Exploring the HCV

Oscar Scholin, Graham Hirsch, and Prof. Elijah Quetin

Pomona College

Abstract

In this project, we attempt to analyze the Hubble Catalog of Variables (HCV), a subset of the Hubble Source Catalog (HSC) [Bon+19]. We pursue two main avenues of research: (1) a restrictive search for a particular astrophysical object and (2) an unrestrictive search designed to identify all unique types of variables within the catalog. For (1), we create an algorithm based around searching for a Tidal Disruption Event (TDE), the result of the overcoming of a star's self-gravity by the extreme gravitational pull of a supermassive black hole ($M_{BH} \lesssim 10^8 M_\odot$) [Gez21]. Our implementation resulted in the detection of two confirmed supernovae as well as two objects previously tagged as transient but their cause unknown near the centers of galaxies–and of these two, one strong TDE candidate. We have questions pertaining to the construction of the HCV whose answers will aid in the verification of this algorithm. For (2), we create a series of functions intended to reflect the variability of an object and implement several unsupervised machine learning algorithms to classify all the objects in the HCV [Sok+17]. While the clustering results were ultimately deemed nonsensical, we believe there are opportunities to further refine our approach. We believe that the HCV is a potentially useful tool that warrants both user forbearance and further analysis.

What is the HCV?

The Hubble Space Telescope has been imaging our universe in a variety of different camera and filter combinations for over 30 years [Bon+19]. During that time, it has undergone numerous upgrades and seen 3 catalog releases (the Hubble Source Catalog, HSC) attempting to identify and track the same objects imaged over time albeit in different filters and from different angles [Bon+19]. Recently a group from Athens, led by Dr. Bonanos, helped to narrow down the third iteration of the HSC from 3.7 million total sources to 84,428 variable candidates [Bon+19]. They accomplished this feat by only including objects that had ≥ 5 recorded observations, a magnitude ≥ 27, and by requiring median absolute deviation (MAD) $> 5\sigma$ as a robust variability index for their detection algorithm [Bon+19].

Spatial Distribution of HCV Objects

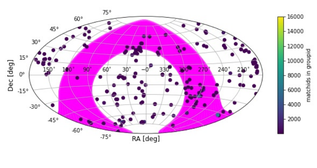

Fig. 1: A plot of all the groupids (group identifiers) in the HCV plotted using the Aitoff all-sky spherical projection, colored to show how many matchids (unique object identifiers) they contain. The magenta illustrates a first order approximation of the plane of the Milky Way, assuming a constant galactic latitude of $|b| \leq 20$ degrees [Fre15].

References

Bla+17 N. Blagorodnova et al. "iPTF16fnl: a faint and fast tidal disruption event in an E+A galaxy". In: The Astrophysical Journal 844.1 (July 2017). arXiv:1703.00965 [astro-ph], p. 46. issn: 1538-4357. doi:10.3847/1538-4357/aa7579. url: http://arxiv.org/abs/1703.00965 (visited on 08/22/2022).

Sok+17 K. V. Sokolovsky et al. "Comparative performance of selected variability detection techniques in photometric time series data". In: Monthly Notices of the Royal Astronomical Society 464 (Jan. 2017). ADS Bibcode: 2017MNRAS.464..274S, pp. 274–292. issn: 0035-8711. doi: 10.1093/mnras/stw2262. url: https://ui.adsabs.harvard.edu/abs/2017MNRAS.464..274S(visited on 08/19/2022).

PSG18 Ilya N. Pashchenko, Kirill V. Sokolovsky, and Panagiotis Gavras. "Machine learning search for variable stars". In: Monthly Notices of the Royal Astronomical Society 475.2 (Apr. 2018). arXiv:1710.07290 [astro-ph], pp. 2326–2343. issn: 0035-8711, 1365-2966. doi:10.1093/mnras/stx3222. url: http://arxiv.org/abs/1710.07290 (visited on 08/19/2022).

Bon+19 A. Z. Bonanos et al. "The Hubble Catalog of Variables (HCV)". In: Astronomy Astrophysics 630 (Oct. 2019). arXiv:1909.10757 [astro-ph], A92. issn: 0004-6361, 1432-0746. doi: 10.1051/0004 - 6361/201936026. url: http://arxiv.org/abs/1909.10757 (visited on 08/11/2022).

Hos+19 Zafiirah Hosenie et al. "Comparing Multiclass, Binary, and Hierarchical Machine Learning Classification schemes for variable stars". In: Monthly Notices of the Royal Astronomical Society 488.4 (Oct. 2019), pp. 4858–4872. issn: 0035-8711. doi: 10.1093/mnras/stz1999. url: https://doi.org/10.1093/mnras/stz1999 (visited on 08/19/2022).

Bed20 Renesh Bedre. DBSCAN in Python (with example dataset). en. Mar.2020. url: https://www.reneshbedre.com/blog/dbscan-python.html (visited on 08/22/2022)

Fre+20 K. Decker French et al. "The Host Galaxies of Tidal Disruption Events". en. In: Space Science Reviews 216.3 (Apr. 2020), p. 32. issn: 0038-6308, 1572-9672. doi: 10.1007/s11214-020-00657-y. url: http://link.springer.com/10.1007/s11214-020-00657-y (visited on 08/22/2022).

Gez21 Suvi Gezari."Tidal Disruption Events". In: Annual Review of Astronomy and Astrophysics 59.1 (Sept. 2021). arXiv:2104.14580 [astro-ph], pp. 21–58. issn: 0066-4146, 1545-4282. doi: 10.1146/annurev-astro-111720-030029. url: http://arxiv.org/abs/2104.14580 (visited on 08/19/2022).

Dev22 Google Developers. Machine Learning Glossary. en. 2022. url: https://developers.google.com/machine-learning/glossary (visited on 08/22/2022).

LJK22 Zheyu Lin, Ning Jiang, and Xu Kong. "The Prospects of Finding Tidal Disruption Events with 2.5-Metre Wide-Field Survey Telescope (WFST) Based on Mock Observations". In: Monthly Notices of the Royal Astronomical Society 513.2 (May 2022). arXiv:2204.01615 [astro-ph], pp. 2422–2436. issn: 0035-8711, 1365-2966. doi: 10.1093/mnras/stac946. url: http://arxiv.org/abs/2204.01615 (visited on 08/22/2022).

CBP2015 cosmos r 20111218 24A. url: http://simbad.cds.unistra.fr/simbad/sim-id (visited on 08/22/2022).

Dev Google Developers. Create a Manual Similarity Measure | Machine Learning. en. url: https://developers.google.com/machine-learning/clustering/similarity/manual-similarity (visited on 08/22/2022).

SKLa SKLearn.sklearn.cluster.DBSCAN.en.url: https://scikit-learn/stable/modules/generated/sklearn.cluster.DBSCAN.html (visited on 08/22/2022).

SKLb SKLearn.sklearn.cluster.KMeans.en.url: https://scikit-learn/stable/modules/generated/sklearn.cluster.KMeans.html (visited on 08/22/2022).

SKLc SKLearn.sklearn.manifold.TSNE.en. url: https://scikit-learn/stable/modules/generated/sklearn.manifold.TSNE.html (visited on 08/22/2022).

BED03 Vasily Belokurov, N. Wyn Evans, Yann Le Du, Light-curve classification in massive variability surveys — I. Microlensing, Monthly Notices of the Royal Astronomical Society, Volume 341, Issue 4, June 2003, Pages 1373–1384, https://doi.org/10.1046/j.1365-8711.2003.06512.x (visited on 08/22/2022)

Fre15 Freeman, K. (2015). The Galactic Bulge. In: Freeman, K., Elmegreen, B., Block, D., Woolway, M. (eds) Lessons from the Local Group. Springer, Cham. https://doi-org.ccl.idm.oclc.org/10.1007/978-3-319-10614-4_3 (visited on 08/25/2022)

Restrictive Search Methodology

Originally, we began by trying to answer whether there were Tidal Disruption Events (TDEs) hidden within the HCV by performing a reductive search. TDEs are the result of a star passing within the Roche radius of a supermassive black hole with $M_{BH} \lesssim 10^8 M_\odot$, which becomes pulled apart due to the massive tidal forces; part of this matter becomes trapped in an orbit around the black hole and is accreted onto a disk, which releases electromagnetic radiation [Gez21]. This radiation has a characteristic shape when plotted as brightness (magnitude) over time: a rapid flare up and exponentially decaying flare down [Gez21]. Crucially, they also occur in the centers of galaxies [Gez21]. To perform our search, we quantified each lightcurve by computing the difference in magnitude between consecutive points and determining whether it exceeded the magnitude error in absolute value. If it did, we assigned it a '+' or a '-' depending on whether the object got brighter or dimmer; if it did not, we assigned it a '0'. We consolidated consecutive events of type flare up, flare down, or flat, and used this information to construct a matrix for every object in the HCV. We then imposed six temporal, spatial, and magnitude restrictions tailored to the search for TDEs, using current theoretical calculations and observed TDEs [Gez21],[LJK22],[Bla+17].

Lightcurve for Detected Supernova

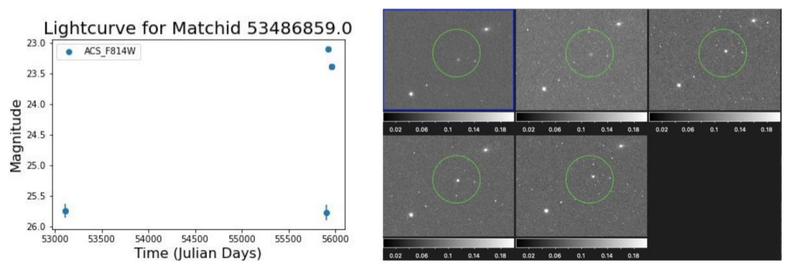

Fig. 2: The left plot shows the lightcurve (magnitude over time) of candidate 53486859, a confirmed supernova detected with our algorithm. On the right is a matrix of the images used to produce the plot on the left, which evolves in time from left to right, top to bottom.

Is the HCV useful for finding TDE candidates?

A paper published in 2020 lists the coordinates (RA and Dec) for 43 known TDE candidates [Fre+20]. As a first step, we investigated whether any of these objects were present in the HCV–we found no data matching any of the coordinates given in the paper. Thus, we sought to find previously undiscovered potential TDE objects using the algorithm described above. In the end, we produced a list of 21 candidates. We were able to manually verify a complete flare up and down on one candidate (53486859), shown above, which when we cross-referenced the ra and dec using SIMBAD turned out to be a confirmed supernova [SIM]. Further research with SIMBAD yielded a potential Cepheid variable, a pulsar, two supernovae, an x-ray source though likely too far away from the target coordinates, several non-variable stars and galaxies (one of which (38467177) does show a flare, but the object practically appears out of nowhere and is inconsistent with reported magnitude values, and another (49953382) that shows no visible flare), and most interestingly two transient events (9095412 and 74947430) located 0.19 and 0.22 arcseconds away from their target coordinates; both situated in galaxies within a galaxy cluster. TDEs are a sub-type of transient, making transients of extreme interest for us [Fre+20]. For candidate 9095412, the galaxy of interest appears be to viewed edge on, and the flare up appears to occur left of center, though it is not entirely obvious. However, object 74947430 illustrates a clear flare down, making it a stronger candidate than 9095412. Therefore these two objects, in particular 74947430, may make fruitful subjects of further investigation and validation. We also note that in the search for black holes, microlensing events, the result of light bending with spacetime due to the presence of the black hole, may be a more lucrative area of research vis-a-vis TDEs as their probability is much higher–instead of the 1 star every 1000 years that becomes disrupted by the black hole, the prerequisite for a microlensing event means that light only needs to pass around the black hole and make it to Earth [Gez21],[BED03].

Potential TDEs

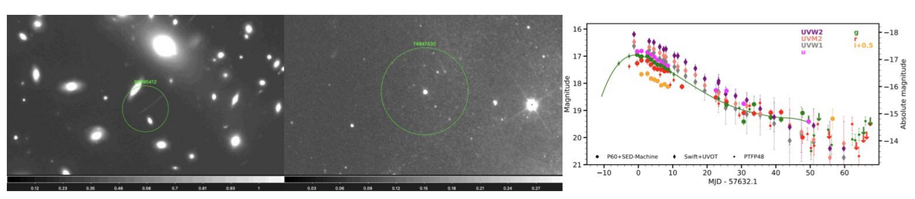

Fig. 4: On the far left is object 90995412 and in the middle is object 74947430, candidates detected via our algorithm and listed as 'transients' by SIMBAD–these are our TDE candidates. On the right is a lightcurve of an actual TDE [Bla+17].

Unrestrictive Search Methodology

With the vast amount of data in astronomy, in particular for the search for variable objects, effective computer algorithms are needed for rapid processing and analysis. An alternative to the more traditional method described to the left, which involves sifting through a set of potential objects to arrive at a more refined candidate list to find a given object, machine learning programs can be employed in an attempt to study all objects in a given dataset simultaneously. Several papers have utilized supervised machine learning to detect and classify variable stars [PSG18], [Hos+19]. The process of *supervised* machine learning requires labeled training data—that is, the users have to feed the computer information with a labeled class, e.g. an image and a label "dog", many times over to produce satisfying results [Dev22].
However, obtaining a sufficient quantity of labeled training data, especially in the field of variable star searching is hard to come by. Hence the appeal of *unsupervised* machine learning, or clustering, which does not require the data to be labeled and instead determines both the number and distribution of clusters [Dev22]. To gather data suitable for clustering, we followed [Sok+17] and [PSG18] and defined 23 variability indices, ranging from more generic statistical measures like MAD and Chi Square statistic to custom-build astronomical functions such as Stetson's J and K. In order to facilitate the learning, we normalized the data and performed the t-SNE dimensionality reduction technique, which is a nonlinear method of projecting higher dimensions to lower ones, in this case two [SKLc]. With this t-SNE data, we could implement clustering. We chose to use two algorithms: KMeans and DBSCAN [SKLb],[SKLa]. Simply put, DBSCAN is superior at identifying irregularly-shaped clusters and even detecting outliers though it requires defining several tunable parameters (called hyperparameters) to be utilized effectively, whereas KMeans tends to identify spherically-symmetrical clusters but us much easuer to use [Bed20]. Both were considered for final analysis.

Graph of T-SNE Plot

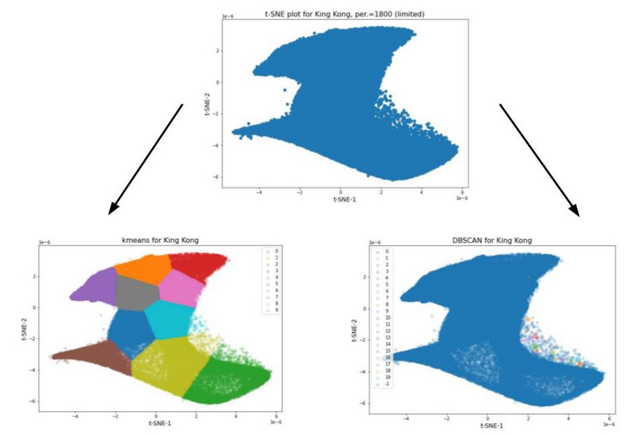

Fig. 3: The top image shows the result of performing a t-SNE dimension reduction from the 16 variability indices selected for further analysis (from the originally chosen 23) with a perplexity value of 1% the size of the starting data (DBSCAN hyperparameter). Moving down, on the left is the result of the KMeans clustering algorithm, and DBSCAN on the right.

Can we use clustering to characterize the HCV?

When we fed a matrix of 23 columns and 179871 rows (one for each matchid and filter combination), we found 10 clusters with KMeans and 20 plus outliers with DBSCAN on the t-SNE data. When we compared sample similar and dissimilar objects (according to the shape of their lightcurves: flare up and down, flare up or down, periodic, etc), we found an accuracy of 2/5 for both KMeans and DBSCAN out of the five sample cases. However, for KMeans each of the test objects, regardless of their lightcurve type, was assigned to different clusters. Moreover, for DBSCAN, each of the 5 test objects was assigned to cluster 0 except for one. In an attempt to improve clustering results, we prepared a subset of features following [PSG18]; here the accuracy improved to 3/5 in both cases. However, again, in KMeans all but one pair was assigned to different clusters, each of approximately equal size (the largest accounts for 15%); in DBSCAN all were assigned to cluster 0, which accounts for 99.8% of the data. Using the full 23 indices and the subset of 16, the conventional metric of similarity of two vectors, Euclidean distance, yielded non-sensible results: 2/3 of similar objects showed high distances and 1/2 of dissimilar objects showed low distances [Dev]. For the t-SNE data, the two dissimilar objects were 3 and 5 times further away than two of the similar objects; however, one of the similar objects was 4 times further away than the other two dissimilar ones.
We can undertake further research to first acquire and test a larger sample of test objects to eliminate the effects of possible outliers, to verify our implementation of the variability indices, and to experiment with different combinations of these indices. It is also possible that with the limited number of data points per light curve (25% of matchid-filter combinations have only 5 observations, 60% have ≤ 10, and only 30% have ≥ 15), many of these variability indices simply do not well reflect the data, in which case we will need to investigate alternate indices and perhaps even a more simple, literal interpretation of the lightcurve along the lines of the restrictive search methodology.