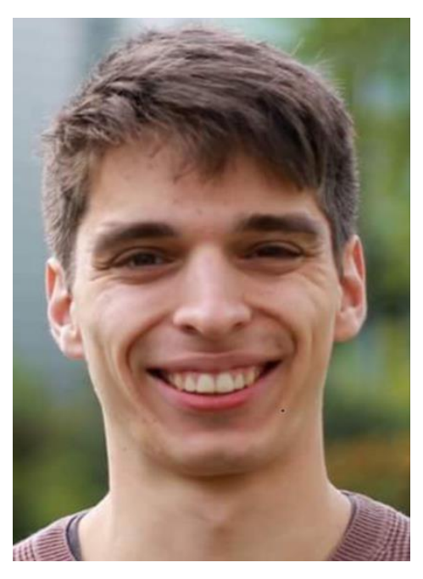
k-means clustering in galaxy feature data from the GAMA survey

Sebastian Turner[1], Lee Kelvin[1], Ivan Baldry[1], Paulo Lisboa[2], Steven Longmore[1], Chris Collins[1]

[1]Astrophysics Research Institute, LJMU, 146 Brownlow Hill, Liverpool, L3 5RF, UK
[2]Department of Applied Mathematics, LJMU, Byrom Street, Liverpool, L3 3AF, UK

✉ **s.turner1@2012.ljmu.ac.uk** **@sebturne**



Abstract

Using an unsupervised machine learning algorithm, we find that galaxies may be meaningfully divided into five distinct groups. We also explore new perspectives on the established bimodality of galaxies. Our approach will be useful for the analysis of the huge data volumes expected from next generation surveys like Euclid, and for new analysis of existing data sets like Galaxy Zoo.

Introduction

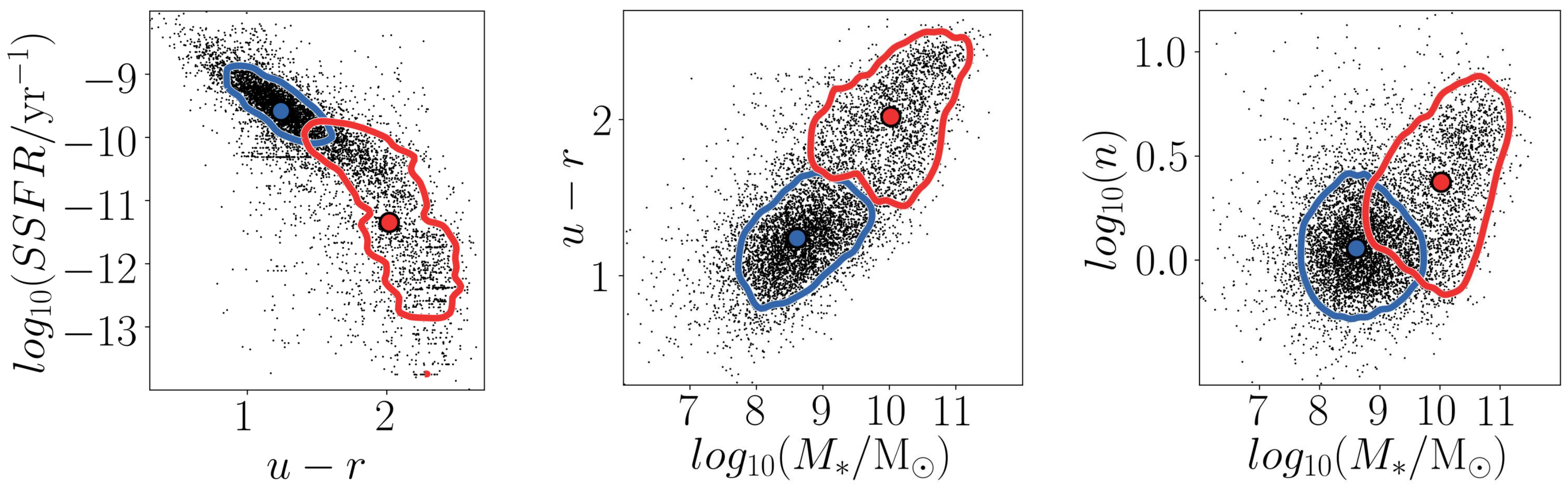
Understanding the diversity of galaxies necessitates a classification scheme that segregates galaxies in a way that reflects their formation and evolution. Galaxies are commonly distinguished as being blue, star-forming, disky, late-type galaxies in low density environments, or red, quiescent, spheroidal, early type galaxies in high density environments. The existence of further, meaningfully distinct subclasses has previously been mostly speculative. We explore this using the *k*-means unsupervised clustering machine learning algorithm.

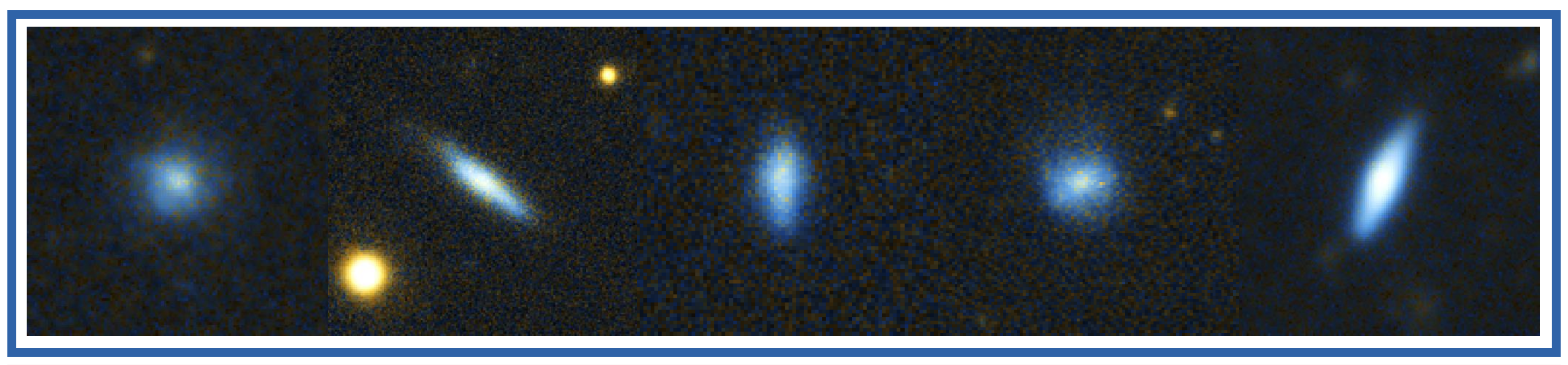
Data

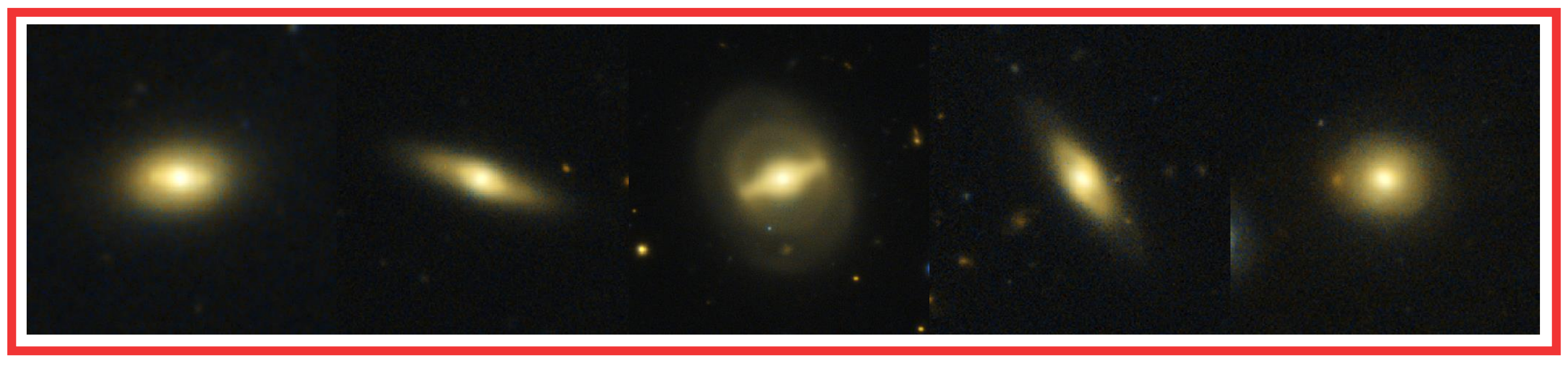
We derive a redshift- ($z < 0.06$) and magnitude- ($r_{PETRO} < 19.8$) limited sample of 7338 morphologically classified galaxies from the GAMA survey. We select a preliminary set of five features to represent our sample: **stellar mass (M_*)**, ***u-r* colour**, **Sérsic index (*n*)**, **half-light radius**, and **specific star formation rate**. We convert all features (apart from *u-r* colour) to logarithmic units, truncate our sample to remove outliers, and Z-score the data to **normalise** it. PCA reveals that stellar mass, *u-r* colour, and specific star formation rate correlate and dominate the variance within our sample. They are therefore expected to dictate most of the clustering outcomes, especially at low *k*. Sérsic index and half-light radius are expected to play a role in the clustering outcomes at higher *k*.

Results: Two Clusters

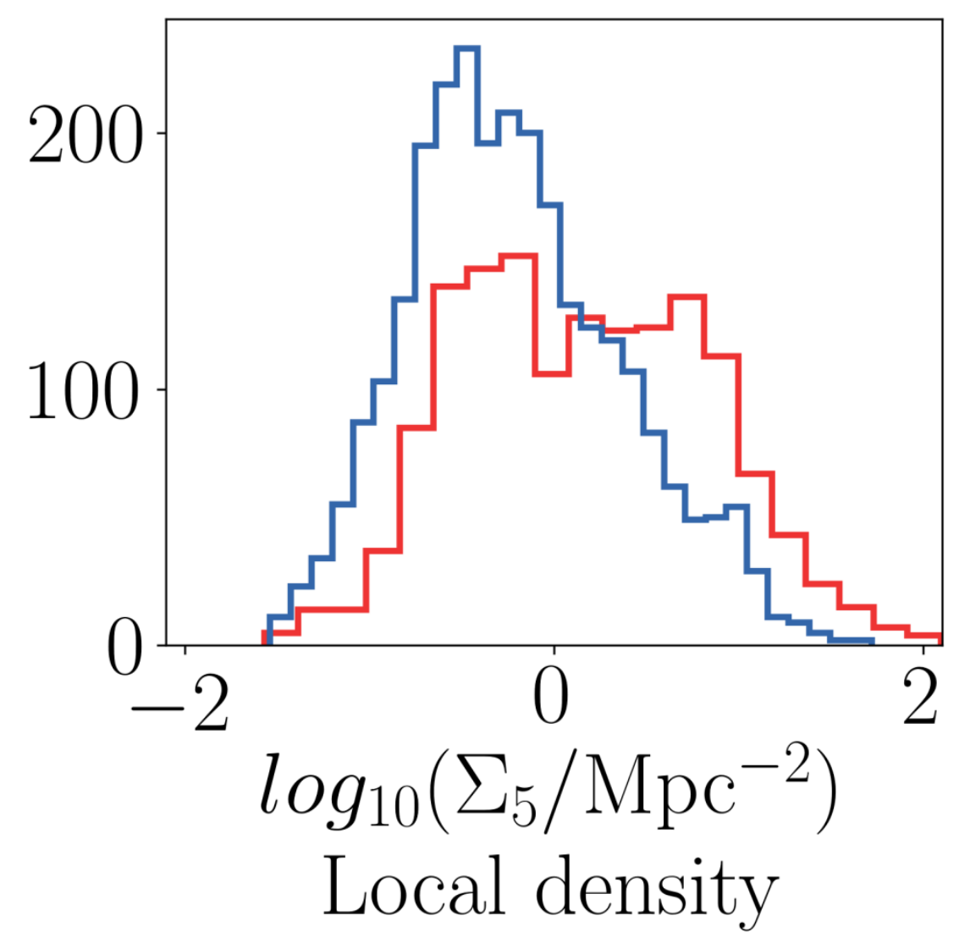
Searching for two clusters in our 5D feature-space means galaxies are distinguished mostly by their masses, colours, and star formation activity, in a manner that is broadly consistent with the established bimodality of galaxies. The difference in the morphologies of galaxies in each cluster arises mostly due to the correlation of morphology itself with mass, colour, and star formation.





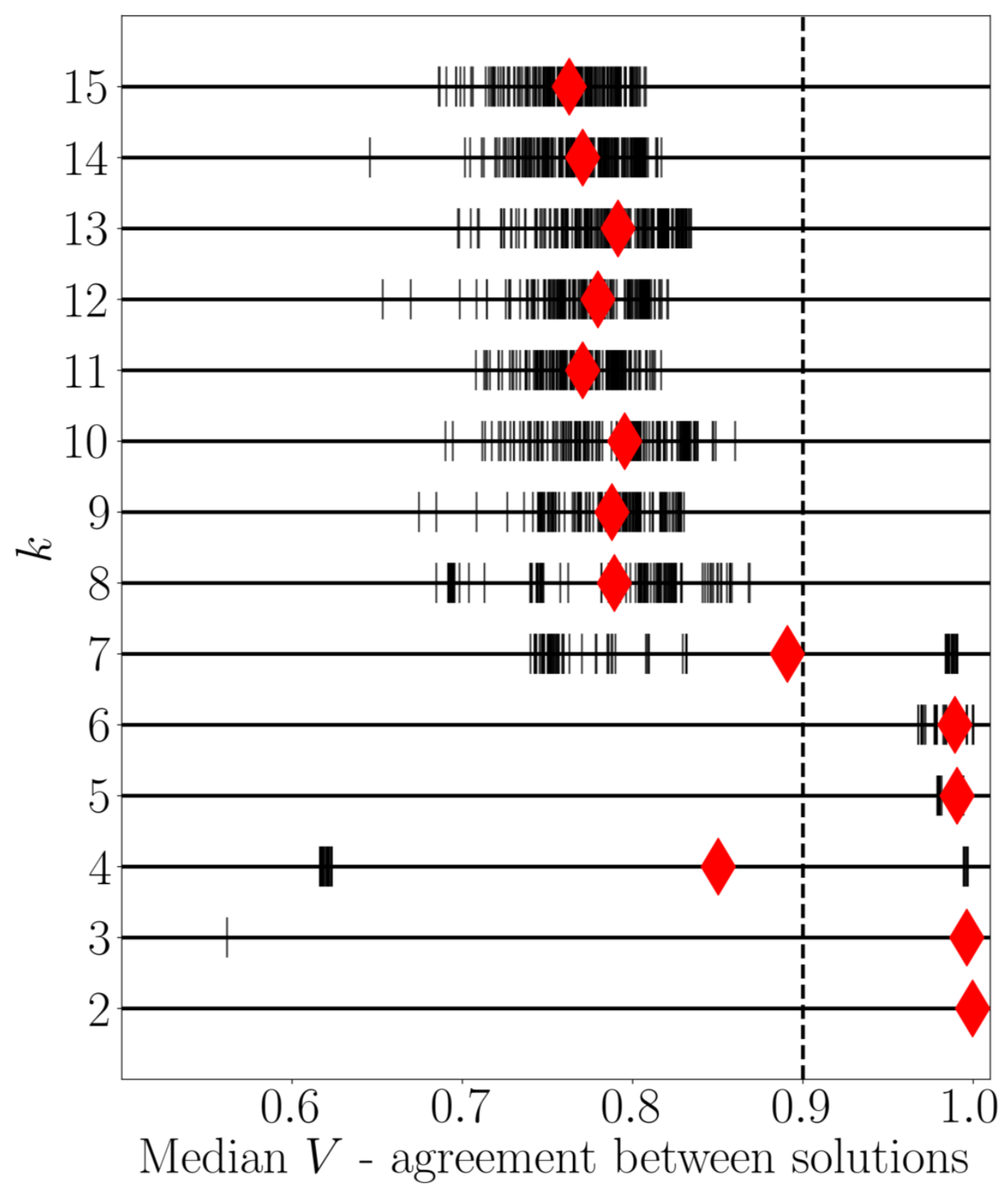


Splitting our sample into two clusters using *k*-means results in a coarse partition of data. This is in part due to the excess of low-mass, blue, star forming galaxies therein. More clusters are needed to properly explore our sample's data structure. The local environmental densities of the galaxies in each cluster are consistent with observed trends of environment with galaxy masses, colours, and star formation activity.



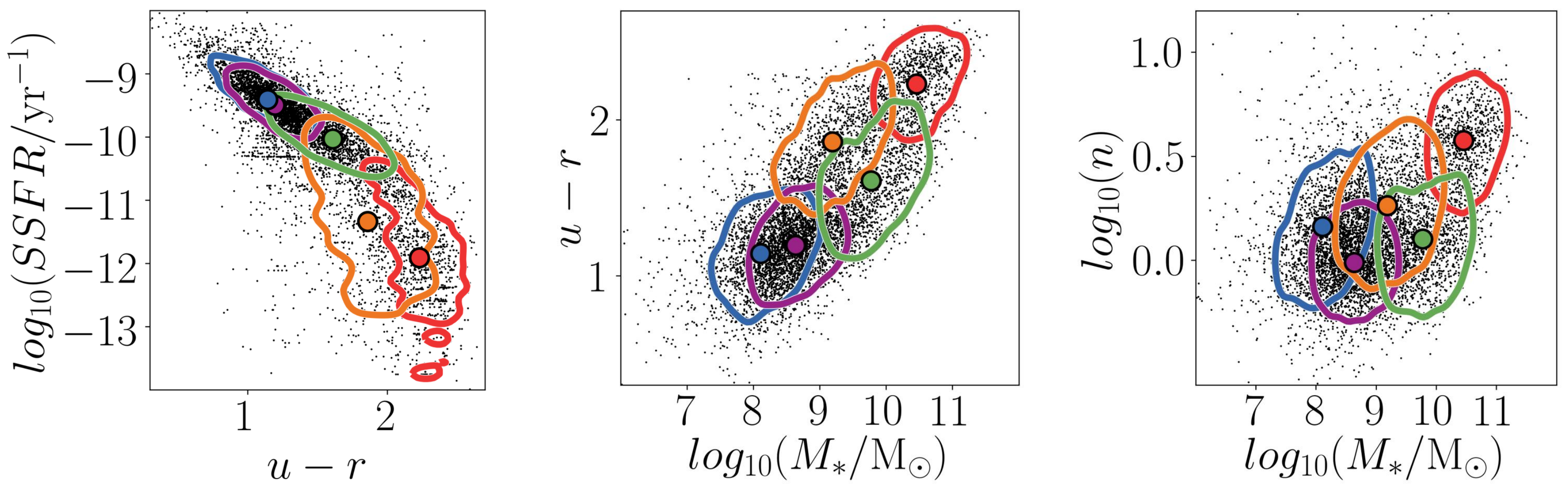
k-means

The *k*-means algorithm partitions our sample into *k* clusters. As a local search heuristic, the outcome of *k*-means is dependent on randomised input initialisations. We sample many varying initialisations to ensure we find globally optimal solutions. Adopting an exploratory approach, we also sample a range of values of *k* and examine whether clustering at each is stable. We focus below on **stable clustering at k = 2 and k = 5**. Clustering is also stable at k = 3 and k = 6.

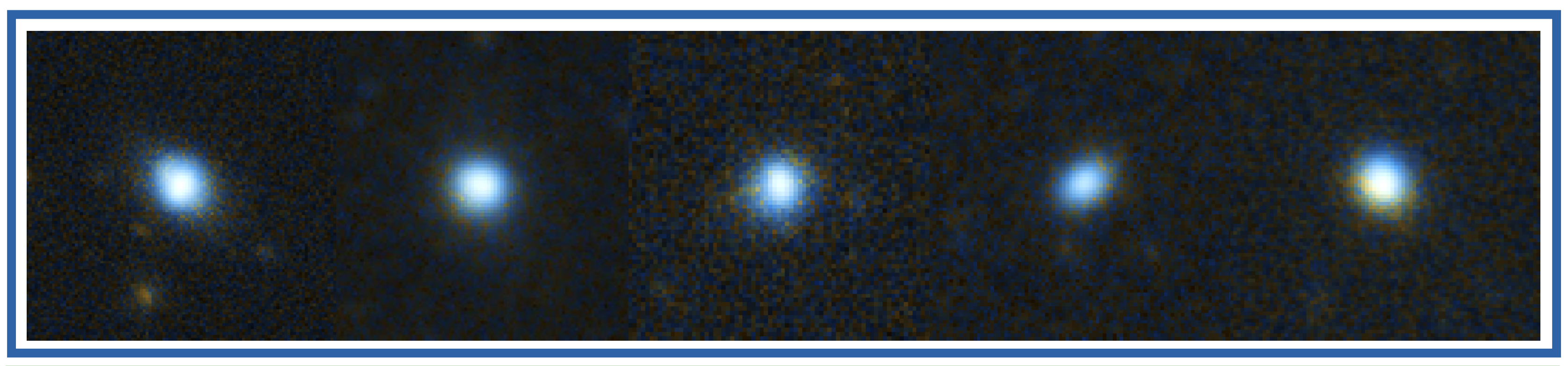


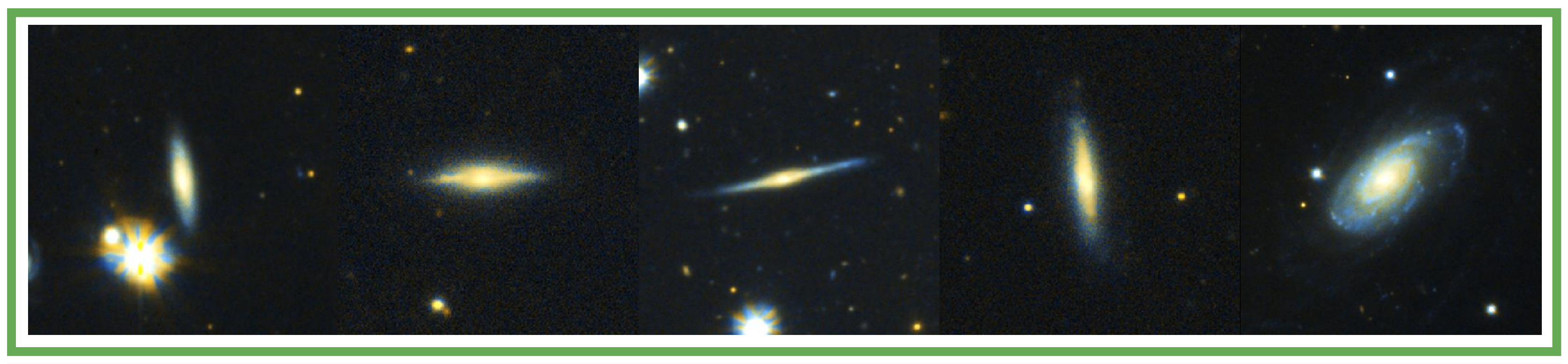
Results: Five Clusters

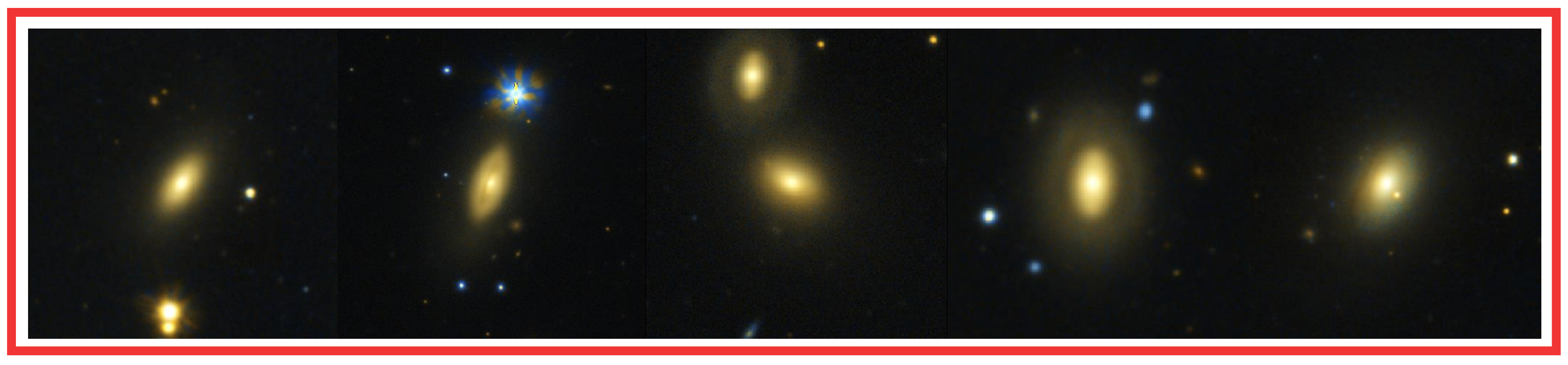
Searching for five clusters enables *k*-means to probe subtler variances in the data. Therefore, it is able to identify **evolved galaxies**, **green valley galaxies**, and **dwarf galaxies**.



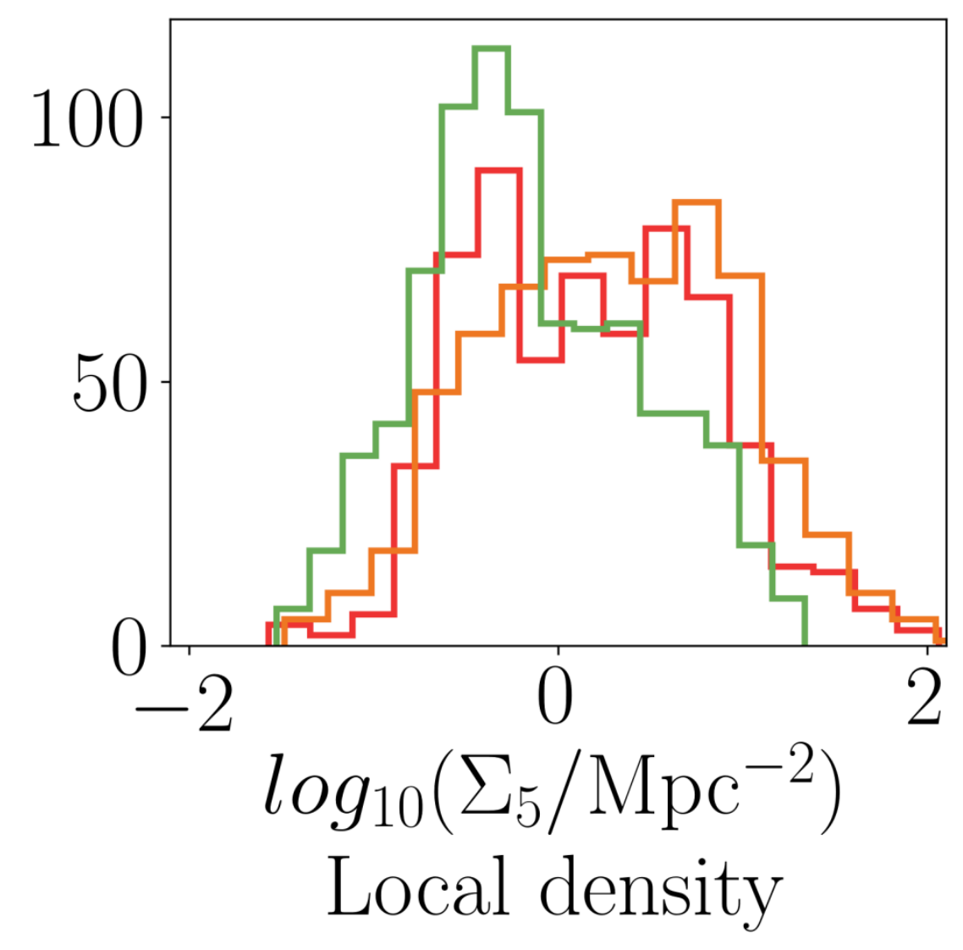
Mass, colour, and star formation still lead the clustering and yield an underlying bimodal structure, but morphological distinctions have now also been made.







Our cluster of **green valley galaxies** occupy uniformly low density environments, while our **other cluster of evolving galaxies** contains galaxies in higher density environments. Our **most evolved cluster** contains a mix of galaxies from both environments, suggesting that galaxies follow different evolutionary pathways depending on their environment, before converging on the red sequence.



References

***k*-means:** MacQueen, J., 1967. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, 1, 281 / Lloyd, S., 1982. IEEE Transactions on Information Theory, 28, 129 / Lisboa, P. J., et al., 2013. BMC Bioinformatics, 14, S8 / **GAMA:** Driver, S.P., et al., 2009. A&G, 50, 5.12 / **Sérsic fitting:** Kelvin, L., et al., 2012. MNRAS, 421, 1007/ **SED fitting:** Driver, S. P., et al., 2016. MNRAS, 450, 1441 / da Cunha, E., et al., 2008. MNRAS, 388, 1595 / Taylor, E. N., et al., 2011. MNRAS, 418, 1587 / **Morphologies:** Kelvin, L., et al., 2014. MNRAS, 439, 1245 / Moffett, A. J., et al., 2016. MNRAS, 457, 1308 / **Environments:** Brough, S., et al., 2013. MNRAS, 435, 2903