Galaxy Morphological Classification via Unsupervised Machine Learning in the Big Data Era led by JWST, EUCLID, LSST and SKA

Ilin Lazar (Univ. of Hertfordshire)

with Garreth Martin (KASI), Sugata Kaviraj (Hertfordshire), Alex Hocking (Microsoft) and Jim Geach (Hertfordshire)

Morphology in the Era of Big Data Surveys

- Traditionally morphology was measured via parametric (e.g. Sersic profiles) or non-parametric (e.g. CAS) methods
- Recent progress involve more accurate measures with the aid of machine learning (ML) (Huertas-Company+2019, Walmsley+ 2020,2023, Cheng+ 2020, Sarmiento+2021, Martin+2020)
- Supervised ML methods are calibrated against visual inspection (Lintott+ 2011) which is highly accurate but time consuming (without transfer learning)
- Big Data Surveys like LSST or JWST will produce tens of billions of objects – will need ML (fast data processing) along with visual inspection (for Supervised ML)
- Unsupervised ML does not require training on labelled data (built to minimize the human intervention)

The Algorithm

Hocking+2018, MNRAS 473, 1108
Martin+2020, MNRAS 491, 1480

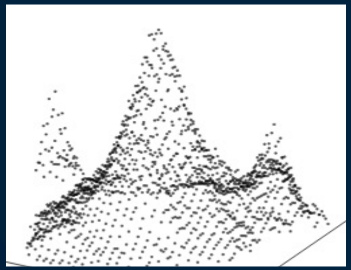
Lazar+ in prep.

1 Patch Extraction

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

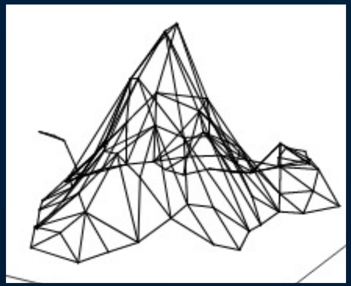
2 Create the feature space

Translate the power spectrums into a data matrix



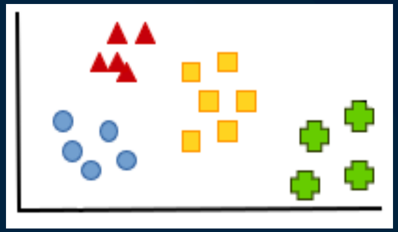
3 Reduce the size of the feature vector

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.



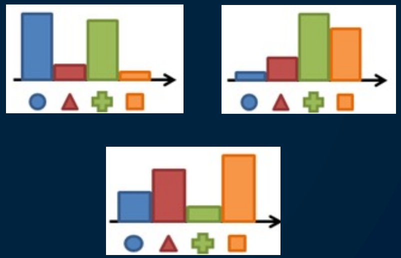
4 Cluster further the reduced dataset

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 .



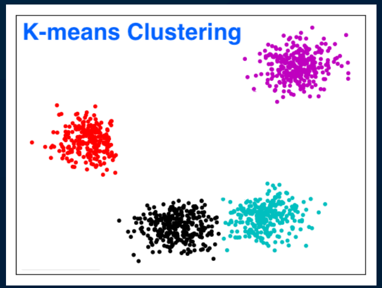
5 Create object sample vectors corresponding to patch “types”

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

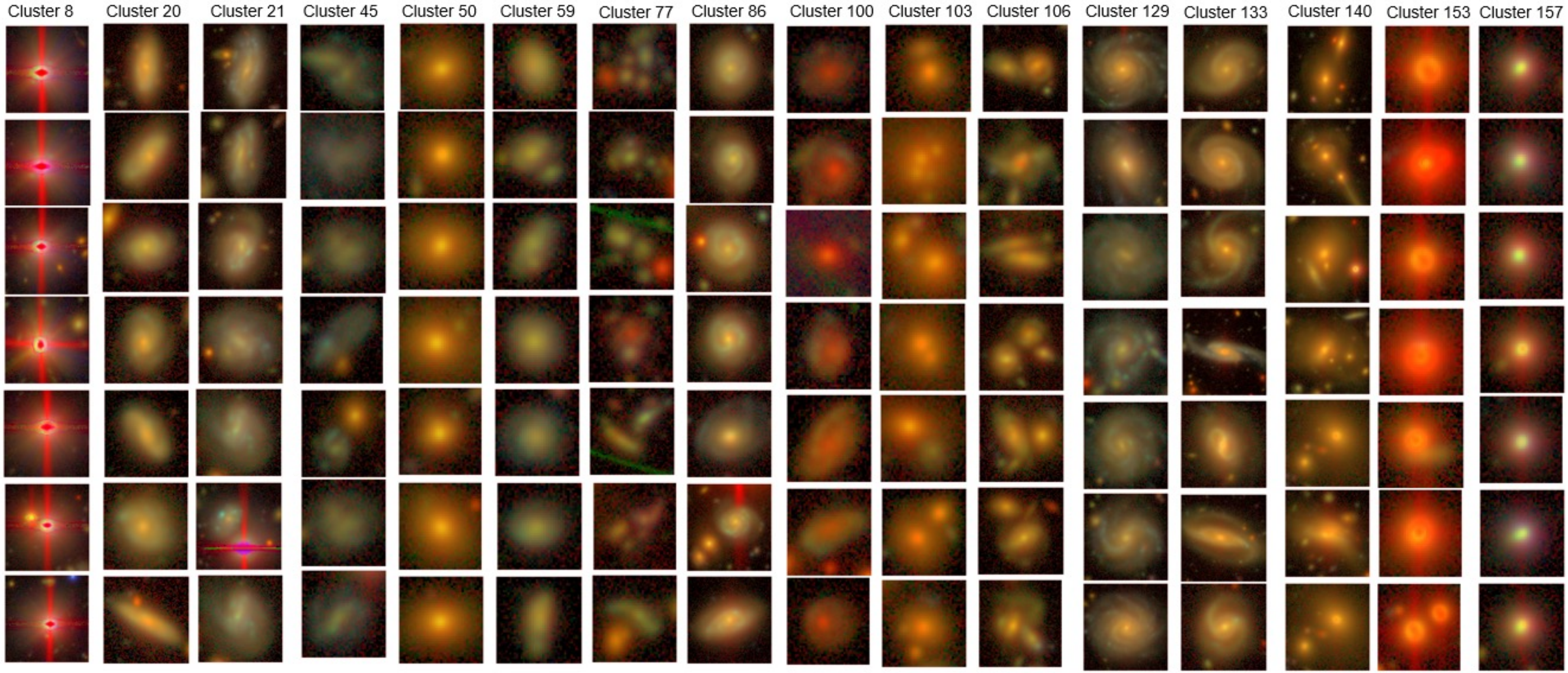
Use the K-means clustering technique to form final object groups .



Credits: Martin+ 2020

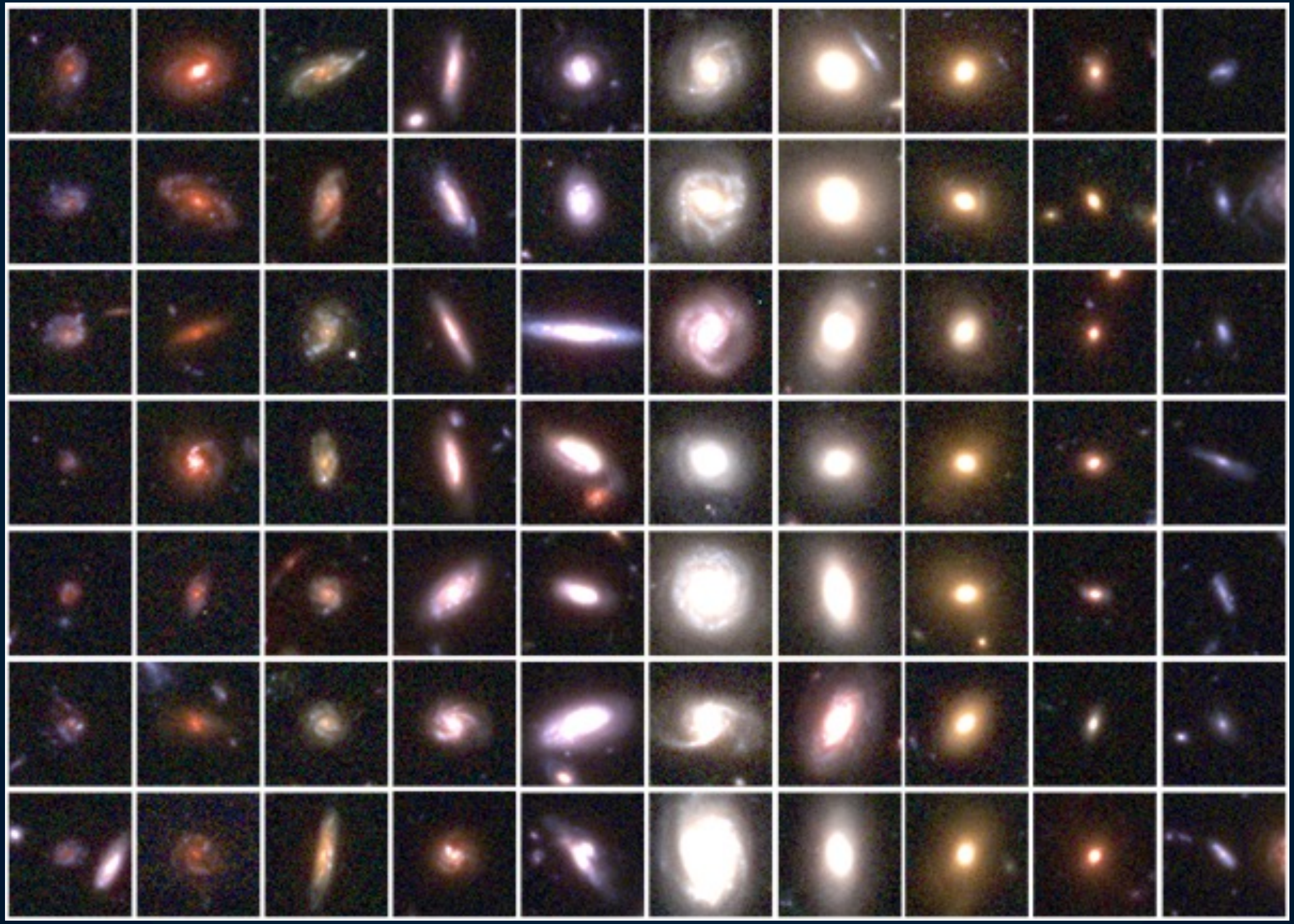
Automatic classification of HSC data

Martin+2020, MNRAS 491, 1480



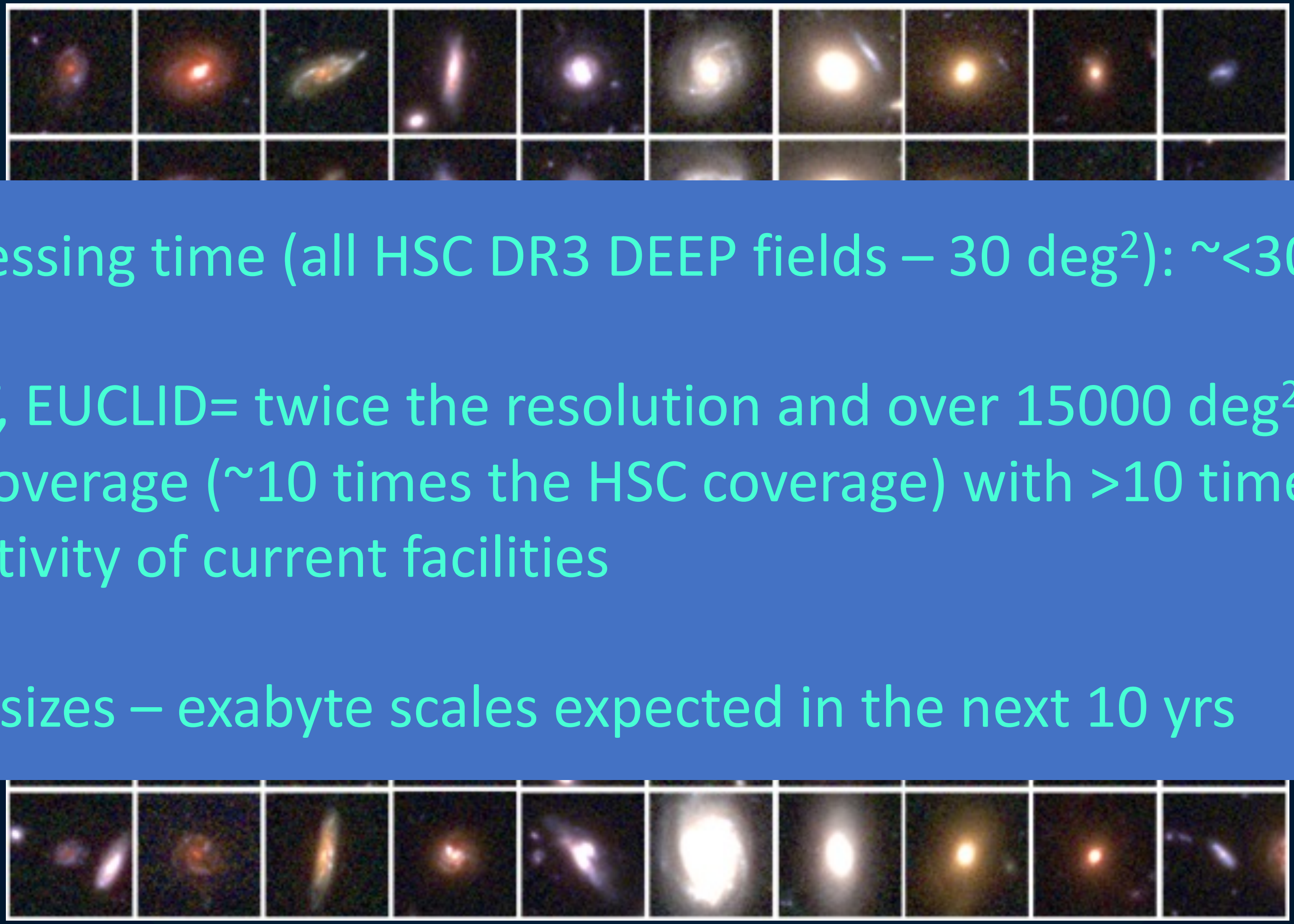
Automatic classification of CANDELS data

Hocking+2018



Automatic classification of CANDELS data

Hocking+2018



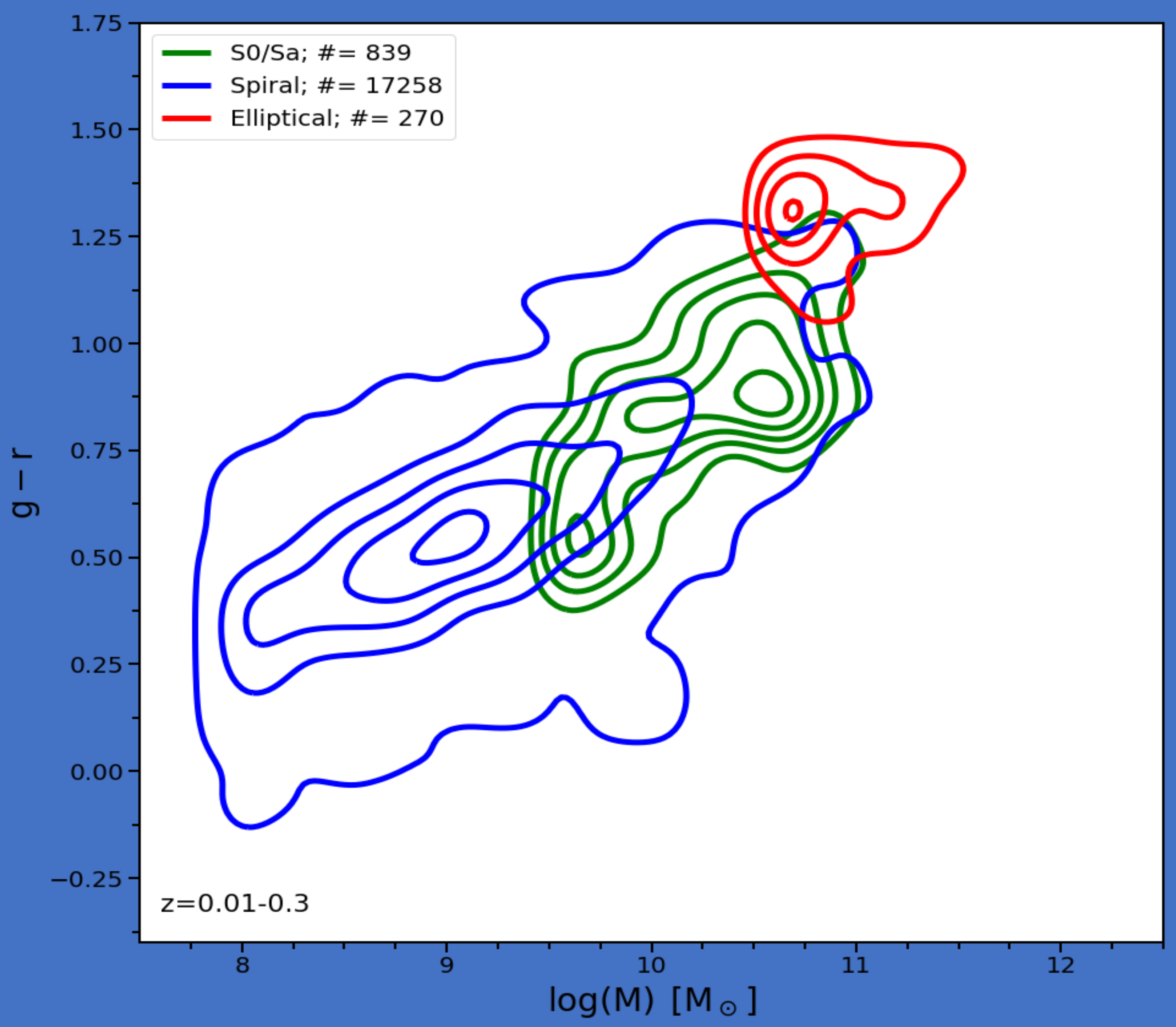
Processing time (all HSC DR3 DEEP fields – 30 deg^2): ~<30 hrs

JWST, EUCLID= twice the resolution and over 15000 deg^2 of sky coverage (~10 times the HSC coverage) with >10 times the sensitivity of current facilities

Data sizes – exabyte scales expected in the next 10 yrs

Results

Color-mass bimodality showing the 'red sequence' and 'blue cloud' is present (e.g. Baum+ 1959, Visvanathan+ 1981) with the S0 'green valley' connecting the two groups

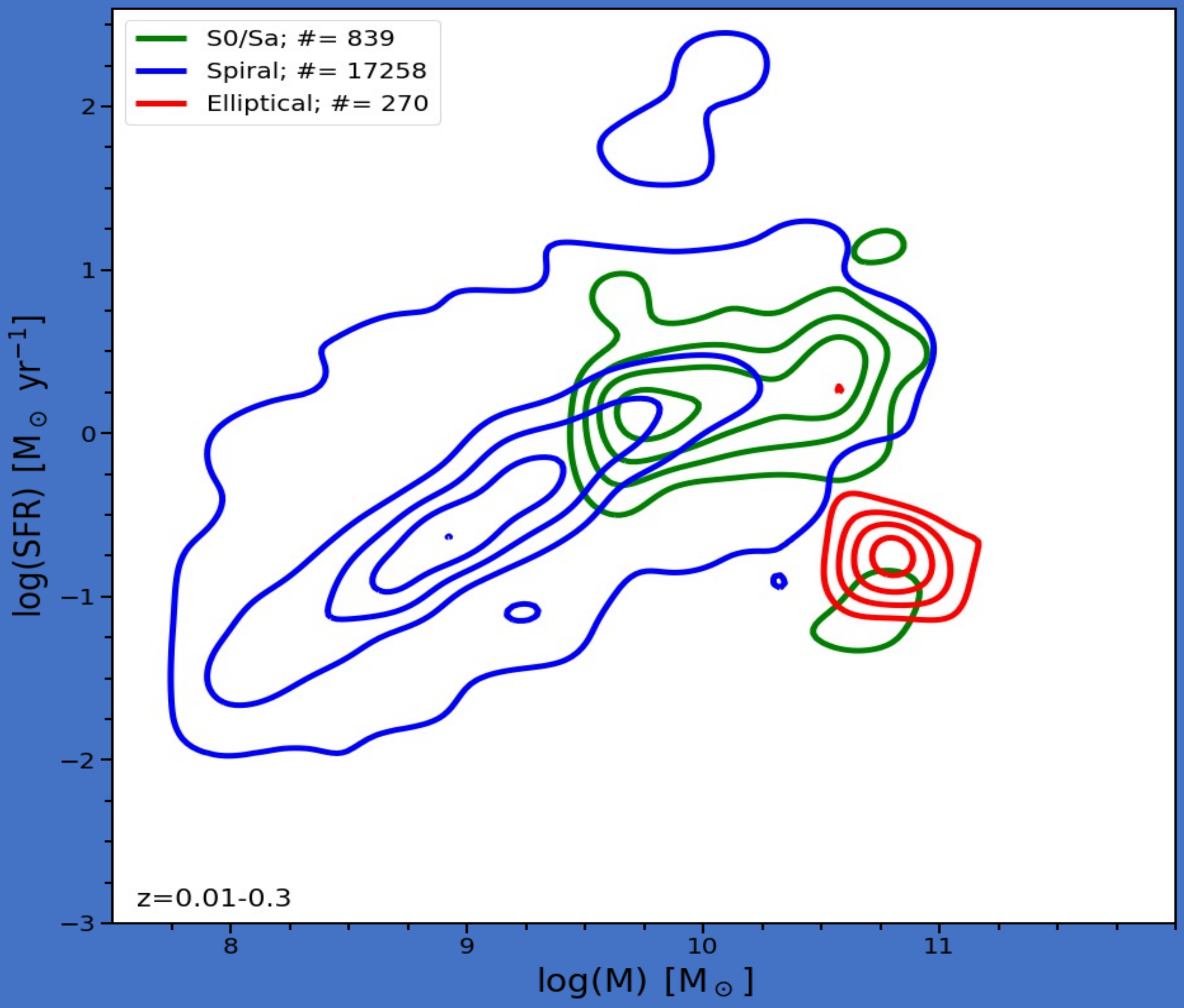


Data origin: HSC-SSP UDEEP DR2 (Lazar+ in prep.)

Results

Color-mass bimodality showing the 'red sequence' and 'blue cloud' is present (e.g. Baum+ 1959, Visvanathan+ 1981) with the S0 'green valley' connecting the two groups

SFR-mass relation: majority of spirals retain high levels of star formation as opposed to ellipticals (e.g. Thronson, Bally & Hacking 1989, Pogge & Eskridge 1993)



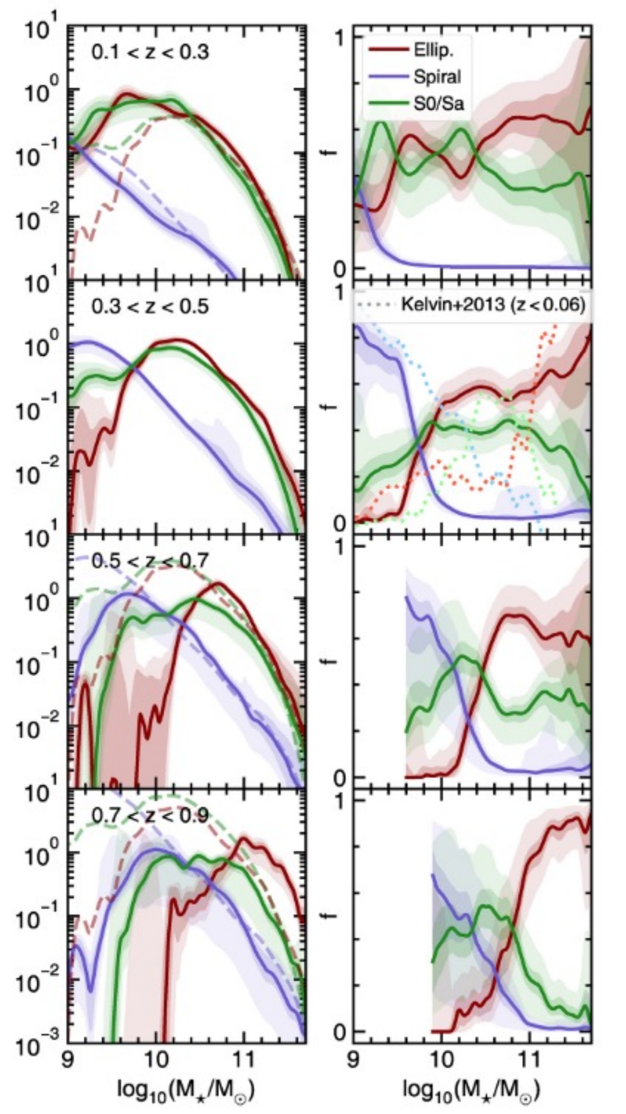
Data origin: HSC-SSP UDEEP DR2 (Lazar+ in prep.)

Results

Color-mass bimodality showing the 'red sequence' and 'blue cloud' is present (e.g. Baum+ 1959, Visvanathan+ 1981) with the S0 'green valley' connecting the two groups

SFR-mass relation: majority of spirals retain high levels of star formation as opposed to ellipticals (e.g. Thronson, Bally & Hacking 1989, Pogge & Eskridge 1993)

Bimodal stellar mass distribution (Vulcani+ 2011, Conselice+ 2008, Kelvin+ 2014)



Data origin: HSC-SSP UDEEP DR1 (Martin+ 2020)

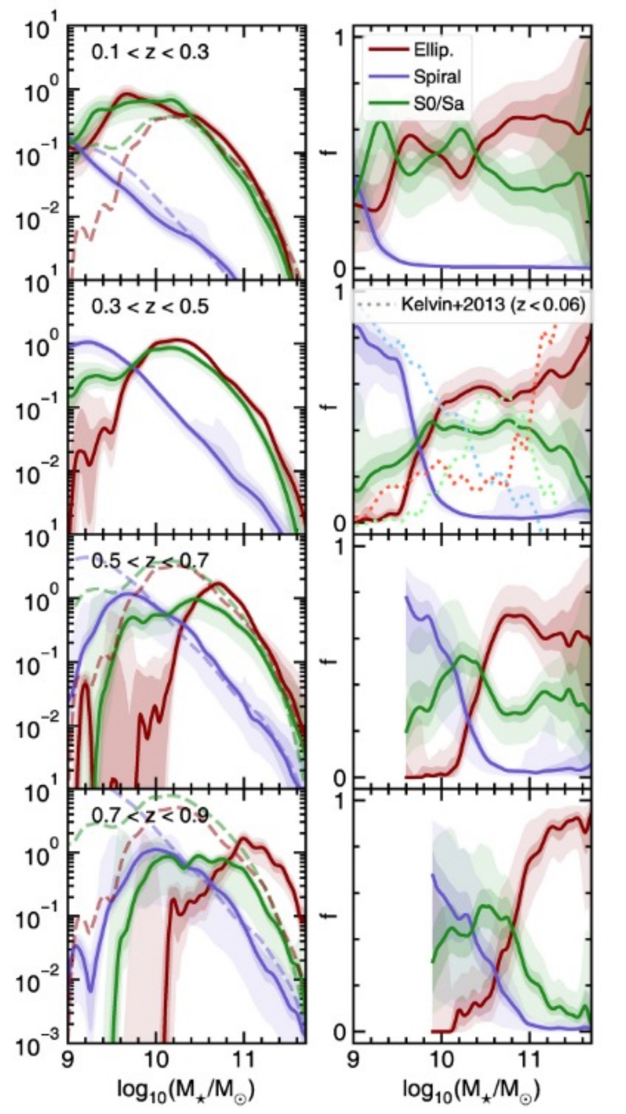
Results

Color-mass bimodality showing the ‘red sequence’ and ‘blue cloud’ is present (e.g. Baum+ 1959, Visvanathan+ 1981) with the S0 ‘green valley’ connecting the two groups

SFR-mass relation: majority of spirals retain high levels of star formation as opposed to ellipticals (e.g. Thronson, Bally & Hacking 1989, Pogge & Eskridge 1993)

Bimodal stellar mass distribution (Vulcani+ 2011, Conselice+ 2008, Kelvin+ 2014)

All known trends in morphology from the literature are reproduced well



Data origin: HSC-SSP UDEEP DR1 (Martin+ 2020)

Automated identification of rare/peculiar objects: low mass Blue Ellipticals

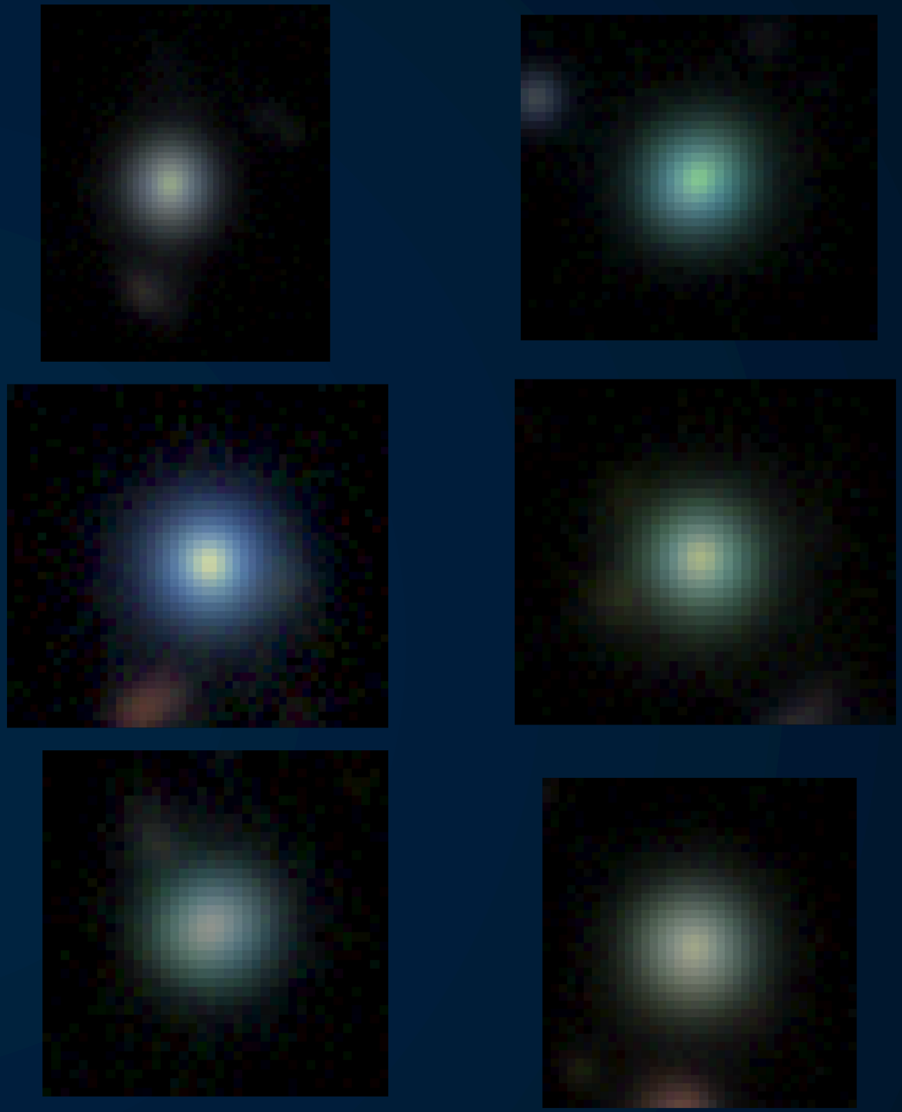
Physical properties – same as spirals
Morphology - same as ellipticals

A sample of 59 Blue ellipticals confirmed by spectroscopic redshifts from an original sample of approx. 100

- Blue Ellipticals – great debate in the literature upon origin of SFR activity (merger or sustained gas accretion events) (Schawinski et al. 2006, 2007,2009, Fukugita et al. 2004, Yi et al. 2005, Kaviraj 2014)

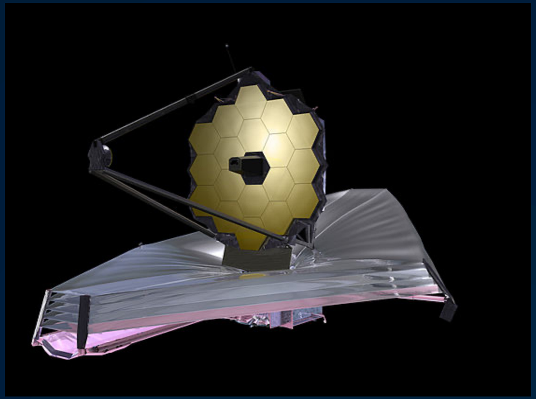
The majority do not show tidal tails – sign for secular accretion evolution history

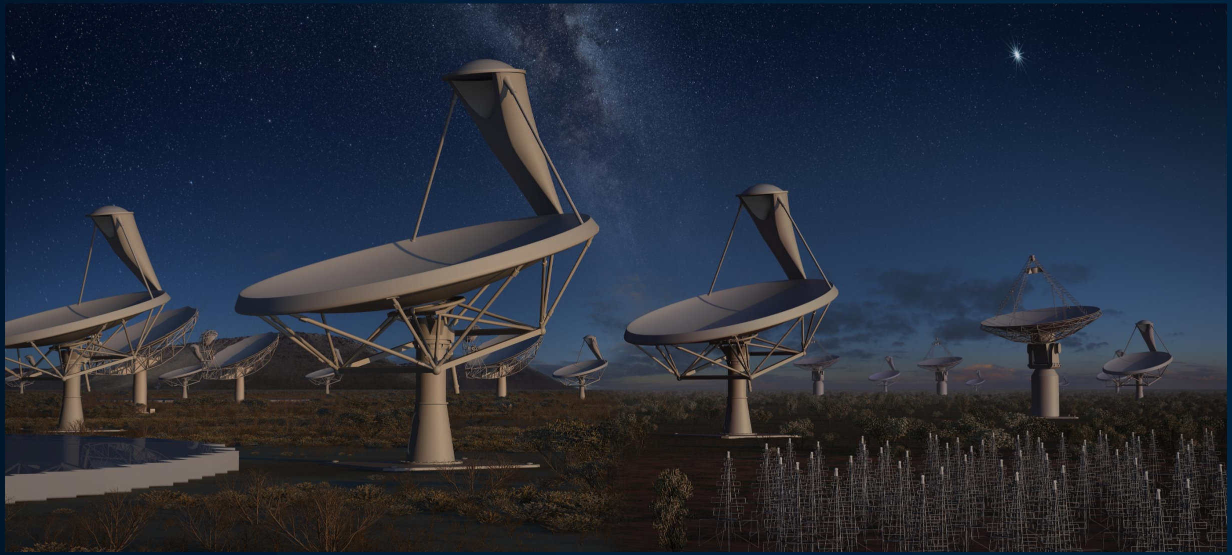
More details – Lazar et al. 2023



Data origin: HSC-SSP UDEEP DR3

Future Plans





Credits: LSST Project/NSF/AURA, SKA Project Dev. Office, NOAO/AURA/NSF, NASA

Contact details: i.lazar@herts.ac.uk

- **Release morphology catalogue for HSC DR3**
- **Use the method on upcoming large data volumes from big data surveys (e.g. EUCLID, JWST, SKA)**
- **Add more bands to the feature space (e.g. IR, UV)**
- **Make the code more accessible to the scientific community**