ISyE 6740 – Summer 2020
Final Report

Team Member Names: **Thomas Cycyota**
Team ID: **160** *(solely responsible for all work completed in project)*

Project Title: **Spectral Fly Fishing**

Problem Statement

In the world of fly fishing, there are thousands of fly options that a fisher-person can choose from to catch the perfect fish[1]. Depending on geographic location, time of year, target fish species, and countless other factors, the skill choosing the ideal fly can take a lifetime of fishing to master.

To add to the challenge of mastering fly selection for the average hobbyist, one must also learn to navigate trends and new arrivals in the wholesale fly market[2]. Meant to mimic bugs or other small creatures in every possible combination, the list of flies available to the recreational fly fisher continues to grow as innovative new patterns, materials, and techniques meet the market demands of fly-fishing consumers.

With this context, this project assesses the viability of machine learning techniques in comparing and contrasting different fishing flies. In particular:

- Assessing *unsupervised clustering techniques* to group flies into commonly-accepted categories
- Using *spectral clustering* to identify important factors and relationships between flies

Data Source

The data source used is obtained from Orvis.com, one of the industry leaders in fly fishing and a well-known source for wholesale fly patterns. This source website was chosen because of its large inventory, as well as because all images for flies are standard in size and have a uniform white background, which simplifies pre-processing steps.

Approach

Using specialized Python libraries BeautifulSoup, URLlib, and PIL, images thumbnails were scraped from the Orvis.com online shop for hundreds of fly patterns. These JPEG images were downloaded locally, then resized and compressed to a more computationally effective format of 64x64 pixels. Finally, two Numpy matrices were created to hold the "flattened" pixel values for these images in both greyscale and colored (RGB) formats. Non-desirable images of fly boxes and multiple flies together were manually removed to produce the final dataset consisting of only single flies on a white or light-grey background.

[1] Henderson, Brody (2013, September 8). *How Do I Choose Which Trout Fly Will Work Best?* Vail Valley Anglers, https://blog.vailvalleyanglers.com/how-do-i-choose-which-trout-fly-will-work-best/

[2] Juracek, John (2016, April 18). *A failure of modern fly design*. Hatch Magazine, https://www.hatchmag.com/blog/failure-modern-fly-design/7713505

Dataset

The final dataset consists of **459** images of flies. This dataset is converted into both a greyscale and RGB matrix formats, where each row represents an image and each column represents a corresponding pixel value for the 64x64 reshaped image:

- Greyscale dataset: `images_grey_final.txt`, shape of (459 rows, 4096 columns)
- Color RGB dataset: `images_rgb_final.txt`, shape of (459 rows, 12,288 columns)

For full code details, please see project file "`00_ImageScraping`".

Sample Data

Fly name: Yellow Humpy
Index in dataset: `117`
Image of fly in (a) full-res original, (b) re-scaled RGB, and (c) re-scaled Greyscale formats:

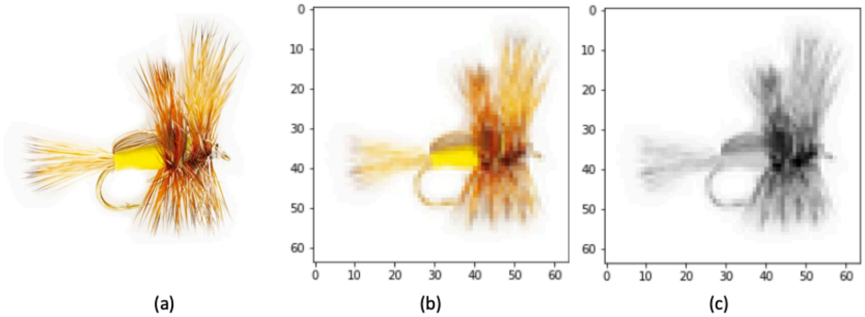

Figure 1. Sample images of a single fly

Methodology

The methodology in this project uses two primary unsupervised algorithms to better understand the relationships between images in this dataset. Analysis consists of a Jupyter notebook for each phase, including:

- Phase 1: K-means clustering on PCA decomposed image data
    - Please see project file "`01_KMeansClustering`"
- Phase 2: ISOMAP algorithm for spectral clustering visualization
    - Please see project file "`02_SpectralClustering`"

Phase 1 Methodology: K-means clustering

The dataset described above does not come with obvious labels for the data. Each fly has a unique name, text description and recommended use according to the website. However, based on external knowledge[3], fly fishing commonly breaks down all flies into three categories:

- Dry Fly: designed to simulate an insect landing or floating on top of the water.
- Nymphs: designed to sink to simulate the mostly underwater lifecycle stage of an insect
- Streamers: designed to mimic larger water and land animals to attract larger fish.

[3] Fly Fishing Flies – The Three Types. (2015, December 02). Blue Ridge Mountain Life, https://blueridgemountainlife.com/types-of-fly-fishing-flies

Because these fly categories are relatively well-defined, but we do not have category labels for each image, the k-means algorithm can help better understand these flies purely based on the information contained in thumbnail images. The color RGB fly image dataset was used in this phase to allow color-related features to act as possible clustering attributes.

Given *m* data points corresponding to each image of a fly, k-means clustering assigns each image a label according to its assignment to cluster *k* by minimizing the distortion function over $\{r^{ij}, \mu^{j}\}$, where $r^{ij}=1$ if x^{i} belongs to the j-th cluster and $r^{ij}=0$ otherwise.

$$J = \sum_{i=1}^{m}\sum_{j=1}^{k} r^{ij}|\, x^{i} - \mu^{j}|^{2},$$

In the case of images, it can be both computationally and mentally difficult to interpret a "cluster center" for a high-dimensional data structure. In the case of the problem at hand, each greyscale image has 4096 dimensions, and each color image has 12,288 dimensions. To simplify this challenge, principle component analysis (PCA) helps us select the first two eigenvectors that maximize the variance of the projected data on a lower dimension.

With *m* data points corresponding to each image of a fly, we estimate the "mean" image x̄ (e.g. average pixel value), and covariance matrix *S* defined by:

$$\bar{x} = \frac{1}{m}\sum_{1}^{\mathrm{m}} x\_i$$

$$S = \frac{1}{m}\sum_{1}^{m}(x_i - \bar{x})(x_i - \bar{x})^{T}$$

We want to visualize these relationships in two dimensions, so the goal is to find eigenvectors w_1 and w_2 that satisfy:

$$maximize \frac{1}{m}\sum_{i}^{m}(w^{T}x_i - w^{T}\bar{x})^{2}$$

Taking the eigenvectors w_1, w_2 corresponding to the first and second largest eigenvalues λ_1 and λ_2, we then get the principle components z_1 and z_2 by normalizing by the standard deviation:

$$z_1 = \frac{w_1^{T}(x_i - \bar{x})}{\sqrt{\lambda_1}}$$

$$z_2 = \frac{w_2^{T}(x_i - \bar{x})}{\sqrt{\lambda_2}}$$

Rather than deriving these algorithms, the associated notebooks implement Scikit Learn's implementation, in which case `KMeans` is an efficient implementation of Lloyd's "expectation maximization" algorithm[4] and `PCA` uses Singular Value Decomposition to project data to a lower dimensional space[5]. With a combination of these techniques, clear distinction and separation of the different fly images is obtained while maintaining interpretability.

[4] https://scikit-learn.org/stable/modules/clustering.html#k-means

[5] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Phase 2 Methodology: ISOMAP Spectral Clustering

While in Phase 1 the approach allowed us to distinguish "categories" of flies, in Phase 2 the goal is to differentiate relationships between flies. This is not only useful in the case of understanding this small dataset of flies from a particular website, but also more generally to help interpret the dimensions and attributes along which flies are made: *if a fish was caught with this fly before, what other flies might lead to similar results?*

The greyscale fly image dataset was used for this phase for computational efficiency and iteration speed as the matrix had one-third the pixels compared to the color dataset.

The ISOMAP algorithm is an implementation of spectral clustering, which allows us to understand data in the context of relationships between points. With m data points corresponding to each image of a fly, we follow this approach:

1. Build the weighted adjacency matrix W using each image's nearest neighbors by Euclidean distance. This combines the steps of calculating inter-node distance (e.g. Matrix A), then selecting the top-n neighboring images for each.
2. Calculate the graph matrix D, which is a square matrix representing the inter-node weights.
3. Calculate the graph Laplacian matrix L
4. Compute the resulting eigenvectors and eigenvalues of L
5. Construct matrix Z as the dot product result of the eigenvectors and diagonal eigenvalues.

The result of this algorithm is a matrix Z which allows us to plot each image in a two-dimensional space. However, as opposed to the PCA components in Phase 1 using Kmeans, these two dimensions do not represent features in the original data, but rather the similarity of each image amongst the sample population.

Spectral Clustering allows us to view the relationships between individual images while also observing interesting trends across the entire dataset. One possible application is to discover new flies that one may not have known about before or identify similar patterns amongst a vast universe of potential flies.

Evaluation and Results

Interpreting the results of the above methodologies is by definition a qualitative, rather than quantitative, approach. Working with images that lack other categorical date (e.g. which class of fly), evaluation criteria were chosen that assess the effectiveness of the above approaches in understanding our dataset; where feasible, other quantitative measures have been added.

Phase 1 Results: K-means clustering

To arrive at a single result depicting this problem, the following were plotted together:

- all 459 PCA-reduced images, each with a blue dot
- a blue "X" representing each of three cluster centers
- brown, yellow, and light-blue "mesh grid" of the cluster boundaries (Voronoi Diagram)
- 80 randomly-selected thumbnail images overlaid on top of the plot

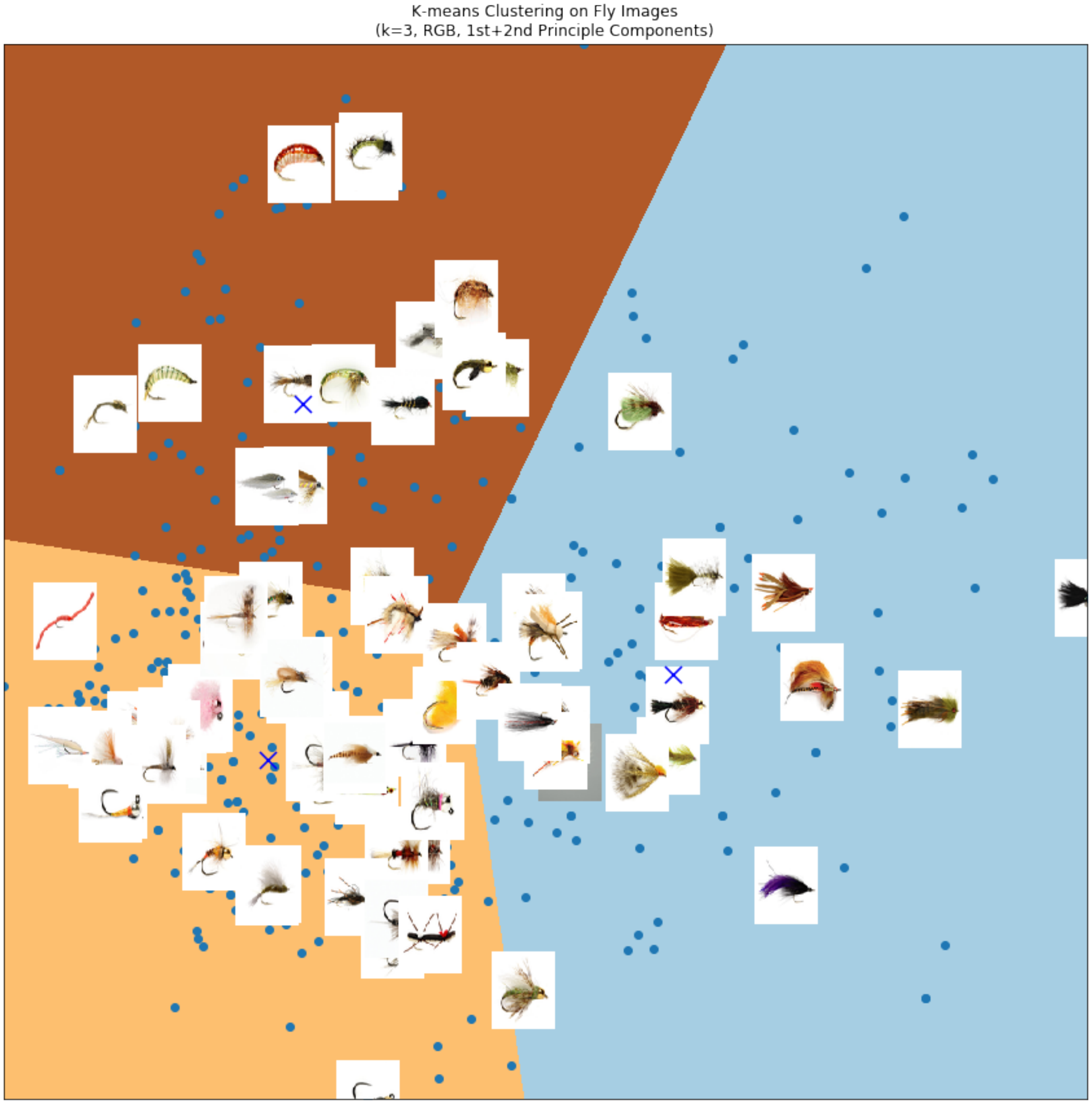

Figure 2. K-means on PCA-reduced RGB fly images

Using Scikit Learn's `KMeans` model fit on PCA reduced data, we get three clusters of the following sizes:

- `Cluster 0 Size: 112`
- `Cluster 1 Size: 244`
- `Cluster 2 Size: 103`

Randomly selecting 36 images from each family and plotting them together, we can see some strong characteristics emerge amongst each cluster. Note that re-running this plot will return different results for each run, but an assessment is conducted below using these plots:

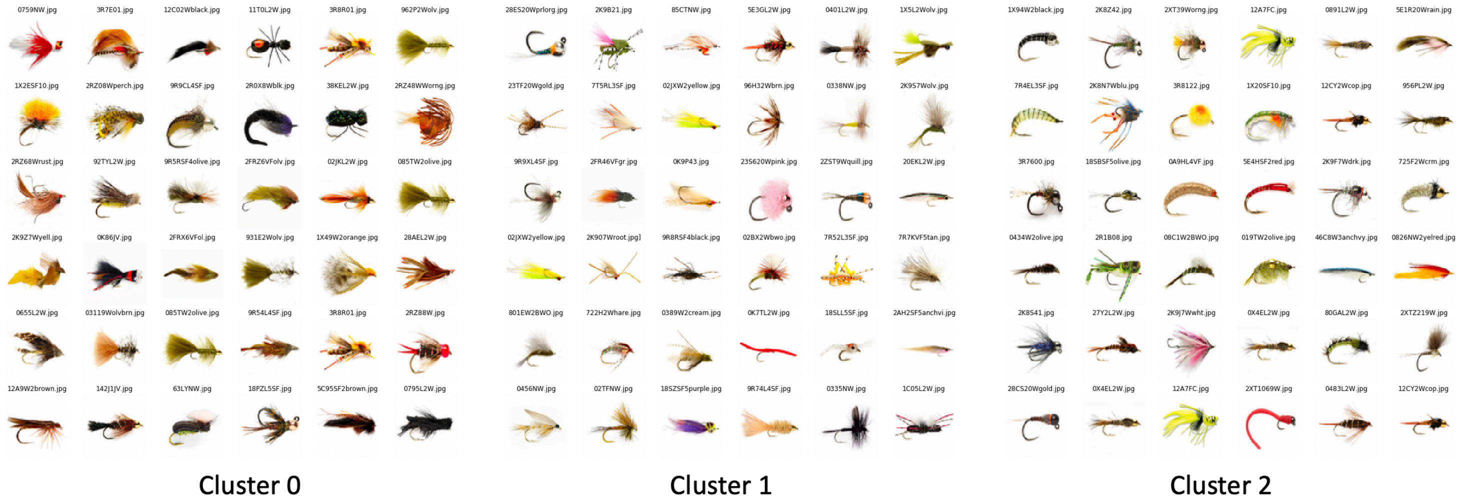

Figure 3. Sample images from each K-means cluster

Qualitative Results

Each cluster has strong visual trends:

- Cluster 0 can be considered the **Streamer** family: these flies take up a large part of the image space, with bold and darkly-colored feather patterns flaring out beyond the back of the hook. The components making up the fly are much larger than the hook.
- Cluster 1 can be considered the **Dry Fly** (e.g. surface fly) family: these are more delicate flies with somewhat translucent "bug wing" patterns emerging towards the front of the hook, helpful in keeping the fly afloat. Many flies also have string-like additions branching off from the main body, similar to the legs or appendages of bugs.
- Cluster 2 can be considered the **Nymph** (e.g. sinking fly) family: the strongest similarity between flies in this family are the curved shape that the materials follow around the hook. Many also have large beads at the head of the fly to make it sink.

While the above qualitative interpretation of these results allows the observer to add experiential context and background knowledge, even observers without fly fishing experience are able to notice differences between the sets of images in each cluster.

Quantitative Results

Using these plots and the fly ID's plotted above each thumbnail, the website was referred to for an accurate description of the fly category[6]. We can then determine which "family" of flies each category represent, and within these plots how many are inaccurately clustered to that fly family:

Cluster	**Interpreted Fly Family**	**# in Family**	**# out of Family**	**Mis-classification Rate**
Cluster 0	Streamer	33	3 Nymph	8.3%
Cluster 1	Dry Fly	17	10 Nymph, 9 Streamer	52.7%
Cluster 2	Nymph	26	8 Streamer, 2 Dry Fly	27.8%

The results above reveal that this PCA and K-means approach is useful in predicting some fly families, but not all. From just an image, it is relatively useful in predicting if a fly is a "streamer", but less so for the "dry fly"

[6] For example, in Cluster 2 we see "2R1B08". Searching online for "Orvis 2R1B08", we find the page describing this is a streamer fly: https://www.orvis.com/p/cohen-s-frog-legged-popper/2r1b

and “nymph” families. Cluster 1, which contains over double the number of flies as either of the other clusters, is less distinguished in the fly attributes it captures, and is not as useful for potential future applications; in other words, if this model were used to classify a new fly, a label of Cluster 1 would be much less reliable for a fisher-person than a label of Cluster 0. This distinction also helps us realize that in fly fishing, many flies look nearly identical and are heavily biased towards looking like the flies in Cluster 1 above.

Phase 2 Results: ISOMAP Spectral Clustering

In computing the necessary matrices for ISOMAP, the 25 nearest neighbors by Euclidean distance were used, which allows us to compare each fly to the most similar 5% of the rest of the flies. This number proved effective computationally and in the resulting depiction of the ISOMAP plot.

Following the ISOMAP method described previously, we are able to visualize:

- all 459 ISOMAP-reduced image locations plotted with a small blue dot
- X and Y axis depicting the plane of relationship dimensions intrinsic to this image dataset
- 200 randomly-selected thumbnails images overlaid on top of the plot

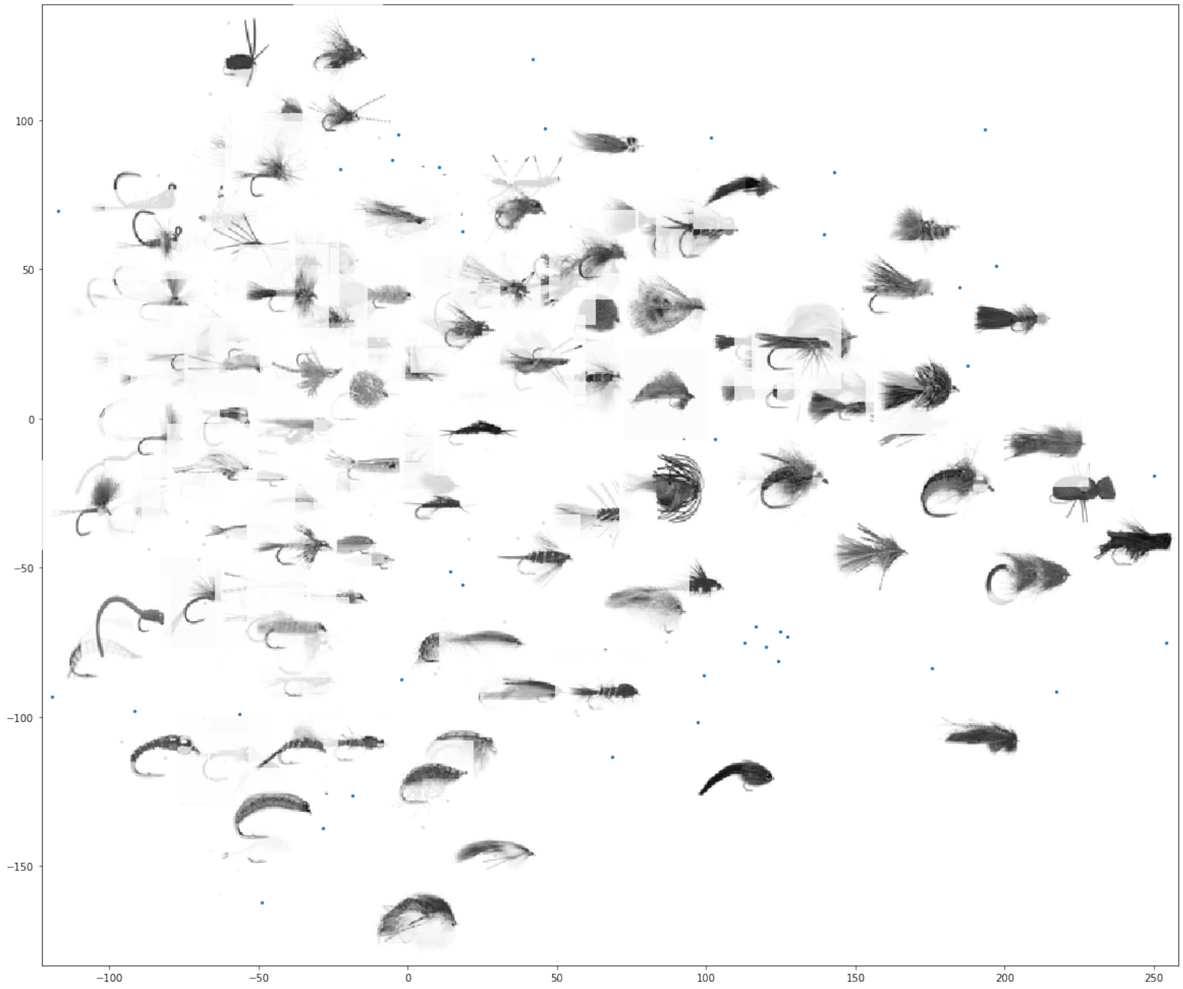

Figure 4. ISOMAP Spectral Clustering on greyscale fly images

From the above plot, some visual trends begin to appear to the observer: images at the right-side of the plot are darker and thicker, while at the left-side are thinner and less fully-pixelated. Images towards the bottom are more compact and tightly curving inwards, while towards the top are radiating outwards and lack an inward-curve.

Because ISOMAP is a technique closely related to PCA, we expect to see somewhat similar results between the two techniques; this is indeed the case, where the PCA-axes closely align to the ISOMAP axes in terms of visual traits of the images. To pick apart the differences between the two approaches, ISOMAP gives us a simple method to pick out similar flies. The below plot depicts the five images at each extreme of the ISOMAP axes.

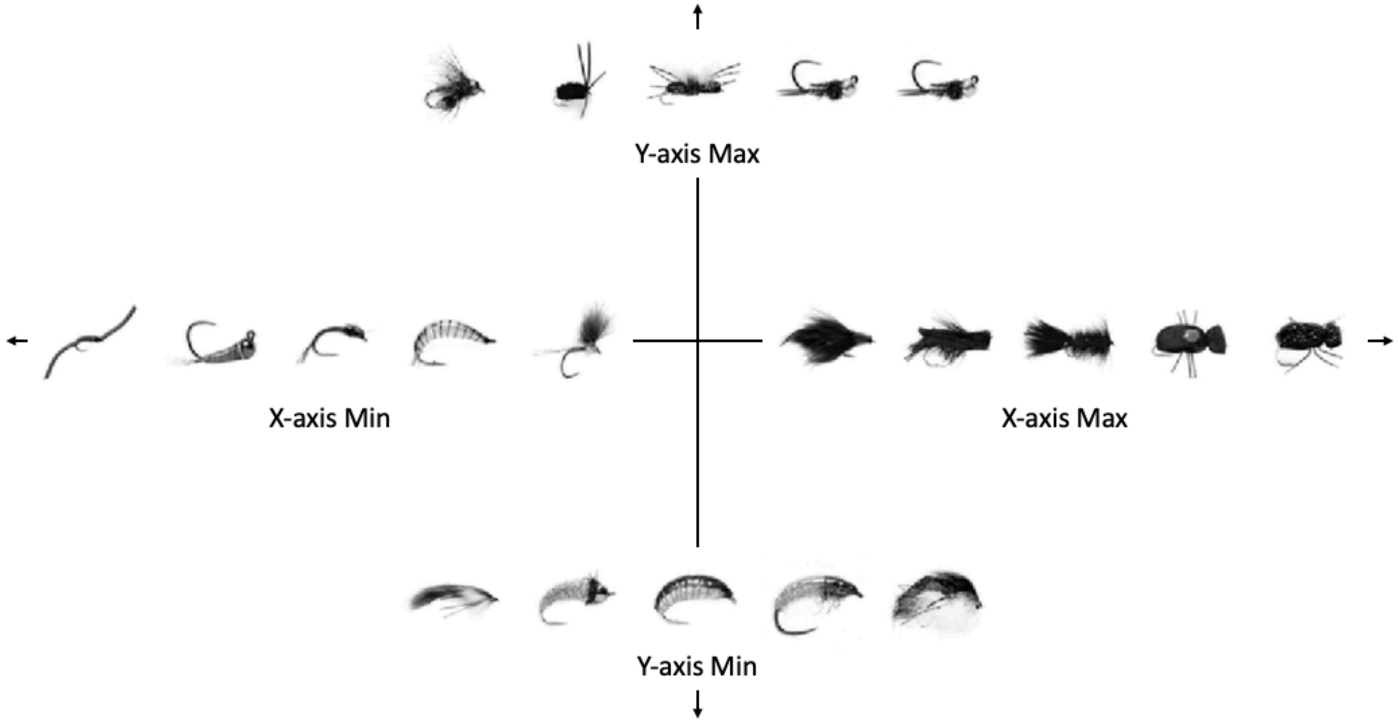

Figure 5. Images at extremes of ISOMAP mapping

While somewhat challenging to pick out from the first plot, the above clearly confirms the initial impressions, namely that we can reason about the X-axis as a thin vs. thick dimension, and the Y-axis as a curved vs. radiating dimension. Upon closer analysis of each of these images we notice how similar some of the flies are: in “Y-axis Min” flies actually seem to mimic different life-cycle stages of a bug, and “Y-axis Max” have two flies that are actually different colors in the original dataset. ISOMAP appears useful to the recreational fisherperson on both an individual level, in the case of identifying similar flies to a known image, and as a whole industry, in the case of identifying larger trends and dimensions of design already present.

Potential Next Steps

While much of fishing is considered generational knowledge passed person-to-person, this project demonstrates that we are able to extract insights simply from images of flies. To continue expanding this body of research, obtaining labeled images of flies would allow supervised learning applications and more detailed statistical analysis of existing fly trends. Additionally, generative adversarial networks (GANs) might allow new flies to be created by using neural networks to generate novel fly patterns.

While the dataset used in this analysis was relatively small, a larger body of images with more comprehensive metadata would provide machine learning applications that could further address the challenges of fly fishing outlined in this project.