Automated morphological classification and discovery for LSST using unsupervised machine learning techniques

Garreth Martin – University of Arizona / KASI

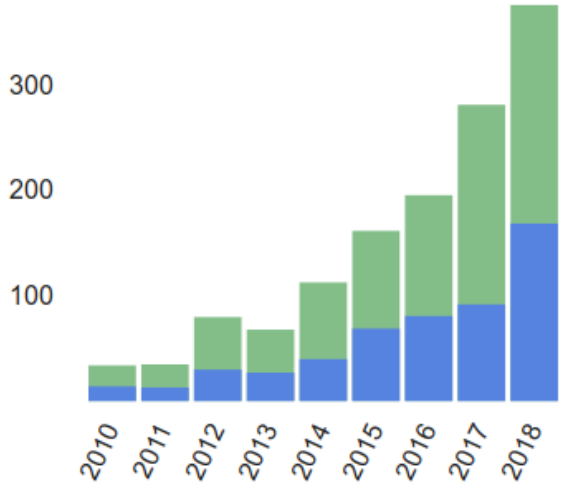
A brief overview of machine learning in astronomy

- Machine learning in astronomy is fairly old (e.g. **Lahav+ 1995**), but has recently become much more commonplace

A wide range of machine learning solutions have been applied to problems in astronomy, predominantly based on supervised techniques

- **Huertas-Company et al. (2015)** convolutional neural networks **Ostrovski et al. (2017)** supervised Gaussian mixture models, **Schawinski et al. (2017)** generative adversarial networks, **Goulding et al. (2018)** random forest classifier **Siudek et al. (2018)** unsupervised Fisher expectation-maximisation

- **Some specific to LSST presented this week: Roussi** Siamese networks (semi-supervised learning) to locate lenses (*and general discovery of rare objects*), **Balaji** Deep learning applied to detect stellar streams



ADS abstracts referencing “machine learning” by year

Morphological classification challenges for LSST

LSST data volume will be unprecedented in astronomy
– Cadence of ~**3 days**

– ~ 30 galaxies / arcmin^2 (~**20 billion galaxies**/17 billion stars in total)

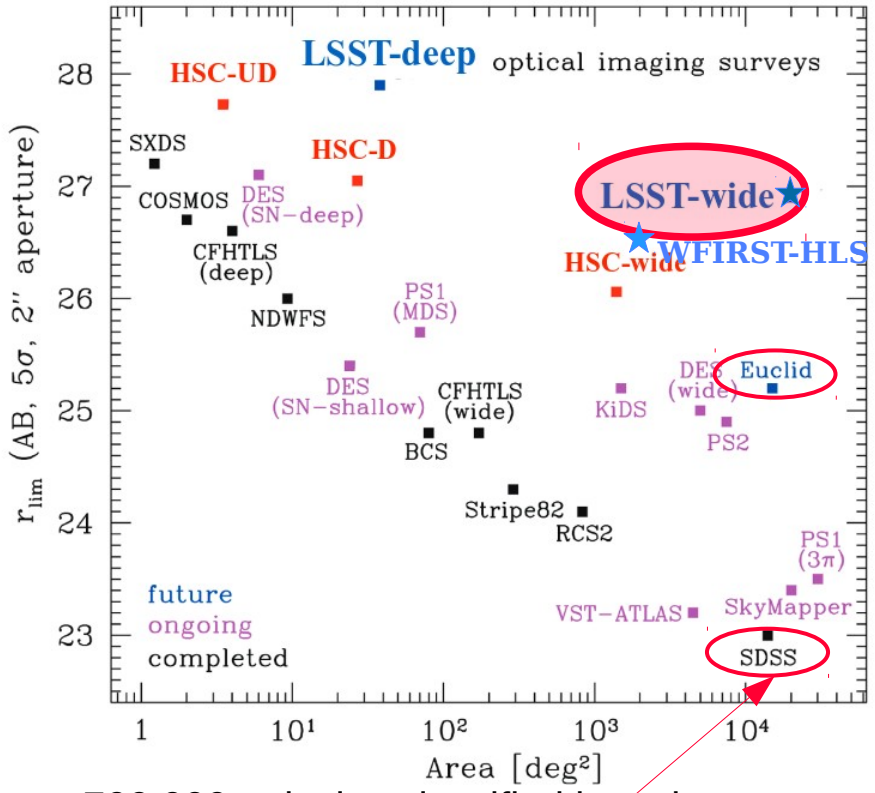
– Full reprocessing of survey data annually, but much more often for some applications

How do we classify all these objects?

– Rapidly changing datasets mean we may need to re-classify co-added data between data releases

– Repeated construction of **unbiased training sets** for high cadence (rapidly changing) data will be difficult

– The large area of LSST will allow the construction of samples of rare/faint types of object, but these object will not have robust training sets available



700,000 galaxies classified by volunteers over 4 years **(Lintott+2008)**

Solutions?

– Classifications based on raw image data using **citizen science** (e.g. galaxy zoo) and **machine learning** based techniques have produced high quality classifications in the past

– But as **data sets continue to grow**, human classification will become less and less viable

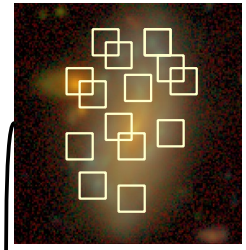
– Some hope that **citizen science combined with machine learning** e.g. **Beck et al. (2018)** can mitigate this, but very large data volumes will still be challenging

– Machine learning techniques are currently the only realistic solution, but high survey cadence and rapidly changing data mean **there are still challenges for supervised techniques**

– **Unsupervised** methods, which **do not require training sets, allow for discovery and do not introduce human bias** in the training stage represent a promising alternative

Application of an unsupervised technique

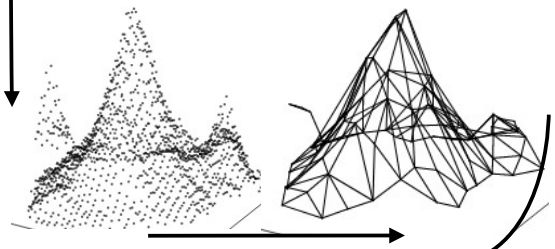
Hocking et al. (2018)



Convert the survey images into a data matrix

– Extract patches at each non-zero pixel in a multi-band image

– The radial power spectrum of a patch produces a rotationally invariant representation of the patch

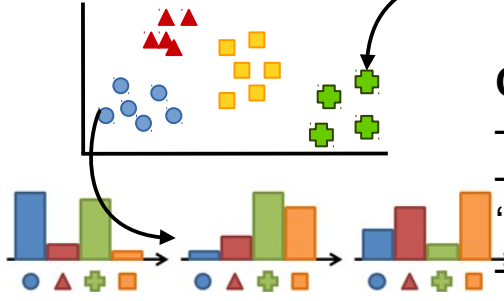


Use GNG and HC to produce a condensed version of the original data set

– Using the output patches, iteratively fit the data using growing neural gas to produce a graph made up of nodes

– Each node in the graph represents a group of similar patches

– By applying hierarchical clustering, we can further reduce the number of groups by reducing them to similar ‘types’ of patches



Create object sample vectors corresponding to patch ‘types’

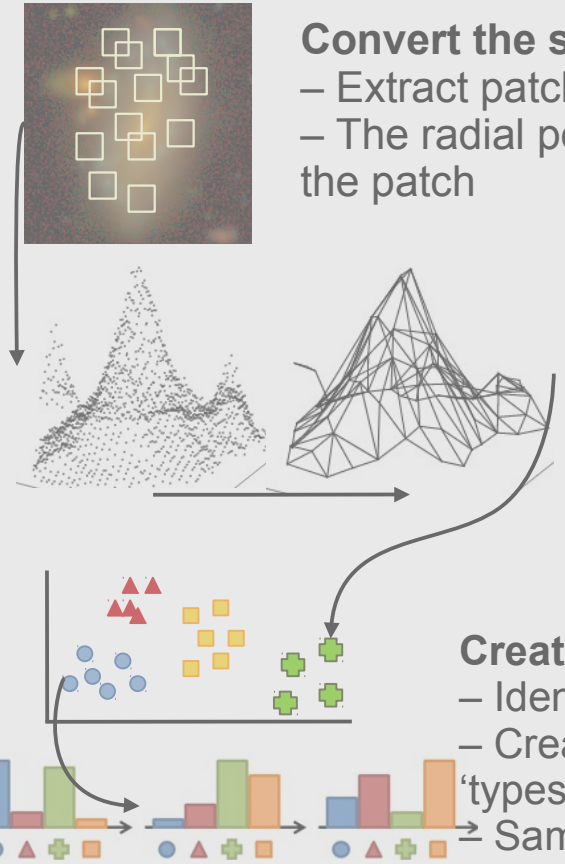
– Identify objects using connected component labelling (or existing segmentation map)

– Create a sample vector for each object, represented by a histogram of the different ‘types’ of patches they are formed from

– Sample vectors are weighted by tf*idf (term frequency-inverse document frequency)

Application of an unsupervised technique

Hocking et al. (2018)



Convert the survey images into a data matrix

– Extract patches

– The radial pow

the patch

Condensed version:

Use clustering techniques (**growing neural gas & hierarchical clustering**) to create a library of pixel ‘types’ based on colour, intensity and ‘texture’

Produce histogram descriptions (‘**feature vector**’) of objects that describe the frequency of each pixel type in that object

Create c

– Identify

– Create

‘types’ o

→ Sample vectors are weighted by tf*idf (term frequency-inverse document frequency)

Application of an unsupervised technique

– Alternative feature extraction approaches may also be effective e.g. using neural networks to select features instead of engineering the features

– The modeling and clustering steps only need to be performed once for a **representative sample of the data (a few thousand objects)**.

– Using the same model we can then apply the algorithm to unseen data.

– **~30- 40 milliseconds per pixel** using a single thread of execution for the first pass modelling / classification step.

– Once the first pass is completed only **1.5 milliseconds per pixel for classification** of subsequent similar data (e.g. more patches of the same survey).

– **SDSS in a few days** on a HPC cluster. **Not yet optimised.**

Application of an unsupervised technique

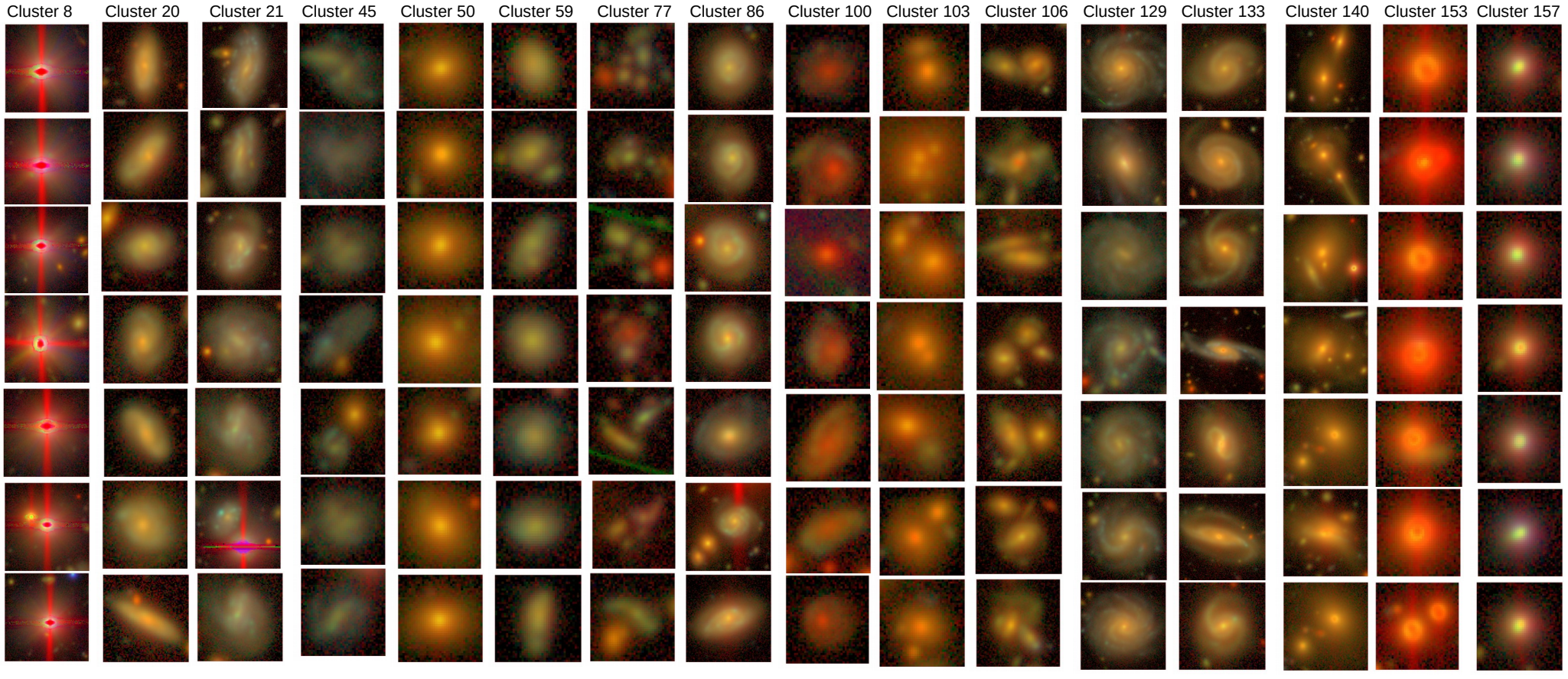
Using each object's feature vector we can:

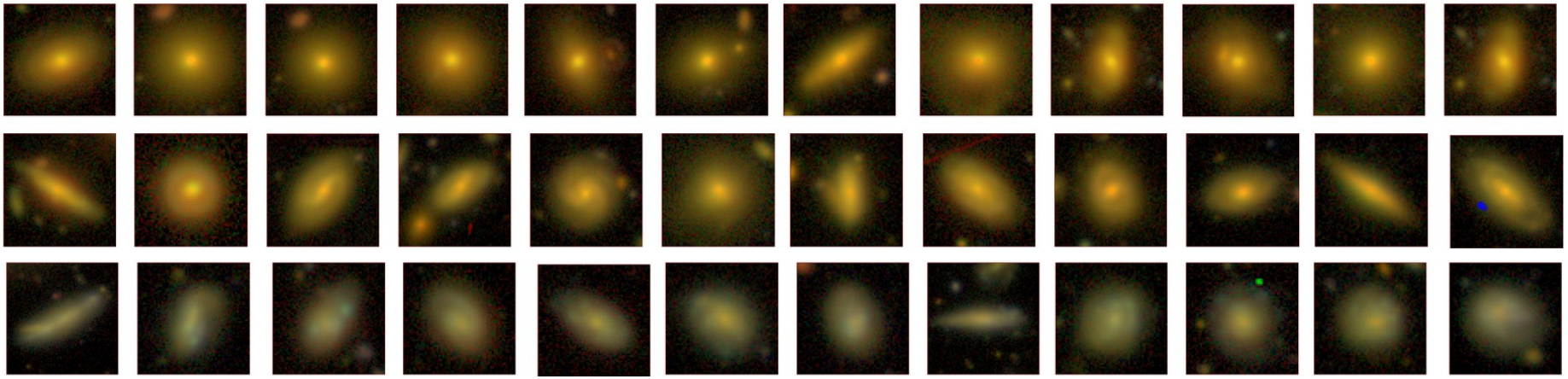
a) **Automatically group objects by using an additional clustering step**

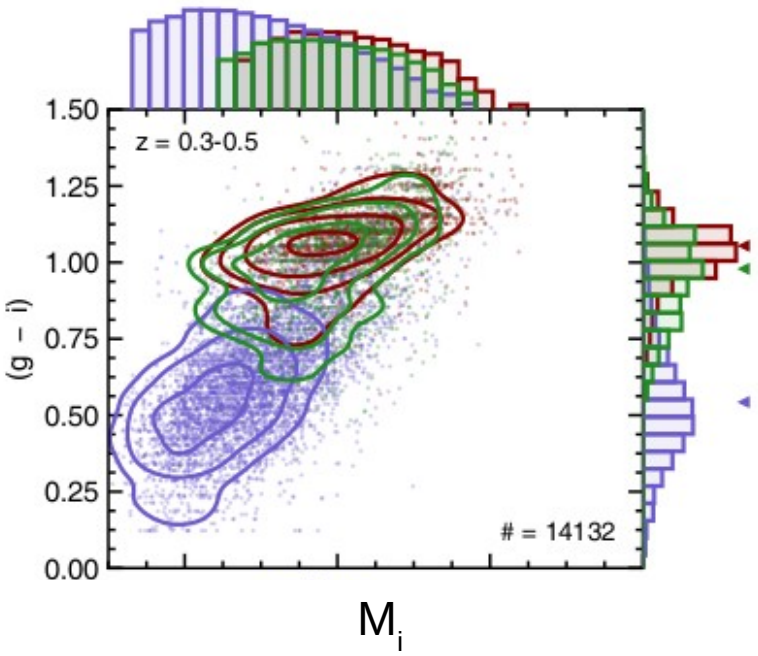
– Can be applied to the the whole set of object sample vectors
– Identify **arbitrary groups** of objects with similar feature vectors
– Groups of like objects have **no semantic labels**

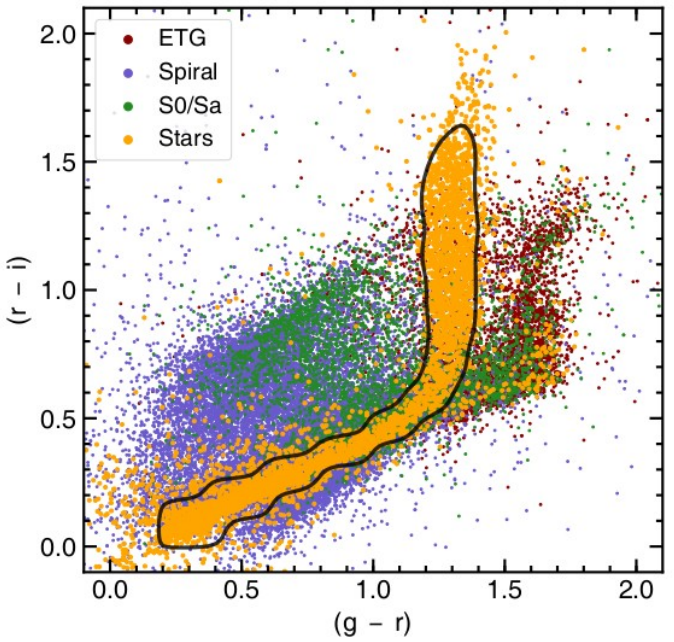
Automatic classification of HSC-SSP *ultra deep DR1* data (LSST-like)

Martin et al. (in press)

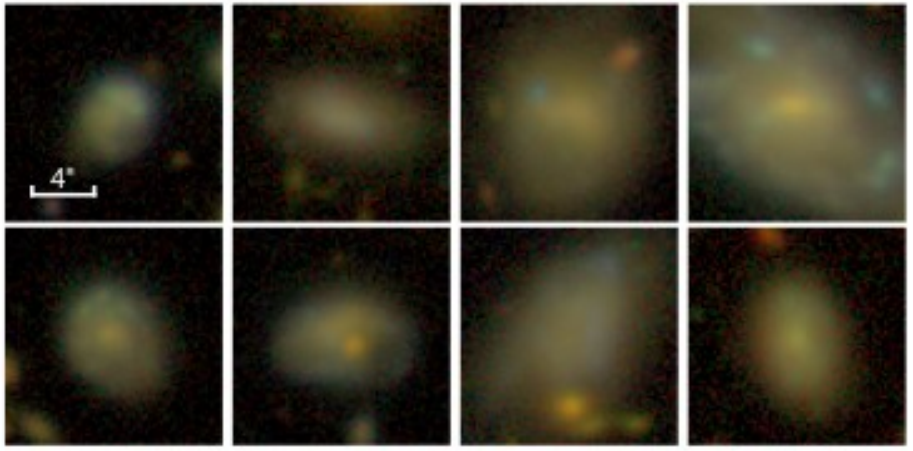




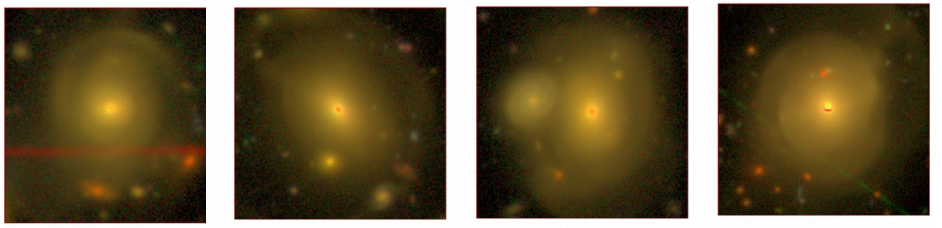




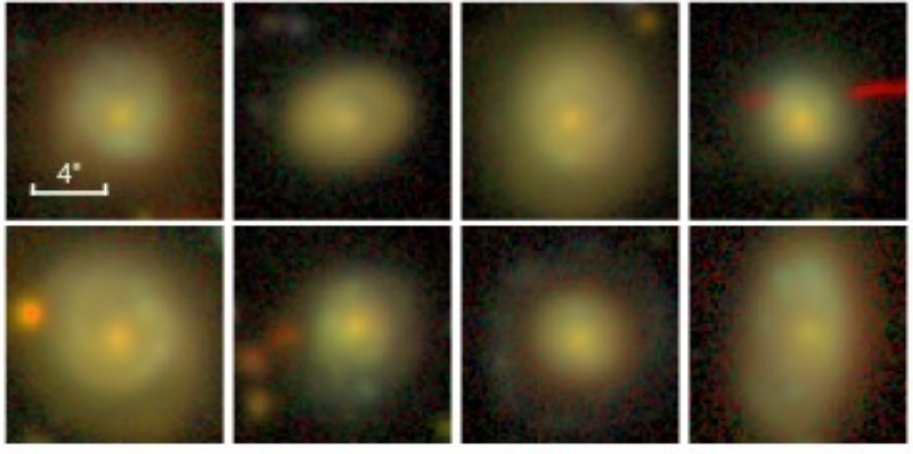
Classifications based on **visual inspection of a small subset** of each group produce expected relations



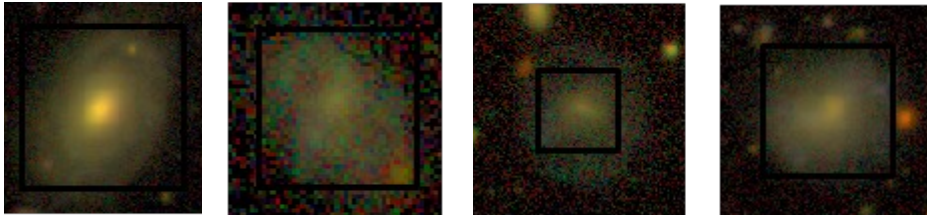
Clumpy discs



Shells



Rings / accretion events



LSB discs

Some more examples of individual clusters featuring rare/specific types of object

(i.e. groups of object with similar feature vectors)

Unsupervised machine learning

Using each object's sample vector we can:

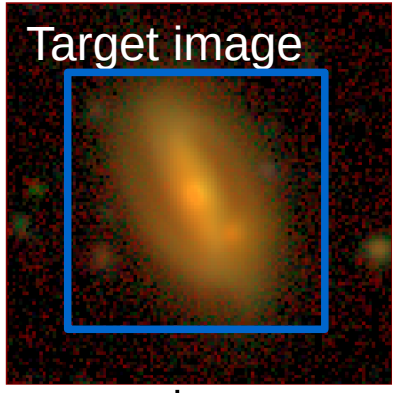
a) **Automatically group objects by using an additional clustering step**

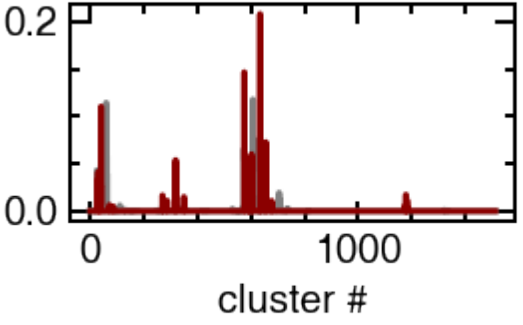
– Can be applied to the the whole set of object sample vectors
– Identify **arbitrary groups** of objects with similar feature vectors
– Groups of like objects have **no semantic labels**

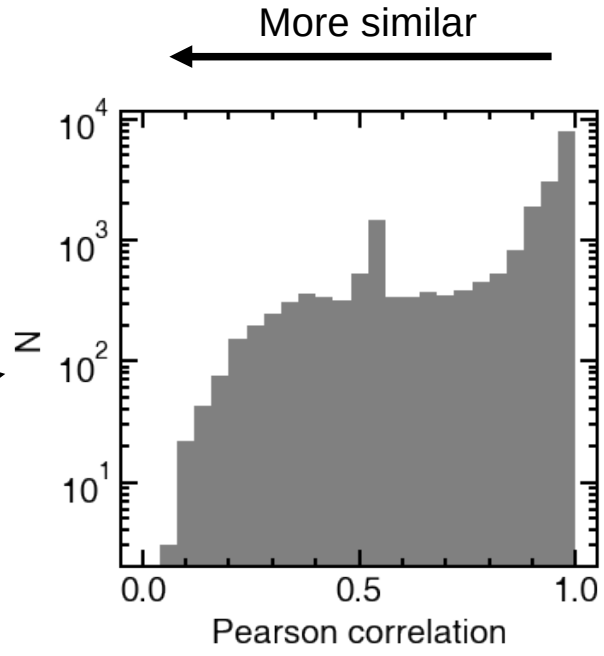
b) **Discover objects similar to a given target object**

– Find the objects whose sample vectors are closest to the target sample vector
– Use some distance measure (Pearson correlation coefficient, Cosine distance, etc)

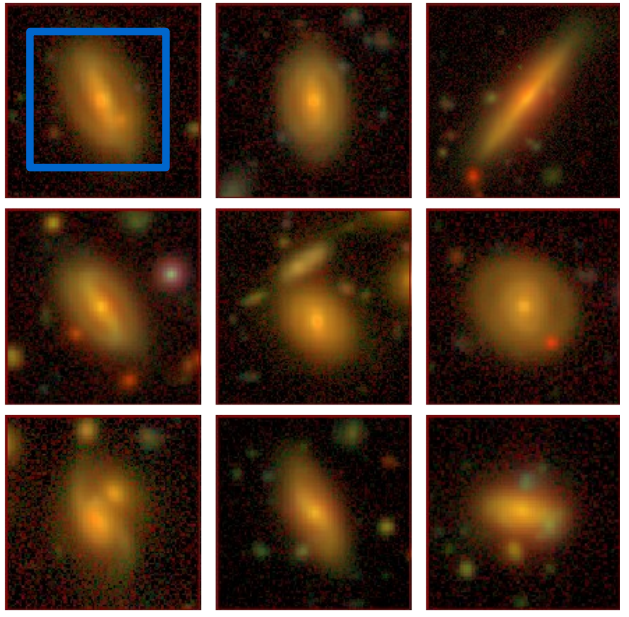
Finding like objects by similarity score



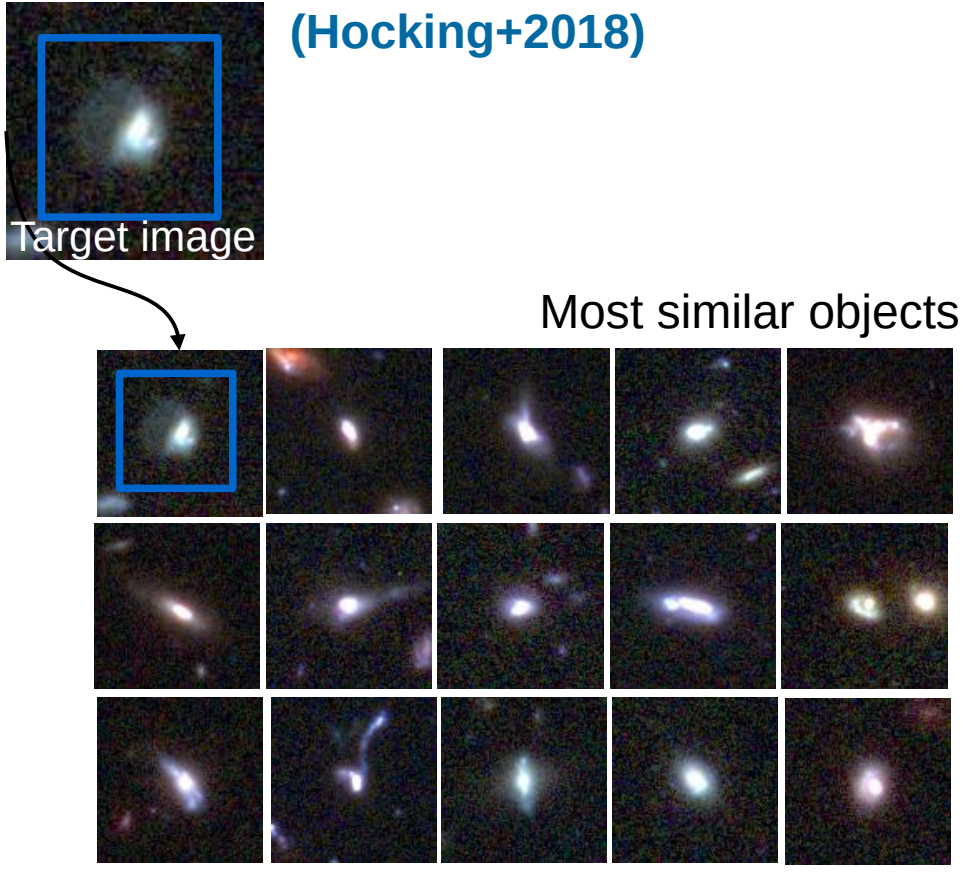




Most similar objects



Finding like objects by similarity score



(Hocking+2018)

Searching for specific features

Similarity searches allow us to auto-detect e.g. tidal features

Searching for the nearest feature vectors allows us to produce a library of similar objects

Summary

– As datasets increase in size, **new solutions need to be developed for classifying objects** from raw data. The data volume expected from **LSST precludes many traditional solutions** (citizen science, supervised learning) for a range of use cases.

– Unsupervised techniques offer a number of advantages over supervised learning:
  – Not reliant on human classifiers
  – Classifications are not based on a training set, allowing for discovery
  – Unsupervised techniques can produce a human usable *description* of an object

– Robust classification and HSC-SSP (*LSST-like*) data.

– Other possible applications include:
  – Extraction of specific combinations of feature vectors in order to select/discover rare objects
  – Linking observations and simulations by visual similarity