Applying Machine-Learning Algorithms to Large Datasets of Galaxy Spectra:

Finding Patterns and Detecting Outliers

Guy Goren
Phsyics Honors Program
Advisor: Prof. Dovi Poznanski

School of Physics and Astronomy, Tel-Aviv University, Tel Aviv 69978, Israel
28 August 2018

Abstract

As large astronomical datasets become greater and more accessible, we must ask ourselves how we utilize them to their fullest – understanding both macroscopic trends, as well as more specific interesting objects. Machine-learning (ML) based algorithms prove themselves as highly efficient in understanding complicated correlations among large datasets with dozens of features that characterize every object. By learning these correlations, we can detect trends, learn about the similarity and dissimilarity between objects, and detect specific outliers – objects that lack common correlations or that are characterized by new and unique correlations.

We present an unsupervised variation of the Random Forest (RF) algorithm, which provides us with a similarity measure between every pair of observations in our data set. We implement it over 150,000 spectra measurements of galaxies obtained from Sloan Digital Sky Survey (SDSS). By analyzing the similarity measures, we discover outlying observations, galaxies that are governed by unusual physical phenomena, as well as visualize the data to reveal underlying structures. As true for many ML algorithms, most of the methods discussed are generic and are suitable for a variety of domains and types of data.

The purpose of this document is to summarize the methods of work used, their advantages and disadvantages, as well as the unusual physical phenomena discovered, all in the purpose for an efficient future research. A demo which includes the relevant Python code is available upon demand.

Keywords

Galaxies, spectra, data analysis, outlier detection, machine learning, random forest.

1 INTRODUCTION

The article "The weirdest SDSS galaxies: results from an outliers detection algorithm" by D. Baron and D. Poznanski [1] presents the results obtained by applying a specific ML algorithm, Random Forest (RF), on a dataset consisting of spectra of over 2 million galaxies obtained from the Sloan Digital Sky Survey (SDSS). 400 outliers were analyzed and most of them were confirmed as galaxies with peculiar properties, such as: unusual velocity structure, unusual emission lines, sodium excess galaxies, galaxies hosting supernovae, galaxy-galaxy gravitational lenses, etc.

This research dives deeper into the capabilities of RF for analyzing the galaxy spectra obtained from SDSS:

- Further understanding how hyper-parameters and feature engineering affect the algorithm’s sensitivity to the properties of a spectrum, such as emission/absorption lines, continuum, noise, etc.
- Combining RF with dimensionality reduction and visualization tools such as t-distributed stochastic neighbor embedding (t-SNE).
- Pinpointing new outlying galaxies and small clusters, analyzing their spectra and understanding the physical phenomena that govern these objects.

In order to achieve these goals, first we must be able to compare spectra of different galaxies. This is achieved by de-reddening the spectra (compensating for light absorbed by dust clouds in our galaxy), shifting the spectra to the galaxies' rest-frame (de-redshift), interpolating the spectra to the same wavelength grid and normalizing. We also remove their continuum by fitting a polynomial to the flux and dividing the flux at each wavelength by the polynomial's value. One can then apply the unsupervised RF algorithm to the data. The output of the algorithm is a dissimilarity matrix which includes the 'distance' between every pair of galaxies in our dataset. By analyzing this matrix, we visualize their distribution in parameter space, find the galaxies furthest (similarity-wise) from all others (outliers), identify clusters, and more.

2 DATA SET & FEATURE ENGINEERING

2.1 Data Set

The data is obtained from the Sloan Digital Sky Survey (SDSS) DR14. Only objects classified with class='GALAXY' where taken (the exact query can be found with the demo). The 1-D version of the spectra was used. Our main sample is the 150,000 objects with the top median SNR. Along the entire document, unless stated otherwise, this is the sample discussed.

The sample is consisted of different types of galaxies. As a very rough approximation it can be estimated that about $\sim 90\%$ are 'red' galaxies and $\sim 10\%$ are 'blue' (see a more detailed 'color' distribution at section 4). This homogeneity of the data often has a significance influence on ML algorithms, as we would like the algorithm to learn the characteristics of a variety of objects, and not just the most common ones.

2.2 Minimum Redshift Threshold

Objects at a very low redshift (z) are often:

- Stars misclassified as galaxies.
- Since those objects are relatively close, often only the center of the galaxy is captured in the fiber.

As we are interested in detecting true outlying galaxies, we would like to minimize indicating bad classifications or bad measurements as outliers. Therefore, objects with $z \leq 0.01$ were discarded (only $\sim 1.23\%$ of the objects):

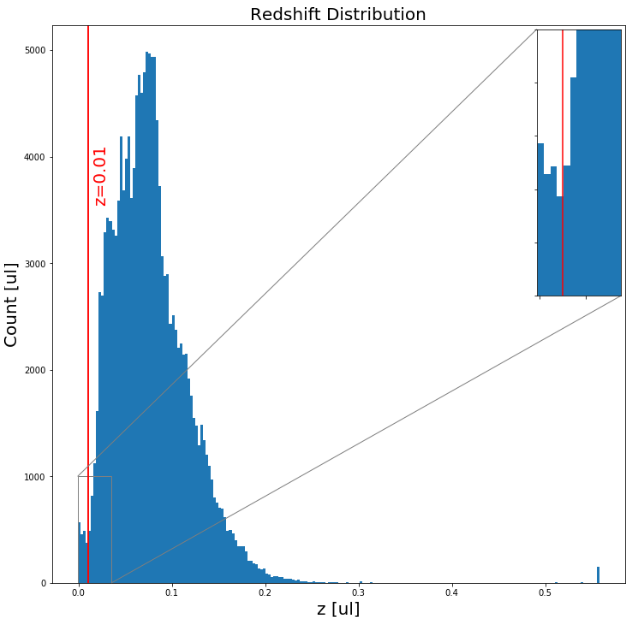

Figure 1: Redshift distribution of the top 150,000 SNR bin. The $z = 0.01$ threshold is marked.

2.3 Bad Pixels

The following flux values were discarded:

- All values in the range $(5565\ \text{Å}, 5590\ \text{Å})$, due to known high sky emissions.
- Values with an uncertainty (standard deviation) larger than the value itself.

2.4 Deredenned Spectrum

Each spectrum was deredenned in order to compensate for extinction in the Milky-Way galaxy. The $\mathrm{E}(B - V)$ values were interpolated based on the dust map obtained by Schlegel, Finkbeiner & Davis (1998) [2]. The extinction model used is based on the Fitzpatric99 reddening law.

2.5 Deredshift and Interpolation

We would like to have comparable features among the different galaxies – their flux values at the same wavelengths in their rest-frames. Therefore, first we must deredshift each spectrum to its rest frame:

$$(1)\quad \lambda_{rf} = \frac{\lambda_{ob}}{1+z}$$

λ_{rf} : Rest-frame wavelength ; λ_{ob}: Observed wavelength ; z: Redshift

Since the objects have a similar observer-frame band, but are at different redshifts, their resulting rest-frame band isn't the same. As we would like to use as much data available as possible, and at the same time not to introduce too many missing values into our sample, we must balance between the two. We chose a mutual wavelength grid of $[3749.7\ \text{Å}, 7896.7\ \text{Å}]$ such that only 1% of the objects have a rest frame band which doesn't include the entire mutual band. Inevitably, this choice results in multiple missing values at the edges of 1% of the objects. Finally, we interpolate each spectrum to the mutual wavelength band at $0.5\ [\text{Å}]$ spaces, resulting in a total of 8,295 flux values.

2.6 Normalization

Most of the information in the spectra is located at the correlations found in the spectra and not in the total flux emitted. Moreover, the flux measured falls as $\sim \frac{1}{r^2}$, where r, our distance to the galaxy, is generally unknown. Therefore, we normalize each spectrum such that its median flux value would be 1. Namely, to obtain the normalized flux, for each galaxy we divided its flux by the median flux.

2.7 Smoothing

The spectra was smoothed by using a $3.5\ [\text{Å}]$ running median filter in order to reduce some of the noise present in the data, especially sky flux that wasn't subtracted properly, resulting in very thin lines $(0.5\ [\text{Å}] - 1\ [\text{Å}])$, thinner than the spectroscopic lines that can be measured, given the spectrometer resolution.

2.8 Handling Missing Values

Missing values are present in our data for numerous reasons: bad exposures, bad pixels flagged by us (see 2.3), and placing the spectra on a mutual grid (see 2.5). As we would like to avoid objects with missing values indicated as outliers, we discard objects that over 15% of their features are missing values.

Since the Random Forest (RF) algorithm doesn't take missing values as an input, we impute missing values by using the median flux at each wavelength among all objects.

2.9 Removing Continuum

By analyzing the results obtained, we concluded that the unsupervised RF algorithm is highly sensitive to the galaxy's continuum (see Section 4). We aimed to balance between the continuum and spectroscopic lines influence on the results. Distinguishing the continuum from the lines isn't a trivial task and can be approached in many way, such as extracting known lines, spectral analysis (for example, by using a LPF to extract the continuum and a HPF for the spectral lines), etc. In our case, the continuum was removed for each spectrum by:

1) Sampling the continuum: Calculating the median flux at intervals of $300\ [\text{Å}]$. An extra sample was taken at each of the edges of the spectrum at a $100\ [\text{Å}]$ window. By doing so we ignore even the very wide lines.
2) Fitting a 5th degree polynomial to the samples.
3) Dividing the flux by the polynomial value for each wavelength.

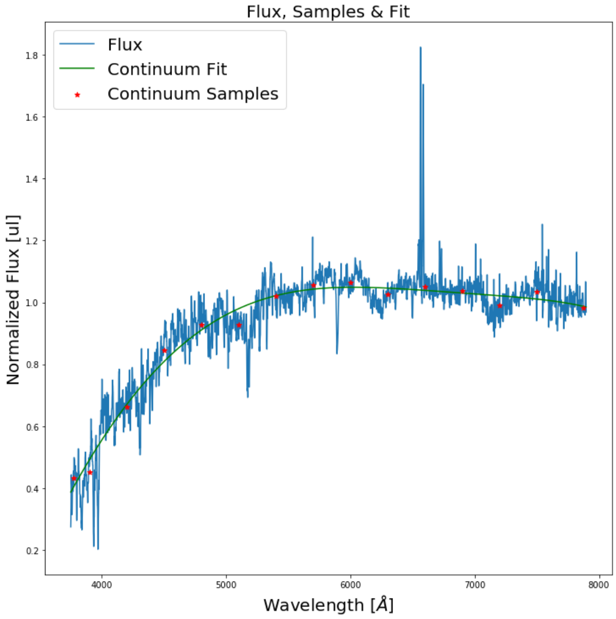

Figure 2a: An example for a flux, the continuum samples and the polynomial fit.

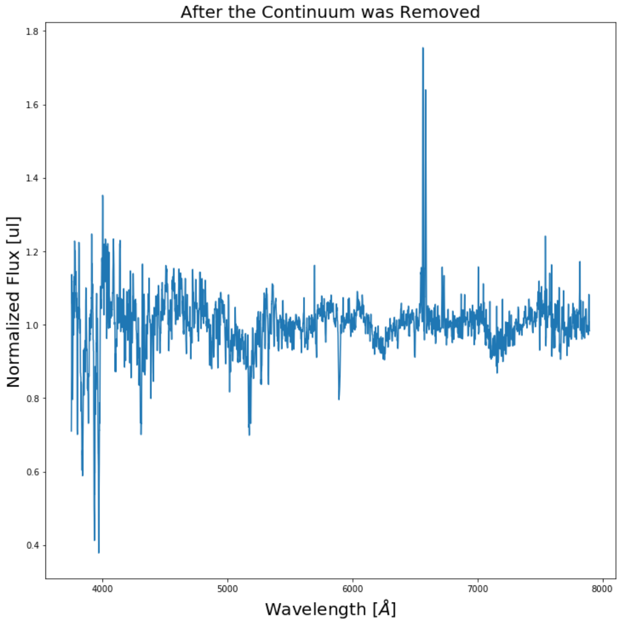

Figure 2b: The flux after its continuum was removed for the example from figure 2a. Spectroscopic lines are clearly visible, whereas continuum measurements all have flux values of $\sim\mathbf{1}$.

It is worth mentioning that the continuum is not removed completely, as different types of continuums are biased differently relative to the fit. Yet, as we'll see later, its influence on the results is dramatically reduced.

3 DISSIMILARITY MEASUREMENT ALGORITHM

3.1 Random Forest (RF) Basics

RF is an ensemble learning method, most commonly used as a supervised method for classification and regression. A matrix is given as an input to the classifier, along with a label vector. The matrix's rows represent different observations (galaxies, in our case) and its columns represent different features (wavelengths, in our case), such that each cell represent the flux at a given wavelength for a particular galaxy. The classifier operates by constructing many decision trees (a forest), where each tree is based only on a sample of observations drawn with replacement from the input. At each node of the tree, a set of features is drawn without replacement. Then, a condition for splitting the observations into two child nodes is calculated – a threshold for a feature, out of those that were drawn, which yields the best separation between the different classes. All of the observations with a value lower or equal to the calculated value are transferred to the left child node and those with a higher value, are transferred to the right child node. The process is repeated recursively for each child node until a stopping criteria is met (for example: maximum leaf depth, minimum number of observations to split, etc.).

RF's 'sklearn' implementation was used with the following hyperparameters: $min\_samples\_split = 10, max\_features = 1500, n\_estimators = 500$.

By the nature of the RF classifier, as trees are grown independently, it is easily parallelizable – making it suitable for working with large sets of data.

3.2 Unsupervised Implementation & Dissimilarity matrix

In our case, no labels are available naturally. Therefore, we perform unsupervised learning, similarly to the one described in Shi & Horvath (2006) [3]. Synthetic data is created with the same dimensions as the original – for each synthetic object, for each feature, a value is drawn with replacement from the same feature within original data (i.e. each feature's column is shuffled). This way, the synthetic data has the same marginal distributions as the original, but lacks the correlations. Our labels for the RF input would be 'real' and 'synthetic'. As a result, by applying the classifier to the data, trees are grown based on the correlations present in the data, which allows the RF algorithm to distinct real and synthetic observations.

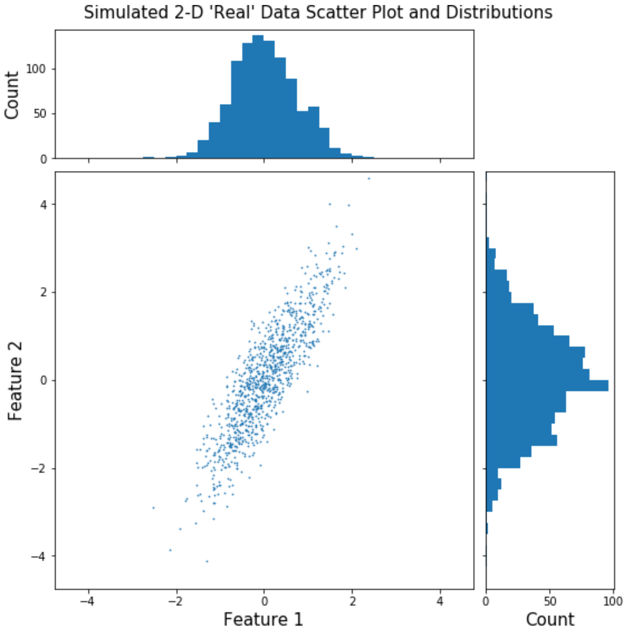

Figure 3a: A scatter plot with the marginal distributions for 1000 simulated observations data taken from a 2-D multivariate random distribution.

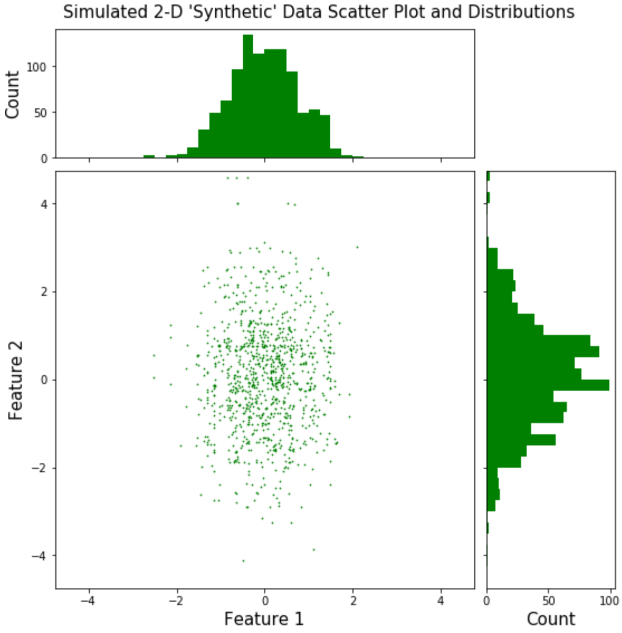

Figure 3b: A scatter plot with the marginal distributions for the 'synthetic' data obtained from the simulated data depicted in figure 3a.

As illustrated in figure 3a & 3b for a simulated data with 2 correlated features (in our case there are 8,295 features), the synthetic data has the same marginal distributions as the real one, but lacks the correlation.

Observations that end up at the same leafs often are assumed to have similar correlations, and hence are similar. We calculate a similarity matrix which measures how similar two galaxies are to each other by the number of trees they ended up at the same leafs and only if they were all classified as real, normalized by the total number of leafs:

$$(2)\quad S_{ij} = \frac{\Sigma_{k=1}^{n_{trees}} L_{ij}^{k}}{n_{trees}} \quad : \quad 0 \le S_{ij} \le 1$$

S_{ij} : Similarity between the galaxies i & j.

n_{trees}: Number of trees.

L_{ij}^{k} is an indicator, which equals 1 if galaxies i & j ended up at the same leaf of the k^{th} tree and both classified as real, and 0 otherwise. The latter condition assumes that trees that classify a real galaxy as synthetic one didn't recognize its correlations successfully.

Lastly, in order to achieve the dissimilarity between each pair of galaxies D_{ij}, we calculate:

$$(3)\quad D_{ij} = 1 - S_{ij}$$

This dissimilarity matrix describes an abstract distance between two objects in parameters space. We expect objects with similar correlations to have a distance close to 0 (always end up at the same leafs) and objects which barely share mutual correlations to be at a distance 1 away (never end up at the same leafs).

Even though we interpret D_{ij} as a distance matrix, we must keep in mind that D_{ij} only partially maintains the properties of a distance matrix:

- The entries on the main diagonal are all zero. i.e. The distance of every object to itself is zero, as it is trivially always end up at the same leaf with itself.
- Not all of the off-diagonal entries are different than zero. i.e. Two galaxies could end up at the same leafs at all trees.
- D_{ij} is indeed a symmetric matrix.
- The triangular inequality doesn't necessarily hold. Specifically, for example, 2 galaxies could be very far from each other, but very close to a third galaxy, each on its own.
- The distances aren't necessarily linear.
- The distances are bounded – the maximal possible distance is 1.

We should keep those properties in mind when trying to interpret and visualize them.

4 CONTINUUM INFLUENCE AND REMOVAL

As mentioned in section 2.9, after analyzing the continuum influence on our results, namely its influence on D_{ij}, we concluded that we should remove the continuum as its influence was too significant, however it doesn't hold most of the interesting physical information. Here we'll present the main measure used to estimate this influence.

In order to quantify the continuum, for each spectrum we calculate the average wavelength (λ), weighted by the measured flux (f) at that wavelength:

$$(4)\quad \bar{\lambda} = \frac{\Sigma_i \lambda_i f_i}{\Sigma f_i}$$

This measure indeed is determined mostly by the continuum, as most of the spectrum is consisted of the continuum. Even though some lines might have significant high/low flux values – it would have a negligible influence on $\bar{\lambda}$ as they span along thin bands.

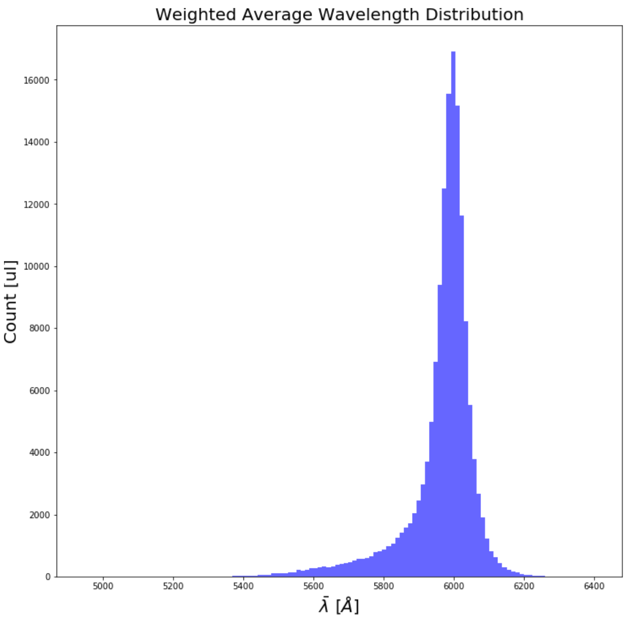

Figure 4: Weighted average wavelength distribution.

Figure 4 depicts the homogeneity of our data: most of our sample consists objects with a relative high average wavelength ('red galaxies'), and much fewer lower wavelength ('blue galaxies'). Extreme objects from both ends of the spectrum aren't common as well.

By comparing $\bar{\lambda}$ and W_{100} we can estimate the continuum influence over the weirdness score, with and without removing the continuum. We create a scatter plot where each point represents a galaxy by its $\bar{\lambda}$ and W_{100} values:

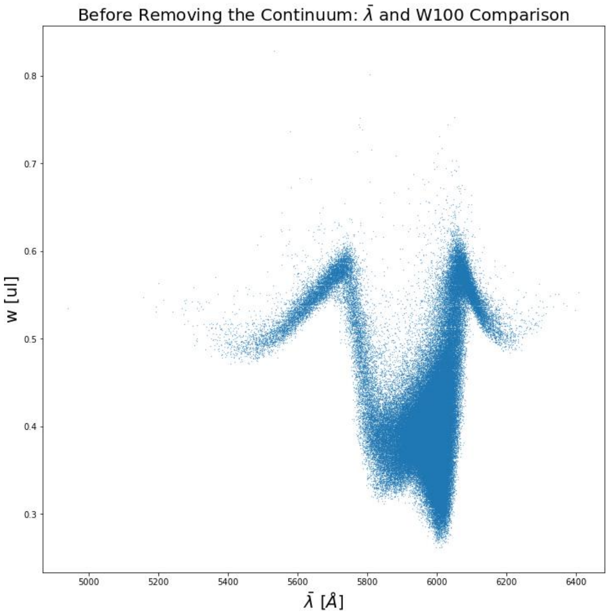

Figure 5a: Before removing the continuum: $\bar{\lambda}$ and W100 scatter plot.

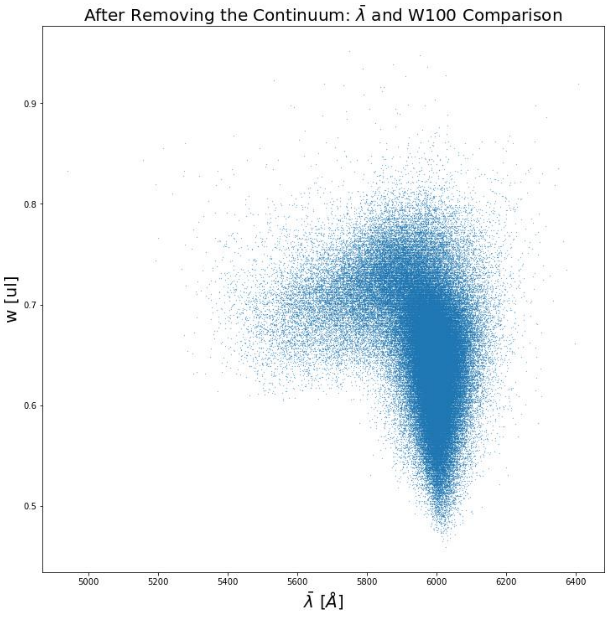

Figure 5b: After removing the continuum: $\bar{\lambda}$ and W100 scatter plot.

We can clearly see in Figure 5a that without removing the continuum, $\bar{\lambda}$ and W_{100} are highly correlated, resulting in a nearly 1-D scatter plot. This isn't a desired result, as knowing $\bar{\lambda}$, which is easily calculated analytically, allows us to determine W_{100} with a relatively low uncertainty, except of some exceptions. That means W_{100} hardly holds any new interesting information, other than $\bar{\lambda}$.

In Figure 5b we see that this 1-D relation doesn't hold anymore, as objects across all $\bar{\lambda}$ range are given a variety of W_{100} values. There is still a strong relation between the two, as the continuum information isn't removed completely:

- Different polynomial fits, as done in the continuum removal process, are biased differently relative to the continuum, depending on its type. i.e. the fit might be above or below the flux values, depending on the continuum itself.
- The continuum is indirectly correlated with other characteristics of the spectrum.

The bottom of the scatter plot, corresponding to low W_{100} scores, matches well with the $\bar{\lambda}$ distribution (figure 4), as one might expect – the most common type of objects in the data set will, most probably, have the nearest neighbors, hence are the least weird.

5 DISSIMILARITY MATRIX VISUALIZATION BY T-DISTRIBUTED STOCHASTIC NEIGHBOR EMBEDDING (T—SNE)

t-SNE is a nonlinear dimensionally reduction technique, which models any high-dimensional object by a 2 or 3 low-dimensional space coordinates, specializing at preserving nearby neighbors. This algorithm is most commonly used by calculating a $n \times n$ Euclidean distance matrix for a set of n observations, and then reduce its dimensionality in order for it to be visualized. Here we use the precomputed dissimilarity matrix D_{ij}, as the distance matrix, and visualize it in a 2-D space.

We must keep in mind that by reducing the dimensionality of the data from 149,999 ($dim = number\ of\ observations - 1$) to 2, a lot of information is lost, such that the exact relation between the galaxies is somehow distorted. Moreover, we should expect further distortion as a result of the fact that the dissimilarity matrix doesn't preserve all of the properties of a distance matrix (see 3.2).

During the research, other algorithms were used as well, including multidimensional scaling (MDS) and principal component analysis (PCA), but failed to provide a result as useful as t-SNE, considering our main goals – detecting outliers and small clusters.

5.1 Distribution and Overall Trends

The scatter and density plots first provide us with a general notion of the distribution of our data – how similar/different galaxies are from each other, how many clusters are present, which objects tend to stay away from most of the sample, etc.

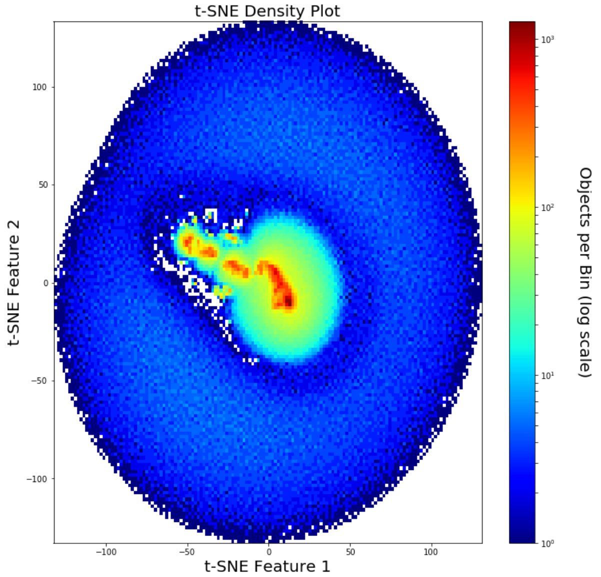

Figure 6a: t-SNE density plot.

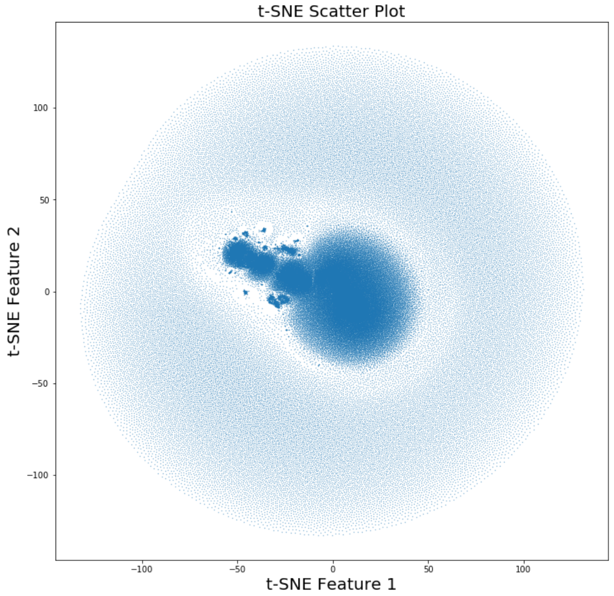

Figure 6b: t-SNE scatter plot.

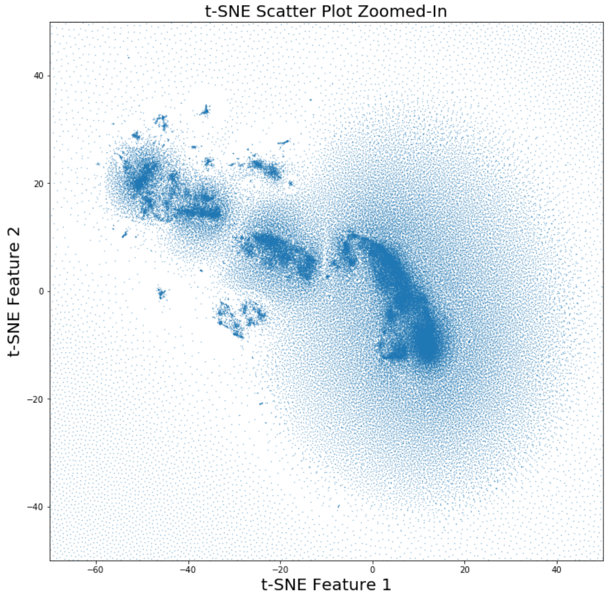

Figure 6c: t-SNE scatter plot, zoomed-in on the main cluster, discarding the circular 'cloud'.

An undesired circular 'cloud' is visible in the plots, which is a result of a failure of the t-SNE algorithm dealing with our dissimilarity matrix, which as explained doesn't preserve all of the properties of a distance matrix. In this section, when analyzing the t-SNE plots, we'll ignore this 'cloud' and limit our discussion to the main cluster of objects, as seen in figure 6c. It is possible to evade this phenomena to some degree by using a higher perplexity as a hyperparameter (perplexity=25 was used), on the expense of risking in losing the isolation of some small clusters, as well as a higher computational time.

Few clusters are seen – areas with a high density of objects, corresponding to similar objects. Some of the areas are also distant from the rest of the sample, suggesting that they share certain characteristics that others don't, as the t-SNE algorithm aims at keeping similar objects by nearby points and dissimilar objects are modelled by distant points with high probability. We'll focus on some of the clusters on section 7.

5.2 Physical Properties

We are interested in knowing how the distances are correlated with the physical properties of the galaxies. First, we look at the continuum ($\bar{\lambda}$; see eq. (4)):

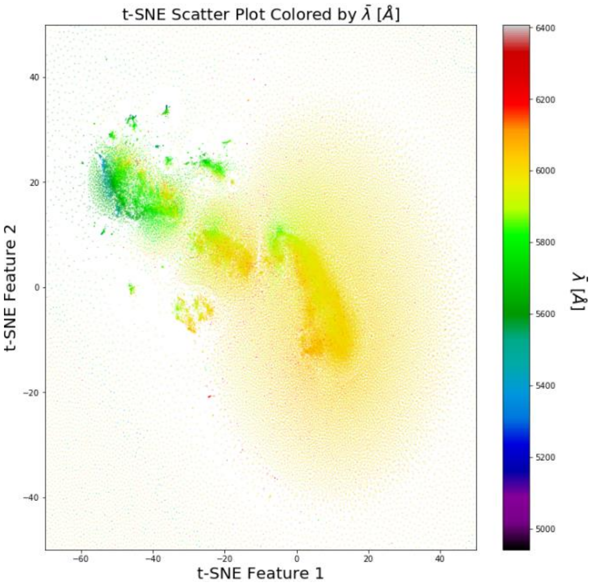

Figure 7a: t-SNE scatter plot colored by $\bar{\lambda}$.

As we see, even though the continuum was removed from the spectra, it is highly correlated with the distances – galaxies of different $\bar{\lambda}$ spread over different areas of the t-SNE map, where the most extreme $\bar{\lambda}$ objects end up at two different edges, with a clear color gradient imbetween.

We present more t-SNE maps colored by a variety of physical properties. Some of the properties are missing for some of the galaxies, hence only galaxies with available values are plotted.

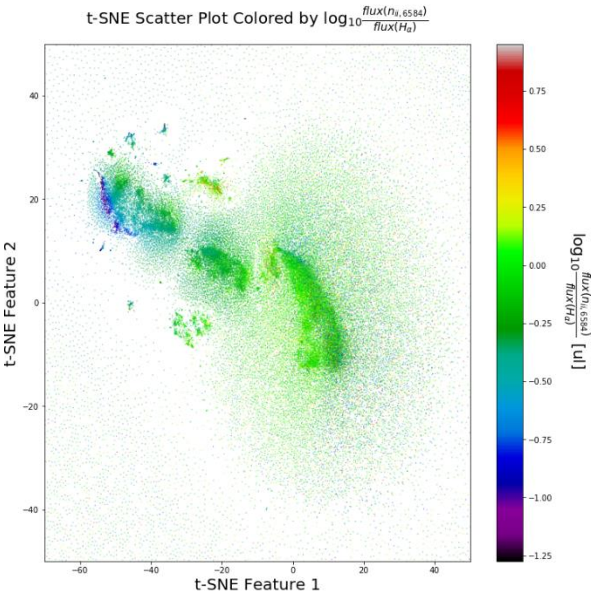

Figure 7b: t-SNE scatter plot colored by $\log_{10}\frac{flux(n_{ii,6154})}{flux(H_\alpha)}$.

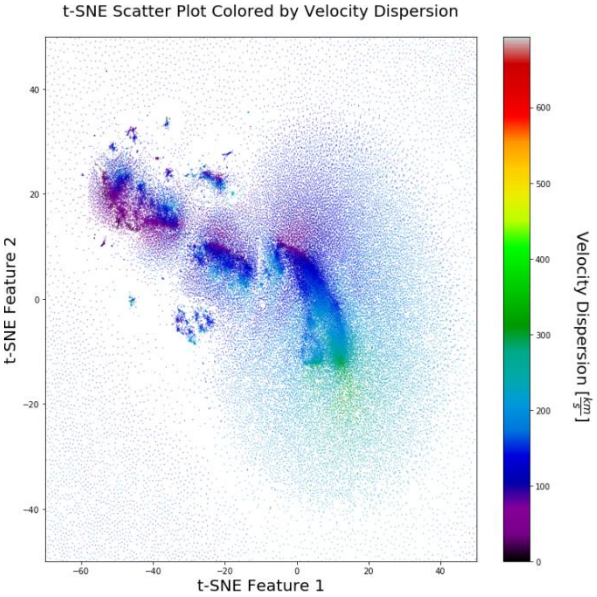

Figure 7c: t-SNE scatter plot colored by velocity dispersion (from Schlegel).

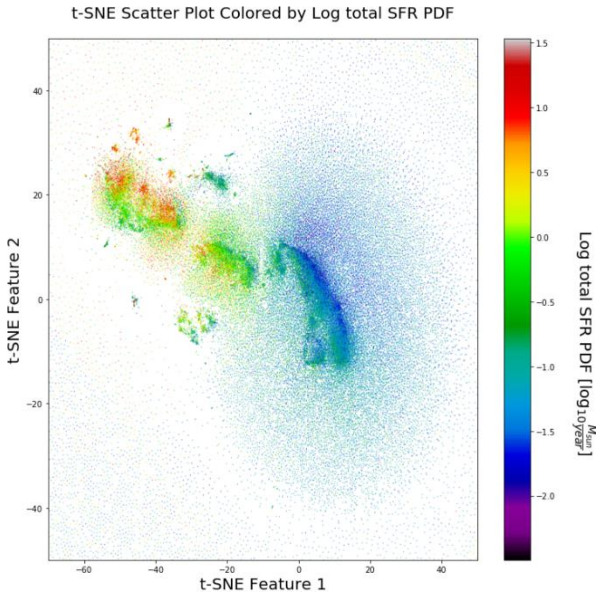

Figure 7d: t-SNE scatter plot colored by Log total SFR PDF.

As seen, the t-SNE features are highly correlated with a variety of physical properties (some of them are cross-correlated). We expect the high-dimensionality dissimilarity-matrix to hold much more information regarding more detailed physical properties.

6 OUTLYING GALAXIES

Outliers aren't objects with specific well-defined characteristics. Yet, a reasonable abstract definition would be objects that share very few characteristics with other objects in the sample, and/or have their own unique characteristics. In our case, we look for objects which are furthest from the rest of the sample, as determined by D_{ij}. We define a weirdness score (W_n) to be the average distance of an observation to its n nearest neighbors: We'll often use $n = 100$ ($\sim 0.066\%$ of the sample), as a reasonable number – we would expect that non-outliers would have much more than 100 very near neighbors (with a low D_{ij}), hence result in a low weirdness score, whereas outliers wouldn't have as many as 100 near neighbors, hence resulting in a high weirdness score.

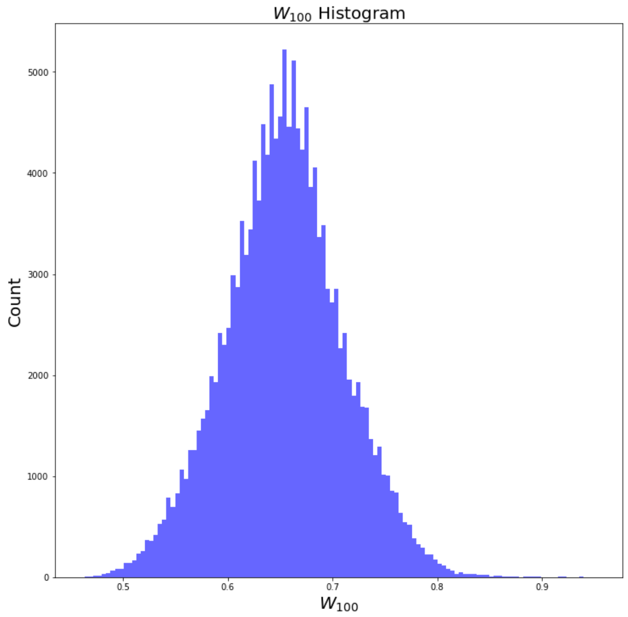

Figure 8: W_{100} histogram.

As depicted in Figure 3, as one might expect, most of the objects have a similar weirdness score, corresponding to the fact that our data is mostly homogenous, consisting of very similar spectra, yet with various degrees of similarity.

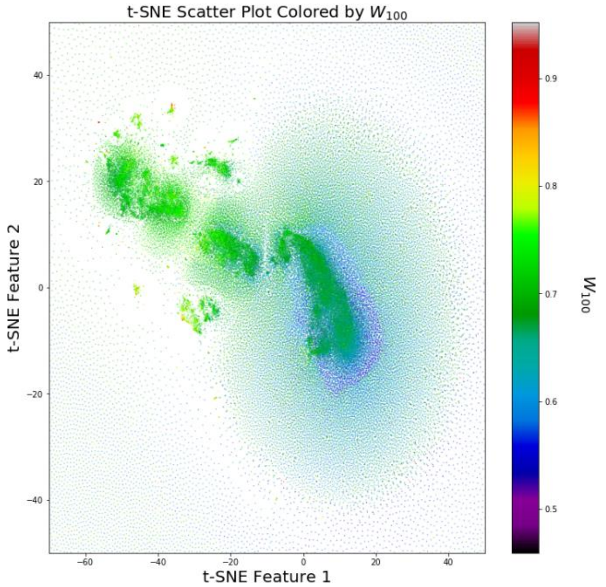

Figure 9: t-SNE scatter plot colored by W_{100}.

Objects with high W_{100} tend to be either true outliers, or a result of bad measurements or a failure within the SDSS pipeline. We present the main types of outliers found, as obtained from analyzing the spectrum and photometry of the 150 objects with the highest W_{100} score (appendix 11.1). The exact W_{100} scores of the entire sample are available upon demand.

6.1 Failed Measurements

6.1.1 Missing a Band of Flux Values

About 15% of the top outliers are measurements where a part of their spectrum is missing. This is despite the fact that objects with a high number of missing values were discarded, and any missing values that were present in the objects that were left were imputed (see 2.8).

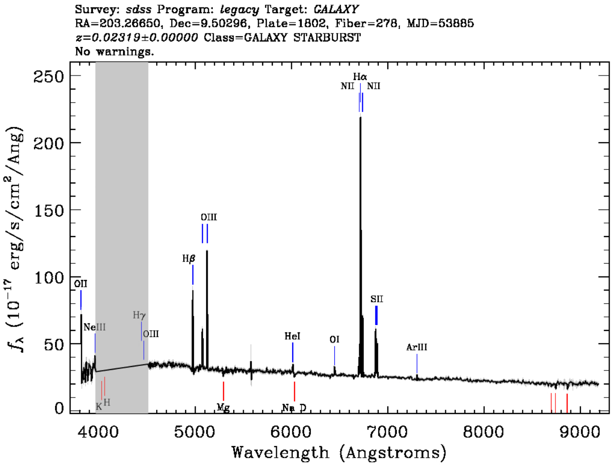

Figure 10: An example for an outlier with a missing band of flux values.

6.1.2 Chance Alignments

About 4% of the top outliers are chance alignments, when two or more objects are aligned on our line of sight (LOS), such that the resulting spectrum is a mixture of their spectra.

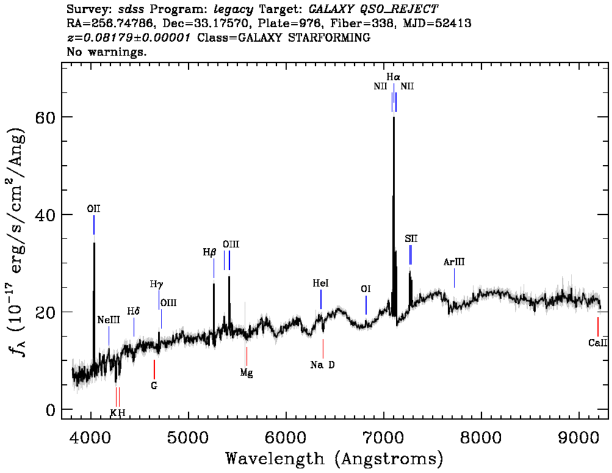

Figure 11a: An example for the spectrum of a chance alignment of a galaxy and a star.

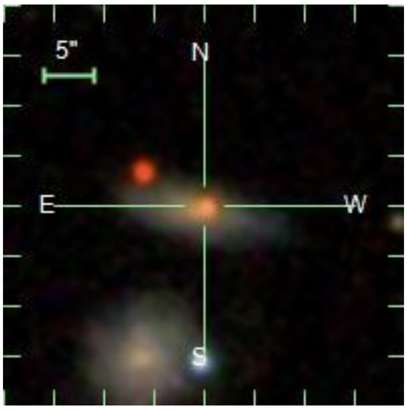

Figure 11b: The photometry of the object from figure 11a. A star on top of a fainter galaxy is seen.

6.1.3 More

There are other types of failed measurements, resulting in an unreliable spectrum.

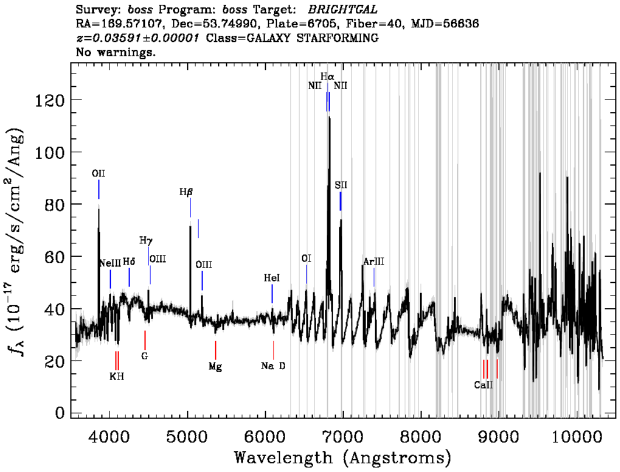

Figure 12: An example for the spectrum of a bad measurement.

6.2 Pipeline Failures

Even though the SDSS pipeline is robust, providing a reliable spectrum and classifying objects properly most of the times, when looking for outliers, specific objects where the pipeline failed show up.

6.2.1 Quasars & Blazars Misclassified as Galaxies

About 10% of the top outliers are quasars & blazars misclassified as galaxies. Since their spectra is highly distinct from galaxies', characterized especially in many broad spectral lines, it is unsurprising that those objects show up as outliers in our data set.

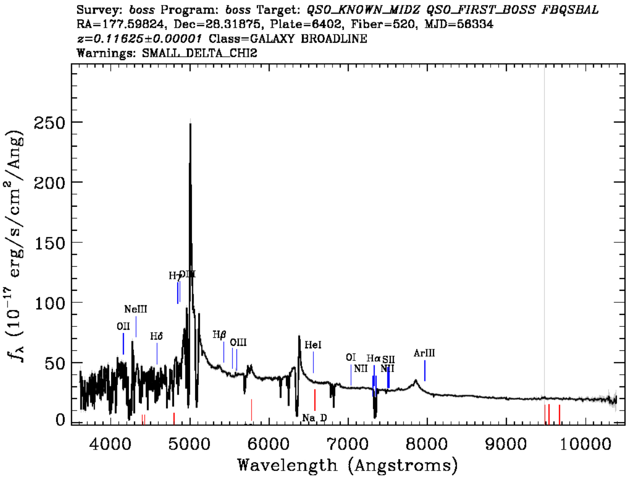

Figure 13a: An example for a quasar misclassified as a galaxy.

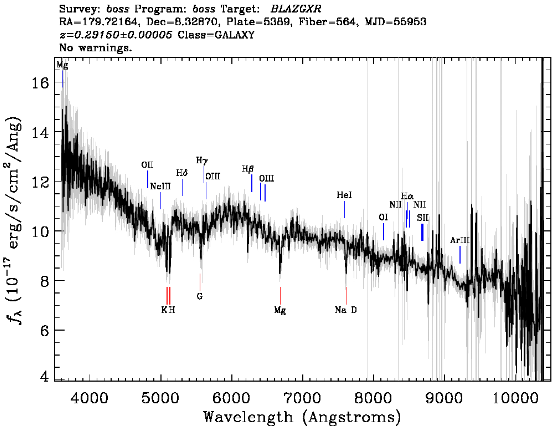

Figure 13b: An example for a blazar misclassified as a galaxy.

6.2.2 White Dwarf (WD) Stars Misclassified as Galaxies

About ~1% of the top outliers are WD stars misclassified as galaxies.

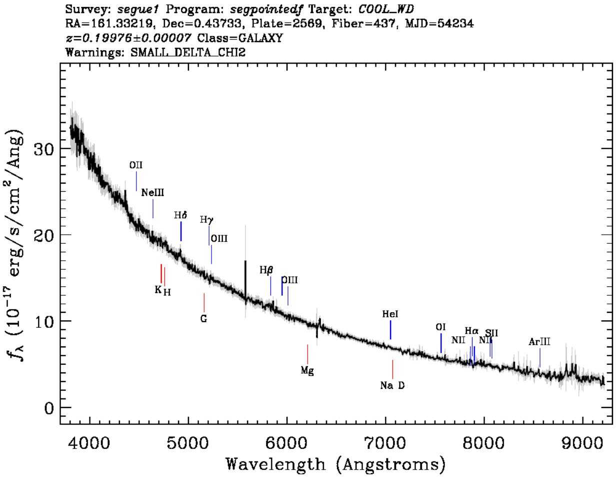

Figure 14: An example for a WD star misclassified as a galaxy.

6.2.3 Bad Sky Emissions Subtraction

About 20% of the top outliers are objects with a wrong sky emissions subtraction done in the SDSS pipeline. For wavelengths where high sky emissions are present, sometimes they are not subtracted properly from the measured flux, resulting in the addition of fake emission/absorption lines in the spectrum. These lines are usually very narrow. This is despite of the fact that the flux was smoothed (see 2.7) in order to avoid this exact problem.

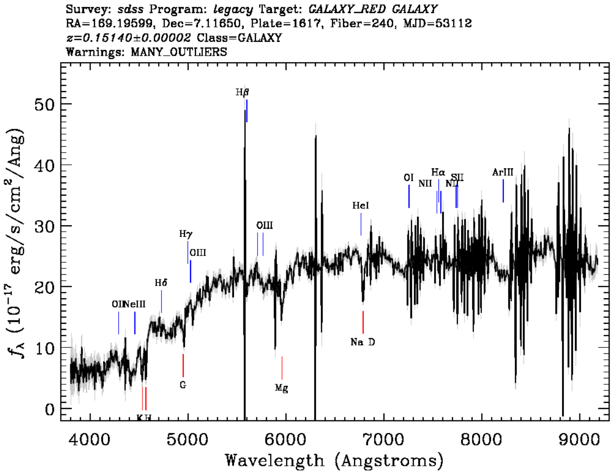

Figure 15: An example for an outlier due to bad sky subtraction. Many false lines are present around $\lambda > 7000$ [Å]

6.3 Outliers

Finally, the most interesting results are galaxies with proper spectrum measurements that contain unusual characteristics. This is about 10% of the top W_{100} objects. We only present the most interesting objects' spectrum. A deeper analysis of their characteristics would be done in future research.

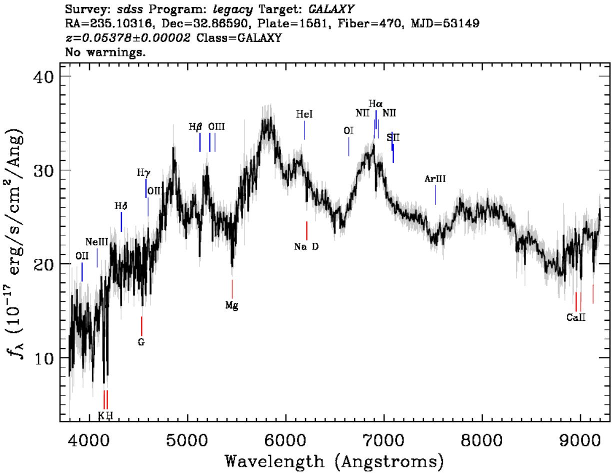

Figure 16a: An outlier's spectrum – a galaxy that hosted a supernova when the spectrum was measured.

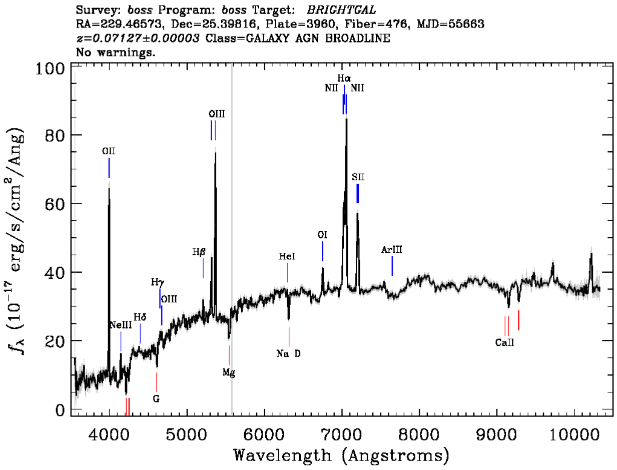

Figure 16b: An outlier's spectrum – double spectral lines appear, probably as a result of two merging galaxies.

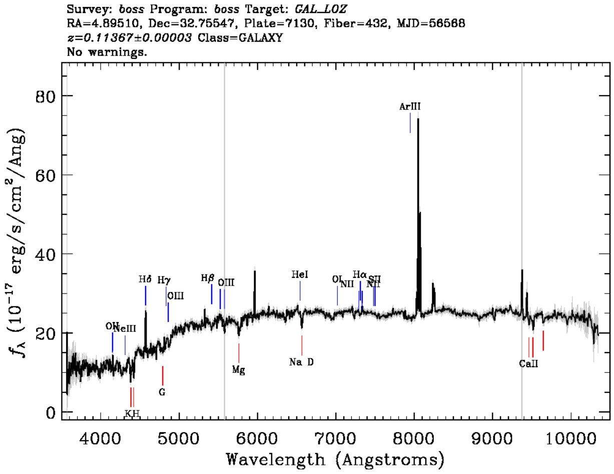

Figure 16c: An outlier's spectrum – unrecognized emission lines appear, corresponding to an object at $z \sim 2 \cdot 0.11 = 0.22$. Probably a lensed system.

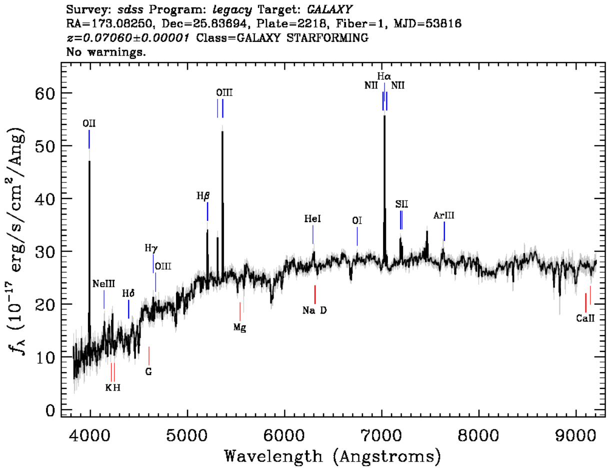

Figure 16d: An outlier's spectrum – unrecognized emission & absorption spectral lines are present.

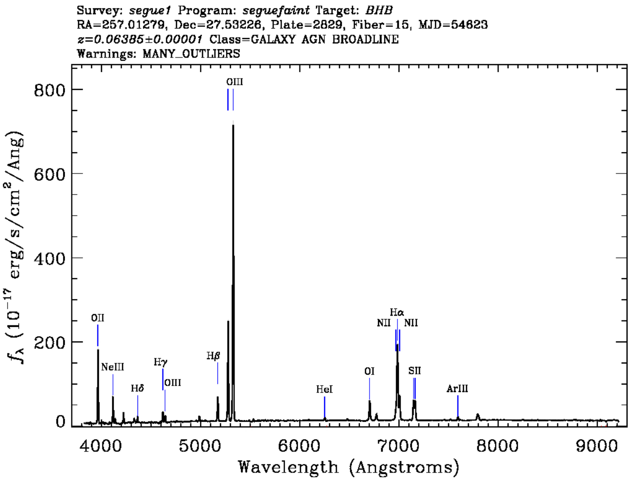

Figure 16e: An outlier's spectrum – very strong emissions. Only 50-80 objects (<0.05% of the sample) are characterized in similar Neon, Argon & Helium emission lines.

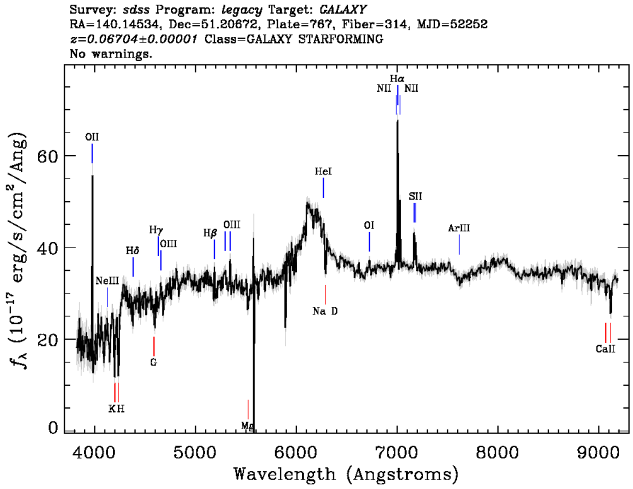

Figure 16f: An outlier's spectrum – unknown structure at $6000 - 6400$ [Å].

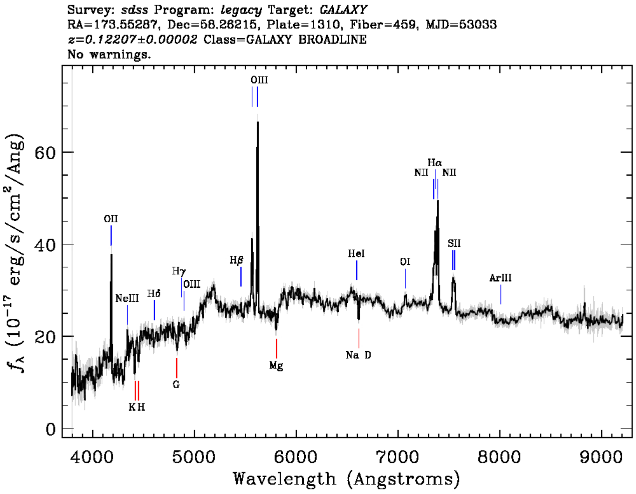

Figure 16g: An outlier's spectrum – very wide emission lines.

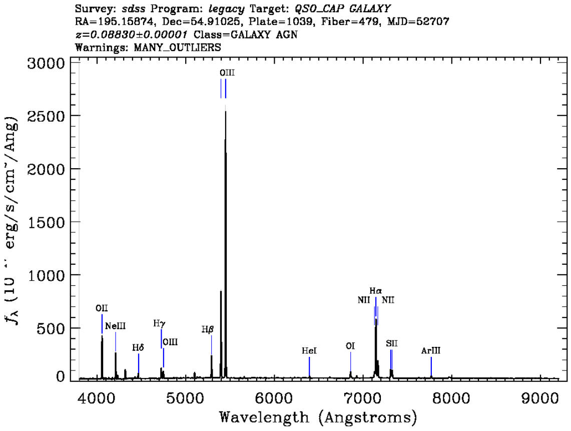

Figure 16h: An outlier's spectrum – unrecognized emission & absorption spectral lines are present.

7 FINDING SIMILAR OBJECTS

When researching an object, often we find it interesting to ask ourselves whether there are similar objects, and if there are – how many? One might consider approaching those questions analytically, determining what are the exact features we would like other objects to have such that we would call them 'similar', quantify them, and find objects with the most similar features. In our case, by calculating the dissimilarity matrix we obtained a generic and unsupervised method that allows us to compare every pair of objects, thus by taking the objects with the lowest dissimilarity to a particular object we would instantaneously find the objects most similar; its nearest neighbors. The exact features that these objects share aren't straight-forward and depend highly on complex processes, such as the way the input data was constructed (see section 2) and the algorithm itself (see section 3).

Moreover, as presented in section 5, we expect objects with similar features to be located at similar coordinates on the low-dimensional t-SNE map. Small clusters that share similar characteristics, but lack most of the features that are mutual to most of the sample, are expected to be tightly clustered and separated from the rest of the sample on the t-SNE map.

7.1 Supernovas

If we would like to find another supernova, similar to the outlier shown in figure 16a, we would just take its nearest neighbor:

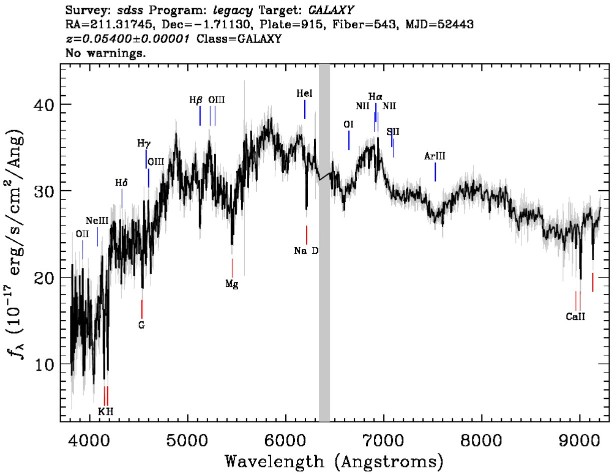

Figure 17: The spectrum of the nearest neighbor to the outlier depicted in figure 16a.

As expected, this nearest neighbor is indeed another supernova, with a very similar spectrum. In fact, at least 7 of its 10 nearest neighbors host a supernova, some of them are reported while others are not. The 50 nearest neighbors can be found at appendix 11.2.1, along with their W_{100} scores.

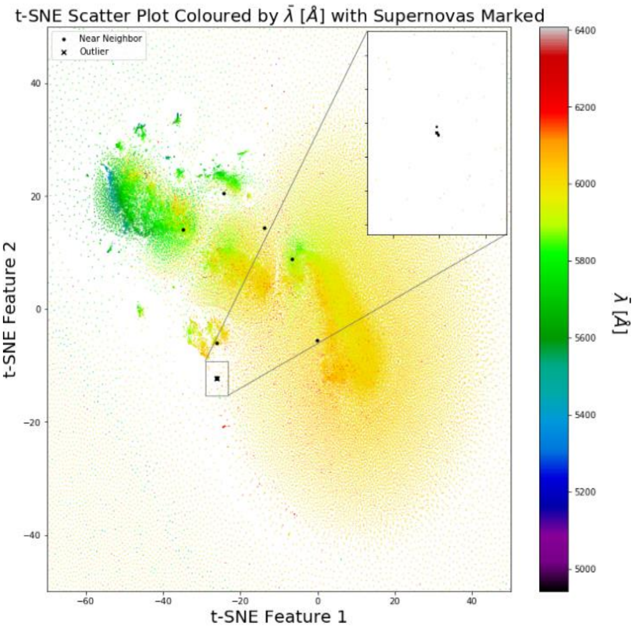

Figure 18: t-SNE map colored by $\bar{\lambda}$ with the outlier from figure 16a and its 10 nearest neighbors marked.

As depicted in figure 12, even though some of the supernovas are spread over the t-SNE maps, 5 of them are tightly clustered and separated from the rest of the sample. It suggests that these 5 supernovas have a lot in common, whereas others have their own, even nearer neighbors.

7.2 Quasars & Blazars

Since quasars and blazars, which are highly distinct from the rest of the galaxies, are falsely present in our data set, we would expect them to be well separated from the rest of the sample.

Similarly to the method described in 7.1, if we would like to find another blazar, similar to the one depicted in figure 13b, we would look at its nearest neighbor:

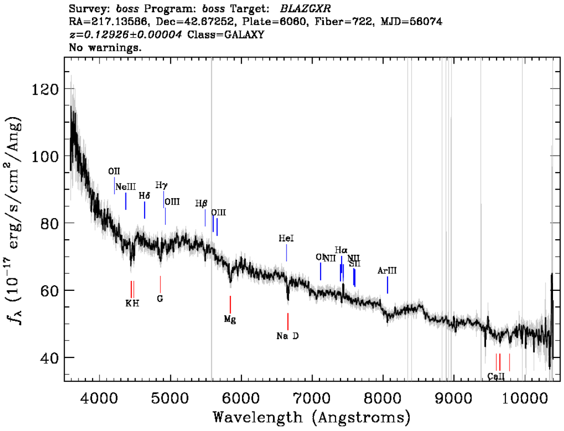

Figure 19: The spectrum of the nearest neighbor to the outlier depicted in figure 13b.

This is indeed a blazar with a similar spectrum.

Let's look at its 30 nearest neighbors:

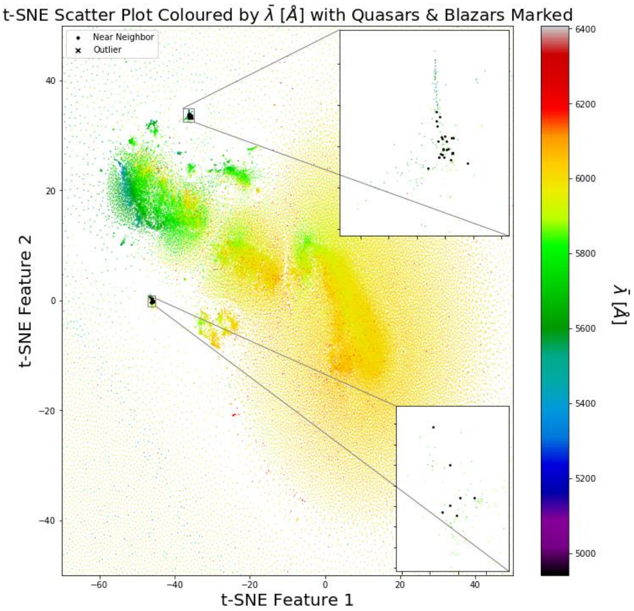

Figure 20: t-SNE map colored by $\bar{\lambda}$ with the outlier from figure 13b and its 30 nearest neighbors marked.

As we see, those objects are nicely clustered at 2 different locations on the t-SNE map, well separated from the rest of the sample. By inspecting their spectra, all of them are active objects, blazars and quasars of different kinds. It is reasonable to assume that the two clusters, despite their strong similarity, have some different characteristics. The 50 nearest neighbors can be found at appendix 11.2.2, along with their W_{100} scores.

7.3 White Dwarfs (WD)

WD are also falsely present in our data set. Similar to the quasars and blazars, we would expect them to be well separated from the rest of the data.

Let's look at the 20 nearest neighbors of the WD depicted in figure 14:

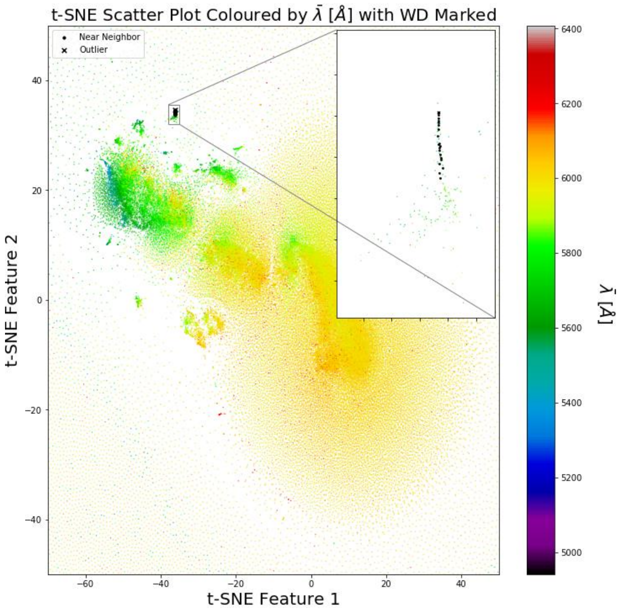

Figure 21: t-SNE map colored by $\bar{\lambda}$ with the outlier from figure 14 and its 20 nearest neighbors marked.

All of its neighbors are located at the same area, very close to the quasars depicted in figure 20. By inspecting their spectra, most of those objects are indeed WD, whereas some are quasars. It is suggested that despite the fundamental difference between WD and quasars, WD's spectral lines have much more in common with active objects, such as quasars and blazars, than with other galaxies. This is also shown by the fact that many of those WD were targeted mistakenly as quasars by the SDSS. The 50 nearest neighbors can be found at appendix 11.2.3, along with their W_{100} scores.

8 SUGGESTIONS FOR FUTURE RESEARCH

8.1 Noise

Despite the great importance of uncertainties in physical measurements, especially in astrophysics where the Signal to Noise (SNR) is relatively low, the algorithm presented doesn't take them into account. It makes since by the method in which the trees are grown in the unsupervised RF algorithm, such that a threshold is set of a specific feature to move measurements to the left or right child node, such that an object with a high uncertainty at this feature would be nearly randomly moved to one of the child nodes. This way, it might be separated from similar objects. As a result, we would expect noisier objects to have a lower similarity to similar objects, as they will end up at the same leaf less often.

In fact, we have already witnessed one such effect, such that, as presented in section 6, about 20% of the top outliers are a result of a bad sky emissions subtractions, and an additional 15% are a result of missing a band of flux values. This is a significant effect, which is part of the very high uncertainties present in our data. It is reasonable to assume that smaller uncertainties also cause a distortion in our results, namely distorting the dissimilarity matrix.

A clear influence can be seen by inspecting the correlation between the median SNR and W_{100}:

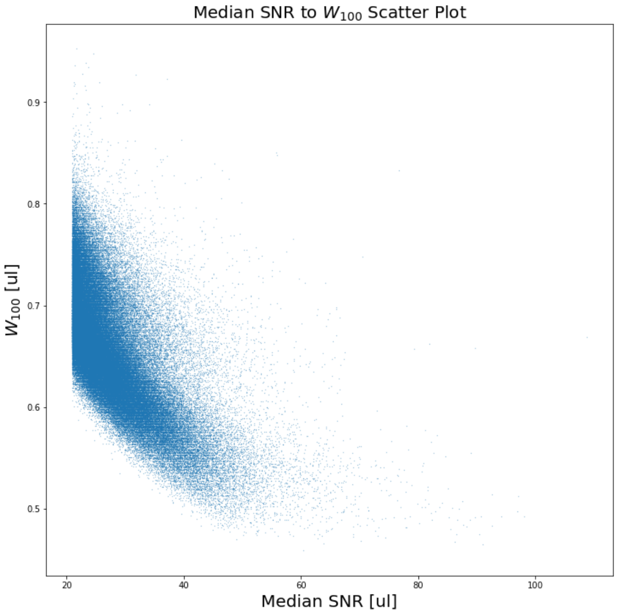

Figure 22: Median SNR to W_{100} scatter plot.

Figure 13 depicts the phenomena discussed clearly – the noisier the object is, the weirder it is, and vice versa.

Another influence is seen by comparing the median SNR distribution of the entire sample to that of the top 10,000 outliers:

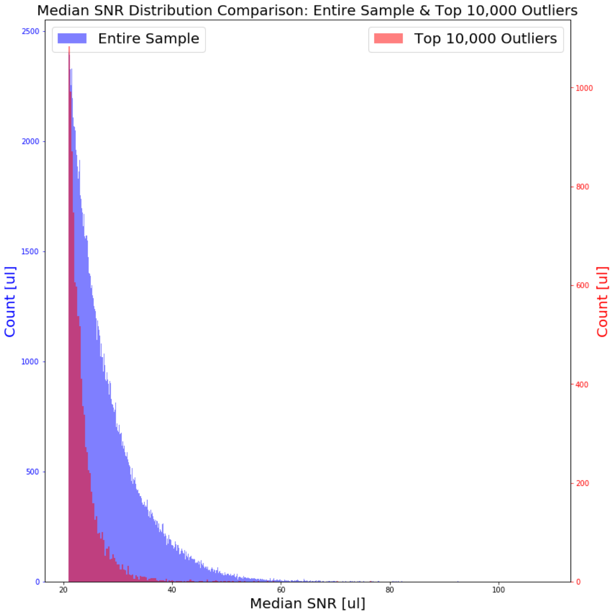

Figure 23: Median SNR Distribution Comparison: Entire Sample & Top 10,000 Outliers.

As seen, the top outliers distribution tends more to the low median SNR, suggesting the same phenomena. Quantifying it, the mean median SNR for the entire sample is ~ 27.96, whereas among the top 10,000 outliers it is ~ 23.65. This phenomena is highly undesired, as we would like the dissimilarity to be determined by the physical properties of the galaxies and not by our ability to measure their flux accurately. A few possible solutions:

- An easy and straightforward solution: removing the weirdness scores to median SNR correlation depicted in figure 22. Although some distortion would be fixed, this is far from an ideal solution as it doesn't into account uncertainties of specific features, but only the median SNR, meaning that the distances will still be distorted.
- Modifying the algorithm to take into account uncertainties, for example, by moving galaxies with an uncertainty larger than a certain value into both child nodes upon growing the trees.
- Weighting the degree of similarity for two galaxies that ended up at the same leaf by the cumulative uncertainties of the features chosen in the path for their leaf.

It is worth noting that by the nature of the median SNR distribution of the entire sample, as depicted in figure 22, the variance in the median SNR within the sample declines if a sample of a similar number of objects with a smaller median SNR is taken (instead of a sample of the objects with the top median SNR), as there are more objects with a low median SNR, than with a high one. Therefore, even though these results aren't presented in this document, we already observed that the SNR influence declines as well for low-SNR samples.

8.2 Synthesizing the Synthetic Data

By the way in which the synthetic data is synthesized (see 3.2), for roughly homogenous data sets, like ours, we would expect the synthetic data to be similar to the common objects in the sample. For example, for a data set with 90% objects of type 1, and 10% objects of type 2, when drawing a value for a feature for a synthetic observation, it would be 90% taken from a type 1 observation. Hence, roughly 90% of its features would be of type 1 and 10% of type 2, resembling mostly a type 1 object. Therefore, upon growing the trees, we expect type 2 objects to be easily separable from the synthetic data, as they are very different. As a result, those objects would end up at leafs at shallower trees depths. This phenomena was verified by inspecting the leafs' depths distributions of different types of galaxies. A possible consequence is that galaxies with a smaller representation in the data set might actually have a higher similarity among them (they end up at smaller leaf depths $\Longrightarrow$ separated less times based on their flux values $\Longrightarrow$ have a higher chance to end up at the same leaf). Moreover, since they are separated less times based on their flux values, their true similarity is actually lower – i.e. the algorithm doesn't 'focus' on the characteristics of less common observations. Furthermore, it further contributed to non-linearity of the dissimilarities. On the bottom line, an undesired phenomena is presented – it is plausible that observations that are less common in the data set that share less mutual characteristics with each other, would be considered as more similar than common objects that share more mutual characteristics.

Possible solutions:

- Synthesize new synthetic data recursively after a determined number of splits at the trees' nodes – this way the algorithm would 'focus' on different kinds of objects, as they are moved to different child nodes.

- Perform a preliminary clustering for the data and synthesizing synthetic data within each cluster. The clustering method could be itself based on the dissimilarity matrix obtained from a preliminary run of the unsupervised RF.

8.3 Similarity Measures

Based on the RF's output, the decision trees, different similarity measurements can be determined. In the implementation presented, counting the number of leafs in which two observations were classified as real and ended up together was taken as the measure for similarity. This similarity is binary in the sense that for each decision tree, the similarity between two observations (L_{ij}^{k} ; see section 3.2) is either 0 or 1. Yet, for example, it is reasonable to treat observations that were separated at the last split, right before a leaf node, as more similar than observations that were separated at the root. Similarly, observations that ended up at deeper leafs are likely to be more similar than observations that ended up at shallower ones. Therefore, it is suggested to try and define more accurate measurements for similarity, for example:

- Use a multi-level similarity measurement for each tree: Define the similarity as the depth of the last node that two observations were separated, divided by the average depths of the leafs that the two observations ended up at. This would result in a value of 1 for observations that ended up at the same leaf, a value of 0 for observations that were separated at the root (defining the root as depth 0), and any value in between for observations that were separated somewhere along the way. This would also result in less quantized distances, as many more levels of similarity would be created.
- Take into account the depth of the leaf at which two observations ended, by weighting each similarity by the depth of the node that determines their similarity. For the above example, it can be done by simply not dividing by the average depths of the leafs that the two observations ended up at. It can potentially solve some of the distortion described in section 8.2, where under-represented observations end up at shallower leafs.

8.4 Data Selection

By preparing the input data for the algorithm (see section 2), a number of observations were discarded: objects at a very low redshift, or objects that had too many missing values. Yet, by analyzing the outliers, it is suggested that many of them were a result of inaccurate measurements or pipeline failures. Those outliers perhaps could have been avoided by further discarding those types of observations upfront. The trade-off between discarding objects and keeping them in the sample is clear: mistakenly discarding interesting observations on the one hand, or having failed observations distorting the results on the other hand. Therefore, a balance must be found – discarding only a small portion of the sample efficiently and based on as accurate measures as possible. A few suggestions:

- Further discarding objects at a low red shift. Objects with $z \leq 0.01$ were discarded. It seems that this is not enough, as even objects with a slightly higher redshift are often galaxies were only the center of the galaxy's spectrum was measured, hence resulting in observations that are fundamentally different. As seen in the redshift distribution (figure 1), higher redshifts can be further discarded without losing a significant portion of the sample. A minimum redshift of $z = 0.03$ is suggested.
- Discarding objects with a missing band of flux values.
- Using a shorter band of wavelengths grid in order to have less missing values introduced at the ends of the spectrum.
- Bad sky subtraction:
    - Detecting and discarding objects with a bad sky flux subtraction. One simple possible solution would be using the fact that a bad sky flux subtraction results in a very sharp change in flux values. Therefore, by calculating the numeric derivative for each object, which is proportional to difference between two adjacent flux values, objects with a high number of derivatives over a certain threshold can be discarded.
    - Further smoothing the spectrum by using a larger median window (at the expense of losing some true thin spectral lines).

8.5 Further Analysis

As presented along the entire document, the dissimilarity matrix holds a vast high-dimensional information regarding both macroscopic and microscopic phenomena. The presented analysis is only the tip of the iceberg, as there are endless possible analyzations one can think of in order to answer a variety of scientific questions. We present a few key ideas for further analysis:

- More observations: Only the SDSS data set contains the spectra of over 4 million objects of different types, whereas only $150{,}000$ galaxies were studied in this research.
- Detecting small clusters of galaxies:
    - Quantifying small clusters based on the dissimilarity matrix by a variety of measures. For example, objects with a high change in weirdness score when comparing W_{10} and W_{100}, namely $W_{100} - W_{10}$, are likely to represent small clusters of $10 - 100$ objects.
    - Detecting small clusters on the t-SNE map and analyzing their physical properties.
- Galaxy evolution: We expect objects on similar locations on the t-SNE map to be similar, whereas further objects to be less similar. By drawing a line that connects two distant locations on the t-SNE map and inspecting the objects along it, we expect to have a continuous change in the galaxies' spectra. Different such lines might represent different galaxies' evolutions.
- Classification: The dissimilarity matrix can be used as a distance matrix for any distance-matrix-based

classification algorithm desired in order to classify the data set.

- Implementing as a tool for researchers, providing them the ability to instantaneously discover similar objects to an examined one.

9 CONCLUSIONS AND DISCUSSION

An unsupervised RF variation was implemented over 150,000 galaxies' spectra, which was manipulated especially to correspond with the algorithm, focusing on the spectral lines characteristics. By doing so, we obtained a similarity measure for every pair of galaxies in our data set, providing us a vast amount of high-dimensional data which allows us to reveal specific outlying objects, as well as underlying macroscopic structures. We successfully detected objects with a variety of unusual physical phenomena and others like them, detected failed spectra measurements and failed data processing, and visualized the data set on a 2-D map. The unusual phenomena pinpointed include galaxies hosting a supernova while their spectrum was measured, two merging galaxies, unrecognized spectral lines, unrecognized structures in the spectrum, and extremely wide and strong emission lines. Those physical phenomena should be further examined and understood.

The richness of information available in the dissimilarity matrix is far from being completely analyzed, and should be further extracted by utilizing it intelligently in order to provide answers for new and diverse scientific questions. Moreover, the algorithm should be applied for larger data sets, providing a higher probability to obtain quality conclusions.

As true for many ML algorithms, most the methods used are generic and suitable for many types of domains and data types, such as tabular data, time-series, imaging, etc. The main computational time bottle-neck is computing the dissimilarity matrix at $O(n_{galaxies}^2 n_{trees})$, but it should only be calculated once. Most of its analysis, such as calculating weirdness scores, can be done at $O(n)$. Furthermore, since only a small number of features are drawn for each node upon growing the trees, it is compatible for working with a large number of features. Yet, the algorithm's accuracy should be further improved, both as a generic ML algorithm (synthesizing the synthetic data and a more accurate dissimilarity measure), and both in order to perform better specifically over spectra (especially accounting for noise).

10 REFERENCES

[1] D. Poznanski, D. Baron, arXiv:1611.07526

[2] D. Schlegel, D. Finkbeiner, M. Davis, 10.1086/305772

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

11 APPENDIX

11.1 Top 150 Outliers

Rank	plate	mjd	fiber	W100	Rank	plate	mjd	fiber	W100
1	1256	52902	21	0.9521	76	647	52553	625	0.8550
2	568	52254	240	0.9478	77	7166	56602	186	0.8550
3	7261	56603	339	0.9385	78	3182	54828	148	0.8550
4	969	52442	447	0.9360	79	2644	54210	318	0.8550
5	1802	53885	278	0.9341	80	1721	53857	76	0.8545
6	2032	53815	240	0.9277	81	5021	55863	878	0.8545
7	2763	54507	279	0.9268	82	4788	55889	822	0.8530
8	6402	56334	520	0.9224	83	2481	54086	278	0.8530
9	895	52581	279	0.9194	84	1939	53389	372	0.8525
10	1257	52944	314	0.9189	85	313	51673	529	0.8521
11	7417	56753	720	0.9175	86	6138	56598	763	0.8511
12	2089	53498	522	0.9160	87	2118	53820	240	0.8506
13	637	52174	524	0.9160	88	266	51630	94	0.8506
14	419	51879	279	0.9116	89	2887	54495	538	0.8506
15	745	52258	280	0.9082	90	6028	56102	322	0.8491
16	5947	56093	875	0.9033	91	1179	52637	129	0.8486
17	1325	52762	240	0.8984	92	2523	54572	135	0.8486
18	555	52266	538	0.8979	93	2948	54553	1	0.8486
19	1816	53919	270	0.8975	94	1043	52465	640	0.8481
20	2784	54529	60	0.8975	95	1179	52637	280	0.8481
21	687	52518	344	0.8960	96	2887	54495	411	0.8477
22	1581	53149	470	0.8940	97	1732	53501	598	0.8477
23	2306	53726	336	0.8940	98	2027	53433	548	0.8472
24	1257	52944	319	0.8916	99	4667	55868	760	0.8472
25	813	52354	524	0.8901	100	2886	54498	228	0.8472
26	1586	52945	278	0.8896	101	1424	52912	515	0.8467
27	1617	53112	240	0.8882	102	1843	53816	612	0.8467
28	7883	57331	834	0.8877	103	767	52252	314	0.8467
29	676	52174	420	0.8862	104	1181	53358	280	0.8467
30	1257	52944	305	0.8848	105	1310	53033	459	0.8467
31	2829	54623	15	0.8804	106	7864	56979	733	0.8467
32	968	52412	11	0.8779	107	1928	53327	544	0.8462
33	434	51885	240	0.8770	108	2613	54481	29	0.8462
34	965	52438	22	0.8765	109	2502	54180	240	0.8457
35	826	52295	331	0.8745	110	1543	53738	580	0.8457
36	566	52238	240	0.8735	111	4078	55358	298	0.8452
37	4182	55446	184	0.8726	112	6873	56541	596	0.8452
38	2758	54523	240	0.8711	113	1867	53317	198	0.8452
39	1876	54464	360	0.8706	114	2742	54233	341	0.8447
40	547	51959	224	0.8691	115	1951	53389	522	0.8447
41	7154	56955	96	0.8691	116	1199	52703	279	0.8447
42	5865	56067	723	0.8682	117	359	51821	574	0.8447
43	1273	52993	78	0.8682	118	967	52636	634	0.8442
44	2172	54230	240	0.8677	119	5649	55912	282	0.8442
45	746	52238	317	0.8672	120	2566	54333	23	0.8442

46	6138	56598	314	0.8672	121	2628	54326	171	0.8438
47	5389	55953	564	0.8662	122	6705	56636	40	0.8438
48	1656	53533	612	0.8657	123	3237	54883	608	0.8438
49	5475	56011	524	0.8633	124	2236	53729	565	0.8438
50	2139	53878	381	0.8628	125	788	52338	177	0.8433
51	1618	53116	240	0.8628	126	2218	53816	1	0.8433
52	7166	56602	274	0.8618	127	6316	56483	678	0.8433
53	4373	55811	580	0.8618	128	1039	52707	479	0.8433
54	2204	53877	83	0.8618	129	960	52425	430	0.8428
55	879	52365	409	0.8618	130	1452	53112	383	0.8428
56	307	51663	360	0.8613	131	1036	52582	389	0.8423
57	2974	54592	60	0.8613	132	571	52286	288	0.8418
58	546	52205	596	0.8608	133	480	51989	24	0.8418
59	7130	56568	432	0.8608	134	3770	55234	404	0.8408
60	853	52374	321	0.8608	135	4372	55541	40	0.8408
61	661	52163	280	0.8604	136	328	52282	578	0.8403
62	2226	53819	396	0.8604	137	2222	53799	178	0.8403
63	2016	53799	180	0.8604	138	7835	56986	479	0.8398
64	4397	55921	328	0.8604	139	5656	55940	264	0.8398
65	2131	53819	240	0.8599	140	2199	53556	78	0.8398
66	1391	52817	267	0.8589	141	5031	56209	774	0.8394
67	1728	53228	614	0.8589	142	1529	52930	318	0.8394
68	355	51788	96	0.8579	143	1939	53389	308	0.8394
69	4082	55367	636	0.8579	144	1038	52673	298	0.8394
70	2758	54523	223	0.8569	145	757	52238	604	0.8389
71	450	51908	523	0.8560	146	830	52293	177	0.8389
72	694	52209	608	0.8560	147	1715	54212	280	0.8389
73	1946	53432	32	0.8560	148	883	52430	309	0.8389
74	2569	54234	437	0.8555	149	360	51780	321	0.8384
75	1850	53786	573	0.8555	150	564	52224	240	0.8384

11.2 Nearest Neighbors

11.2.1 Supernova's Nearest Neighbors

The following objects are the 50 nearest neighbors of the supernova depicted in figure 16a:

Rank	plate	mjd	fiber	W100	Rank	plate	mjd	fiber	W100
1	915	52443	543	0.7490	26	1222	52763	358	0.6265
2	1298	52964	304	0.7998	27	2601	54144	561	0.6055
3	2744	54272	561	0.6167	28	783	52325	514	0.5967
4	291	51928	76	0.7280	29	1872	53386	456	0.6069
5	4381	55824	798	0.7700	30	901	52641	267	0.5693
6	1059	52618	144	0.7485	31	879	52365	278	0.5723
7	1803	54152	260	0.6191	32	1749	53357	360	0.5620
8	2419	54139	50	0.6187	33	6494	56363	518	0.5869
9	5007	55710	956	0.6177	34	1358	52994	580	0.5728
10	2768	54265	233	0.6436	35	1179	52637	103	0.7446
11	1985	53431	502	0.5815	36	5298	55979	964	0.6348
12	2575	54085	267	0.5986	37	1390	53142	361	0.5718
13	2519	54570	499	0.5703	38	2744	54272	552	0.5952
14	666	52149	475	0.6851	39	2514	53882	521	0.6113
15	2022	53827	286	0.6982	40	2439	53795	352	0.5444
16	2156	54525	47	0.6372	41	2758	54523	453	0.5786
17	1791	54266	583	0.7896	42	1842	53501	535	0.6069
18	4805	55715	950	0.6060	43	366	52017	632	0.6372
19	783	52325	448	0.7524	44	1590	52974	555	0.6104
20	6039	56099	476	0.5952	45	1179	52637	316	0.7852
21	2238	54205	563	0.5825	46	428	51883	376	0.5747
22	2522	54570	72	0.6299	47	2500	54178	259	0.5918
23	356	51779	408	0.6343	48	1499	53001	181	0.5527
24	592	52025	308	0.5801	49	358	51818	616	0.6411
25	1927	53321	379	0.6040	50	918	52404	583	0.5962

11.2.2 Blazar Nearest Neighbors

The following objects are the 50 nearest neighbors of the blazar depicted in figure 13b:

Rank	plate	mjd	fiber	W100	Rank	plate	mjd	fiber	W100
1	6060	56074	722	0.7266	26	508	52366	245	0.7412
2	954	52405	429	0.6841	27	970	52413	413	0.7886
3	541	51959	428	0.7183	28	2133	53917	629	0.7861
4	2128	53800	523	0.7686	29	1704	53178	611	0.6885
5	2495	54175	380	0.6978	30	1569	53168	380	0.7222
6	6424	56272	880	0.6914	31	2919	54537	616	0.7275
7	453	51915	209	0.7500	32	972	52435	595	0.7549
8	1712	53531	182	0.7715	33	384	51821	55	0.7505
9	5979	56329	790	0.7275	34	5966	56101	512	0.7661
10	5287	55952	868	0.7358	35	394	51812	442	0.7266
11	492	51955	391	0.7275	36	394	51913	453	0.7339
12	1395	52825	150	0.7666	37	4456	55537	532	0.8149
13	907	52373	578	0.7124	38	4709	55720	498	0.7891
14	7873	57307	438	0.7524	39	1325	52762	634	0.6914
15	312	51689	173	0.7710	40	4977	56047	748	0.7510
16	1771	53498	569	0.7896	41	5427	56001	998	0.7930
17	4643	55946	750	0.7417	42	2230	53799	638	0.7622
18	2520	54584	244	0.7671	43	5878	56033	547	0.7158
19	6701	56367	866	0.7642	44	1755	53386	586	0.7939
20	1186	52646	261	0.7358	45	1924	53330	454	0.7910
21	601	52316	226	0.7241	46	429	51820	210	0.7593
22	6044	56090	174	0.7607	47	658	52146	397	0.7842
23	2349	53734	202	0.7495	48	644	52149	93	0.7715
24	2658	54502	164	0.7515	49	1768	53442	169	0.7769
25	2427	53815	365	0.7612	50	329	52056	237	0.7925

11.2.3 WD's Nearest Neighbors

The following objects are the 50 nearest neighbors of the WD depicted in figure 14:

Rank	plate	mjd	fiber	W100	Rank	plate	mjd	fiber	W100
1	2305	54414	115	0.8325	26	3182	54828	148	0.8550
2	2628	54326	171	0.8438	27	6673	56419	400	0.8320

3	2226	53819	396	0.8604	28	4450	55540	404	0.7681
4	1237	52762	621	0.8276	29	1703	53799	108	0.7715
5	712	52199	123	0.7676	30	5699	55953	962	0.8228
6	937	52707	486	0.7588	31	2421	54153	509	0.7695
7	659	52199	101	0.8281	32	1288	52731	335	0.7876
8	1679	53149	616	0.8325	33	286	51999	236	0.8110
9	1805	53875	205	0.8008	34	907	52373	578	0.7124
10	1186	52646	261	0.7358	35	4643	55946	750	0.7417
11	2663	54234	56	0.7437	36	2517	54567	430	0.7778
12	1751	53377	536	0.7993	37	6010	56097	796	0.8101
13	2230	53799	638	0.7622	38	5878	56033	547	0.7158
14	508	52366	245	0.7412	39	954	52405	429	0.6841
15	5352	56269	292	0.8057	40	1939	53389	308	0.8394
16	5979	56329	790	0.7275	41	5287	55952	868	0.7358
17	2658	54502	164	0.7515	42	902	52409	65	0.8071
18	1999	53503	546	0.7705	43	8731	57416	880	0.8369
19	440	51885	113	0.7793	44	1437	53046	517	0.8169
20	1975	53734	50	0.7974	45	6986	56717	892	0.8374
21	782	52320	48	0.7866	46	2761	54534	610	0.7280
22	1325	52762	634	0.6914	47	4654	55659	190	0.8345
23	7688	57360	608	0.7949	48	362	51999	270	0.7744
24	1935	53387	94	0.8071	49	3966	55571	482	0.7622
25	1721	53857	100	0.8262	50	6424	56272	880	0.6914