Personal Statement

of Zihua Liu (M.S. Computer Science program applicant for 2018 Fall)

As a student majoring in Computer Science, I am committed to addressing social progress using my professional skills. Especially, I am fascinated with data mining and machine learning techniques since today's world has stepped into the era of big data. Accordingly, improved methods for data mining and machine learning significantly benefit society and save time and resources. Driven by my fascination to these areas, I aspire to use my computer science knowledge to develop effective data mining and machine learning methods to promote efficiency and convenience.

In my sophomore year, a simple project in my Web Data Mining Course inspired me to specialize in these areas. This project aimed to identify the types of flight safety issues involved in a huge number of given aviation safety report documents. The most challenging component of this project was dealing with the huge quantity of English words with different tenses and forms. Although I was new to natural language processing at that time, I introduced the TF-IDF algorithm and word2vec model to build a lexicon before I applied a support vector machine for classification. The classification precision of over 85% gave me a sense of fulfillment given that this project was my first opportunity to solve practical issues. I realized the broad application prospect of data mining and machine learning techniques in the real world. At that moment, my aim was to pursue data mining and machine learning as my working focus. Therefore, I joined the research group in the Software Engineering Institute of Peking University under Professor Ge Li. Accordingly, I delved into data mining and machine learning methods, such as deep learning, pattern recognition and optimization problem solution, among other techniques.

Although my academic and research endeavors in the university strengthened my foundation in data mining and machine learning, I realized that these fields has an inseparable relationship with the real world. Thus, research in these areas could not be separated from reality. In my junior year, I worked as an intern in the Software Analytics Group in Microsoft Research Asia, thereby enabling me to expand my knowledge and experiences in industrial areas.

My internship in Microsoft Research focused on the "Auto Insight" project, the primary goal of which was to establish a research framework for automatic mining and recommendation of various insights from multi-dimensional data. My responsibility was to implement insights for time series, sequential data, and the pivot table recommended in this project. The task was challenging given that this project was for commercial use and adhered to strict standards for response time and calculation accuracy. When I attempted to fit an isotonic regression curve for time series data, the processing time of the traditional algorithm was far beyond the limitation. To address this problem, I traded time with space and presented a prefix isotonic regression algorithm that decreased the response time 50 times compared with the original algorithm. The team pushed this project, which comprised over 20000 lines of code, to the market and achieved success. Moreover, this project has been integrated into the Quick Insights module of Microsoft's Power BI, helping Microsoft demonstrate industry-leading vision and technical strengths in the Business Intelligence market. It has also been embedded into the next version's Office365 to allow more convenient data analysis in Excel, empower Excel's 30 million users to more efficiently analyze their data.

Meanwhile, my internship in Microsoft enabled me to learn how scientific research and industrial products complemented each other. During the embedding process of Auto Insights to Excel, the latter has a function for time series data prediction but the prediction results of Excel were inaccurate when the known data sequences are noisy. After investigating into the core algorithm of the Excel's forecasting compute unit, I figured that Excel uses the entire user data to construct a regression model. Therefore, the prediction performance would be influenced by the noise hidden in training data. This type of prediction algorithms is a common approach, but the academia lacks research on eliminating noise in the prediction process. Thus, I launched a research project that proposed a novel algorithm that imported a generative model to the traditional prediction algorithm inspired by the ability of the generative model to depict data from an overall angle. Consequently, I increased the prediction accuracy by 50% compared with the algorithm in Excel and have submitted my paper to IJCAI 2018. Given my accomplishments in academic research and industrial engineering in Microsoft, I inevitably acquired a substantially broad and realistic perspective of computer science.

As an intern in Microsoft, I obtained profound awareness that solid professional knowledge and excellent programming skills are crucial in my field. Given that my university has provided me with a comprehensive education, I can confidently say that I have developed an extensive understanding of computer systems and architectures. In the Operating System course, I implemented an operating system, such as UNIX OS on Nachos, using advanced data structure and efficient algorithm to complete thread scheduling and synchronization, file system, and other modules. In the Computer Architectures course, I wrote a simulator of RISC-V ISA with C++, using virtual data structures to simulate a computer hardware, such as registers, caches, virtual memory, and program counter. With my comprehensive system knowledge, I have the ability to overcome the challenges and difficulties I meet in my work and optimize the programs and algorithms to save time and space.

In the future, I intend to become a technical leader in a renowned IT company and develop cutting-edge software and products. Moreover, obtaining and utilizing large-scale industrial data will enable me to develop advanced technologies to benefit more people. I will definitely be honored if the world becomes considerably efficient with my contribution.

I am well aware that the computer science education in Columbia University is highly esteemed and the most competitive in the US. I am attracted to expand my understanding in data mining algorithm and machine learning models through your well-designed courses like "Machine Learning for Data Science" and "Cloud Computing and Big Data". In addition, the outstanding career statistics are the most assertive evidence of the program quality, thereby inspiring me to furtherly consolidate my professional knowledge in computer science and enhance my academic abilities to be better equipped in here to fulfill my career target in the future.

Finally, I would like to mention that I am enthusiastic to exchange ideas with academic elites all over the world and I am full of energy to solve the projects and labs in more challenging graduate courses. Despite that the school setting would provide an important geographical and cultural change for an international student such as me, I know this change will be beneficial for me in many ways, academically, linguistically and emotionally. I am confident that I have the determination, intelligence, and strength to succeed in this exciting scholastic experience. Lastly, I am full of hope and expectation that you will give me the privilege of continuing my education in your distinguished institution.