I'm currently working on a project that focuses on the creation of a machine learning algorithm to categorize the scents of chemicals based on their gas chromatography signals. I started working on this project last spring with Dr. Novikov. The project would require chemicals to be purchased and run through a gas chromatography (GC) apparatus, graciously supplied by the Applied Physics Institute (API), in order to create a basis for an algorithm that would take chromatograms for unknown scents and generate a verbal description of their odor. We applied for a KY NSF EPSCoR grant, and for that, I worked on drafts of the request. We did end up receiving support from them.

In the beginning, I focused on Andrew Dravniek's *Atlas of Odor Character Profiles* (1985). The Atlas records the applicability of 146 distinct odors for 160 different chemicals. Scents are usually grouped into families dependent on the verbal description. To start, I categorized these chemicals into scent families.

I then modeled the Atlas applicabilities in Python. I started looking into different order reducing techniques because plotting 146-dimensional data is not readily possible. The goal of this was to show a correlation between families. I started with Linear Discriminant Analysis (LDA) which takes two of the multidimensional data and plots each data point as a point of only those two dimensions, color-coded by family. For example, I plotted all of the chemicals based on only the descriptors Lemon and Cadaverous. I moved to Parallel coordinates, which worked as a graph with a plot for each chemical with applicability on the y-axis and the descriptor on the x-axis, with the plots color-coded for family. I then moved to actual dimension reduction. Using t-SNE (t-distributed stochastic neighbor embedding), I found that the reduced data set did not have the distinct clouds for each family I was looking for. Using PCA (Principal Component Analysis), I found similar results.

This was around the time of my first presentations, so I worked on some visualizations for those. I made spider plots for each family, and then word clouds so there would be a better visual understanding of their similarities and differences.

During the modeling period, I also looked for the chemicals that corresponded the highest with the verbal description of their family so that there would be a basis for the algorithm. With input from Dr. Novikov and Adam Emberton from API, I settled on three chemicals from three scent families. I chose to focus on the aromatic, green, and floral families. While waiting for the chosen chemicals to arrive, I was trained on the GC API allowed me to use for the project. I ran acetone, one of the descriptors from the Atlas, at different concentrations so that I could show how concentration relates to amplitude in a chromatogram.

Once the chemicals were received, I ran them each to check for the detection time. This dictates how long to run the GC for samples of different chemicals. I found that most of the scents from the same family tend to be detected at similar times.

Currently, I am working on finding the detection limits, or the lowest concentration that can be detected and characterized as a certain chemical, for each. I am also working on some data analysis for the WKU Student Research Conference. I am also trying to use Umap visualizations of the Atlas data, which is another dimension reduction tool.