Statement of Purpose

The classic *Algorithms+Data Structures=Programs* by Niklaus Wirth pointed out the fundamental role of data in forming the CS discipline. As a self-motivated software engineering student at Xi'an Jiaotong University (XJTU), I perceive how Data Science involves using efficient data structures to organize, store, and manage data and using data mining and machine-learning models for effective analysis. The data deluge generated by nowadays biomedical systems has the potential to transform the way healthcare is carried out. The application of Data Science in the Biomedical discipline will lead to a better understanding of the mechanisms of disease and ultimately improve the delivery of healthcare to people. However, too few possess the skills needed to use automated analytical tools and cut through the noise to create knowledge from big data. This situation signals the urgent demand for Healthcare Data Science development and drives me to conduct advanced research by applying for an MS program in Healthcare Data Science at USC.

In my previous projects, I have gained a lot of experience and insights in coding and Data Science. For instance, I joined the group of Prof. Wang Weiping—a famous expert in Cybersecurity at Central South University—for summer research. I was assigned to the topic *Intelligent Disposal Technology for Malicious Traffic*, aiming to realize a new method for intelligent disposal of malicious traffic based on knowledge-based decisions. My role is to provide the data support for this technology by building models to extract *Structured Threat Information eXpression* (STIX) entities from the massive unstructured *Cyber Threat Intelligence* (CTI) reports and to store and maintain them in an efficient data structure. For flexibility, scalability, and usability, I chose *knowledge graph*, a powerful Data Science technique for mining information from text, to organize extracted entities. The primary challenge was accurately determining the semantic correlations between distant entities, like one being at the beginning and another at the end of the text. As traditional rule-based techniques for CTI report analysis cannot adequately tackle this problem, I turned to a deep learning model, *Sentence-BERT* (SBERT), and customized it accordingly to achieve the task with 82.8% precision as opposed to 72.1% precision by a non-AI, rule-based algorithm. Through this research, I improved my coding skills in processing datasets to support data-driven systems, and realized the potential for enhancing dataset formation through the use of ML algorithms.

What's more, I also have experience in modeling and analyzing big data using deep learning algorithms. I joined Dr. Peng Zhang's cybersecurity research group at the Chinese Academy of Sciences (CAS) as an intern and investigated URL-based phishing website detection. Since most available datasets of phishing website detection are not sufficiently robust, as most phishing sites have been blocked and inactive, I developed a real-time updating dataset using web crawlers and necessary data processing, including 564,434 latest benign and phishing URLs. During the experiment, I determined that although SOTA models report impressive detection results, their performance is usually accompanied by massive resource requirements and long training and inference times. Therefore, I built an efficient model based on my dataset, with the advantages of parallel CNNs, GRU, and the multi-head attention mechanism. With the URL as input, leveraging parallel CNNs enabled the model to extract local features from various receptive fields and effectively optimize training and detection times. Next, the output of the convolutional layers is put into a GRU layer to shorten the time with efficient contextual information extraction. Finally, a multi-head attention layer was applied for weighted enhancement, followed by a fully connected layer for result output. Consequently, I accomplished a significant speed-up of 34.93% with an accuracy of 98.3%.

I also interned at Chengdu Suncaper Data Company to acquire extensive training in big data technologies, such as building Hadoop clusters in Ubuntu, using Hive for large-scale data storage and management, performing big data processing and analysis using PySpark, and implementing interactive data visualization with Zeppelin. Our team also developed a recommendation feature for a website, in which I applied the K-Nearest Neighbors algorithm (KNN) to support the function of similar friend recommendations, contributing to teamwork progress. At present I am working in Prof. Wei Ke's group as a research assistant at XJTU, endeavoring to lightweight large vision models based on ViT.

In the future, I hope to further my studies with a master's degree in Healthcare Data Science and eventually be a professional data scientist in healthcare industry. I hope to contribute to innovative solutions, drive data-driven decision-making in healthcare, and make meaningful impacts in the healthcare industry. The MS program in Healthcare Data Science at USC offers exciting courses like Integration of Medical Imaging Systems (BME 527), Data Science Professional Practicum (DSCI 560), Machine Learning for Data Science (DSCI 552), and Advanced Biomedical Imaging (BME525). The comprehensive coursework and practice opportunities at MS Healthcare Data Science will prepare me to become a biomedical data scientist professionally. I believe that the exceptional collegial and collaborative environment at the Keck School of Medicine and the Viterbi School will provide a perfect platform to fulfill my objectives.