Happy Accidents

Group Members

Ally Warner, Jonathan Dunstan, Alex Bigelow

Overview

Bob Ross is the inspirational PBS painter that we all know and love from the 1980's and the 1990's. Ross made painting fun and relaxing with his easy going attitude and light hearted vocabulary. Bob Ross didn't believe in mistakes, only happy accidents which encouraged many people to start painting. Ross' purpose with his show "The Joy of Painting" was to spread the love and happiness that comes from painting beautiful landscapes of all kinds.

From these wonderful landscapes, we wanted to learn more about the amazing Bob Ross. Some initial ideas were to study his speech, his choice of colors, the similarities or differences between all of his paintings, and if he changed his style throughout the seasons of his show.

Data Collection

PAINTING ELEMENTS
Our main dataset that we used for this project is a very high-dimensional table created by Walt Hickey [1], available on GitHub [2]. Each column corresponds to elements of Bob Ross' paintings, and the rows correspond to specific TV episodes. With the exception of the title and episode number, cells contain a 1 or a 0, indicating whether the column element exists in the painting produced in the row episode. We used this dataset to find similarities, differences and changes over time of Ross' paintings.

EPISODE TEXT
We also attempted to obtain text transcriptions of Bob speaking as he worked. We first attempted to use video archives [3] of the actual episodes, but in spite of trying three different python libraries, and researching many more, we were not successful. While one low-level technique was able to yield some results [4], Google's servers keep timing out because the audio in each episode was too long.

A greater concern, however, was the terrible quality of transcription. One of our ideas was to use the frequency of specific color names in the text as a proxy for the colors in the painting – we even found CMYK values for the colors that Bob uses [5] for this purpose. However, even if the python timeout problems had been resolved, the transcriptions of color names were unusable. Some are somewhat rare in English (e.g. "Pthalo Blue"), and the results were far too noisy for us to extract any meaningful color space features.

Instead, we tried to obtain the official closed-captioning data – we contacted the local PBS station here in Salt Lake, KUED, and they directed us to the station that filmed "The Joy of Painting". This station lost all of its closed captioning data during a system update for the show. We then contacted the official Bob Ross Website [6] and have yet to receive an answer. In conclusion, we had no choice but to abandon analysis on the text and colors in each painting.

Singular Value Decomposition

Our main dataset only contains zeros and ones, indicating whether the painting in an episode (the rows) had a particular element (the columns, e.g. "Trees"). We considered the singular value decomposition (SVD) of the matrix to get an idea for the richness of the dataset.

If Bob Ross only had a few typical painting types, such as the typical Bob Ross winter scene, and the typical Bob Ross beach scene, we should expect to see only a few high singular values, with the rest close to zero. From the figure below, we can see that the singular values do not immediately drop off which implies that, while he is known for mainly painting landscapes, there is great diversity in the elements of those landscapes. Bob Ross was not a one-type-of-painting kind of guy!

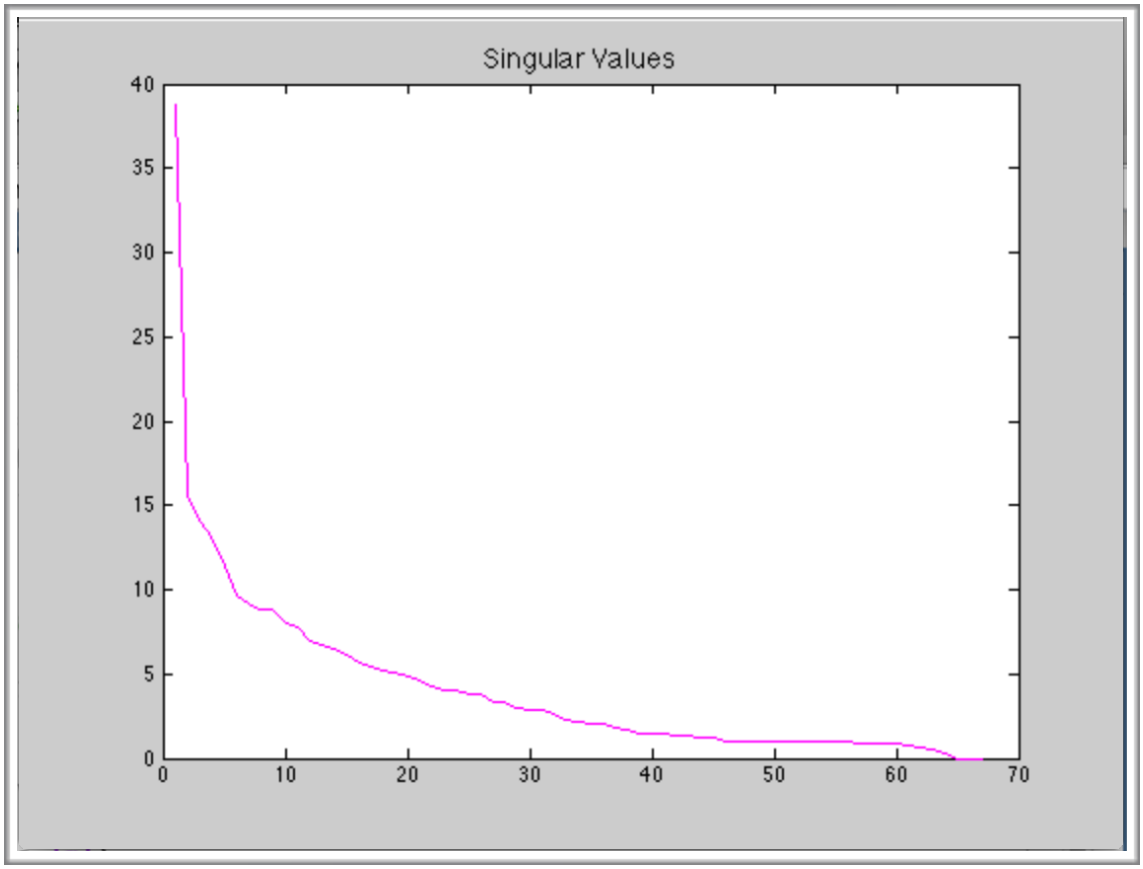


Figure 1: Singular Values

The rest of the Singular Value Decomposition (see Appendix, "Singular Value Decomposition") also reflects the variety of Bob's paintings; each vector in V transpose is fairly meaningless. The first vector with a very high singular value is close to zero for almost every attribute save two; "Wooden Frame" and "People". These features are very rare in Bob's paintings, and the paintings that do have these features are easily identified through them, but none of the remaining paintings are clearly identified by any of the other features. This makes sense since each painting only has a few elements, and there is no predominant element in the majority of paintings.

Distance Metric

This lack of obvious structure in the SVD suggests that clustering Bob's work will be difficult – it is unlikely that it will be possible to find very clean cluster separations, but this does not mean that there is nothing to learn.

To cluster paintings, we are using a variation to the Jaccard similarity to provide a notion of "distance" between any two paintings. This would simply be the size of the intersection of features, divided by the size of the union of features. The clustering results that we provide in the next section are using this metric. This metric has helped find meaningful intersections that will be shown in the next section.

Hierarchical Clustering

We began with hierarchical clustering, using one minus our Jaccard painting similarity as a distance metric. We used single-link, complete-link, and mean-link cluster distance approaches (see Appendix, "Hierarchical Clustering").

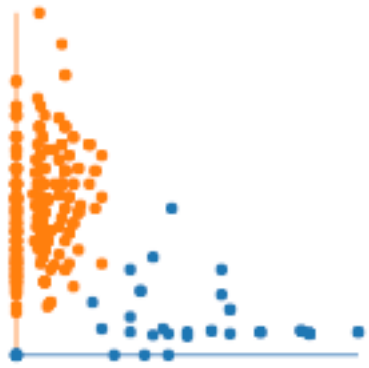
We also clustered the transpose of the matrix, treating each painting element as a data item, and episodes as attributes.

The results show that the mean-link distance metric is most effective; one small cluster of ocean-themed paintings stands out in the dendrogram / heat map visualization. Other structure may exist, but this is less clear.

Center-Based Clustering

With this distance metric, we also applied distance-based clustering approaches, such as Gonzalez and kMeans++ algorithms, that use a specific painting as cluster centers. We felt it would not be appropriate to try to use an approach like true k-Means clustering because the "mean" of two paintings using our distance metric is not well defined.

Instead, we refined both the Gonzalez and kMeans++ clustering results using the k-Medioids algorithm. All results, initial and refined clusters, can be viewed on the final page of the Appendix, and an interactive version that supports hovering over paintings is available online [7].



As we could see from the hierarchical clustering results (Figure 2), there is one clear cluster of ocean-themed paintings, and the other clustering algorithms were able to find this when the number of centers was as low as 2. Even though we tried up to 10 clusters, no other major structure could be observed in either the paintings or the painting elements with the small exception of desert-themed paintings. There was a very small cluster of three desert-themed paintings that the kMeans++ algorithm revealed with nine clusters (see the interactive online plots [7]).

Figure 2. Hiercharchial Clustering Results

Visualization Technique

To validate and compare the results of our clustering algorithms, we extend a visualization technique mentioned in [8].

The Jaccard similarity-based distance function presents a problem for visualization. Euclidean-based coordinate systems can naturally and appropriately be represented using scatterplots that use the native coordinate system. Again, with the Jaccard similarity, there is no coordinate system; only distances.

To get an idea about the tightness and separation of our clusters, we employ a barycentric coordinate-inspired approach [8]. We define a vector for each cluster, representing the cluster center. Then, each point in the dataset is plotted as the sum of all vectors, each weighted by that point's distance to the corresponding center.

For two or three centers, the resulting plots are unambiguous – they use true barycentric distances. For greater dimensions, the ordering of vectors is arbitrary; a different ordering may show a different plot. Future work exploring this ambiguity would be interesting visualization research, and is beyond the scope of this project.

Note that even cluster centers may not lie directly on their respective axes; they can be influenced by other centers if the Jaccard similarity between different centers is nonzero. This is informative, in that it shows some of the uncertainty in the clustering.

Figure 3: Visualization Technique

Temporal Analysis

After doing analysis with each individual episode, we wanted to wanted to look into Ross' paintings over time on a larger scale. So we used a variation of our distance metric previously mentioned on the location index of 1's in the dataset for each season. Figure 4 is a heat map of the Jaccard similarities, darker colors indicate higher similarity, of Ross' paintings by season.

We were excited to discover a shift at season 9 where the seasons become more similar. Ordering the heatmap that we used for hierarchical clustering by episode instead of cluster order (see Appendix, "Sorted by episode"), we can see that Ross started to paint ocean scenes with more regularity. It seems that, while the paintings remained diverse, Bob seems to have settled into a "rhythm" for *when* he painted different scenes. This is a very subtle trend – the most extreme Jaccard similarity values still vary roughly between 0.6 and 0.8 – but it is there.

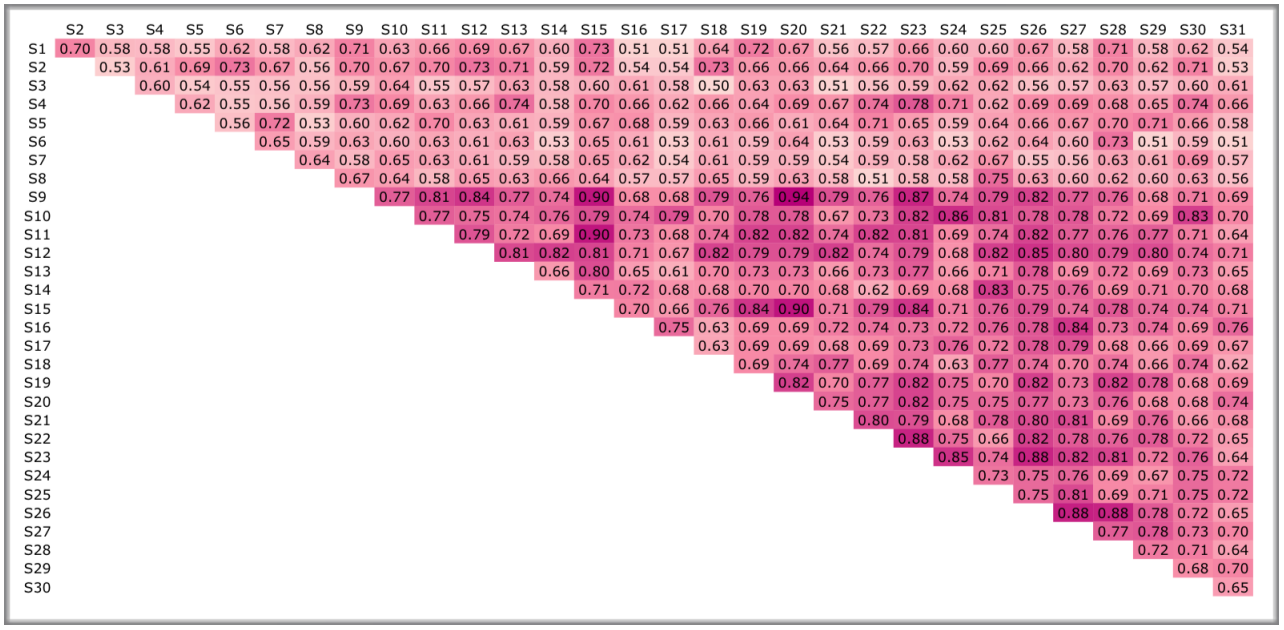


Figure 4: Heatmap of Jaccard Similarities by Season.

Conclusion

Bob Ross has given many people inspiration to get involved with painting and to find it fun and relaxing. After conducting this analysis on Bob Ross' work, we have discovered that Bob Ross is a "hard-to-cluster" kind of artist. Although Ross primarily created landscapes on the show, his work is diverse and full of different kinds of outdoor life. We can see this from every data mining avenue we explored.

We would like to be able to compare Ross with other painters because his work is so diverse within landscapes. We would like to see if other artists can also keep up this kind of diversity over a period of ten years. We would also like to continue the quest of finding the closed captioning data or another text transcription software to better capture what Bob Ross says during his shows. Not only could this give us more insight to Bob Ross himself, but could help with the comparison to other artists.

"In painting, you have unlimited power. You have the ability to move mountains. You can bend rivers. But when I get home, the only thing I have power over is the garbage." - Bob Ross

Appendix

References

[1] http://fivethirtyeight.com/features/a-statistical-analysis-of-the-work-of-bob-ross/

[2] https://github.com/fivethirtyeight/data/blob/master/bob-ross/elements-by-episode.csv

[3] https://archive.org/search.php?query=subject:%22The+Joy+of+Painting%22

[4] http://codeabitwiser.com/2014/09/python-google-speech-api/

[5] http://www.art-paints.com/Paints/Oil/Bob-Ross/

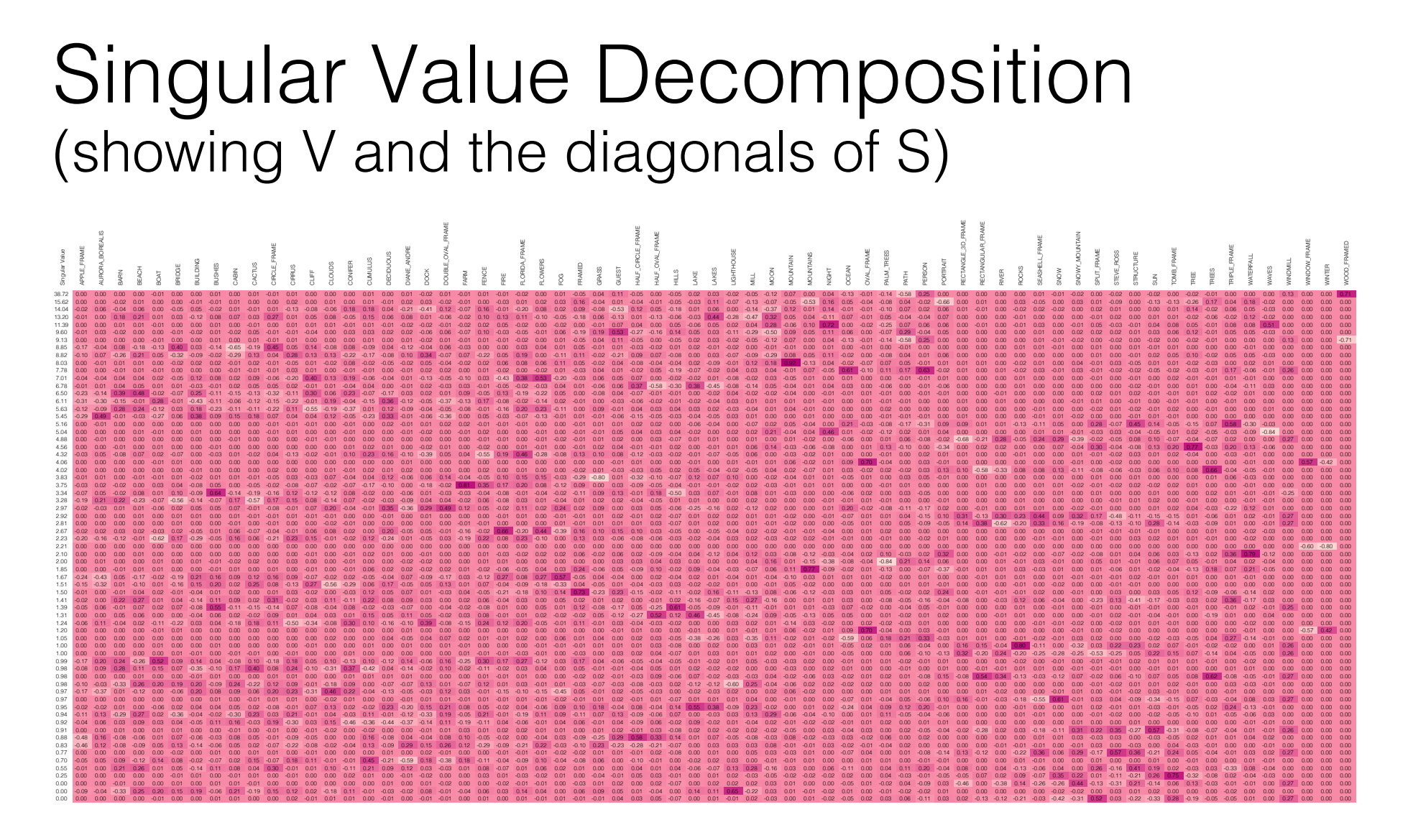
[6] http://www.bobross.com/

[7] https://dl.dropboxusercontent.com/u/64657676/Project/barycentric.html

[8] Charles D. Hansen, Min Chen, Christopher R. Johnson, Arie E. Kaufman, Hans Hagen, *Scientific Visualization: Uncertainty, Multifield, Biomedical, and Scalable Visualization*, Springer-Verlag London, 2014, p.180 - 182

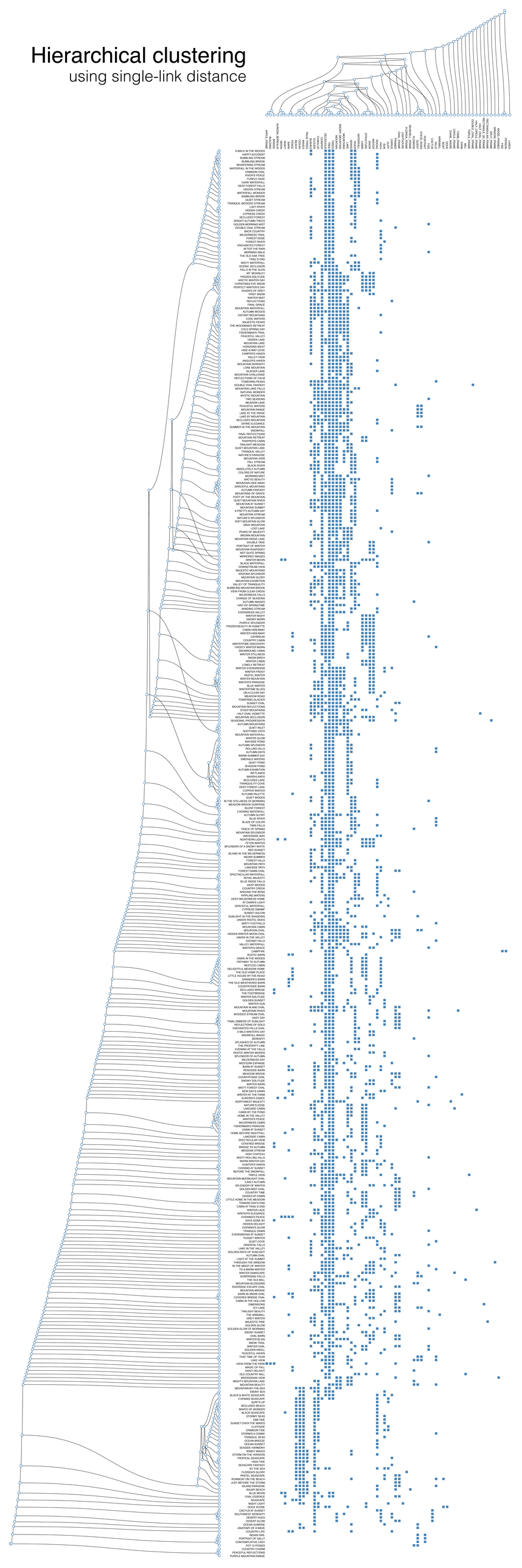
Singular Value Decomposition

(showing V and the diagonals of S)



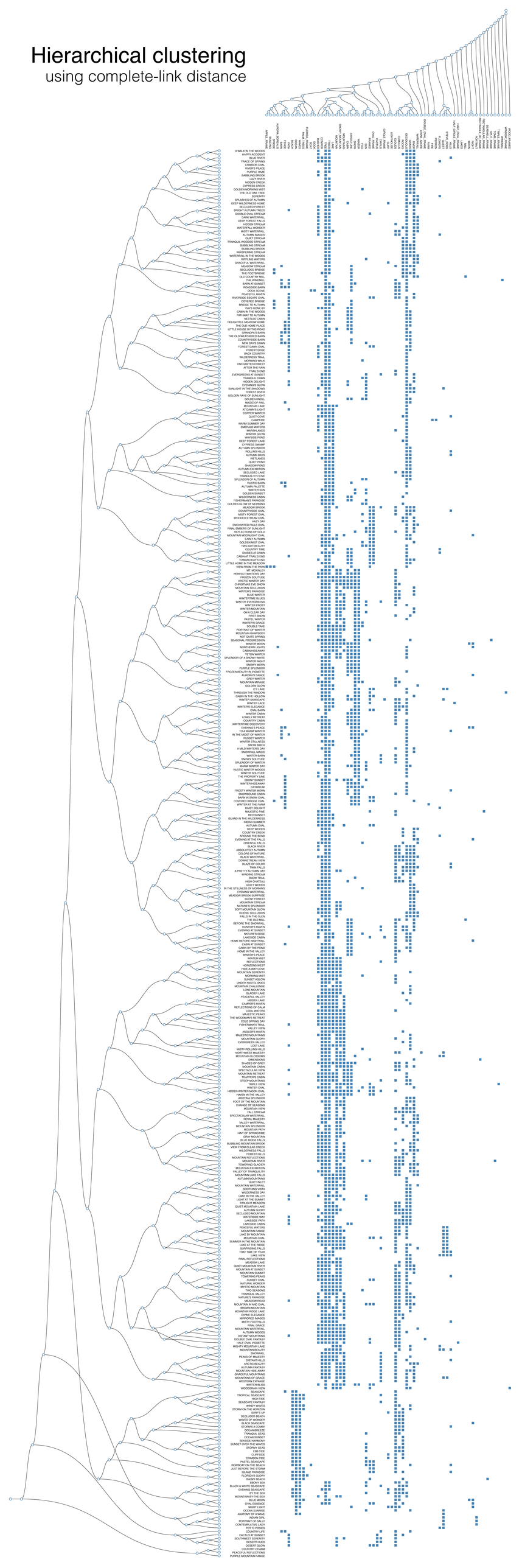
Hierarchical clustering

using single-link distance



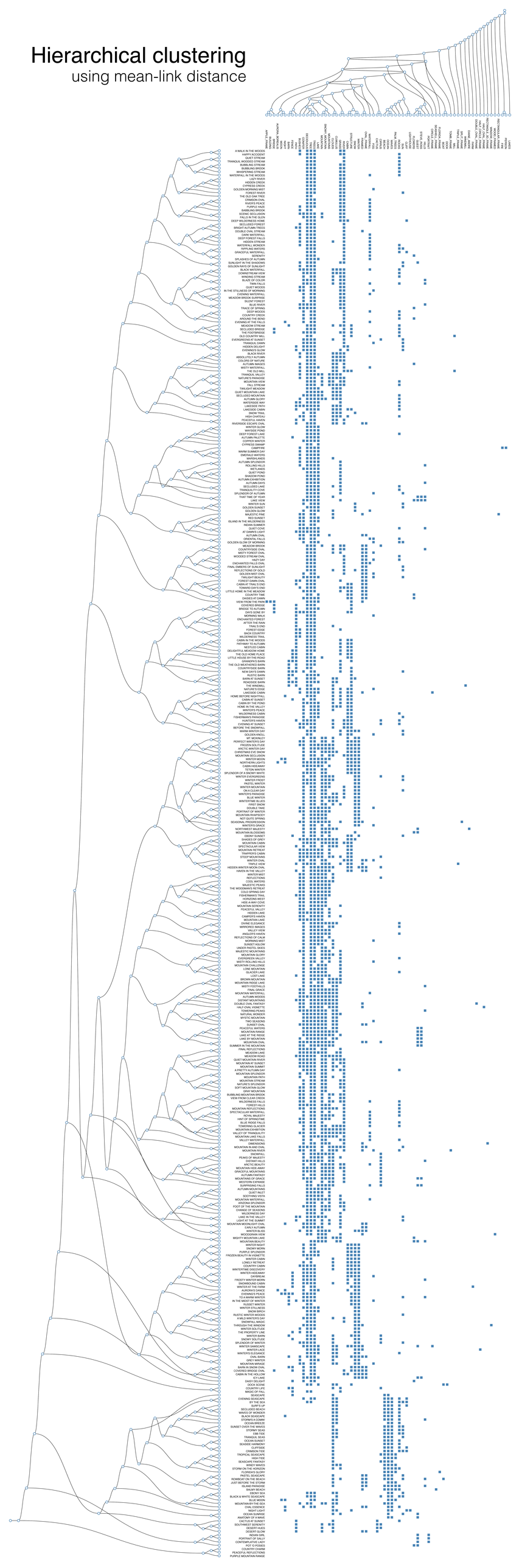
Hierarchical clustering

using complete-link distance



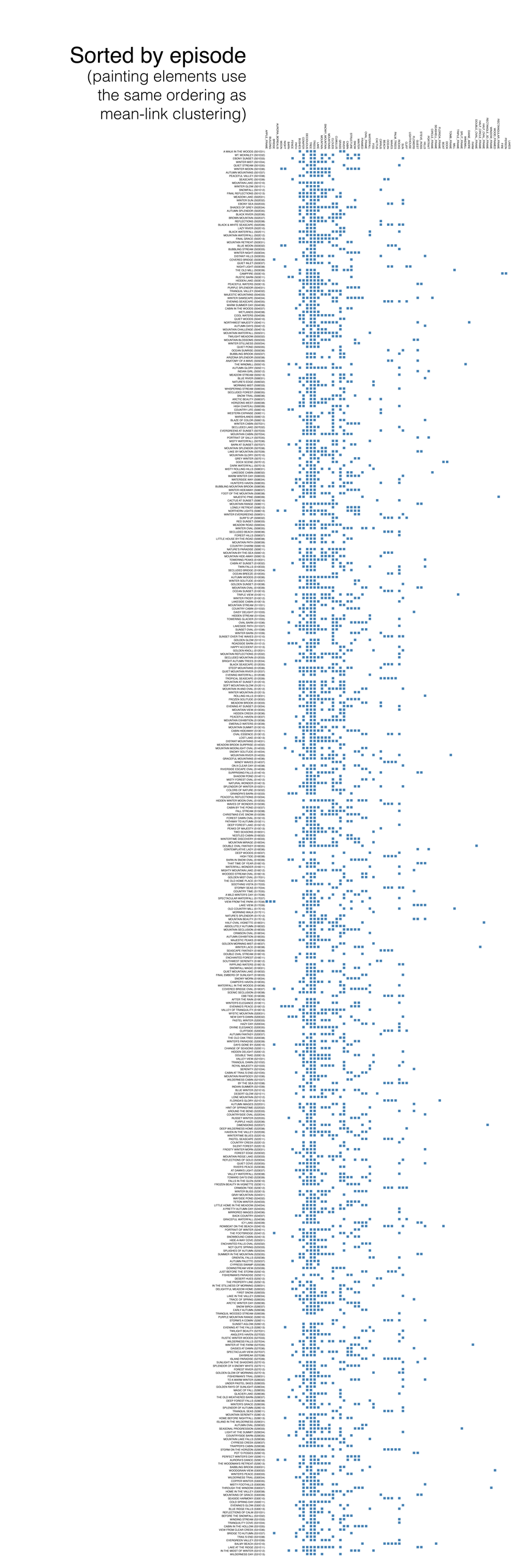
Hierarchical clustering

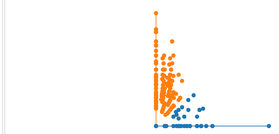
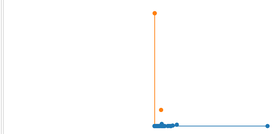
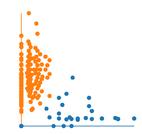
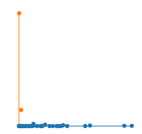
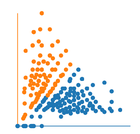
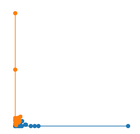
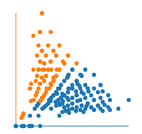
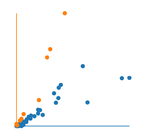
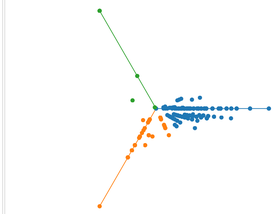
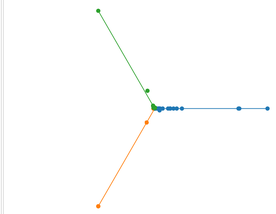
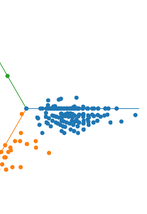
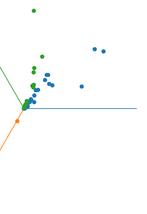
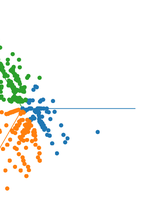
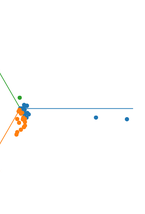
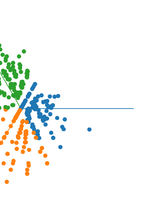
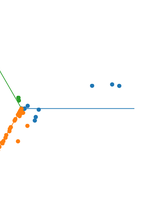
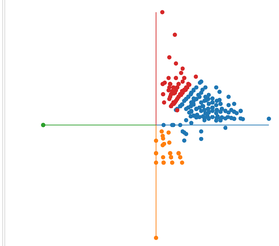
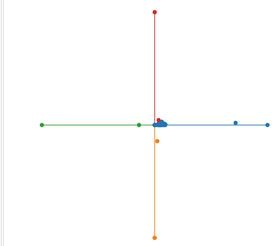
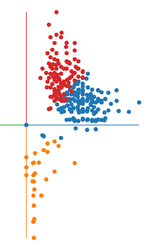
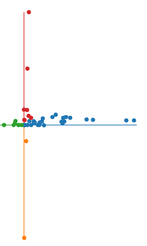
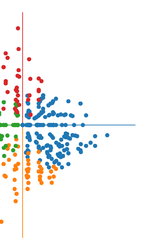
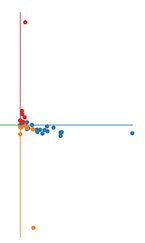
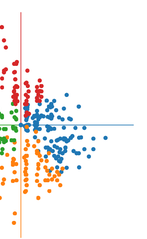
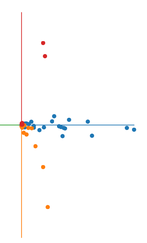
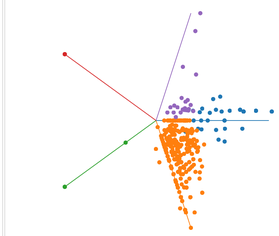
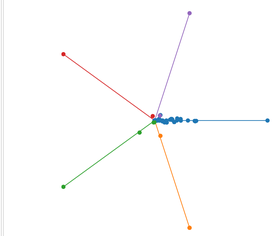
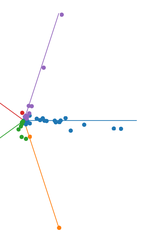
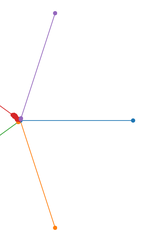
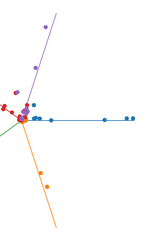
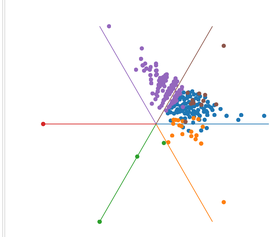
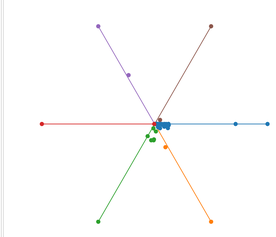
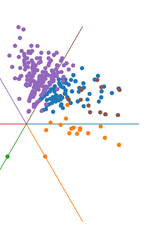
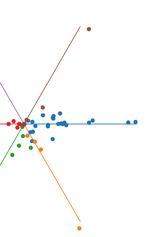
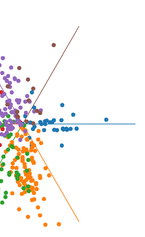
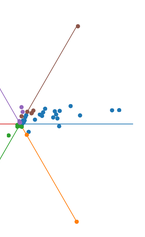
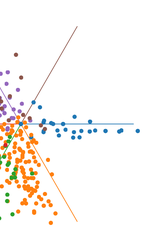
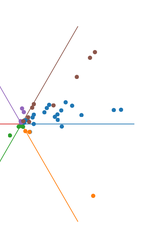
using mean-link distance



Sorted by episode

(painting elements use the same ordering as mean-link clustering)



	Algorithm: gonzalez		Algorithm: gonzalezMedioids		Algorithm: kMeansPlusPlus		Algorithm: kMeansPlusPlusMedioids	
Clusters:	paintings	elements	paintings	elements	paintings	elements	paintings	elements
2	 Total Cost: 258.25	 Total Cost: 61.76	 Total Cost: 229.61	 Total Cost: 56.65	 Total Cost: 229.89	 Total Cost: 62.70	 Total Cost: 218.07	 Total Cost: 57.31
3	 Total Cost: 263.13	 Total Cost: 59.66	 Total Cost: 226.28	 Total Cost: 55.74	 Total Cost: 216.00	 Total Cost: 58.92	 Total Cost: 208.51	 Total Cost: 54.39
4	 Total Cost: 244.98	 Total Cost: 58.85	 Total Cost: 202.46	 Total Cost: 53.08	 Total Cost: 209.61	 Total Cost: 58.21	 Total Cost: 203.53	 Total Cost: 54.49
5	 Total Cost: 238.70	 Total Cost: 56.44	 Total Cost: 208.86	 Total Cost: 52.02	 Total Cost: 204.17	 Total Cost: 60.61	 Total Cost: 201.63	 Total Cost: 51.15
6	 Total Cost: 252.67	 Total Cost: 56.25	 Total Cost: 201.91	 Total Cost: 49.57	 Total Cost: 201.16	 Total Cost: 52.24	 Total Cost: 178.38	 Total Cost: 50.59
7	Total Cost: 265.82	Total Cost: 55.98	Total Cost: 178.65	Total Cost: 49.63	Total Cost: 198.30	Total Cost: 51.81	Total Cost: 181.97	Total Cost: 48.57
8	Total Cost: 244.55	Total Cost: 52.80	Total Cost: 189.87	Total Cost: 47.43	Total Cost: 210.04	Total Cost: 51.56	Total Cost: 180.61	Total Cost: 46.86
9	Total Cost: 216.42	Total Cost: 52.95	Total Cost: 178.34	Total Cost: 46.09	Total Cost: 195.18	Total Cost: 48.71	Total Cost: 164.33	Total Cost: 46.61
10	Total Cost: 218.48	Total Cost: 51.19	Total Cost: 184.25	Total Cost: 45.44	Total Cost: 182.05	Total Cost: 50.16	Total Cost: 162.52	Total Cost: 44.15