Statement of Purpose for Ph.D. Application

As I take a break for a scroll through my social media feed, a strange feeling often arises. Somewhere between the second or third targeted ad, I begin to feel that my phone remembers more of my conversations than I do. While I used to think I was the only one who felt this way, this has become a well-known phenomenon that many accept as part of the contract one makes when purchasing and subscribing to modern day technology. Only a few truly seek to dig deeper to discover who exactly is taking note of our searches, and what their goal is. I am curious and want to obtain answers. How do “they” manage to store and learn from all this data that grows larger by the second? Where do we draw the line between ethical gathering of information for research and abuse of personal data? Most importantly, how can this overload of knowledge be taken beyond advertisements and be used to create a positive change in the world? After four years of studying data science and applied mathematics at Cal Poly, I have gathered the knowledge to answer some of these questions, but certainly not all.

During my first year as an applied math major, it was my prerogative to find a concentration I felt passionate about, so I began to explore potential paths by taking various computer science classes and statistical modeling courses. I enjoyed my time in these classes so much that I struggled to choose just one to focus on. When I discovered data science - where statistics, mathematics, and computer science collide, I knew that this would be the perfect opportunity to receive an interdisciplinary education where I could apply theory to solve real world problems.

While my undergraduate degree provided me with a solid mathematical and statistical background, my goal is to focus on computer science to complete my understanding of the data science world. The more I learn about the field, the more I gather that it is only the beginning of a big change to follow. It is evident that the amount of potential data that is being collected continuously exceeds any pre-existing capacities to process it. Have we learned how to sufficiently process this data? Or have we merely touched a number of individual problems, like image, text, and speech processing, and are only scraping the surface of other areas of application? Are the same algorithms being correctly applied to different problems or are there unexplored models better suited for some? I aspire to continue my journey to answer these questions in a highly academic setting where I can gain a deeper understanding of machine learning and data engineering, learn the theory behind the algorithms, and obtain the programming proficiency necessary to improve upon them.

After graduating Summa Cum Laude and receiving an award for Graduating Senior Mathematical Excellence, I became a full-time data scientist at Specific Diagnostics, a biomedical company that focuses on early diagnostics of antibiotics rejections. My industry experience as well as my undergraduate research, inspired me to pursue a graduate degree in the data science field.

The first step I took in realizing my long-term educational goal of pursuing a doctorate was diving into a summer research project. I won a prestigious grant to build a bioacoustics app, using the programming language R on the Shiny-App platform, with a top data science and statistics scholar. The goal of the project was to create a free, interactive website for marine biology students with which they can upload sound files and visualize the audio using graphs. While I was unfamiliar with the bioacoustics field prior to beginning this research project, I was excited to develop a new software and apply data science techniques to data outside the traditional table format.

I began creating the first tab in the app, where the user can upload a .wav or .mp3 file or choose from ten preloaded audio files. The first challenge I encountered was the complexity of sound data. Given the variety of input formats, sound is traditionally represented through different graphs, combining the time, frequency, and amplitude of a signal into a spectrogram, an oscillogram, or a spectrum graph. While R has dozens of packages that are built for bioacoustics graphs, I wanted to find a way to make those graphs more interactive and user friendly so that future students have an easier and more engaging experience. In Shiny, you can create input windows which allow the user to input different values for the frequency, amplitude, and time, which in turn change the graphs being displayed. While that method was certainly a sound solution, it forced the user to toggle the inputs one at a time, slowing down the app significantly. After researching graphing tools in R, I found ggplotly, which allows for zooming in and out of a graph, and it inspired me to create an interactive graph in which input windows were no longer needed. Since each of the three graphs shared an axis with another graph, I realized this could be leveraged so that when the user zooms into one graph, the other two will change simultaneously, completely eliminating the need for an outside menu while also reducing loading time.

The second feature I created is the segmentation algorithm that can distinguish between marine animal sounds and background noise. This tool highlights audio segments which show marine animal communication, so the student can study the patterns different animals make. The student may choose to click on one or more segments and see the frequency, amplitude, and the elapsed time of each call segment, all of which are downloadable for future use.

What started as a summer research opportunity turned into two years of hard work as I was the sole student researcher. As a result, my website application has become an official part of the curriculum for introductory bioacoustics courses.

Knowing that my work will actually benefit students propelled my desire to further explore the global impact of data science and software. For my senior project, I wanted to continue to use my skills and passion to help create positive change while strengthening my knowledge of machine learning. Worldwide, car accidents and road crashes are the leading cause of death among young people. Accident rates can be reduced by conducting road safety assessments, but data collection for these assessments requires constant care and thus can be too expensive in countries with severe resource limitations. Advised by a scholar, and in collaboration with Amazon Web Services and the World Bank, I devoted six months to this cause. As a solution, I set up the infrastructure to help reduce the cost and time associated with road assessments through the use of street view images that are captured by cameras posted at intersections. The World Bank provided me with 20,000 of these images and AWS trained me to use their cloud instances, which allowed me to develop an image classification algorithm that automatically identifies key road features such as pedestrian facilities, lanes and traffic signals.

First, using CityVista software, I segmented the images so that each object (car, stop sign, pedestrian, etc.) is assigned a color. Then, I transferred the RGB values of each pixel to a data frame where a list of objects was assigned either 1, if that color existed in the image, or 0 otherwise. In terms of road safety assessment, it was especially important to take into account the number of certain entities present in each image, in addition to indicating whether they exist or not. To accommodate this, I created a non-binary feature detection algorithm for specific entities, such as people and cars, so that later those counts could help evaluate how crowded the particular intersection is. After putting together my feature set for each image and researching the most salient models for this kind of data, I chose a convolutional neural network (CNN) as my image classification algorithm. I optimized the architecture of the CNN as well as the learning rate, and as a result, the final model performed with 95.6% validation accuracy. After months of development, I was thrilled by the final performance, but I was curious to find out if the results were actually sufficient. Although my supervisors at AWS as well as my advisor were very pleased and sent the project forward, it was difficult to shake my own inclination that anything below 100% is not good enough. Unlike building a model for an AI course where the only thing at stake is my grade on the assignment, I felt like my work had real weight to it and could impact the lives of many people. I would like to continue studying machine learning so that I can understand where we draw the line to say “accurate enough” and where we have to continue pushing for higher accuracy.

While there is significant value in one-on-one work with professors, my capstone project for the data science minor enhanced my ability to bounce ideas off of other students and create meaningful results as a team. Similar to my other projects, it continued fulfilling my quest to find ways to use data science to bring positive change into the world. I put together a team of students to partner with an organization that is leading the nation in data analytics techniques designed to stop human trafficking in the United States.

We were tasked with creating a model that will be able to assign illicit massage businesses in California with a risk score that reveals the likelihood of that specific business engaging in human trafficking. We started gathering data by scraping reviews from Yelp and Google

Reviews using the Beautiful Soup package and Selenium, as well as obtaining data from a website called RubMaps which is meant to rate prostitutes at illicit massage parlors. Using the organization's human trafficking vocabulary, we were able to perform text analysis on each review and count the frequency of triggering words. After solidifying our feature set, we proceeded with KMeans clustering, an unsupervised algorithm, as we did not have ground truth as to which businesses were involved in trafficking and which were not. We validated our model by testing it on a similar dataset that was created in other states such as Florida, which has already gathered some ground truth regarding those businesses. Since it correctly classified the known illicit massage businesses in Florida, we used it to compile a list of businesses in San Luis Obispo with probability greater than 85% to engage in human trafficking. This list along with a detailed report was sent to the district attorney in San Luis Obispo, which helped push the district to begin taking action against human trafficking.