the feature space; this reduction via feature extraction/selection can be supervised or unsupervised (our choice in unsupervised).
2. **Agglomerative clustering**: both distance definition and a linkage model (simple, average, complete, Wards, etc.) need to be provided to perform clustering.

A high number of initial latent bases (i.e. clusters from PPS) is good for almost all applications (empty clusters, if any, can be discarded); critical values for D_{th} are classically determined by two similar methods both embodying a **stability criterion**:

1. **Plateau analysis**: final number of clusters N(D) is calculated over a large interval of D, and critical value(s) D_{th} are those for which a plateau is visible.
2. **Dendrogram analysis**: the stability threshold(s) D_{th} can be determined observing the number of branches at different levels of the graph.

The algorithm as a whole

1. **PPS** determines a large number of distinct groups of objects: nearby clusters in the colours space are mapped onto the surface of a sphere.

2. **NEC** aggregates clusters from PPS to a (a-priori unknown) number of final clusters.

3. These clusters are examined and "interesting" ones are selected through **the Base of knowledge.**

Two free parameters to be set are the number of latent variables for PPS ("resolution" of the initial clustering) **and the critical value(s) of dissimilarity threshold D_{th} for NEC.**

A high number of initial latent bases (i.e. clusters from PPS) is good for almost all applications (empty clusters, if any, can be discarded); critical values for D_{th} are classically determined by two similar methods both embodying a **stability criterion**:

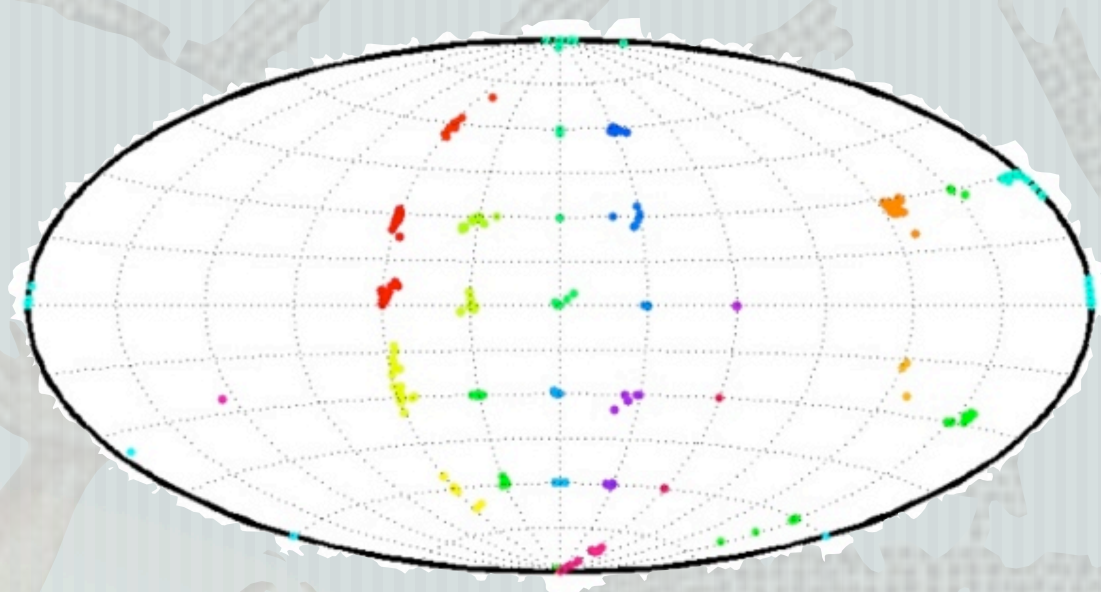
1. **Plateau analysis**: final number of clusters N(D) is calculated over a large interval of D, and critical value(s) D_{th} are those for which a plateau is visible.

2. **Dendrogram analysis**: the stability threshold(s) D_{th} can be determined observing the number of branches at different levels of the graph.

Unsupervised clustering for QSOs

PPS

The Probabilistic Principal Surfaces model [*Chang 2000*] belongs to the family of the so called "latent variables" methods and can be regarded as an extension of the Generative Topographic Mapping.



$$p(\boldsymbol{t}, \boldsymbol{x}) = p(\boldsymbol{x})\, p(\boldsymbol{t} \,|\, \boldsymbol{x})$$

$$\boldsymbol{t} = y(\boldsymbol{x}; \boldsymbol{y}) + \mathbf{u}$$

$y(\boldsymbol{x}; \boldsymbol{y})$ defines a manifold in the data space given by the image of the latent space.

NEC

Unsupervised clustering method based on "negative entropy", an inverse measure of the gaussianity of a distribution.

$$\mathrm{NegE}(p_x) = S(\phi_x) - S(p_x)$$

$$S(p_x) = -\int p_x(u) \log(p_x(u))\, du$$

For each couple of contiguous clusters **A** and **B** in the sample, these two relations are checked. Iff at least one is true, **A** and **B** are replaced by **C** = **A** ∪ **B**.

$$\mathrm{NegE}(A \cup B) < \mathrm{NegE}(A) + \mathrm{NegE}(B)$$

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

Tuning of the method

Once partition of colours space is completed (as a function of D_{th}), **clusters mainly populated by QSOs** (according to the knowledge-base at our disposal) **are selected and informations about these clusters are exploited for selection of QSOs candidates.**

To determine the critical dissimilarity D_{th} threshold we rely not only on a stability requirement.

[**cluster is "successful"**] $\overset{\text{Def}}{\Longleftrightarrow}$ [**its fraction of confirmed QSO is higher then a fixed value**]

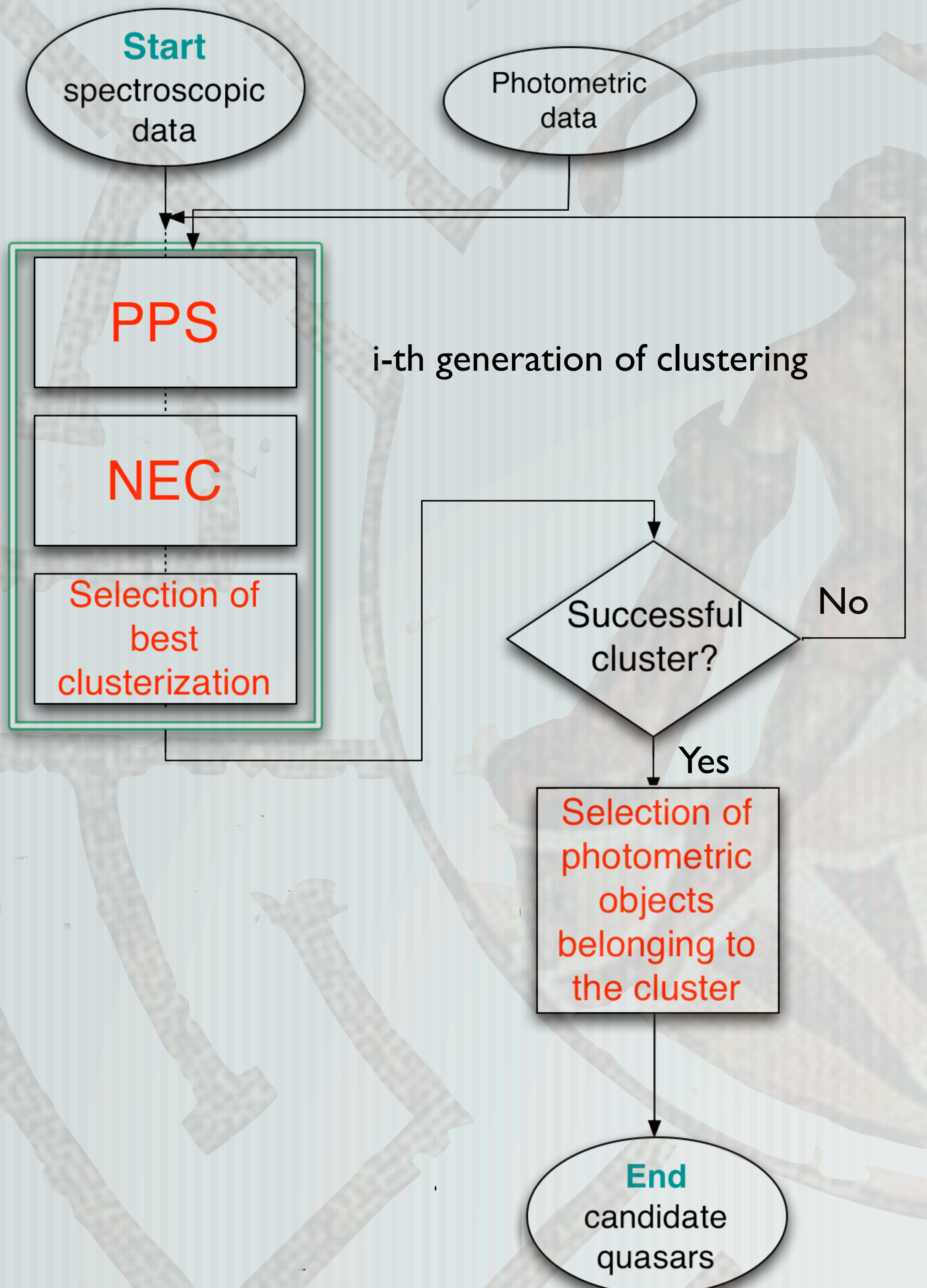
we ask D_{th} to maximise the **Normalised Success Ratio** (NSR):

$$\text{NSR} = \frac{\text{Number of successful clusters}}{\text{Number of total clusters}}$$

The process is recursive: feeding merged unsuccessful clusters in the clustering pipeline until no other successful clusters are found. The overall efficiency of the process e_{tot} is the sum of weighed efficiencies e_i for each generation:

$$e_{tot} = \sum_{i=1}^{n} e_i$$

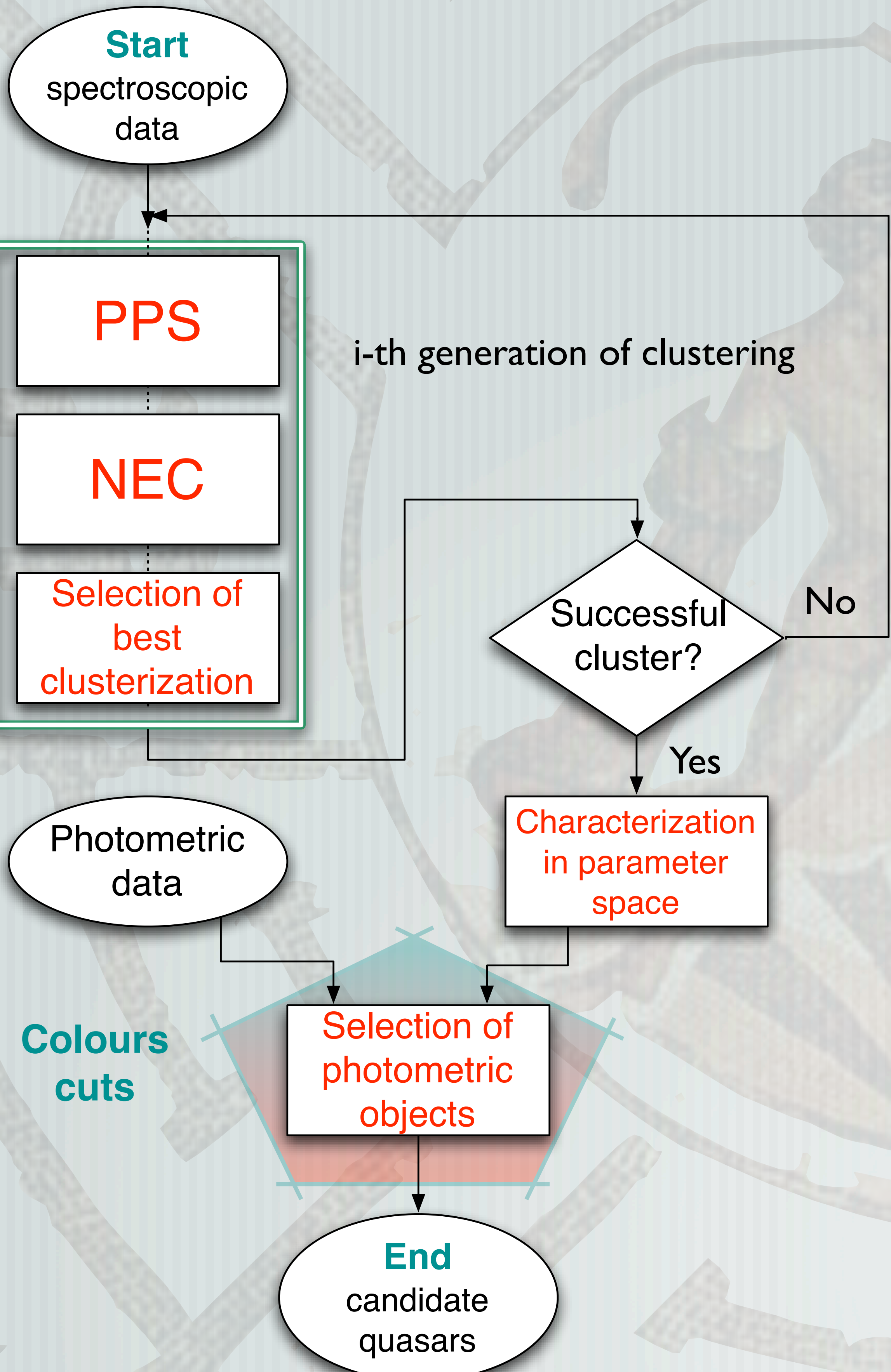
Selection of new candidates



Different methods for the extraction of QSOs candidates

- **"Re-labelling"**: both spectroscopic and photometric objects put into the same clustering process: candidate QSOs are selected as those objects belonging to clusters where spectroscopic confirmed QSOs ("tracers") are found.
- **"Photometric cuts"**: "goal-successful" clusters are described in terms of their colours distribution; associated cuts are applied to photometric sample for candidate selection.
- **"Mahalanobis' distance"**: it is used to measure the distances of a given photometric object from each cluster; the object is assigned to the nearest "goal-successful cluster" or rejected.

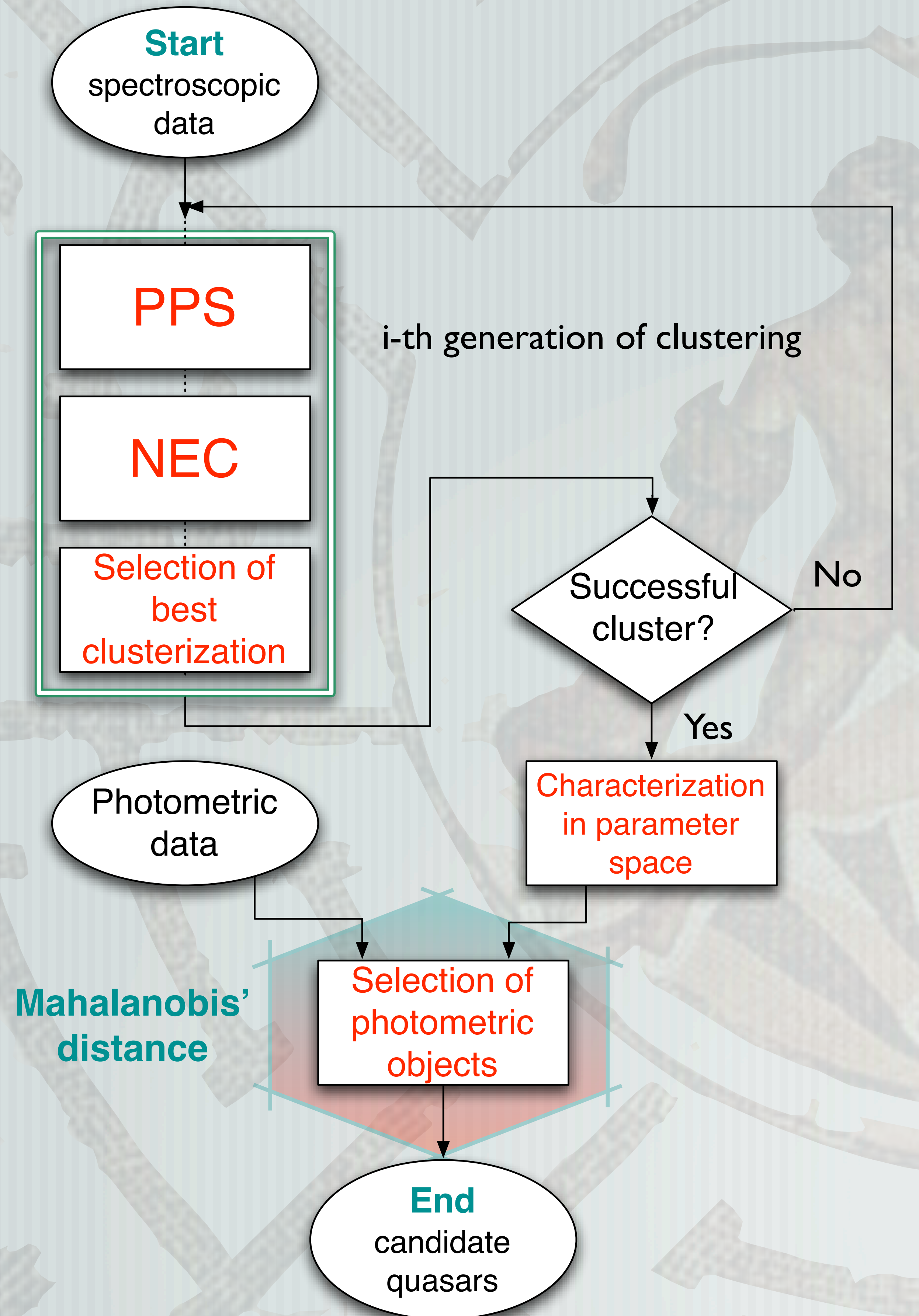
Selection of new candidates



Different methods for the extraction of QSOs candidates

- **"Re-labelling"**: both spectroscopic and photometric objects put into the same clustering process: candidate QSOs are selected as those objects belonging to clusters where spectroscopic confirmed QSOs ("tracers") are found.
- **"Photometric cuts"**: "goal-successful" clusters are described in terms of their colours distribution; associated cuts are applied to photometric sample for candidate selection.
- **"Mahalanobis' distance"**: it is used to measure the distances of a given photometric object from each cluster; the object is assigned to the nearest "goal-successful cluster" or rejected.

Selection of new candidates

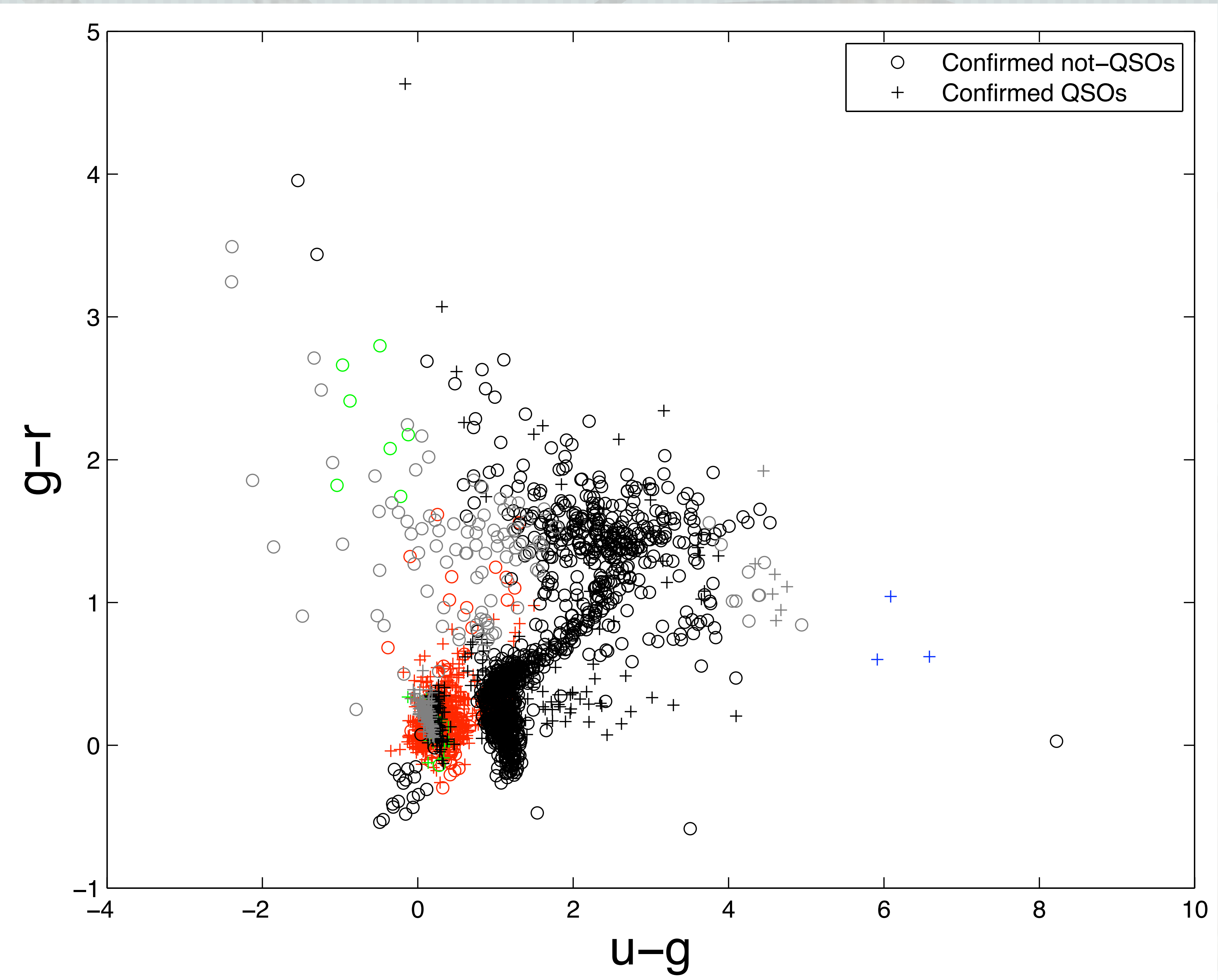


Different methods for the extraction of QSOs candidates

- **"Re-labelling"**: both spectroscopic and photometric objects put into the same clustering process: candidate QSOs are selected as those objects belonging to clusters where spectroscopic confirmed QSOs ("tracers") are found.
- **"Photometric cuts"**: "goal-successful" clusters are described in terms of their colours distribution; associated cuts are applied to photometric sample for candidate selection.
- **"Mahalanobis' distance"**: it is used to measure the distances of a given photometric object from each cluster; the object is assigned to the nearest "goal-successful cluster" or rejected.

Comparison with a colour-colour plot

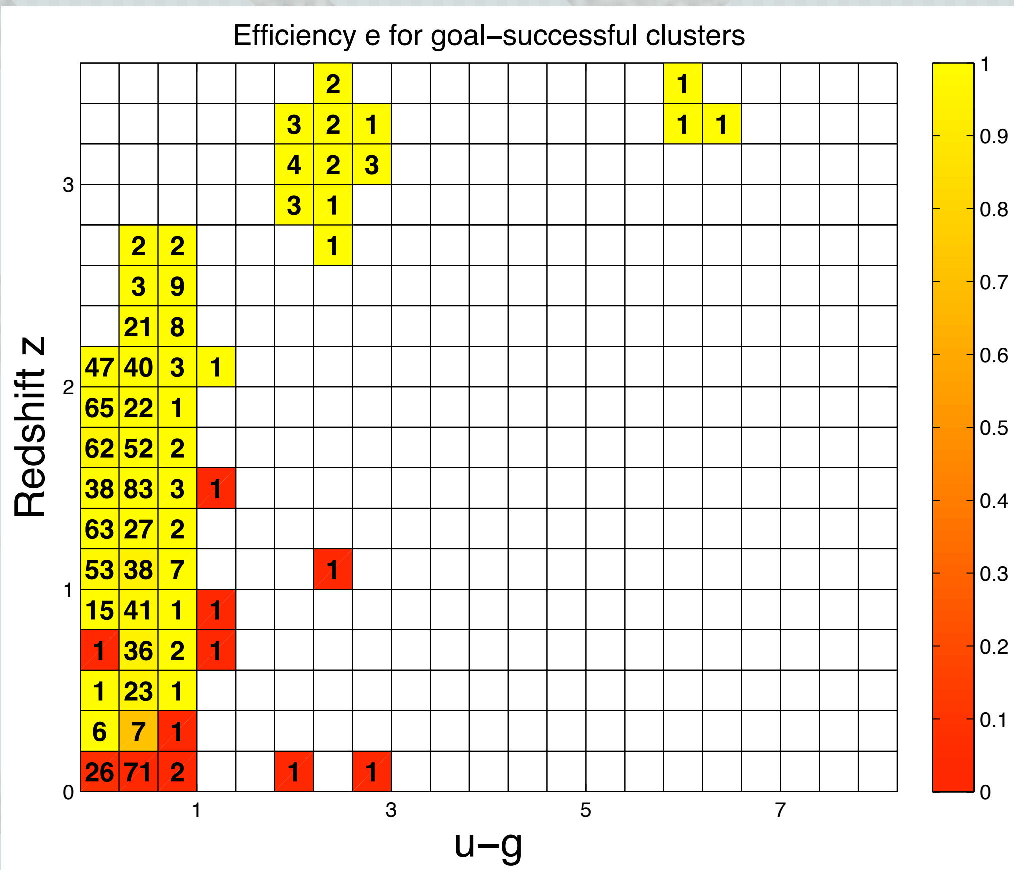
u - g vs **g - r**



Only a fraction (43%) of these objects have been selected as candidate QSOs by SDSS targeting algorithm: the remaining sources have been included in other spectroscopic programmes (mainly stars and galaxy).

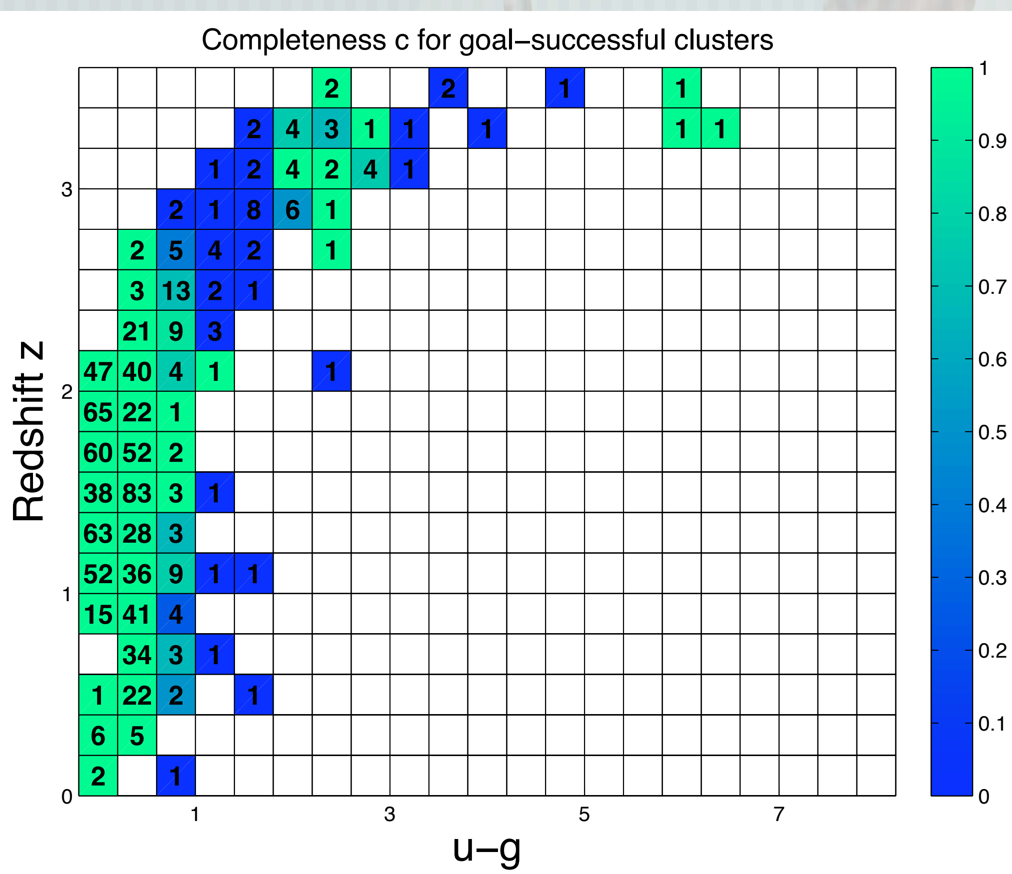
In this experiment the clustering has been performed on a sample of stellar objects observed in optical **and** infrared observations, but using **only** optical colours. The BoK is the SDSS spectral classification index “specClass”.

Exp 3: local values of *e* and *c*



Local values of the efficiency *e* in the (z vs u-g) planes.

Details on "how many" wrong objects we are selecting and "where" in the parameter space (and if we are smart enough, on why we are selecting them...)



Local values of the completeness *c* in the (z vs u-g) planes.

Details on "how many" right objects we are losing and "where" in the parameter space (and if we are smart enough, on why we are selecting them...)

Results (I)*

Sample	Parameters	Labels	e_{tot}	c_{tot}	n_{gen}	n_{suc_clus}
Optical SDSS QSOs candidates	4 optical colours	'specClass'	83.4 % (± 0.3 %)	89.6 % (± 0.6 %)	2	(3,0)
Optical + NIR star-like objects	4 optical colours + 3 infrared colours	'specClass'	91.3 % (± 0.5 %)	90.8 % (± 0.5 %)	3	(3,1,0)
Optical + NIR star-like objects	4 optical colours	'specClass'	92.6 % (± 0.4 %)	91.4 % (± 0.6 %)	3	(3,0,1)

* [*D'A. et al.*, submitted to MNRAS.]

Classification of Active Galactic Nuclei

Most galaxy classifications are based on morphological informations, which only partly reflect the physical differences between different class of objects. One clear example is represented by **galaxies containing AGNs, which do not fit comfortably inside any morphological classification known (except weak correlations).**

A combination of unsupervised and supervised classification methods might work also in this case...

Selection of active of galaxies in terms of a minimal set of spectroscopic and photometric parameters embodying the physical differences of their nuclei as closely as possible.

The complex and more general problem of the physical classification of galaxies lurking in the darkness...

Classifying with NNs

For photometric redshift data mining reduces to interpolation, and NNs are best interpolators. This approach exploits the existence of a “knowledge base”, i.e. of subsample of the data for which spectroscopic redshift are available, to be set as priors.

In a suitable parameter space (photometric space), a neural network is trained on a subsample of objects with reliable classification (“targets”). This trained network is used to select photometric candidates.

Knowledge base

Training, Validation, Test

Parameters + Target

Dataset

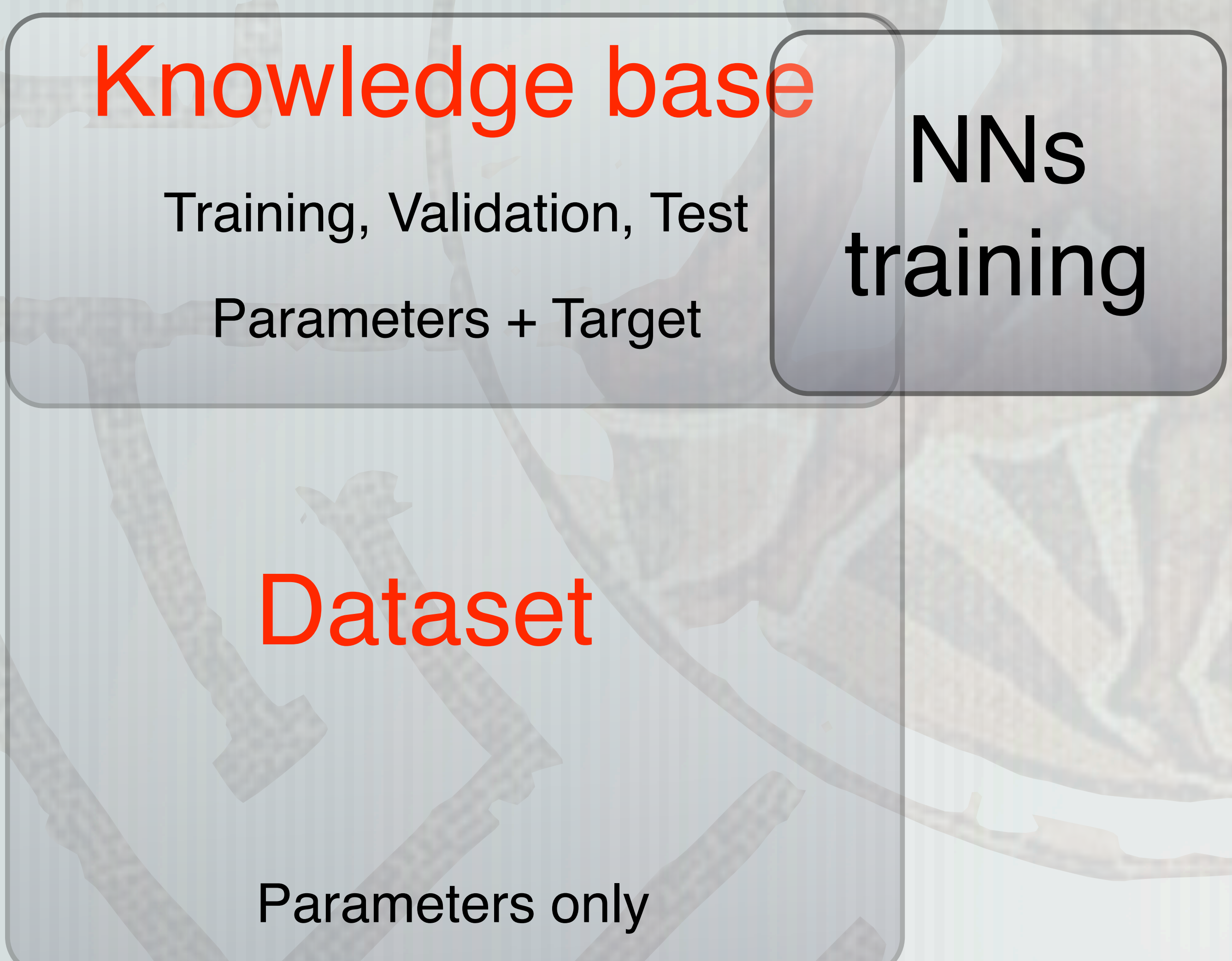
Parameters only

NNs

Classifying with NNs

For photometric redshift data mining reduces to interpolation, and NNs are best interpolators. This approach exploits the existence of a “knowledge base”, i.e. of subsample of the data for which spectroscopic redshift are available, to be set as priors.

In a suitable parameter space (photometric space), a neural network is trained on a subsample of objects with reliable classification (“targets”). This trained network is used to select photometric candidates.



Classifying with NNs

For photometric redshift data mining reduces to interpolation, and NNs are best interpolators. This approach exploits the existence of a “knowledge base”, i.e. of subsample of the data for which spectroscopic redshift are available, to be set as priors.

In a suitable parameter space (photometric space), a neural network is trained on a subsample of objects with reliable classification (“targets”). This trained network is used to select photometric candidates.

Knowledge base

Training, Validation, Test

Parameters + Target

Dataset

Parameters only

trained NNs

Classifying with NNs

For photometric redshift data mining reduces to interpolation, and NNs are best interpolators. This approach exploits the existence of a “knowledge base”, i.e. of subsample of the data for which spectroscopic redshift are available, to be set as priors.

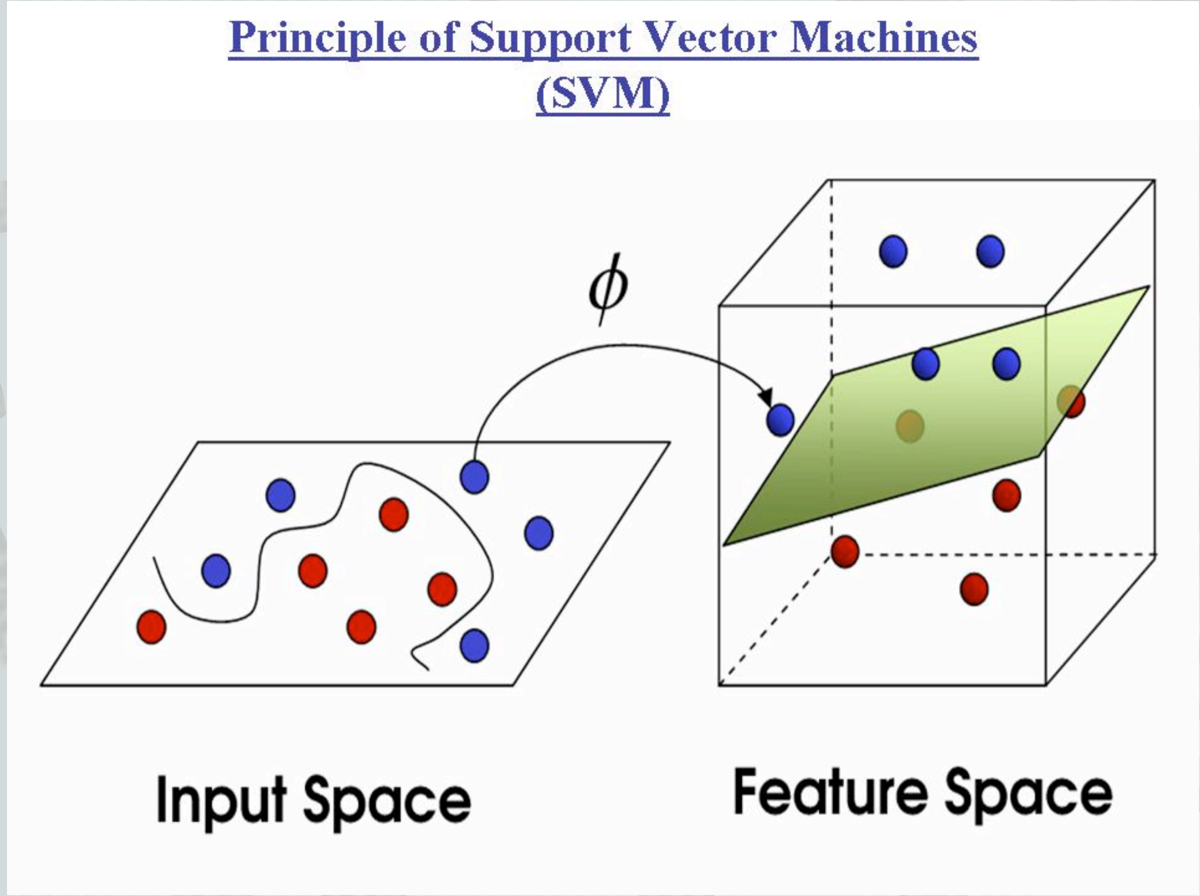
In a suitable parameter space (photometric space), a neural network is trained on a subsample of objects with reliable classification (“targets”). This trained network is used to select photometric candidates.

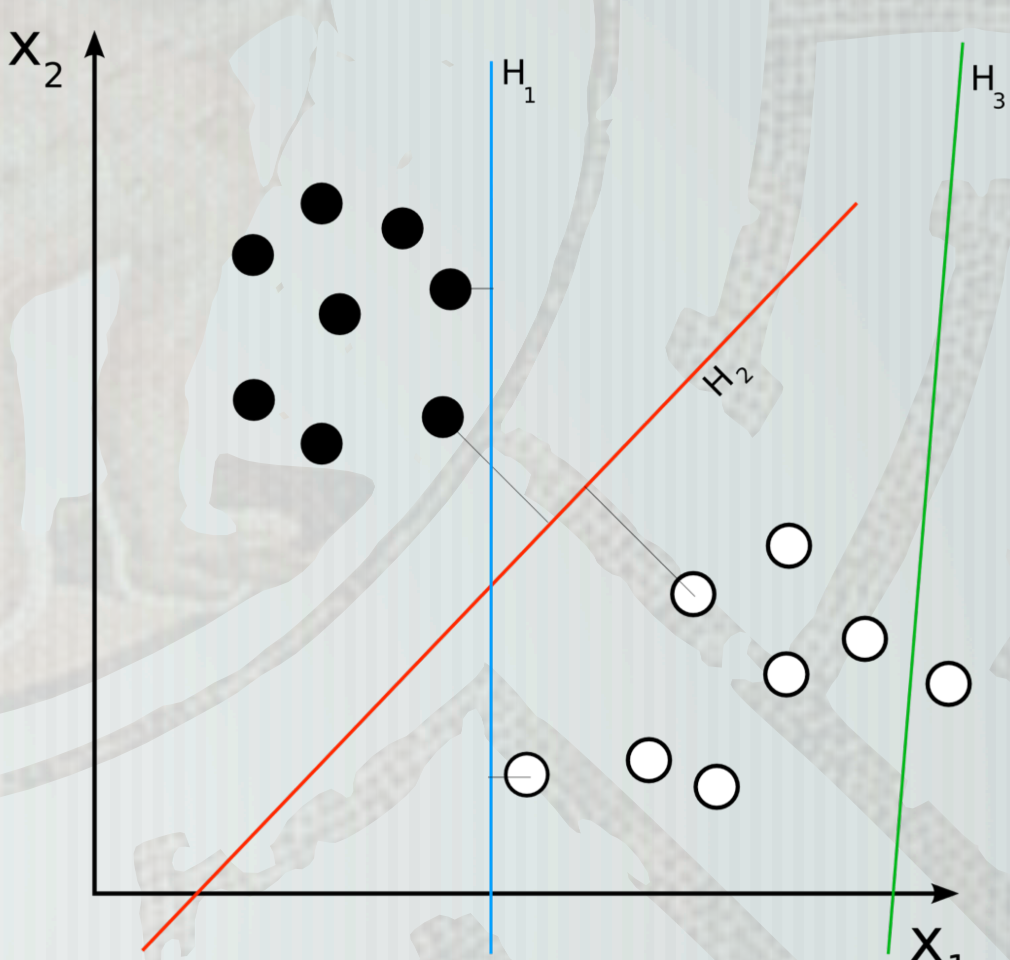
Support Vector Machine

Support vector machines (SVMs) [*Bennet & Campbell 2000*] **are a set of related supervised learning methods used for classification and regression.**

SVMs map input vectors to a higher dimensional space where a maximal separating hyper-plane is constructed. The "kernel function" of the SVM in the "C-Support Vector Classification" implementation [*Boser et al. 1992*], depends on two parameters, one in the model (C) and the other in the "kernel function" (γ):

$$K(x_i, x_j) = e^{(-\gamma \| x_i - x_j \|^2)}$$





SVM parameter space exploration strategy

The sampling strategy of the 2-dim parameter plane (γ, C) for the SVM (proposed by [Hsu et al.]) consists in running different jobs on a grid whose knots are spaced by a factor 4 on both parameters ($\gamma = 2^{-15}, 2^{-13} \ldots 2^{3}$, $C = 2^{-5}, 2^{-3}, \ldots 2^{15}$).

Cross-validation of results and “folding” (5 subsets) of the dataset are used for all experiments.

Heavy computing tasks, mainly independent and parallelization-prone (according to the “batch-parameter” paradigm). → **Grid computing**

The SVM experiments for different couples of values of the parameters (γ, C) have been run on a 112 knots grid infrastructure of the SCOPE Virtual Organization.

SVM parameter space exploration strategy

The sampling strategy of the 2-dim parameter plane (γ, C) for the SVM (proposed by [Hsu et al.]) consists in running different jobs on a grid whose knots are spaced by a factor 4 on both parameters ($\gamma = 2^{-15}, 2^{-13} \ldots 2^{3}$,

Cross-validation of results and “fo

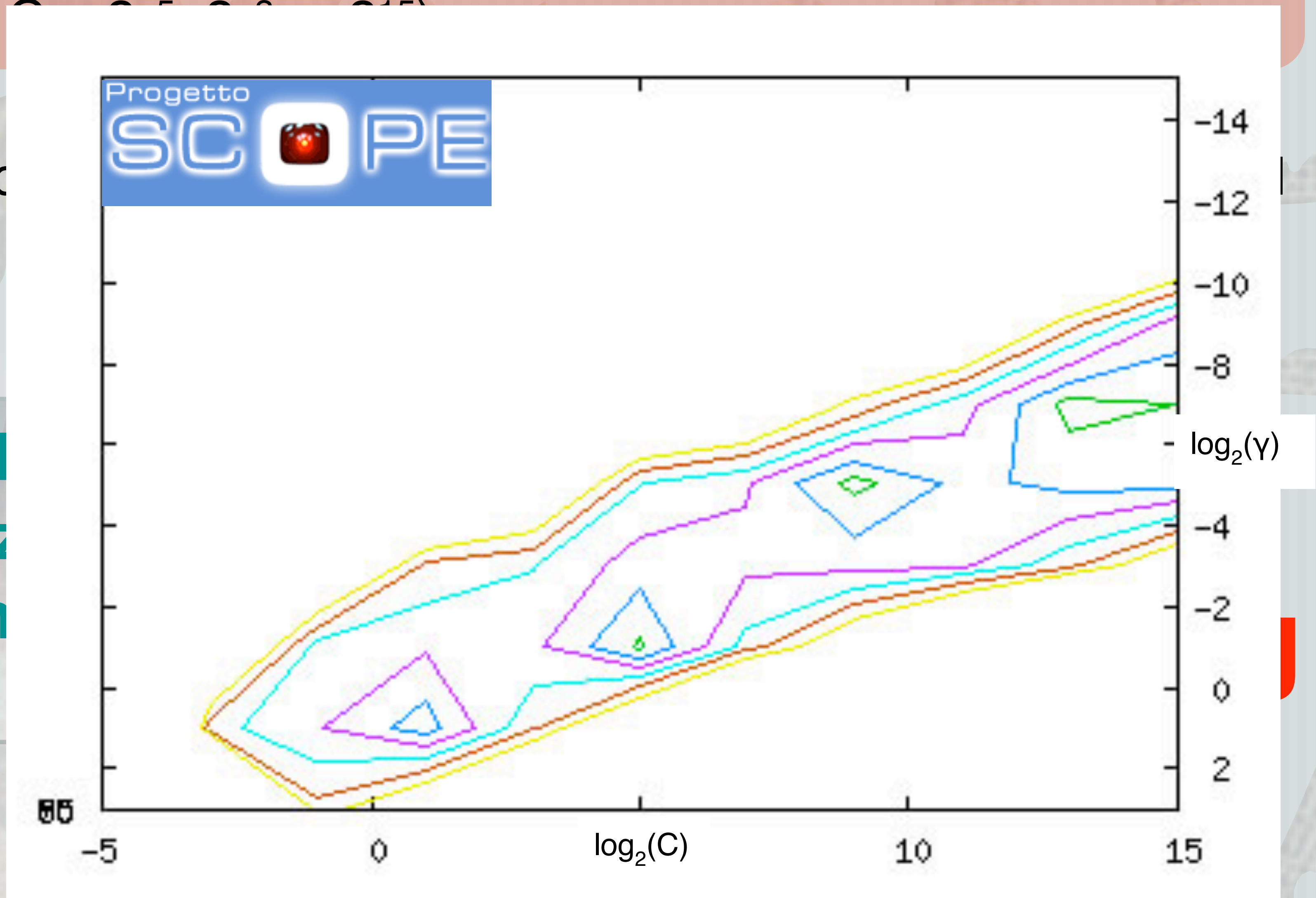
experiments.

Heavy computing tas

independent and paralleliz

(according to the “batch

paradigm).

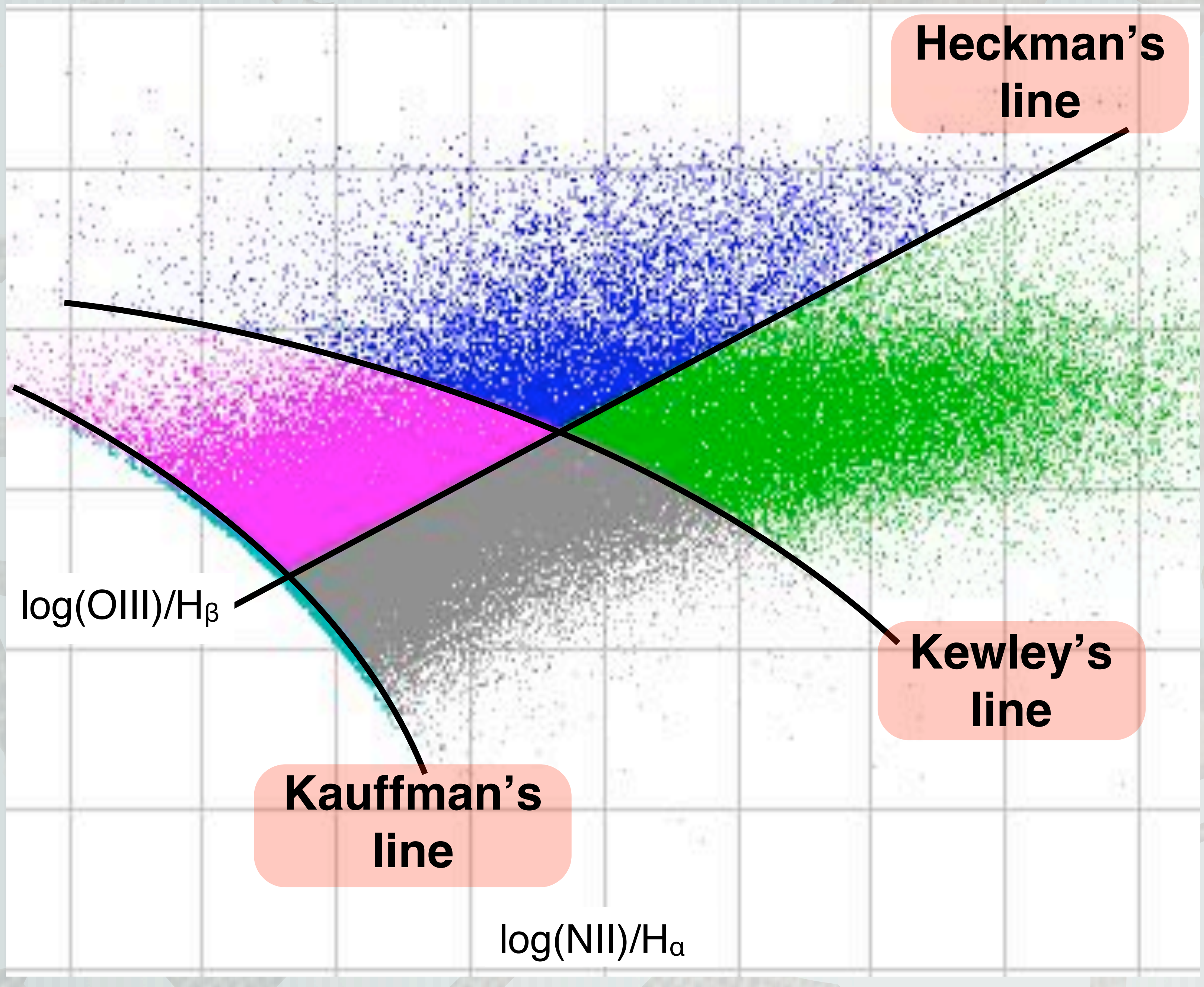


The SVM experiments for different

run on a 112 knots grid infrastructure of the SCOPE Virtual Organization.

A 2-dim BoK (I)

Diagnostics plots based on spectroscopic observables are useful to disentangle AGNs from other galaxies: the luminosity of the [OIII]λ5007 emission line is a tracer of the strength of activity in the nuclear region, so A BPT plot [*Baldwin et al. 1981*] based on the emission lines flux ratios OIII/H_β ratios NII/H_α can be used to separate SB galaxies and AGNs [*Kewley et al. 2001*].



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

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

Parameters of the experiments

Photometric parameters

PhotoObjAll table (SDSS-DR5) → petroR50_u, petroR50_g, petroR50_r, petroR50_i, petroR50_z
concentration_index_r
fibermag_r
$(u-g)_{dered}$, $(g-r)_{dered}$, $(r-i)_{dered}$, $(i-z)_{dered}$
dered_r

Photometric redshifts [*D'A. et al. 2007*] → photo_z_corr

Targets

1° Experiment: AGNs vs SB Galaxies:

AGNs -> 1, Mixed -> 0

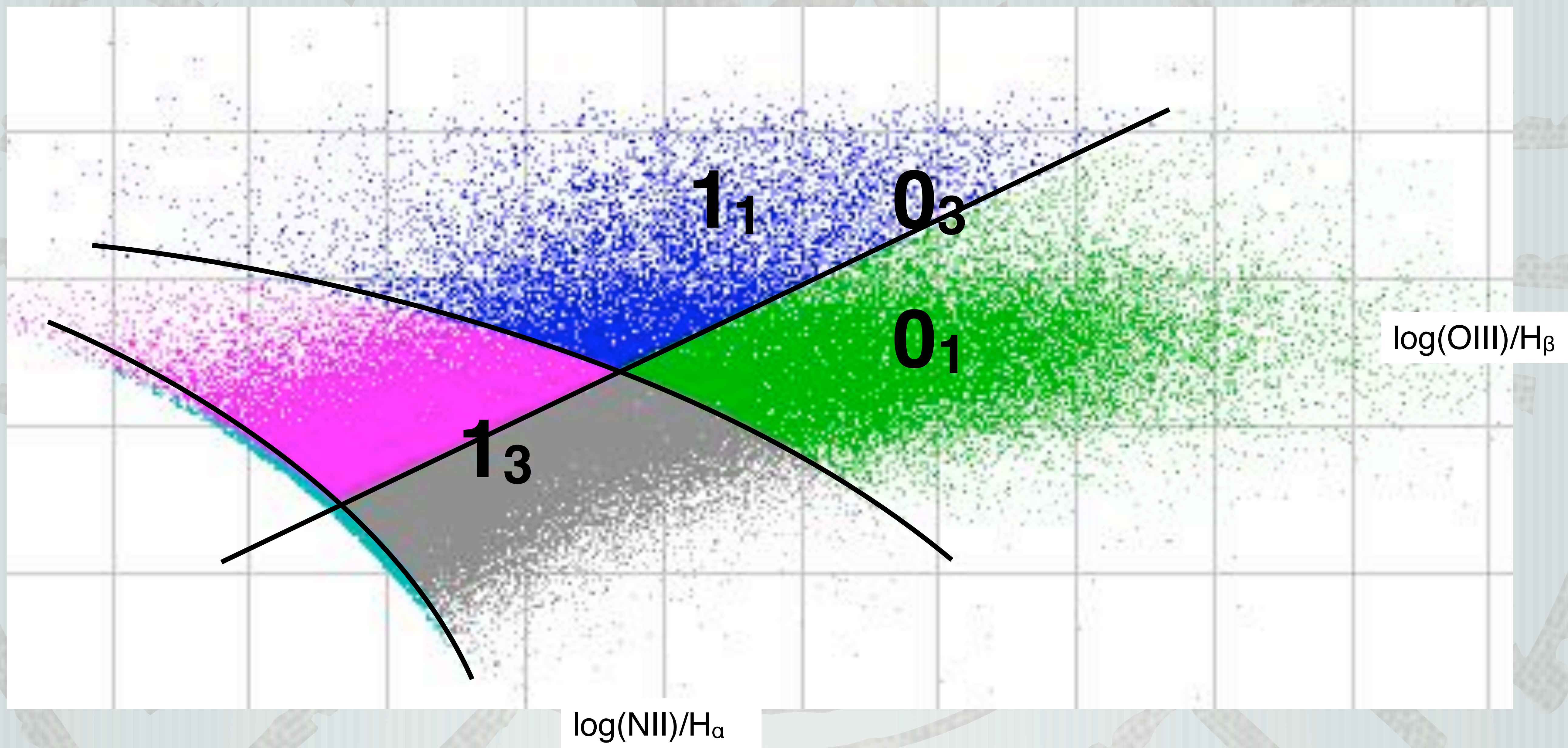
2° Experiment: Type I AGNs vs Type II AGNs:

Type 1 -> 1, Type 2 -> 0

3° Experiment: Seyferts vs LINERs:

Seyfert -> 1, LINERs -> 0

A 2-dim BoK (II)



Results (II)*

Sample	Parameters	BoK	Algorithm	e_{tot}	c_{tot}
SDSS galaxies (1)	SDSS photometric parameters + photo redshift	**BPT plot + Kewley's line**	***SVM*** / ***MLP***	~74% / ~76%	~55% / ~54%
SDSS galaxies (2)	SDSS photometric parameters + photo redshift	**BPT plot + Kewley's line + FWHM requirements**	***SVM*** / ***MLP***	e_1~82% e_0~86% / e_0~98% e_1~95%	~98% / ~100%
SDSS galaxies (3)	SDSS photometric parameters + photo redshift	**BPT plot + Heckman's+ Kewley's lines**	***SVM*** / ***MLP***	~78% / ~80%	~89% / ~92%

* *[Cavuoti, D'A., Longo, 2008], in prep.*

Improving AGN classification

How can we improve AGNs selection and classification?

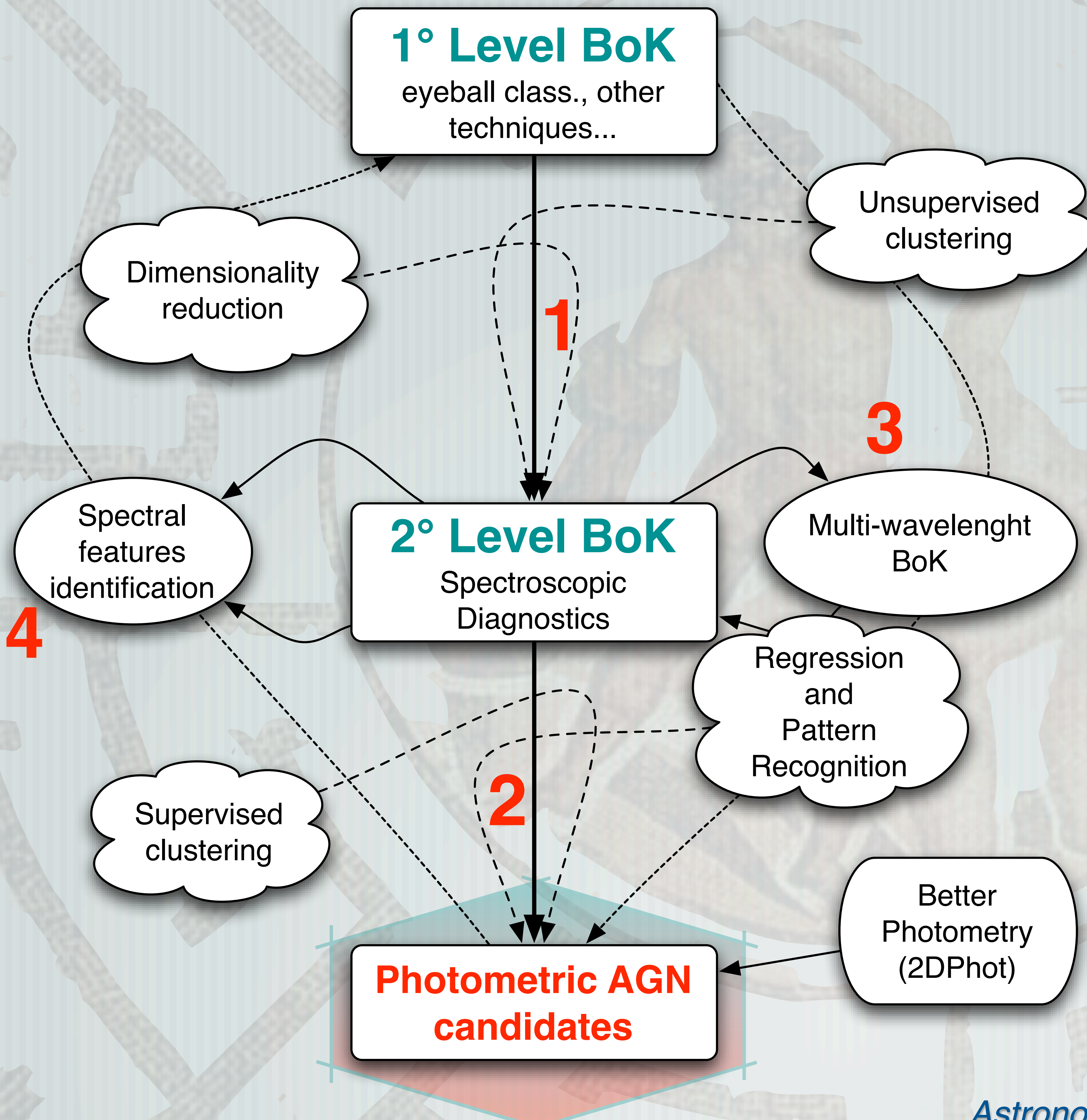
Improving the accuracy of the photometric parameters measuring them with specifically tailored tools (in coll. with De Carvalho and La Barbera - OAC).

Improving the efficiency of separation between different families using better spectroscopic diagnostics, i.e. **using a better BoK** (in coll. with Rafanelli and Benvenuti - Università di Padova)

Customizing the algorithms for specific tasks in order to determining more efficient criteria/ strategies for the fine tuning of the results (better exploration of plane (γ, C) for SVM, for instance.)

The improvement of the Base of Knowledge can be accomplished not only enhancing the quality of spectroscopic classification, but also **by enlarging the wavelength range whence the BoK is extracted**. This is a possible approach to connect differently selected AGNs (X-ray ⇔ optical/infrared)

A long-haul endeavour



1. Determination of a more efficient separation of different types of AGNs in a multi-dimensional spectral diagnostics space.
2. Selection of photometric candidates exploiting a BoK based on the previous point.
3. Selection of AGNs from optical/NIR photometric data using a not-optical BoK.
4. Automatic identification of spectral patterns and features using .

Conclusions

- Unsupervised clustering methods have proved to be efficient and reliable for QSOs photometric candidate selection. Refinements of BoK will be key for further enhancements.

- The integration of supervised classification algorithms for photometric data and unsupervised selection of spectroscopic BoK for AGNs classification looks promising.

- Syncretism of expertise: astrophysics, statistics, data mining, distributed computing together. Who could ask for more?

- VO provides powerful and flexible tools for massive data gathering and analysis. Our algorithms have been deployed inside the Astrogrid VO environment and are available for everyone. For details, documents, papers browse the link:

http://people.na.infn.it/~astroneural/