Accelerating the Search for Rare Gems in Big Data: A Zooniverse Case Study using Citizen Science & Machine Learning

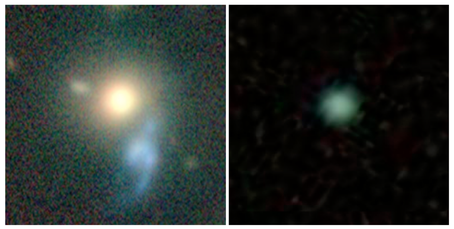
Kameswara Bharadwaj Mantha[1], Hayley Roberts[1], Lucy Fortson[1], Chris Lintott[2], Hugh Dickinson[3], William Keel[4]
Ramanakumar Sankar[5], Coleman Krawczyk[6], Brooke Simmons[7], Mike Walmsley[8], Izzy Garland[7], Jason Shingirai Makechemu[7], Laura Trouille[9], Cliff Johnson[9]

University of Minnesota [1] University of Minnesota-Twin Cities, [2] University of Oxford, [3] Open University, [4] University of Alabama, [5] University of California Berkeley, [6] University of Portsmouth, [7] Lancaster University, [8] University of Toronto, [9] Adler Planetarium & Zooniverse.

Finding Scientifically Interesting and Rare Objects

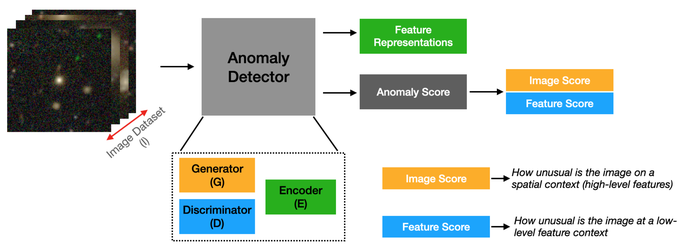
Crowdsourced research through citizen science (CS) platforms such as Zooniverse have been successful in reaching millions of individuals in an effort to help with astronomy projects over the past decade and enabled public engagement with a wide variety of data products from pioneering scientific instruments. The Galaxy Zoo project enabled the discovery of the then unknown-unknown astrophysical objects such as the Hanny's Voorverpje (Lintott+09; left figure below) Green Pea galaxies (Cardamone+09; right figure below), highlighting the role of humans in the identification and characterization of such unknown-unknowns in a dataset.

1. Artificial Intelligence (AI) strategies have become the new norm in several data processing pipelines, especially in the ever-increasing big-data astronomy paradigm.
2. Recent unsupervised AI approaches such as anomaly detection (AD) and recommendation engines (RE) have become popular to efficiently explore the unknown/rare discovery space in big data.
3. However, AI alone lags behind human-driven approaches in several different avenues, highlighting the strong need for efficient strategies to involve the crowd and AI together for this purpose.
4. But first, there is a strong need to first understand relationship betwen human & machine identified anomalous samples.

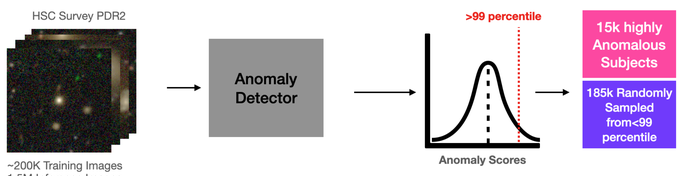


Unsupervised Anomaly Detection Framework

We designed an unsupervised anomaly detection model taking inspiration from a network called fast-AnoGAN (Schlegl et al., 2019), a Generative Adversarial Network (GAN; Goodfellow et al., 2014) framework applied to medical imaging and also astronomical images (Storey-Fisher et al., 2021). This model comprises two separate models: A Wasserstein GAN with gradient penalty (wGAN-GP) and an Encoder as illustrated in the figure below



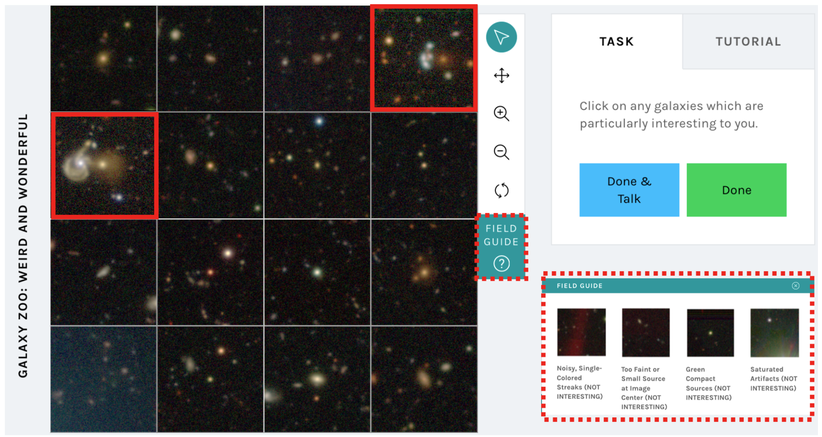
We trained our anomaly detection model on $\sim$ 200,000 images containing galaxies ($18.5 < i < 21.5$) from the Subaru Hyper-Suprime Cam Survey (HSC) and subsequently applied it on 1.5 Million images. As such, we then constructed a sample of $\sim$ 200,000 images such that 15,000 were deemed highly anomalous by our model and the remaining randomly sampled from the overall dataset.



The Galaxy Zoo: Weird & Wonderful Project

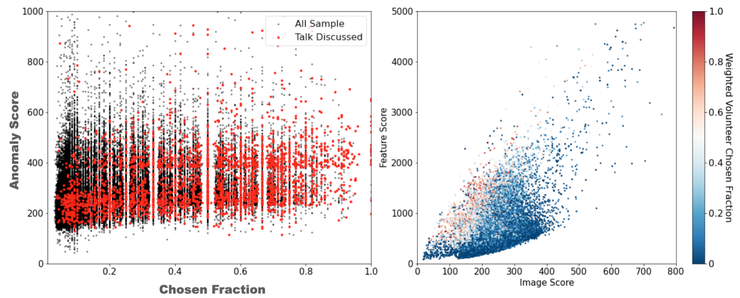
To understand synergy between the machine and human based identification of anomalous objects (that may be of scientific interest) in large datasets, we launched and finished the Galaxy Zoo: Weird & Wonderful citizen science project on Zooniverse.

Overall $\sim$ 2000 citizen science volunteers participated in this project, where each of them were shown a grid of 16 randomly pooled images at a time and are asked to select any that they find to be interesting (see illustration of the interface below). Volunteers have an option to further discuss any seen images in the "Talk" discussion boards and provide #tags.

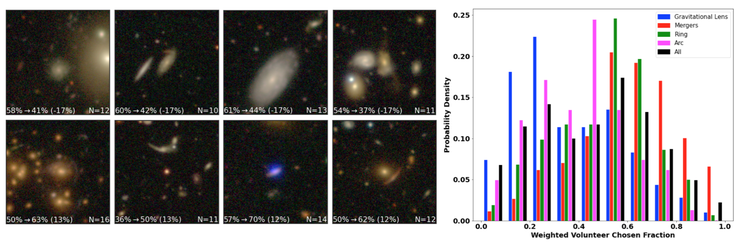


We computed the **Chosen Fraction (CF)** as the ratio of number of volunteers selecting a particular image to the total number of volunteers that have seen it. We assessed the relationship between CF and the machine based image, feature, and total anomaly scores (see the following figure below).

We find no appreciable correlation between the anomaly score and chosen fraction. However, interestingly, images with higher feature scores tend be preferentially selected by the volunteers.



We found that objects/images containing different complex characteristics had **a lower CF** compared to others as a consequence of different levels of volunteer experience. As such, to account for this, we computed the **experience-weighted chosen fraction (w.CF)** by up-weighting the selections by volunteers who had substantial participation (> 100 classifications) in the original Galaxy Zoo project.

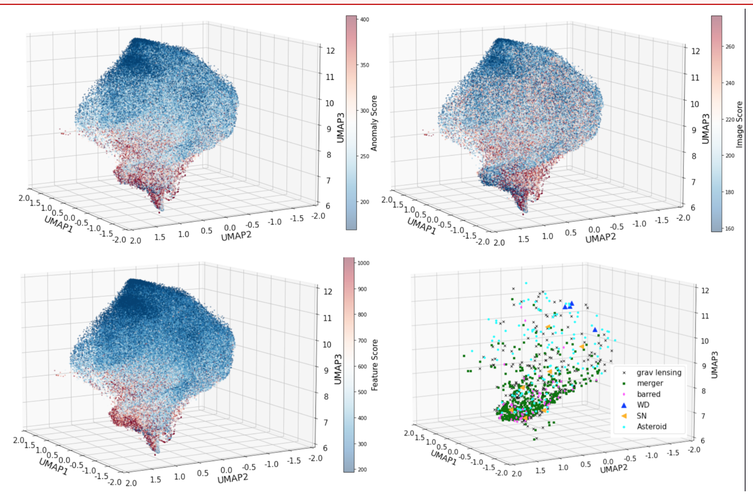


Experience weighting of the chosen fraction yielded in an up-weighting of certain astrophysically relevant characteristics (see bottom panel in the collage)

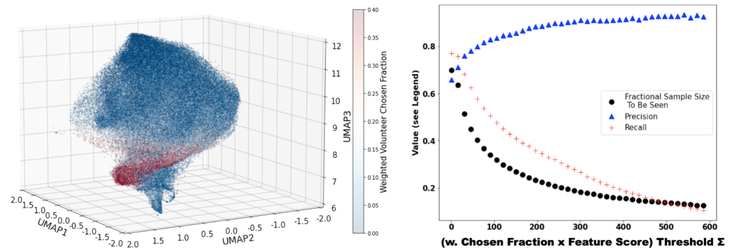
Comparing ML and GZ:W&W Outcomes

We processed the anomaly detection model based feature representations of the $\sim$ 200,000 images using dimensionality reduction and projection method called UMAP and assessed the various machine model and GZ:W&W derived metrics (anomaly scores, w.CF) in the following illustration.

We notice that the subjects with relatively high anomaly scores preferentially populate the lower portion of the UMAP space. However, it is worth noting that images with high feature scores and image scores occupy different locii in UMAP space, where high feature score images being more tighly distributed.



We explored using the weighted chosen fraction and feature scores as metrics to preselect from a previously unseen dataset if an image is to be considered interesting or not. We find that the product of the feature score and chosen fraction serves as a better metric to distinguish images in the feature space, while also ensuring that the selected number of images by the decision boundary is small ($\sim$ 20% of a total sample).



Summary

In the era of the ever-growing big-data problem of astronomy, the gap between data collection and analysis becomes a significant barrier, especially for teams searching for rare scientific objects. While machine learning (ML) methods serve as a swift means to automatically parse these large data sets, they lag behind in robustly identifying those that are scientifically interesting, a task which humans are inherently skillful. Human-in-the-loop (HITL) strategies that effectively combine the relative strengths of citizen science (CS) and ML offer an effective path forward, but employing such methods first requires us to better understand the relationship between human vs. machine identified samples. We present a case study of successfully using ML and CS through the Galaxy Zoo: Weird & Wonderful project. Using insights from this work, we present recommendations on operationalization of human-machine collaborative frameworks towards addressing the challenges of big data in astronomy.

References

- Cardamone et al., 2009, MNRAS, 399, 1191.
- Lintott et al., 2009, MNRAS, 399, 129.
- Goodfellow et al., 2020, ACM, 63, 139.
- Schlegl et al., 2019, Med. Image Analysis, 54, 30.
- Storey-Fisher et al., 2021, 508, 2946.

Acknowledgements

KBM, LF, HR would like to acknowledge partial support for this work from the following grants: NSF IIS 2006894 and NASA Award 80NSSC20M0057. CJL acknowledges support from the Sloan Foundation. RS acknowledges partial support from NASA Award 80NSSC22K0804. We also acknowledge the tremendous work done by the volunteers on the Galaxy Zoo: Weird & Wonderful project.