Previous Research

Matthew Le

For the past year, I have been involved in two research projects, both of which are funded by the NSF Research Experience for Undergraduates (REU) program. The first project is under the guidance of Professor Eric Van Wyk, dealing with finding ways to design programming languages, such that they can easily be extended to support additional constructs. The second project that I have been working on deals with investigating the effects of Pacific Ocean warming on Atlantic hurricane activity. For this project, I have been working under the mentorship of Professor Vipin Kumar while working closely with one of his graduate students and a senior research scientist in Professor Kumar's laboratory. Through these research opportunities, I have been able to produce open-source software, present my research to the general public, and collaborate with leading experts.

Domain specific languages (e.g. R, MATLAB, SQL, *etc.*) have become popular because they allow the programmer to code at a high level of abstraction using constructs relevant to their discipline (e.g. matrices, database tables, *etc.*) while enjoying the benefits of domain specific optimizations. However, the cost of designing, maintaining, and implementing domain specific languages is often expensive[2]. Our research aims to allow programming language experts to extend general purpose programming languages such as C or C++ into a domain specific language in a modular and composable manner (*i.e.* changes to the host languages are automatically propagated to the custom extensions).

The extension that I have been working on provides support for matrices and the operations that accompany them. As datasets continue to increase in size and dimensionality, matrices have become a intuitive data structure for large multi-dimensional data. One of the main goals of my extension is to seamlessly integrate matrices into the host language. To do this, we first added a new type to the language's type system, which then gets reduced, or "forwards" to a type that already exists in the host language, in this case structs (the underlying representation we have chosen for matrices). This allows the programmer to declare a matrix variable in the same manner they would create any other type of variable without worrying about the underlying implementation of a matrix object. Next, we overloaded the arithmetic operators, so that matrix arithmetic can be done by using the basic operators such as +, -, *, and /. This allows us to abstract away the details of complicated matrix arithmetic algorithms. Matrix multiplication is a perfect example of this. To multiply two matrices together, is a fairly complex task, but we are able to generate the code to do this for the programmer, as well as perform optimizations such as parallelizing the loops used to compute the result. In addition to this, we have also overloaded the array reference and assignment operators, so that the programmer can also reference multiple elements, as well as assign to multiple elements of a matrix in one line of code. This extension effectively increases programmer efficiency as they do not have to worry about mundane yet complex arithmetic and assignment operations and focus on greater problems instead. Another advantage of using our matrix extension is that the source code can be optimized to run in parallel, effectively increasing its performance.

As of now, we currently have a functioning language extension that supports just about everything one would be able to do in MATLAB – a costly closed-source programming language. I am in the process of implementing domain specific optimizations such as parallelization. My work will serve as a case study for a broader project led by Professor Van Wyk. In Spring 2012, I presented my results at the University of Minnesota Undergraduate Research Symposium. This was an interesting experience because we were not just presenting to computer scientists. It proved to be a rather difficult task to explain what we were working on to the general public who had never been exposed to computer science. I am

now more thoughtful about my communication as I value the importance of communicating my results to the broader community.

A second research opportunity that I am currently working on is in the field of climate data mining, under the guidance of graduate student James Faghmous and his advisor Professor Vipin Kumar. Warming along the equatorial Pacific, known as The El-Niño Southern Oscillation (ENSO), is a phenomenon that has been linked to annual Atlantic hurricane counts [1]. However, the exact pathway by which ENSO influences Atlantic hurricanes is not well documented. Our research aims to formalize the ENSO-Atlantic hurricane relationship using data-driven methods.

When I began working on this project, James had already developed a method that explained the Pacific-Atlantic relationship more accurately than ENSO. Traditionally, climate scientists have expressed the impact of the Pacific Ocean on Atlantic hurricanes by monitoring fixed regions in the ocean and recording their sea surface temperature. James proposed that instead of monitoring the temperature of the Pacific Ocean, one can monitor the location of the warmest region in the equatorial Pacific ocean. Our index, known as Spatial ENSO (S-ENSO) correlated with Atlantic hurricanes much more than traditional ENSO indices. My contributions to this project were to extend S-ENSO and testing its statistical significance.

Once the original S-ENSO was proposed, I went on to experiment with variations of the index to increase its accuracy. Many of these variations involved incorporating additional variables in addition to sea surface temperature. After a few months, I had come up with an index of my own that performed better than the original S-ENSO with an increase from 0.6 to 0.8 correlation between S-ENSO and Atlantic hurricane counts. This version of the index has remained our top performing predictor of Atlantic hurricane activity when we restrict ourselves to the Pacific Ocean. Over the past few months we have been compiling our results and testing the validity and significance of the index. Since we have a relatively limited amount of data (about 30 years) and the auto-correlated nature of climate data (data that are close in space and time tend to be related), we could not use traditional significant tests (*e.g.* t-test) to ensure that the high correlation between S-ENSO and Atlantic hurricane activity is not due to random chance. Instead, I helped implement a variety of experiments that use randomization and cross validation (regression analysis where portions of data are excluded from training) to verify the significance of our results. Our preliminary results have been so promising that we have started collaborating on a publication with world-renown hurricane expert Kerry Emanuel from the Massachusetts Institute of Technology (MIT). This new collaboration has been beneficial as I had to become comfortable compiling and presenting my results to domain experts. Additionally I was able to present a poster on my work at the annual NSF Expeditions Grant Workshop, where I interacted with a broad range of scientists, faculty, and students from over 10 universities.

I have had a great time working on these two projects, and feel that they have stimulated my interest in conducting research in the future. I certainly have a different perspective on conducting research now that I have had some exposure to it, and feel that I am much more prepared than previously in carrying out research at the graduate level.

References

[1] M.C. Bove, J.J. O'Brien, J.B. Elsner, C.W. Landsea, X. Niu. Effect of El Nino on U.S. Landfalling Hurricanes, Revisted. *Bulletin of the American Meteorological Society*, Volume 79, Number 11

[2] A. van Deursen, P. Klint, J. Visser. Domain-Specific Languages. *ACM SIGPLAN Notices Volume 35 Issue 6, June 2000*