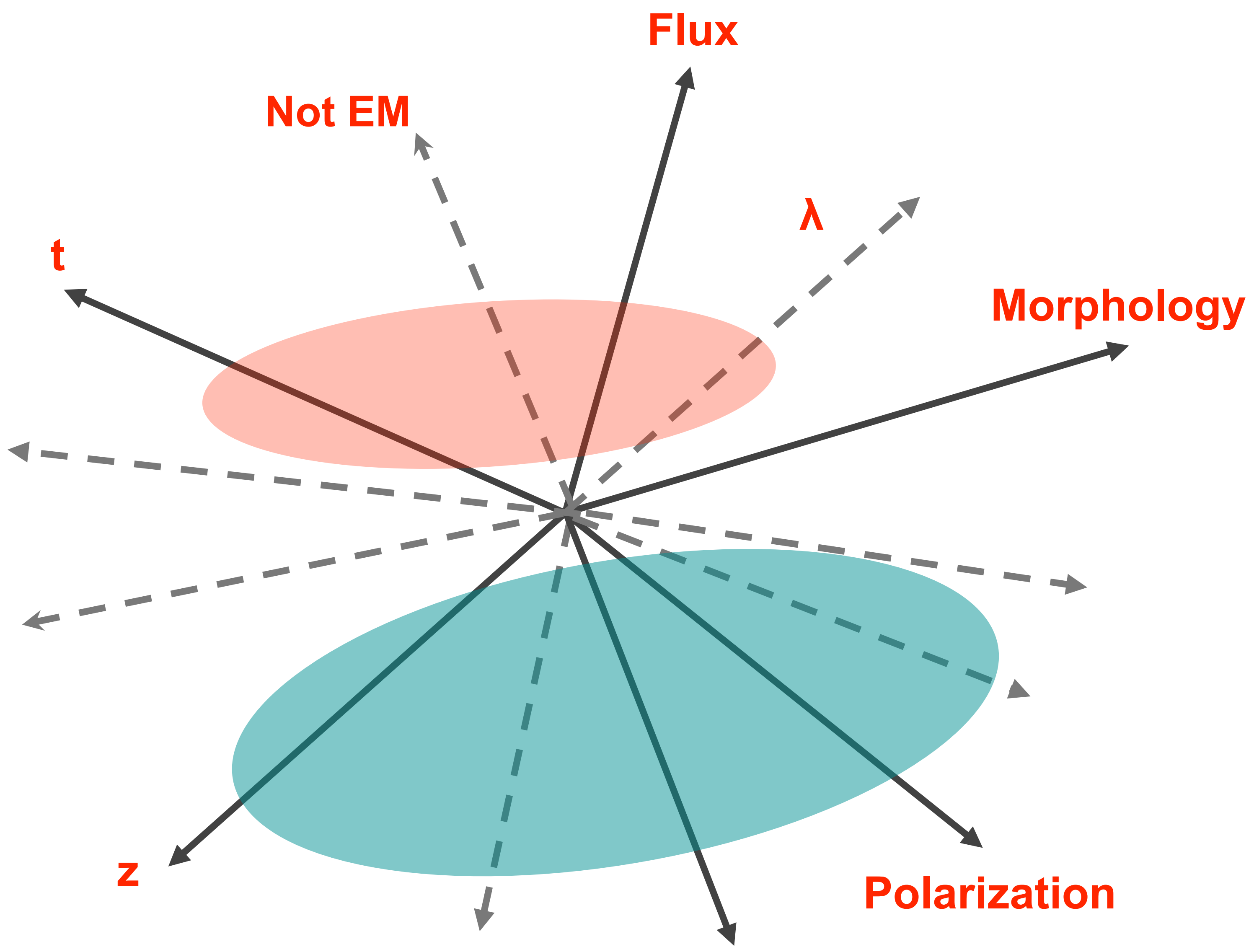
CLaSPS: Knowledge Discovery for the exploration of complex multi-wavelengths astronomical datasets. Applications to CSC+, a sample of AGNs built on the Chandra Source Catalog and to a Blazars sample.

R. D'Abrusco

Harvard-Smithsonian Center for Astrophysics

Motivations

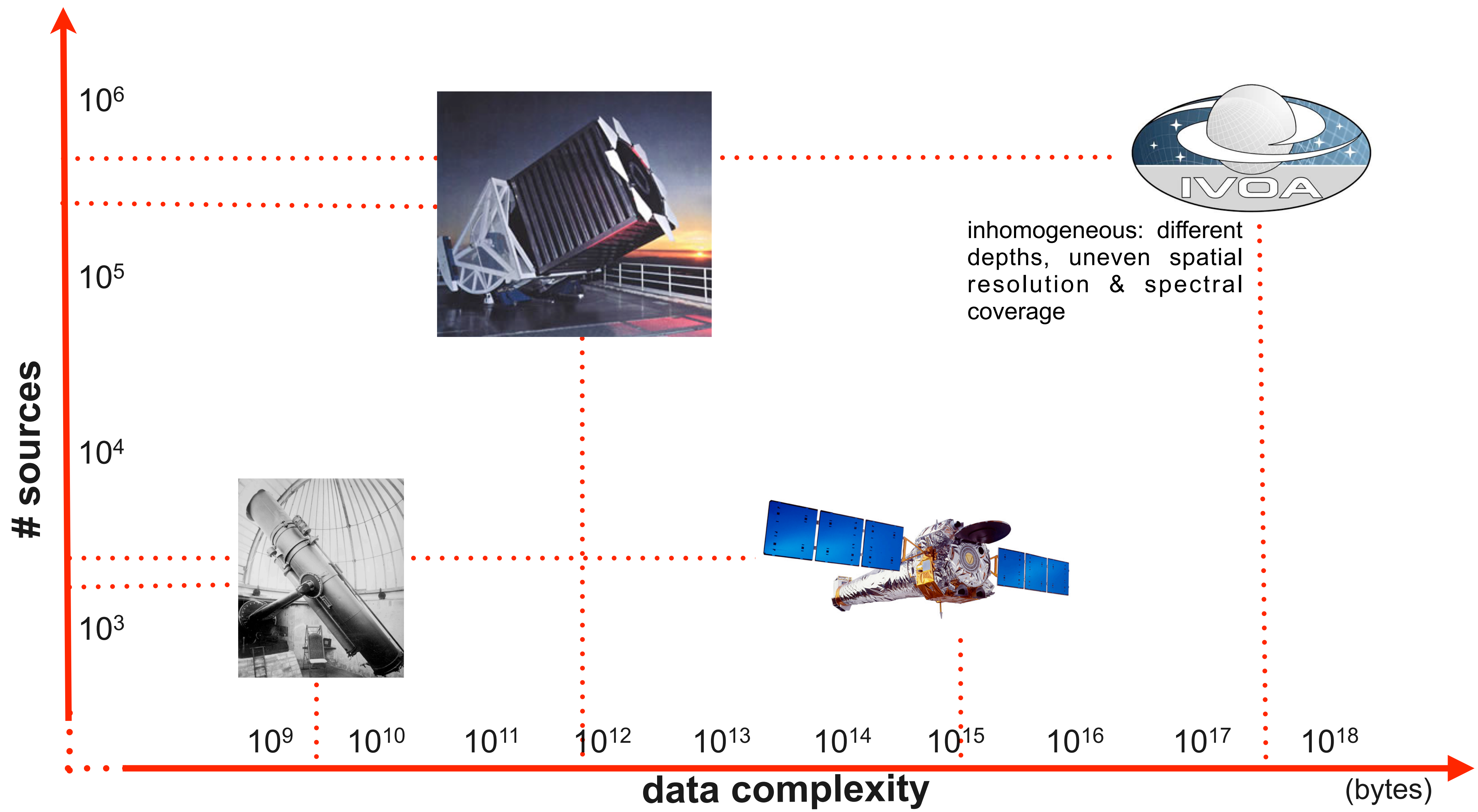
The characterization of the distribution of complex astronomical dataset in a high-dimensionality parameter space can reveal new patterns and correlations.



Most of the discoveries in astronomy have so far taken place in very low dimensional (2, 3 dimensions) projections of the observable space.

The data deluge (cit.)

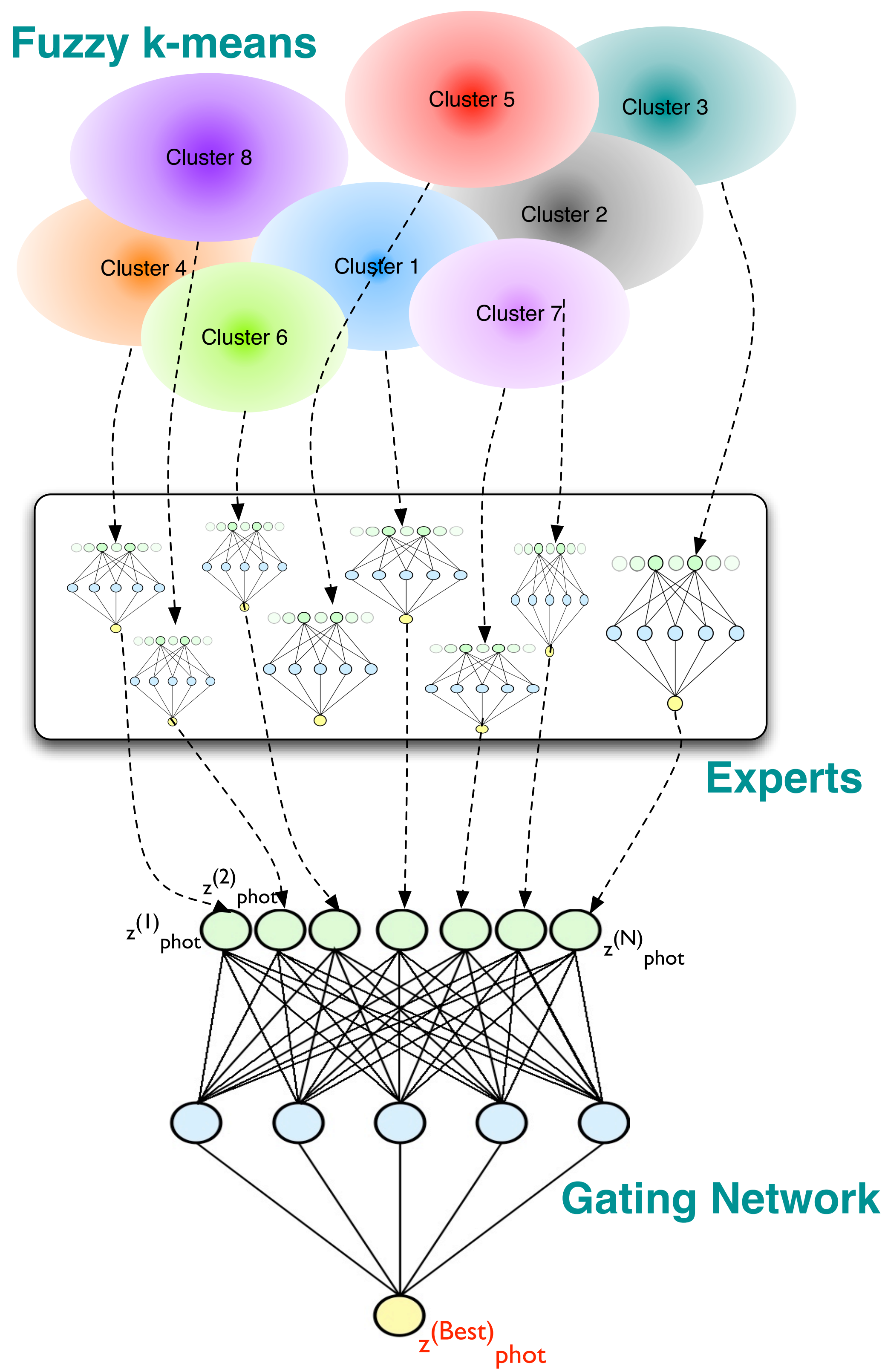
Knowledge Discovery (KD) techniques can tackle the challenge of the massive and/or complex astronomical datasets.



Not even all correlations in low-dimensionality *feature* spaces have been explored yet.

An example of KD workflow

Clustering of the optical-UV *feature* spaces of galaxies and quasars improved the accuracy of the z_{phot} reconstruction (Laurino et al. 2011, in pub. MNRAS)



A new method

Clustering-Labels-Scores Patterns Spotter (CLaSPS)

• Unsupervised Clustering (UC) algorithms used to produce groupings of the sources in the *feature* space associated to their observables;

• Additional observables (*labels*) are used to identify interesting clusterings:

- extract new patterns that could not be determined in low-D projections of the *feature* space;
- expand known correlations among *features* and/or *labels to* high-D spaces;
- spot unusual behaviors (e.g. outliers);

Statistical issues

Few points for a large space

>10 dimensional *features* space
populated by $10^2 \sim 10^3$ sources

Upper limits & Clustering

Inclusion of upper limits as *features* of
the distribution of sources

Clusters vs Outliers

Well populated, homogeneous clusters
oriented vs small clusters/singletons
(outliers).

Statistical issues

Few points for a large space

>10 dimensional *features* space populated by $10^2 \sim 10^3$ sources

→ **Low specific density, a.k.a. "curse of dimensionality"**

Upper limits & Clustering

Inclusion of upper limits as *features* of the distribution of sources

Clusters vs Outliers

Well populated, homogeneous clusters oriented vs small clusters/singletons (outliers).

Statistical issues

Few points for a large space

>10 dimensional *features* space
populated by $10^2 \sim 10^3$ sources

→ **Low specific density, a.k.a. "curse of dimensionality"**

Upper limits & Clustering

Inclusion of upper limits as *features* of the distribution of sources

→ **A general theory not available**

Clusters vs Outliers

Well populated, homogeneous clusters oriented vs small clusters/singletons (outliers).

Statistical issues

Few points for a large space

>10 dimensional *features* space populated by $10^2 \sim 10^3$ sources

→ **Low specific density, a.k.a. "curse of dimensionality"**

Upper limits & Clustering

Inclusion of upper limits as *features* of the distribution of sources

→ **A general theory not available**

Clusters vs Outliers

Well populated, homogeneous clusters oriented vs small clusters/singletons (outliers).

→ **Ensemble of UC methods**

UC methods

Dimensionality reduction	UC	Dimensionality expansion
	K-means	Principal Component Analysis (PCA)
	Hierarchical Clustering (HC)	
Self-Organizing Maps (SOM)		Support Vector Machine (SVM)
Principal Probabilistic Surfaces (PPS)		

complexity

UC methods

Dimensionality reduction

UC

Dimensionality expansion

K-means

Principal Component Analysis (PCA)

Hierarchical Clustering (HC)

Self-Organizing Maps (SOM)

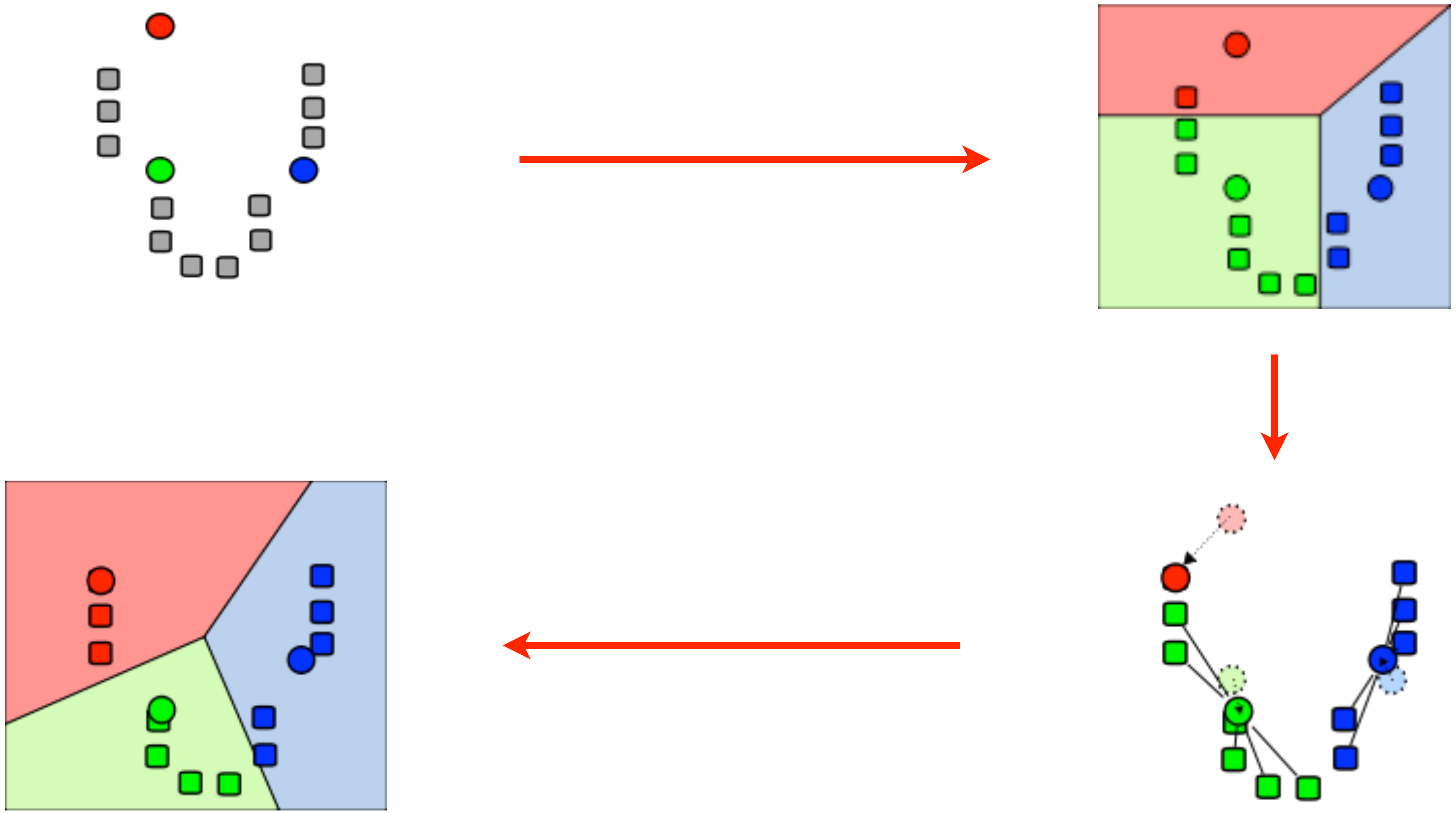
Support Vector Machine (SVM)

Principal Probabilistic Surfaces (PPS)

complexity

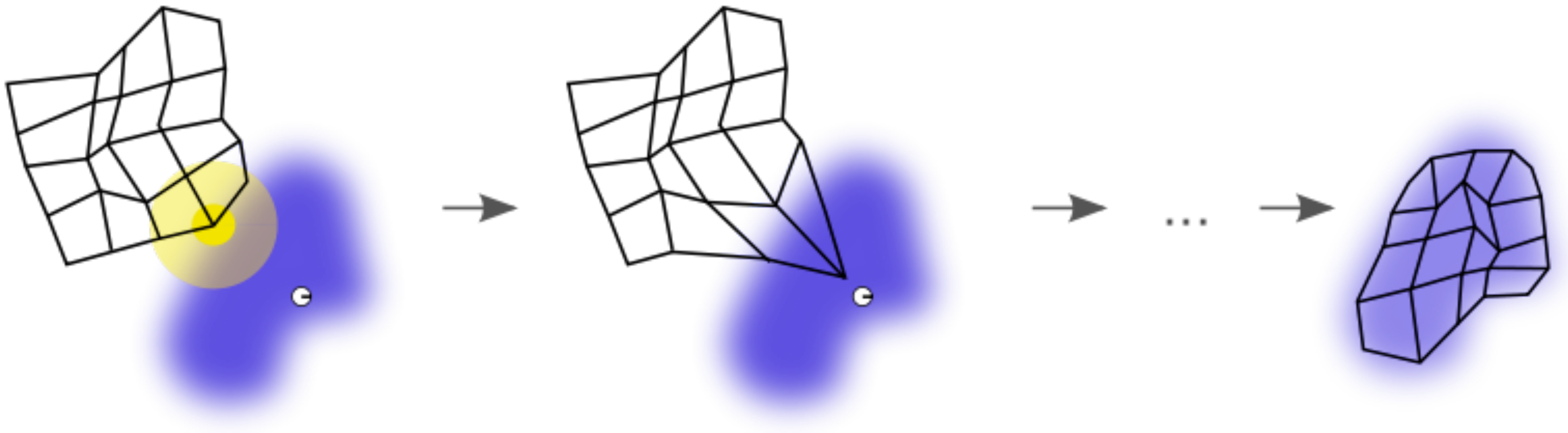
K-means

Iterative descent method, employs euclidean distance.
The number of clusters k is a parameter that needs to be specified.



SOM

Constrained version of the K-means -prototypes are encouraged to lie on a 2-d manifold, which is adjusted to the distribution of the “training” points in the high-dimensional *features* space.



SOM can work as an algorithm for UC and as a supervised classifier.

Hierarchical Clustering

Generalized K-means

HC does not require *k* to be fixed, as all clusterings with different values of *K* are produced, once assigned a *measure of dissimilarity*, based on pairwise dissimilarities between members of the clusters.

dissimilarity ≡ (metric, linkage strategy)

Hierarchical Clustering

Generalized K-means

HC does not require k to be fixed, as all clusterings with different values of K are produced, once assigned a *measure of dissimilarity*, based on pairwise dissimilarities between members of the clusters.

dissimilarity ≡ (metric, linkage strategy)

Metrics

Euclidean, Manhattan, Mahalanobis, maximum, ...

Linkage strategies

Single linkage $d(C_1, C_2) = \min(d_{ij}) \quad \{i \in C_1, j \in C_2\}$

Complete linkage $d(C_1, C_2) = \max(d_{ij}) \quad \{i \in C_1, j \in C_2\}$

Group linkage $d(C_1, C_2) = \frac{1}{N_{C_1} N_{C_2}} \sum_{i \in C_1} \sum_{j \in C_2} d_{ij} \quad \{i \in C_1, j \in C_2\}$

UC and visualization

Effective visualization techniques are required in order to *grok* the results of UC in high-dimensional space. Visualization techniques for multi-variate datasets are often used as exploratory techniques in KD.

K-means

Hierarchical Clustering (HC)

Self-Organizing Maps (SOM)

complexity
of the UC method

UC and visualization

Effective visualization techniques are required in order to *grok* the results of UC in high-dimensional space. Visualization techniques for multi-variate datasets are often used as exploratory techniques in KD.

K-means

Scatterplots
Boxplots
Violin plots
Parallel Coordinates plot

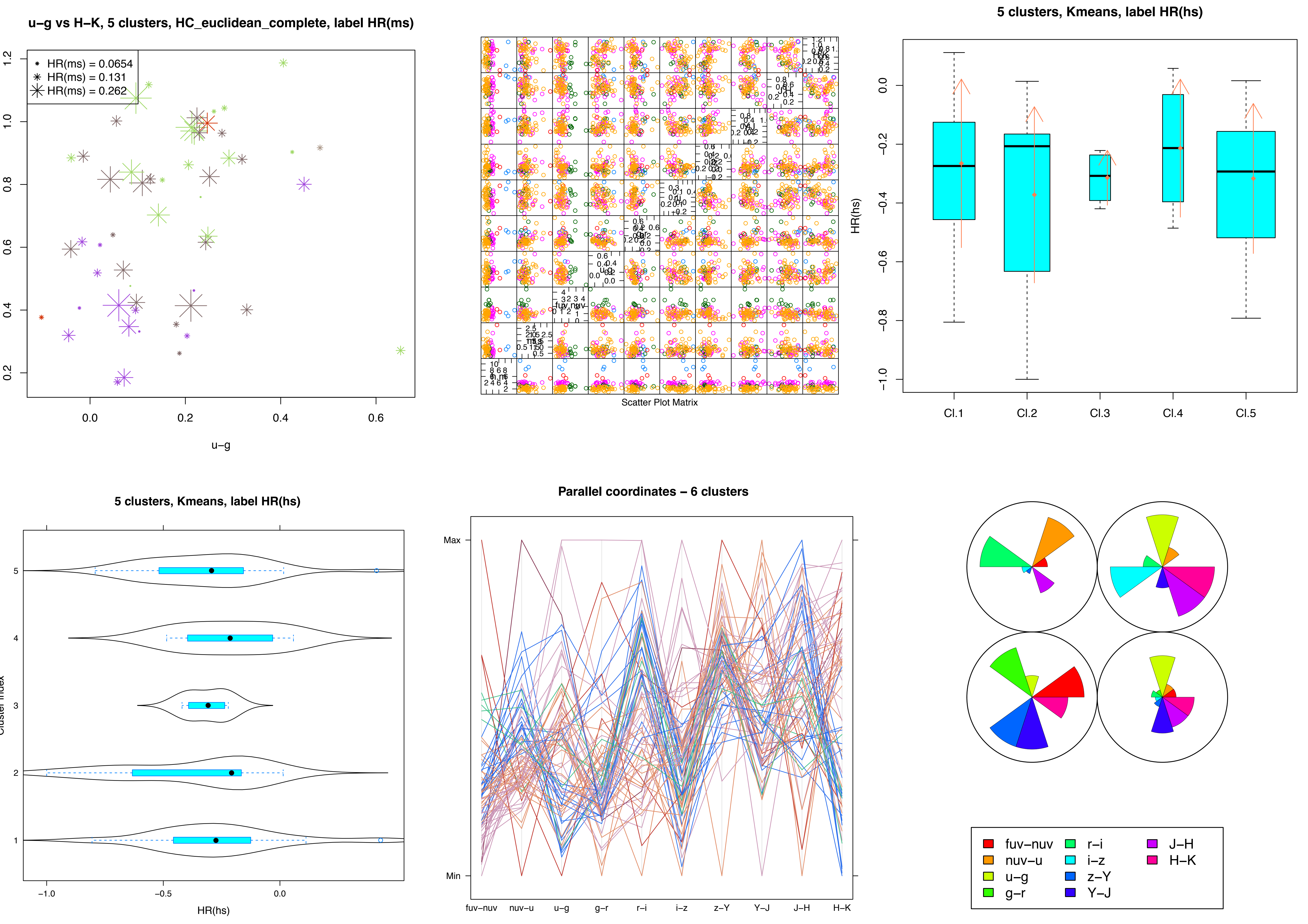
Hierarchical Clustering (HC)

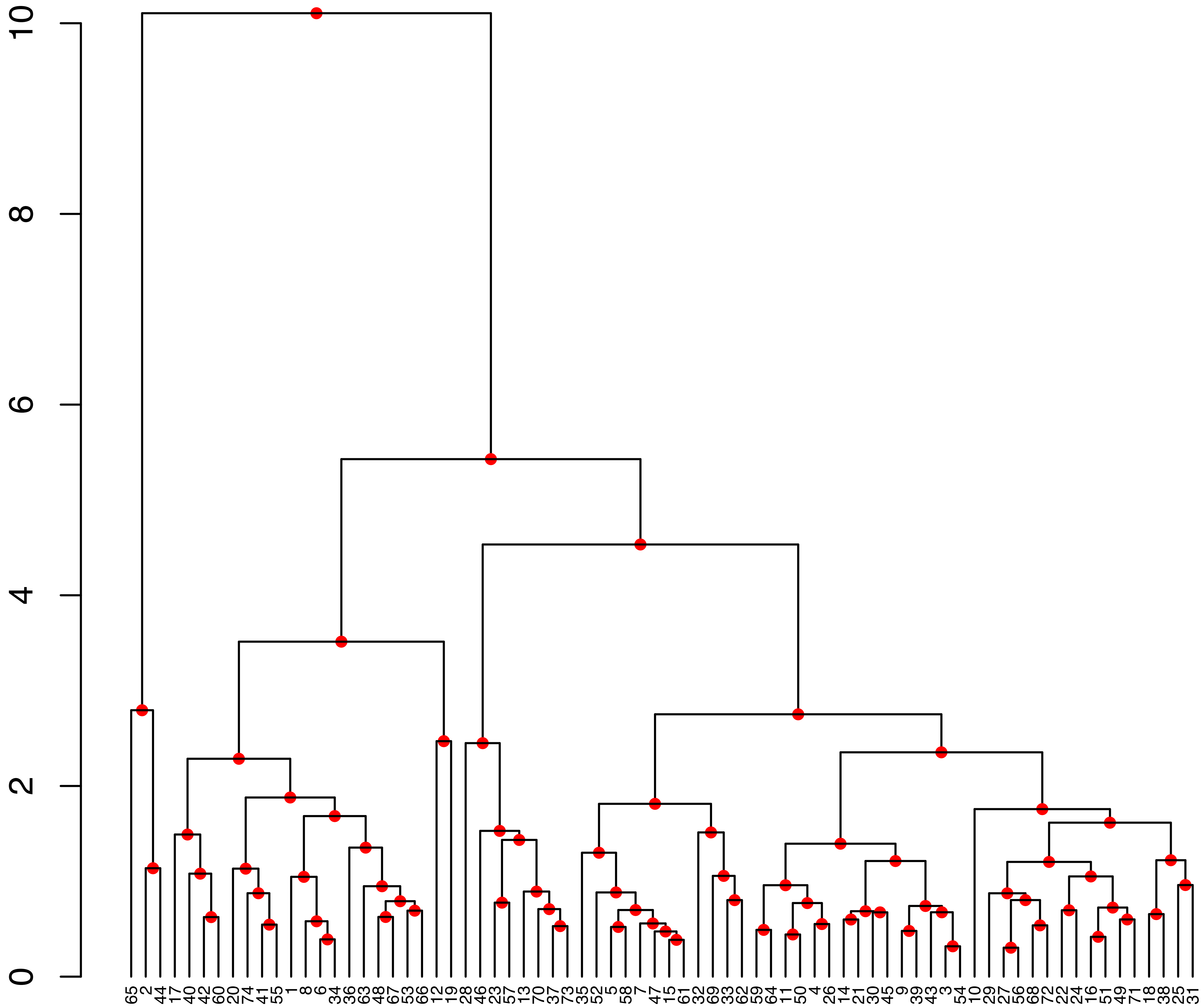
"Dendrograms"
"Heatmaps"
Linear plots

Self-Organizing Maps (SOM)

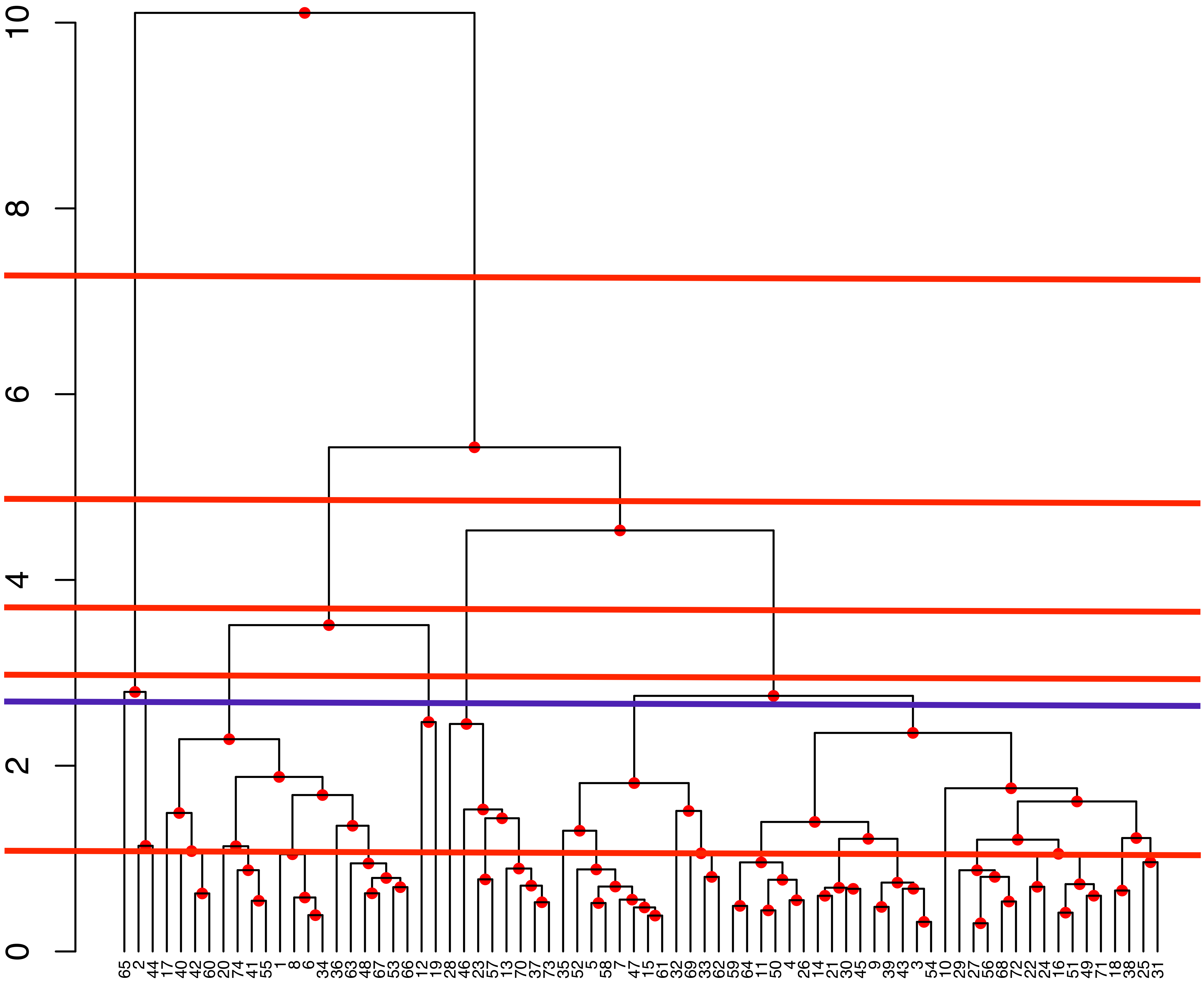
"Heatmaps"
"Codebook" plots

complexity
of the UC method


u–g vs H–K, 5 clusters, HC_euclidean_complete, label HR(ms)
HR(ms) = 0.0654
HR(ms) = 0.131
HR(ms) = 0.262
u–g
Scatter Plot Matrix
5 clusters, Kmeans, label HR(hs)
HR(hs)
Cl.1
Cl.2
Cl.3
Cl.4
Cl.5
5 clusters, Kmeans, label HR(hs)
Cluster index
HR(hs)
Parallel coordinates – 6 clusters
Max
Min
fuv–nuv
nuv–u
u–g
g–r
r–i
i–z
z–Y
Y–J
J–H
H–K
fuv–nuv
nuv–u
u–g
g–r
r–i
i–z
z–Y
Y–J
J–H
H–K

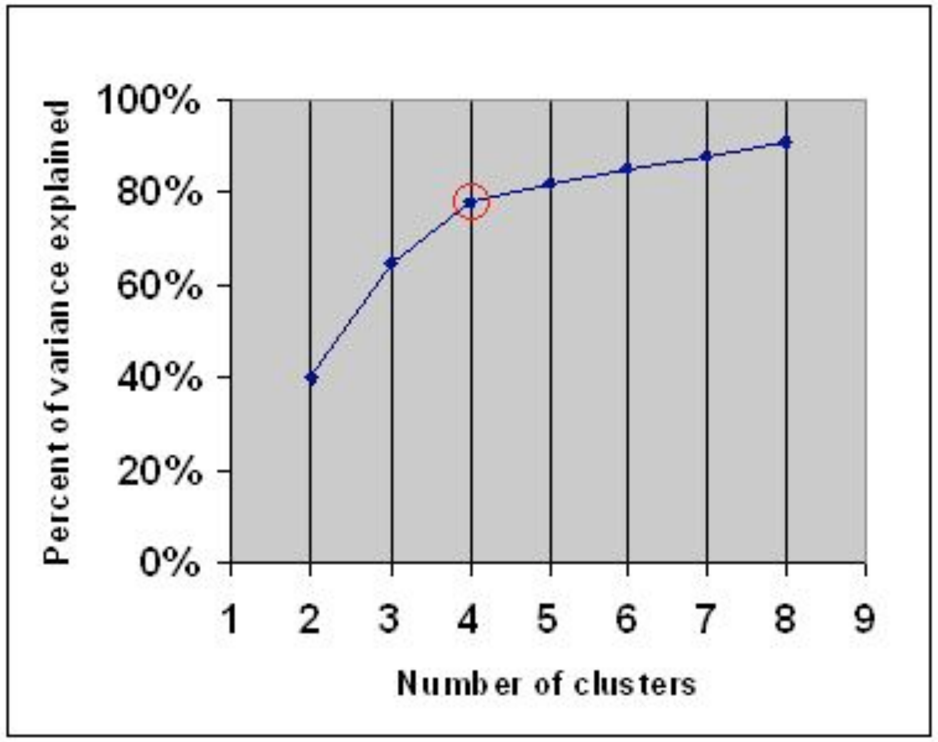


0
2
4
6
8
10
65 2 44 17 40 42 60 20 74 41 55 1 8 6 34 36 63 48 67 53 66 12 19 28 46 23 57 13 70 37 73 35 52 5 58 7 47 15 61 32 69 33 62 59 64 11 50 4 26 14 21 30 45 9 39 43 3 54 10 29 27 56 68 72 22 24 16 51 49 71 18 38 25 31

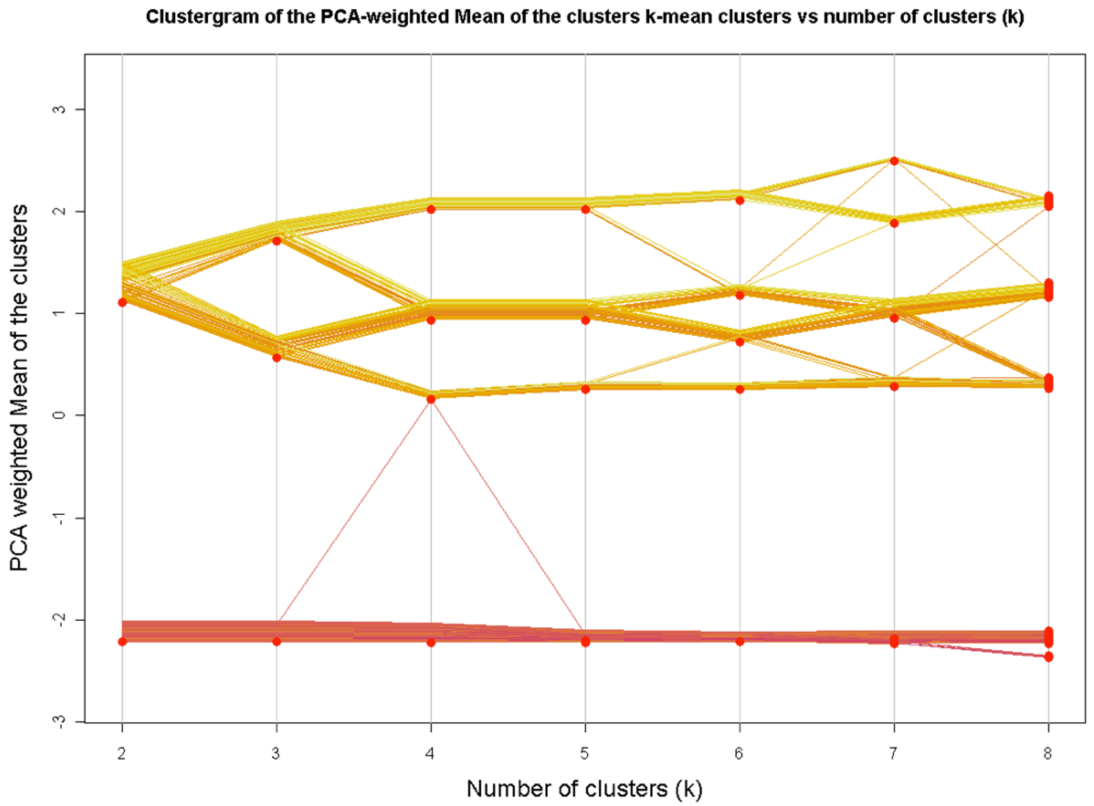


10
8
6
4
2
0
65 2 44 17 40 42 60 20 74 41 55 1 8 6 34 36 63 48 67 53 66 12 19 28 46 23 57 13 70 37 73 35 52 5 58 7 47 15 61 32 69 33 62 59 64 11 50 4 26 14 21 30 45 9 39 43 3 54 10 29 27 56 68 72 22 24 16 51 49 71 18 38 25 31

Number of clusters



$$k \approx \sqrt{n/2}$$



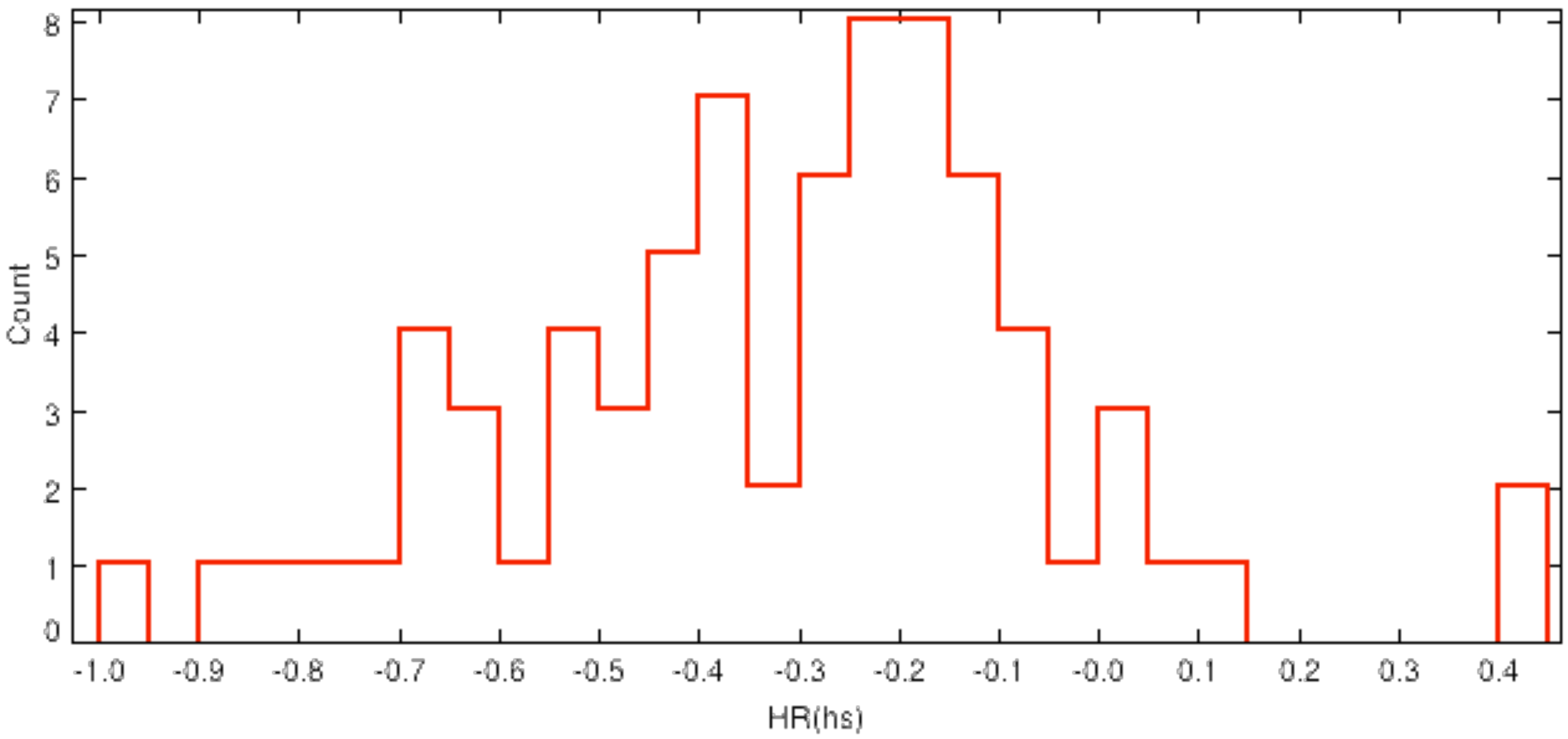
Choice of the "optimal" *k* can be not based on the statistical characteristics of the *features* used for the clusterings.

A different approach exploits the availability of external information (*labels*) to characterize the content of the clusters for each value of *k*, and evaluate a "figure of merit" based on the *labels* distribution.

Labels

Some observed quantities, called *labels* (either continuous or categorial) are used to pick those clusterings whose clusters are most correlated with the *label*(s), i.e. the clusterings where sources labeled with different values of the label are most separated.

L, f, colors, n_h, time variability indices, morphology, classification flags, etc.



Binning of the *label* values, i.d. the determination of *label classes* is crucial for the selection of the clusterings

The scores

Diagnostics that express the level of correlation between clusterings membership and one *label* class distribution.

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

$$S'_{\mathrm{tot}} = \frac{1}{N_{\mathrm{clust}}} \frac{\sum_{i=1}^{N_{\mathrm{clust}}} N_i \cdot S_i}{\sum_{i=1}^{N_{\mathrm{clust}}} N_i} = \frac{1}{N_{\mathrm{clust}}} \frac{\sum_{i=1}^{N_{\mathrm{clust}}} N_i \cdot S_i}{N_{\mathrm{tot}}}$$

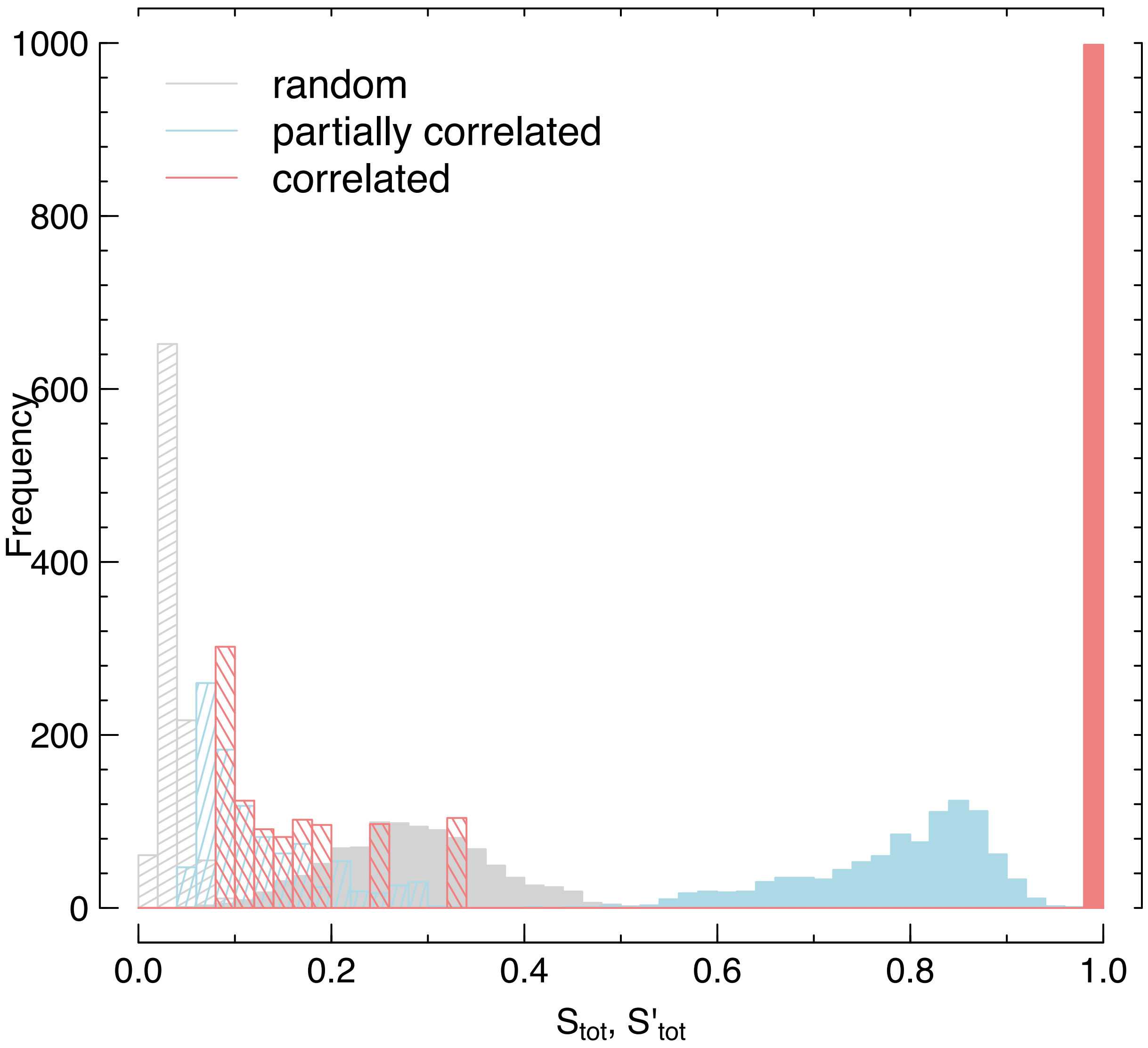
[**optimal clustering**] $\xleftrightarrow{\text{Def}}$ [**largest value of the *scores***]

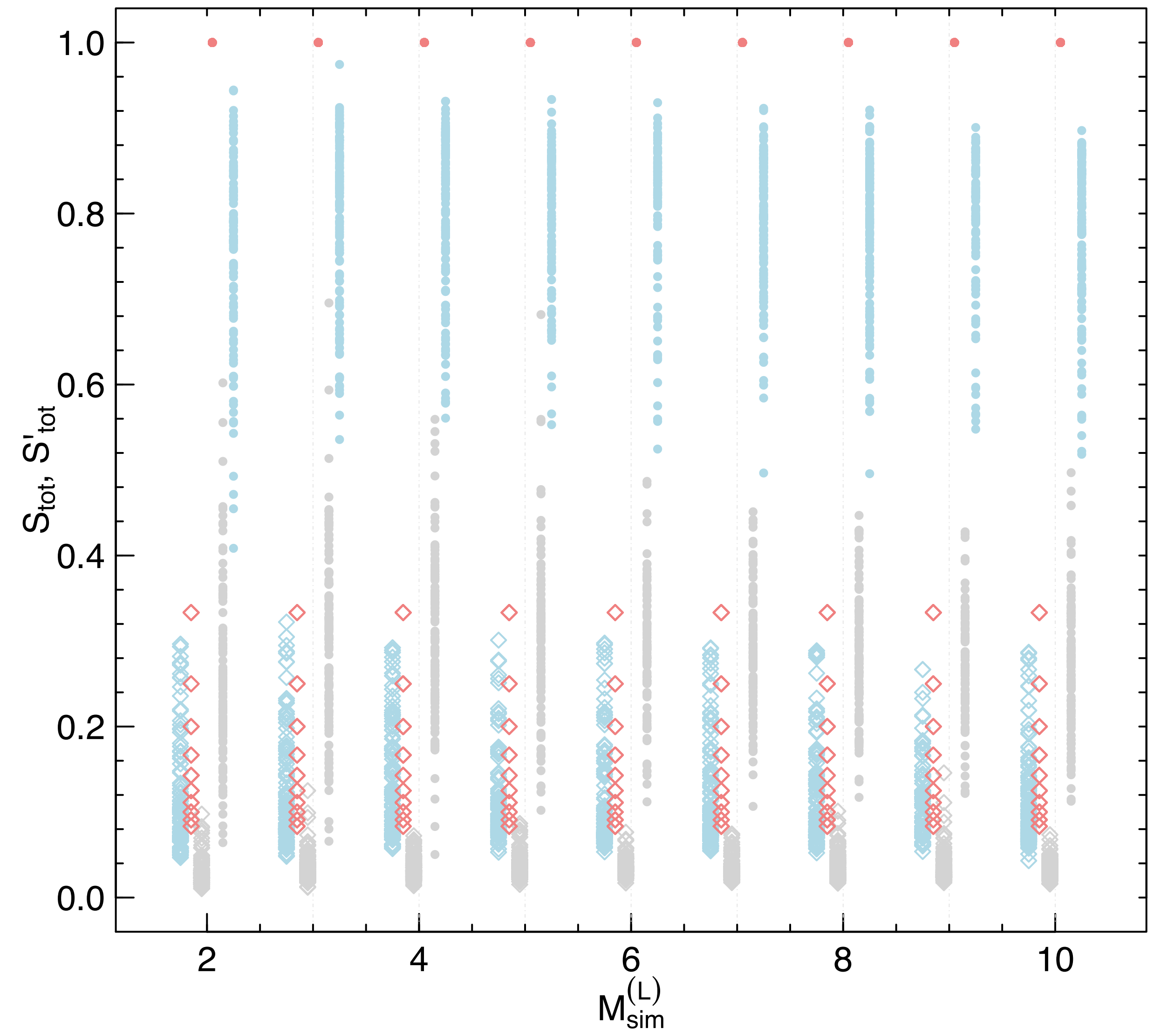
The "optimal" clustering is, by choice, the cluster whose *scores* values are the largest, since for these clusterings the degree of correlation between cluster membership and *labels* values is maximum.

The score can be evaluated for both continuous and categorial values.

Validation of the score

The score has been validated through simulated clusterings with varying degree of correlation with labels, number label classes, number of clusters and total number of observations





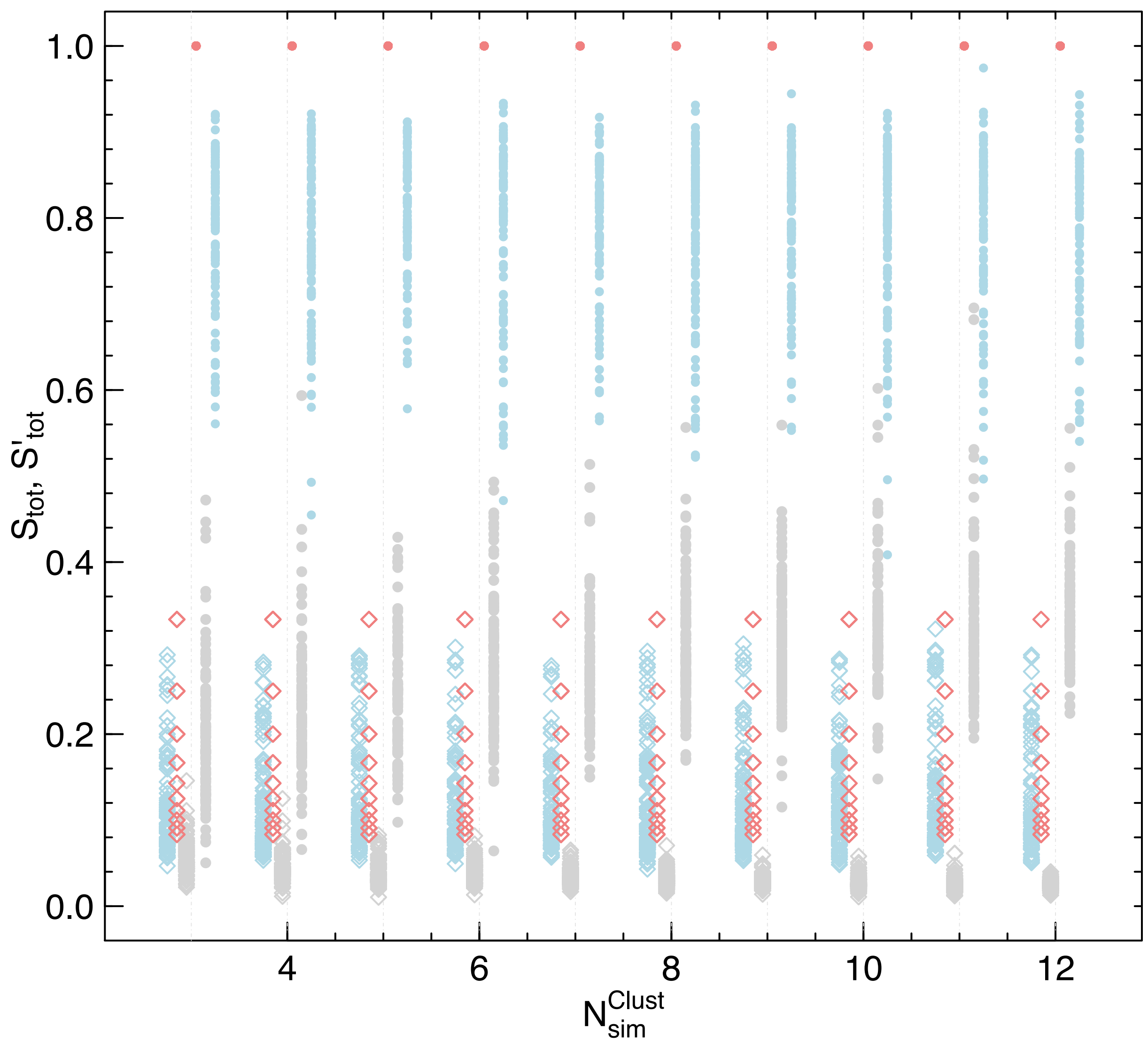
Random — from 100% to 80% random, remainder lin. assigned to clusters

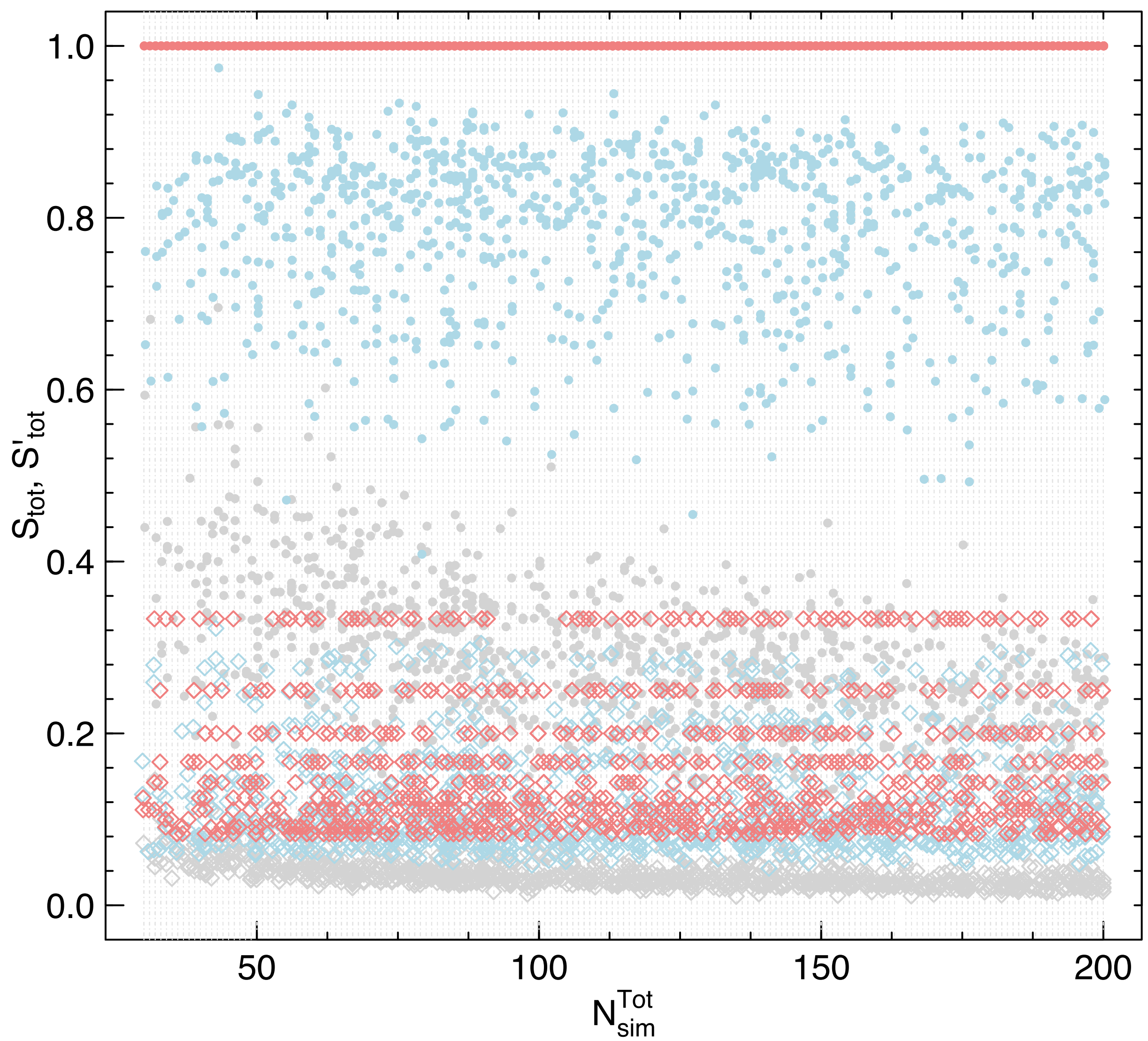
Partially correlated — from 20% to 50% random, remainder lin. assigned to clusters

Correlated — from 20% to 0% random, remainder lin. assigned to clusters

Validation of the score

The score has been validated through simulated clusterings with varying degree of correlation with labels, number label classes, number of clusters and total number of observations





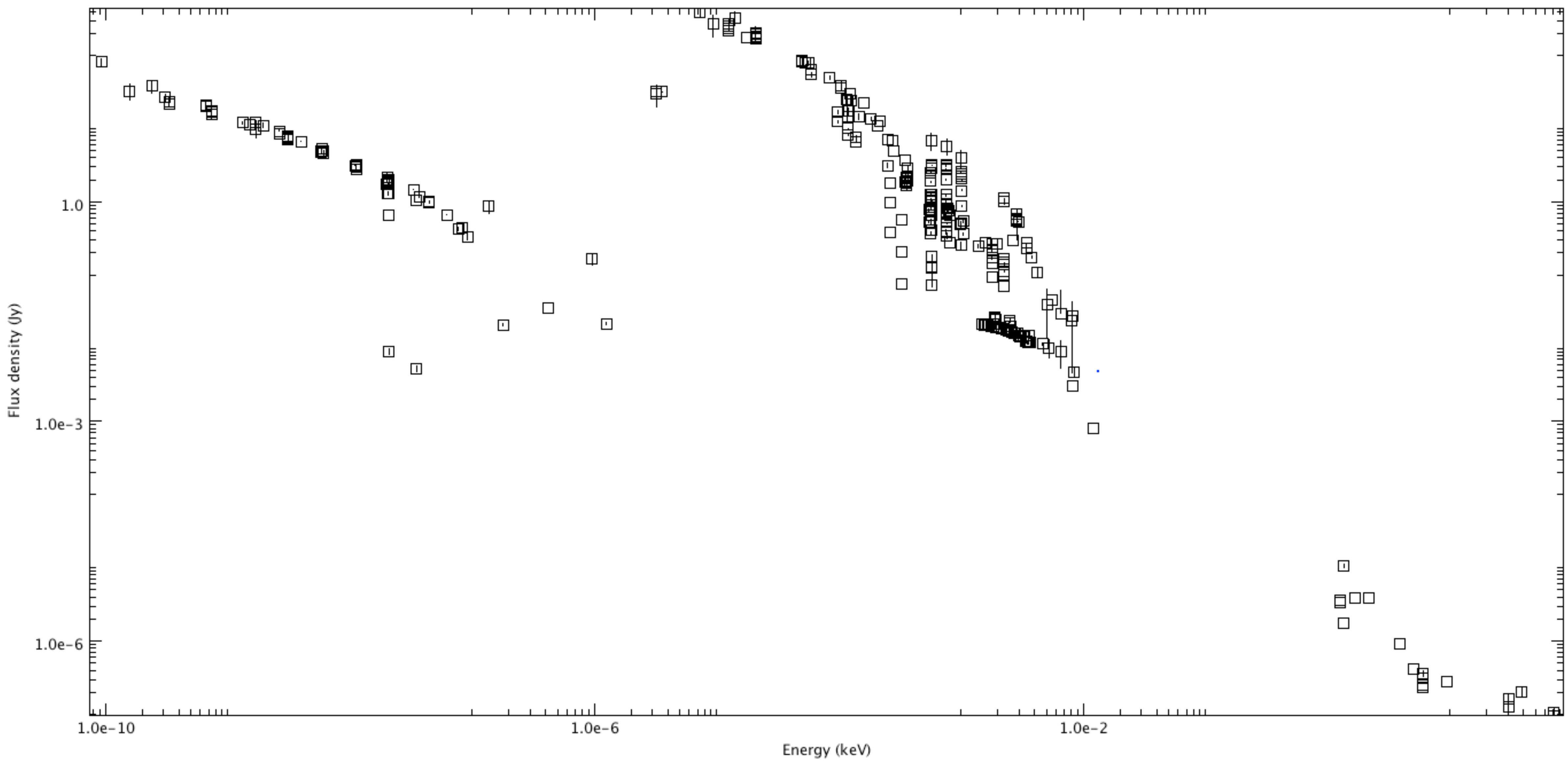
Random — from 100% to 80% random, remainder lin. assigned to clusters

Partially correlated — from 20% to 50% random, remainder lin. assigned to clusters

Correlated — from 20% to 0% random, remainder lin. assigned to clusters

One project

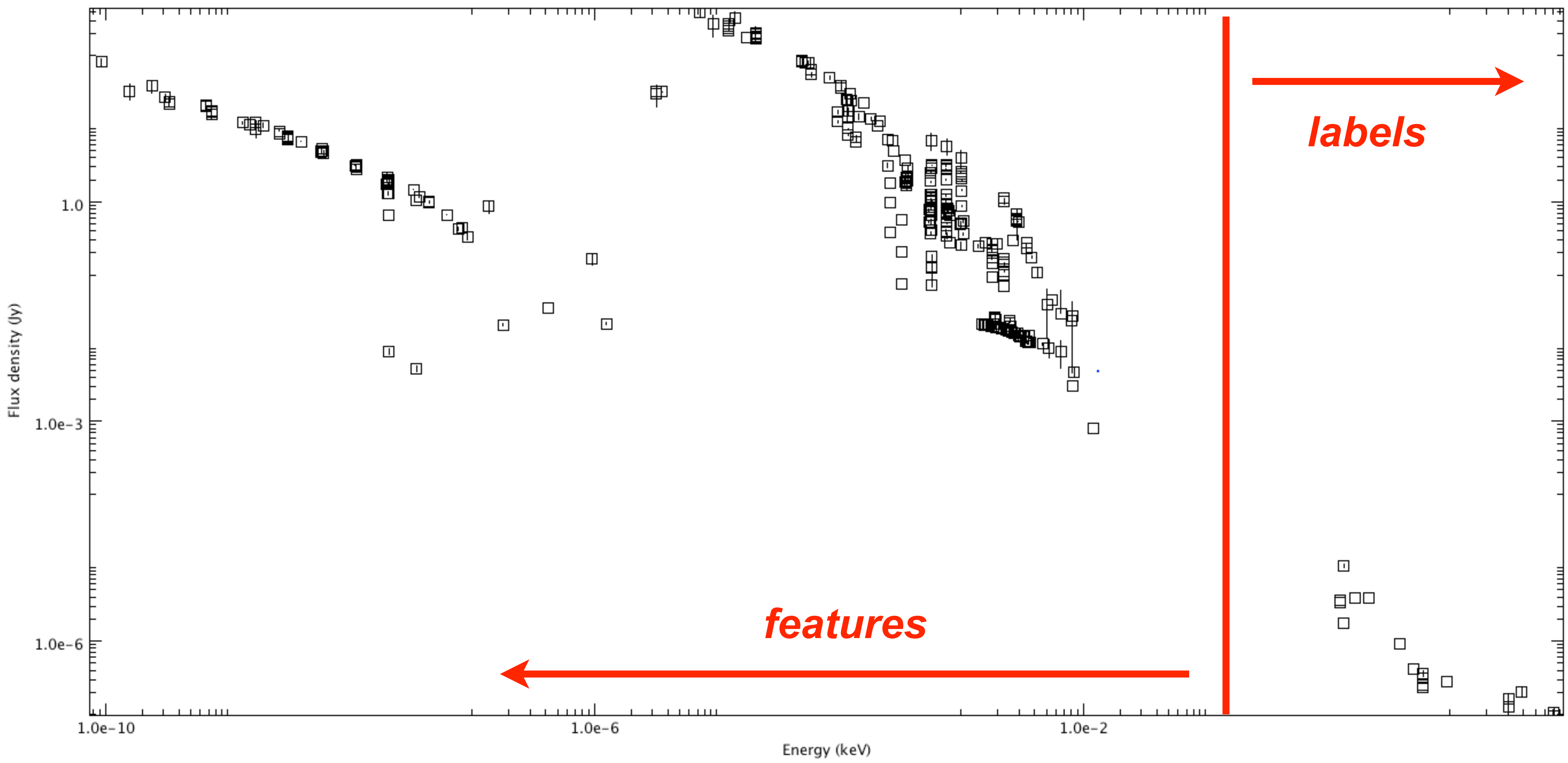
Characterization of the distribution of optically selected AGNs in the multi-wavelength photometric *features* space, using their X-ray properties as labels.



The primary purpose is to obtain a possible census of AGN behavior in the 13−dimensional *features* space of X-UV-optical-IR-Radio photometry and to constrain their X-ray properties with their other photometric observables, and select outliers (if any).

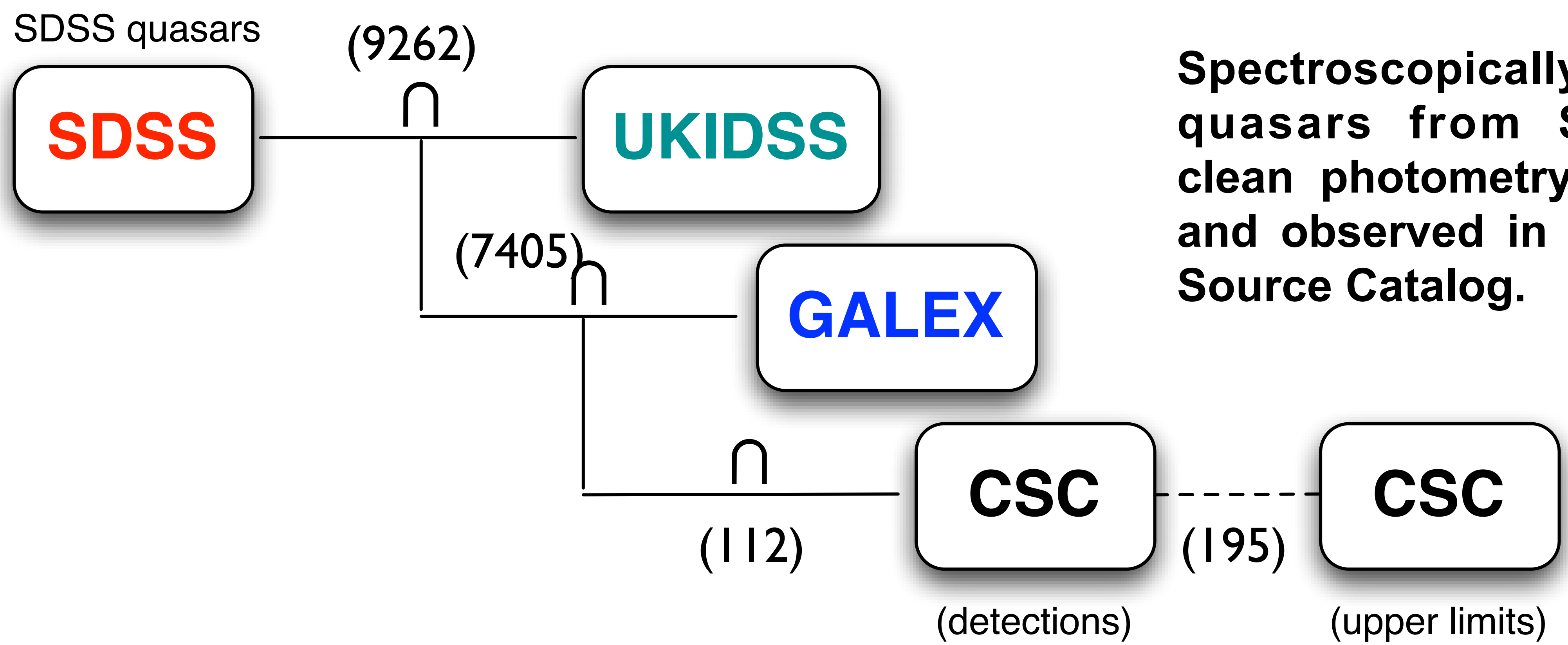
One project

Characterization of the distribution of optically selected AGNs in the multi-wavelength photometric *features* space, using their X-ray properties as labels.



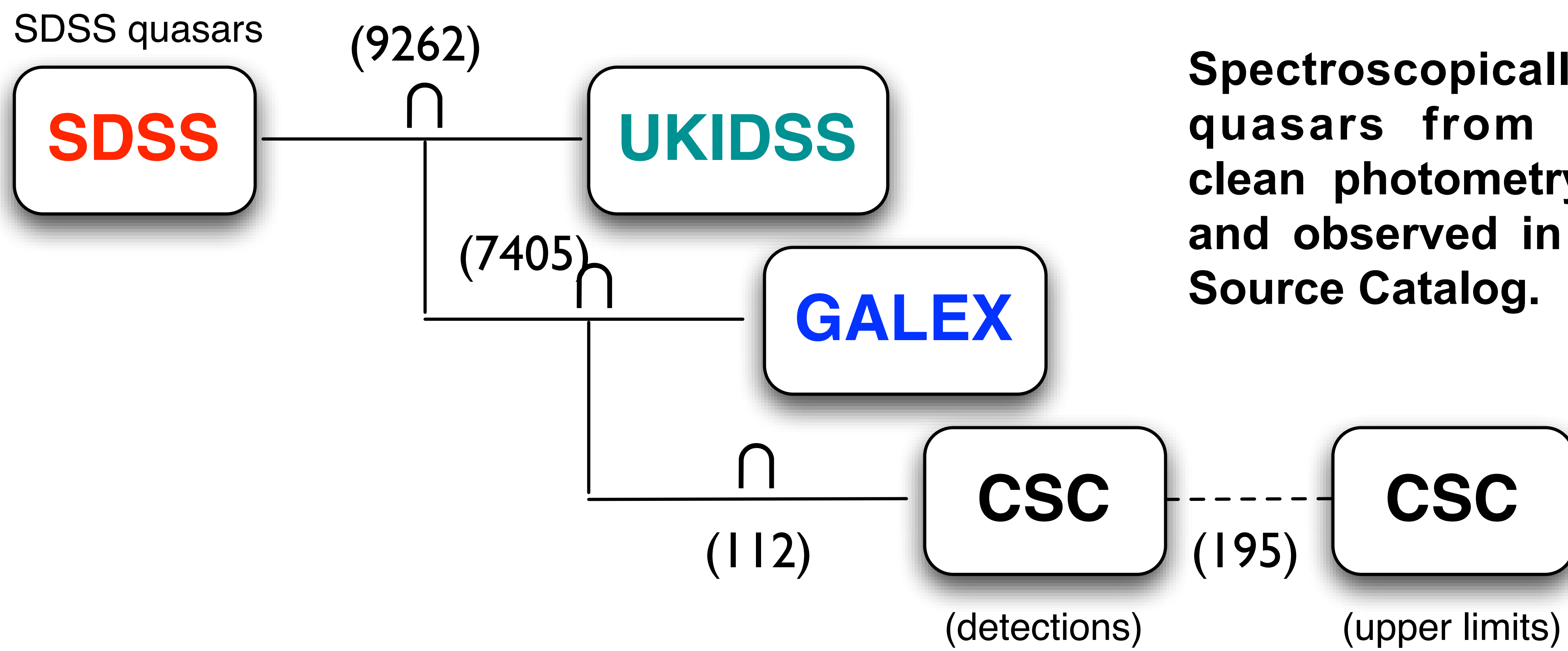
The primary purpose is to obtain a possible census of AGN behavior in the 13−dimensional *features* space of X-UV-optical-IR-Radio photometry and to constrain their X-ray properties with their other photometric observables, and select outliers (if any).

CSC+

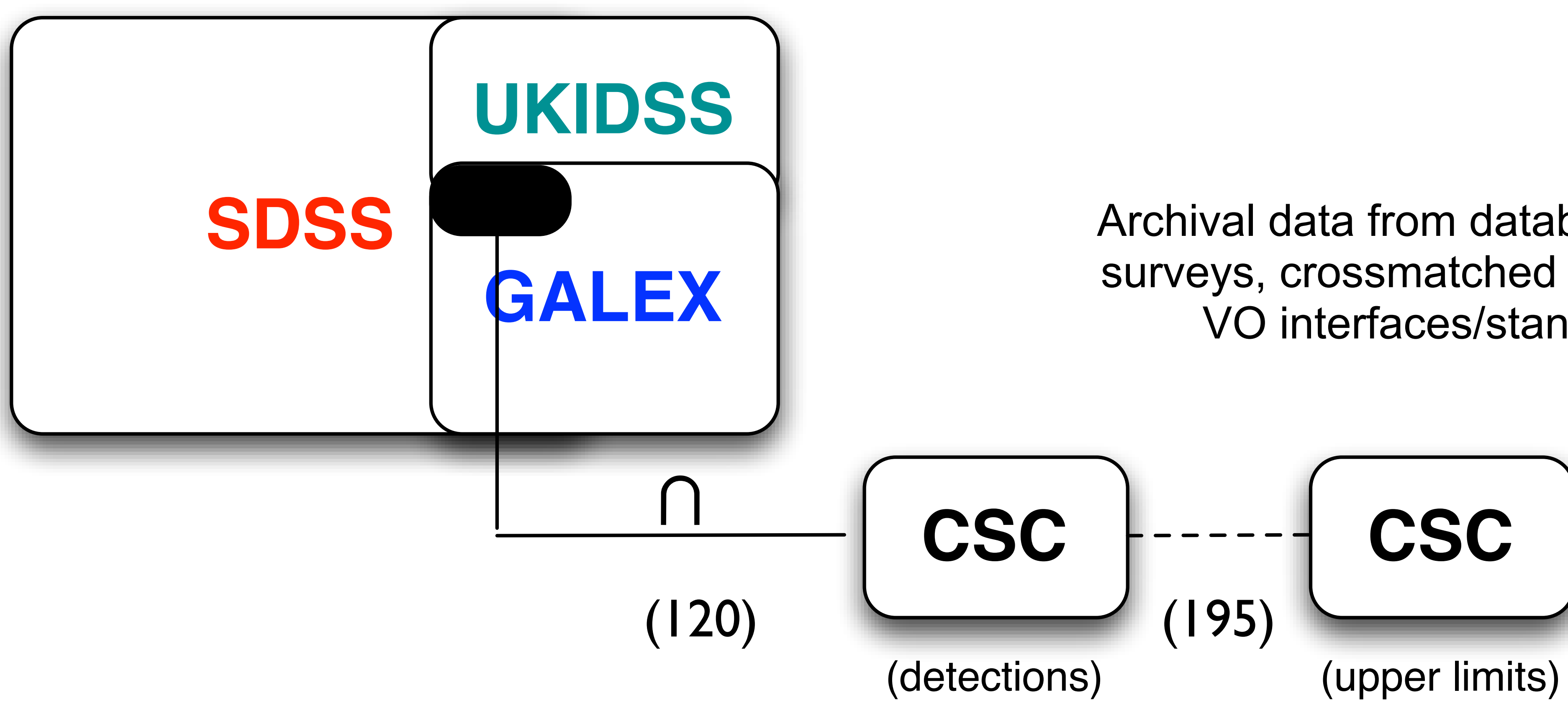


Spectroscopically confirmed quasars from SDSS, with clean photometry in NIR, UV and observed in the Chandra Source Catalog.

CSC+

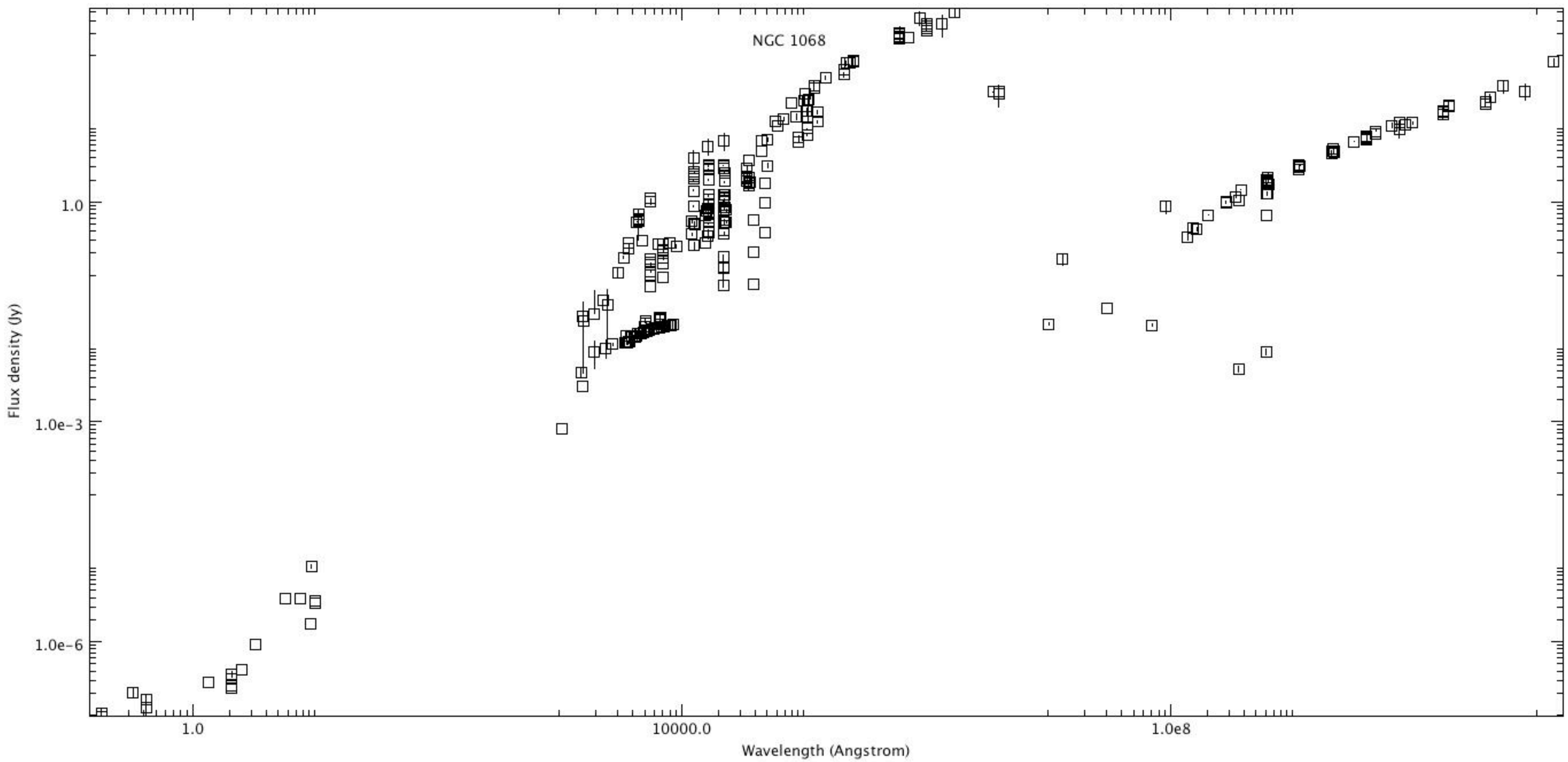


Spectroscopically confirmed quasars from SDSS, with clean photometry in NIR, UV and observed in the Chandra Source Catalog.



Archival data from databases of the surveys, crossmatched catalogs by VO interfaces/standards.

Features and labels for CSC+



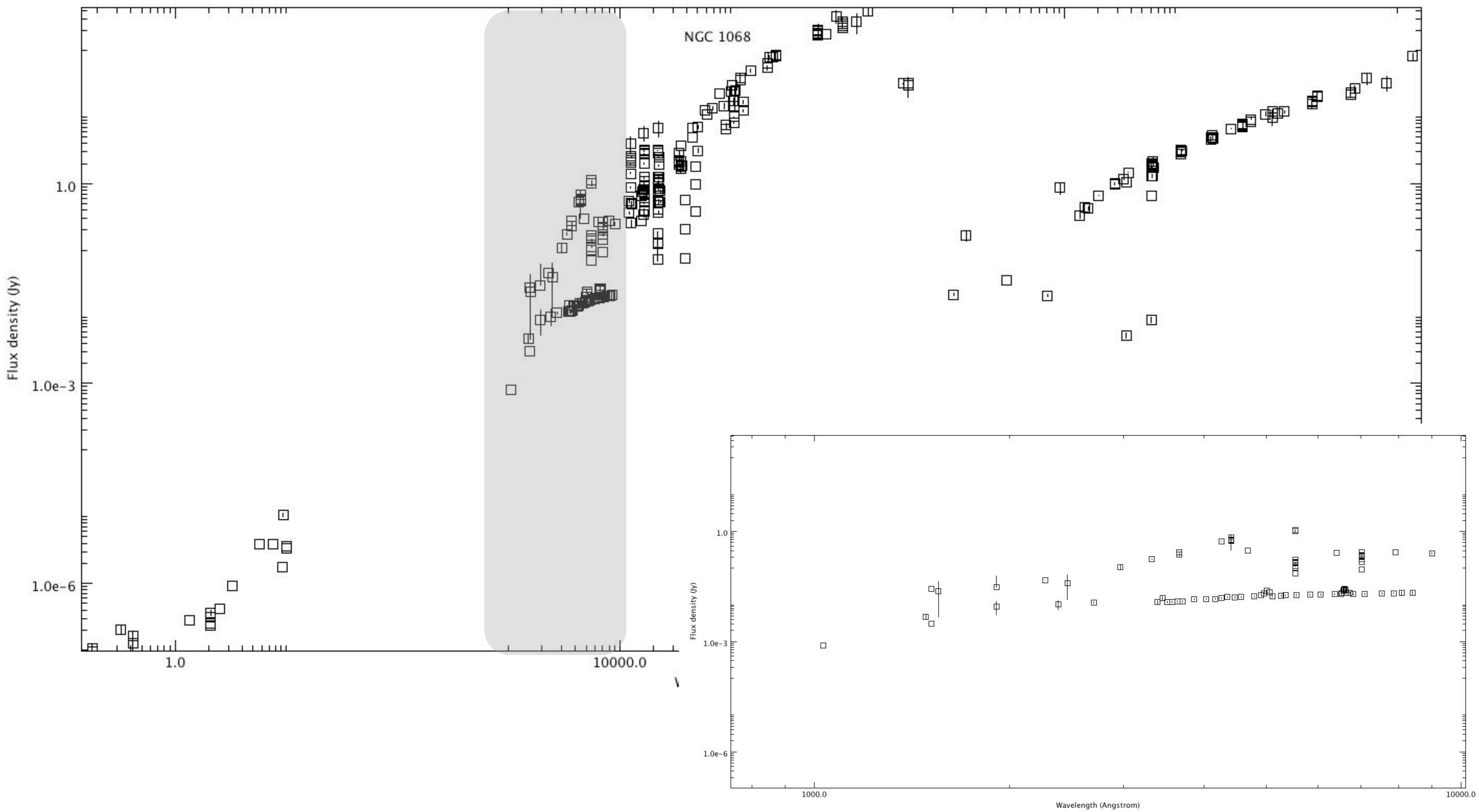
Features

***{fuv-nuv, nuv-u, u-g, g-r,
r-i, i-i, i-Y, Y-J, J-H, H-K, radio}***

Labels

{L_B, HR_{HS}, HR_{MS}, z}

Spectral coverage CSC+



Features

{fuv-nuv, nuv-u, u-g, g-r, r-i, i*-i, i-*Y, Y-J, J-H, H-K}

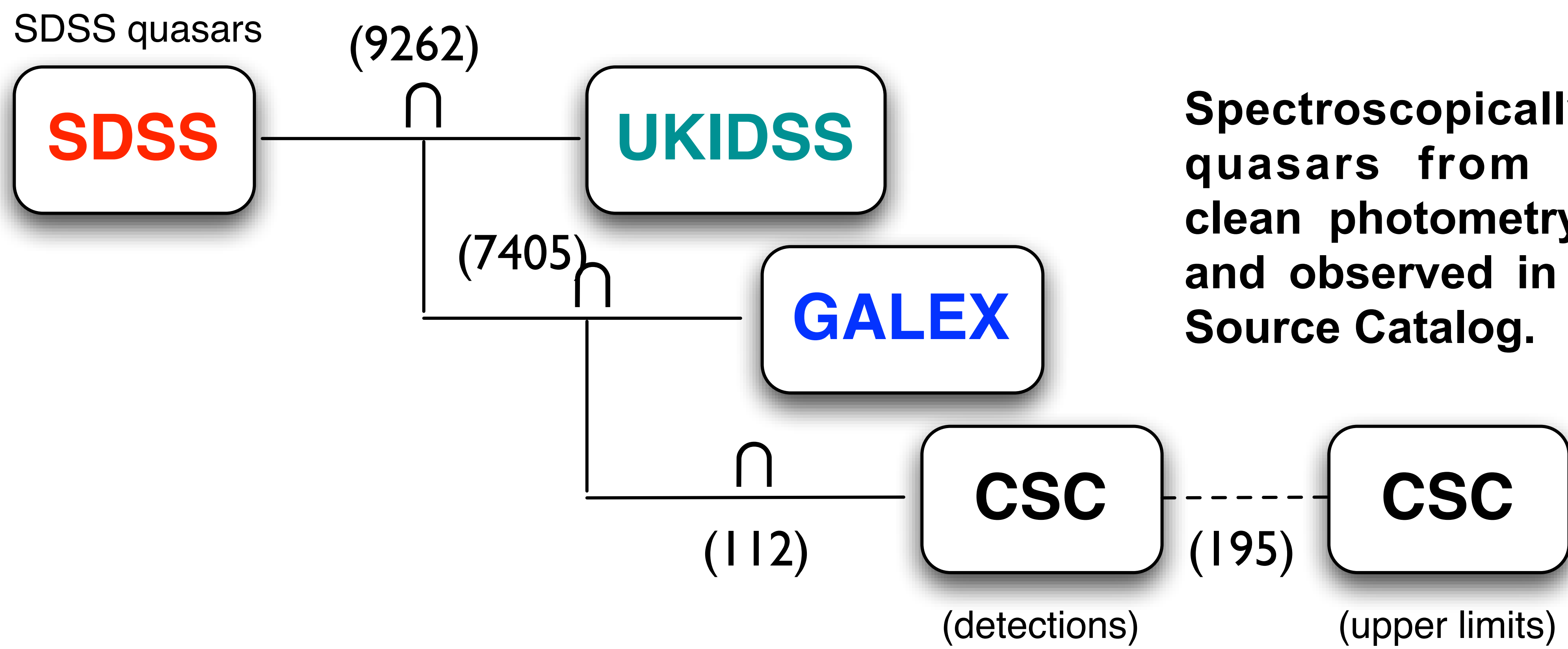
No radio data in VLA-First/NVSS.
No IR in Spitzer, SWIRE.

Labels

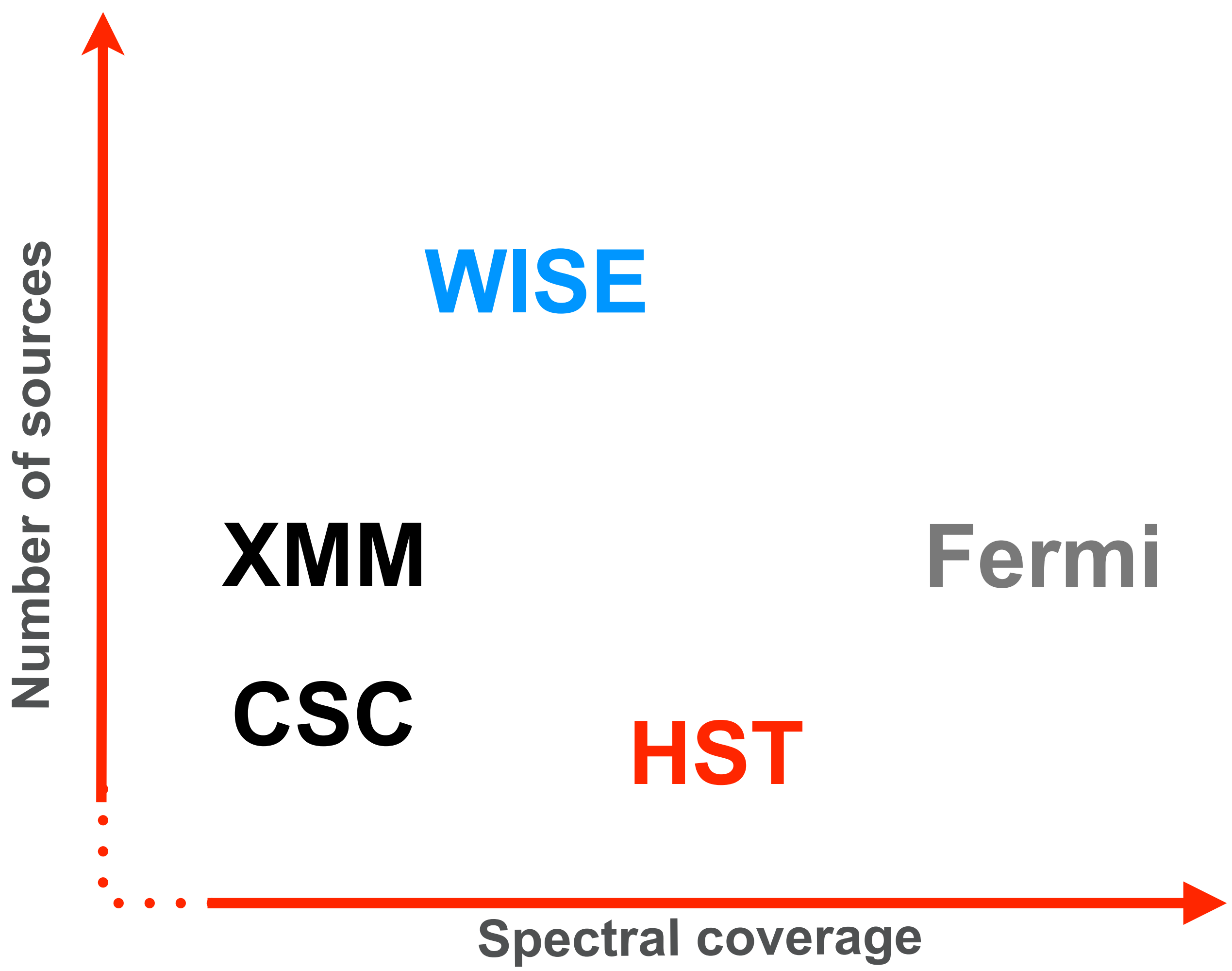
{L_B, HR_{HS}, HR_{MS}, z}

More *labels* to come:
L_H, Γ, α_{ox}, X-ray variability,...

CSC+ small sample

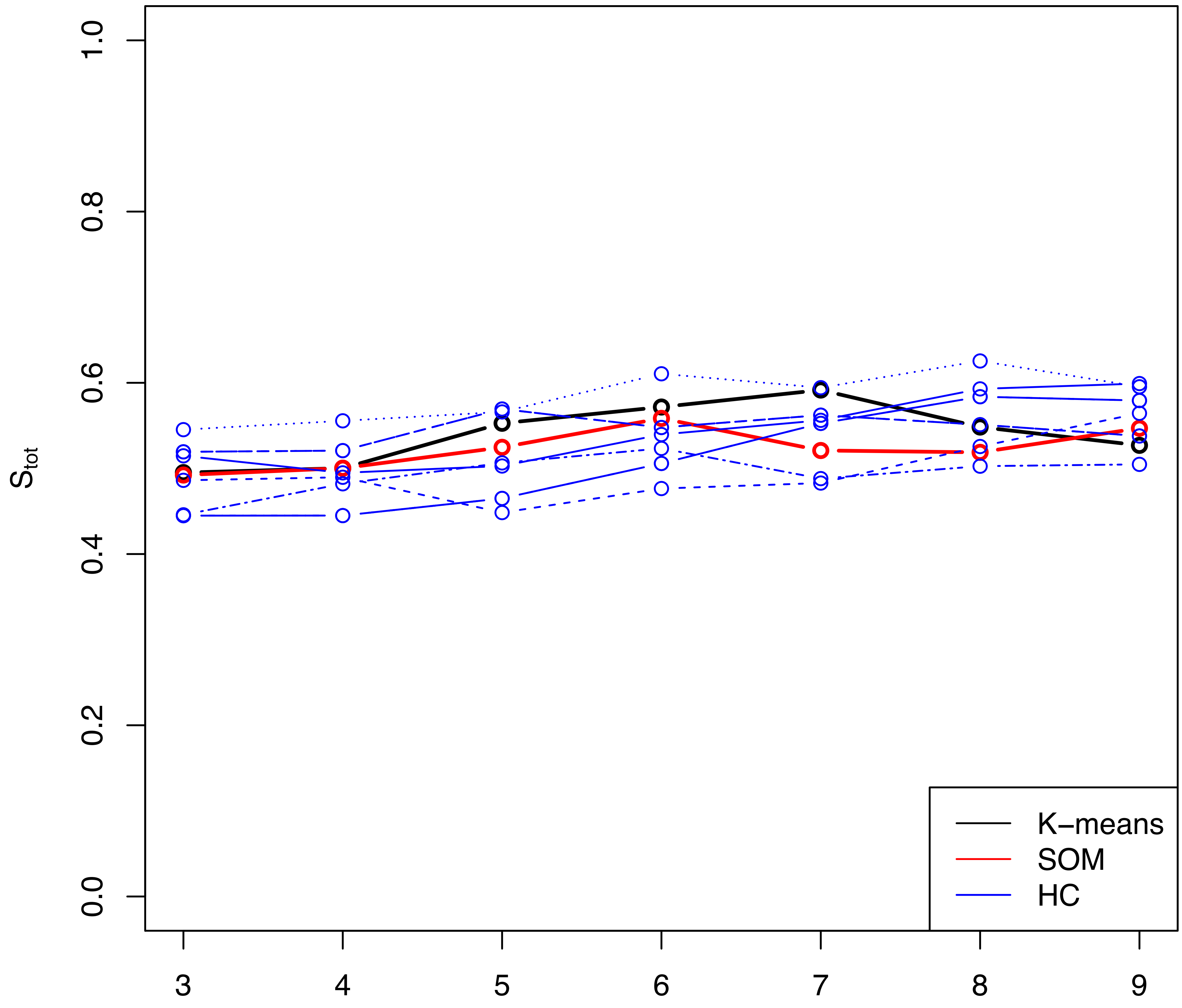


Spectroscopically confirmed quasars from SDSS, with clean photometry in NIR, UV and observed in the Chandra Source Catalog.

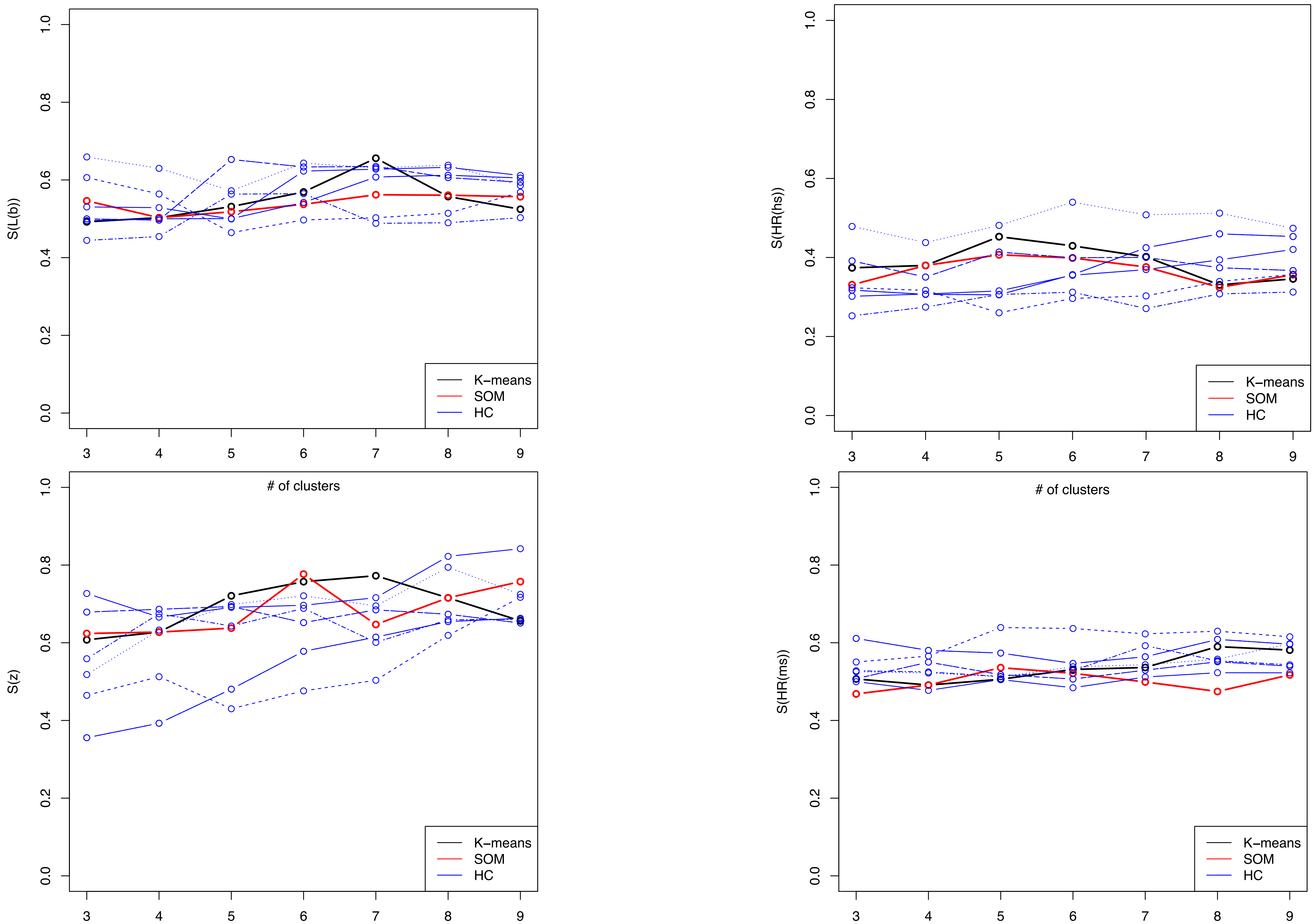


Other observations can be used to improve the spectral coverage and obtain a larger dataset.

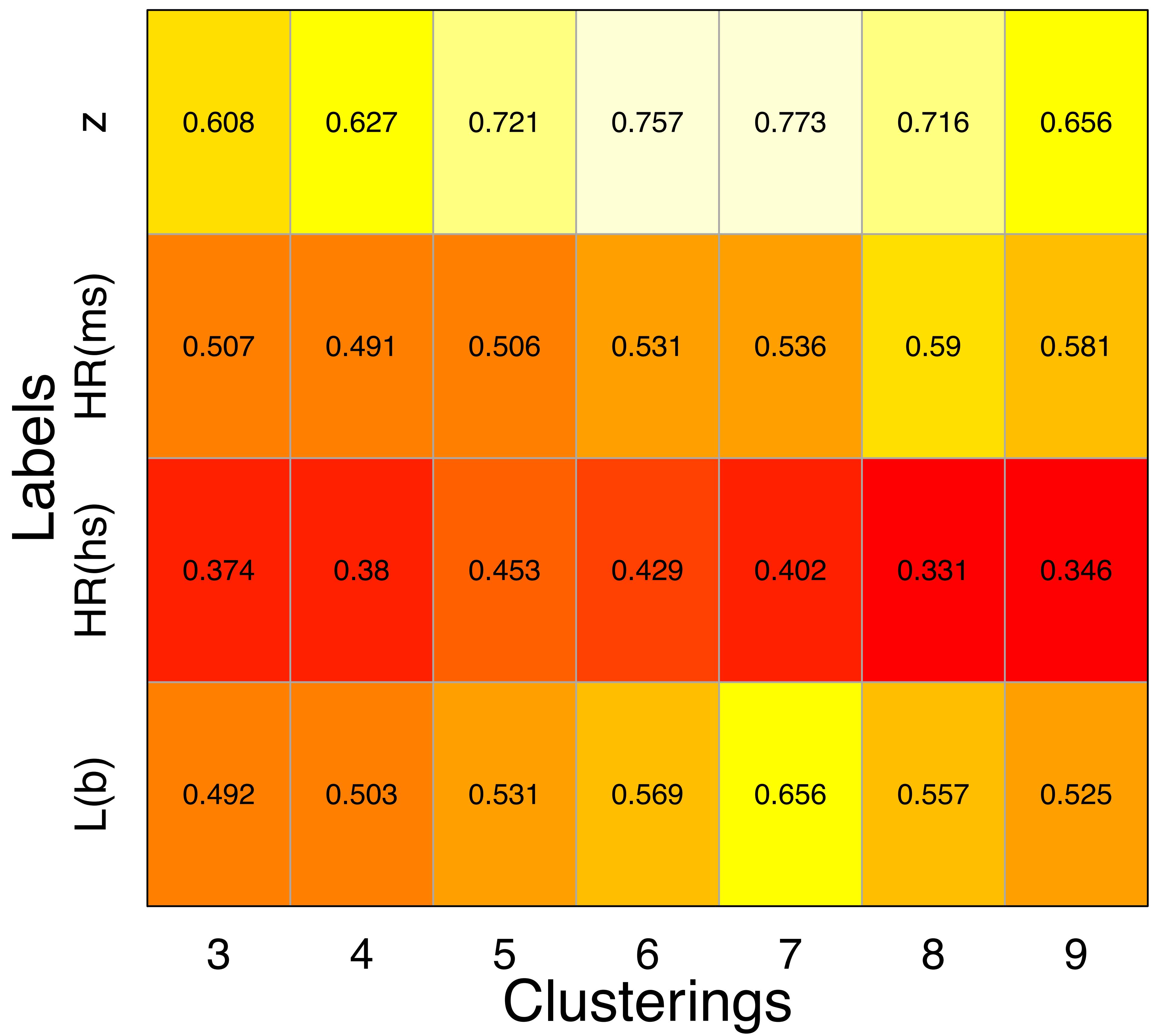
Selecting the UC method



Single labels



Selecting clusterings



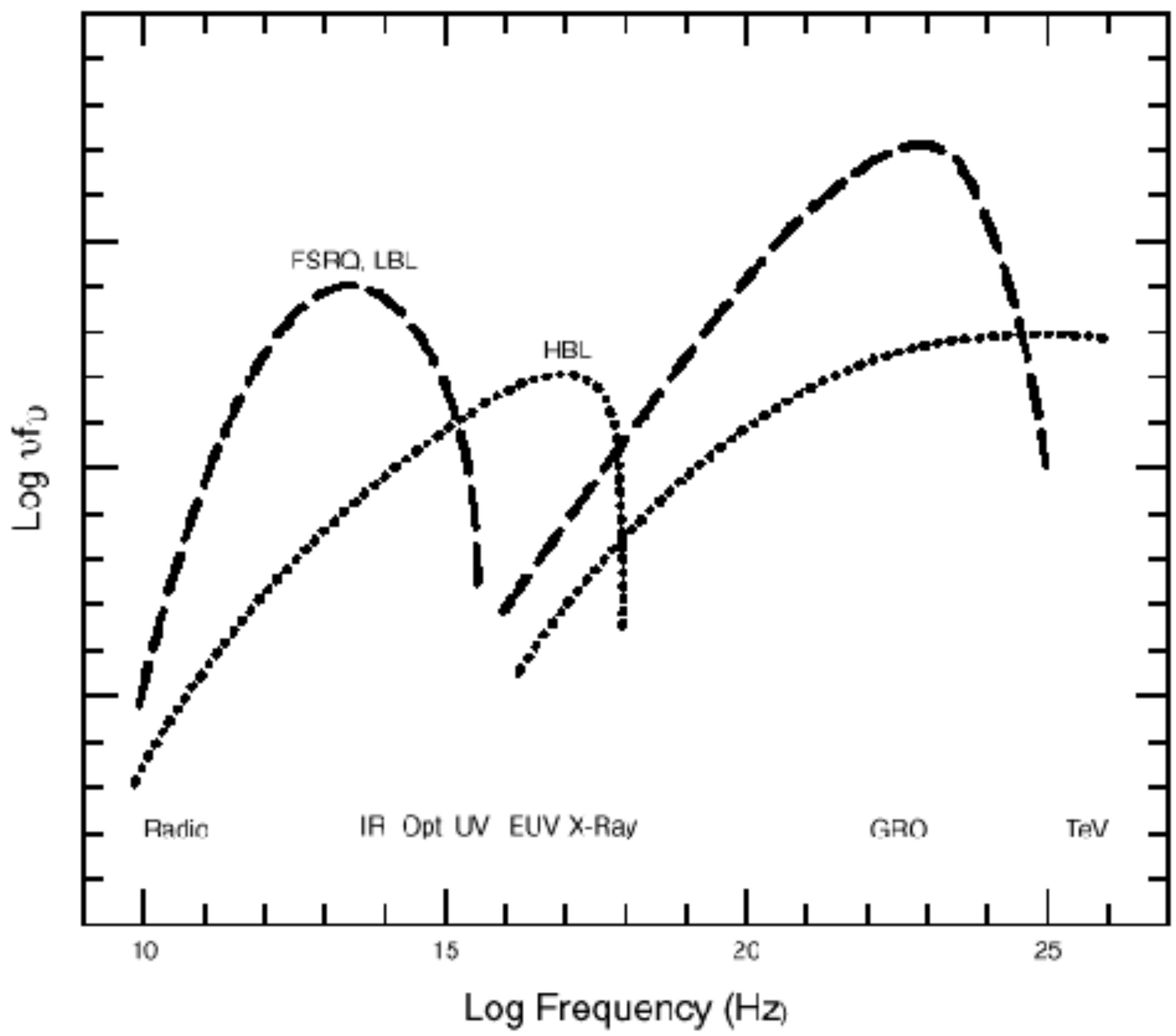
Examining the cluster(s)

Clusters	3	4	5	6	7	8	9
Total	0.492	0.503	0.531	0.569	0.656	0.557	0.525
9	0	0	0	0	0	0	0.714 (7)
8	0	0	0	0	0	0.833 (6)	0.474 (19)
7	0	0	0	0	0.714 (7)	0 (1)	1 (3)
6	0	0	0	0.714 (7)	1 (3)	0.5 (6)	0.5 (6)
5	0	0	0.438 (16)	0.467 (15)	0.538 (26)	0.538 (26)	0.833 (6)
4	0	0.333 (6)	0.75 (4)	0.5 (4)	0.467 (15)	1 (3)	0.3 (10)
3	0.5 (10)	0.75 (4)	0.548 (31)	0.538 (26)	0.5 (4)	0.333 (12)	0 (1)
2	0.556 (45)	0.537 (41)	0.421 (19)	0.75 (4)	0.538 (13)	0.714 (7)	0.4 (10)
1	0.421 (19)	0.391 (23)	0.5 (4)	0.444 (18)	0.833 (6)	0.538 (13)	0.5 (12)

Blazars

Blazars are AGNs observed down their relativistic jet!

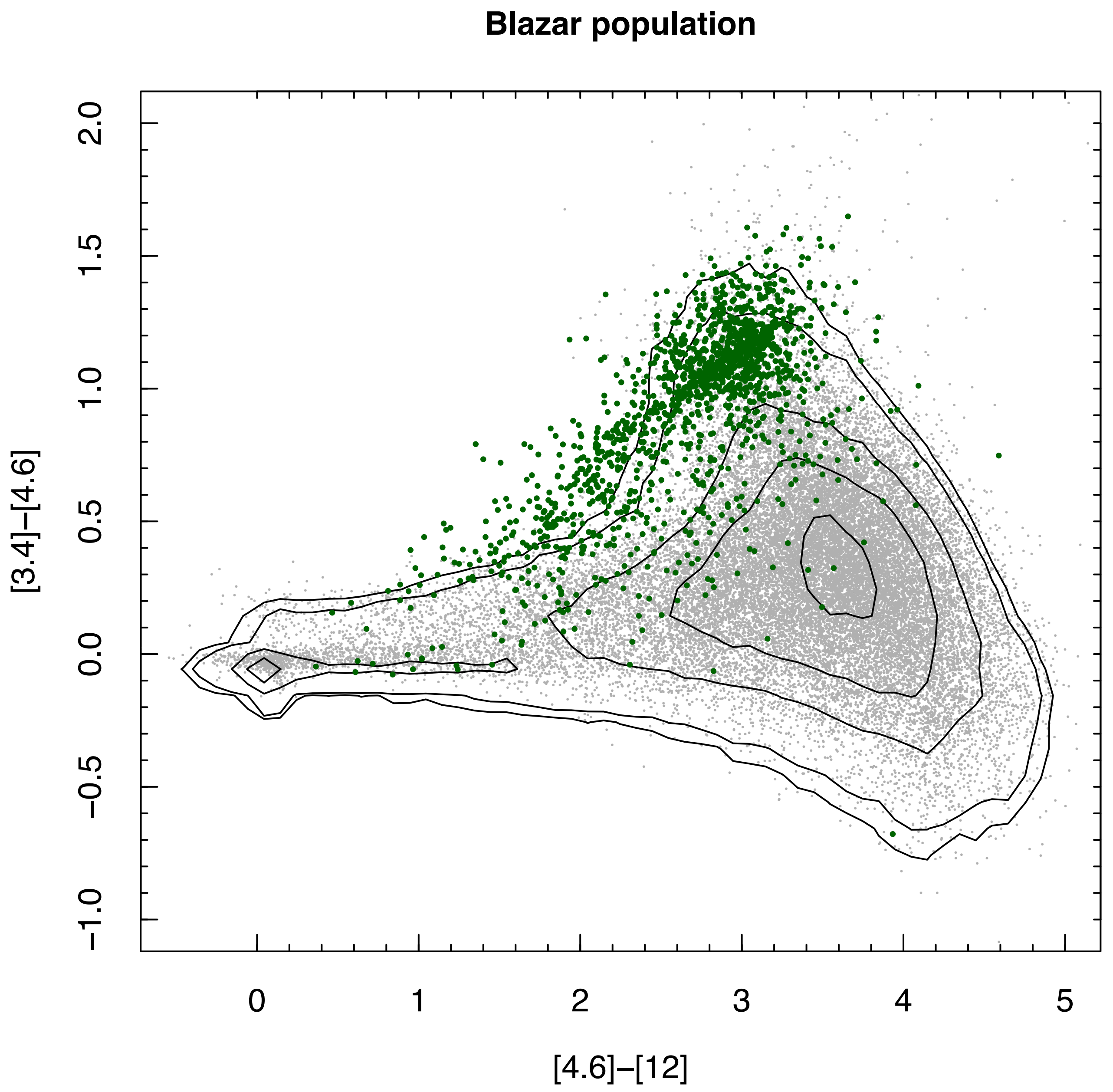
- Useful for the understanding of the emission mechanism at the very centers of AGNs
- Rarest class of AGNs but several sub-classes in terms of spectral characteristics have been observed
- γ-ray emission dominates their energy output



An interesting by-product

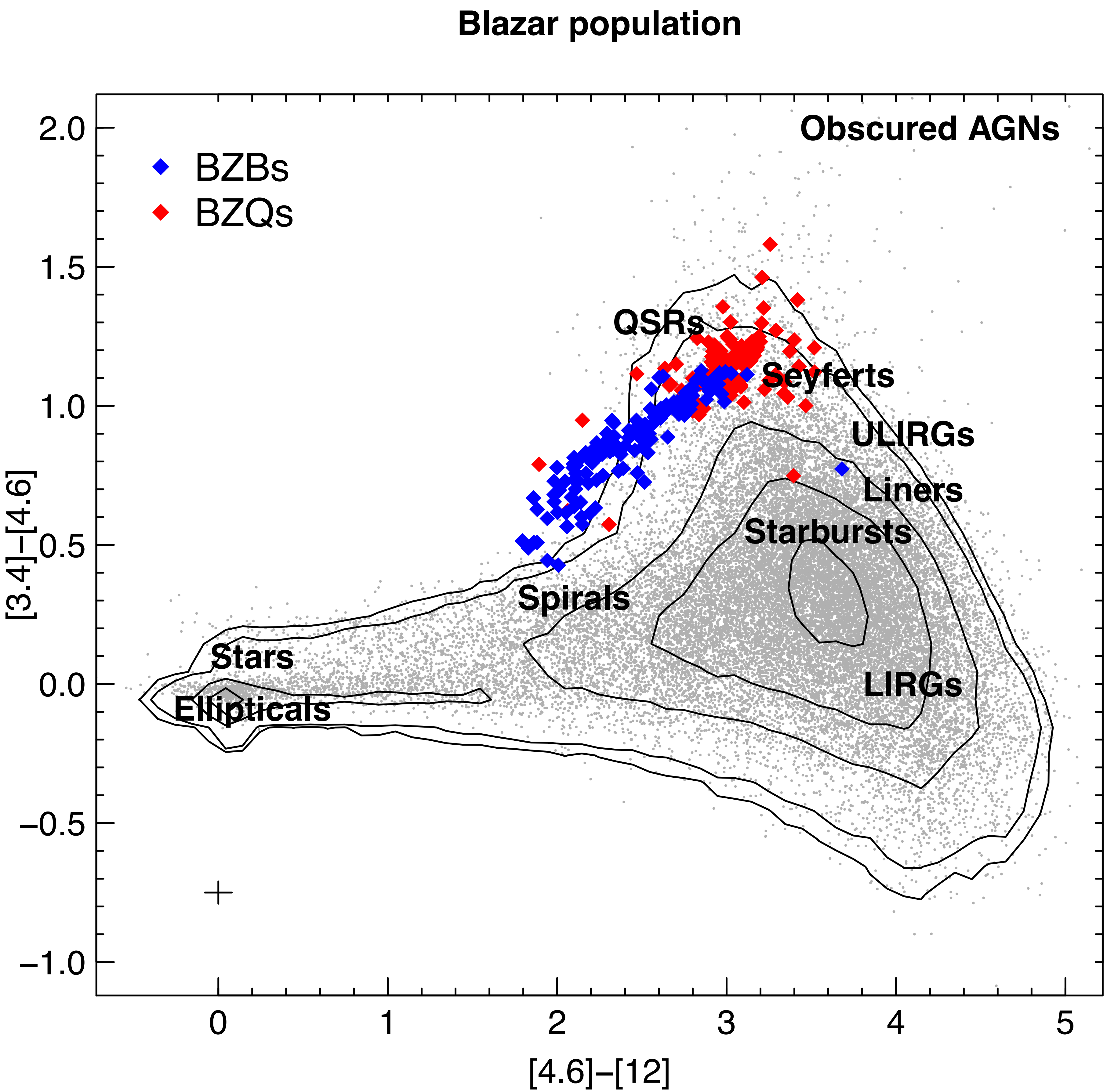
CLaSPS has been applied to a sample of AGNs selected with different techniques within the largest multi-wavelength *feature* space available from large area astronomical surveys, spanning from MIR to UV

Dataset	$\rightarrow$	*WISE* sources;
Features	$\rightarrow$	UV(*Galex*)+Optical(*SDSS*)+ NIR(*UKIDSS*)+IR(*WISE*)
Label	$\rightarrow$	Blazars spectral classification (ROMA-BZCat), γ-ray emission



A clear peak in the score values for few clusters has triggered more extensive investigation

Fermi results



Method: what's next?

Inclusion of upper limits in the clustering can follow two different approaches:

- upper limits are replaced by multiple realizations of their value according to a model of the observable, then distinct clusterings are performed and statistically combined (**conservative, but need a model!**)
- the upper limits are replaced by values obtained by interpolation (or extrapolation) of the detected values in the same dataset (**risky!**).

The clusterings can be used to "train" a classification tool and extract sources based on the distribution of the *labels*

Data-driven consistent binning for continuous *labels* (co-clustering)

A slightly different approach that does not employs *labels*:

different clusterings of the same dataset obtained using all the observables as *features* or previous labels are compared, and single sources are used as "tracers" of interesting properties.

Conclusions

A *serendipitous* finding obtained using CLaSPS on the Blazars population, reliably connecting for the first time, non-thermal emission and IR observations.

CSC+ sample is a typical example of the datasets that will become widespread with large area surveys and VO technology. In the working:

comparison with similar results from similar dataset
do "not X-ray" observables trace the X-ray properties of AGNs?
can classification of AGNs be achieved using the available *features*?

Homogeneous datasets?

C-COSMOS: unmatched wavelength coverage, tailored for the investigation of AGNs-galaxy connection as a function of the environment;
SWIRE: mostly optical and IR coverage, focused on the relation of the SFR with nuclear activity;