In my years of industry experience shipping machine-learned models to hundreds of millions of users, I have been consciously crafting my career to push the envelope of what applied models can accomplish. I have inched closer to research with every move and have made an impressive impact in applied industry research, but feel I have reached a ceiling in the work I can do without the separation of product and research and without the formal training and mentorship of a PhD program. I am a unique Ph.D. candidate who can perform compelling research immediately relevant to real-world problems. My sojourn to become an academic researcher has not been a direct shot but provides me with unique and invaluable experience in real-world applications.

The summer before my senior year of high school, I managed to sweet-talk my way into a software internship with Microsoft in Mexico City despite having no software experience. I worked on a project exploring public sentiment on twitter for the Mexican government. This project piqued my interest in using language as a proxy for some unknown latent variables.

Inspired by my first higher-level CS class, Prof. Malik Magdon-Ismail's Machine Learning, I took a semester off to work as an ML researcher at Basis Technology. During my 5 months in this lab, I worked on scaling Named Entity Resolution models from English to Chinese, Russian, French, Spanish, and Farsi. By moving to a neural network, scaling the language-specific unsupervised corpora, and countless tweaks, I improved multilingual accuracy from less than 20% to English parity. My work on these systems taught me how tiny details in task formulation and modeling could cause considerable differences in downstream performance.**Moreover, this project ignited a passion for building systems that scale across language, tasks, and domains.**

When I returned to school in the fall, Prof. Stacy Patterson introduced me to formal computer science research. Under her guidance, I worked on my first research project 'Experiments in Inferring Social Networks of Diffusion’[[1]](#endnote-1) in which focused on inferred hidden relationships in social data through the mining of weak links. In this research, I applied well-established virology algorithms to the metadata of images shared on Reddit. This method identified users who bridged disjoined topics and caused viral events. Prof. Patterson proposed this as a compelling master’s topic and encouraged publishing the work formally. At the time, I found myself firmly planted in the ‘applied’ industry camp and was eager to join the workforce. Knowing me better than I know myself, she hypothesized I would thrive in research after I worked for a few years and found my path.

Upon graduating, I joined Microsoft in Redmond and **once I was ramped up and slightly bored with my daily operational responsibilities in software translation, I took the initiative to explore opportunities to increase efficiency.** I hypothesized I could save our translators much time by pre translating content using Neural Machine Translation (NMT.) I began to read every paper and textbook I could get my hands and talked to the plethora of experts at Microsoft Research. Once I had a firm grasp on the subject, I built a pipeline that would take the existing state of the art models and then fine-tune on my task-specific data, which accumulated over the decades. To prove my model’s value, I designed task-specific metrics that measured the effect on translators: edit distance, edit time, BLEU, ROUGE-L, and human annotations for fluency and accuracy. My pipeline scaled to 38 languages and allowed translators to work over 30% faster, which I used to renegotiate our contract resulting in $1,800,000 of savings per year.

**Inspired by this success, I itched to move closer to the research side of things, so I made the jump over to Microsoft AI & Research** to work on Metrics and Applied Science on Bing Core Relevance. In this role, I have created and designed new metrics that track and analyze the ranking model’s performance on relevance, locality, freshness, and authority. I have scaled our human reference system (HRS) to 16 languages, which generate over a million labeled data points a week. This data is used to train and evaluate ranking models for all of Bing. In my most significant project, I overhauled our data creation pipeline to be a streaming system. The system I designed samples user traffic in real-time, generates ranking candidates for the thousands of experiments running, generates an HRS label and provides various metrics that measure relevance all in less than 8 hours. In the previous implementation, the latency was over a year. This system has enabled the broader ranking team to surpass their yearly relevance improvement goal by over 130% in 5 months**. I have learned how vital data is to computer science. When data is well-curated, and metrics are stable, and representative impressive new models can be studied and deployed at a lightning pace.**

**Independent of my day-to-day responsibilities at Microsoft, I have been performing applied research** as MS MARCO[[2]](#endnote-2). Under this banner, I have collaborated to release 8 datasets with strong baseline models that have gotten thousands of downloads, hundreds of official model submissions. In the MSMARCO QnA and NLGEN[[3]](#endnote-3) (80 submissions), I generated the data, designed the evaluation framework, and implemented baseline models. In the MSMARCO Passage Ranking[[4]](#endnote-4) (56 submissions), I repurposed the QnA dataset to create a ranking dataset, built baselines, and built evaluation metrics. This dataset is the focus of the 2019 and 2020(WIP) NIST’s TREC Deep Learning Track[[5]](#endnote-5) and the main element my colleagues, and I have used to teach Deep Learning in search 2019 and 2020(WIP) ACM SIGIR/SIGKDD Africa Summer School on Machine Learning for Data Mining and Search[[6]](#endnote-6). My most recent work in which I designed the metrics and studied system performance, Open Domain Web Keyphrases Extraction Beyond Language Modeling[[7]](#endnote-7), was accepted to EMNLP 2019 as an oral session**. In creating and evangelizing these datasets, I have learned how vital documentation, narrative, and cohesion is to research. Formalized problems with high-quality data empower the research community cohesively focus on improving performance at an astounding pace.**

Parallel to my work at Microsoft, I also started a master's in Computational Linguists at the University of Washington to formalize my understanding of the Fundamentals of Applied statistics for NLP and ground this understanding in traditional Linguistics. The combination of human-generated and statistical methods has taught me how to deeply understand phenomena in language and the importance of strong baselines in research. In a class on Neural Text Representation with Prof. Yan Song, I worked on techniques to distill the knowledge of complex models such as BERT and GPT-2 into more straightforward representations like Glove and Word2Vec. In my work, I have been able to improve word similarity representations by 15% on research benchmarks[[8]](#endnote-8) without increasing training time or inference speed. In my thesis, ‘Progressive Training of Text Representations,’ I am continuing and expanding this research and will be submitted for peer review in 2020.

My experiences in researching, applying, and communicating science has motivated me to become a researcher. I aspire to be a Ph.D. student and later complete a post-doctorate honing my research before I become a professor of computer science with a specialization in learned models. I seek to use computers to discover previously unknown relationships and use my experience and energy to inspire and empower the next generation of students and scientists

Given my foundation in Computer Science from RPI, Computational Linguists from UW, and applied research at Microsoft, I believe I have a solid background that empowers me to thrive as I focus on research. In the coming years, I seek to research the relationships of complex systems using unlabeled data to build efficient and robust representations. If admitted to your program, I would like to focus on model learning efficiency, robustness, and scalability. My experiences have taught me there is a tremendous wealth of labeled and unlabeled data, which I hope to leverage to solve challenging problems like multimodal and multi-item inference. As an ardent follower of the current bleeding-edge research, it's clear to me that current approaches learned models have proven to be wildly successful leveraging massive corpora and extended training. Still, I believe we can improve even more. The current state of the art systems takes weeks to create and leverage enormous corpora, which makes them difficult to recreate and unsuitable for latency bound use cases. I believe there is much work to be done to understand how models can learn more efficiently, and datasets can be made smaller and cleaner. With improvements in learning, we will build a truly intelligent artificial agent.

1. <https://arxiv.org/abs/1910.04277> [↑](#endnote-ref-1)
2. <http://www.msmarco.org/> [↑](#endnote-ref-2)
3. <https://arxiv.org/abs/1611.09268> and <https://github.com/microsoft/MSMARCO-Question-Answering> [↑](#endnote-ref-3)
4. <https://arxiv.org/abs/1611.09268> and <https://github.com/microsoft/MSMARCO-Passage-Ranking> [↑](#endnote-ref-4)
5. <https://trec.nist.gov/tracks.html> and <https://microsoft.github.io/TREC-2019-Deep-Learning/> [↑](#endnote-ref-5)
6. <http://sigir.org/afirm2019/> and <http://sigir.org/afirm2020/> [↑](#endnote-ref-6)
7. <https://github.com/microsoft/OpenKP> and <https://arxiv.org/abs/1911.02671> [↑](#endnote-ref-7)
8. <http://alfonseca.org/eng/research/wordsim353.html> and <https://fh295.github.io/simlex.html> [↑](#endnote-ref-8)