Elijah Gunther, University of Pennsylvania

Scientific internship with Prof. Brenda Wilson, University of Illinois
(This project was joint with Ella Hiesmayr)

The goal of our project was to identify proteins likely to be cytotoxic necrotizing factors (CNFs), a type of cell-killing toxin produced by bacteria such as E. coli. The only way to definitively identify such proteins is through laboratory testing, but there are billions of known protein sequences, so our goal was to narrow these down to a limited set of likely candidates. Proteins consist of a chain or chains of amino acids folded into specific conformations that are key to their function. In our research we use several different tests in series to narrow down our list of sequences, each one using a higher level of protein structure than the previous, thus being more selective but also more computationally intensive. First, we searched for sequences with a few key amino acids in a certain configuration that are found in all known CNFs. Then we used a software called HMMER, which uses profile hidden Markov models, a statistical model common in bioinformatics, to identify proteins with sequences similar to those of known CNFs. We then built a statistical model to identify sections of proteins with beta sheets, a structure inside a protein known to be present close to the active site in CNFs. Finally, we used a software package called MODELLER to attempt to fold sequences into confirmations similar to those of known CNFs, then assess the quality of the folding as a proxy of the likelihood the protein actually is a CNF.

Ella Hiesmayr, University of California at Berkeley

Scientific internship with Prof. Brenda Wilson, University of Illinois
(This project was joint with Elijah Gunther)

The goal of our project was to identify proteins likely to be cytotoxic necrotizing factors (CNFs), a type of cell-killing toxin produced by bacteria such as E. coli. The only way to definitively identify such proteins is through laboratory testing, but there are billions of known protein sequences, so our goal was to narrow these down to a limited set of likely candidates. Proteins consist of a chain or chains of amino acids folded into specific conformations that are key to their function. In our research we use several different tests in series to narrow down our list of sequences, each one using a higher level of protein structure than the previous, thus being more selective but also more computationally intensive. First, we searched for sequences with a few key amino acids in a certain configuration that are found in all known CNFs. Then we used a software called HMMER, which uses profile hidden Markov models, a statistical model common in bioinformatics, to identify proteins with sequences similar to those of known CNFs. We then built a statistical model to identify sections of proteins with beta sheets, a structure inside a protein known to be present close to the active site in CNFs. Finally, we used a software package called MODELLER to attempt to fold sequences into confirmations similar to those of

known CNFs, then assess the quality of the folding as a proxy of the likelihood the protein actually is a CNF.

Nirjal Shrestha, University of Florida

Scientific internship with Prof. Paul Bonthuis, University of Illinois

Animal behavior includes all the ways animals interact with other organisms and the physical environment. Behavior can also be defined as a change in the activity of an organism in response to a stimulus, an external or internal. To fully understand a behavior, we want to know what causes it, how it develops in an individual, how it benefits an organism, and how it evolved. Some behaviors are innate, or genetically hardwired, while others are learned, or developed through experience. In many cases, behaviors have both an innate component and a learned component [1]. Behavior is shaped by natural selection. Useful behaviors directly increase an organism's fitness, that is, they help it survive and reproduce. Ethology is a field of basic biology, like ecology or genetics. It focuses on the behaviors of diverse organisms in their natural environment. Many problems in human society are related to the interaction of environment and behavior, and in the context of human health there is interest in the contribution of genetics to influence behavioral health and the efficacy of distinct therapeutics [2]. The fields of socioecology and animal behavior deal with the issue of environment behavioral interactions both at an evolutionary level and a proximate level. Increasingly social scientists are turning to animal behavior as a framework in which to interpret human society and to understand possible causes of societal problems. Careful behavioral data allow neurobiologists to narrow the scope of their studies and to focus on relevant input stimuli and attend to relevant responses. In many case the use of species-specific natural stimuli has led to new insights about neural structure [2]. Different animal behavioral models are used to empirically answer specific biological questions. Animal behavior experiments traditionally require labor intensive scoring of behaviors by researchers, where subjective researcher biases can influence variability and reproducibility of the findings. Thus, supervised machine learning of behavior videos has the potential to solve many of these problems. Supervised machine learning is a subcategory of machine learning and artificial intelligence that use labeled datasets(behavioral) to train algorithms which classify data/behaviors (in our case) or predict outcomes/behaviors accurately [3]. Specifically for behavioral analysis, deep learning models is used for pose estimation and behavior tracking to analyze videos and extract behavioral data. Pose estimation is a computer vision technique that predicts and tracks the location of a person or object. In this project we use the pose estimation, and behavioral analysis applications that use machine learning to build classifiers to track mouse body point movements and score complex social interactions like aggression, mating, parental care etc. For the pose estimation, we used DeepLabCut (DLC)and Social LEAPEstimates Animal Poses (SLEAP), where LEAP stands for Latent Encoding of Atypical Perturbation. LEAP maps system structure changes to neural net structure changes using structural actionable variables [23]. For the behavioral analysis, we used Simple Behavioral Analysis (SimBA). DeepLabCut is an efficient method for 2D and 3D markerless (no reflective markers present to assist computer for tracking) pose estimation based on transfer learning with deep neural networks that achieves excellent results (i.e. you can match human labeling accuracy) with minimal training data (typically 50-200 frames)per project[4].SLEAPis an open source deep-learning based framework for multi-animal pose tracking[7].SimBA is an open source toolkit for computer classification of complex social behavior with the flexibility for users to define their own classifiers[6].The basic outline for the project is to import the videos for a single behavioral model paradigm, extract frames from videos, label the frames for pose estimation/behavioral analysis, train the machine learning model(s),predict on new data, extract behavioral measures from each subject, and apply different visualization technique if required. The pose estimation and behavioral analysis have completely different training steps, labeling and training models. For instance, each unique behavior experimental paradigm, like aggression versus mating, requires training of unique pose estimation and behavioral analysis machine learning models. This process of extracting behavior data is as equally or more accurate than that done by experimenter observations with less man hours of work. If the same machine learning models trained to acquire data for a specific behavioral paradigm is used by a different experimenter or on a unique set of animals to answer a different question, it also solves the problem of experimenter biased scoring of the data that may confound the interpretations of the results.

Shannon Weed, University of Notre Dame

Industrial Internship with Eric Weisstein at Wolfram|Alpha

Wolfram|Alpha is a computational engine with a broad knowledge base. One target audience for this software is students. Currently, students can use Wolfram|Alpha to compute solutions to many types of problems, including basic calculus problems involving limits, derivatives, integrals, and more. The goal of our project was to expand the results for calculus on Wolfram|Alpha to help further students' education. In addition to being able to compute solutions to problems, students will be able to search for calculus terms and theorems and access relevant information and interactive demonstrations. My work involved writing this information in a way that was both precise and student friendly. Once this project is finalized and published, students will be able to access the majority of Calculus I and Calculus II terms and theorems through Wolfram|Alpha in an interactive way conducive to learning.

Jingyang Judy Zhang, Northwestern University

Industrial Internship with Bradley Janes at Wolfram|Alpha

I worked with the Wolfram content developing team on developing step-by-step solutions for statistical contents. My projects are creating detailed explanation and step-by-step instructions for conducting two widely used hypothesis tests, two-sample t-test and F-test of equal variances. Both projects are written in Wolfram Language and will be added to the Wolfram|Alpha site, a powerful computational knowledge engine that provides answers to users' queries. The step-by-step solution is an add-on feature that presents users with easy-tofollow steps leading to the final answer of the problem. Therefore, the content developing team must carefully consider how these steps are constructed and presented in the step-by-step solution so that the content is not only helpful for understanding the problem-solving process but also clear and concise to avoid overwhelming users, especially those who are new to the topic.

Summer 2021

Mauricio Campos, University of Illinois Urbana-Champaign
Scientific internship with Prof. Alex Leow at The Collaborative Neuroimaging Environment for Connectomics at the University of Illinois Chicago (CoNECt@UIC)

The activity of neuronal populations is dynamic, continuously fluctuating and can reveal information about disease in human brain. Using functional MRI (fMRI), people are able to measure how different regions in the brain are connected functionally. With the help of diffusion spectrum imaging tractography, people can also have access to the structural connectivity. It is widely understood that the functional connectivity is informed by the structural one. However, the underlying structure to function relationship has not been fully studied. In this project, we construct a graph-based encoder-decoder system to learn the mapping from structural connectome to functional connectome. We implement a graph convolution network (GCN) in the encoder and integrate information from neighbor brain regions to get the low-dimensional embedding. The embedding shows the relationship and the regions that play important roles in the mapping.

Surya Teja Eada, University of Connecticut

Scientific internship with Prof. Justin McGrath at the United States Department of Agriculture

The USDA has an R private package called “BioCro” that uses various crop and soil related parameters and inputs changing weather conditions to model growth of a plant for various types of crops such as Soybean, Sorghum, Miscanthus, and more. The modeling uses multiple parameters to virtually model the growth of the crop. The model estimates dynamic values such as areas, mass of leaf, stem, grain, root based on the changing weather conditions with the idea that the thermal time has a strong relationship to the growth of any plant. The model currently has many versions according to the modules that are used to dynamically grow the plant and each version of the model has many parameters. The model version we worked on was using a partitioning logistic module and the crop we worked on to start off with is sorghum. In this sorghum, the yield is evaluated as the weight of leaf per unit area. The aim of the project was to find appropriate actual values of yields, areas to compare the model with the actual and in some way find the parameters that are optimal (gives the best fit). In some sense, we also sought to find the appropriate loss definition minimizing which gives a reasonable fit. We started off with squared error loss, absolute loss but then stuck with a normed Mahalanobis loss that considers the different variances in various outputs and the number of data points at a particular time of growth. The project also identified datasets with measurements throughout the growing period and a dataset with only tractor harvesting measurements and intended to find if using just the tractor harvesting measurements may give a good fit. We obtained that using the data from the entire growing season will allow for improved normed Mahalanobis statistic. Also using this statistic is very important in comparison to others since using other statistics might bias towards harvest season due to its yield values which are high. If the final yield is only what we intend to predict, then one can maybe use the harvest data only but since the importance of the BioCro model is also drawn due to its outputs throughout the season it makes more sense to find parameters that fit the entire curve well which is only done by the normed Mahalanobis statistic or by using equally spaced equal number of data points throughout the growing season. Better conclusions can also be made using subject matter judgment giving priority of yield and leaf area index (LAI) in some numerical fashion. Currently, ‘more data’ seems better overall however inclusion of LAI data from the growing season can give scores that are reasonable. In such case, gathering data at least for partitioning sake and biomass yield during the season can be avoided.

Rebekah Eichberg, Indiana University

Scientific internship with Prof. Maryanne Alleyne at the University of Illinois

This project focused on using a combination of data/computer science techniques to more efficiently study a class of molecules called brochosomes that cover the surface of Leafhopper insects. The overarching goal of the research itself is to better understand how the design/form of brochosomes influence their function. Our specific focus was on making data processing and analysis more efficient and straightforward for the frontline researchers. The main effort during this summer’s internship focused on building a structured outline for how brochosome diameters could be obtained and easily processed without manual measurements. This work involved learning the ins-and-outs of using a modern neural network processing technique that can be used to highlight the outlines of the brochosomes in images obtained from the leafhoppers. Over the course of the summer, we also worked on various minor projects including implementing statistical analysis of brochosomes on a wing position/species basis, contact angle measurements of nanodroplets on various surfaces, and data visualization of brochosome images. We found the project fascinating in the multidimensional scope it required. In our short time, we used traditional computer science techniques for data/image processing, modern day neural network structures, and numerical/statistical software packages to approach the problems encountered.

Stark Ledbetter, University of Washington

Scientific internship with Prof. Justin McGrath at the United States Department of Agriculture

The USDA has developed a software called BioCro, which models crop growth based on weather conditions in each season. The version of the model that we worked with takes 81 parameters. Parameters are constants in equations, like pi in the equation for the area of a circle. Some of these 81 parameters can be obtained from known facts about a given crop (for example, the coldest and hottest temperatures at which the crop can survive). These work like pi. They have been measured by scientists, and we have a good idea of their value. However, there are 11 parameters that are prohibitively difficult to measure directly. Instead, processes they affect are measured, and their values are inferred by fitting them to a model for each crop. To fit these parameters, we consider real data by measuring crops throughout a season (this part was done by others before the project started). Then we run the model with many hypothetical sets of values for the parameters, optimizing according to model fit criteria, until we find the set that gets the model predicted values closest to the measured values. We considered several statistical criteria to measure "closeness," and one of the main goals of this project was to find an effective definition of "closeness" for this type of modeling. Projects like this have been done before with other crop models. The most common way to do this is to fit the model to end-of-year yield (the weight of usable crop that gets harvested). BioCro predicts a lot of values other than end-of-year yield. In fact, it predicts the expected weight and size of various parts of the crop at every hour of the growing season. Therefore, it made sense to collect more measurements than just end-of-year yield, giving a more accurate set of fitted parameters. One big way to tell that the parameters are more accurate is that after fitting, the model still performs well when run with an entirely different (test) data set.

Xi Ning, University of North Carolina at Charlotte

Industrial internship with Jerome Ng at AbbVie

We finished the project "Chronic Migraine patients treatment choice Analysis using Machine Learning predictive models". A migraine is a neurological disorder characterized by recurrent headache attacks of moderate to severe pain. Chronic migraine (CM) can be diagnosed when patients experience 15 or more headache days per month for more than 3 months with at least 8 days of migraines features. In this project, we applied four machine learning models, regularized logistic regression, Random Forest, XGboost and traditional logistic regression to investigate the real-world characteristics of CM patients who are persistent and exclusive user of two different migraine treatments. Our goal is to explore the important factors that contribute to the drug choice of CM patient. Besides, we compared the strengths and limitations of these four models for predicting a binary patient related outcome in general.

Claire Plunkett, University of Utah

Scientific internship with Prof. Alex Leow at The Collaborative Neuroimaging Environment for Connectomics at the University of Illinois Chicago (CoNECt@UIC)

I worked with the CoNeCT lab and the University of Illinois Chicago on their BiAffect project, which uses typing dynamics collected via a custom keyboard to analyze study participants' mood and cognition. I implemented multiple types of hidden Markov model to investigate if we could tell what type of cognitive process a user is in based on the order of keys they press on their keyboard. These models used machine learning in Python to fit the models to the data of an individual. We found that there are multiple types of cognitive processes that can be detected, such as typing-heavy phases and correction-heavy phases. These models can now be fit to the typing dynamics of other participants to make comparisons between and within participants.

Mengchen Wang, University of Illinois at Urbana-Champaign

Scientific internship with Prof. Alex Leow at The Collaborative Neuroimaging Environment for Connectomics at the University of Illinois Chicago (CoNECt@UIC)

The activity of neuronal populations is dynamic, continuously fluctuating and can reveal information about disease in human brain. Using functional MRI (fMRI), people can measure how different regions in the brain are connected functionally. And with the help of diffusion spectrum imaging tractography, people can have access to the structural connectivity. It is widely assumed that the functional connectivity reflects the structural one. However, the underlying structure to function relationship has not been well studied. In this project, we construct a graph-based encoder-decoder system to learn the mapping from structural connectome to functional connectome. We implement a graph convolution net- work (GCN) in the encoder and integrate information from neighbor brain regions to get the low-dimensional embedding. The embedding shows the relationship and the regions that play important roles in the mapping.