I want to pursue a PhD in Computer Science with a focus onNatural Language Processing (NLP) and computational linguistics. I am currently doing research with both Capital One and Monitaur where I work on information extraction and explainability in machine learning respectively. During my final two years of my Bachelor’s I conducted research in NLP on the stylometric analysis of language and modeling human behavior through language. I also researched topic modeling and sentiment analysis as a data science intern at Dell to create time-series analyses of public opinions of their consumer product lines. Having worked done two years in both research and industry, I can confidently say that I want to pursue research and I am excited to now be applying for a PhD. My main goal is to take on problems in NLP from the perspective of a computer scientist, a linguist, and a cognitive scientist; I want to become a full-time researcher and create meaningfully different outcomes in people's lives through the use of computational models of human language.

During my undergraduate degree at Stony Brook University, I dedicated myself to work with the intersection between Natural Language Processing and Cognitive Science. After two years in the Machine Learning Club and Professor Gregory Zelinsky’s cognitive science research lab, I was ready to tackle my own independent research. I joined Professor Ritwik Banerjee for two year long projects as a two-person research team. Our first research topic was on demonstrating how online predators use language to abuse their victims. This research gave insights into how systems could be built to detect and preemptively stop online predation in chat rooms. In our second year we worked on a research paper that explored how to use linguistic features to learn about the social demographics of writers in the domain of scientific writing. The model I built for this research was able to predict the gender of the authors in an ACL paper with 95% accuracy, based only on the written content of their paper. As one of the less than 10 students who successfully completed the Honors CS program in my year, I proudly presented my second year-long research project as my honor’s thesis. We then submitted our findings in a research paper that is currently in the review stage for EMNLP. I am very fond of the work I did, and I am happy to say that Ritwik and I are now close friends.

After completing my B.S. I joined industry to learn how machine learning is used in a production setting. Today I work as a senior data engineer at Capital One with my team, where we use machine learning in all steps of our pipeline to make credit limit decisions based on credit card statements. My contributions to the team center around conducting research into image processing and natural language processing pipelines and then implementing them. I replaced our previous vendor Optical Character Recognition (OCR) system with an internally built platform that uses Google’s open-source OCR tool Tesseract to create transcripts of scanned documents with over twice the accuracy as before. After improving our transcriptions I applied information extraction techniques from NLP to automate the process of extracting data from our input that is fed directly into financial algorithms for determining credit limits. The recognition and positive feedback we received for this work was what allowed our team to be given the bandwidth to become an enterprise platform research team for all things in the OCR and information extraction domain. Ultimately, these projects reinforced my desire to do research towards solving meaningful and complex problems.

This led me to begin working for Monitaur part-time as their first and now lead research data scientist. The CTO Andrew Clark and I had previously met through a series of data science conferences. We bonded over wanting to see a more accountable and less hyperbolic approach towards the machine learning industry research. Our goal at Monitaur is to remove the black box ambiguity around models used in medical and financial machine learning so that they can be used in professional settings where they are normally banned for being unexplainable. The work I have done on the Monitaur platform has made it possible to generate explanatory high-precision rules or anchors (*Anchors: High-Precision Model-Agnostic Explanations[[1]](#footnote-0)*) for a variety of machine learning models. One such model I built and anchored is used for making medical diagnoses with convolutional neural networks (*CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning[[2]](#footnote-1)*), which validated that my process scaled up to complex and intensive models. The work I do with Monitaur is documented towards research papers in AI explainability for publication.

After reading about the research done by Professor Yejin Choi of the Allen Institute for Artificial on fake news generation and detection I was resolved to apply to the University of Washington. In *Defending Against Neural Fake News* Professor Choi’s team built out a system called Grover that not only could generate fake news but more importantly detect it. What intrigued me in particular was using adversarial networks to build both a powerful faker and a detector at the same time. This work could not have been released at a better time as AI Safety specialists have been increasingly worried about the rise of malicious usage of GPT2. I would be interested to see if this research could be expanded by exploring the generation of irregular language and slang that might be present on Twitter or Facebook fake news stories. I would also be interested in doing research with Tim Althoff, Hannaneh Hajishirzi, Noah Smith, Luke Zettlemoyer and working with the Allen School's Natural Language Processing group.

I love doing research honestly and truly, and I would feel unfulfilled if I could not do more. I really care about NLP and computational linguistics and work hard to keep up with papers and press material in the field. Since the day I started doing research with Professor Banerjee I have felt a genuine urge to immerse myself in my research and I would be honored to continue it at the University of Washington.

**Looking for 500-750 words (ctrl + shift + c)** Resources:

* <https://www.cs.umd.edu/grad/writing-statement-of-pupose>
* <https://www.quora.com/How-should-a-statement-of-purpose-for-Masters-in-Data-science-be>
* <http://www.cs.cmu.edu/~harchol/gradschooltalk.pdf>

University of Washington PhD

After reading about the research done by Professor Yejin Choi of the Allen Institute for Artificial on fake news generation and detection I was resolved to apply to the University of Washington. In *Defending Against Neural Fake News* Professor Choi’s team built out a system called Grover that not only could generate fake news but more importantly detect it. What intrigued me in particular was using adversarial networks to build both a powerful faker and a detector at the same time. This work could not have been released at a better time as AI Safety specialists have been increasingly worried about the rise of malicious usage of GPT2. I would be interested to see if this research could be expanded by exploring the generation of irregular language and slang that might be present on Twitter or Facebook fake news stories. I would also be interested in doing research with Tim Althoff, Hannaneh Hajishirzi, Noah Smith, Luke Zettlemoyer and working with the Allen School's Natural Language Processing group.

1. <https://homes.cs.washington.edu/~marcotcr/aaai18.pdf> [↑](#footnote-ref-0)
2. <https://arxiv.org/pdf/1711.05225.pdf> [↑](#footnote-ref-1)