# Algorithmic and machine learning approaches to automatic identification of peculiar galaxies in large astronomical databases

Lior Shamir
Kansas State University
XXXII Astronomical Data Analysis Software and Systems (ADASS), 2022

**Abstract**

Modern autonomous digital sky surveys can image many millions, and in some cases billions, of extra-galactic objects. While the vast majority of these objects are galaxies of well-known morphological types, it is expected that a small portion of these objects are rare or unseen galaxies. These objects can provide important information about the history, present, or future of the Universe, and can be of paramount scientific interest. Due to the extreme size of these databases, even a rare one-in-a-million extra-galactic object is expected to appear 1000 times in a database of one billion objects. But since these objects are hidden among a very high number of regular objects, finding these peculiar galaxies is impractical to perform manually, and requires automation. Here I discuss the application of different approaches to automatic detection of peculiar galaxies. These include supervised machine learning and model-driven algorithms for the detection of known rare objects, or unsupervised machine learning for the detection of unusual objects that are not yet known. Because of the large size of these databases, even a small false positive rate might lead to a large number of false detections, making the algorithm impractical. Therefore, one of the practical challenges of algorithms for automatic detection of peculiar objects in real-world data is the ability to control the trade-off between the completeness of the detection and the false positive rate. Another challenge is the ability to avoid false positives driven by artifacts, saturated pixels, and other bad image data. The algorithms and different approaches to peculiar galaxy detection are described, as well as their application to large dataset from DES, SDSS, and HST. These activities led to the detection of a large number of peculiar extra-galactic objects that would be impractical to detect manually.

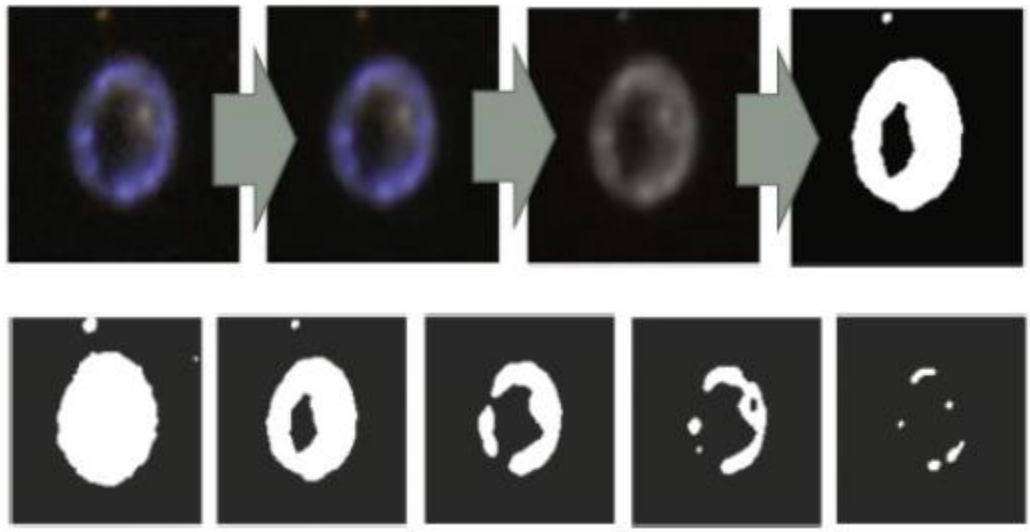
## Model-drive: Detection of known morphological types

When the galaxy has known morphology, it is possible to develop specific model-driven or data-driven (machine-learning) algorithms that can identify that specific morphological type. A simple example is ring galaxies. Ring galaxies are known but relatively rare galaxies. Because their shape is known, they can be detected by training a machine learning algorithm, or by using a model-driven approach.

This is an example of a simple model-driven algorithm that can detect ring galaxies. The algorithm applies dynamic thresholding, separates the foreground from the background, and then analyzes the shape at different background threshold to identify a ring.

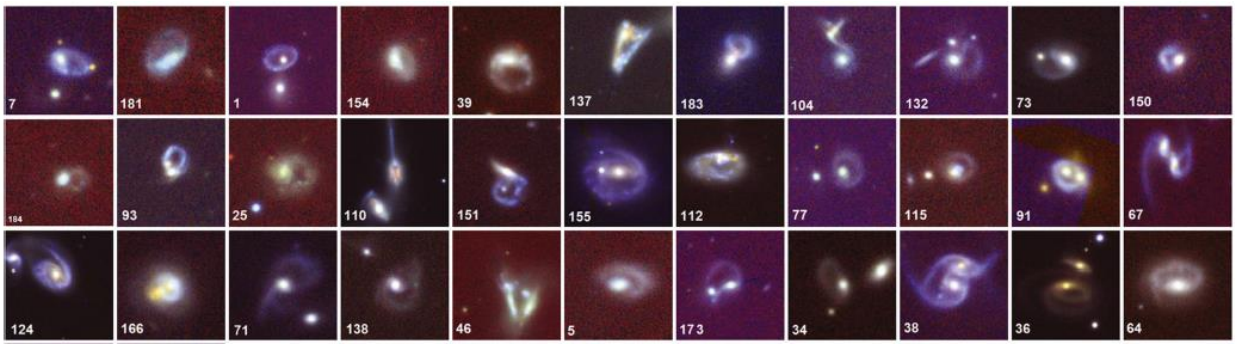
Examples from the ring galaxy catalog of (Shamir, 2020)

For every threshold, the algorithm checks for the existence of a relatively large region that is inside foreground pixels. After adjustment of parameters, the algorithm is able to detect ring galaxies. Full information about the algorithm is in (Timmis & Shamir, 2017; Shamir, 2020).



Two ring galaxy catalogs were prepared using the method, one from Pan-STARRS (Timmis & Shamir, 2017), and from SDSS (Shamir, 2020). The algorithm is not complete, and does not identify all ring galaxies. When the databases generated by autonomous digital sky surveys are very large, applying the algorithm to a large number of galaxies identifies a large number of ring galaxies. Catalogs were released (Timmis & Shamir, 2017; Shamir, 2020).

Previous catalogs of ring galaxies were also released, but compiling them required a substantial amount of labor from a relatively large number of participants (e.g., Buta, 2017). The catalogs described here were completed within a reasonable/manageable amount of labor invested in the process. Catalogs prepared manually take a long time, often several years, to prepare. The method shown here allowed




Examples of ring galaxies from the automatically-generated catalog of (Timmis & Shamir, 2017).

## Identification of unknown morphological types- Algorithmic approach

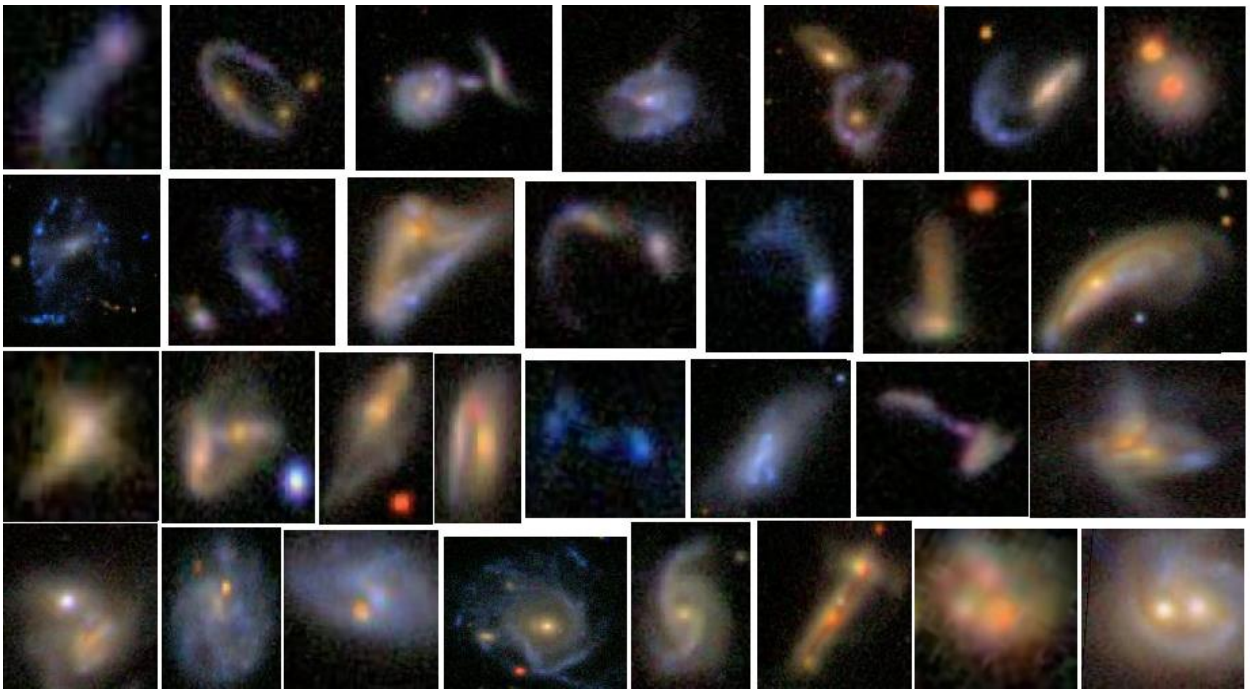
In very large databases generated by autonomous digital sky surveys, galaxies of unknown types might exist. These galaxies are rare, and therefore difficult to find, but these unknown types can provide unique information about the evolution of galaxies, as well as the past, present, and future of the Universe. Because they are not known, it is not possible to train a supervised machine learning system to identify them. Approaches to identify novelty galaxies of unknown shapes can rely on deep learning or shallow learning. For deep learning, autoencoders is the common approach for novelty detection. An example of such experiment is (Margapuri et al., 2021). The reconstruction loss of the autoencoder can be used to identify a galaxy image that is not common in the dataset, and therefore can be considered an outlier (Shamir, 2012).

Another approach is to apply outlier detection algorithms to features computed from each image. That approach allows to control the type of information being used to detect the outliers. An important advantage of this shallow learning approach is its ability to provide good control of the false-positive rate to completeness ratio. Due to the size of the data, even a small rate of false-positives will lead to an unmanageable output size. For instance, 1% false positives applied to a dataset of 10B objects will provide output of objects that need to be scanned manually.

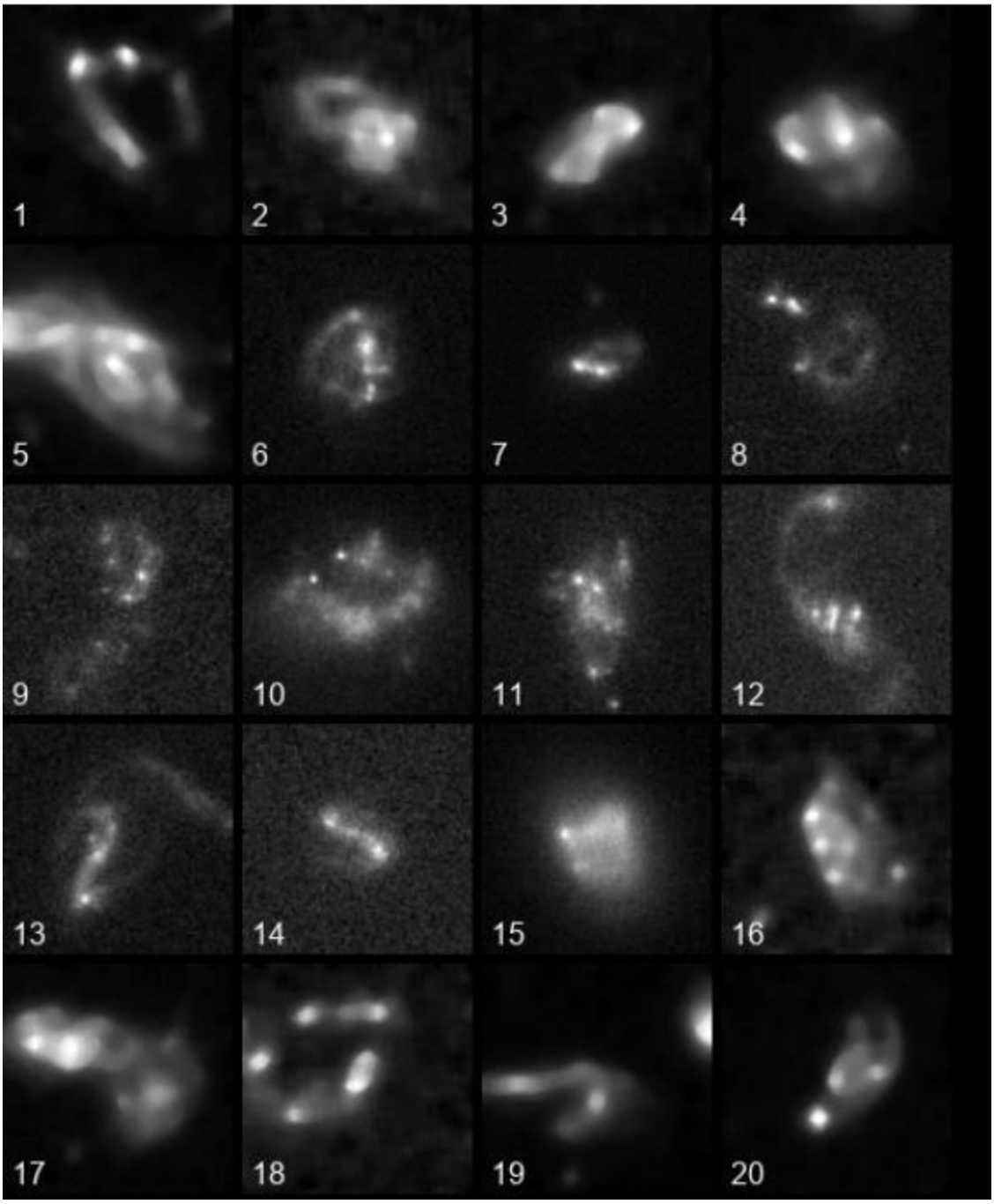
Below are examples of the application of a novelty galaxy detection algorithm to galaxies from Hubble Space Telescope (Shamir, 2021), SDSS (Shamir & Wallin, 2014), and the Dark Energy Survey.

The algorithm is based on a pre-defined feature set (Shamir, 2010) of numerical image content descriptors. Then, a distance metrics based on the Earth Movers Distance is applied to measure the distance between all pairs of galaxies. The distances are then ranked for each galaxy. The galaxies with the highest Kth distance from the other closest galaxies to them are determined as novelty galaxies.

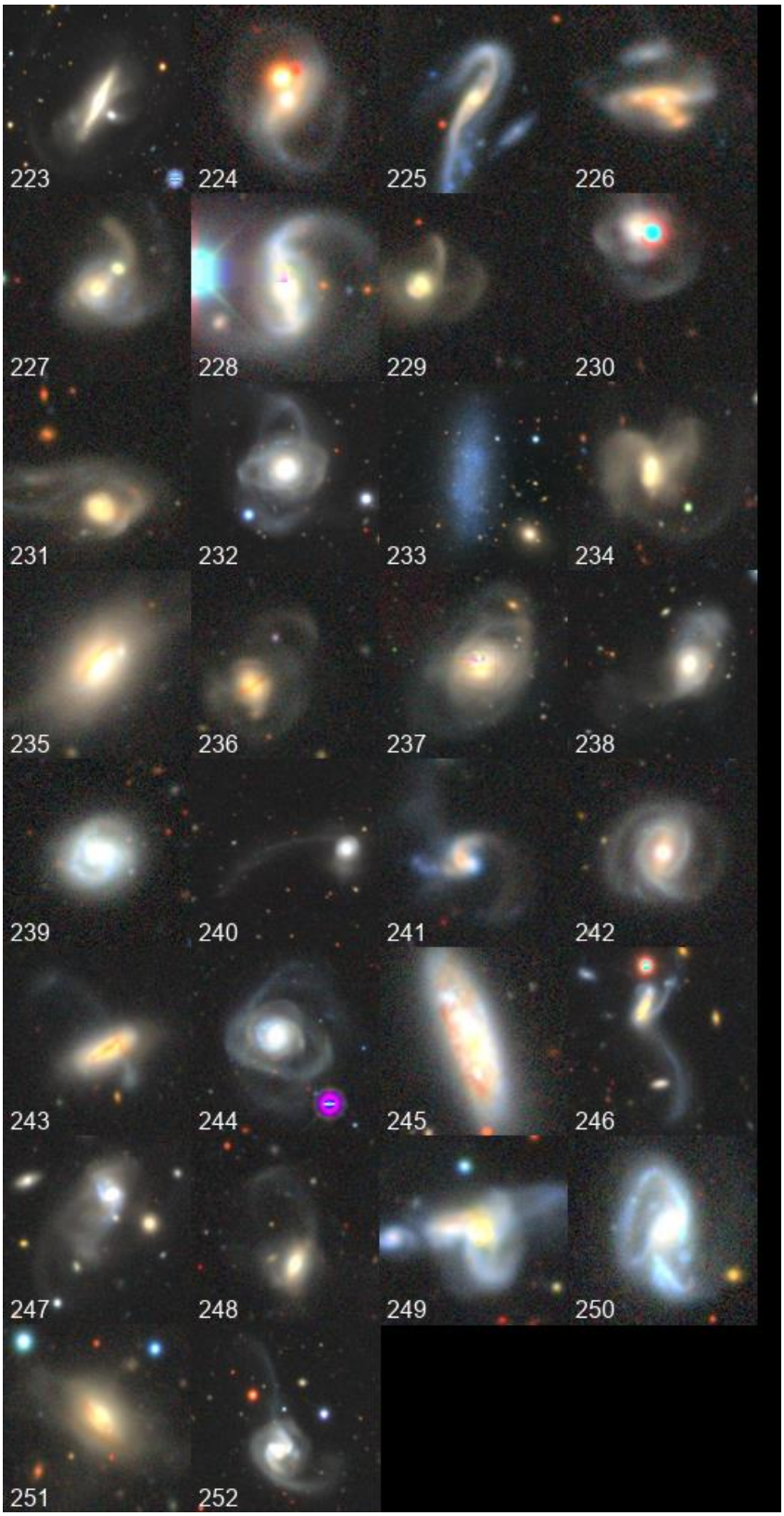
The algorithm is not perfect, and still provides a relatively high number of false-positives. Still, the algorithm reduces the data by three orders of magnitude, and allows for manual cleaning of the data to search for novelty galaxies in the data (Shamir, 2012,2020).



Examples of peculiar galaxy pairs detected in SDSS (Shamir & Wallin, 2014)

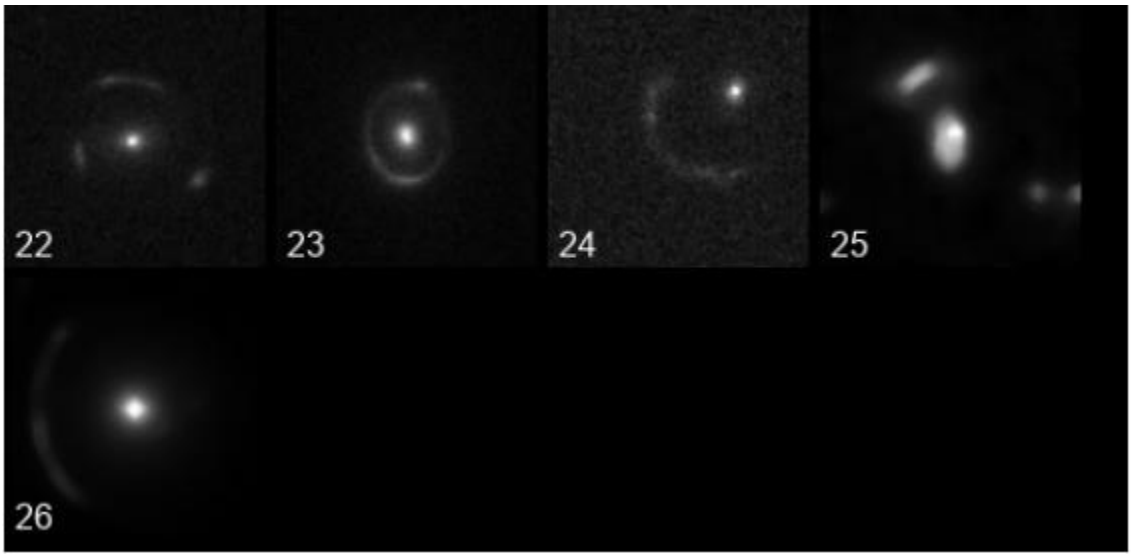



Peculiar galaxies identified by the algorithm in HST (Shamir, 2021)




Peculiar galaxies detected in DES

The algorithm can also detect gravitational lenses. Gravitational lenses are not necessarily peculiar galaxies, but their distorted shape makes them unusual compared to other galaxies. The following figure shows some examples of high redshift gravitational lenses detected in the HST CANDELS fields (Shamir, 2021).

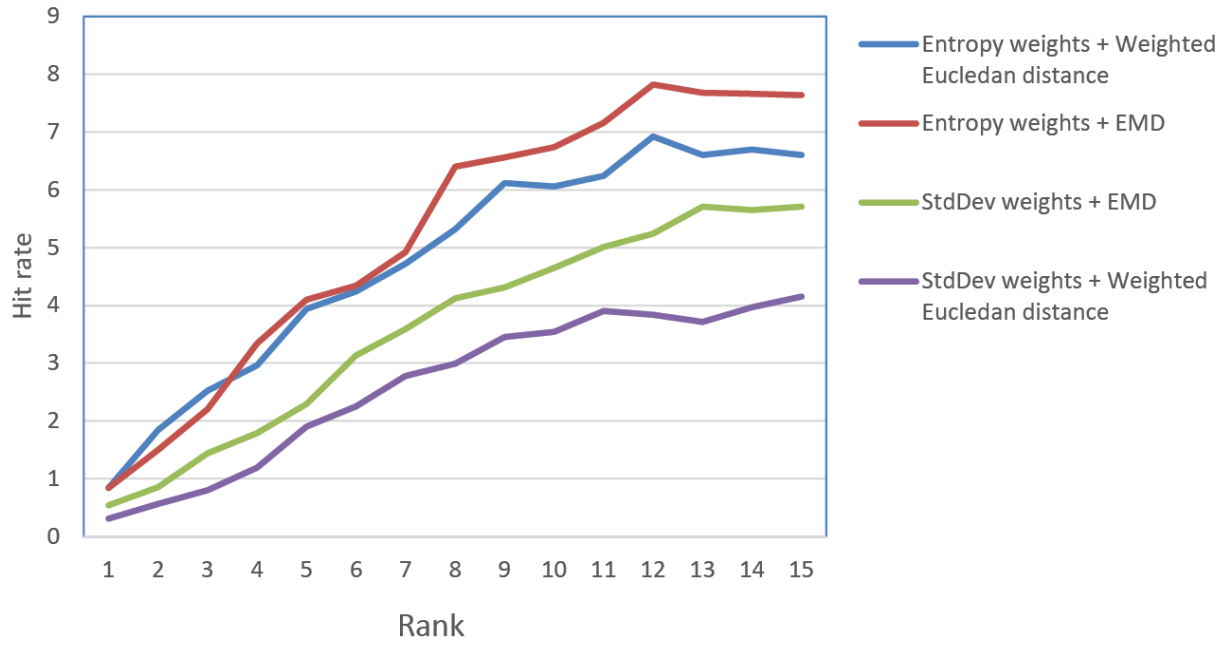



Examples of gravitational lenses detected by an automatic process in CANDELS

## Empirical testing of the performance of galaxy outlier detection algorithms

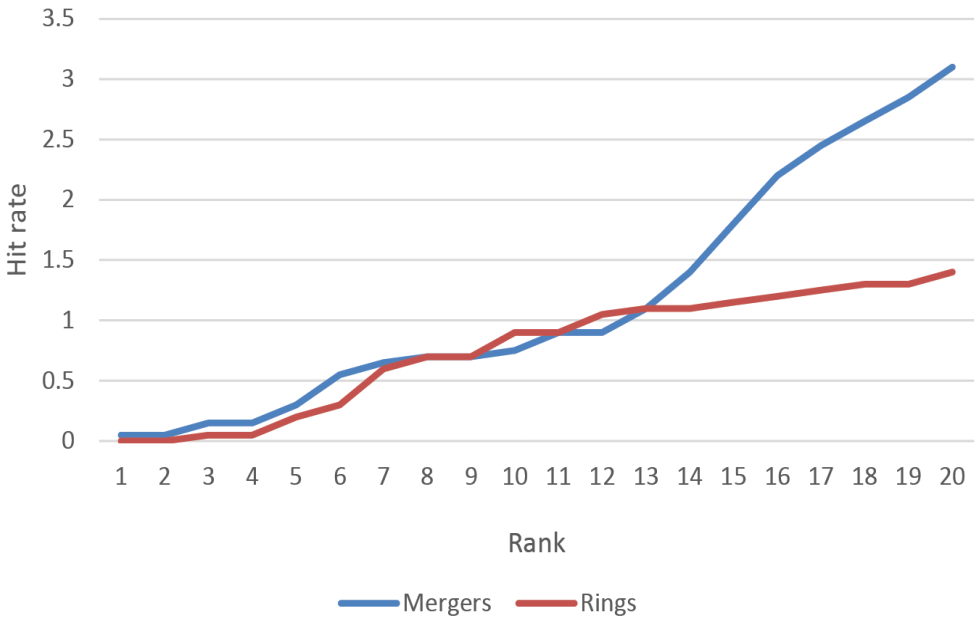
Testing the performance of outlier detection algorithm needs to account for the ability of the algorithm to detect novelty galaxies, but also the false-positive rate. Because databases collected by astronomical digital sky surveys are so large, even a small rate of false-positives will make the algorithm practically useless. That is because the number of false positives can be too high to clean manually. An algorithm for outlier galaxy detection should therefore be able to control the false-positive rate, and balance between the completeness of the algorithm (its ability to detect all peculiar galaxies) and the false-positives. Obviously, the definition of a “peculiar” galaxy is not clear, and therefore no algorithm can detect all peculiar galaxies and no false-positive, since the distinction between a false-positive and a peculiar galaxy is not fully defined.

A basic test is to combine a small number of elliptical galaxies and a larger number of spiral galaxies. In a universe of just spiral galaxies, and elliptical galaxy would be considered peculiar. Applying the algorithm can test the number of elliptical galaxies returned by the algorithm as peculiar. In each run, the algorithm returns a rank-K galaxies as the most peculiar. The rank K can be increased to test for the number of true positives among the total number of galaxies returned By the algorithm (most of them are false positives).

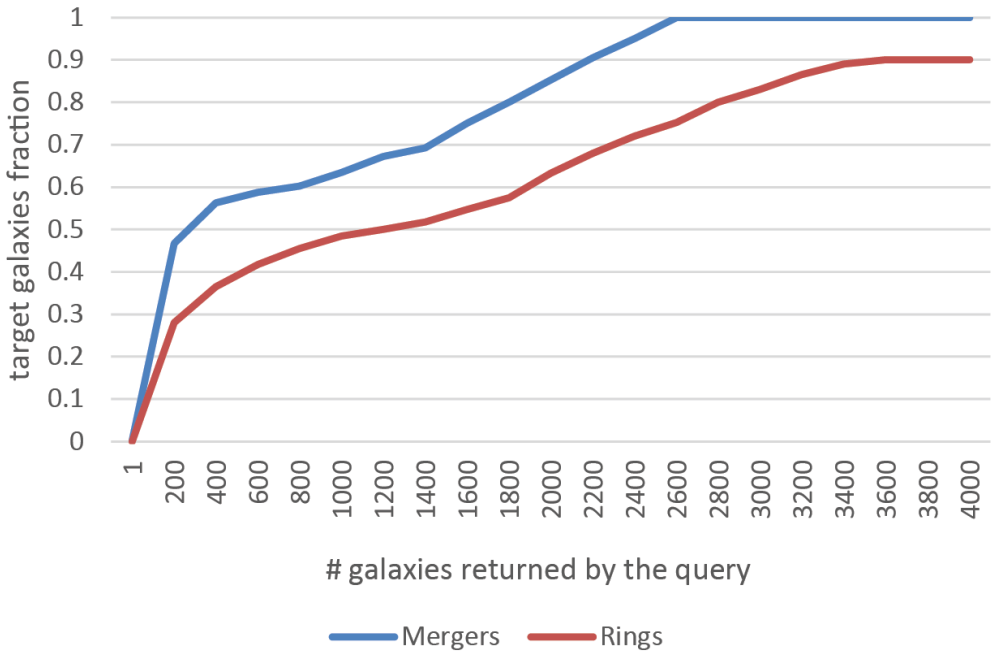



Ten elliptical galaxies among 100 spiral galaxies

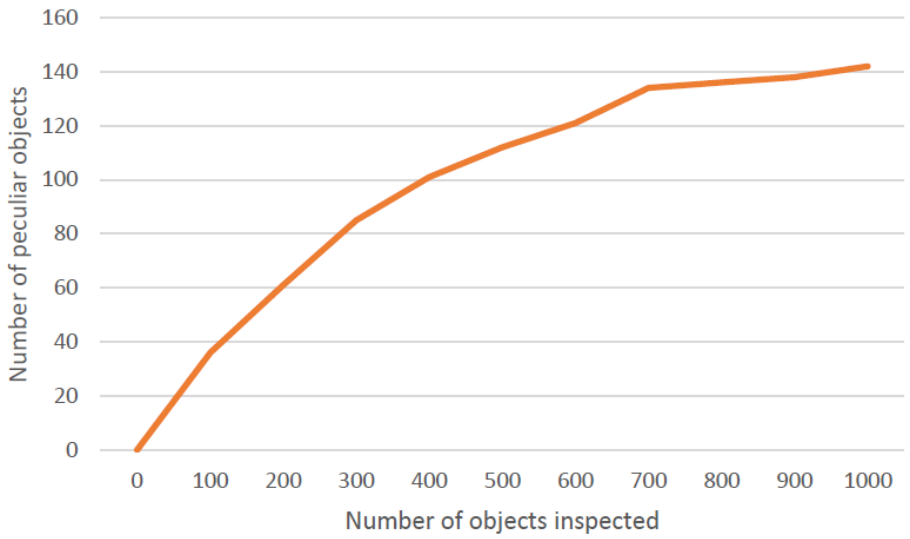
The examples show the results of testing  the performance of an algorithm for automatic detection of peculiar galaxies. The experiment combines a few ring galaxies in a much larger set of galaxies, and tests how many ring galaxies are detected by the algorithm. The detection is done by testing the number of ring galaxies among the top K galaxies returned by the algorithm as the most likely to be peculiar. It also does the same thing for merging galaxies among a dataset of non-merging galaxies.




Twenty ring galaxies and 20 interacting galaxies among 10,000 galaxies (i<18)




Completeness of the catalog. Twenty ring or interacting galaxies among 10,000 galaxies




Peculiar objects detected manually among the peculiar objects detected by the algorithm. It is clear that the vast majority of the objects detected by the algorithm are not really peculiar, but the algorithm reduces the data by ~3 orders of magnitude.

**References**

Shamir, L., Automatic detection of peculiar galaxies in large datasets of galaxy images, *Journal of Computational Science*, 3(3), 181-189, 2012.

Buta, R.J, Galactic rings revisited – I. CVRHS classifications of 3962 ringed galaxies from the Galaxy Zoo 2 Database, *Monthly Notices of the Royal Astronomical Society* 471, 4027–4046, 2017.

Timmis, I., Shamir, L., A catalog of automatically detected ring galaxy candidates in PanSTARRS, *Astrophysical Journal Supplement Series*, 231(1), 2, 2017.

Shamir, L., Automatic detection of full ring galaxy candidates in SDSS, *Monthly Notices of the Royal Astronomical Society*, 491(3), 3767-3777, 2020.

Shamir, L., Automatic identification of outliers in Hubble Space Telescope galaxy images, *Monthly Notices of the Royal Astronomical Society*, 501(4), 5229-5238, 2021.

Margapuri, V., Thapa, B., Shamir, L., Automatic detection of novelty galaxies in digital sky survey data, *International Journal of Computer Applications*, 28(1), 2021.

Shamir, L., Morphology-based query for galaxy image databases, Publications of the Astronomical Society of the Pacific, 129(972), 024003, 2017b

Acknowledgement: The project is funded by NSF grant number AST-1903823