Galaxy Morphological Classification via Unsupervised Machine Learning: A Way Forward in the Exascale Era of Big Data Surveys

Ilin Lazar[1], Garreth Martin[2,3], Sugata Kaviraj[1], Alex Hocking[4], Jim Geach[1]

[1]University of Hertfordshire, [2]University of Arizona, [3]Korea Astronomy and Space Science Institute, [4]Microsoft

Abstract

The morphological properties of galaxies are important tracers of the physical processes, e.g. minor/major mergers, gas accretion and tidal interactions, that have shaped their evolution.

We present an unsupervised machine learning algorithm, that utilizes hierarchical clustering and growing neural gas networks to group together survey image patches with similar visual properties, followed by a clustering of objects (e.g. galaxies) that are reconstructed from these patches. We implement the algorithm on the Deep layer of the Hyper Suprime-Cam Subaru-Strategic-Program, to reduce a population of hundreds of thousands of galaxies to a small number (≈100) of morphological clusters, which exhibit high purity. These clusters can then be rapidly benchmarked via visual inspection and classified by morphological type.

Using the morphological clusters obtained by the algorithm, we successfully reproduce many known trends of galaxy properties (e.g. stellar-mass functions, rest-frame colours) as a function of morphological type, which demonstrates the efficacy of the method.

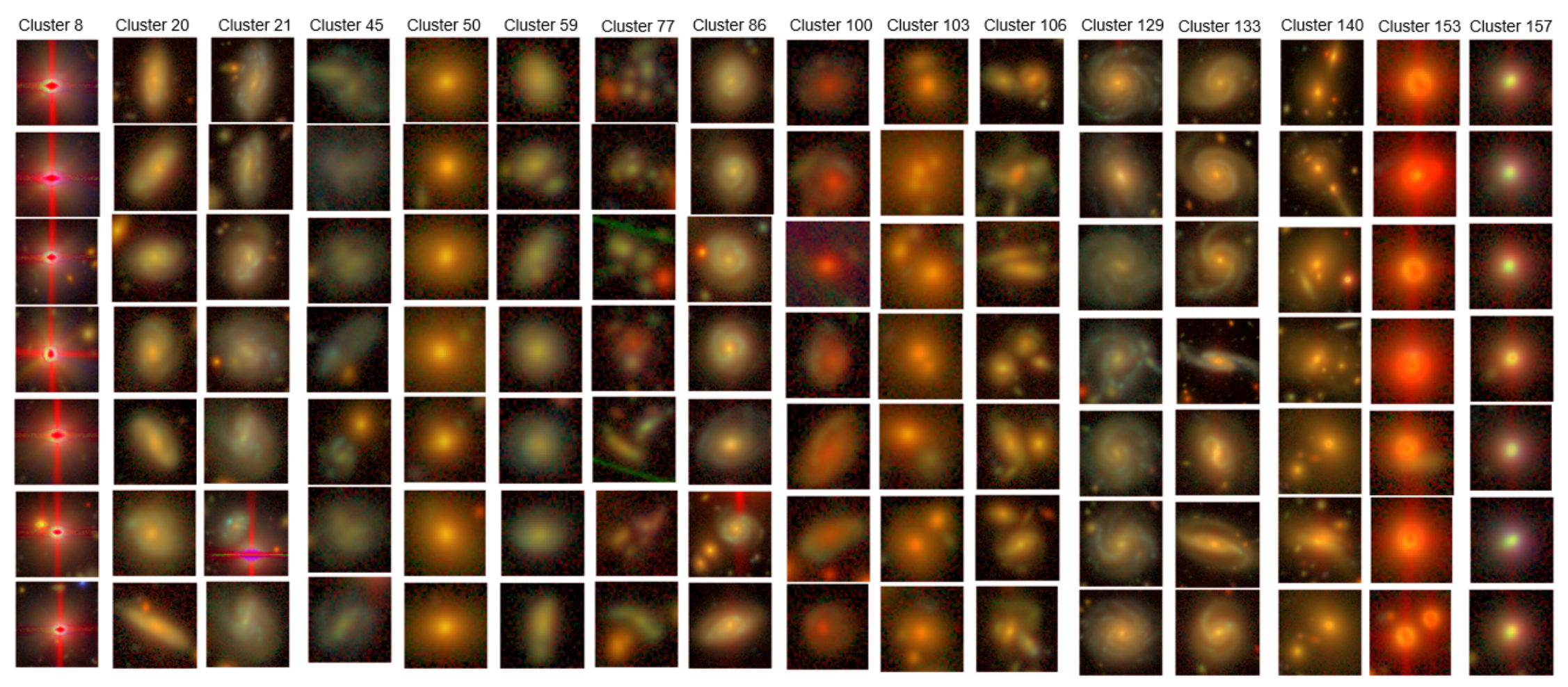
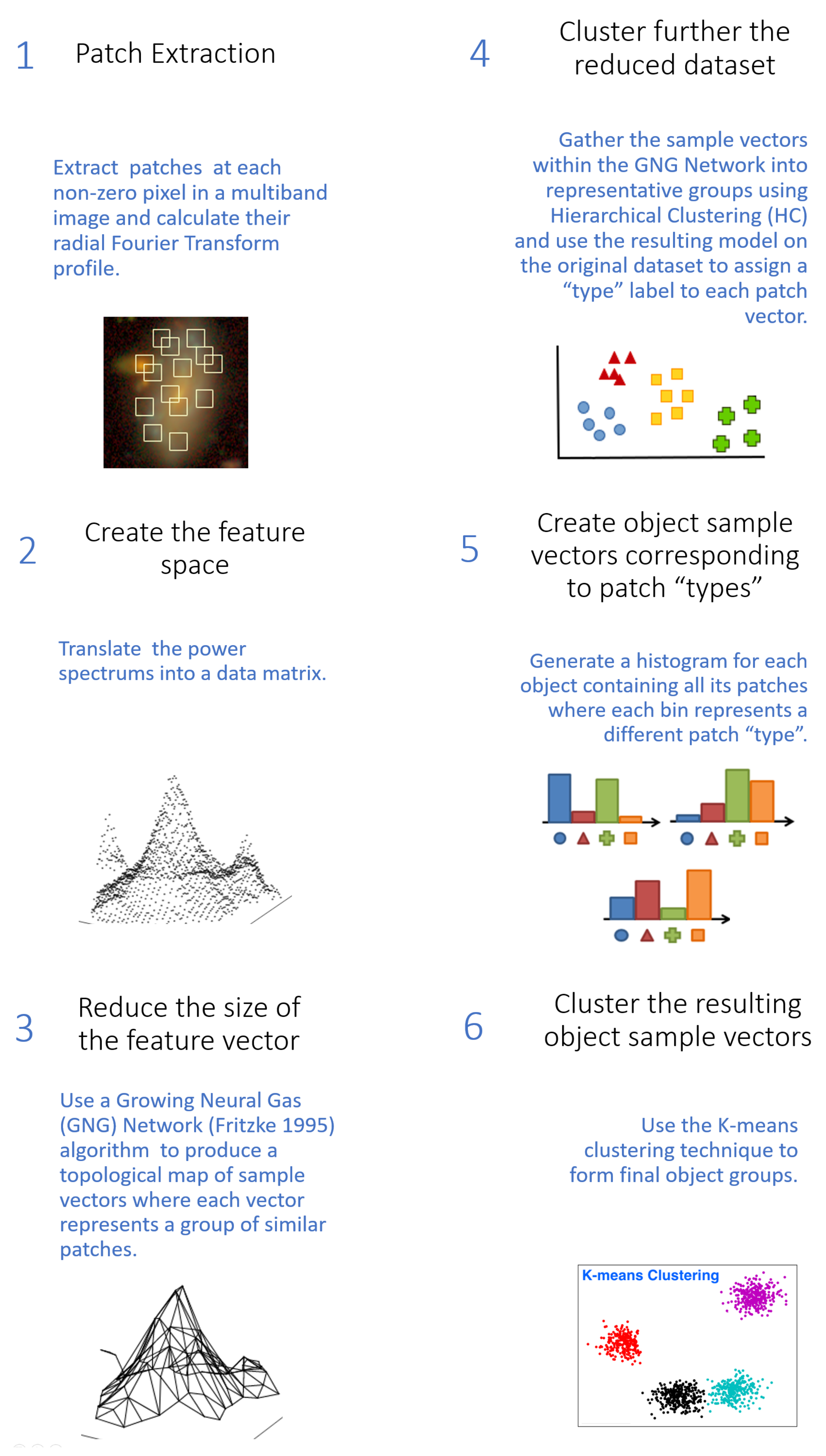


Fig. 1: Representative object samples from some of the morphological clusters. (Martin+2020, MNRAS 491, 1480)

Why use Unsupervised ML?

- Traditionally morphology was measured via parametric (e.g. Sersic profiles) or non-parametric methods (e.g. CAS) methods
- Starting with the arrival of big data surveys there has been an ongoing movement towards machine learning in the scientific community for this classification problem (Martin+2020, Walmsley+2020, Spindler+2020, Cheng+2020)
- Supervised ML methods are calibrated against visual inspection (Lintott+2011) which is highly accurate but time consuming
- Big Data Surveys like LSST or JWST will produce tens of billions of objects. Hence we will need ML but also a significant amount of human resources to label the training sets (for Supervised ML)
- Unsupervised ML does not require training on labelled data

How does it work?



1 Patch Extraction

Extract patches at each non-zero pixel in a multiband image and calculate their radial Fourier Transform profile.

2 Create the feature space

Translate the power spectrums into a data matrix.

3 Reduce the size of the feature vector

Use a Growing Neural Gas (GNG) Network (Fritzke 1995) algorithm to produce a topological map of sample vectors where each vector represents a group of similar patches.

4 Cluster further the reduced dataset

Gather the sample vectors within the GNG Network into representative groups using Hierarchical Clustering (HC) and use the resulting model on the original dataset to assign a "type" label to each patch vector.

5 Create object sample vectors corresponding to patch "types"

Generate a histogram for each object containing all its patches where each bin represents a different patch "type".

6 Cluster the resulting object sample vectors

Use the K-means clustering technique to form final object groups.

Fig. 2: Algorithm schematic. (Martin+2020, MNRAS 491, 1480)

Labelling by morphological type of each object group from the final K-means model is done by inspecting visually a representative sample (see Figure 1) of objects from each cluster.

Results

- A wide variety of galaxy morphological types are being clustered correctly by the algorithm from general (e.g. ellipticals/spirals) to more detailed (e.g. clumpy discs/high-z mergers) morphologies

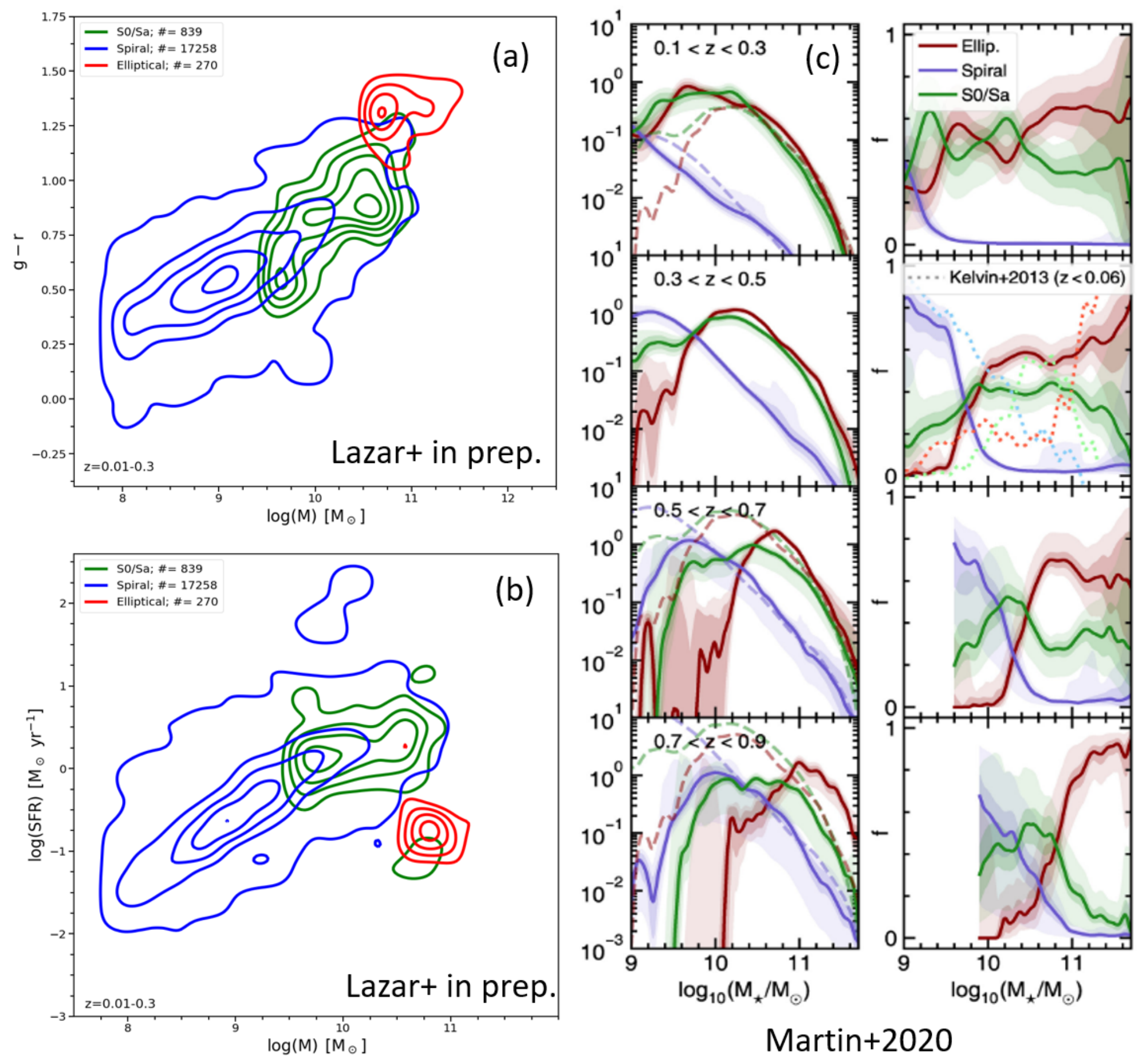


Fig. 3: (a) Color-mass diagram for spirals, ellipticals and S0/Sa types. (b) SFR-mass diagram.(c) Stellar mass distribution as a function of redshift.

- Color-mass bimodality is being retrieved, showing the "red sequence" (populated by ellipticals) and the "blue cloud" (populated by spirals) both connected by the S0/Sa "green vallley". (e.g. Visvanathan+1981)
- The majority of spirals retain increased star formation as opposed to ellipticals which increase in fraction as redshift decreases, both populations being constrained by a low-high mass bimodality over cosmic time. (e.g. Conselice+2008)

Future Plans

- Release morphology catalogue for HSC DR3 (Lazar+ in prep)
- Use the method on the upcoming big data surveys (e.g. LSST, EUCLID, JWST, SKA)

References

Martin+ 2020, 2020 MNRAS 491 1480; Walmsley+ 2020, 2020 MNRAS 491 1554W; Spindler+ 2020, arXiv 2009 08470; Cheng+ 2020, arXiv 1908 03610; Lintott+ 2011, 2011 MNRAS 410..166L; Visvanathan+ 1981, 1981 AA 100L 20V; Fritzke+ 1995, MIT Press, p.625-632.; Conselice+ 2008 2008 MNRAS 386-909C.