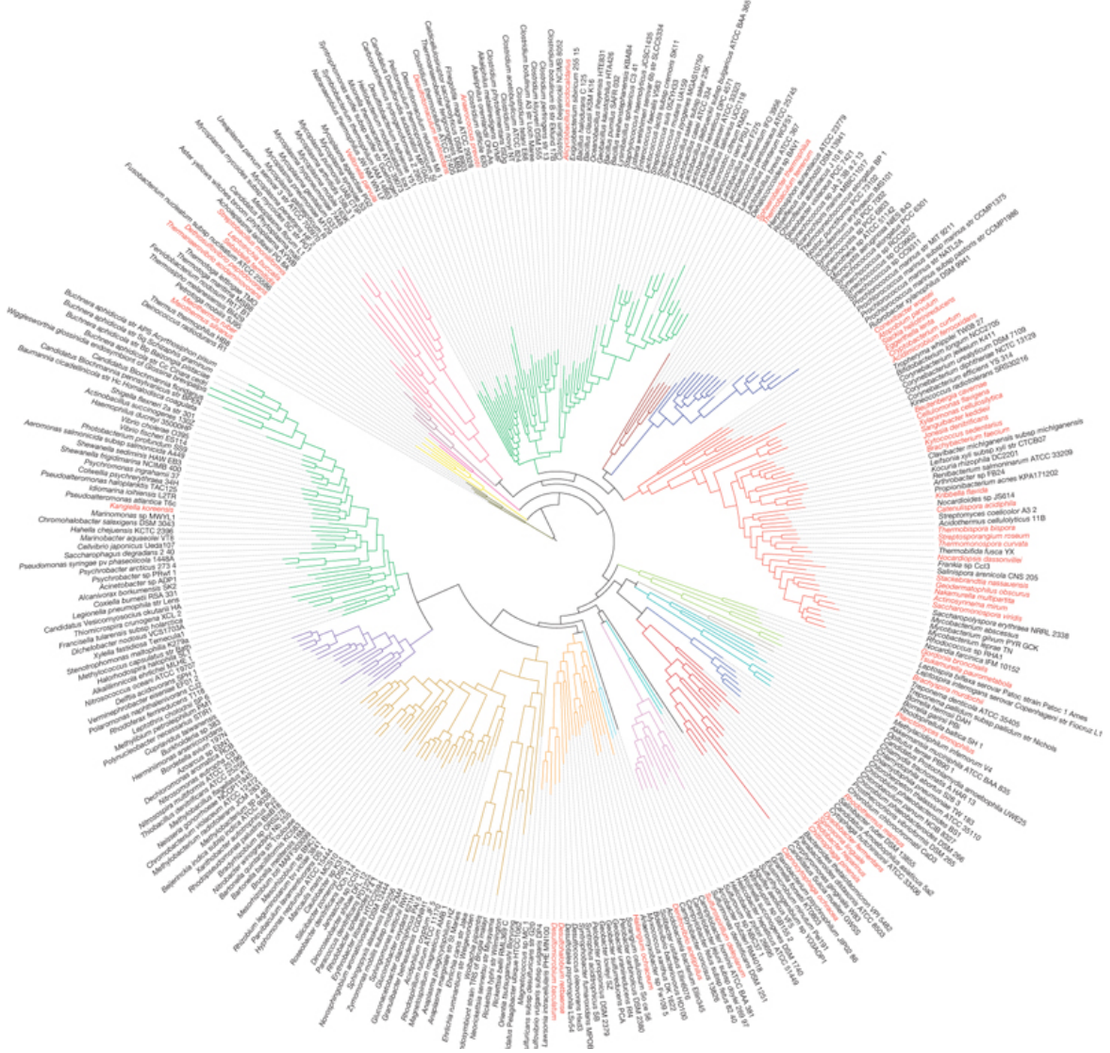


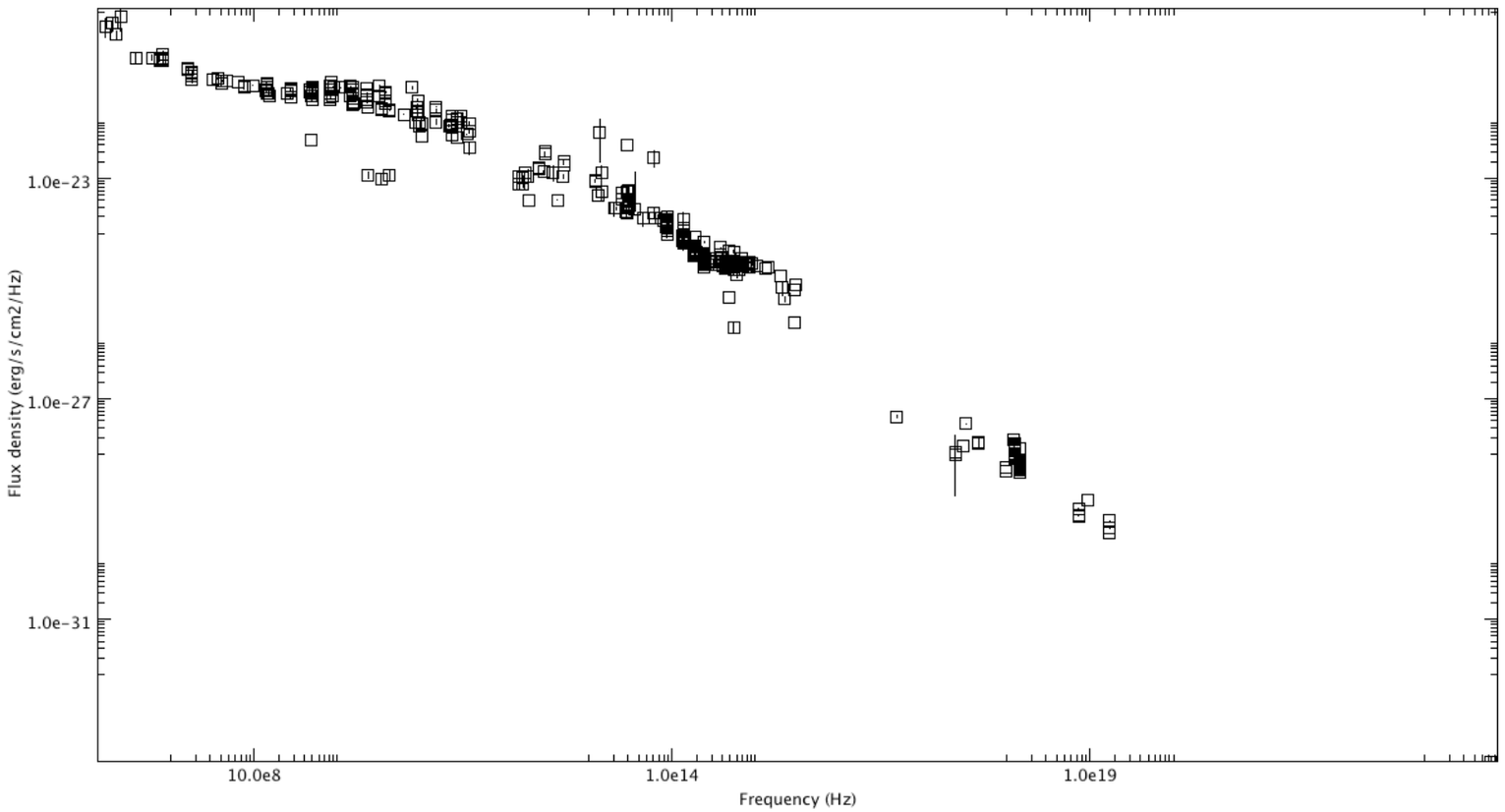
The exploration of multi-wavelength astronomical datasets: the case of AGNS in the Chandra Source Catalog and unsupervised clustering.

R. D'Abrusco

KISS Study Program - Digging Deeper: Algorithms for Computationally limited searches in Astronomy - 6/9/11 - Pasadena

Motivations

Characterization of the distribution of AGNs in a high dimensionality parameter space obtained by combining multi-wavelength data and study of their X-rays properties.



The primary purpose of this study is to obtain a possible census of AGN behavior in the 13-dimensional features space of X-UV-optical-IR-Radio photometry, and pick up outliers and constrain their nature.

The approach

Unsupervised clustering in a high dimensional features space

AND

use additional information (labels) to identify interesting cluster(ing)s

IN ORDER TO

derive new high dimensional correlations and/or expand known correlations to other bands, and/or spot unusual behaviors (**outliers**).

Statistical issues

Few points for a large space

> 10 dimensional features space
10^2~10^3 sources

Very low specific density

Upper limits & Clustering

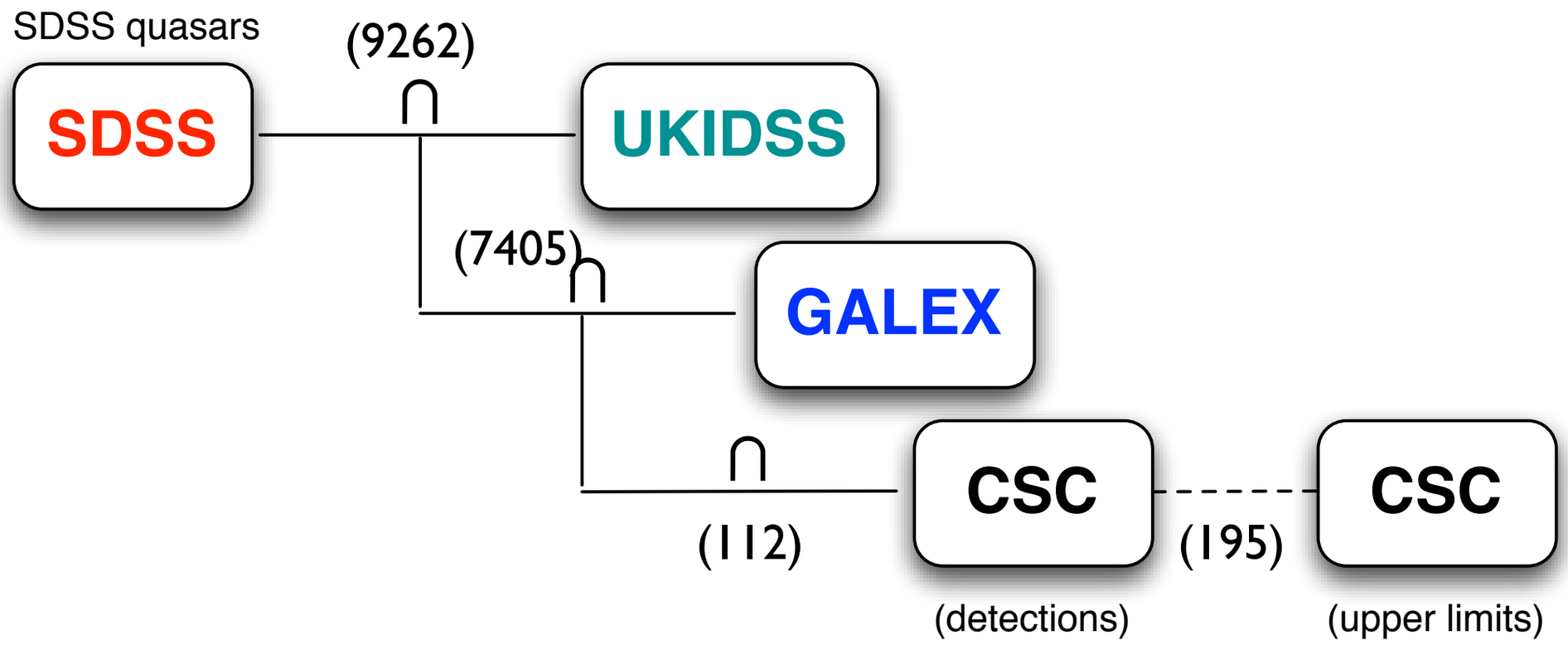
Inclusion of sources with no detections but reliable upper limits.

Outliers vs Clusters

Most clustering methods tends to prefer the selection of either well populated homogeneous clusters or to sparse clusters/singletons (outliers).

The datasets

1) **“Large area surveys” sample**

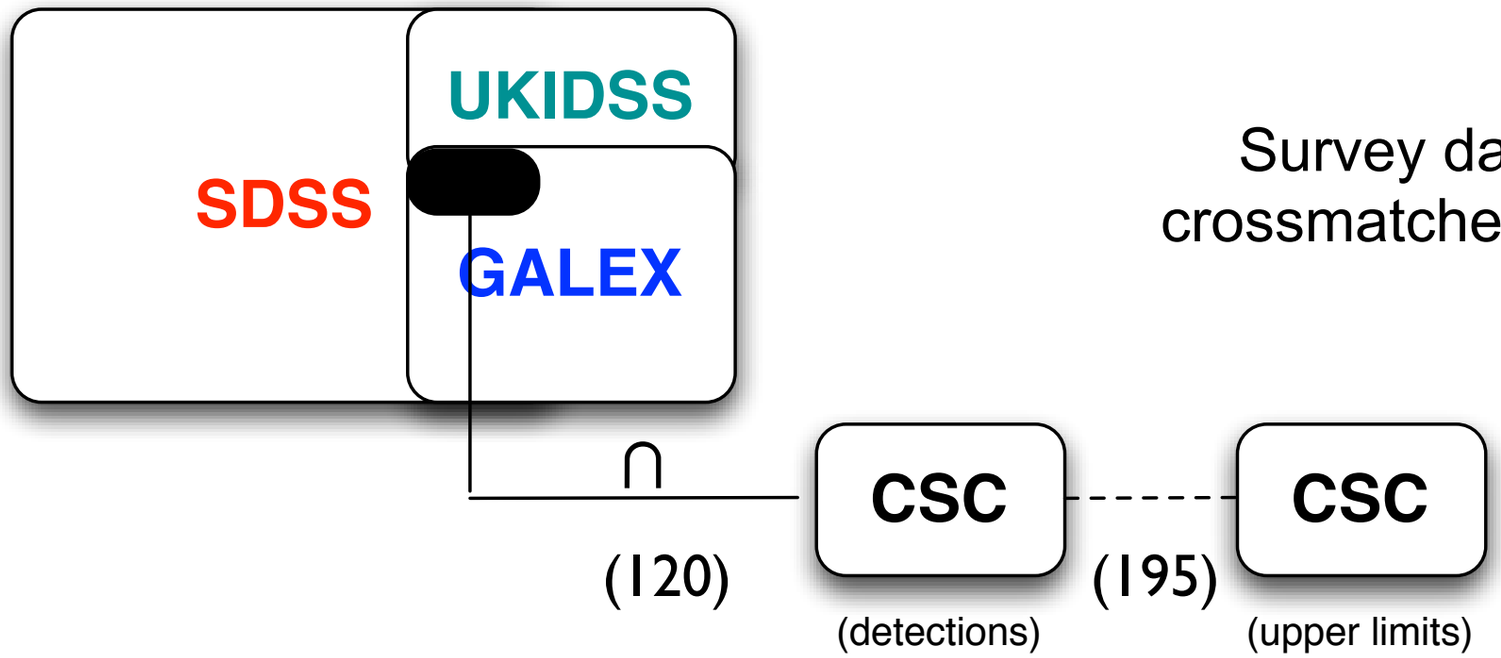


2) **Chandra COSMOS X-ray survey**

3) **SWIRE**

The datasets

1) **“Large area surveys” sample**



Survey databases with crossmatched catalogs, TAP

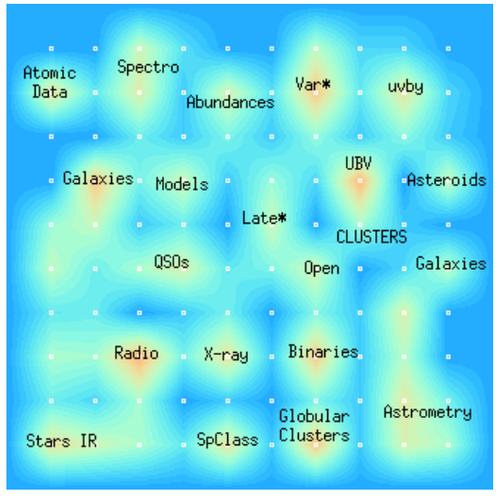
2) **Chandra COSMOS X-ray survey**

3) **SWIRE**

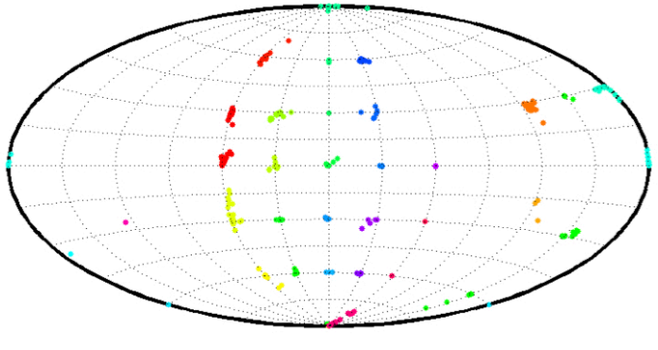
Multiple techniques

Hierarchical Clustering (K-means)

Self-Organizing Maps (SOM)

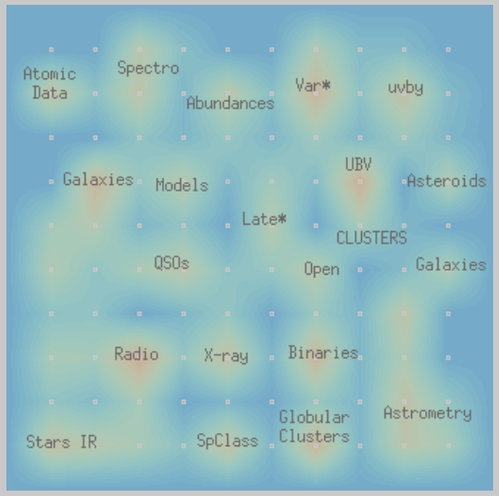


Principal Probabilistic Surfaces



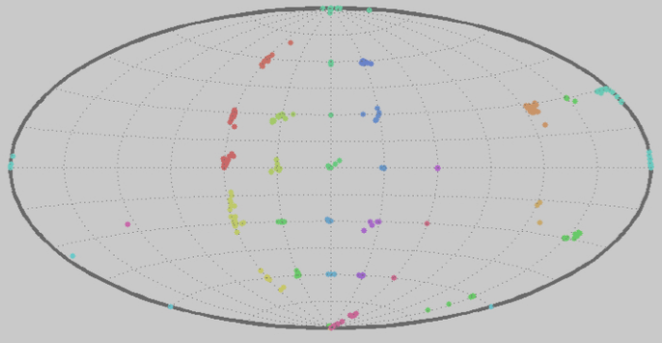
Multiple techniques

Hierarchical Clustering (K-means)



Self-Organizing Maps (SOM)

Principal Probabilistic Surfaces



Dendrograms

Representation of hierarchical structure - HC tree.

HC does not require a fixed number of clusters and produces all possible clustering, given a *measure of dissimilarity* (distance). Every generation of clusters maximizes the between-group dissimilarity.

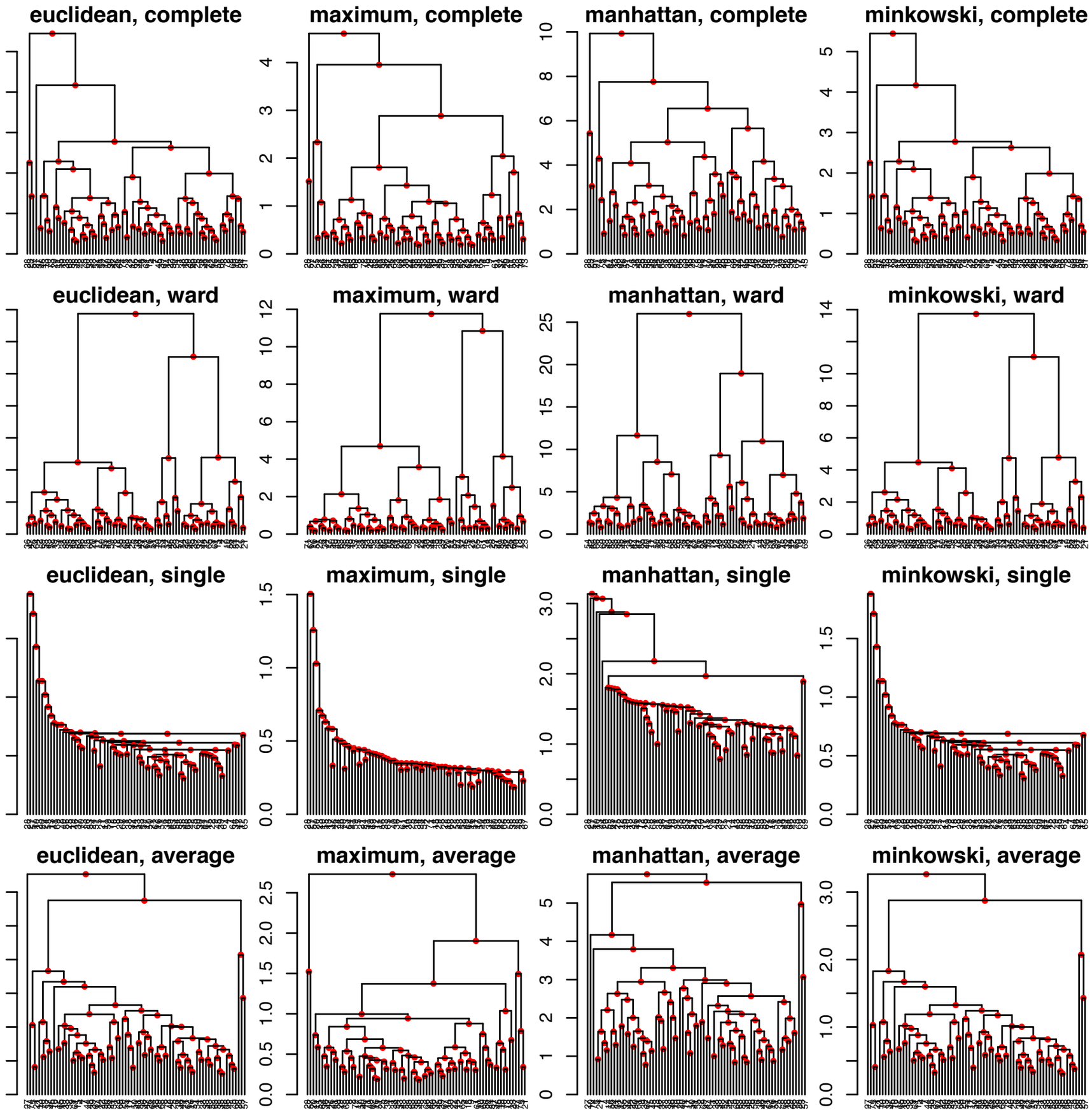
dissimilarity ≡ (metric, linkage strategy)

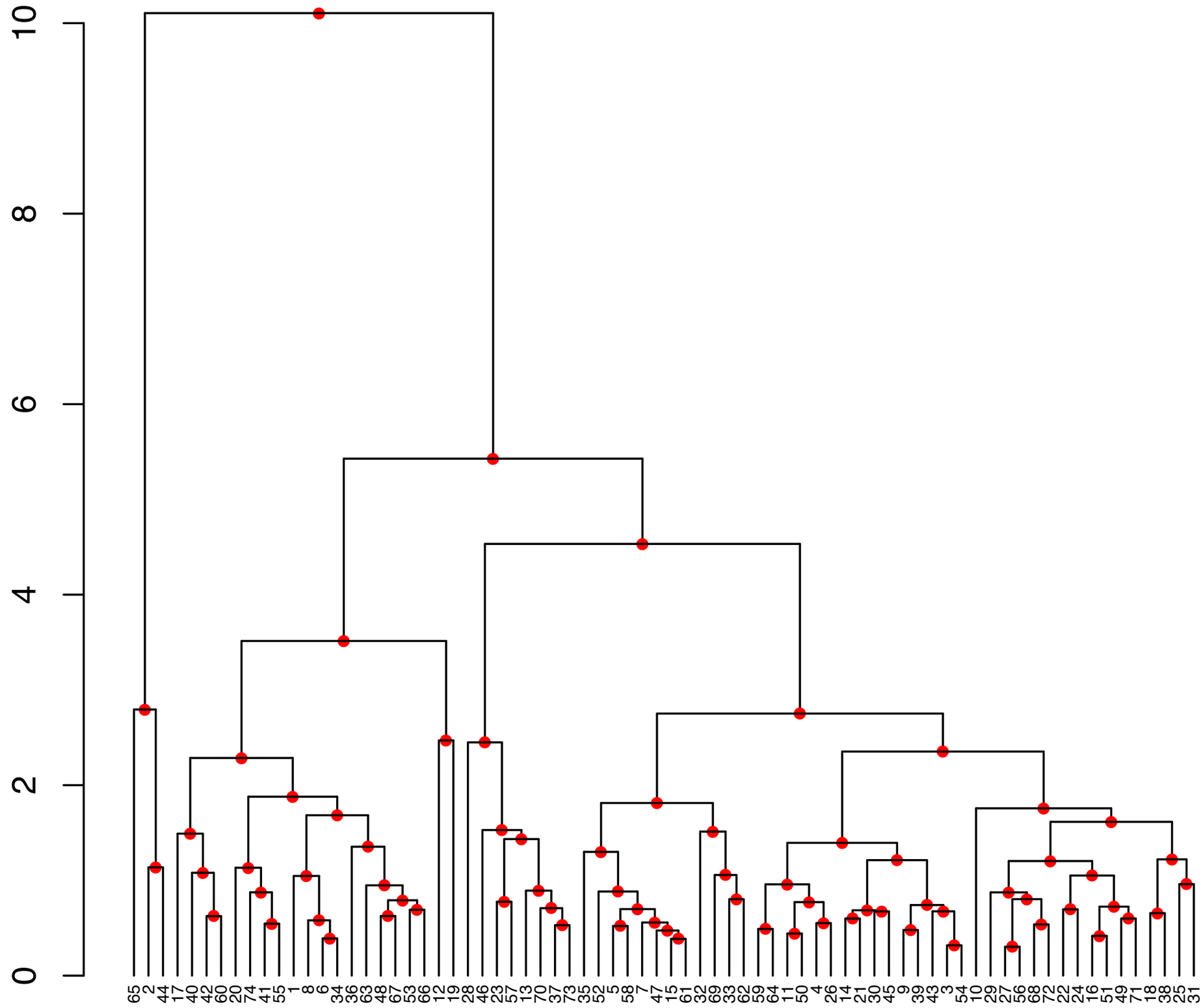
Metrics:

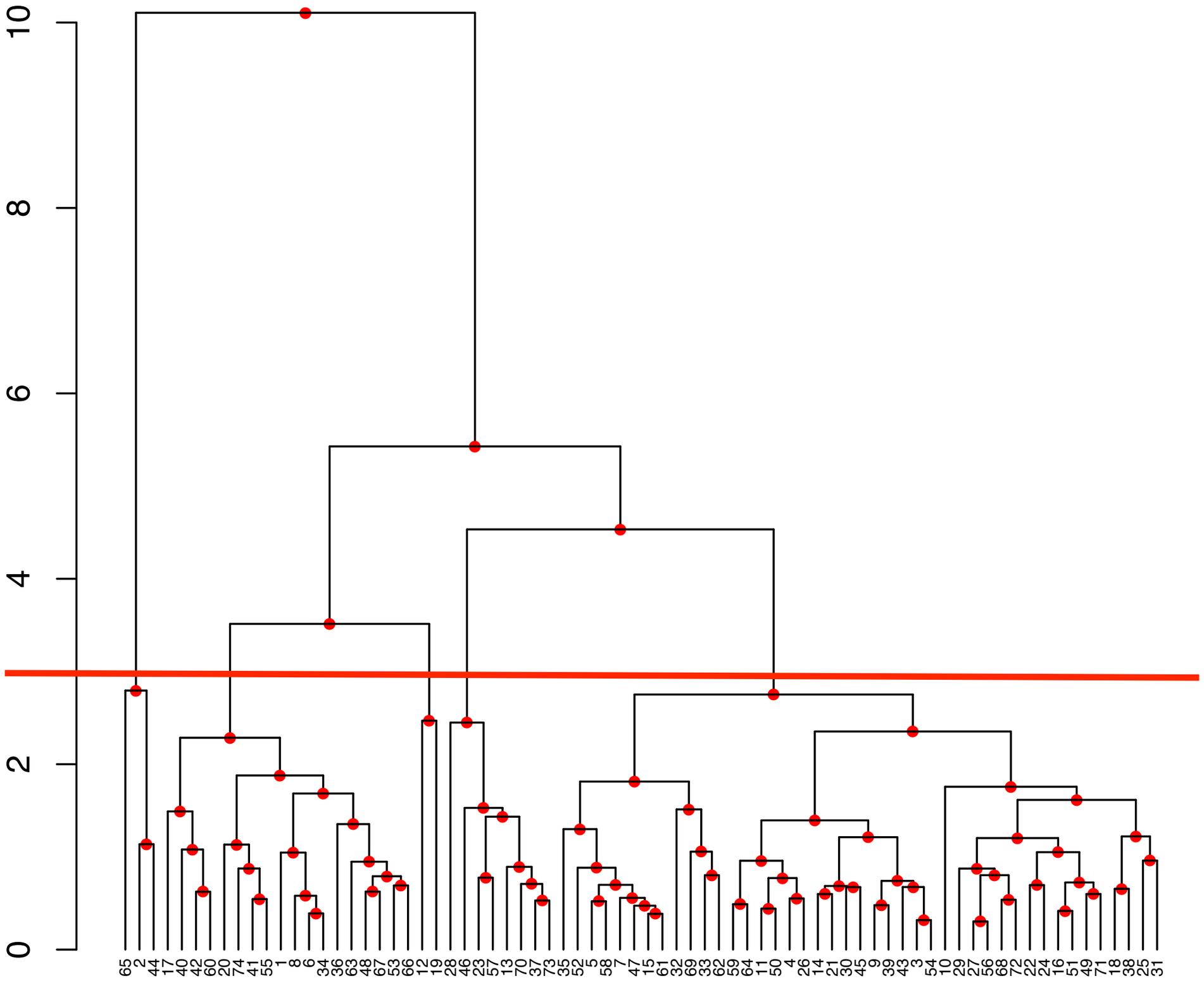
Euclidean, Manhattan,
Mahalanobis, maximum, etc.

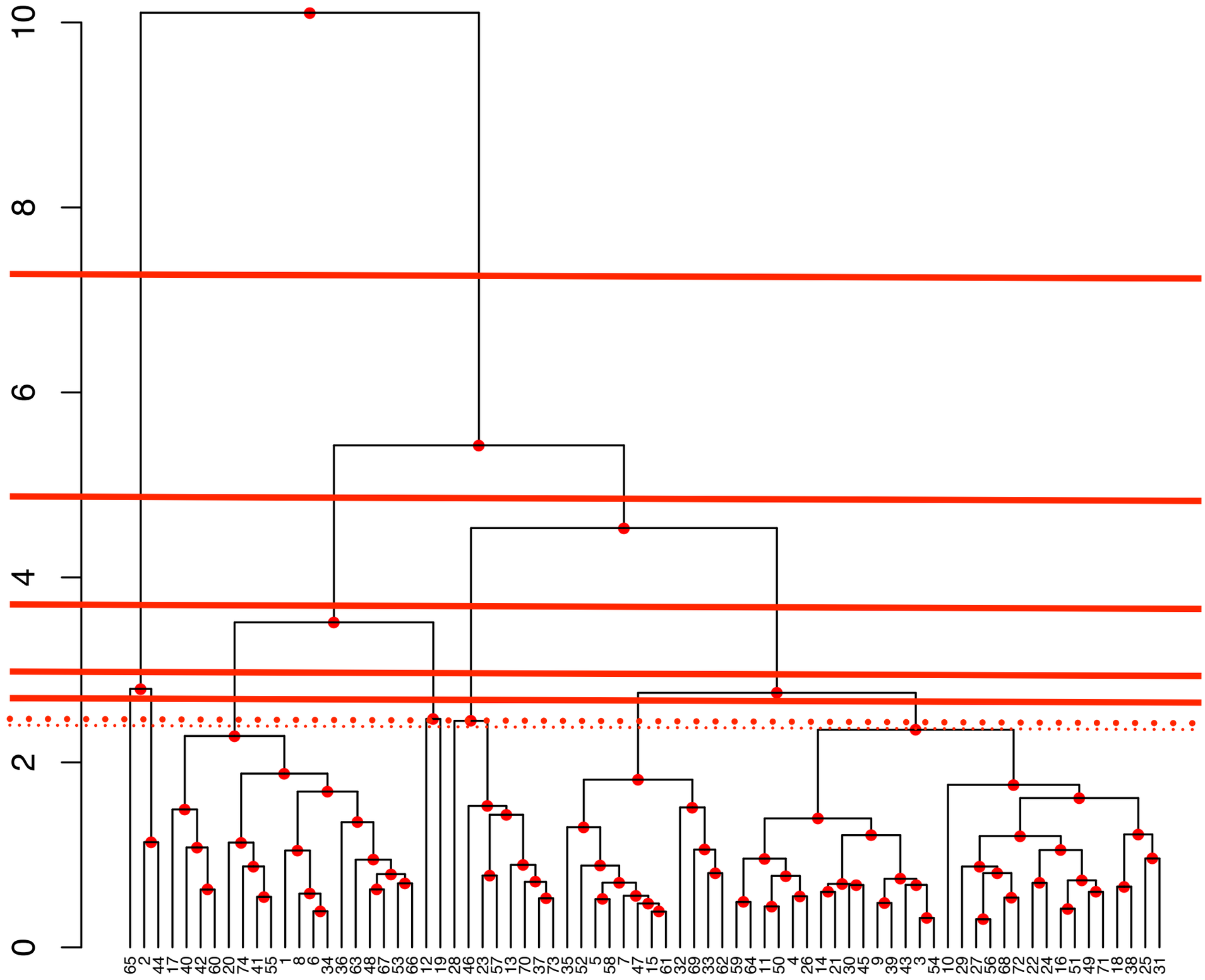
Linkage strategies:

complete, single,
average, etc.


euclidean, complete
maximum, complete
manhattan, complete
minkowski, complete
euclidean, ward
maximum, ward
manhattan, ward
minkowski, ward
euclidean, single
maximum, single
manhattan, single
minkowski, single
euclidean, average
maximum, average
manhattan, average
minkowski, average


Distance: euclidean; linking strategy: complete
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

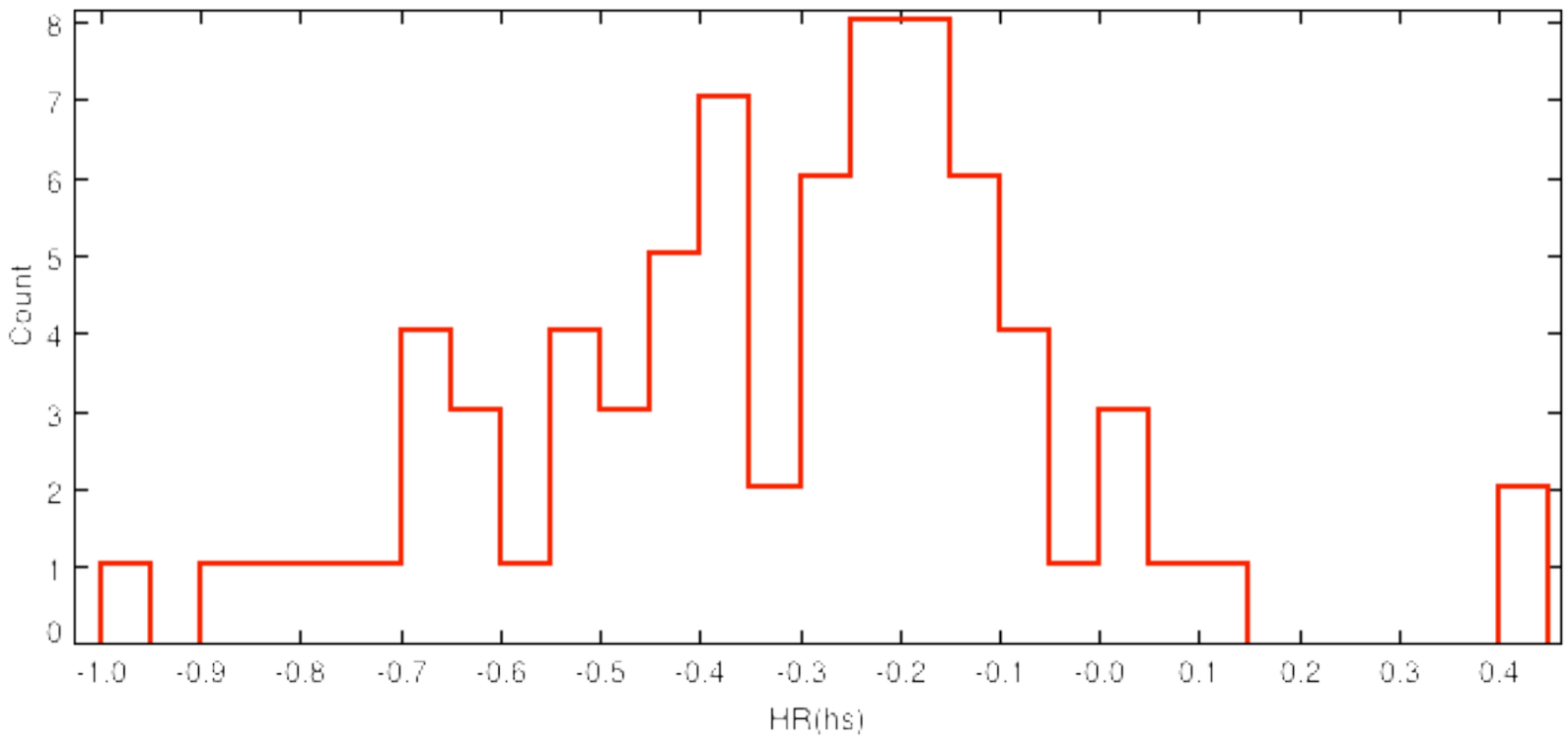

Distance: euclidean; linking strategy: complete
10
8
6
4
2
0
65 2 44 17 40 42 60 20 74 41 55 1 8 6 34 36 63 48 67 53 66 12 19 28 46 23 57 13 70 37 73 35 52 5 58 7 47 15 61 32 69 33 62 59 64 11 50 4 26 14 21 30 45 9 39 43 3 54 10 29 27 56 68 72 22 24 16 51 49 71 18 38 25 31


Distance: euclidean; linking strategy: complete
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

Labels to pick clusters

Other measured quantities, called “labels” (continuous or discrete) are separated into bins (if continuous) and used to pick those cluster(ing)s which seem interesting.

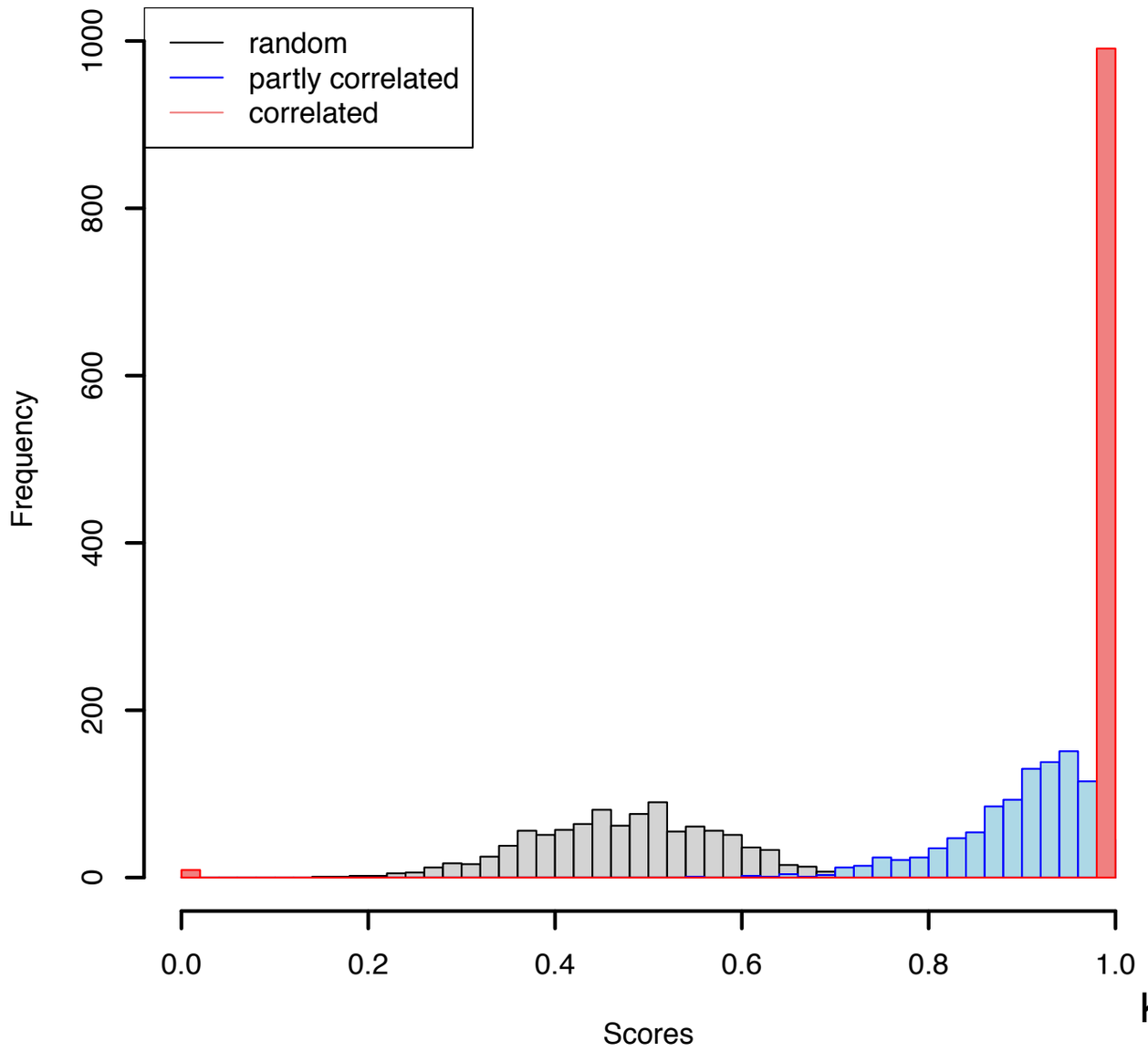
L_X, HR, Γ, n_h, X-ray/optical ratio, X-ray time variability, radio morphology (if available)



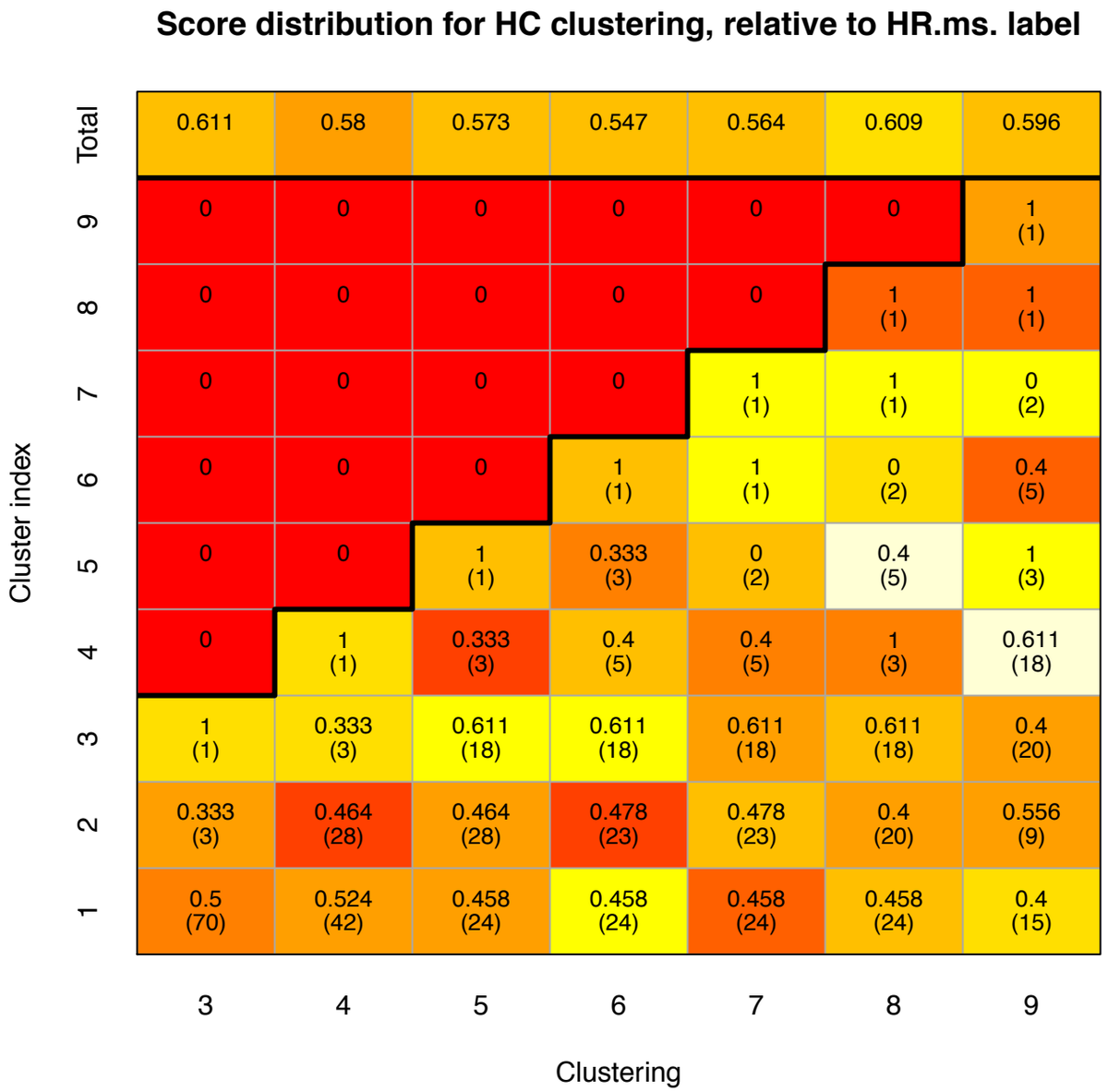
Discriminating clusterings

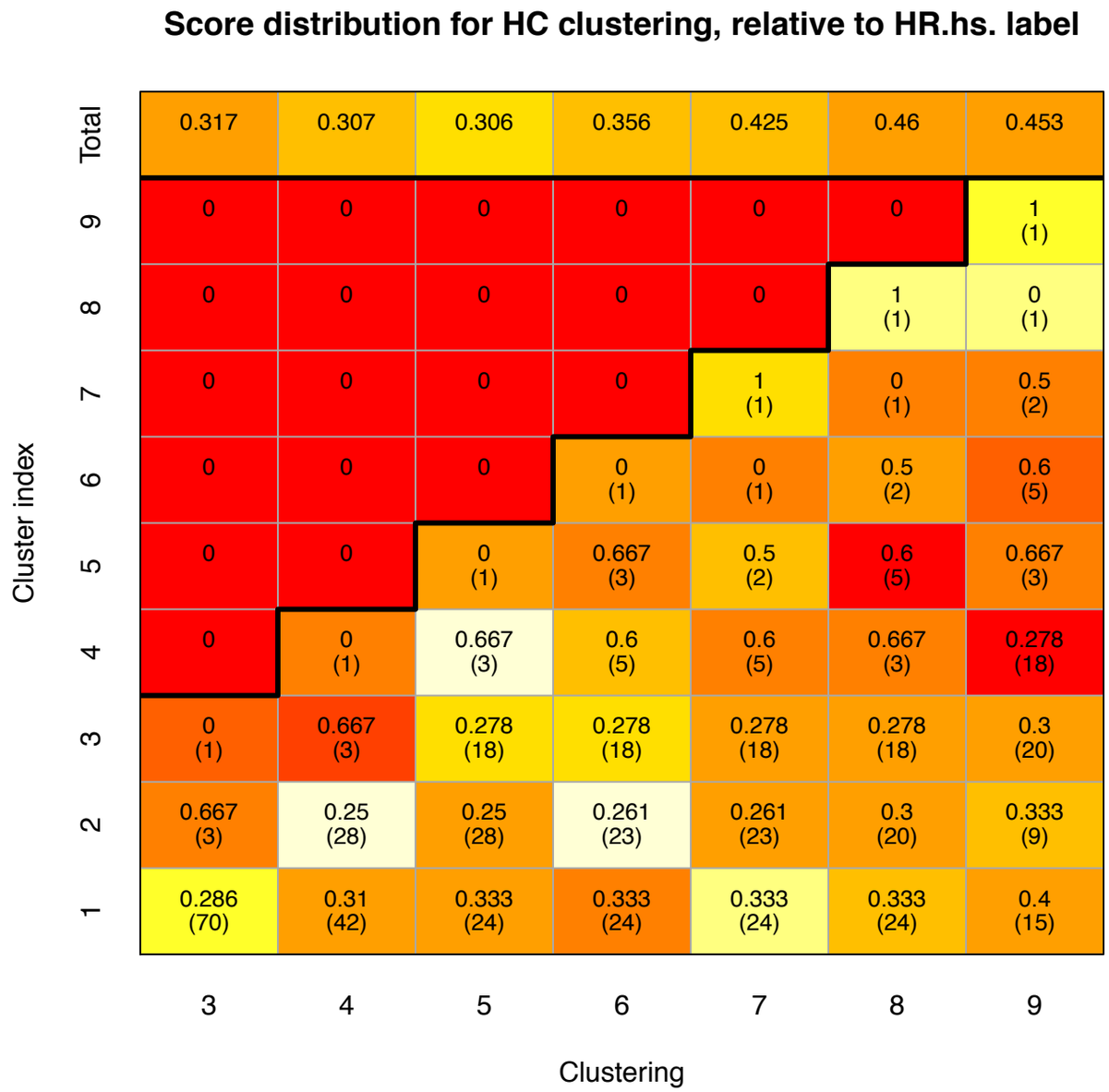
The score, a modified version of the “total variation” of a feature vector:

$$S_{TOT} = \frac{\sum_{i=1}^{N_{Cl}} S_i}{N_{Cl}} = \frac{\sum_{i=1}^{N_{Cl}} \left(\sum_{j=1}^{K-1} ||f_{ij} - f_{i(j+1)}|| \right)}{N_{Cl}}$$

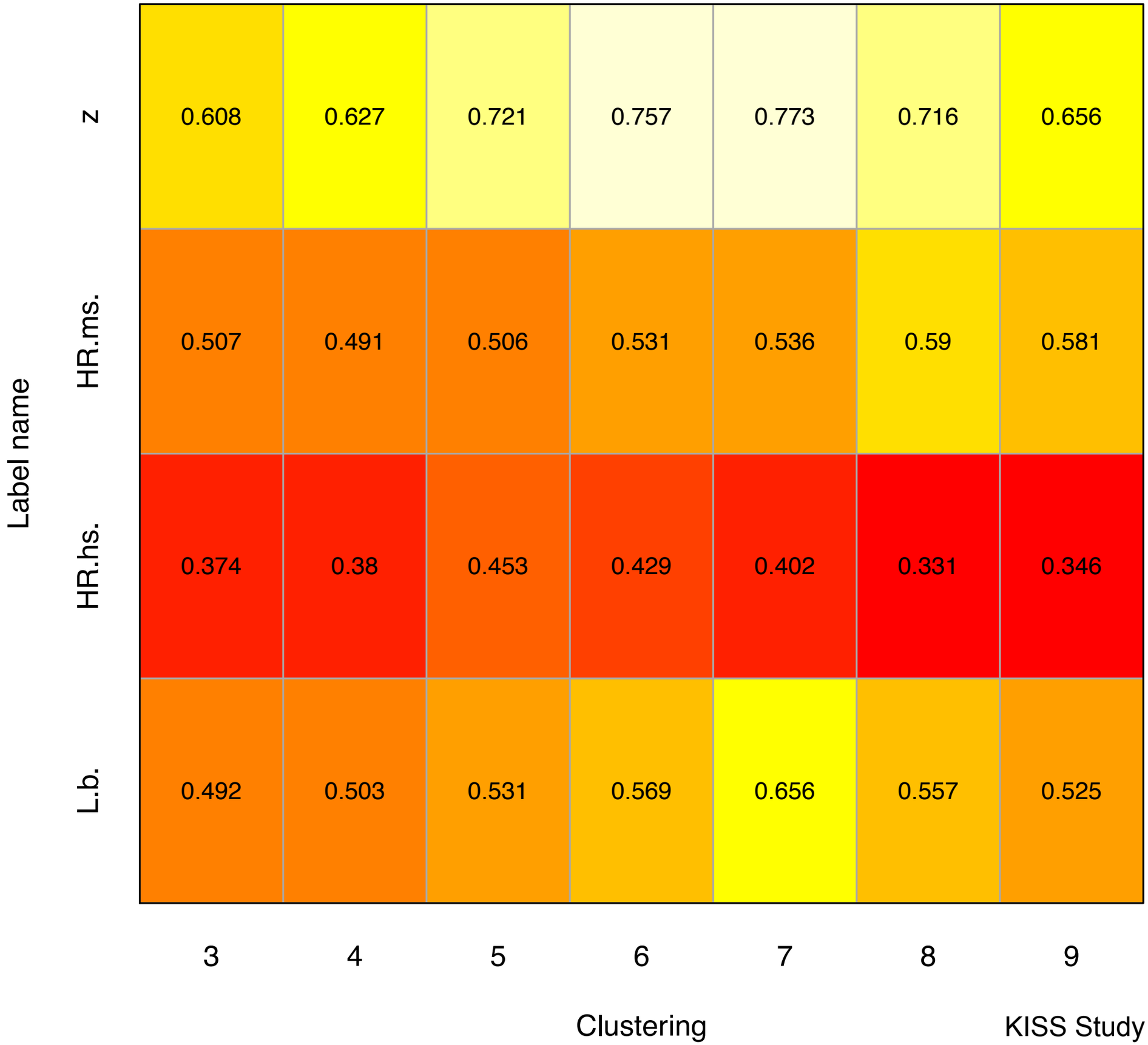


Scores for single labels and clusters

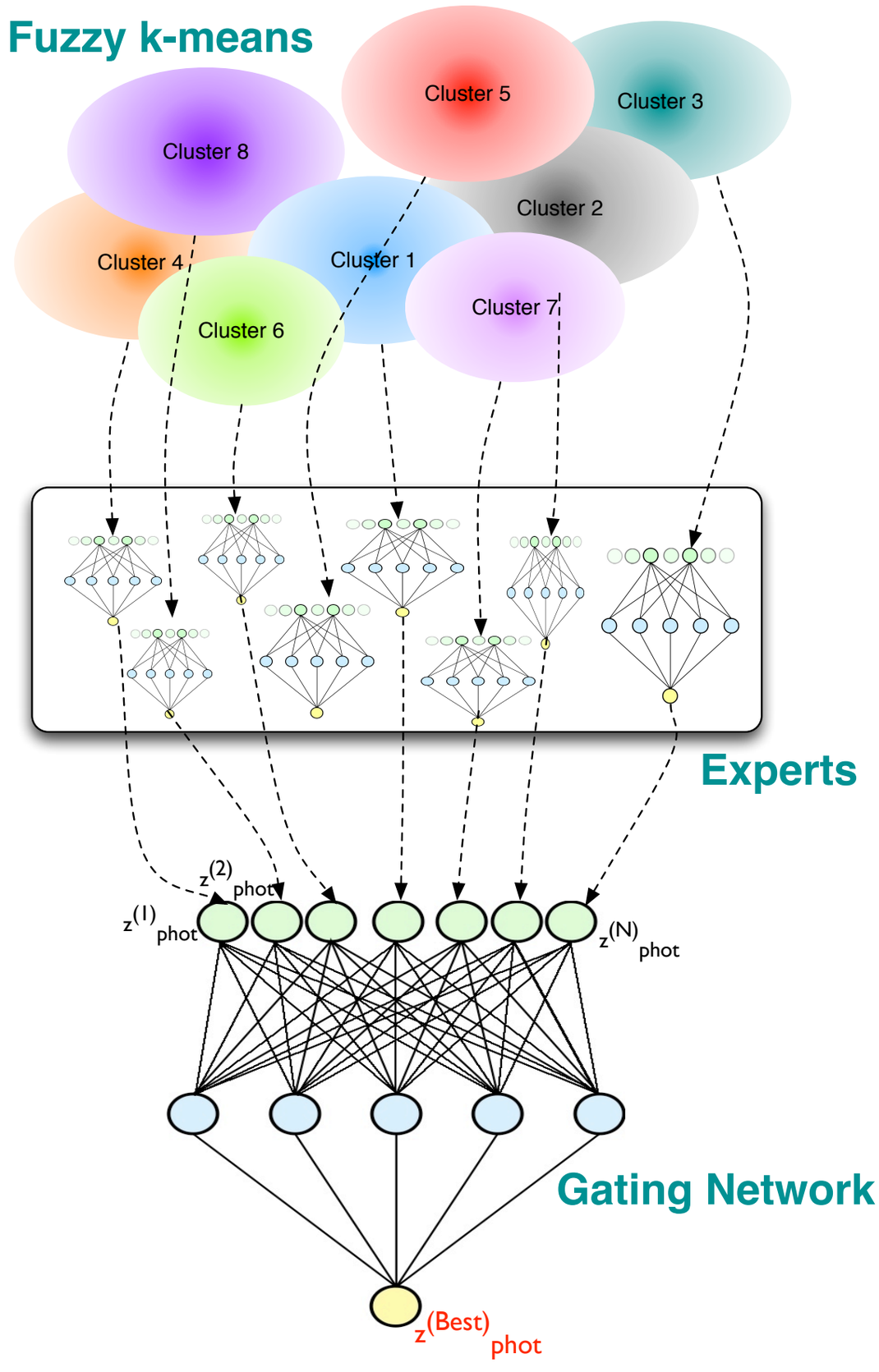




Scores for clusterings

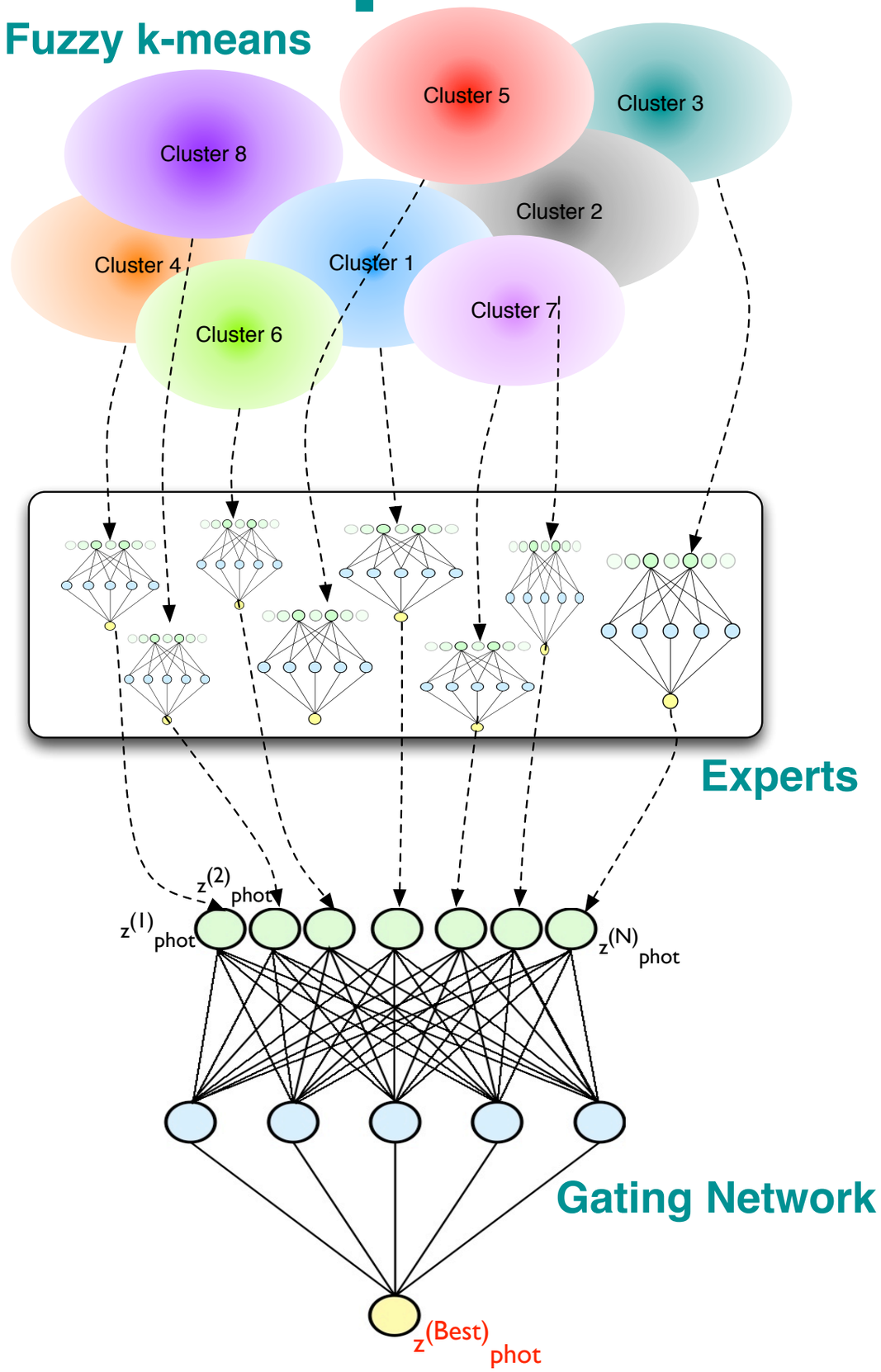


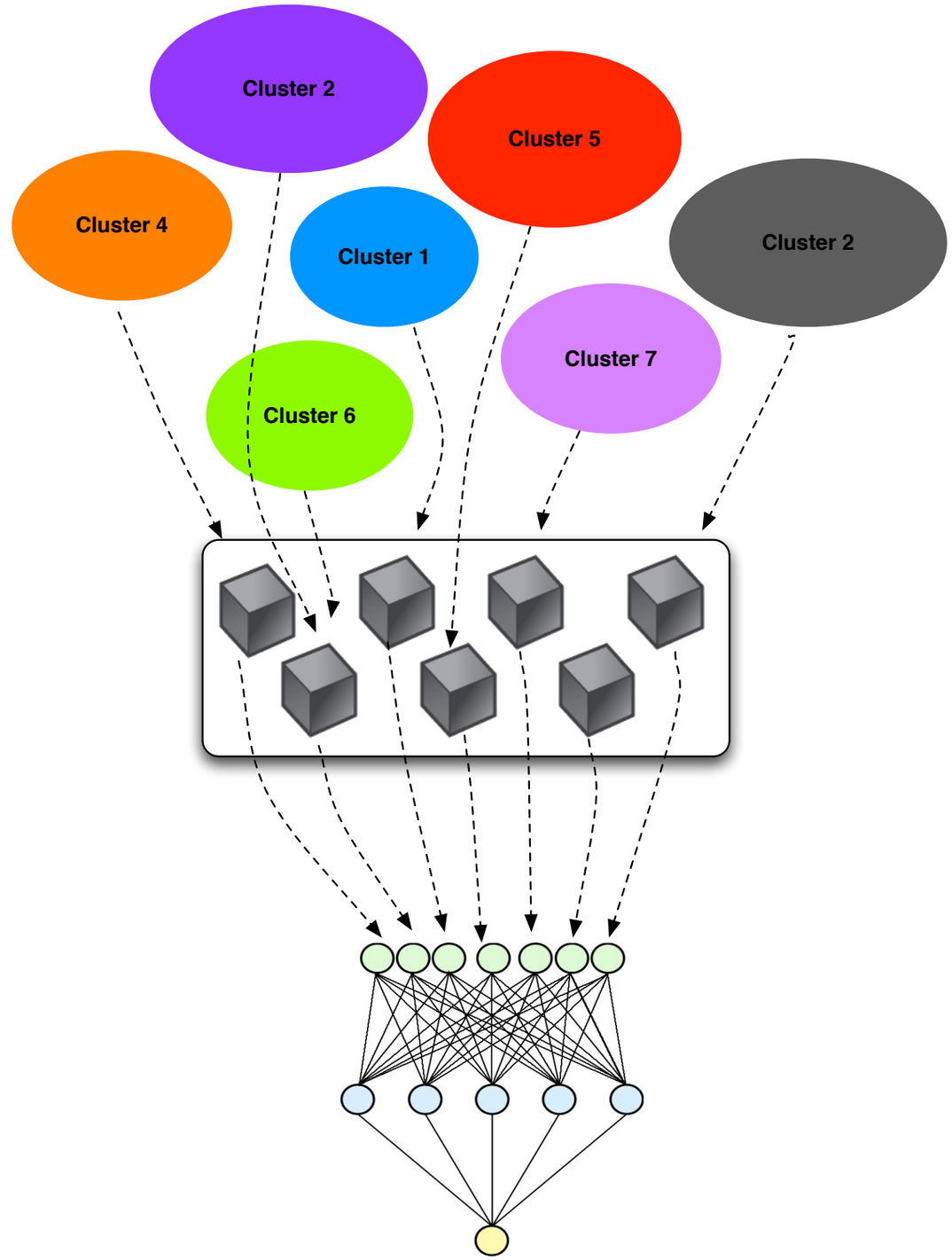
Experts for z_{phot}



Experts for other problems

Fuzzy k-means





Conclusions

Results for the first dataset and Chandra COSMOS survey with HC

Open issues:

clustering with upper limits/no detection/missing data (model based? simulations?)

labels “binning”: is there a data-driven way to accomplish this? Co-clustering is being explored.

Integration of different methods in a Gated Expert model:

example 1: template fitting and machine learning experts for z_{phot} calculation.

example 2: extraction of candidate quasars from optical photometric quasars.