Effectively using unsupervised machine learning in next generation astronomical surveys

Itamar Reis[1]⋆, Michael Rotman[2], Dovi Poznanski[1], J. Xavier Prochaska[3,4], and Lior Wolf[2,5]

[1] *School of Physics and Astronomy, Tel-Aviv University, Tel-Aviv, 69978, Israel*
[2] *School of Computer Science, Tel-Aviv University, Tel-Aviv, 69978, Israel*
[3] *UCO/Lick Observatory, University of California, 1156 High Street, Santa Cruz, CA 95064, USA*
[4] *Kavli Institute for the Physics and Mathematics of the Universe (WPI), University of Tokyo, Kashiwa 277-8583, Japan*
[5] *Facebook AI Research*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

In recent years many works have shown that unsupervised Machine Learning (ML) can help detect unusual objects and uncover trends in large astronomical datasets, but a few challenges remain. We show here, for example, that different methods, or even small variations of the same method, can produce significantly different outcomes. While intuitively somewhat surprising, this can naturally occur when applying unsupervised ML to highly dimensional data, where there can be many reasonable yet different answers to the same question. In such a case the outcome of any single unsupervised ML method should be considered a sample from a conceivably wide range of possibilities. We therefore suggest an approach that eschews finding an optimal outcome, instead facilitating the production and examination of many valid ones. This can be achieved by incorporating unsupervised ML into data visualisation portals. We present here such a portal that we are developing, applied to the sample of SDSS spectra of galaxies. The main feature of the portal is interactive 2D maps of the data. Different maps are constructed by applying dimensionality reduction to different subspaces of the data, so that each map contains different information that in turn gives a different perspective on the data. The interactive maps are intuitive to use, and we demonstrate how peculiar objects and trends can be detected by means of a few button clicks. We believe that including tools in this spirit in next generation astronomical surveys will be important for making unexpected discoveries, either by professional astronomers or by citizen scientists, and will generally enable the benefits of visual inspection even when dealing with very complex and extensive datasets. Our portal is available online at `galaxyportal.space`.

Key words: keyword1 – keyword2 – keyword3

1 INTRODUCTION

Every so often in the history of astronomy, visual inspection of data has led to an unexpected scientific discovery. As the volume of data that astronomical surveys gather increases, the presence of increasingly rare phenomena within these datasets is essentially unavoidable. It is worth noting that in astrophysics, what are observationally rare phenomena can actually be quite common and important but short lived (e.g., supernovae that only occur a handful of times per century in a given galaxy but play a crucial role in its chemical evolution). However, the likelihood of detecting such rare phenomena is decreasing, since much of the data we now gather cannot realistically be visually inspected.

The Sloan Digital Sky Survey (SDSS; Eisenstein et al. 2011) has obtained spectra of ∼ 3M galaxies and quasars. Since these were accumulated over nearly two decades, and intensively studied by the community, a non negligible fraction of the (high signal to noise ratio; SNR) data was in fact visually inspected by different people. This will surely not be the case for next generation surveys, that will generate ∼ 10 times more data in ∼ 1/10 of the time. For example, the Dark Energy Spectroscopic Instrument (DESI, Levi et al. 2013) which has begun commissioning, will observe ∼ 30M galaxies, and SDSS-V (Kollmeier et al. 2017) plan to obtain ∼6M stellar spectra and ∼25M spectra of the Milky Way interstellar medium (ISM). While the data rates increase a hundredfold, the number of experts available to inspect them remains more or less constant.

Since we will not be able to examine every spectrum these surveys will generate, we propose to enlist the help of unsupervised Machine Learning (ML) in order to prioritize. Unsupervised ML is the name of a broad family of tools that could be used to detect rare

⋆ E-mail: itamarreis@mail.tau.ac.il

objects or trends. These tools include: (i) Clustering algorithms, which are used to detect groups of objects sharing similar features, (ii) anomaly detection algorithms, which are used to detect unusual objects, and (iii) dimensionality reduction algorithms. The latter family of tools facilitates clustering, trend finding, and anomaly-detection analyses, by representing a dataset in a low dimensional space that can be better visually (or otherwise) inspected. Since these algorithms are data rather than model driven, they have the potential to detect patterns in the data that we did not know existed, and therefore would not have searched for directly.

A wide variety of approaches to perform each of these tasks has been developed by the ML community. `Python` implementations of many of the most common approaches are publicly available in `scikit-learn`. These are relatively easy to use, with conveniently homogenized user interfaces. These tools are flexible and could be tuned to produce useful results from most datasets. Many examples of application of these methods to astronomical data are available in `AstroML` (VanderPlas et al. 2012).

In this work we make use of both anomaly detection and dimensionality reduction, in the following subsections we briefly review both, focusing on recent usage in astronomy.

1.1 Anomaly detection

Anomaly detection algorithms typically rank objects based on a definition of abnormality, where many such definitions exist. Three common general approaches are: (i) objects that are the least similar to other objects in the data, (ii) objects that reside in low density regions of the data, and (iii) objects that are not well reconstructed by a model of the data. However, there are many ways to measure similarity, a variety of methods to define and measure the density, and obviously, endless approaches to modeling data, lending to the richness of existing anomaly detection algorithms. As could be seen in the examples below, all these approaches were successfully applied to astronomical datasets.

Applications of anomaly detection to spectroscopic data include Boroson & Lauer (2010) who detected anomalies in SDSS quasar spectra using the reconstruction error of a Principal Component Analysis (PCA) model of the data, i.e., the residual between the data and best fitting model. A more recent use of reconstruction-based anomaly detection is Ichinohe & Yamada (2019), where the model of the data (in this case X-ray spectra), was built using a variational auto encoder. Meusinger et al. (2012) used self organizing maps (SOM, Kohonen 1982) for anomaly detection in SDSS quasar spectra. Their unusual quasars were defined to be objects residing in low density regions of the 2D embedding of the data created via SOM. Distance based anomaly detection, was applied to SDSS galaxy and APOGEE stellar spectra by Baron & Poznanski (2017) and Reis et al. (2018b) respectively, using an unsupervised Random Forest distance (Shi & Horvath 2006).

For light curve data, distance based anomaly detection was used by Protopapas et al. (2006) and Richards et al. (2012) with different definitions of similarity. Protopapas et al. (2006) worked with raw light-curve data and used the cross correlation distance. Richards et al. (2012) worked with extracted features and used Random Forest (in this case it was supervised, trained on labeled data). In an example that does not directly fit into any of the general approaches described above, Nun et al. (2014) used a supervised Random Forest to predict the class of unlabeled objects, and anomalies were detected as objects having unusual voting distributions. Nun et al. (2016) detected anomalies using an ensemble of anomaly detection methods.

An example of anomaly detection on imaging data is Shamir & Wallin (2014) who found peculiar SDSS galaxy pairs. As we discuss below, interpreting and understanding why an object was selected as anomalous by a given algorithm is often not trivial when dealing with spectroscopy or light curves. However, when inspecting the images of the objects detected in Shamir & Wallin (2014) their unusual features are manifest even to non experts in galaxy morphology.

1.2 Dimensionality reduction

Dimensionality reduction algorithms can be divided into types according to the quantity they try to preserve when representing the data in a lower dimensional space. Some algorithms, such as Multi Dimensional Scaling (`MDS`) try to preserve the similarity between all the objects. Other algorithms only try to only preserve the neighborhood, or nearest neighbors, but not the actual similarity values, e.g., `t-SNE` (van der Maaten & Hinton 2008). Auto-Encoders try to preserve information, in the sense that the data itself could be reproduced from the low dimensional representation, typically referred to as the latent space. In this work we use the Uniform Manifold Approximation and Projection (`UMAP`) algorithm (McInnes et al. 2018), which tries to preserve the topology of the manifolds on which the objects lie between the low dimensional representation and the original data. See Gisbrecht & Hammer (2015) for a review of dimensionality reduction algorithms.

In recent years a number of works applied dimensionality reduction techniques to astronomical datasets and showed that the resulting embeddings contain useful information (in der Au et al. 2012; Meusinger et al. 2012; Jofré et al. 2015; Traven et al. 2017; Anders et al. 2018; Reis et al. 2018b). For example, Reis et al. (2018b) created an embedding of the APOGEE (Majewski et al. 2016) infrared stellar spectra dataset using `t-SNE`. They showed that the location of a star on such a map contained information about its effective temperature, surface gravity and metallicity. In addition, groups of peculiar stars such as Be and carbon stars were clustered in specific locations on the map.

While it is clear that such maps can contain non-trivial information, it is not obvious how we can extract this information and potentially learn something new about the data. We suggest, as also done by in der Au et al. (2012), that one way to do this is by using the maps in an interactive way. Selecting and inspecting objects directly from the map enables studying the sample in detail. We will use such interactive maps of the data in the portal we present in Section 3.

1.3 Outline

In Section 2 we perform a comparison between various anomaly detection methods on a dataset of galaxy spectra from the SDSS. From the method comparison we draw conclusions that are general to any unsupervised ML application, and we discuss a number of challenges to the effective incorporation of these methods into our workflow with future surveys. In Section 3 we present the data portal that we are developing. We use unsupervised ML to construct as many useful but different human-inspectable summaries of the data as possible, and gather them in an intuitive and easy to use interactive portal. We showcase this approach with SDSS galaxy spectra. We summarize in Section 4.

2 THE CHALLENGES WITH ANOMALY DETECTION

2.1 Anomaly detection method comparison

The fact that many unrelated methods were successfully applied to astronomical data raises the question of how to choose which algorithm to use for a given project, or specifically, with next generation spectroscopic surveys. Is there, for a specific dataset, a single algorithm that is optimal? To try to answer this question we use a sample of 150,000 SDSS galaxy spectra, and apply four different anomaly detection algorithms; PCA reconstruction error, unsupervised Random Forest, Fisher Vector based anomaly detection, and Isolation Forest.

For PCA reconstruction error and Isolation Forest (Liu et al. 2008) we use the `scikit-learn` implementation. For unsupervised Random Forest we use the `scikit-learn` implementation of Random Forest, and our own code (available at `github.com/ireis/unsupervised-random-forest`) for calculating the anomaly score. We add to the comparison the results of the same algorithm from Baron & Poznanski (2017). For Fisher Vector based anomaly detection (Rotman. et al. 2019) we use our own implementation. The anomaly detection methods are described in more detail in Appendix B.

The galaxy spectra were obtained from the 14th data release of the SDSS SDSS DR14 (Abolfathi et al. 2017). We selected objects with `Class = GALAXY` from the `SpecObj` table and used only galaxies for which the rest frame spectrum contained flux values in the wavelength range of 3700Å $< \lambda <$ 8000Å. Out of these galaxies we selected the 150,000 with the highest SNR (according to the `SNMedian` field in the `SpecObj` table). The preprocessing stage consisted of removing flux values marked as bad by the SDSS pipeline (i.e., flux values with inverse variance of 0), normalizing the spectra by the median, shifting the spectra to the rest frame according to the SDSS pipeline redshift, and interpolating the spectra to a fixed wavelength grid. We note that, as shown for example in Baron & Poznanski (2017), objects with incorrect redshifts can be found as outliers in this scheme. We use the normalized flux values as features for all the algorithms.

By definition, unsupervised tasks are challenging to optimize. What constitutes a successful application of anomaly detection? In science we typically aim to detect a large variety of anomalies, and the sole detection of objects of a single kind is considered non-satisfactory. In all cases significant tuning of the hyper-parameters or implementations (the difference between implementation decisions and hyper-parameters is an implementation decision) was required to obtain satisfactory results.

As an example of hyper-parameter tuning, with Isolation Forest we used rank values instead of the normalized flux values (that is, we strip the marginal distribution of each feature), to get satisfactory results. Without this modification we obtained only objects with extreme emission line strengths. Due to a small difference between the `scikit-learn` implementations of Random Forest and Isolation Forest, Random Forest is not sensitive to the marginal distributions of the features while Isolation Forest is. This is because in `scikit-learn`, at each node of a Random Forest tree, the best split search grid is constructed from the feature values (of the objects in the node) themselves, while for an Isolation Forest tree the grid is a linearly spaced set of values, between the minimum and maximum feature values. This example illustrates how seemingly insignificant implementation decisions can completely change the output of anomaly detection algorithms.

Additional examples for implementation decisions and hyper-parameters that we identified as having a major effect on the results include: (i) The properties of the synthetic data in Unsupervised Random Forest. This algorithm involves comparing the data to synthetic data that needs to be created by the user. There is infinite freedom in constructing synthetic data, and naturally, this affects the results. We were able to obtain useful results with a number of different synthetic data types. See also Shi & Horvath (2006) who compared two relatively similar methods of creating synthetic data and obtained very different results; (ii) The number of objects used to train a single tree in Isolation Forest and Random Forest. This hyper-parameter has a major effect on the behavior of these algorithms, and yet it is not built into `scikit-learn`, and requires wrapping around their implementation; (iii) The number of components in the Gaussian mixture model for the Fisher Vector method; (iv) The details of the reconstruction error calculation for PCA reconstruction error. We were only able to obtain satisfactory results when calculating the reconstruction error separately on relatively small regions of the spectra, and inspecting objects with the largest errors in different regions.

2.2 Consistency and significance of the discovered anomalies

After tuning the different methods we were able to obtain satisfactory results with all four of them. Considering all the methods uncovered a satisfactory list of anomalies, it is not trivial to pick the best one. Worst, we found that there is little overlap between the anomalies detected by the different methods and the mean overlap between the top 500 objects from two different algorithms is $\sim$ 20. The overlap between the anomalies detected with variations of the same method is also small, and in fact not different from the overlap between anomalies from completely different methods. Visually inspecting the detected anomalies showed that every method missed some interesting objects found by others. This suggests that a single anomaly detection algorithm does not produce in practice the most unusual objects in the data, rather the output should more appropriately be considered as a small subset of the unusual objects.

The overlap between the anomalies detected by different methods can teach us about the statistical significance of the anomalies. This is illustrated in Fig. 1. The top panels show Venn diagrams of two randomly drawn subgroups of size $g = 500$, which were drawn from a parent population of size $n = 1,000$ (right) or size $n = 10,000$ (left). The bottom panel show Venn diagram of $g = 500$ anomalies detected by the two implementations of Random Forest, chosen here for example. The extent of the overlap between the two subgroups suggests that there are $n \sim 10,000$ anomalies in the data.

This can be done quantitatively in the following way. Let us assume that there are n anomalies in the data. Let g be the number of anomalies detected by each method, and $k_{i,j}$ the overlap between the anomalies detected by methods i and j. Further assuming that these sub-samples were uniformly drawn, we can estimate n given g and $k_{i,j}$, for each i, j. To calculate the expected k given n and g consider having a group of objects of size n, and a subgroup of size g. We are now randomly choosing, from the original group, another subgroup of size g. If $n \gg g$ the probability of a single randomly chosen object to be in the first subgroup is g/n. Choosing g such objects, the expected overlap is g^2/n. In this case we can easily estimate the number of anomalies with

$$n \sim \frac{g^2}{k}. \quad (1)$$

See Appendix A for the calculation without assuming $n \gg g$. Given the formula, an intersect of $k = 20$ between $g = 500$ anomalies

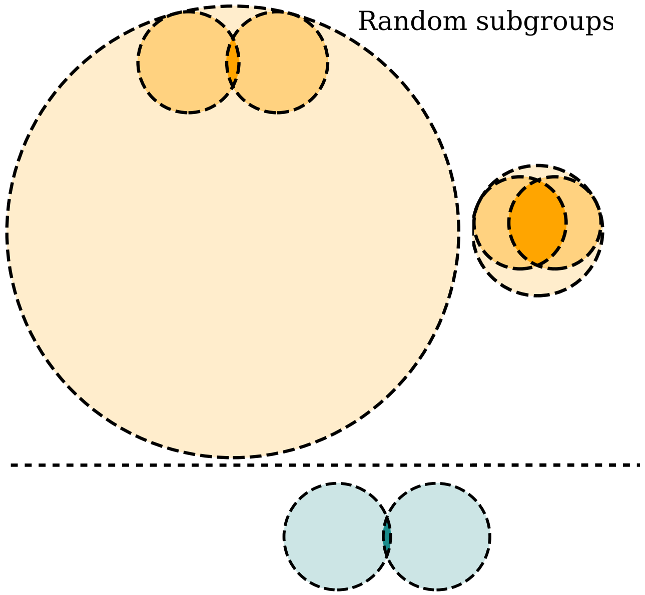


Figure 1. An illustration of how the total number of anomalies in the data could be estimated from the overlap between the anomalies detected by different methods. The top panels show Venn diagrams of a group of a given size ($n = 10,000$ on the left panel, and $n = 1,000$ on the right panel) and 2 randomly drawn subgroups of size $g = 500$. One can see the dependence of the overlap between the subgroups on the size of the original group. The bottom panel show a Venn diagrams of 2 sets of 500 anomalies, detected by the two implementations of Random Forest. The overlap suggests we are in a situation similar to the one in the top left panel, where there is a large number of anomalies in the data, ~ 10,000 in our case, of which we only detect small fractions. A quantitative estimation of the number of anomalies in the data is shown in Fig. 2.

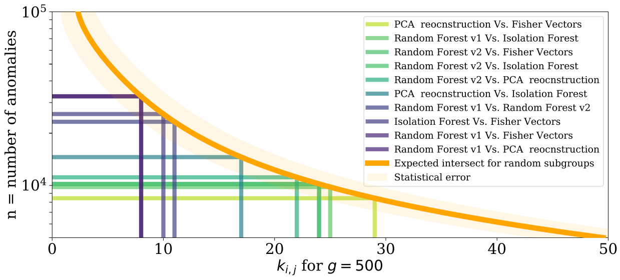


Figure 2. The overlap ($k_{i,j}$, x-axis) between the top $g = 500$ detected anomalies of different anomaly detection methods, and the resulting estimated total number of anomalies in the data (n, y-axis). The orange line is the expected size of a parent population given an overlap between two randomly drawn subgroups of size 500. If one assumes simplistically that there is a well defined group of anomalies from which different detection method pick random sub-samples, one can measure the size of the underlying group of anomalies. Using it only as a ballpark estimate, it seems that the underlying group is of order 10^4, which is a few percent of the data. Random Forest v1 and v2 are two implementations of the same algorithm. The statistical uncertainty in the expected intersect is shown in light orange, while the Poisson uncertainty is omitted.

detected by different methods suggests $n \sim 12,500$ anomalies in the data.

In Figure 2 we show the results of this calculation for each pair of anomaly detection methods we applied. $k_{i,j}$ is shown on the x-axis, and the resulting n is shown on the y-axis. The orange line represents $n(k)$ for random groups and the shaded orange region is its standard deviation, calculated numerically, by randomly drawing many groups and calculating the standard deviation of their intersects for different values of n. The largest overlap is between the PCA reconstruction and Fisher Vector methods (yellow-green line), for which $k_{i,j} = 29$ and $n = 8,525$. The lowest overlap is between one of the Random Forest implementations and PCA reconstruction (dark blue line), for this case $k_{i,j} = 8$ and $n = 31,716$. Note that the agreement between the two Random Forest implementation is lower than that of completely different methods. While not shown in the figure, this is also true for different implementations of any of the other methods we applied.

Taking this result at face value means that our dataset contains $O(10^4)$ anomalies that would be picked up by either of these algorithms. As a consequence, by choosing only one algorithm, even if optimized, and visually inspecting a few hundred candidates, a common and manageable number, we will only assess $O(5\%)$ of the true anomalies. This fraction is expected to further decrease as the number of objects worth noting will increase with the size of the data, but the number of objects inspected by a single person is not. Next generation spectroscopic surveys will contain one or two orders of magnitude more objects than what we have considered here, thus requiring a different approach.

For the sake of the discussion above we have made the assumption that each anomaly detection method detects a random group of anomalies. If this assumption were true, then the distribution of the anomalies detected by any single method would have been representative of the overall distribution of anomalies in the data. In such a case each single method would have the potential to detect all types of anomalies. This would be surprising, and indeed does not hold in practice, as can be seen for example in Figure 3, where we show the SNR distribution of the top 100 anomalies from three methods. The fact that the three distributions are so different from each other makes it clear that the different methods are biased towards different types of objects. Note that if two methods are biased towards the same types of objects their intersect will be larger than that of randomly drawn groups and decrease our estimate of the number of anomalies in the data, and vice versa. The biased nature of the detected anomalies should thus increase the scatter in Figure 2 but not necessarily change the mean value. To get an unbiased sample, that has a better chance of detecting all types of anomalies, it is necessary to use an ensemble of different methods. This however does not obviate the need to examine many more objects than typically done.

Also, we used 500 objects per algorithm, but our results above are largely insensitive to that choice, in that as long as we use a few hundred objects per algorithm, we consistently get $n = O(10^4)$. With more than a few hundreds, n starts to increase, likely due to false positives. Using less than ~ 200 objects per algorithm creates zero overlaps between the subsamples (and brings the difficulties of small number statistics).

2.3 Interpretability of the anomalies

Interpreting anomalies can be an even greater challenge than detecting them. All anomaly detection methods discussed above only produce a list of unusual objects and do not provide any indication as to what is unusual about a given object. This critical task is very

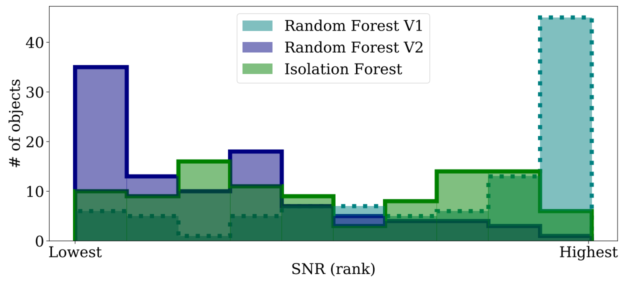


Figure 3. The Signal to Noise Ratio (SNR) distribution of the top 100 anomalies from 2 implementations of unsupervised Random Forest, and an implementation of Isolation Forest, on the SDSS galaxy spectra dataset. While one implementation of unsupervised Random Forest favors high SNR objects, the other favors low SNR object. This Isolation Forest implementation does not seem to show any SNR preference. This demonstrates how different detection methods or specific implementations are biased differently and produce inherently different outcomes. The x-axis is the SNR rank and not the SNR value, otherwise the SNR distribution dominates the graph.

time consuming, and prone to errors as there is no guarantee that our interpretation will be correct, i.e., that we indeed found the reason that an object was tagged as anomalous.

Interpretation becomes even more challenging when a single object is too complex to be easily inspected in a glance. This would be the case for spatially resolved spectroscopy, spectroscopic time series, or high resolution spectra, to give examples from the world of spectroscopy. As the complexity of the data increases, the question of what is unusual about a given object becomes more challenging but also more interesting. Furthermore, with the complexity, more objects will show some unusual features, or as often coined in the ML literature: when the complexity of the data is high enough every object is an anomaly. All the anomaly detection methods we have discussed are inherently designed for low dimensional data in which the anomalies are obvious and only need to be quickly and automatically detected. Progress will therefore come from a more streamlined interpretation, rather than more clever detection algorithms.

2.4 Unsupervised ML for high dimensional data

The challenges we discussed are both stemming from the complexity (or high dimensionality) of datasets and not their size. These challenges are not special to astronomy, and in fact anomaly detection in high dimensional data is an active ML field of research (Aggarwal & Yu 2001; Zhang & Zhao 2004; Müller et al. 2008; Kriegel et al. 2009; Müller et al. 2010; Keller et al. 2012, listing a few examples). See also Zimek et al. (2012) for a ML oriented review of this topic.

Before discussing possible approaches for handling high dimensional data we would like to emphasize that while we focus on anomaly detection, other unsupervised ML applications, such as dimensionality reduction and clustering, suffer from the same issues. For example, in high dimensional data there could be many different but informative ways to divide objects into clusters, and different clustering algorithms can produce different but valid outcomes. This is illustrated in Figure 4 for the case of data with only three independent features. We see that these data can be divided into clusters in a number of different ways, without any single way being the best one.

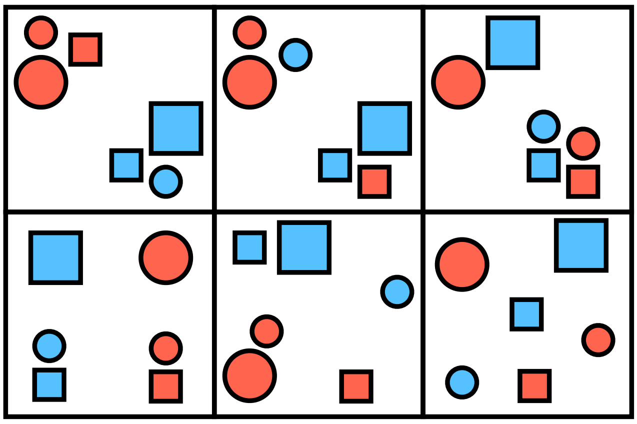


Figure 4. Six different ways of clustering the same data, none of which can be qualified as optimal. As the complexity of data increases, and each object is represented with more features, there could be many ways to divide the data into clusters (or similarly to define anomalies). While different from each other, all are conceivable. This is an illustration of the fact that in unsupervised ML there could be many possible answers to the questions we are asking. We should not look for the best answer, but instead find and inspect as many valid ones as possible.

The reason that anomaly detection, clustering, and dimensionality reduction all suffer from the same issues, could be reduced to the definition of a pair-wise distance between objects in the data, which is the basis of many unsupervised ML methods. In complex data the relationship between two objects cannot be described by a single distance. If two objects are similar in one feature and different in another, there is no good way to include the information in a single number. As exemplified in Figure 4, is a red square more similar (i.e., closer) to a red circle or a blue square? Different definition of distance will give more weight to some features and less to others, and thus produce different outcomes for any unsupervised ML application based on this distance. Each different outcome has the potential of providing useful information.

A common approach for handling high dimensional data is working with subspaces, each subspace containing all the objects, but only a subset of the features. The advantage of this approach is that with small enough subspaces the issues arising due to the high dimensionality of the data disappear. In a small subspace the definition of distance becomes unique so there will be a single way to cluster the data, and define anomalies. The results we obtain are also easier to interpret, as we know why an object is an anomaly, or what is the common feature of cluster members, according to the subspace in which the anomaly or cluster were detected.

By working with subspaces one can resolve the two issues discussed above. One can scan many solutions, each from a given algorithm in a given subspace with its own merit, and with an easier path to interpretation, since an outlier in a limited subspace would be easier to understand, as we indeed show below. However, the unavoidable cost of having many algorithms, run within many interesting subspaces, is that there are many results to vet, more than can be realistically done by a single human.

3 FRAMEWORK FOR THE EXPLORATION OF UNSUPERVISED ML RESULTS

One way to address the difficulty of having many different but useful outcomes, is collecting them in an easy to use tool so that they could be inspected by the community. In order to allow large numbers of people from the community to explore current and future datasets through the unsupervised ML lens, we are building an interactive graphical portal to large and complex datasets.

Data exploration portals are common in astronomy, a few examples are: `SIMBAD` (Wenger et al. 2000), `Galaxy Zoo` (Lintott et al. 2008), The Open Supernova Catalog (Guillochon et al. 2017), `ESASky` (Baines et al. 2017), `Marvin` (Cherinka et al. 2018), and `SkyPortal` (van der Walt et al. 2019). An upcoming portal for anomaly detection in astronomical data is `Astronomaly` (Lochner et al. 2019). These portals are generally designed to inspect single objects in a convenient way, and thus cannot be used to explore large numbers of sources.

The key novel feature of our portal is machine-learned two-dimensional embeddings (or maps) of the data, that automatically group sources which are similar to each other, and from which objects can be interactively selected and inspected. The interactive maps should be used with the simple notion that similar objects are located close to each other. This makes detecting potentially interesting phenomena quite intuitive. Objects that are isolated on a map are interpreted as objects that are not similar to any other object in the dataset. Any structure in the ordering of objects suggests a continuous change in the properties of objects along the structure. A compact group of objects implies the objects share some common properties. While the maps are built using ML their usage is intuitive and does not require any prior knowledge in this field. Furthermore, by working in subspaces of the data, we can produce a number of such maps, each containing different information. As discussed above, when working with such subspaces, the interpretation of any unsupervised ML method, and specifically the 2D maps, is also made easier.

To illustrate the possible use cases of such a portal we apply it to the SDSS galaxy spectra dataset (same dataset as described in Section 2.1). In the next subsection we discuss the details of our portal as currently implemented for this dataset. A screenshot of the portal is presented in Figure 5. The portal itself is available online at `galaxyportal.space`, a user manual is available at `toast-docs.readthedocs.io/en/latest/`.

3.1 Implementation details

Our implementation for the SDSS galaxies is intended to showcase the general approach described above, and the details can change in other implementations or in future versions of this portal. The main challenge in constructing the portal was creating interactive linked graphs containing large amounts of data. Since showing hundreds of thousands of points on a graph is both prohibitive and pointless, we implemented an adaptive graph that shows a random subset (with a fixed upper limit size) of the objects in the current frame, where more objects in a specific region could be seen by zooming in. Our code is based on the `Bokeh` library[1].

To create the maps of the data we first divide the spectra to a number of wavelength regions which are manually defined according to the locations of common emission and absorption lines.

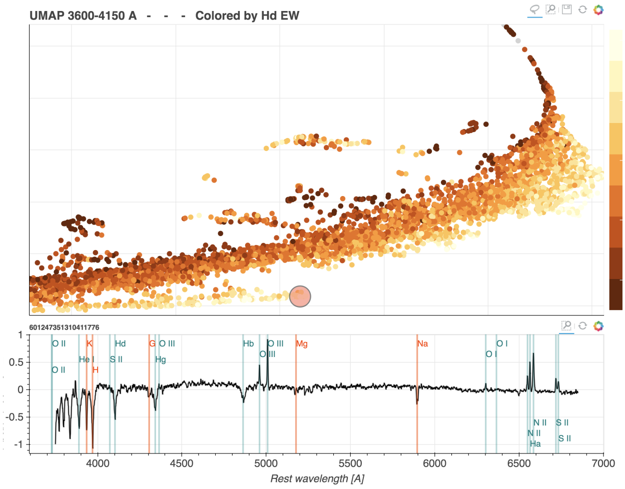


Figure 5. A screenshot from our data portal (`galaxyportal.space`), where one can study the SDSS galaxy spectra through an unsupervised ML lens. In the top panel, each point represents a galaxy, and galaxies with similar spectral properties are located close to each other in this abstract plane. In this specific embedding the similarities are based on the wavelength region 3600 – 4150[Å]. The map is further colored by the equivalent width of the H_δ absorption line. This map and coloring are just examples of the many that we made available. The user of the portal can easily toggle between various embeddings and coloring schemes. In the bottom panel we see the spectrum of the galaxy that has been selected interactively in the top panel, where it is marked with a red circle.

These wavelength regions are examples of subspaces of the data. In each region the spectrum is normalized by the median flux value in the region. Next, Euclidean distances are calculated between the objects (i.e., a sum of the squared differences of normalized fluxes). Finally, we apply the `UMAP` algorithm to the distances to obtain the maps.

To illustrate the advantages of using maps created in subspaces of the data, let us consider a map constructed from the region of the spectrum containing the NaI D doublet. An isolated group of galaxies on this map will most likely contain galaxies with unique NaI D line profiles. On the other hand, on a map constructed from the entire spectrum, (i) there is no guarantee the same group of galaxies with unique NaI D profile will appear as an isolated group, since on such a map the galaxies could be grouped according to some other more prominent feature, and (ii) given an isolated group on the map it is hard to determine what is the unique feature shared by the objects in the group. Figure 6 shows an example of the first point, where objects clustered on a map created using the NaI D region are no longer clustered on a map constructed from a different wavelength region.

In addition to the machine learned maps, we uploaded to the portal common galaxy diagnostics such as the BPT diagrams (Baldwin et al. 1981) from which objects can be interactively selected as well. All the embeddings in the portal are linked which allow the user to select objects on one map and display them on another. With this, it is easy to check, for example, where galaxies with unusual NaI D profiles lie on the BPT diagram. This example is shown in the bottom left panel of Figure 6.

With the map displaying 2D information, it is easy to add a third one via color. The maps can be colored by various properties of the galaxies such as line ratios, the star formation rate, and the velocity

[1] https://bokeh.pydata.org/en/latest/

dispersion. Most properties are taken from the SDSS value added catalogs, the rest were calculated by us. Coloring the maps is useful when learning the general location of different types of galaxies on the map. For example see Figure 5 where the `UMAP` is colored by the H_δ equivalent width (EW). In this example the coloring can guide the user to the locations of post-starburst galaxies. The bottom panel of Figure 5 shows one such post-starburst galaxy, located at the end of a one dimensional structure with increasing H_δ EW.

We included in the portal the results of all the anomaly detection algorithms that we used for the method comparison in section 2.1. The location of the anomalies on the different maps is useful for investigating the reason a given object was detected as an anomaly. It is also possible to order the anomalies according to either their location on the map, or any of the available galaxy properties, instead of viewing them ordered by their abnormality score, which is effectively quite random. This allows for a more efficient visual inspection process, as similar objects can be inspected and classified together.

Since anomaly detection is only one of the goals of our portal, we also include other features, that can be used when searching for trends. For example, the user can bin and stack spectra of objects on the fly, by selecting them on the maps, and binning them according to their coordinates or any of the available galaxy properties. A few example use cases for these features are presented below, more extensive instructions on how to use all the features are available in the online documentation at `toast-docs.readthedocs.io/en/latest/`.

3.2 Use case I: Anomaly detection

As a demonstration of the capabilities of the portal for anomaly detection, we focus here on a single map, constructed from the $\lambda = 5680 - 6120$[Å] region. As we show below, this restricted view alone uncovers multiple interesting phenomena. We detect galaxies as anomalies by manually inspecting the map. Namely we select and inspect all objects that are either isolated or located at extreme ends of the 2D distribution of objects. The map showing the detected groups is presented in Figure 6 and a number of example objects are shown in Figure C1. A full description of the findings is given below.

In this wavelength region, the strongest features are the NaI D absorption doublet (to which both cold stars and the ISM contribute), and the HeII emission line. We find the following groups of unusual objects (for the more interesting groups we list a few examples in a table, more examples can be obtained from the portal itself):

- As could be expected, galaxies with extremely strong NaI D absorption are located at an extreme end of the 2D distribution and are easily detected. They are shown in yellow in Figure 6.
- Similarly expected, galaxies with extreme HeII emission are found at another edge. They are shown in dark green.
- Galaxies with strongly blueshifted NaI D are grouped together at a few locations. We mark them in blue. The largest concentration contains a few hundred galaxies, where the objects with the largest blueshifts are also the farthest from the bulk of the galaxies. Another small group of galaxies with blueshifted NaI D contains a few galaxies in which the absorption is also stronger. The last group of these is characterized by a bluer continuum. Table C1 lists a number of examples. Blueshifted NaI D is a signature of outflowing gas. A sizable fraction of the objects we find are star-forming face-on spirals such as the ones discussed in (Heckman et al. 2000; Bae & Woo 2018). Another dominant population is composed of post-starburst galaxies. These are found using the portal by inspecting the locations of the strongly blueshifted NaI D galaxies on the `UMAP` constructed from a region of the spectrum containing the H_δ absorption, where post-starburst galaxies cluster, or by coloring the NaI D map with the H_δ EW.
- Galaxies with multiple component NaI D are located in a single cluster on the map. They are shown in light red. Most galaxies in this group seem to be well described by two velocity components, but some show evidence of three. The images of the objects in this group all show that these objects are blended, suggesting the multiple components are coming from different galaxies at similar redshifts. Some of these objects appear in catalogs of galaxy clusters. Table C2 lists a number of examples.
- We find a number galaxies with redshifted NaI D absorption. Four such objects are located in a small cluster on the map. These are shown in dark red in Figure 6. Three additional objects are found in the region of the map containing galaxies with strong NaI D absorption (shown in yellow on the map). All these objects are listed in Table C3. Inspecting the individual objects, it seems that different reasons cause the apparent redshifted NaI D line. In SDSS J211635.95-004613.1, the emission lines are redshifted by the same velocity as the NaI D line, while the *H* and *K* lines are centered on the systematic redshift. On the other hand, in SDSS J141518.01+230841.0, the *H* and *K* line are redshfited with the NaI D line, while the emission lines are blueshifted from the systematic redshift. In the second case a natural interpretation is an offset AGN (e.g, Comerford & Greene 2014).
- Another cluster on the map is composed of galaxies with an unusual slopes of their SED in this wavelength region. They are shown in purple. Many of these objects were mistakenly classified as galaxies by the SDSS pipeline but are in fact white dwarfs and quasars. The objects that are indeed galaxies, such as SDSS J092034.87+511224.1 and SDSS J091959.69+513346.7 show broad emission features that we think are due to some error in the data acquisition or reduction.
- Galaxies hosting type Ia supernova features in their spectra are also found in a small cluster. We mark them in light green. 8 of these galaxies, already found by Graur & Maoz (2013), are listed in C4. Additional candidates could be found by inspecting the rest of the objects in the cluster. The clustering of these galaxies is due to the fact that the tell-tale feature of these supernovae, the deep Si absorption, falls partly within this wavelength region. Maps constructed with different wavelength regions, in which there are no prominent supernova spectral features, do not contain such a cluster.
- Additional clusters (not shown in color) and objects that are isolated on the map show a variety of unusual behaviors. We find a small cluster of objects with chance alignment with brown dwarfs, objects with the wrong redshift determined by the SDSS pipeline, objects with bad sky lines subtraction, and some with unexplained absorption lines (likely due to a foreground or background source). Some of these galaxies are listed in Table C5.

The bottom left panel of Figure 6 shows the locations of the detected anomalies on the BPT diagram. Note that some of the anomalous galaxies do not have detected emission lines and are thus not shown on the diagram. 3 groups of unusual NaI D region galaxies also cluster on the BPT diagram: (i) Objects with strong NaI D absorption are located in the AGN region of the diagram. (ii) Many of the objects with strongly blueshifted NaI D are outliers in the BPT diagram, and have unusually low OIII/H_β line ratios given

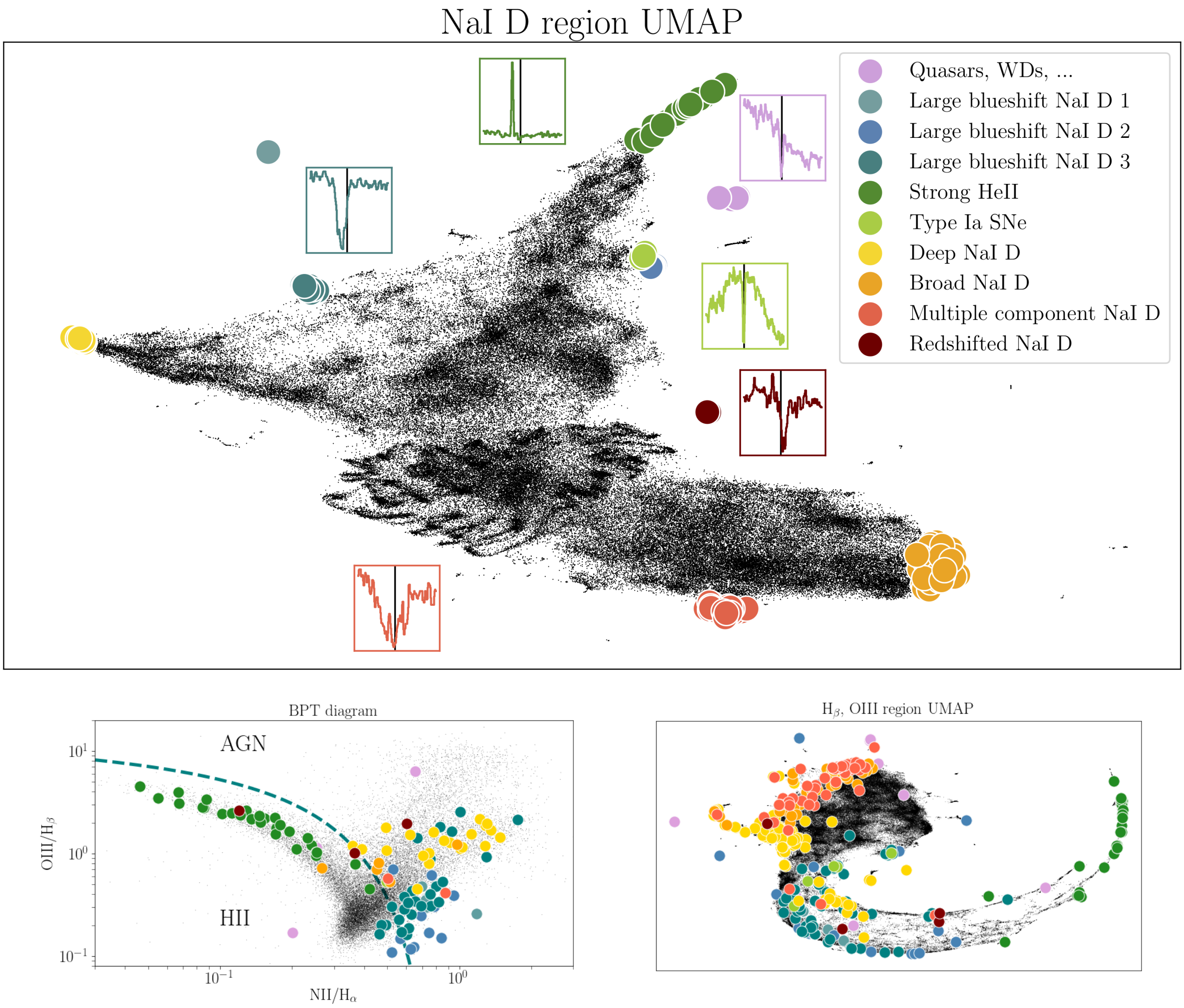


Figure 6. Top panel: A UMAP of SDSS galaxies constructed from the $\lambda = 5680-6120$[Å] region of the spectrum. Groups of unusual objects detected from the map are shown in color. Inserts containing the NaI D part of the spectrum for an example object are shown for select groups. **Bottoms panels:** The locations of the same galaxies with unusual NaI D properties as in the top panel, on different embeddings. The bottom left panel shows the BPT diagram which separates AGNs from star-forming galaxies. The bottom right panel shows a UMAP constructed from the $\lambda = 4700-5100$[Å] region of the spectrum, containing the H$_\beta$ and OIII lines. In the bottom left panel we can see that some of the unusual NaI D groups lie in specific regions of the BPT diagram. The interesting finding is the galaxies with deep NaI D absorption, who seem to be preferentially located in the AGN region of the BPT, while many galaxies with strong blueshifted NaI Dseem to be outliers on the BPT diagram. In the bottom right panel we see that the unusual NaI D objects could not have been detected using this $\lambda = 4700-5100$[Å] map, illustrating that different maps contain different information. The locations of the unusual NaI D objects here are also not random, suggesting correlations between various NaI D and H$_\beta$ and OIII properties. Interactive versions of all these maps, from which the unusual NaI D objects and others were selected and inspected, is available at galaxyportal.space.

their (high) NII/H$_\alpha$ ratios. (iii) Objects with strong HeII emission are located in the star forming branch of the diagram.

To illustrate that different maps contain different information we show the locations of the same unusual groups of objects discussed above on a UMAP constructed from a different wavelength region in the bottom right panel of Figure 6. This map shows the $\lambda = 4700-5100$[Å] region, containing the H$_\beta$ and OIII lines. The unusual NaI D groups are no longer clustered or isolated and would not be detected on this map.

3.3 Use case II: Trend detection

One method to detect trends is by inspecting how objects change along structures on the maps. In this example we use the $\lambda = 6400-6700$[Å] region map. This region contains the H$_\alpha$ and NII lines. Objects that lie along geometrical structures on the map are expected to show continuous changes in these features. The user of our portal can select a specific region on the map and a specific direction, and stack the galaxies in bins along the chosen direction. This procedure takes a few button clicks on the portal. Figure 7 shows an example. What is found in this case is that the galaxies

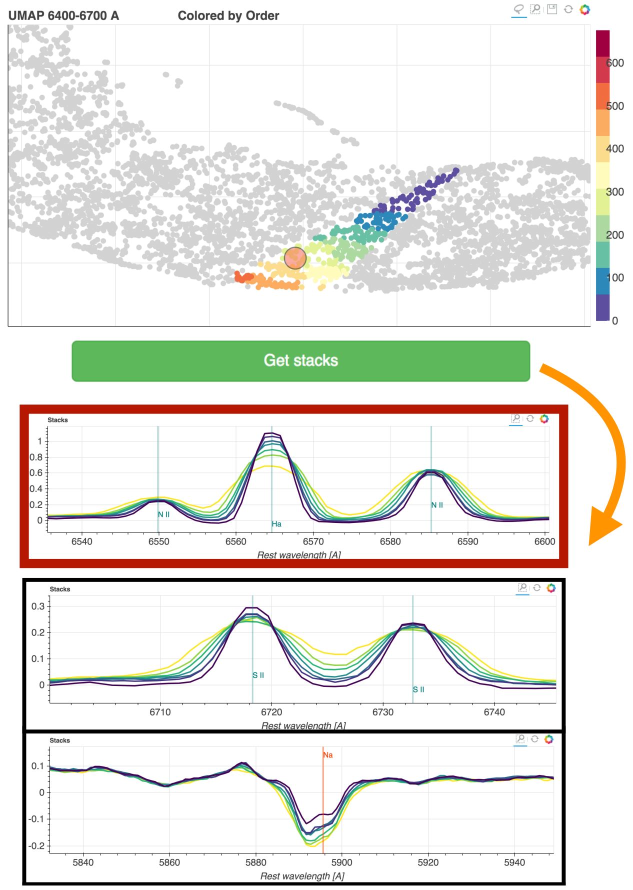


Figure 7. An example of trend exploration. The top panel shows a part of the $\lambda = 6400 - 6700$[Å] map, from which some of the galaxies are selected (the ones that are colored). The galaxies are ordered according to their location on the map, with color following the order. The bottom panels show different wavelength regions of the stacked spectra of the galaxies binned according to this order. In the example shown here the galaxies seem ordered by a combination of the line width and line ratio. The yellow lines have more AGN-like emission features; broader emission lines and higher NII to H_α line ratio. In this example we also see a non-trivial correlation with the NaI D absorption line that seems stronger in the AGN-like galaxies.

are ordered according to a combination of the emission line width and the H_α to NII amplitude ratio. In general the emission features are more AGN-like towards the stack colored in yellow. Trends can be discovered by looking for correlated changes in other regions of the spectrum. With well known samples or object types, one will mostly find trivial trends (e.g., emission line amplitudes that are correlated). In this example, however, we find that the equivalent width of the NaI D absorption correlates with the width of the H_α -NII complex. We see that the more AGN-like stacks have stronger NaI D absorption. Note that the bottom left panel of Figure 6 shows a similar trend as objects with deep NaI D absorption lie in the AGN region of the BPT diagram.

4 SUMMARY

We apply various anomaly detection methods to the same dataset of SDSS galaxy spectra, and show that while they all succeed, they disagree on most of the top few hundred outliers. This naively surprising result is a natural manifestation of the fact that for high dimensional data there could be many different yet reasonable answers to the questions we ask of any unsupervised ML algorithm. As a consequence, any single method will only output a small subset of the answers we wish for.

In order to increase our chances of making discoveries, a practical approach is to accumulate many different results of unsupervised ML and make them available for inspection by the community in an easy and intuitive way. For this, we are developing an exploration tool for astronomical data, that brings together interactivity and unsupervised ML. The main feature of the portal is 2D embeddings, or maps, of the data, where objects that are similar in a given subspace are located near each other.

We demonstrate and develop our portal using the dataset of SDSS galaxy spectra. For this dataset we built several maps, each constructed using a different wavelength region. We also include several additional embedding (e.g, the BPT diagrams), various coloring schemes using metadata, and the results of a number of anomaly detection methods.

We show how such tools and approach can allow for a more effective discovery process when interested in anomaly detection or when searching for trends and correlations. Additional uses include the search for objects of interest via similarity (‘find me more objects like this one’; Reis et al. 2018a), or for studying an object in context of its peers (‘how similar is this objects to others from other perspectives?’). We believe such methods will soon be indispensable.

ACKNOWLEDGEMENTS

This research was supported by Grants No. 2014413 and 2018017 from the United States-Israel Binational Science Foundation (BSF), and Grant No. 541/17 from the Israeli Science Foundation (ISF). JXP acknowledges support to the Applied Artificial Intelligence Initiative by UC Santa Cruz. This research made use of: `Bokeh`, `scikit-learn` (Pedregosa et al. 2011), `SciPy` (including `pandas` and `NumPy`, Jones et al. 01), `IPython` (Pérez & Granger 2007), `matplotlib` (Hunter 2007), `Astropy` (Astropy Collaboration et al. 2013), `Numba` (Lam et al. 2015), the SIMBAD database (Wenger et al. 2000), and the NASA Astrophysics Data System Bibliographic Services.

This work made extensive use of SDSS data. Funding for the Sloan Digital Sky Survey IV has been provided by the Alfred P. Sloan Foundation, the U.S. Department of Energy Office of Science, and the Participating Institutions. SDSS-IV acknowledges support and resources from the Center for High-Performance Computing at the University of Utah. The SDSS web site is www.sdss.org.

SDSS-IV is managed by the Astrophysical Research Consortium for the Participating Institutions of the SDSS Collaboration including the Brazilian Participation Group, the Carnegie Institution for Science, Carnegie Mellon University, the Chilean Participation Group, the French Participation Group, Harvard-Smithsonian Center for Astrophysics, Instituto de Astrofísica de Canarias, The Johns Hopkins University, Kavli Institute for the Physics and Mathematics of the Universe (IPMU) / University of Tokyo, Lawrence Berkeley National Laboratory, Leibniz Institut für Astrophysik Potsdam (AIP), Max-Planck-Institut für Astronomie (MPIA Heidelberg), Max-Planck-Institut für Astrophysik (MPA Garching), Max-Planck-Institut für Extraterrestrische Physik (MPE), National Astronomical Observatories of China, New Mexico State University, New

York University, University of Notre Dame, Observatário Nacional / MCTI, The Ohio State University, Pennsylvania State University, Shanghai Astronomical Observatory, United Kingdom Participation Group, Universidad Nacional Autónoma de México, University of Arizona, University of Colorado Boulder, University of Oxford, University of Portsmouth, University of Utah, University of Virginia, University of Washington, University of Wisconsin, Vanderbilt University, and Yale University.

REFERENCES

Abolfathi B., et al., 2017, preprint, (arXiv:1707.09322)

Aggarwal C. C., Yu P. S., 2001, SIGMOD Rec., 30, 37

Anders F., Chiappini C., Santiago B. X., Matijevič G., Queiroz A. B., Steinmetz M., Guiglion G., 2018, A&A, 619, A125

Astropy Collaboration et al., 2013, A&A, 558, A33

Bae H.-J., Woo J.-H., 2018, ApJ, 853, 185

Baines D., et al., 2017, PASP, 129, 028001

Baldwin J. A., Phillips M. M., Terlevich R., 1981, PASP, 93, 5

Baron D., Poznanski D., 2017, MNRAS, 465, 4530

Boroson T. A., Lauer T. R., 2010, AJ, 140, 390

Breiman L., Friedman J. H., Olshen R. A., Stone C. J., 1984, -

Cherinka B., et al., 2018, arXiv e-prints,

Comerford J. M., Greene J. E., 2014, ApJ, 789, 112

Eisenstein D. J., et al., 2011, AJ, 142, 72

Freund Y., Schapire R. E., 1997, J. Comput. Syst. Sci., 55, 119

Gisbrecht A., Hammer B., 2015, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 5, 51

Graur O., Maoz D., 2013, MNRAS, 430, 1746

Guillochon J., Parrent J., Kelley L. Z., Margutti R., 2017, ApJ, 835, 64

Heckman T. M., Lehnert M. D., Strickland D. K., Armus L., 2000, ApJS, 129, 493

Hunter J. D., 2007, Computing In Science & Engineering, 9, 90

Ichinohe Y., Yamada S., 2019, MNRAS, 487, 2874

Jofré P., Mädler T., Gilmore G., Casey A. R., Soubiran C., Worley C., 2015, MNRAS, 453, 1428

Jones E., Oliphant T., Peterson P., et al., 2001–, SciPy: Open source scientific tools for Python, http://www.scipy.org/

Keller F., Muller E., Bohm K., 2012, in 2012 IEEE 28th International Conference on Data Engineering. pp 1037–1048, doi:10.1109/ICDE.2012.88

Kohonen T., 1982, Biological Cybernetics, 43, 59

Kollmeier J. A., et al., 2017, preprint, (arXiv:1711.03234)

Kriegel H.-P., Kröger P., Schubert E., Zimek A., 2009, in Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. PAKDD '09. Springer-Verlag, Berlin, Heidelberg, pp 831–838, doi:10.1007/978-3-642-01307-2_86, http://dx.doi.org/10.1007/978-3-642-01307-2_86

Lam S. K., Pitrou A., Seibert S., 2015, in Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC. LLVM '15. ACM, New York, NY, USA, pp 7:1–7:6, doi:10.1145/2833157.2833162, http://doi.acm.org/10.1145/2833157.2833162

Levi M., et al., 2013, preprint, (arXiv:1308.0847)

Lintott C. J., et al., 2008, MNRAS, 389, 1179

Liu F. T., Ting K. M., Zhou Z.-H., 2008, in Proceedings of the 2008 Eighth IEEE International Conference on Data Mining. ICDM '08. IEEE Computer Society, Washington, DC, USA, pp 413–422, doi:10.1109/ICDM.2008.17, http://dx.doi.org/10.1109/ICDM.2008.17

Lochner et al., 2019

Majewski S. R., APOGEE Team APOGEE-2 Team 2016, Astronomische Nachrichten, 337, 863

McInnes L., Healy J., Saul N., Grossberger L., 2018, The Journal of Open Source Software, 3, 861

Meusinger H., Schalldach P., Scholz R.-D., in der Au A., Newholm M., de Hoon A., Kaminsky B., 2012, A&A, 541, A77

Müller E., Assent I., Steinhausen U., Seidl T., 2008, in 2008 IEEE 24th International Conference on Data Engineering Workshop. pp 600–603, doi:10.1109/ICDEW.2008.4498387

Müller E., Schiffer M., Seidl T., 2010, in Proceedings of the 19th ACM International Conference on Information and Knowledge Management. CIKM '10. ACM, New York, NY, USA, pp 1629–1632, doi:10.1145/1871437.1871690, http://doi.acm.org/10.1145/1871437.1871690

Nun I., Pichara K., Protopapas P., Kim D.-W., 2014, The Astrophysical Journal, 793, 23

Nun I., Protopapas P., Sim B., Chen W., 2016, The Astronomical Journal, 152, 71

Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825

Pérez F., Granger B. E., 2007, Computing in Science and Engineering, 9, 21

Protopapas P., Giammarco J. M., Faccioli L., Struble M. F., Dave R., Alcock C., 2006, MNRAS, 369, 677

Reis I., Poznanski D., Hall P. B., 2018a, MNRAS,

Reis I., Poznanski D., Baron D., Zasowski G., Shahaf S., 2018b, Monthly Notices of the Royal Astronomical Society, p. sty348

Richards J. W., Starr D. L., Miller A. A., Bloom J. S., Butler N. R., Brink H., Crellin-Quick A., 2012, ApJS, 203, 32

Rotman. M., Reis. I., Poznanski. D., Wolf. L., 2019, in Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR,. SciTePress, pp 124–134, doi:10.5220/0008163301240134

Shamir L., Wallin J., 2014, Monthly Notices of the Royal Astronomical Society, 443, 3528

Shi T., Horvath S., 2006, Journal of Computational and Graphical Statistics, 15, 118

Traven G., et al., 2017, The Astrophysical Journal Supplement Series, 228, 24

VanderPlas J., Connolly A. J., Ivezic Z., Gray A., 2012, in Proceedings of Conference on Intelligent Data Understanding (CIDU. pp 47–54 (arXiv:1411.5039), doi:10.1109/CIDU.2012.6382200

Wenger M., et al., 2000, A&AS, 143, 9

Zhang Y., Zhao Y., 2004, A&A, 422, 1113

Zimek A., Schubert E., Kriegel H.-P., 2012, Statistical Analysis and Data Mining, 5, 363

in der Au A., Meusinger H., Schalldach P. F., Newholm M., 2012, A&A, 547, A115

van der Maaten L., Hinton G., 2008, -

van der Walt S. J., Crellin-Quick A., Bloom J. S., 2019, Journal of Open Source Software, 4

APPENDIX A: NUMBER OF ANOMALIES GIVEN THE OVERLAP

To calculate the probability of obtaining an overlap of k without assuming $n \gg g$, consider a specific order of randomly choosing the objects, where the first three objects are in the subgroup, and the next two are not. The probability for this scenario is

$$\frac{g}{n} \times \frac{g-1}{n-1} \times \frac{g-2}{n-2} \times \frac{n-g}{n-3} \times \frac{n-g-1}{n-4}. \tag{A1}$$

We see that for any order in which the objects are chosen, the probability for choosing k objects form the first subgroup and $g-k$ from the rest of the group is

$$\frac{g(g-1)\dots(g-k)\times(n-g)(n-g-1)\dots(n-g-(g-k))}{n(n-1)(n-2)\dots(n-g)}. \tag{A2}$$

Multiplying by the number of possible orders to choose the objects, the probability to obtain an overlap of k can be written as

$$\frac{g!}{(g-k)!}\frac{(n-g)!}{((n-g)-(g-k))!}\frac{(n-g)!}{n!}\binom{g}{k}. \tag{A3}$$

APPENDIX B: ANOMALY DETECTION METHODS

In this section we briefly describe the four anomaly detection methods that we used in the method comparison, applied to the galaxy spectra, and uploaded to our portal.

(i) PCA reconstruction error. This is an example for model-based approach for anomaly detection. PCA is used to model the data, and the anomaly score is defined to be the χ^2 between the model and the data. Additional examples of ways to model the data that could be used instead of PCA include independent component analysis (ICA), non-zero matrix factorization (NMF), Auto-Encoder, and physically motivated models.

(ii) Unsupervised Random Forest. This is an example for distance based approach for anomaly detection in which the definition of distance definition is based on an ensemble of classification trees. The similarity between two objects is defined to be the number of trees in which the objects end up on the same terminal node when propagated through the tree. The distance is the inverse of the similarity. An unsupervised Random Forest, which is a Random Forest (Breiman et al. 1984) trained to distinguish between real and synthetic data, is an example for an ensemble of classification trees. Other tree ensembles that could be used with the same distance definition include supervised Random Forest, Extremely Randomized Trees (ERTs), and boosted trees (e.g AdaBoost, Freund & Schapire 1997). Many other distance definitions exist, the simplest one being Euclidean distance.

(iii) Isolation Forest. This algorithm does not fall into any of the general three approaches given in Section 2. In Isolation Forest the inverse of the anomaly score is defined to be the number of nodes an object goes through before being isolated (i.e, found to be the only object on a node), summed over all the trees in the ensemble. Isolation Forest is commonly used with ERTs, but in principle could be used with other types of classification tree ensembles.

(iv) Fisher Vector based anomaly detection. This algorithm is related to the density based approach for anomaly detection. Instead of using the density itself, the anomaly score is defined to be the contribution of an object to the Fisher Information the data holds about the parameters of the density distribution model. We used a Gaussian Mixture Model (GMM) to model the density of the data, as in this case the gradients of the density with respect to its parameters, which are needed for the calculation of the anomaly score, have an analytical formula.

APPENDIX C: NaI D ANOMALIES SPECIFIC EXAMPLES

In this section we provide examples from the various types of galaxies detected as anomalies from the NaI D `UMAP`, as described in Section 3.2. Figure C1 shows example spectra of a number of different types of such anomalies. For each object the left panel shows the entire SDSS spectrum, and the right panel shows the zoom in on the NaI D region which was used to detect the objects. A number of example objects for the different types of detected anomalies listed in the following tables: Table C1: galaxies with large blueshift NaI D. Table C2: galaxies with multiple component NaI D. Table C3: galaxies with redshifted NaI D. Table C4: galaxies hosting a type Ia supernova. Table C5: Various additional types of anomalies.

This paper has been typeset from a TEX/LATEX file prepared by the author.

index	SDSS name	comments
1	SDSS J142812.98+611115.6	SF
2	SDSS J094630.90+345500.6	wolf-rayet galaxy, SF, FOS
3	SDSS J104230.55+003441.9	E+A
4	SDSS J073856.16+320317.4	SF, E+A
5	SDSS J125427.34+022059.3	SF
6	SDSS J140621.04+252846.9	SB, ionized outflows
7	SDSS J125427.34+022059.3	E+A
8	SDSS J122715.39+062757.2	SF, E+A
9	SDSS J235047.12+143617.5	SF, ionized outflows
10	SDSS J141943.23+491411.9	SF, FOS
11	SDSS J083950.75+230836.1	SF, FOS
12	SDSS J025600.55+013829.5	SF, FOS
13	SDSS J031034.09+002938.7	SF, FOS

Table C1. Examples for galaxies with extreme NaI D blueshifts. Additional features are included in the comments column: SF = star forming emission lines, SB = star burst, FOS = face on spiral.

index	SDSS name	number of components
1	SDSS J084344.28+385340.6	2
2	SDSS J211138.95+044126.8	2
3	SDSS J103219.59+194052.6	3
4	SDSS J110534.89+410524.2	2
5	SDSS J125856.36+385053.4	2

Table C2. Galaxies with multi component NaI D profile.

index	SDSS name	absorption	emission
1	SDSS J155157.88+203056.9	redshifted	blueshifted
2	SDSS J234028.01-090945.0	redshifted	blueshifted
3	SDSS J120525.71+510611.1	redshifted	multi-component
4	SDSS J125553.16+581948.6	redshifted	multi-component
5	SDSS J005555.93+003940.2	multi-component	redshifted
6	SDSS J141518.01+230841.0	redshifted	blueshifted
7	SDSS J211635.95-004613.1	multi-component	redshifted

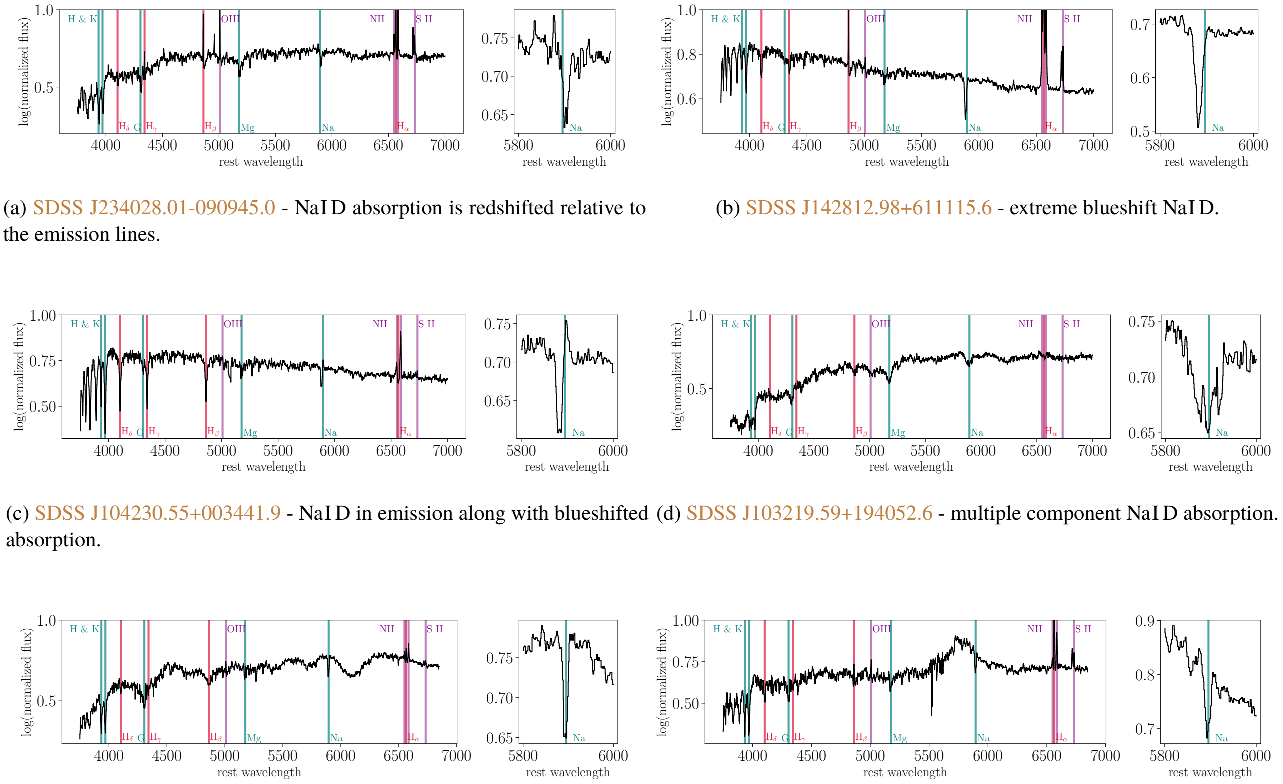
Table C3. Galaxies showing redshifted NaI D absorption. The absorption and emission columns refers to the location of the features relative to the SDSS systematic redshift. The absorption column refers to the other absorption lines in the spectrum, e.g, the H and K lines.

index	SDSS name
1	SDSS J080821.09+005035.3
2	SDSS J142608.24+152501.9
3	SDSS J095153.06+010605.8
4	SDSS J132301.39+243023.6
5	SDSS J140309.73+060754.3
6	SDSS J140237.96+034231.7
7	SDSS J091337.33+295958.4
8	SDSS J154024.75+325157.2

Table C4. Galaxies showing type Ia supernova features in their spectra.

index	SDSS name	Comments
1	SDSS J073714.26+431414.9	shifted set of absorption lines
2	SDSS J152613.25+495322.5	shifted set of absorption lines
3	SDSS J083316.31+152314.6	shifted set of absorption lines
4	SDSS J142812.98+611115.6	extreme blueshift NaI D
5	SDSS J143815.47+570445.1	chance alignment, brown dwarf
6	SDSS J052223.70+005916.4	unknown
7	SDSS J135124.77+054903.1	bad redshift

Table C5. Galaxies that are isolated on the NaI D `UMAP`, suggesting unique properties, not similar to any other galaxy in the sample.



(a) SDSS J234028.01-090945.0 - NaI D absorption is redshifted relative to the emission lines.

(b) SDSS J142812.98+611115.6 - extreme blueshift NaI D.

(c) SDSS J104230.55+003441.9 - NaI D in emission along with blueshifted absorption.

(d) SDSS J103219.59+194052.6 - multiple component NaI D absorption.

(e) SDSS J091337.33+295958.4 - type Ia supernova.

(f) SDSS J092034.87+511224.1 - unidentified broad feature.

Figure C1. Examples of galaxies with unusual NaI D line profiles detected with our portal. The blue vertical lines mark the locations of common absorption lines, the purple vertical lines mark common emission lines, and the red vertical lines mark the Balmer series that can be seen in both absorption and emission.