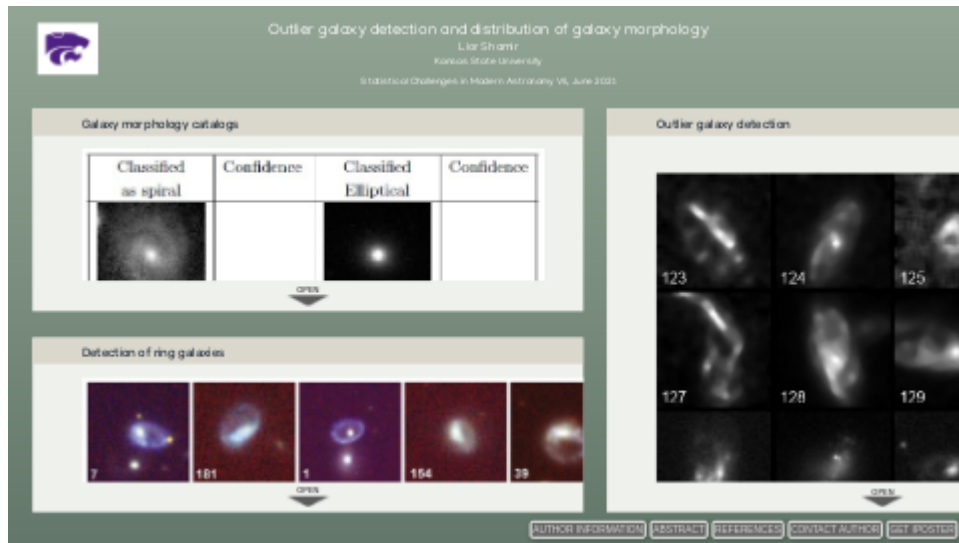


Outlier galaxy detection and distribution of galaxy morphology



Lior Shamir

Kansas State University

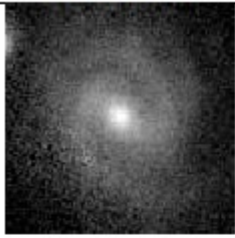
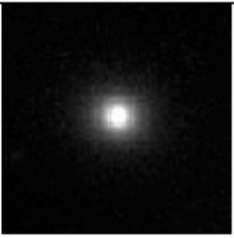
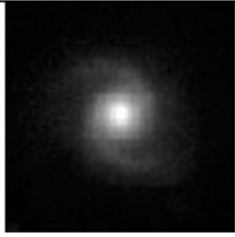

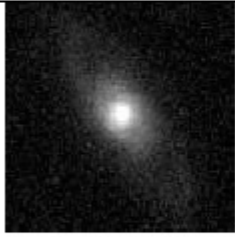
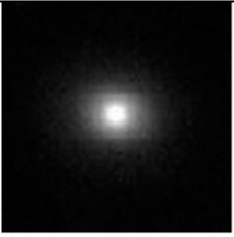




Statistical Challenges in Modern Astronomy VII, June 2021

PRESENTED AT:

**Statistical Challenges in
Modern Astronomy VII**

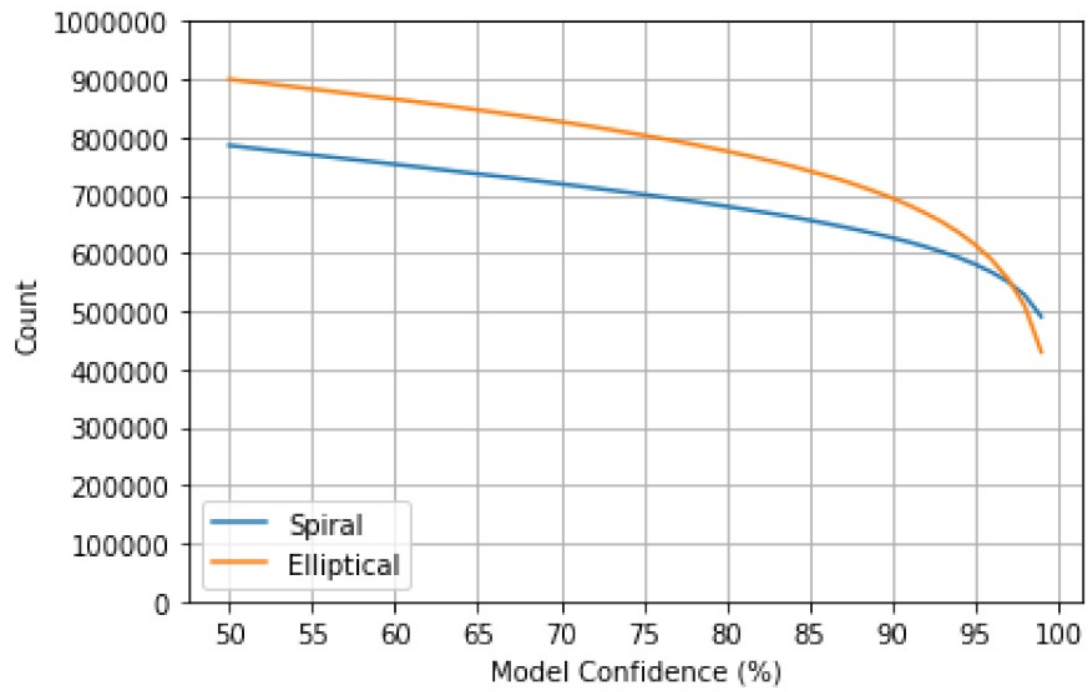


GALAXY MORPHOLOGY CATALOGS

Classified as spiral	Confidence	Classified Elliptical	Confidence
	0.9999		0.9999
	0.9988		0.8799
	0.9718		0.7568
	0.7446		0.6780
	0.5163		0.5637

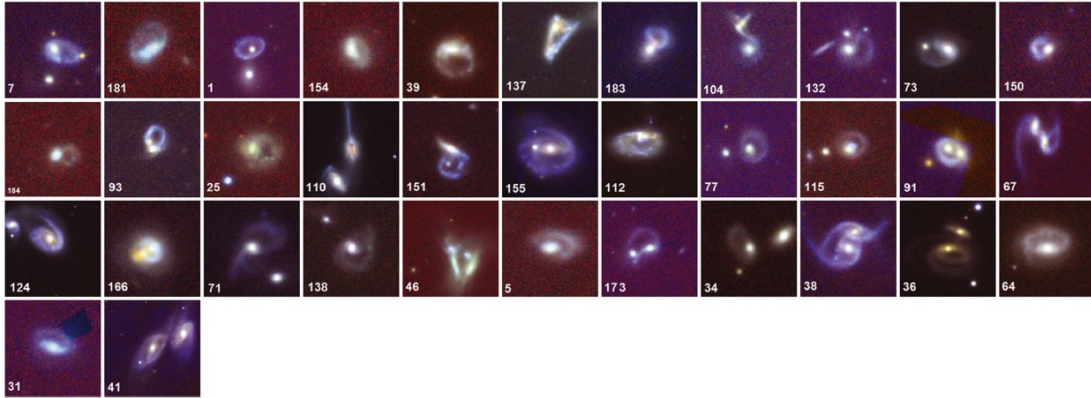
With $>10^9$ galaxies, Vera Rubin Observatory galaxy data will need to be annotated automatically. An example of a catalog of galaxies annotated by using deep learning is a catalog of $\sim 1.7 \cdot 10^6$ galaxies classified by their broad morphology (Goddard & Shamir, 2020). The catalog provide 95% accuracy compared to SDSS galaxies classified by manually by Galaxy Zoo, and considered unbiased "superclean" by Galaxy Zoo criteria.

The neural network is a simple LeNet-5 architecture with ReLU activation functions. The neural network was trained with $\sim 2 \cdot 10^5$ galaxies annotated manually by Galaxy Zoo as "superclean", and happens to also be imaged by SDSS. The training images are also mirrored to reduce human perception bias, but some bias might still exist due to the manual annotation of the training data (Goddard & Shamir, 2020).

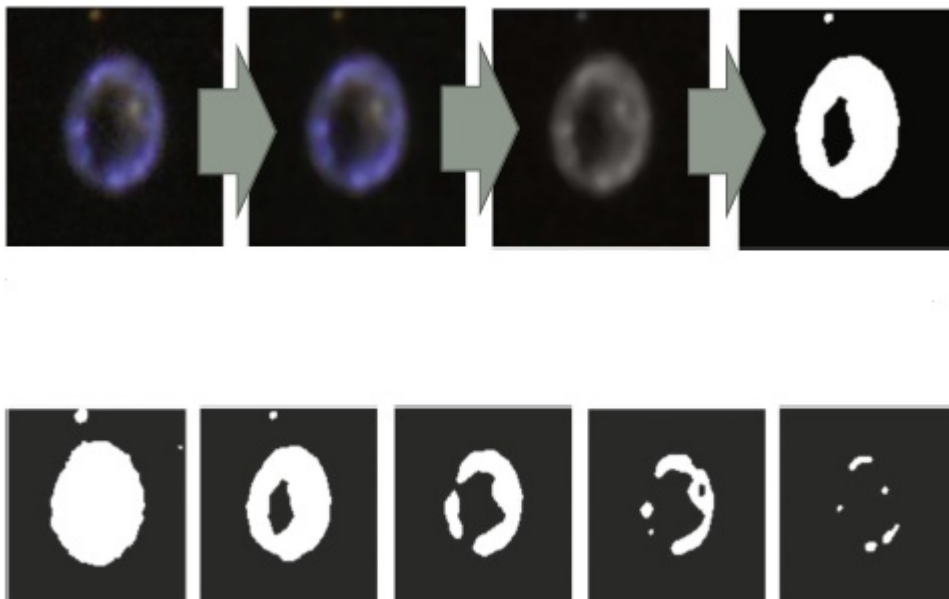


The catalog is available for download at: [https://figshare.com/articles/PanSTARRS DR1 Broad Morphology Catalog/12081144](https://figshare.com/articles/PanSTARRS_DR1_Broad_Morphology_Catalog/12081144) (Goddard & Shamir, 2020)

DETECTION OF RING GALAXIES



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.



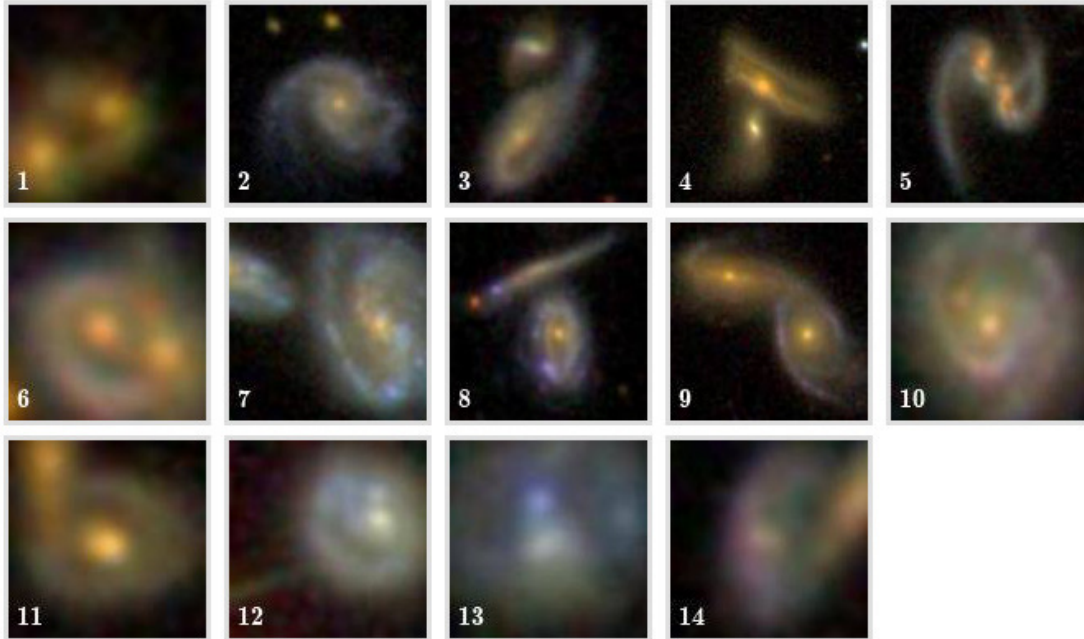
The algorithm applies dynamic thresholding, separates the foreground from the background, and then analyzes the shape at different background threshold to identify a ring. Two ring galaxy catalogs were prepared using the method, one from Pan-STARRS (Timmis & Shamir, 2017), and one from SDSS; Shamir, 2020).

It is interesting that these catalogs were smaller than catalogs prepared manually by crowdsourcing (Butas, 2017). But the crowdsourcing took several years and thousands of people, while the preparation of the automatic ring catalogs were far less labor-intensive, normally one person working a few hours per week. It is clear that with the scales of data generated by Vera Rubin Observatory human analysis will not be practical for full utilization of the data.

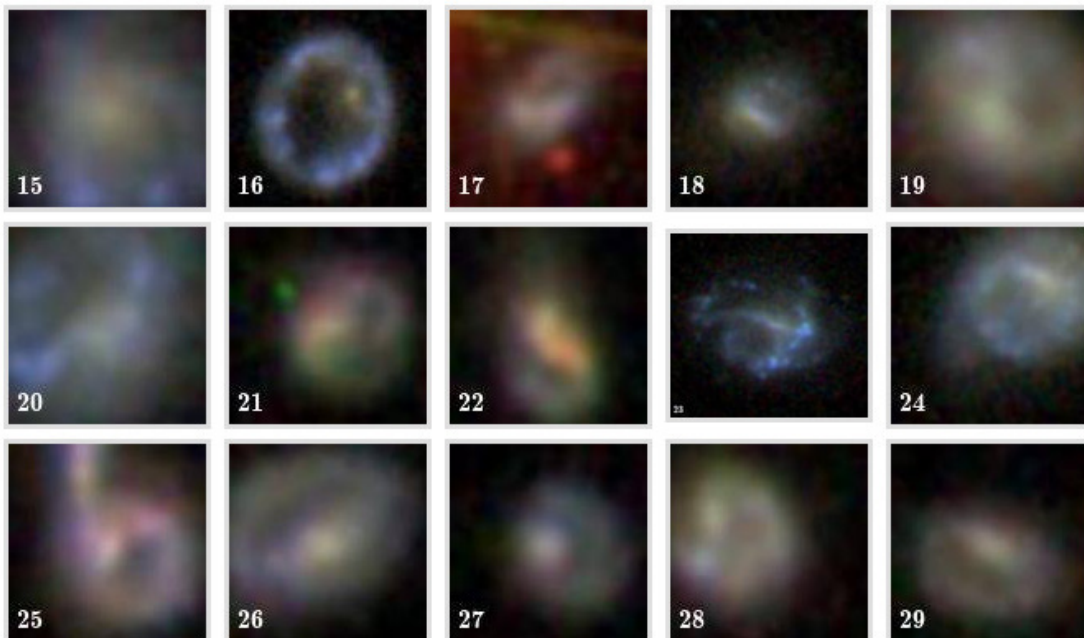
Additionally, most galaxies in automatically-generated catalog were not included in the Butas (2017) catalog or other manually prepared catalogs. While the crowdsourcing has a much higher detection rate, the automatic catalog can "cover" a far larger initial set of galaxies.

Examples of ring galaxies detected by the algorithm

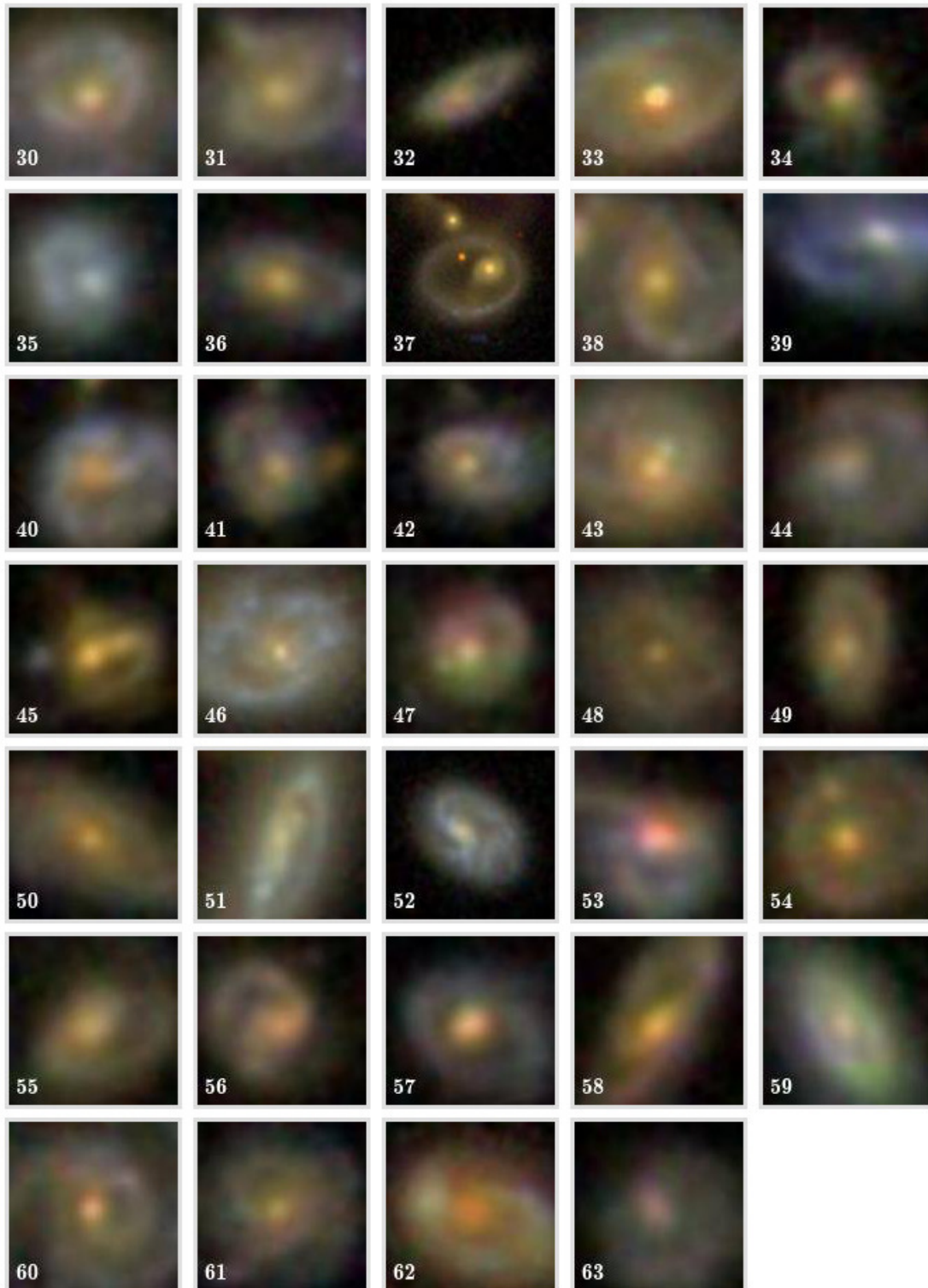
Collisional rings

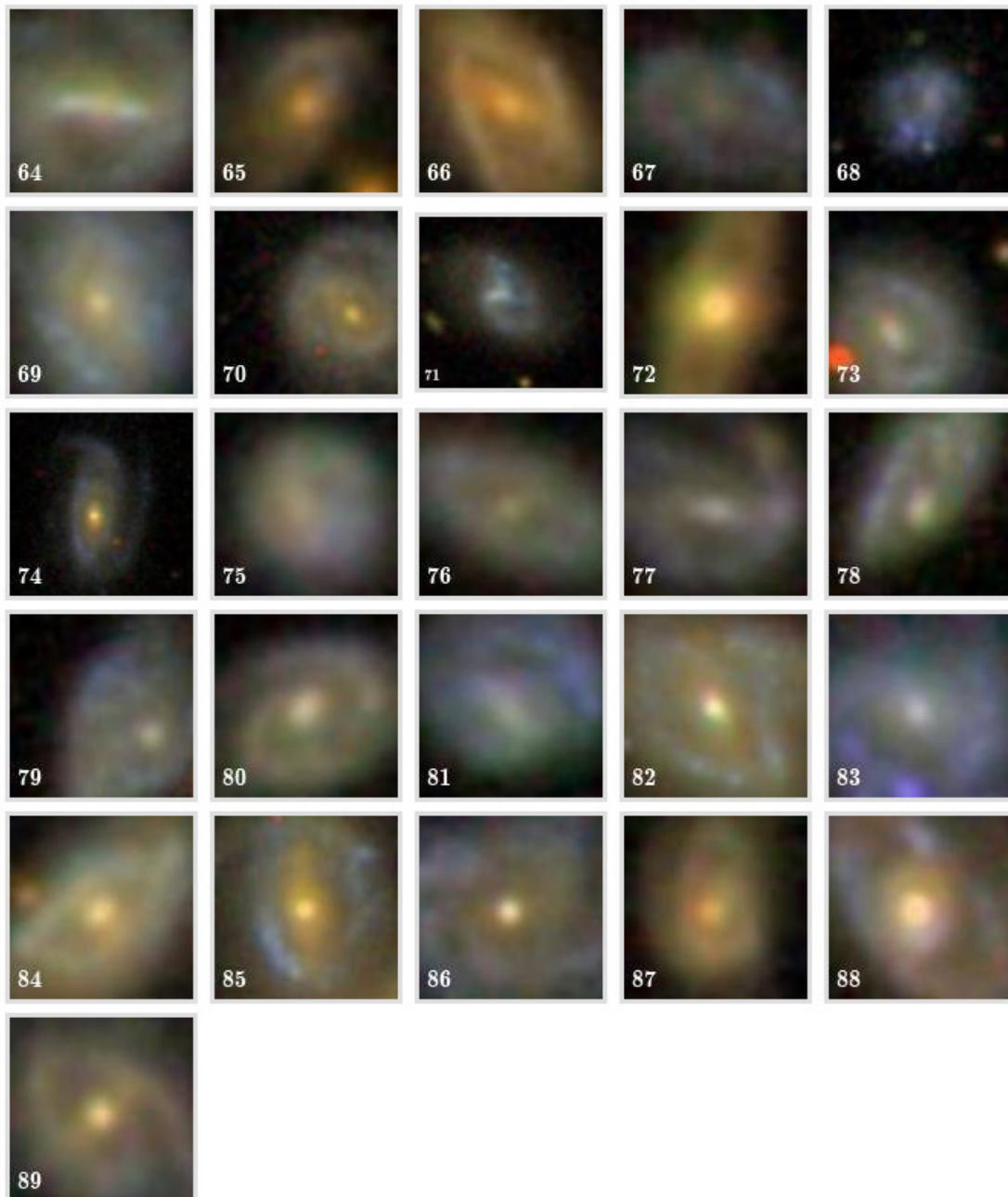


Rings with no nuclei



Rings with off-center nuclei





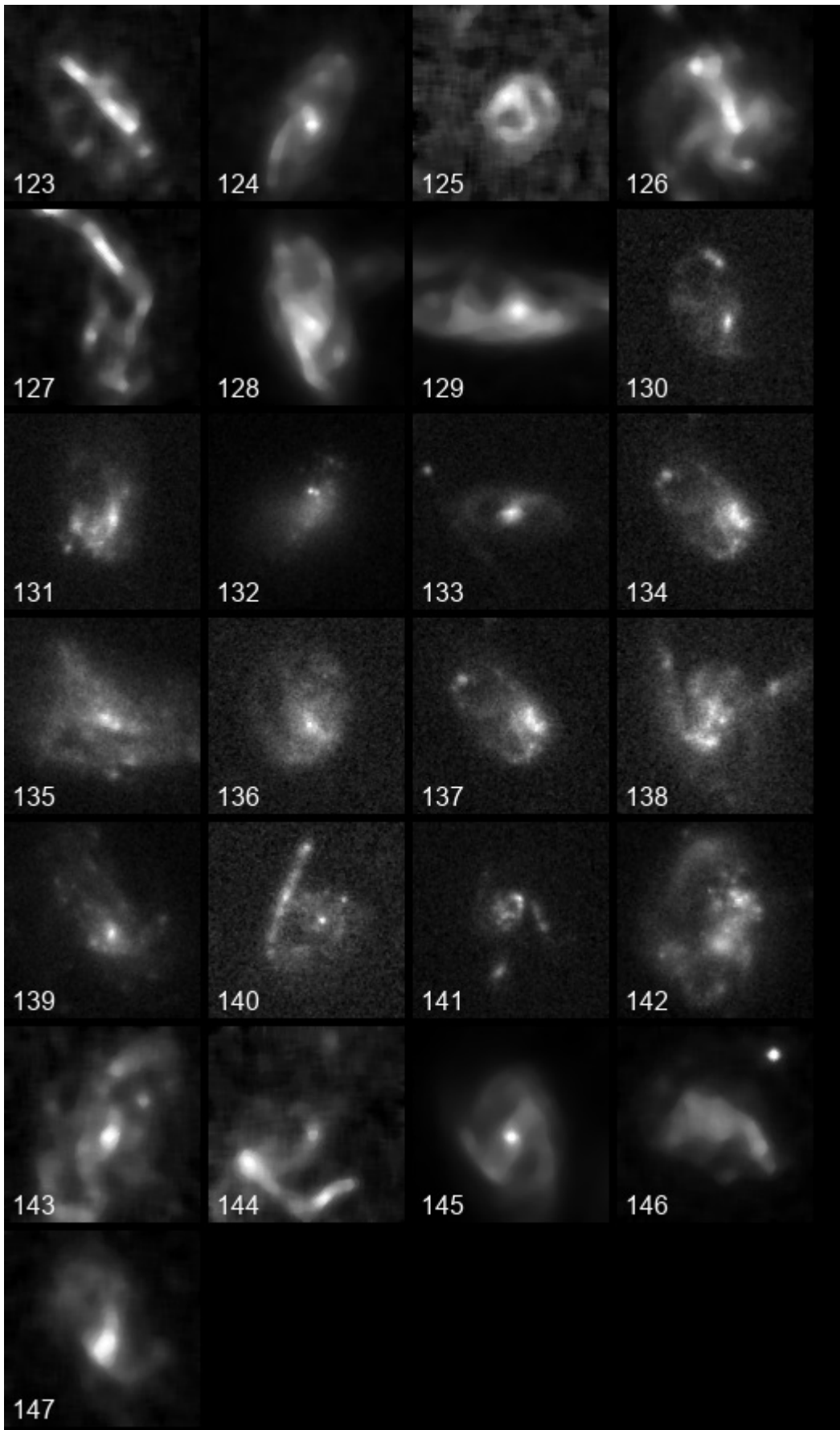
It is obvious that manual detection of all these galaxies would only be possible by using a very large number of people. When the scales of the Vera Rubin Observatory data, manual analysis will become impractical.

Catalogs:

A catalog of ring galaxies in SDSS - (Shamir, 2020).

A catalog of ring galaxies in Pan-STARRS (Timmis & Shamir, 2017).

OUTLIER GALAXY DETECTION



Some rare galaxies, such as ring galaxies, have known morphology. But with a database of the size of the Vera Rubin Observatory, it can be assumed that very many galaxies of unknown morphology will be imaged. Even rare one-in-a-million type of galaxy will be present in the Vera Rubin Database $\sim 10^4$ times. Identifying these galaxies is a difficult task because the morphology of the galaxy is unknown.

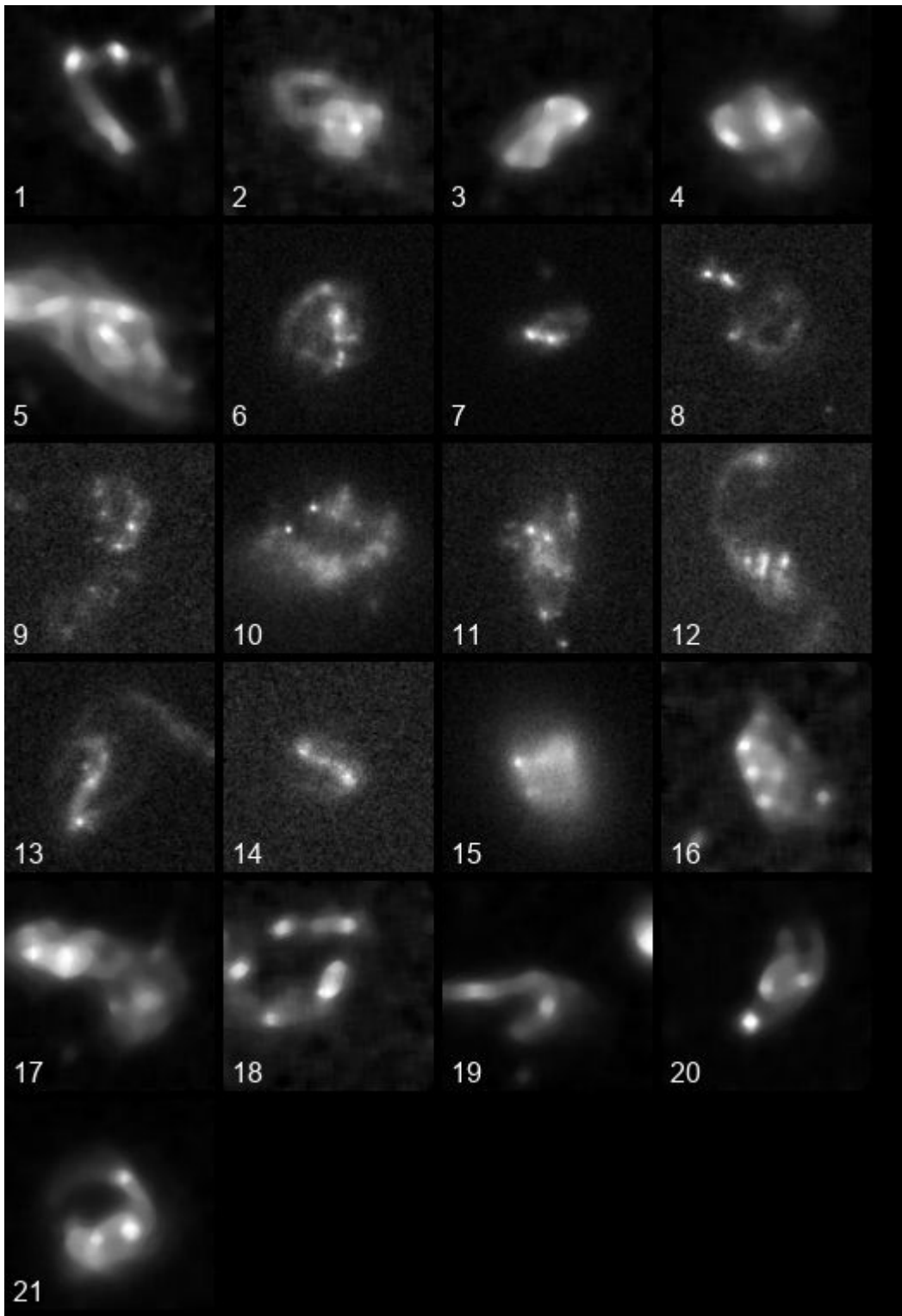
Solution to that problem can be based on unsupervised machine learning. One approach is the use of deep learning autoencoders. An example is (Margapuri, Thapa & Shamir, 2021). The reconstruction loss of the autoencoder can be used to identify a galaxy image that is not common in the dataset, and therefore can be considered an outlier.

Other solutions can be based on "shallow learning" algorithms (Shamir, 2012). The key for practical outlier detection is low false-positive rate. Due to the size of the data, even a small rate of false-positives will lead to unmanageable output. For instance, 1% false positives applied to a dataset of 10^8 objects will provide output of 10^6 objects that need to be scanned manually.

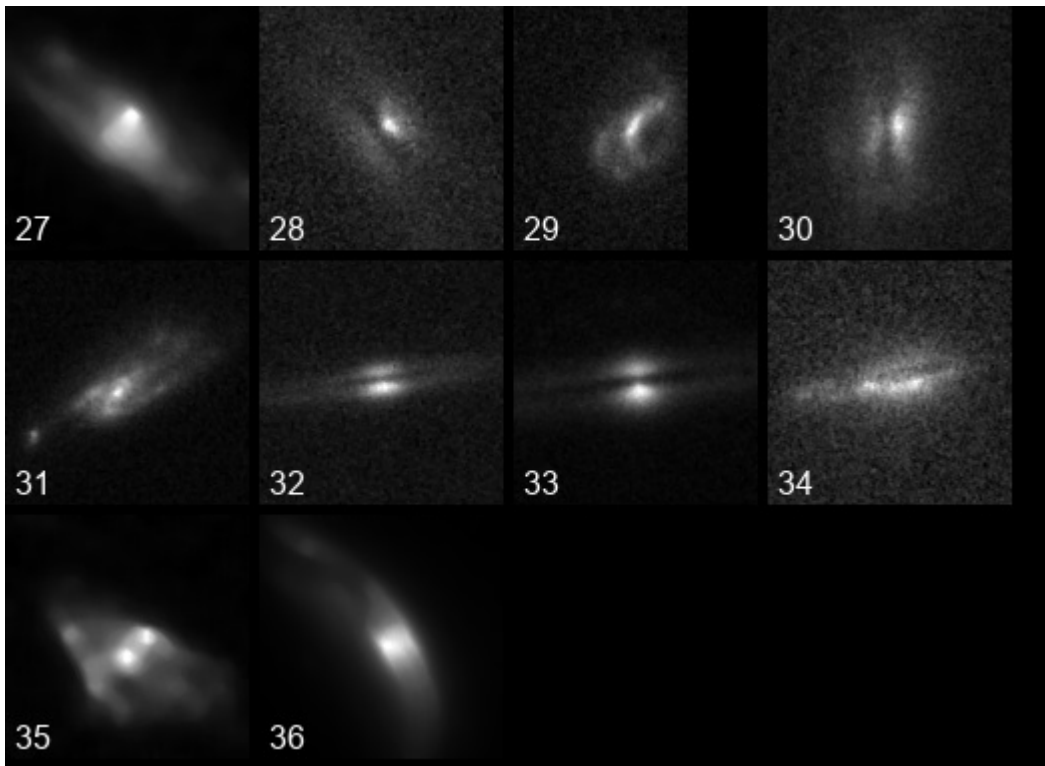
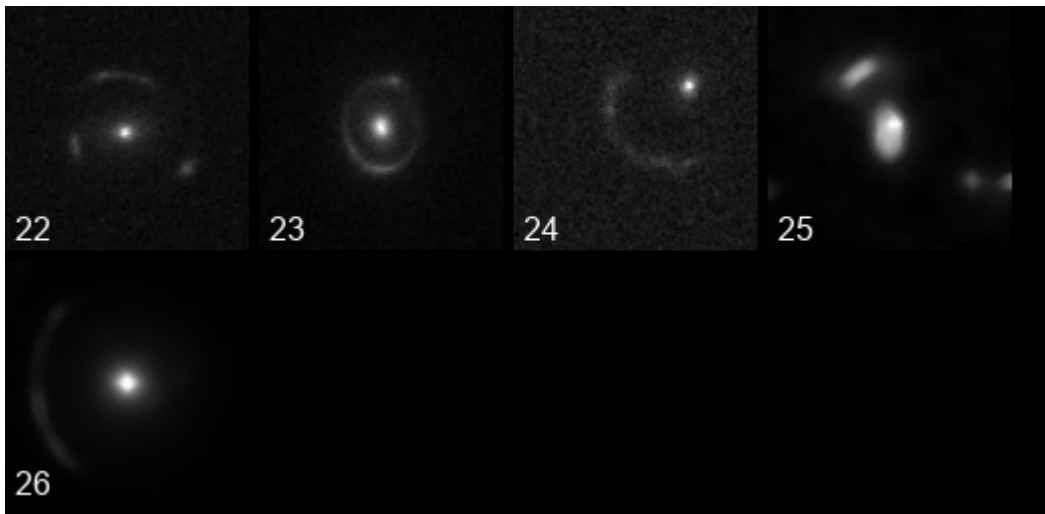
Deep learning autoencoders have a relatively high false-positive rate, Shallow learning algorithms allow to control the sensitivity and specificity and can therefore allow to control the size of the output (Shamir, 2012).

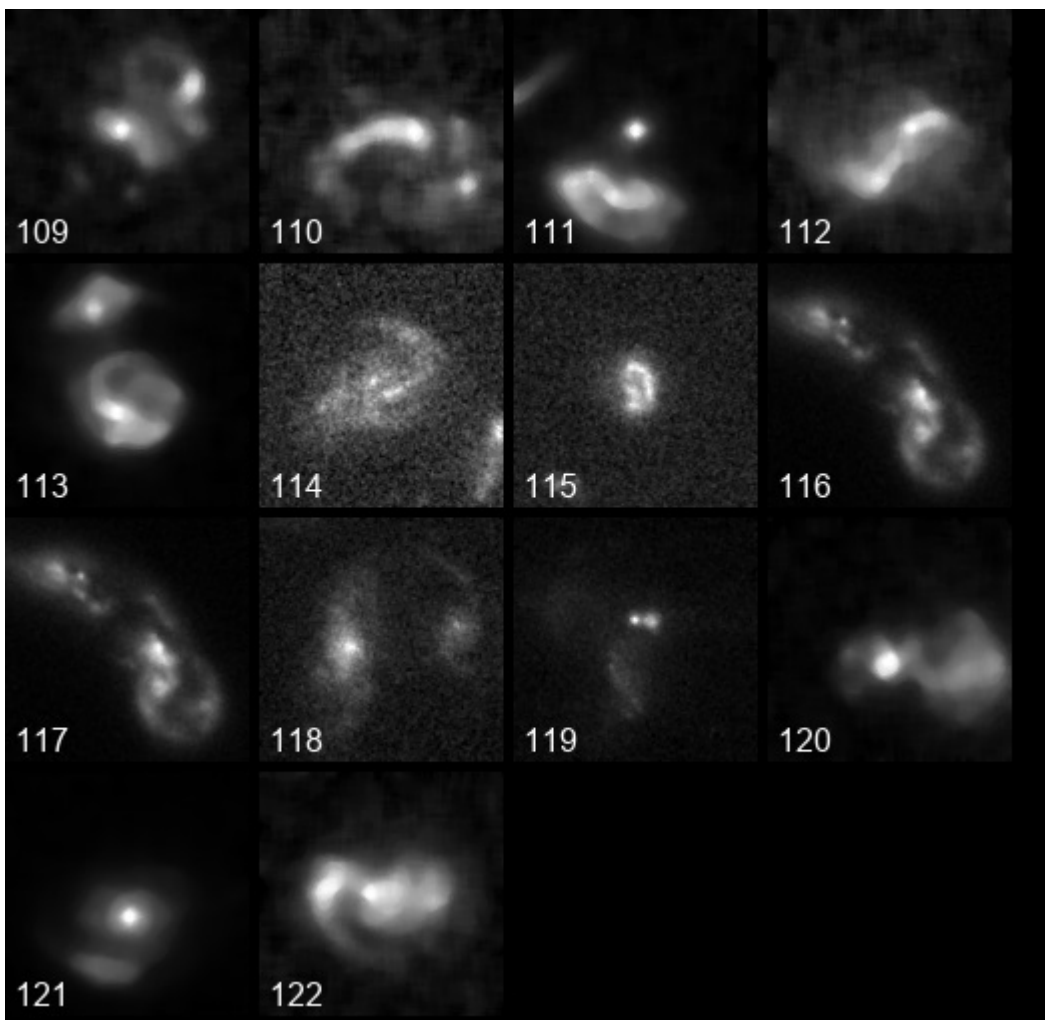
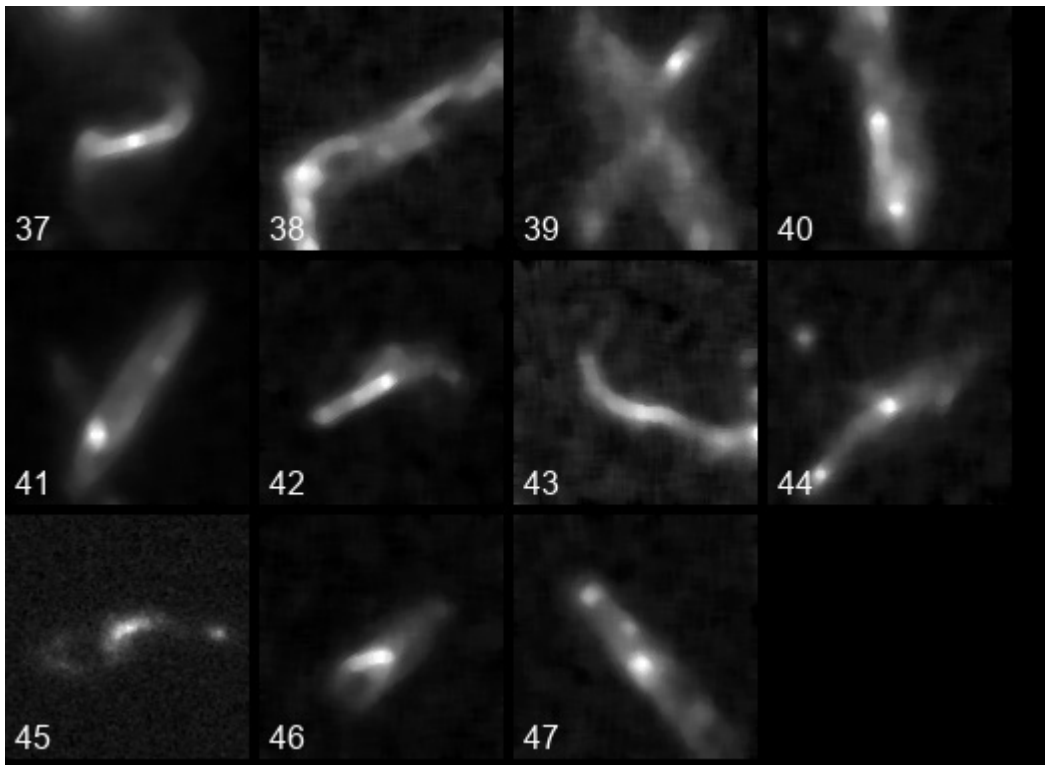
The examples below show peculiar galaxies identified automatically by applying a shallow learning algorithm to $\sim 3 \cdot 10^5$ Hubble Space Telescope galaxies. (Shamir, 2021a).

The current version of the algorithm is not perfect, and produces a large number of false positives (Shamir, 2021a). But it allows to reduce the data by ~ 3 orders of magnitude. Manual examination of the data becomes much more practical with the reduced dataset. The reduced dataset has far higher frequency of peculiar objects, and therefore scanning it manually becomes much more fruitful compared to scanning the original data without any filtering. Still, in its present version it might not provide an optimal solution to the data coming from Vera Rubin Observatory.

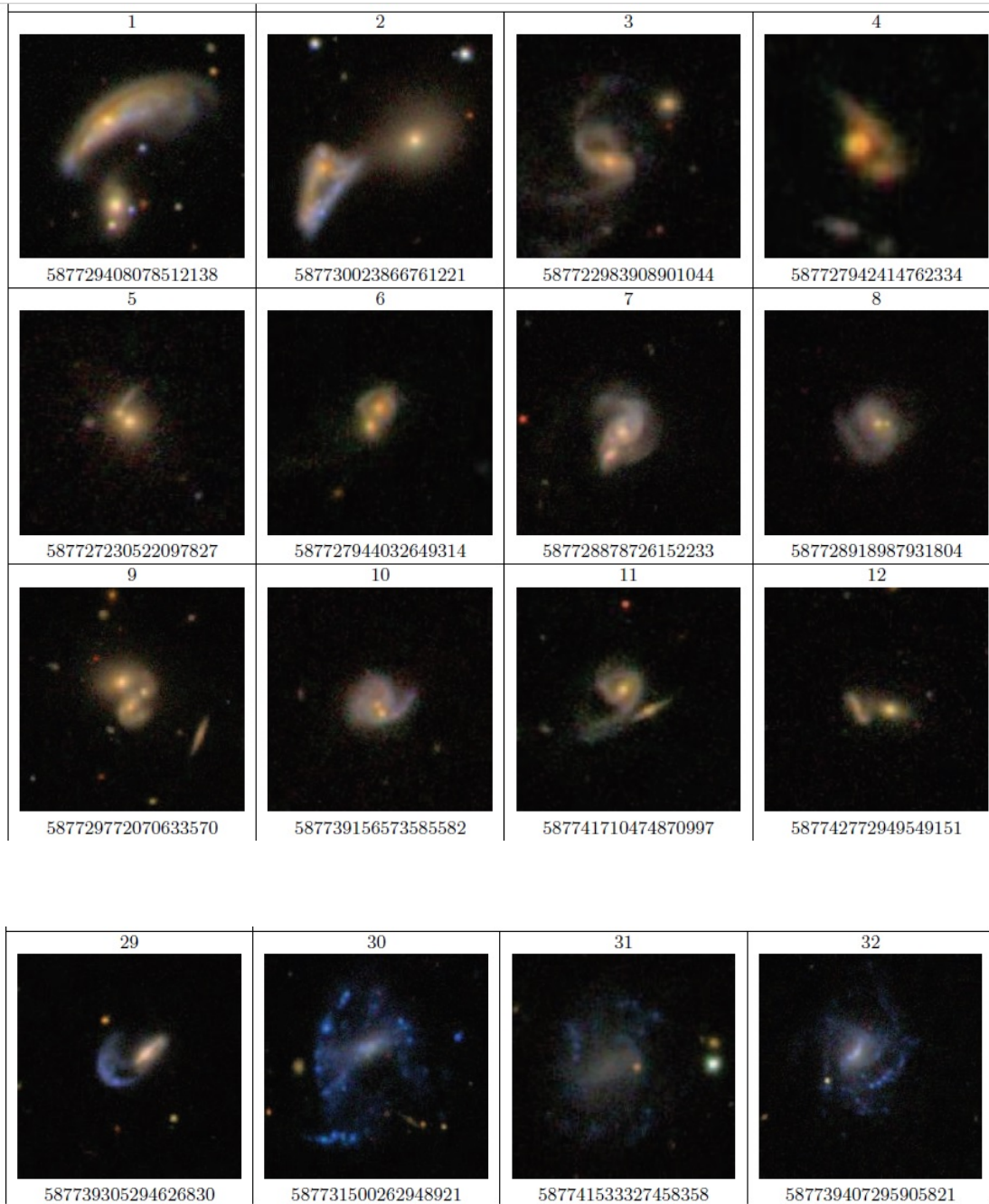


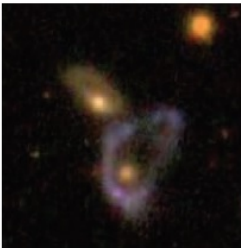




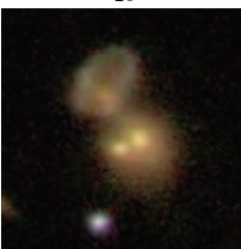
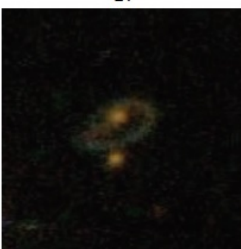

Gravitational lens:



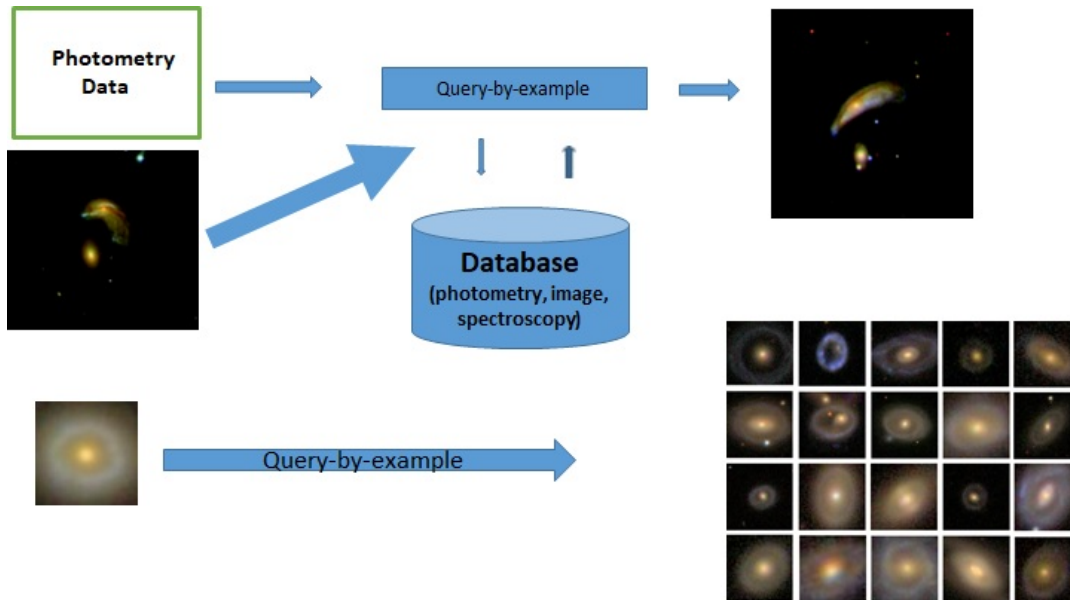


The following galaxies mergers are examples of peculiar galaxy pairs detected automatically in SDSS (Shamir & Wallin, 2014).



<p>21</p>  <p>587730774407840452</p>	<p>22</p>  <p>587742631737229751</p>	<p>23</p>  <p>587730845814751853</p>	<p>24</p>  <p>587728308567015452</p>
<p>25</p>  <p>587729233595859458</p>	<p>26</p>  <p>587736976890134917</p>	<p>27</p>  <p>587731187810238739</p>	<p>28</p>  <p>587725550139277460</p>

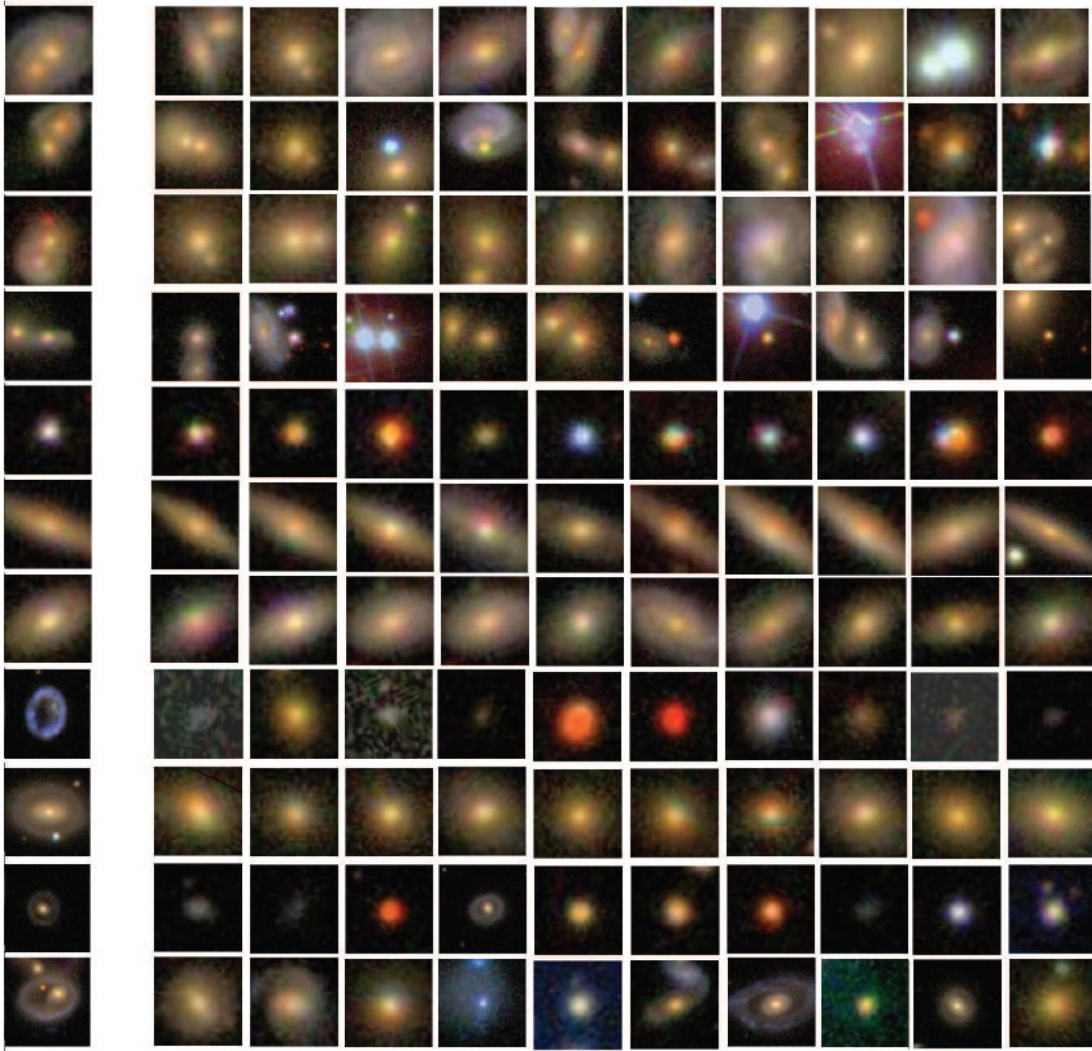
GALAXY QUERY-BY-EXAMPLE



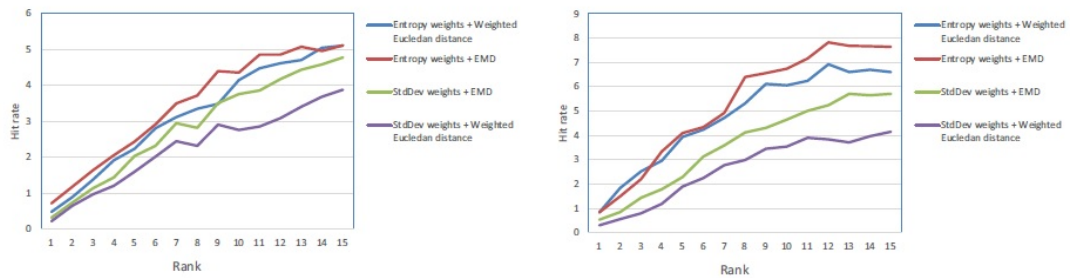
To study a certain type of rare galaxies, sometimes a population of galaxies of the same type is needed. That is because it is difficult to make general conclusions when $N=1$. The galaxy query-by-example algorithm receives a galaxy image as input, and returns a collection of galaxies that are the most similar to the query galaxy (Shamir, 2017b). For that purpose, both photometry data and image data are combined (Shamir, 2017b).

For instance, if the researcher finds a peculiar galaxy of interest and wishes to study it, she can query that image to receive a list of galaxies similar to that specific type.

Below are examples of the query galaxy, and the ten most similar galaxies identified automatically in a dataset of $\sim 10^4$ SDSS galaxies. The algorithm is not perfect, but it is clear that the images fetched by the algorithm are more similar to the query galaxy than random.



The figure below shows an experiment of combining 10 elliptical galaxies with 100 spiral galaxies (left), and 10 spiral galaxies with 100 elliptical galaxies (right). The detection rate of the "peculiar" galaxies is not perfect, but far higher than random based on the query galaxy (Shamir, 2017b).

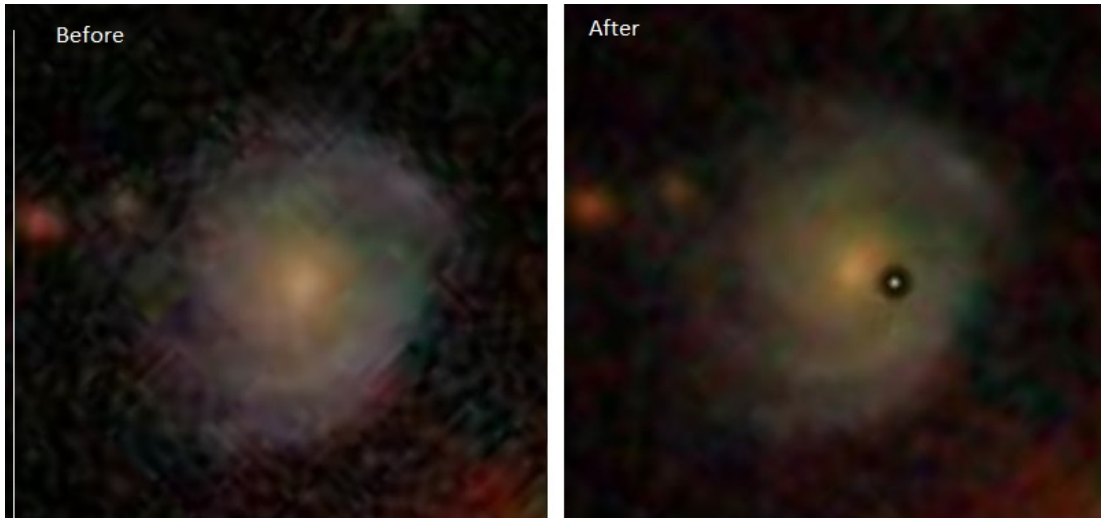


Until now the algorithm was only used in controlled experiments. No catalog or other data products were produced with that algorithm so far.

"CHANGING" GALAXIES

DR7 (before)

DR8 (after)

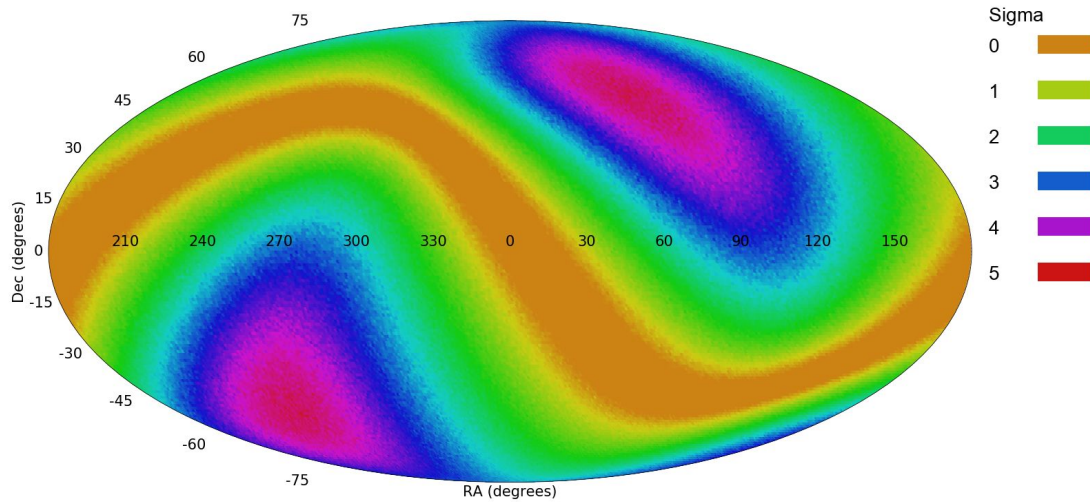


Extended extra-galactic objects that can be visible from Earth are large, and therefore the shape of extra-galactic objects is not expected to change between different data releases. However, such assumptions still need to be tested. With the amounts of data that will be acquired by the Vera Rubin Observatory, algorithms that compare objects in different data releases can identify possible rare cases in which an extra-galactic object imaged in one data release is different from the same object imaged a few years later. While substantial attention has been given to transient detection, this form of "extra-galactic transients" can potentially expand on the existing knowledge.

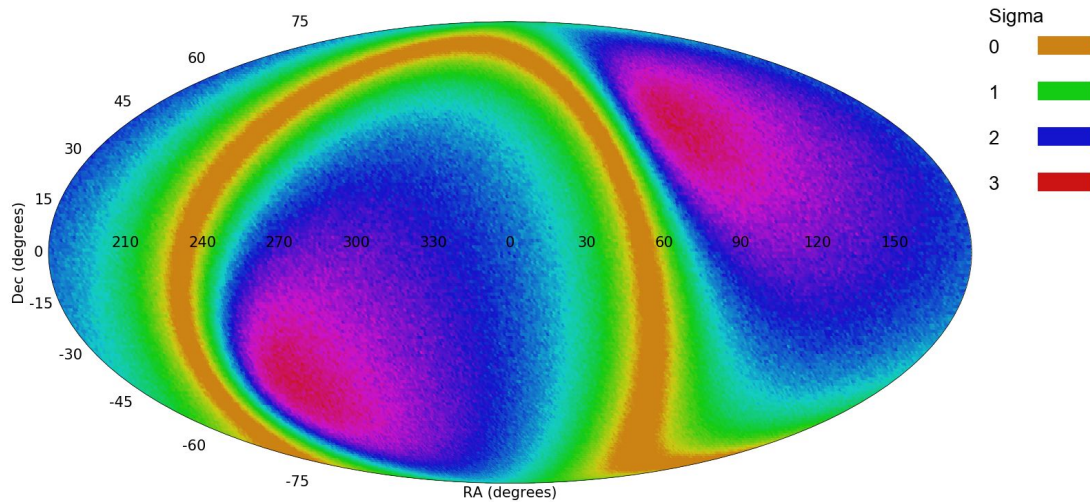
The example above shows an extra-galactic object that "changes" between data releases in SDSS. The reason is not clear. Obviously, galaxies cannot change in a manner that is noticeable within a few years. If the reason for the "change" is astronomical, a possible explanation would be that a foreground nova somehow eclipses the galaxy.

LARGE-SCALE DISTRIBUTION OF GALAXY SPIN DIRECTIONS

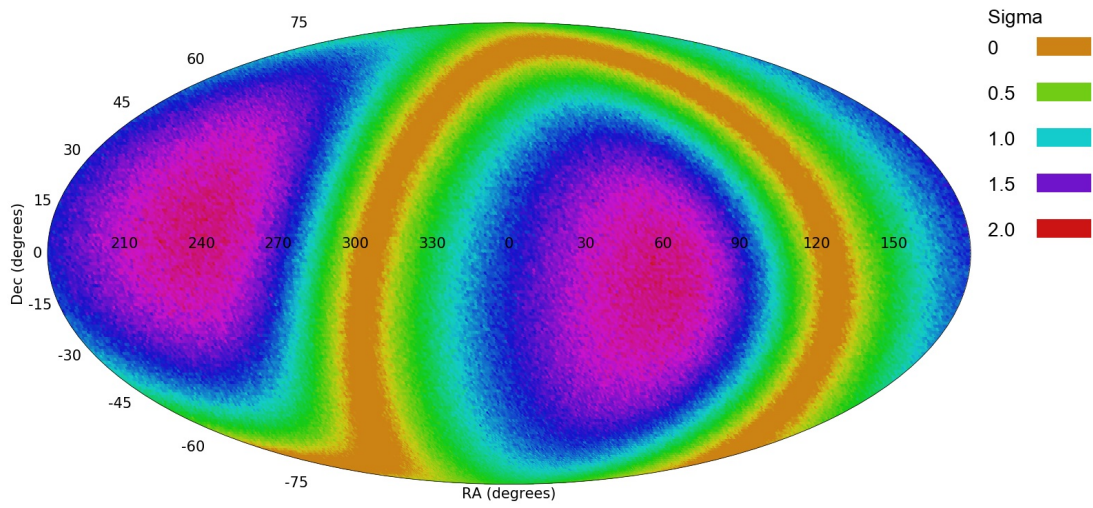
The ability to analyze a large number of galaxies allows to study questions that were very difficult to solve without using digital sky survey and automatic analysis. One of these question is the large-scale distribution of spin direction of spiral galaxies. Below is the spin direction distribution from four different sky surveys: SDSS DR14, Pan-STARRS DR1, DESI Legacy Survey, and HST COSMOS:



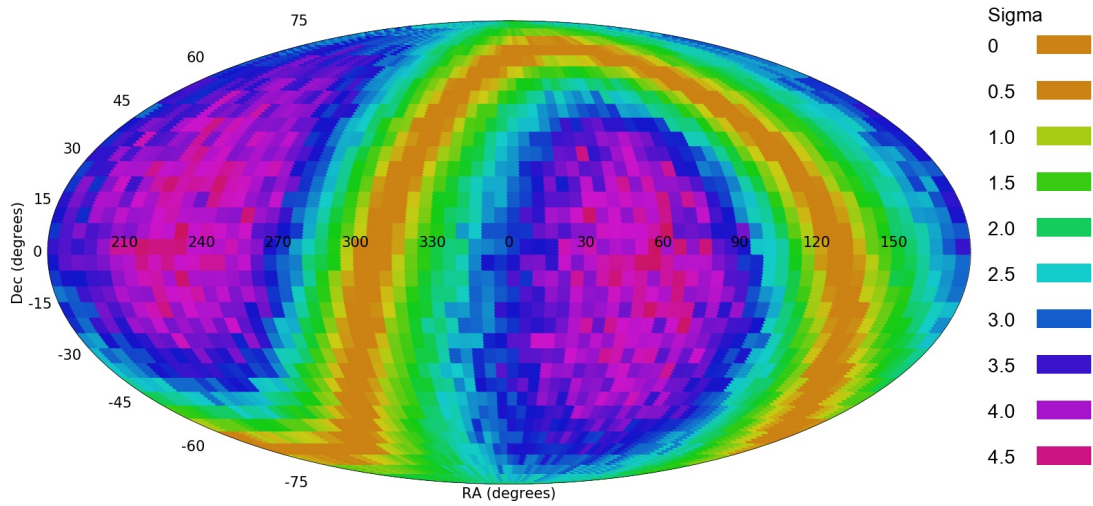
Probability of a dipole axis in galaxy spin direction distribution in SDSS DR14 (Shamir, 2020b). The dataset includes $6.4 \cdot 10^4$ galaxies separated automatically by their spin direction (clockwise or counterclockwise).



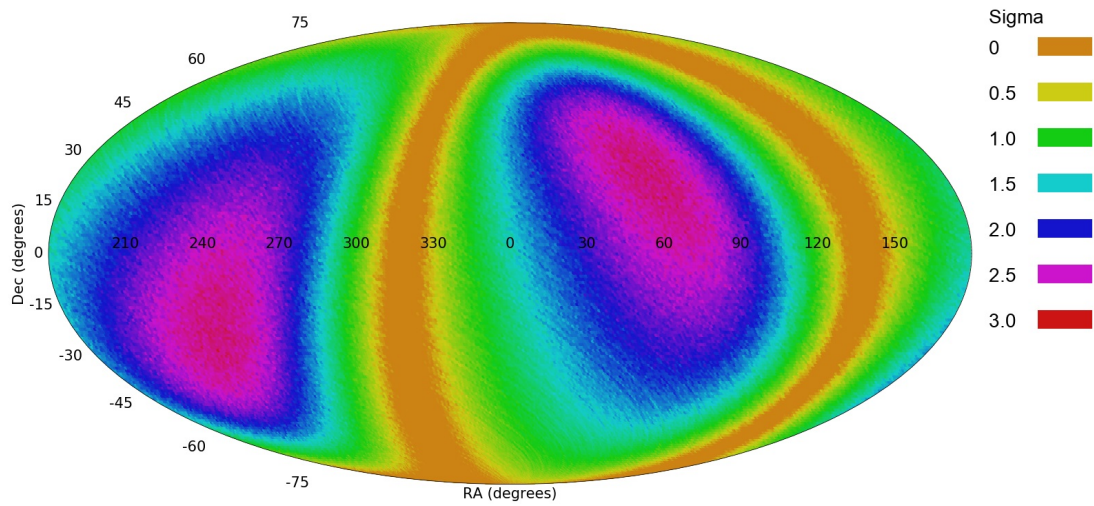
Analysis of $\sim 8.7 \cdot 10^3$ galaxies imaged by HST. The galaxies were annotated manually, and half of the galaxies were mirrored to ensure no human bias (Shamir, 2020a).



The likelihood of a dipole axis in all possible (RA,Dec) combinations in $3.8 \cdot 10^4$ Pan-STARRS galaxies (Shamir, 2020b).



The likelihood of a dipole axis in all possible (RA,Dec) combinations in $8 \cdot 10^5$ galaxies from DESI Legacy Survey. The most likely dipole axis has statistical strength of 4.7sigma. Interestingly, it peaks very close to the CMB cold spot.



The likelihood of a dipole axis in all possible (RA,Dec) combinations in $3.3 \cdot 10^4$ SDSS galaxies that have the same redshift distribution as the Pan-STARRS galaxies above (Shamir, 2020b).

Multiple experiments showed that the location of the most likely dipole axis depends on the redshift of the galaxies (Shamir, 2020a; Shamir, 2020b). When the redshift distribution is normalized, different sky surveys show very similar locations of the most likely dipole axis.

All of these sky surveys also show a higher number of galaxies spinning clockwise in one hemisphere, and a higher number of galaxies spinning counterclockwise in the opposite hemisphere. For instance, the following table shows the distribution of the galaxies in the DESI Legacy Survey.

	Hemisphere 1	Hemisphere 2
# Clockwise	264707	139719
# Total	527266	280632
Asymmetry	0.004	-0.004
P	0.0015	0.01216

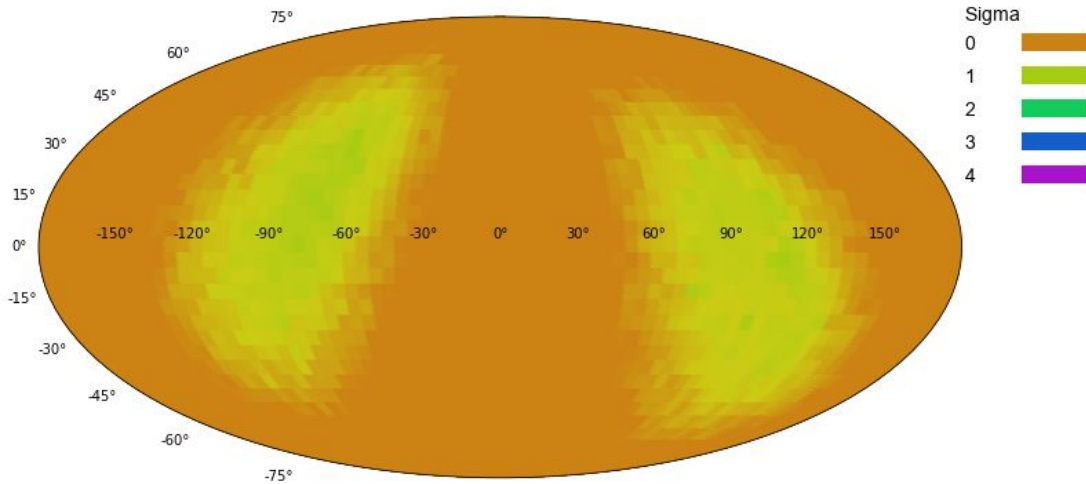
The distribution shows that in the hemisphere (0° - 150° V 330° - 360°) there is a statistically significant higher number of clockwise galaxies, and in the opposite hemisphere (150° - 330°) there is a statistically significant higher number of counterclockwise galaxies.

The DESI Legacy Survey is by far larger than all previous studies, with $\sim 8 \cdot 10^5$ galaxies classified by their spin direction. It also provides data mainly from the Southern hemisphere.

Possible causes of error

The algorithm used to annotate the galaxies is mathematically symmetric. It is based on clear rules, and does not rely on machine learning or complex data-driven rules that are determined automatically during the training process, and therefore there is no guarantee these rules are symmetric.

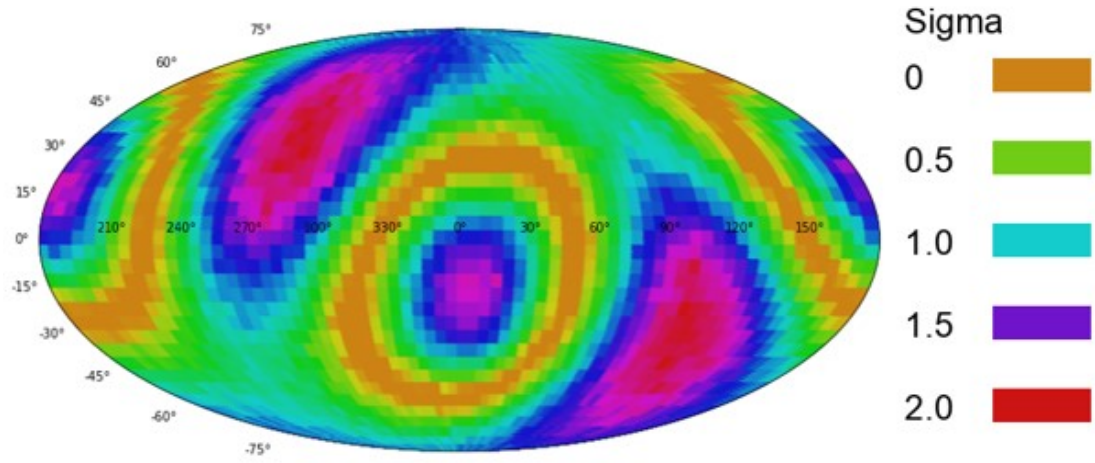
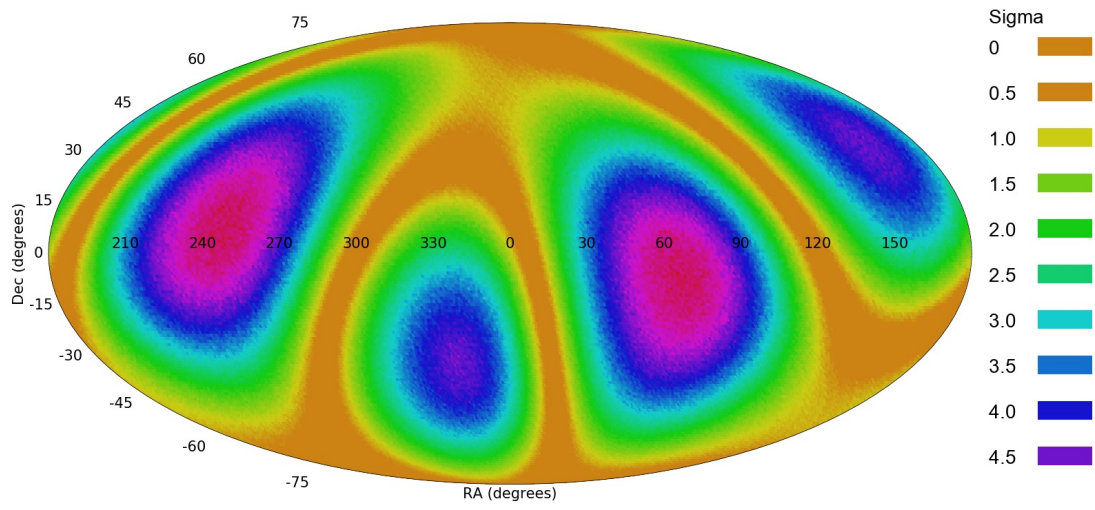
Repeating the experiments after mirroring all galaxy images led to the exact same results, but inverted. The following figure shows an analysis such that the galaxies were assigned with random spin direction. That response is consistent with all algorithms.



The annotation of all galaxies was done with a model-driven deterministic algorithm (Shamir, 2011) that is not based on machine learning or any other form of complex non-intuitive data-driven rules. The algorithm is mathematically symmetric (Shamir, 2011). All galaxies were annotated using the same code and the same computer to ensure no differences in the libraries can have an effect.

Analysis of possible errors can be found in (Shamir, 2021b). Because the annotation algorithm is symmetric, an error in the classification is expected to be distributed equally between clockwise and counterclockwise annotations. In any case, an error is expected to be consistent throughout the sky, and definitely not expected to "flip" in opposite hemispheres. That has been studied theoretically, but also empirically by adding intentional error to some of the annotations. All galaxies and all hemispheres were analyzed by the same code and the exact same computer to avoid any unexpected differences. A complete analysis is described in (Shamir, 2021b).

The distribution can also fit in quadrupole alignment. The following figures show the quadrupole alignment of SDSS and Pan-STARRS galaxies. The profile of the asymmetry distribution is in very close agreement in both telescopes.



AUTHOR INFORMATION

Lior Shamir, Kansas State University

ABSTRACT

Vera Rubin Observatory will generate visual data of extra-galactic objects at unprecedented scales. Since manual analysis of these data using manual analysis is highly impractical, turning these data into knowledge requires efficient algorithms that can practically handle real-world data. Here I show several different tasks related to automatic visual analysis of extended extra-galactic objects, and demonstrate the real-world efficacy of these methods using data from existing instruments such as SDSS, Pan-STARRS, HST, and DESI Legacy Survey. Tasks include basic classification of galaxy morphologies to produce large catalogs, automatic identification of outlier galaxies, automatic visual query-by-example of galaxies, and analysis of large-scale distribution of the spin directions of spiral galaxies. The application of these methods to data from existing digital sky surveys led to data products and the identification of objects that are virtually impossible to identify and collect without using automation.

REFERENCES

- 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.
- Goddard, H., Shamir, L., A catalog of broad morphology of Pan-STARRS galaxies based on deep learning, *The Astrophysical Journal Supplement Series*, 251(2), 28, 2020.
- Kuminski, E., Shamir, L., Computer-generated visual morphology catalog of ~3,000,000 SDSS galaxies, *The Astrophysical Journal Supplement Series*, 223(2), 20, 2016.
- Shamir, L., Asymmetry between galaxies with clockwise handedness and counterclockwise handedness, *Astrophysical Journal*, 823(1), 32, 2016.
- 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., Automatic identification of outliers in Hubble Space Telescope galaxy images, *Monthly Notices of the Royal Astronomical Society*, 501(4), 5229-5238, 2021a.
- Shamir, L., Analysis of the alignment of non-random patterns of spin directions in populations of spiral galaxies, *Particles*, 4(1), 11-28, 2021b.
- Shamir, L., Galaxy spin direction distribution in HST and SDSS show similar large-scale asymmetry, *Publications of the Astronomical Society of Australia*, 37, e053, 2020a.
- Shamir, L., Patterns of galaxy spin directions in SDSS and Pan-STARRS show parity violation and multipoles, *Astrophysics and Space Science*, 365, 136, 2020b.
- Shamir, L., Large-scale asymmetry between clockwise and counterclockwise galaxies revisited, *Astronomical Notes*, 341(3), 324-330, 2020c.
- Shamir, L., Automatic detection of full ring galaxy candidates in SDSS, *Monthly Notices of the Royal Astronomical Society*, 491(3), 3767-3777, 2020d
- Shamir, L., Asymmetry between galaxies with different spin patterns: A comparison between COSMOS, SDSS, and Pan-STARRS, *Open Astronomy*, 29, 2020e.

Shamir, L., Large-scale photometric asymmetry in galaxy spin patterns, *Publications of the Astronomical Society of Australia*, 34, e044, 2017a.

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

Shamir, L., Colour asymmetry between galaxies with clockwise and counterclockwise handedness, *Astrophysics and Space Science*, 362, 33, 2017c.

Shamir, L., Wallin, J., Automatic detection and quantitative assessment of peculiar galaxy pairs in Sloan Digital Sky Survey, *Monthly Notices of the Royal Astronomical Society*, 443(4), 3528-3537, 2014.

Shamir, L., Handedness asymmetry of spiral galaxies with $z < 0.3$ shows cosmic parity violation and a dipole axis, *Physics Letters B*, 715, 25-29, 2012.

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

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

Shamir, L., Ganalyzer: A tool for automatic galaxy image analysis, *The Astrophysical Journal*, 736(2), 141, 2011.