A Granular Method for Finding Anomalous Light Curves and their Analogs

Kushal Tirumala* Caltech
J. Rafael Martínez-Galarza Harvard
Federica B. Bianco University of Delaware
Dennis Crake Harvard

Ashish A. Mahabal Caltech
Matthew J. Graham Caltech
Daniel Giles The SETI Institute

Abstract

Anomalous light curves indicate rare and as yet unexplainable phenomena associated with astronomical sources. With existing large surveys like the Zwicky Transient Facility (ZTF) [1], and upcoming ones such as the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST) [2] that will observe astrophysical transients at all time scales and produce archival light curves in the billions, there is an immediate need for methods that reveal anomalous light curves. Previous work explores anomalous light curve detection, but little work has gone into finding analogs of such light curves. That is, given a light curve of interest, can we find other examples in the dataset that behave similarly? We present such a pipeline that (1) identifies anomalous light curves, and (2) finds additional examples of specific rare classes, in a large corpora of light curves. We apply this method to *Kepler* data, finding around 5000 previously unknown anomalies, and present a subset of these anomalies along with their potential astrophysical classification.

1 Introduction

Our understanding of the Universe has always profited from serendipitous, unexpected astronomical discoveries. When studying stars, astronomers frequently turn to the star's light curve which is a time series of an object's source brightness. A particular area of interest in modern astronomy is finding anomalous light curves, which indicates rare and interesting objects. Recent examples include the discovery of peculiar light curves in *Kepler* data, such as KIC 8462852, commonly known as Boyajian's star [3], the discovery of the interstellar object 1I/'Oumuamua [4], and the detection of quasi-periodic oscillations in the X-ray light curve of galaxy G 159 [5], among many others. These unexpected discoveries often challenge current theories and push us to form new hypotheses.

Significant previous work has gone into time domain-based studies of astrophysical phenomena. Several unsupervised and semi-supervised algorithms for astrophysical anomaly detection have been used, including approaches that use Euclidean proximity/clustering information to isolate anomalies [6–8], and approaches that leverage more complex representations of the data using neural networks [9–12], as well as Gaussian processes [13]. Notable effort has gone into anomaly detection in supernova surveys, in particular by the Supernova Anomaly Detection (SNAD) group, which uses Isolation Forest and active learning to boost the discovery of unusual objects [14–16]. Previous work also investigates finding analogs of a specific anomalous object such as Boyajian's star [6, 17–19].

In this work, we present a method to systematically discover anomalous light curves and their analogs in unlabeled data, and apply it to a set of 147,036 *Kepler* light curves. To the best of our

*This work was done while the author was a student at Caltech.

Fourth Workshop on Machine Learning and the Physical Sciences (NeurIPS 2021).

knowledge, no reproducible methods exist with the goal of finding analogs to anomalous light curves. We use this method to produce a set of previously unknown light curve anomalies, and provide potential anomaly types for a subset of these light curves. Results can be reproduced via code in `https://github.com/kushaltirumala/WaldoInSky`, and full details are in [20].

2 Data and Preprocessing

We use 160,000 light curves from the Kepler telescope archive, obtained from the Mikulski Archive for Space Telescopes (MAST). Each light curve was recorded from the Kepler spacecraft, with one photometry point obtained every 30 minutes. We first preprocess the original set of light curves by normalizing the fluxes to have a mean value of 1 (we divide each light curve by its mean value). We also find a number of artifacts such as single-pixel spikes or dips that can occur due to extraneous events such as cosmic rays and hot pixels. To combat this we directly replace data points outside a 3σ deviation range from the light curve, with the mean from a rolling window of 15 data points (which corresponds to 7.5 hours), which is a common preprocessing step for anomaly detection [21]. We remove light curves that we could not preprocess, leaving us with a total of 147,036 light curves. The preprocessed set covers approximately 85 days of observations, for a total of 3520 measurements per light curve. We refer to this dataset as the *dense set*. We also produce a sparse version of the dataset by uniformly random subsampling 10% of the original time stamps, and looking at the data from only those timestamps across all the light curves (we use the same timestamps for all the light curves). We refer to this dataset as the *sparse set*. With this, we can evaluate our pipeline on data representative of ground-based surveys such as LSST, which often produce unevenly sampled and sparse light curves.

3 Method

Step 1: Find groups of anomalies. We first represent the light curves as a feature vector of the light curve points themselves concatenated with values of their power spectrum. We compute the power spectrum using the Lomb-Scargle method. For the dense set, we construct periodograms using 3000 discrete frequencies, covering a logarithmic range of frequencies corresponding to periods between one hour (twice the Kepler cadence) and 90 days (approximately the duration of the observations). This range covers the typical timescales of different stellar variability phenomena [22, 23]. For the sparse set, we use 300 discrete frequencies, covering frequencies ranging from 4 hours to 90 days. Using these input features, we use the unsupervised random forest (URF) [24] method to perform anomaly detection. URF is an adaption of random forests — first, we generate a synthetic dataset by independently sampling the marginal distributions of each feature in the original dataset. Then, we train a Random Forest classifier to distinguish between the original and synthetic dataset, retrieve class predictions for the original dataset, and measure what proportion of the time data points are both in the same leaf node and the same class, across all trees in the forest. Using this as a similarity metric $S(x_i, x_j)$ between two data points x_i, x_j, we compute the average dissimilarity $1 - S(x_i, x_j)$ of a point x_i to all other data points x_j. We tune the hyperparameters of the Random Forest classifier via gridsearch on a 80-20 training-test split to maximize accuracy (exact hyperparameters are given in table 1). We then chose a threshold of 0.85 in the resulting distribution of anomaly scores (see figure 1), and take all points with scores above 0.85 as candidate anomalies[1]. We chose the cutoff score by looking at the distribution of URF scores (see figure 1), and selecting a value that separates all known anomalies from other light curves.

Step 2: Find analogs of specific light curves. We first transform all input light curves using the *DMDT* approach first introduced in [25]. In this approach, we represent the light curve as a 2D histogram by binning pairs of points into difference in magnitude (DM) and difference in time (DT) bins. We use 21 DM bins and 19 DT bins, making the output feature space a $\mathbb{R}^{21\times19}$ matrix (the specific bins we use are mentioned in Section A.1), which we flatten into 399-sized vectors. We pass these vectors to both t-Stochastic Neighbor Embedding (t-SNE) [26], and Uniform Manifold Approximation and Projection (UMAP) [27], which maps the representation into 2D space. We concatenate these vectors to get a 4D representation of the light curves. We concatenate t-SNE and UMAP embeddings as opposed to using them separately, because the two embeddings seem to capture different trends about

[1] As mentioned previously, we define "anomalies" as light curves that exhibit unexpected astrophysical behavior, not necessarily those that are statistically rare.

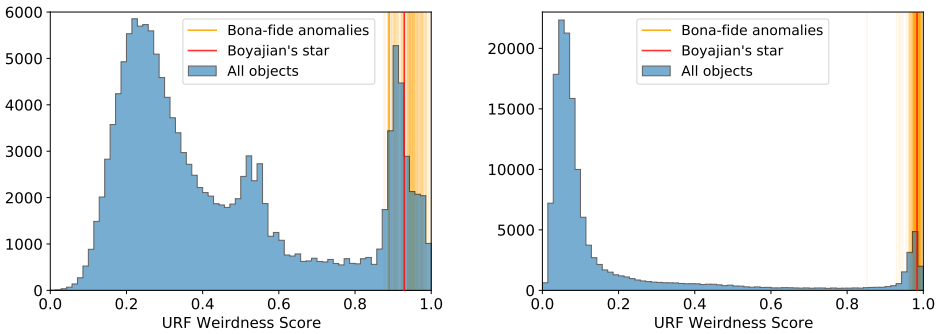

Figure 1: Histograms of URF anomaly scores measured for dense (left) and sparse (right) Kepler dense light curves. Indicated in orange and red are the scores of the known anomalies and Boyajian's star, respectively.

input light curves. As shown in figure 3, t-SNE fragments input light curves into disconnected regions in 2D space, while UMAP projects light curves into a continuous distribution.

To find analogs of a given light curve, we inspect the nearest neighbors (in Euclidean norm) in this 4-dimensional space. Notably, we found that the location of a light curve in this 4D space correlates with the independently measured URF score (we discuss this more in section 4). We find analogs using nearest neighbors because t-SNE and UMAP are formulated to preserve information about nearest neighbors. Due to lack of expert labels, it is difficult to benchmark the decision of using Euclidean distance against other time series similarity metrics; however, as we discuss in section 4, we find that Euclidean distance can differentiate between classes of objects, and previous work has established that Euclidean distance remains an accurate measure of similarity when compared to other metrics [28].

4 Results and Discussion

By definition, anomalies are unknown and not well-defined, which makes benchmarking anomaly detection methods difficult due to lack of agreed-upon baselines. Luckily, *Kepler* data has a known set of rare objects [29] we can benchmark with. These objects were found by an comprehensive literature search to identify objects across 35 variability classes. Note that these points are only "rare" with respect to the *Kepler* dataset, not necessarily with respect to the entire population of stars in our galaxy.

We measure how well our anomaly score from Section 3 is able to identify these rare objects, and summarize the results in figure 1. For the dense set, we observe that all of the rare objects of interest fall within the anomalous peak of the distribution, in that they are all in the top 18% of scores (Boyajian's star is in the top 7%). We determined the uncertainty in the score by running the method 10 times for a subset of about 20,000 objects, each time with different seeds and therefore a different generated synthetic dataset. We find that variances are typically small — for example, objects with mean anomaly scores above 0.85 have standard deviations around 2%-3%. We also note from figure 1 that even with sparse data, the URF scores still recovers known rare objects. Remarkably, about *5000 previously unidentified light curves* also fall within this anomalous peak (i.e. light curves that are anomalous and for which a class has not been assigned). We take members of known rare classes that appear in this peak, and find previously unidentified analogs. We focus on δ-Scuti stars, oscillating binary stars, eruptive RGB stars, and Long Period Variable stars, and report results in tables 2, 3, 4, 5 respectively.

It is clear that the URF score can identify objects from known, rare classes. To investigate why, we consider a Hertzprung-Russell (HR) diagram which plots the inferred luminosity of stars against their inferred temperature, and are commonly used to classify stars according to their astrophysical properties. We infer the temperature and brightness from *Gaia* measurements [30, 31] for all objects in our dataset, and color-code the resulting HR diagram according to the URF score in figure 2. We note that objects with similar URF scores are grouped into similar regions of the HR diagram, which indicates that the URF score captures underlying physical properties. The objects with the highest

URF scores are in a region of the HR called the instability strip, where we know instabilities in stars lead to periodic oscillations of different types due to having exhausting most of their hydrogen, for example δ-Scuti stars, γ-Doradus stars, and RR Lyrae stars. A majority of known rare objects lie along the main sequence (MS), which is the diagonal strip going from the top left to the bottom right. At the lower part of the MS, we see high URF scores in dwarf stars of low luminosity that show flaring behavior due to magnetic activity. Surprisingly, stars that live in the upper MS where we do not expect significant oscillations or flares also have high URF scores (this is also where Boyajian's star lies), indicating it might be a good region to look for anomalies.

Using classifications of objects from the SIMBAD database [32], we see significant overlap of the URF score between different rare classes. Among the most anomalous objects are δ-Scuti stars, for which the mode of the URF score is 0.99 (top 0.6%). Eruptive stars have a similar mode, but have a larger fraction of members with much lower anomaly scores. Eclipsing binaries have a score mode of 0.975 (top 1.6%) and so on. The overlap of URF scores between different rare classes prevents us from finding specific instances of rare objects based on URF score alone. This is where we leverage the t-SNE and UMAP embeddings — inspecting t-SNE and UMAP 2D embeddings in figure 3, we note a striking relationship between the location of a light curve in 2D space and the independently derived URF score. Members of the anomalous peak of the URF distribution occupy a well defined area in both the t-SNE and UMAP embedding plots. More importantly, when we zoom into the "anomalous areas" of the plots we see that known rare objects that exhibit similar behavior are grouped together with respect their Euclidean distance in this representation. This provides a way to differentiate between different classes of rare objects, breaking the degeneracy of the one-dimensional URF score. We also try using t-SNE and UMAP embeddings to find outliers directly by looking at Euclidean distances between points as described in section A.3. We found this does not identify members of known rare classes (such as δ-Scuti stars) like the URF score does, but it does isolate members of unknown classes (such as Boyajian's star, as shown in figure 3).

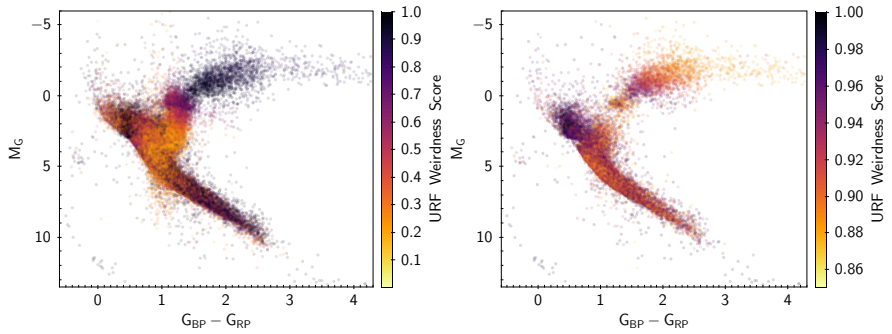

Figure 2: *Left*: The HR diagram for stars in our dataset, color-coded by the URF score for the dense set. The x-axis indicates inferred temperature and the y-axis indicates inferred luminosity. *Right*: The same diagram for objects with URF score larger than 0.85

.

We also investigate the effect of using the sparse set instead of dense set. One explanation for why sparsity affects candidate anomalies is that sparsity dampens large-scale differences in the spectral power distribution (SPD) between high frequencies and low frequencies. We see in figure 4 that the power spectrum for sparse light curves is relatively flat. Therefore, objects such as KIC 6266324 in figure 4 that have an uneven SPD might only be identified in the dense set, whereas others like KIC 9096191 that exhibit unusually high amplitude variability might still be flagged in the sparse set.

We find the relationship between the URF anomaly score, low-dimensional embeddings via t-SNE and UMAP, and astrophysical properties to indicate a possible path toward new discoveries. We are currently performing this analysis on *TESS* data to investigate how well our method generalizes to different light curve datasets.

5 Broader Impact

Within astronomy, this work is useful as a pipeline to detect groups of anomalous light curves and/or specific classes of light curves in unlabeled data. To the best of our knowledge, this is the first method

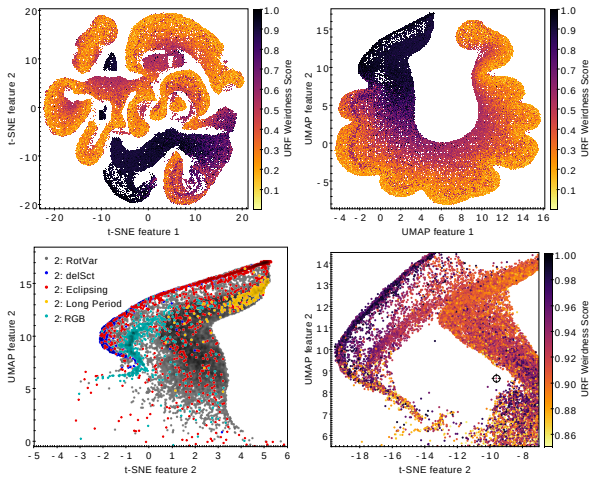

Figure 3: *Top*: The 2D embedding of the dense light curves produced by t-SNE (left) and UMAP (right). *Bottom Left*: The anomalous tip of the UMAP embedding map, with different classes of anomalous objects (rotating variables, δ-Scuti stars, eclipsing binaries, long period variables and oscillating RGB stars) indicated. *Bottom Right*: A particular projection of the 4D embeddings, in a region close to the location of Boyajian's star which is indicated by the cross mark. Shown are only points corresponding to anomalies.

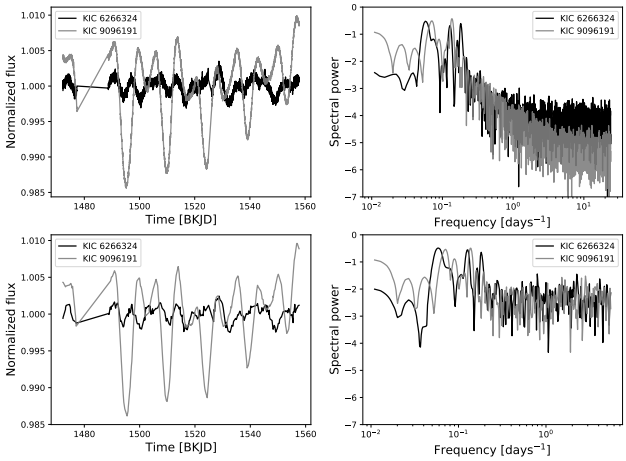

Figure 4: Light curves (left) and periodograms (right) for two objects, both the dense (top) and sparse (bottom) versions. Gray line is KIC 9096191 which is identified as an anomaly in both the dense and sparse sets. Black line is KIC 626632, which is identified as an anomaly only in the dense set. Note that the power spectrum is flatter for the sparse version of both light curves.

that incorporates finding specific classes of anomalies. This method can potentially be useful as a way to find instrumental or data reduction artifacts, as not all anomalies are astrophysical in nature. The pipeline we present can also be applied to any dataset of time-series (not necessarily astronomical data), even if the data irregular (i.e. unevenly sampled or sparse data). However, we have not applied it to different types of input data, and investigated any unintended consequences the method may have. There are potential societal harms of the method, which we share with other time-series anomaly detection methods. For example, applying the method on social network data can cause potential surveillance harms / security concerns (i.e. finding users that exhibit similar behavior or detecting anomalous user behavior). Another example is using this method to analyze time-series climate data could potentially enable certain environmental harms (i.e. finding similar geographic regions to extract resources from). We also note that using individual parts of the method does have limitations. As we describe in section 4, using only the URF score does not allow us to differentiate between different classes of stars, which is why we include the t-SNE and UMAP portion; however, using only t-SNE/UMAP embeddings does not recover known rare objects as we describe in A.1. Only together do these two methods produce good anomalous candidates for further inspection.

References

[1] Eric C. Bellm et al. The Zwicky Transient Facility: System Overview, Performance, and First Results. , 131(995):018002, January 2019.

[2] Željko Ivezić et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. , 873(2):111, March 2019.

[3] T. S. Boyajian et al. Planet Hunters IX. KIC 8462852 – where's the flux?†. *Monthly Notices of the Royal Astronomical Society*, 457(4):3988–4004, 01 2016.

[4] Karen J. Meech et al. A brief visit from a red and extremely elongated interstellar asteroid. , 552(7685):378–381, December 2017.

[5] G. Miniutti et al. Nine-hour X-ray quasi-periodic eruptions from a low-mass black hole galactic nucleus. , 573(7774):381–384, September 2019.

[6] Daniel Giles et al. Systematic serendipity: a test of unsupervised machine learning as a method for anomaly detection. *Monthly Notices of the Royal Astronomical Society*, 484(1):834–849, 2019.

[7] H. Dutta et al. *in: Proceedings of the 2007 SIAM International Conference on Data Mining*. SIAM, 2007.

[8] Marc Henrion et al. Casos: a subspace method for anomaly detection in high dimensional astronomical databases. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(1):53–72, 2013.

[9] Dalya Baron et al. The weirdest sdss galaxies: results from an outlier detection algorithm. *Monthly Notices of the Royal Astronomical Society*, 465(4):4530–4555, 2017.

[10] Alessandro Druetto et al. A deep learning approach to anomaly detection in the gaia space mission data. In *International Work-Conference on Artificial Neural Networks*, pages 390–401. Springer, 2019.

[11] Petr Škoda et al. Active deep learning method for the discovery of objects of interest in large spectroscopic surveys. *arXiv e-prints*, page arXiv:2009.03219, September 2020.

[12] Berta Margalef-Bentabol et al. Detecting outliers in astronomical images with deepgenerative networks. *arXiv preprint arXiv:2003.08263*, 2020.

[13] Haoyan Chen et al. Anomaly detection in star light curves using hierarchical gaussian processes. 04 2018.

[14] M. V. Pruzhinskaya et al. Anomaly detection in the Open Supernova Catalog. , 489(3):3591–3608, November 2019.

[15] Emille E. O. Ishida et al. Active Anomaly Detection for time-domain discoveries. *arXiv e-prints*, page arXiv:1909.13260, September 2019.

[16] Patrick D. Aleo et al. The Most Interesting Anomalies Discovered in ZTF DR3 from the SNAD-III Workshop. *Research Notes of the American Astronomical Society*, 4(7):112, July 2020.

[17] Daniel K. Giles et al. Density-based outlier scoring on Kepler data. , 499(1):524–542, November 2020.

[18] Martin Schmidt. Functorial Approach to Graph and Hypergraph Theory. *arXiv e-prints*, page arXiv:1907.02574, July 2019.

[19] M. Lochner et al. ASTRONOMALY: Personalised active anomaly detection in astronomical data. *Astronomy and Computing*, 36:100481, July 2021.

[20] J Rafael Martínez-Galarza, Federica B Bianco, Dennis Crake, Kushal Tirumala, Ashish A Mahabal, Matthew J Graham, and Daniel Giles. A method for finding anomalous astronomical light curves and their analogues. *Monthly Notices of the Royal Astronomical Society*, 508(4):5734–5756, 2021.

[21] Umaa Rebbapragada et al. Finding Anomalous Periodic Time Series: An Application to Catalogs of Periodic Variable Stars. *arXiv e-prints*, page arXiv:0905.3428, May 2009.

[22] Charlie Conroy et al. A Complete Census of Luminous Stellar Variability on Day to Decade Timescales. , 864(2):111, September 2018.

[23] Laurent Eyer et al. Variable stars across the observational HR diagram. In *Journal of Physics Conference Series*, volume 118 of *Journal of Physics Conference Series*, page 012010, October 2008.

[24] Tao Shi et al. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1):118–138, 2006.

[25] Ashish Mahabal et al. Deep-learnt classification of light curves. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8. IEEE, 2017.

[26] Laurens van der Maaten et al. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[27] Leland McInnes et al. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv e-prints*, page arXiv:1802.03426, February 2018.

[28] Joan Serra and Josep Ll Arcos. An empirical evaluation of similarity measures for time series classification. *Knowledge-Based Systems*, 67:305–314, 2014.

[29] J. Debosscher et al. Automated supervised classification of variable stars. I. Methodology. , 475(3):1159–1183, December 2007.

[30] Gaia Collaboration et al. The Gaia mission. , 595:A1, November 2016.

[31] Gaia Collaboration et al. Gaia Data Release 2. Summary of the contents and survey properties. , 616:A1, August 2018.

[32] Marc Wenger et al. The simbad astronomical database-the cds reference database for astronomical objects. *Astronomy and Astrophysics Supplement Series*, 143(1):9–22, 2000.

[33] F. Pedregosa and others. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A Appendix

A.1 Implementation Details

Table 1: Hyperparameter configurations for different methods

Part	
Method, Dataset	Hyperparameter Settings
URF, dense set	`n_tree`=700, `max_depth` $= 100$, `max_features` $= \log_2(6520)$
URF, sparse set	`n_tree`=150, `max_depth=default`, and `max_features` $= \sqrt{632}$
t-SNE, dense set	`perplexity`=200, `learning_rate`=50.0, `early_exaggeration`=5.0
t-SNE, sparse set	`perplexity`=200, `learning_rate`=50.0, `early_exaggeration`=20.0
UMAP, dense set	`n_neighbors`=200, `min_dist`=0.4, `learning_rate`=0.25
UMAP, sparse set	`n_neighbors`=200, `min_dist`=0.1, `learning_rate`=0.8

For the URF method, we use the scikit-learn implementation of Random Forest [33] and follow the algorithm described in [9]. We also use the scikit-learn implementation of t-SNE [33], and the UMAP python implementation provided in [27]. All these implementations are under a BSD License. The full dataset we use is available at `https://drive.google.com/drive/folders/1Sd31lSXkrRL2Nno30xTKd8bcD7WT7z2F` under a CC-BY 4.0 license. For *DMDT* generation, the bins we use are:

```
dm_bins = [ -0.05, -0.02, -0.01, -0.006, -0.005, -0.004,  -0.0012, -0.001,
            -0.0006, -0.0003, 0,  0.0003, 0.0006, 0.001, 0.0012, 0.003,
            0.004, 0.005, 0.006, 0.01, 0.02, 0.05 ]
dt_bins = [ -1.0/145, 1.0/47, 2.0/47, 3.0/47, 4.0/47, 6.0/47, 10.0/47,
            15.0/47, 20.0/47, 30.0/47, 40.0/47, 1.0, 1.2, 1.4, 1.5,
            1.7, 2, 2.25, 2.5, 3.0 ]
```

A.2 Processing Requirements

The machine that ran the URF method used a 2.4 GHz 8-Core Intel Core i9 Processor with 16 GB of free memory. The machine that ran the t-SNE and UMAP portion used a a GNU/Linux OS, x86-64 architecture (Linux kernel version 3.10.0-957.el7.x86-64) with 31 GB of memory. Running the URF method took around 3 hours. *DMDT* generation for dense light curves took around 36.76 hours, and *DMDT* generation for sparse light curves took around 0.23 hours. Generating t-SNE embeddings for dense light curves took about 1.04 hours and UMAP embeddings for dense light curves took around 2.36 hours. Generating t-SNE/UMAP embeddings for sparse light curves took around 2 hours. In terms of storage, the dense and the sparse *DMDT* data sets occupy ~ 450 MB each. Storing the t-SNE/UMAP embeddings for dense and sparse light curves took around 2MB each.

A.3 Using only t-SNE/UMAP embeddings to detect anomalies

We attempt using only the t-SNE and UMAP embeddings described in section 3 to recover known rare objects. We build a distribution based on the pairwise Euclidean distances between points in the 4-dimensional space of the combined embeddings. For each point, we consider its distance to the nearest neighbor. We then consider the distribution formed by normalizing these distances and take all points that are at least 3.4σ's from the mean as anomalies. This threshold was set by by inspecting the number of anomalies as a function of the threshold, and choosing the value at which the second derivative of this curve was closest to zero for both the t-SNE and UMAP. As seen in figure 5 below, the majority of our known anomalies (in fact, all of them except for Boyajian's star) fall outside of the anomaly area, with many of them having fairly low scores.

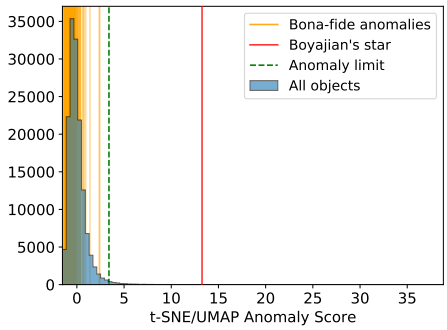

Figure 5: Histograms of anomaly scores derived as described in Section A.3. Indicated in orange and red are the scores corresponding to known rare objects, including Boyajian's star, as well as the anomaly limit (in green).

A.4 List of Full Results

Here we present a subset of the previously unknown anomalies we find, along with their predicted anomaly type. The KIC number corresponds to the Kepler ID. The G magnitude and colors have been obtained from the *Gaia* archive. Blank entries occur when we are not able to compute these values from *Gaia* data.

Table 2: Sample of unclassified anomalies similar to δ-Scuti stars

KIC	G mag	$G_{\rm BP} - G_{\rm rp}$	t-SNE f1	t-SNE f2	UMAP f1	UMAP f2	URF score
100003119			−6.247	−18.123	−1.654	11.323	1.000
100003307			−6.742	−17.711	−1.439	11.610	0.998
100003116			−6.203	−18.156	−1.710	11.328	0.998
8093353			−5.770	−18.456	−1.846	11.076	0.998
100003338			−5.307	−18.781	−1.884	10.773	0.996
10678547	0.776	0.341	−3.247	−19.316	−1.991	9.824	0.996
10612592			−6.116	−18.222	−1.652	11.294	0.995
2695999	4.145	0.994	−0.532	−18.305	−1.642	8.727	0.995
5450881	0.860	0.046	−0.669	−18.538	−1.707	8.741	0.995
10203328			−6.723	−17.734	−1.494	11.628	0.995
12603159	1.807	0.257	−4.512	−19.094	−2.071	10.403	0.994
100003285			−5.818	−18.417	−1.802	11.044	0.994
8640132	2.492	0.548	−5.998	−18.306	−1.829	11.231	0.994
11657371	2.914	0.522	−3.388	−19.384	−2.046	9.852	0.994
7467547	3.773	0.827	−2.966	−19.303	−1.966	9.713	0.994
9410674	6.803	1.383	−6.272	−18.086	−1.542	11.323	0.993
10975463			−6.054	−18.263	−1.697	11.230	0.993
100003367			−3.332	−19.340	−2.061	9.847	0.993
9897683			−5.782	−18.436	−1.767	10.998	0.993
6311520	2.722	0.807	−1.522	−19.019	−1.900	9.110	0.993

Table 3: Sample of unclassified anomalies similar to oscillating binary stars.

KIC	G mag	$G_{\rm BP} - G_{\rm rp}$	t-SNE f1	t-SNE f2	UMAP f1	UMAP f2	URF score
9594654			−7.582	10.908	5.213	17.082	0.999
9776888			−7.629	10.846	5.149	17.092	0.999
100003189			−7.667	10.798	5.091	17.053	0.997
9594468			−7.649	10.820	5.104	17.094	0.997
9956596			−7.705	10.748	4.984	16.999	0.993
100004178			−7.581	10.905	5.239	17.096	0.992
9471705			−7.698	10.757	5.028	16.999	0.992
4049858			−7.611	10.868	5.196	17.095	0.989
100004179			−7.636	10.838	5.125	17.077	0.985
9836795	11.232	2.449	−7.719	10.730	5.061	17.073	0.985
8719524	8.991	1.969	−7.606	10.876	5.223	17.093	0.985
7025613	13.028	3.764	−7.593	10.893	5.198	17.080	0.984
9716337			−7.758	10.674	4.907	16.936	0.983
9541127	2.840	0.594	−7.641	10.888	5.024	17.040	0.980
9722737	2.187	0.707	−7.717	10.721	4.960	17.000	0.972
10284901	2.987	0.459	−7.684	10.779	5.039	17.039	0.966
8087649	2.528	0.555	−7.711	10.722	5.019	16.985	0.961
10350225	8.473		−7.694	10.737	4.968	16.916	0.959
8963394	3.674	0.827	−7.673	10.826	4.928	16.953	0.958
10454962	6.719	1.467	−7.740	10.872	4.968	17.029	0.957

Table 4: Sample of unclassified anomalies similar to eruptive RGB stars

KIC	G mag	$G_{\rm BP} - G_{\rm rp}$	t-SNE f1	t-SNE f2	UMAP f1	UMAP f2	URF score
5014753	5.441		−4.154	−16.684	−0.903	10.464	0.986
8195444	3.473	1.687	−4.507	−14.705	−0.102	11.566	0.980
8649496	0.358	2.009	−1.839	−17.647	−0.982	9.116	0.975
8450468	3.916	0.783	−2.562	−17.238	−1.230	9.658	0.975
7200934	−0.375	1.600	−2.407	−17.495	−0.930	9.427	0.973
10416390			−4.119	−15.529	−0.191	10.818	0.972
8836489	−0.861	1.801	−3.178	−17.401	−0.974	9.807	0.970
5219663	−1.139	1.770	−2.830	−17.494	−1.010	9.605	0.968
1865744	6.551		−3.613	−17.096	−0.925	10.136	0.968
7984243	0.079	1.796	−3.227	−17.062	−0.769	9.880	0.966
8145759	−0.065	1.474	−1.822	−17.744	−1.054	9.124	0.965
4587051	−0.069	1.703	−1.940	−17.529	−0.851	9.155	0.964
10991892	−0.085	1.530	−2.453	−17.553	−0.965	9.401	0.963
6849861	−0.866	1.669	−4.125	−16.036	−0.507	10.685	0.963
9301126	0.484	1.965	−3.414	−17.166	−0.948	10.009	0.962
6025983	−0.401	1.588	−3.542	−16.890	−0.757	10.127	0.961
10447681	0.008	1.569	−3.086	−17.456	−1.024	9.715	0.961
7336419	−0.140	1.504	−3.196	−16.873	−0.430	9.743	0.961
7779434	0.594	1.978	−3.456	−17.093	−0.836	10.057	0.960
6020718	0.078	1.487	−2.775	−17.514	−0.952	9.567	0.960

Table 5: Sample of unclassified anomalies similar to Long Period Variable stars

KIC	G mag	$G_{\rm BP} - G_{\rm rp}$	t-SNE f1	t-SNE f2	UMAP f1	UMAP f2	URF score
11082175	4.482		−9.066	0.093	4.844	14.527	0.912
11554998	1.394	1.415	−9.076	0.162	4.859	14.580	0.908
10483262	−0.907	3.141	−8.948	−0.193	4.790	14.154	0.905
4677837	0.025	1.568	−9.091	0.153	4.678	14.576	0.899
10157826	−1.709	2.390	−9.011	−0.030	4.806	14.335	0.898
7909956	1.326	1.520	−9.025	0.003	4.690	14.395	0.896
8376357	−1.945	3.278	−9.051	0.230	4.946	14.603	0.893
11122913	−5.435	1.729	−9.005	−0.046	4.869	14.345	0.892
12072767	−1.326	3.220	−9.068	0.255	4.793	14.666	0.892
5733729	−2.055	2.574	−9.075	0.322	5.001	14.735	0.892
11033884	−2.470	2.379	−8.947	−0.166	4.862	14.168	0.891
6020264	−1.178	4.010	−9.033	0.004	4.691	14.427	0.891
7739645	−2.122	2.060	−9.042	0.076	4.876	14.503	0.889
100004284			−9.133	0.332	4.619	14.780	0.888
5640488	−1.360	3.255	−9.111	0.321	4.775	14.762	0.888
9820825	−2.267	2.667	−9.112	0.290	4.684	14.687	0.887
5219922	−2.209	2.738	−9.084	0.154	4.684	14.578	0.887
4551712	−1.890	2.276	−8.941	−0.195	4.920	14.170	0.884
4919121	−1.423	3.215	−9.045	0.027	4.774	14.416	0.880
6605787	−1.644	3.006	−9.055	0.148	4.923	14.553	0.880

B Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] In the abstract and introduction, we claim our contribution is that we devise a new pipeline that (1) finds anomalous light curves, and (2) can find analogs to specific light curves, in a large unlabeled corpora of light curves. We also claim that we find 5000 previously unknown anomalies, and that we present a subset of these along with potential astrophysical class.
   In Section 3 we explain the pipeline and in Section 4 we demonstrate that the pipeline recovers known rare objects, by looking at the anomalous peak of anomaly scores we produce. We also demonstrate that there are around 5000 objects that are not yet classified, that also appear in the anomalous peak. We present the a subset of these 5000 in tables 2, 3, 4, and 5. In Section 4, we show that given a light curve of interest, we can find similarly behaved light curves by cross-referencing with known astrophysical classifications from the SIMBAD database. We provide potential astrophysical classifications (in the table headers) for the subset of unknown anomalies we find.
   (b) Did you describe the limitations of your work? [Yes] See Section 5, where we summarize the limitations described more in detail in Section 4, stating that the URF method by itself is unable to differentiate between specific classes of anomalies, whereas the t-SNE/UMAP method by itself is unable to recover known rare objects.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] While we are not aware of any negative societal impacts of our work, in Section 5 we briefly talk about potential negative societal impacts that can come from applying our method on different types of data i.e social network data
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See the end of Section 1, which is a link to a Github repository that has documented code to reproduce all results and figures in the paper. In A.1, we include a URL to the full dataset we used.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All hyperparameters settings and implementation details for the entire pipeline are specified in Section A.1
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] The portion of our pipeline that is randomized is the URF method, and in Section 3 we explain how we run the portion of the pipeline producing URF scores 10 times, averaging the scores across the seeds and reporting the uncertainty in URF scores we got for anomalies
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We provide a complete description of the resources used in Section A.2. We also include the total compute time and storage requirements for the different steps in our method.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We note all existing libraries/implementations we use and cite them in Section A.1
   (b) Did you mention the license of the assets? [Yes] For all previous libaries/implementation we use, we note their license in Section A.1. We note the license of the full dataset we provide in Section A.1. The license for our code is in the specified in the linked Github repository.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We provide a link to new code at the end of the introduction in Section 1. We provide a link to the full dataset we used in Section A.1.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]