My research experience is in natural language processing, machine learning, and astronomy. All of it is in industry or as an undergraduate. These experiences have taught me how to research independently and in groups while preparing me for advanced study.

At Google, I developed the scoring algorithm for a search feature on Picasa Web Albums, Google's photo hosting site. The challenging part, and what differentiates it from other Google searches, is the personal nature of text that users provide with the photos. Since the purpose of the feature is to help anonymous users discover each others work, it was critical to identify parts of text that users find most interesting. So I employed natural language processing to extract content and context from descriptions and assign it higher importance. This was difficult due to the extremely short text that users provide, but by testing multiple algorithms and customizing I was able to produce a good scoring system. In developing the scoring system, I worked closely with two others to understand the capabilities and architecture of Google's search infrastructure. To develop and promote a complete feature, I worked with product management, production engineers, other developers on the team, and a user interface designer. I learned how to do natural language processing on a massive scale and to take responsibility for a feature. Unfortunately, Google is quite secretive about the specifics of any scoring algorithm, another reason why I want to work in academia so my work can be published. Despite this, I am communicating results internally to improve web search and will lecture at MIT on public results as part of a cluster computing course.

In 2006, I interned with the research division of Infosys, India's second largest software outsourcing company. Infosys was looking for ways to automatically reorganize legacy source code to allow easier understanding and adaptation. Working in the Bangalore office, I developed, implemented, experimented with, and reported on one such method. A relatively recent approach used linguistic information such as variable, function, type, and base file names. I examined many natural language models and settled on Latent Dirichlet Analysis (LDA) in part because it supports multiple topics per document. This model is more appropriate for source because code often deals with multiple business topics and modules to accomplish a task. As an example, my method automatically detected that the Apache web server's HTTPS logging code is one third HTTPS and two thirds logging. Significant customization and experimentation was required to attain this accuracy.

Names found in source code have a number of differences from text. One such difference is the propensity of programmers to abbreviate names, a special case of synonymy. LDA is designed to deal with synonymy by placing words that occur in similar contexts into the same topic. However, I observed many abbreviations were not being detected, most likely due to the small amount of text per file depriving LDA of context. I concluded that acronyms, letter deletion, and well-known abbreviations cover the majority of abbreviations. Introducing probable expansions of abbreviations helped LDA understand when different sets of words belong to the same topic. This visibly improved correspondence of topics with the business concepts that I expected to find. With enhancements such as this, my method was able to attain 46.4% similarity with the existing modularization of Postgresql and 83% similarity with Apache's directory structure. Though I was primarily responsible for this approach, my work benefited greatly from discussions with other members of the research group. We helped analyze each others data, discussed approaches to use, and cooperated on common problems such as term weighting and example source code. The experience introduced me to natural language processing, software engineering research, and India.

Starting with a 2005 Summer Undergraduate Research Fellowship, I worked on manifold learning in Netlab, Caltech's networking laboratory known for setting Internet speed records. Professor Steven Low gave me some problems, one of which was doing something interesting with manifold fitting to data points in a high dimensional space, on which I was to work very independently. It was my first exposure to the field beyond simple forms of regression. Many of the papers discussed in detail how to fit their model and indicators that the fit is succeeding or failing. However, with the notable exception of models expressed probabilistic terms, there was little about how certain the model was with each prediction it made. I set out to fix that for one model, kernel principal component analysis (kPCA) which interested me by naturally extending a linear model to non-linearity. One of the challenges I encountered was tractability: kPCA models lie in an abstract space whose dimensionality is too high to be tractable. So, like the algorithm itself, I had to phrase everything in terms of dot products accessible through kernel functions. This guided me to think like the makers of the algorithm and to reduce everything to dot products. At the end of the summer, I distributed a paper and presented at Caltech. Professor Low hired me to continue work and I applied the error model to detect unusual traffic on computer networks. This project sparked my interest in machine learning research.

As a 2004 Summer Undergraduate Research Fellow, I worked with the Galaxy Evolution Explorer Project, a NASA sponsored ultraviolet satellite observatory. My goal was to find previously unknown flare stars and asteroids. Flare stars are usually invisible to ultraviolet sensors but occasionally flare to become brighter than an average star. Asteroids appeared as streaks produced as the asteroid moved during the observation. The difficult part of the problem was finding and verifying objects in the presence of detector problems that produce bright spots and streaks. With careful analysis and corroboration, I identified over 84 variable objects. In collaboration with astronomers at Caltech and Berkeley, we presented the objects in two posters[3, 1] and a paper[2]. It taught me about astronomy and verification in a research setting.

Experience has given me passion and a background for research at an advanced level. The right place for me to do this is a doctoral program where I will be able to study, think, collaborate, and report in depth. The NSF Graduate Research Fellowship will allow me the freedom to choose problems of significant benefit to society.

Publications

[1] Browne, S., Wheatley, J., Welsh, B., Seibert, M., Heafield, K., Rich, R., and the GALEX Science Team. RR Lyrae stars in the far ultraviolet: GALEX observations compared with theoretical predictions. In *Bulletin of the American Astronomical Society, Poster Sessions* (2006), vol. 37.

[2] Welsh, B., Wheatley, J., Heafield, K., and Seibert, M. The GALEX ultraviolet variability catalog. *The Astronomical Journal 130* (2005), 825–831.

[3] Welsh, B., Wheatley, J., Heafield, K., Seibert, M., Browne, S., and the GALEX Science Team. The flaring UV sky. In *Bulletin of the American Astronomical Society, Poster Sessions* (2005), vol. 36.