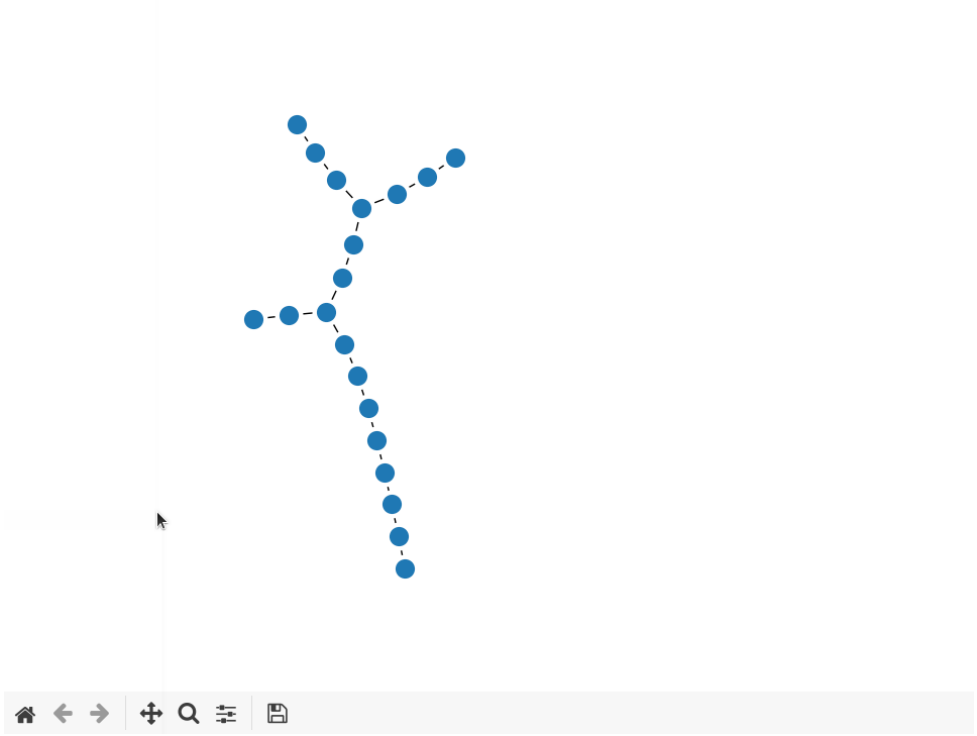
Pranav Rajan
09/15/2021
CS 6965 - Advanced Data Visualization
B. Wang

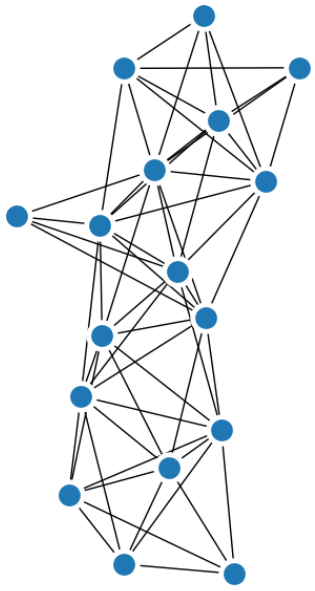
Kepler-Mapper + UMAP

Mapper

Cat Exploration



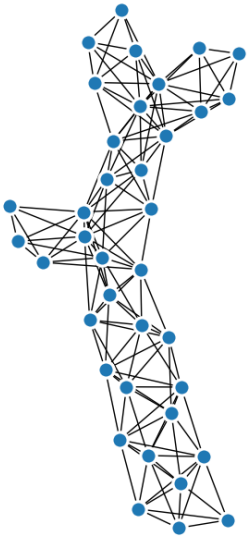
Description: Visualization of the cat data set with overlap parameter = 20%, number of intervals = 15



Description: Visualization of the cat data set with overlap parameter = 80%, number of intervals = 15

Question 1: What is the effect of increasing interval overlap parameter on the final graph in the visualization?

By increasing the overlap parameter, we are increasing the connectivity of the clusters (Professor Wang Lecture 5: Mapper). Comparing the two visualizations with the overlap parameter changed we see that increasing the overlap parameter results in more connections. Based on the Mapper Lecture, this can be interpreted as a higher number of intersections between the different set coverings (represented by the blue points in the visualization) which gives more information about the structure about the data.



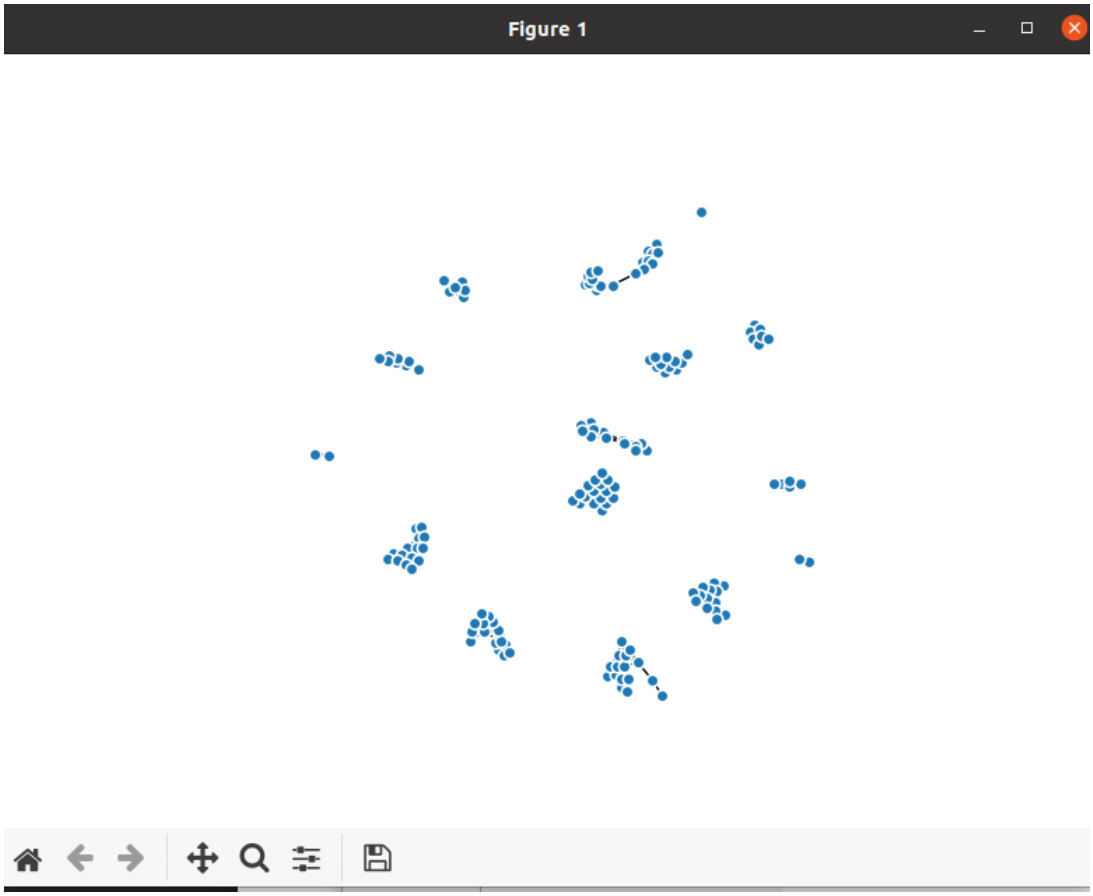
Description: Visualization of the cat data set with 30 intervals, overlap: 80%

Question 2: What is the effect of increasing number of interval parameter on the final graph in the visualization?

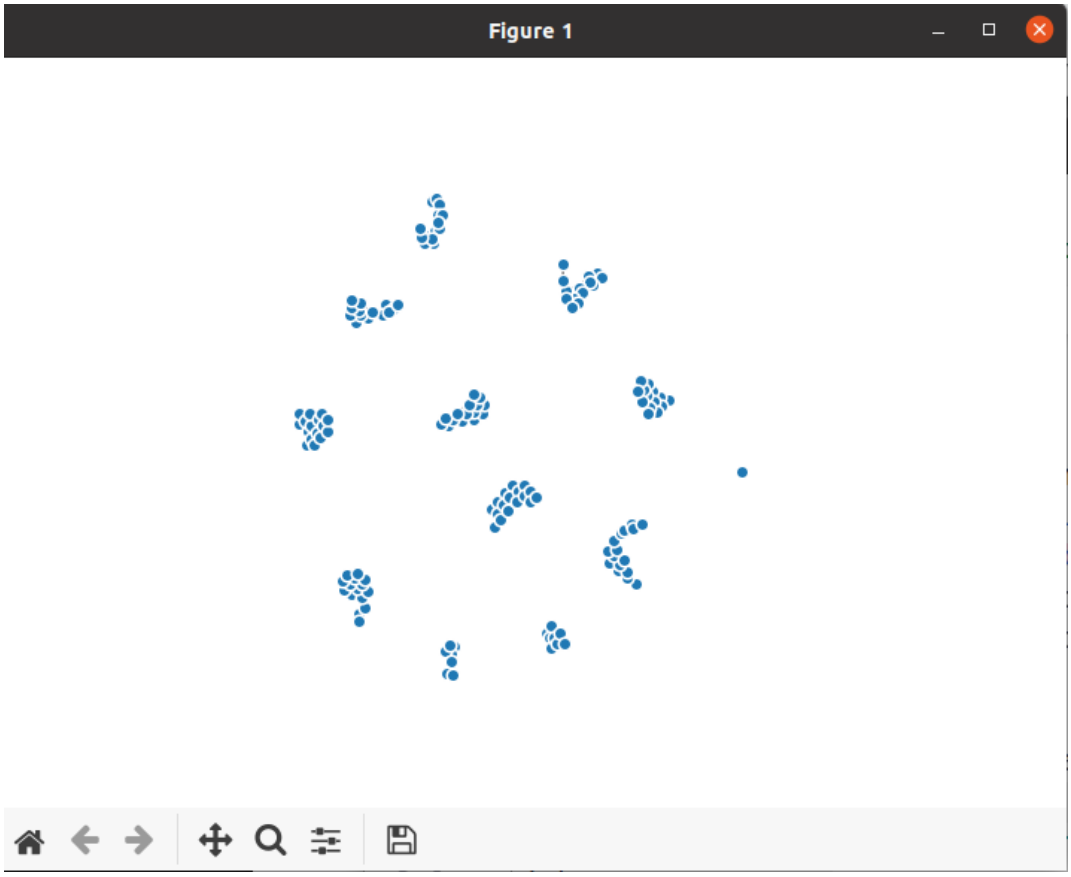
Increasing the number of intervals results in an increase in the number of clusters that we see in the visualization (Professor Wang, Lecture 5: Mapper). Creating more clusters improves the resolution of the data for identifying finer features in the visualization and structure. From a user point of view, the interval parameter can be interpreted as a control for how zoomed in / zoomed out we want to explore the data. Comparing the first visualization of the cat with 15 intervals to the above visualization of the cat with 30 intervals, we have a better image of the structure of the data.

Mapper Digits Exploration

Description: Visualization of the digits data set with a single run using mapper



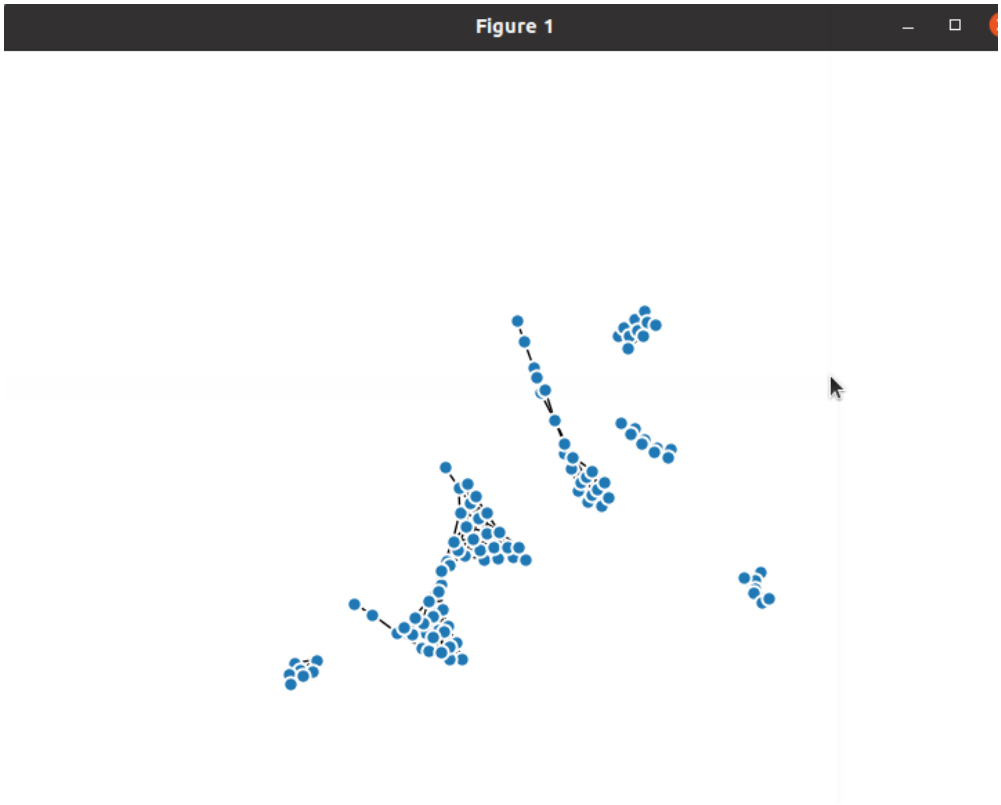
Description: Visualization of the digits data set with a single run using mapper



Description: Visualization of the digits data set with a single run using mapper

Question 3: Why are the results not necessarily identical?

The results of T-SNE with the default parameters will be different every time because of the way T-SNE constructs probability distributions for the pairwise affinities step in the pseudocode algorithm (Professor Wang, Lecture: T-SNE + DR). Each time T-SNE is run, it will generate a probability distribution that can change for each run which affects the result of performing the Kullback-Leibler divergence.

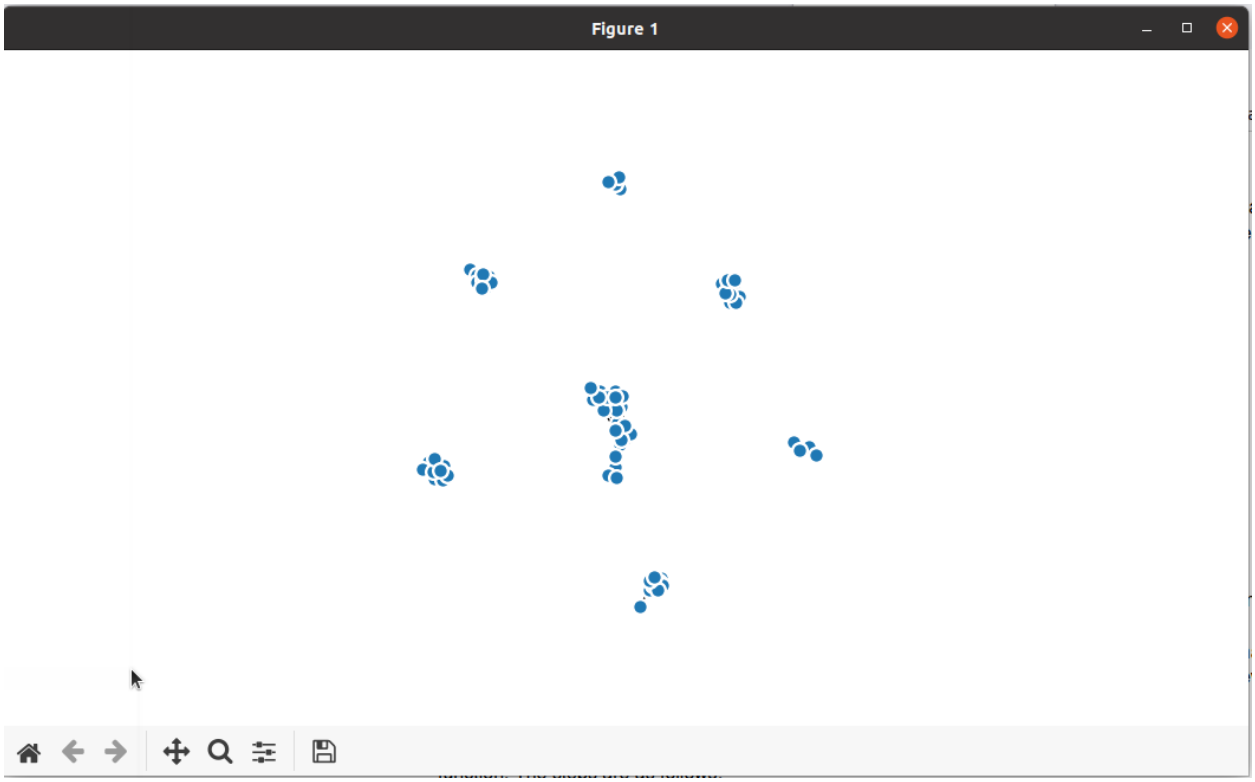


Description: Visualization of digits data using spectral embedding, n_components = 2, random_state = 0, and eigen_solver arpack

Question 4: What is the difference between the results using Spectral Embedding in comparison to the results using T-SNE?

Based on the first 3 visualizations that use T-SNE compared to the visualization that just uses SNE we can see that the clusters are well-defined for T-SNE but ill-defined for SNE. SNE is implemented using a normal distribution whereas T-SNE uses the t-distribution when projecting the data from high-dimensional space to low-dimensional space. From the visualizations we can see that the choice of probability distributions has an effect on how well defined the clusters are.

Resource used to understand T-SNE vs. SNE:
https://www.linkedin.com/pulse/visualization-method-sne-vs-t-sne-implementation-using-tandia

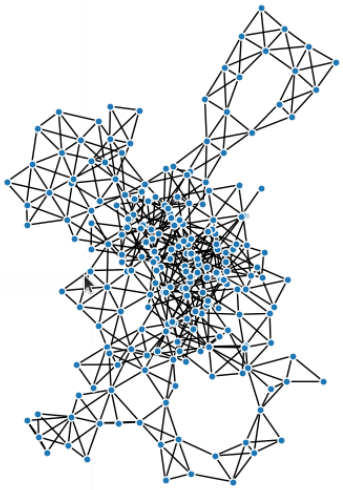


Description: Visualization of digits data using spectral embedding, n_components = 3, random_state = 0, and eigen_solver arpack

Question 5: What is your modification and its effect on the data?

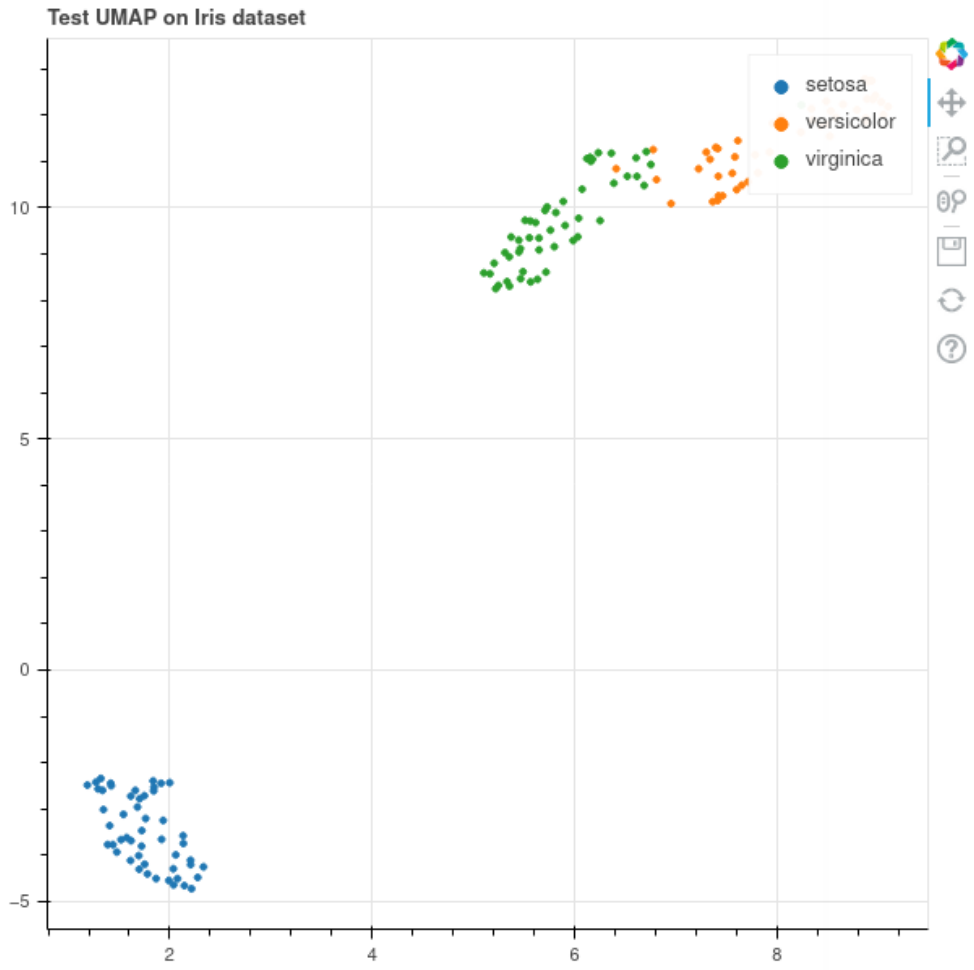
My modification was changing the dimension of the embedded space from 2 to 3. This can be interpreted as changing the projection space from the 2d plane to the 3d plane. Changing the dimension produces better separability but from an interpretation point of view, this seems harder to analyze compared to the 2d plane.

My Dataset with Kepler-Mapper

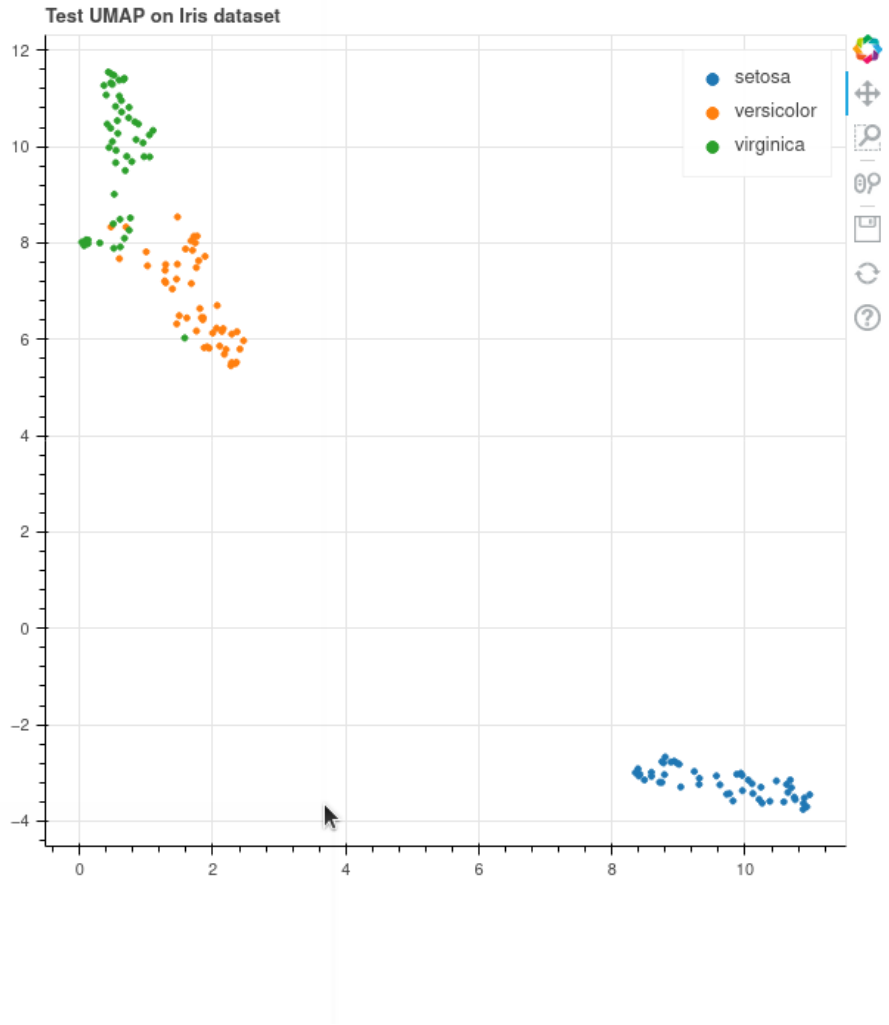


This visualization is of the KDD Cup ‘99 dataset which was created by MIT Lincoln Laboratory in 1998 by processing tcpdump portions of the DARPA IDS Evaluation. The subset of the data used was the SMTP data for the visualization which contains 95,373 samples with 3 dimensions and continuous values. What we can see in this visualization is that there are lots of connections between the cover sets which indicates that a lot of the information shares similar properties.

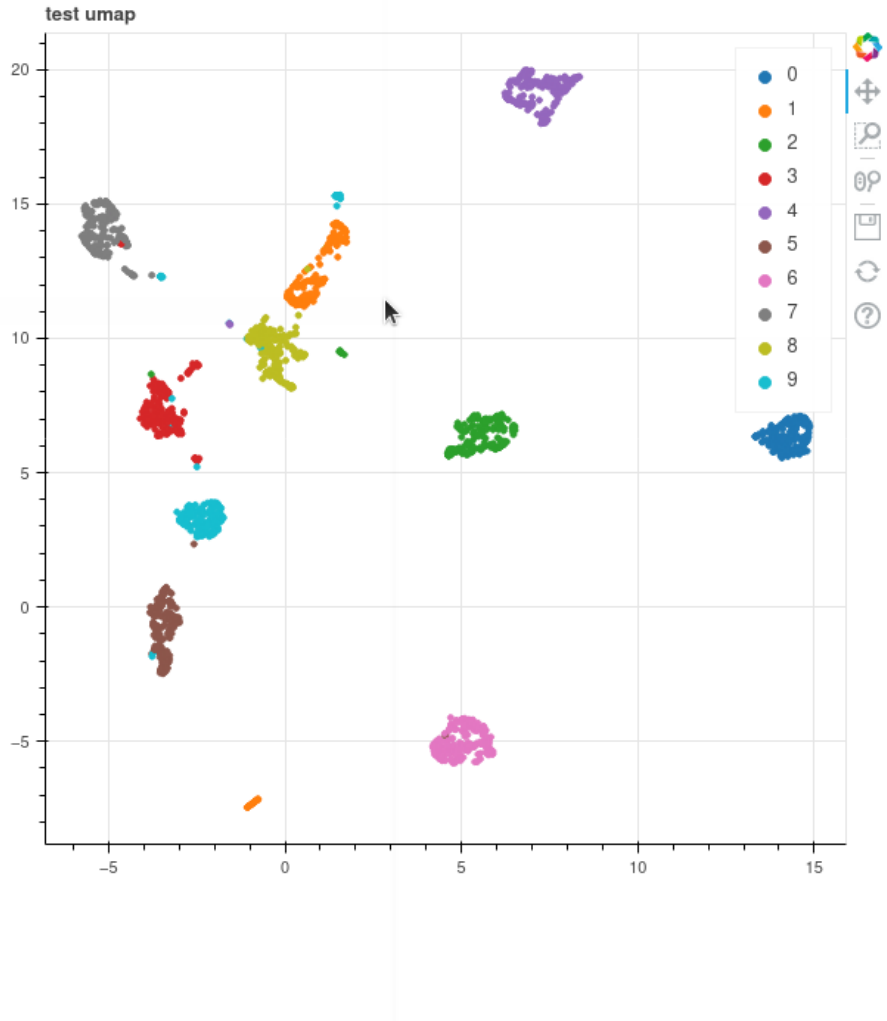
UMAP



Description: Iris dataset with UMAP. Default parameters are used. Number of neighbors = 50

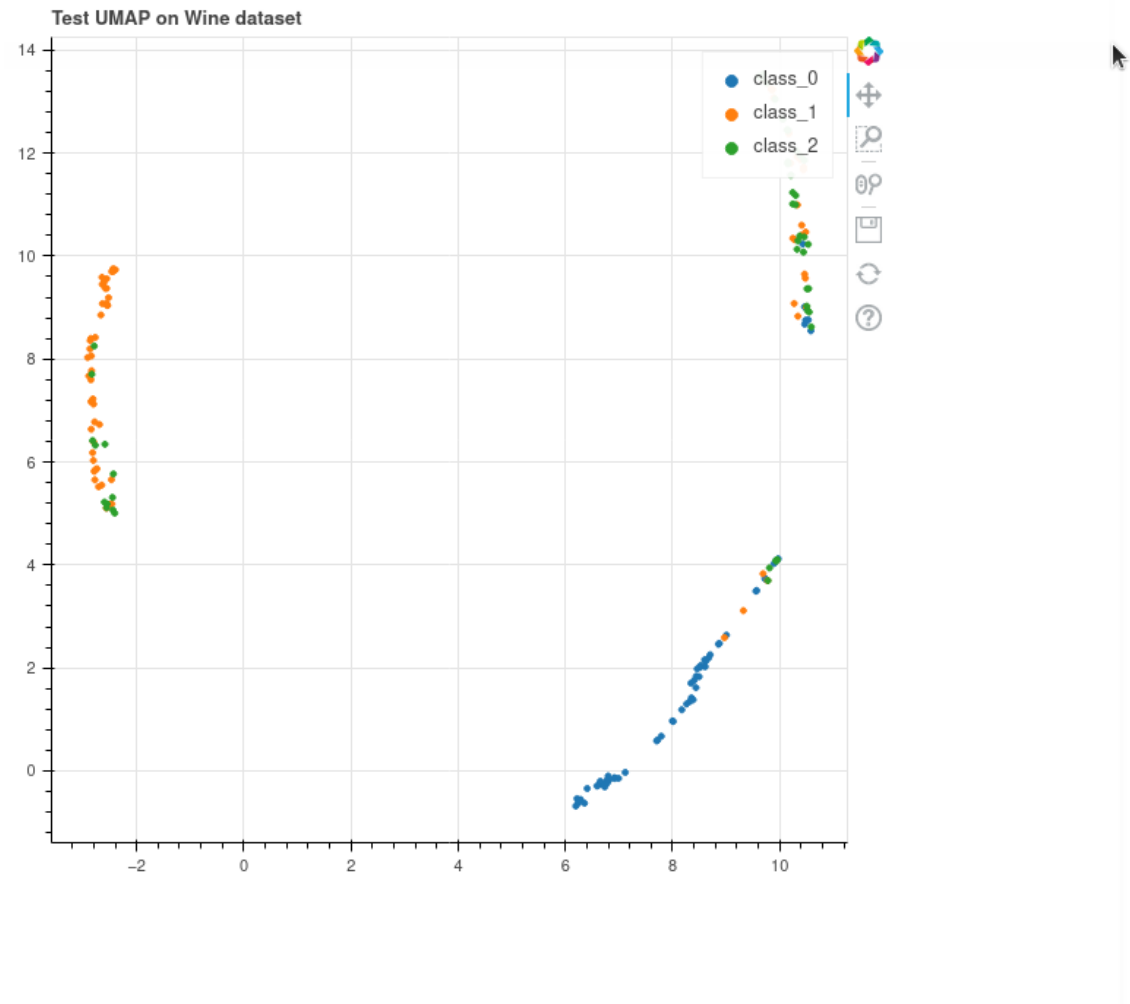


Description: Iris dataset with UMAP. Number of Neighbors = 30



Description: Digits dataset with default parameters.

My Dataset with UMAP



Description: scikit-learn Wine Dataset featuring 13 different classes run with UMAP

Question 1: How different are the results?

According to Adam Coenen and Adam Pearce from Google PAIR, the n_neighbors parameter is the neighbor of approximate nearest neighbors used to construct the HD graph structure. This parameter controls how UMAP focuses on either the local or global structure where a high n_neighbor value produces representing the overall structure but lose fine details. Oppositely a low n_neighbor value focuses on the local structure which may include the finer details. For the iris dataset, this is hard to identify between changing the n_neighbor value since the data seems 2 dimensional. To me, reducing the number of neighbors seemed to produce the same visualization except with a series of image transforms from linear algebra.
Resources used: https://pair-code.github.io/understanding-umap/

Question 2: What's the main difference between UMAP and other non-linear DR techniques such as T-SNE?

UMAP seems better at preserving the overall global structure than T-SNE and other non-linear DR techniques. My understanding of overall global structure is that for large datasets, there is a minimum amount of information lost when the high dimensional dataset is projected down to a lower subspace.

Resources used:
https://towardsdatascience.com/tsne-vs-umap-global-structure-4d8045acba17
https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668
https://towardsdatascience.com/why-umap-is-superior-over-tsne-faa039c28e99
https://pair-code.github.io/understanding-umap/

Question 3: Describe five ideas of combining interactive visualization with UMAP.

1) Exploratory Data Analysis and interaction (using Shneiderman's Principles of interaction). When using Shneiderman's principles such as overview and direct manipulation to explore a dataset, UMAP allows for users to examine vast amounts of data such as the Pixplot project from Yale which would be impossible with traditional methods of viewing photos online at once (ie. google photos, facebook etc)
2) Neural Network Activation Functions - UMAP can be used to understanding how neural networks behave and to visualize what is happening at each layer. Chris Olah, Google and Open AI created the Activation Atlas to understand how modern neural networks process images and how they sometimes produce baffling mis-classifications.
3) NLP - relationships between text - With NLP and machine learning, UMAP can be used to study and visualize the relationships and similarity between texts and create artistic visualization such as Open Syllabus Galaxy.
4) Understanding BERT - Martin Wattenberg, Fernanda Viegas and their colleagues at Google PAIR applied UMAP to visualize and study context sensitive word embedding and how neural networks determine the sensitivity of the embedding. Neural Networks are thought of as black-boxes so UMAP and visualization can help improve visual machine learning interpretability.
5) Knowledge Discovery - Vast amounts of papers are published in academia and it is difficult to track recent advances and find papers that are relevant in research. Orion Search, is a tool that aids user in knowledge search and is an example of

projecting high dimensional data to a lower subspace and making it interactive to make the search experience better for users.

Resources: https://umap-learn.readthedocs.io/en/latest/interactive_viz.html

Describe five best practices involving UMAP by reading the following papers and blogs

1) Choose hyperparameters carefully. Different hyperparameters produce different results and it is easy to misinterpret the visualization based on constraints involving the data and how accurate the project must be. UMAP is fast which allows for better iteration to produce an accurate visualization.
2) Be aware of spurious clusters. As the number of n_neighbors decreases the overall global structure is lost which can result in misinterpretation of the final visualization and the overall structure of the data.
3) Experiment and visualize multiple hyperparameters. Visualizing runs with multiple visualizations allows for better understanding of how to tune UMAP for a particular dataset and what are the optimal parameter values for accuracy and reproducibility.
4) Ignore the size of clusters in UMAP visualizations. Under the hood UMAP uses a notion of distance that can be customized by the user to produce a particular result.
5) Ignore (potentially) cluster distances. Similar to cluster size, since UMAP uses its own notion of distance, the final result may have a warped distance in a lower dimensional subspace.