Statement of Purpose

The classic *Algorithms+Data Structures=Programs* by Niklaus Wirth signified 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 engineering involves using efficient data structures to organize, store, and manage data. With the emergence of *Machine Learning* (ML) algorithms, extensive datasets are utilized to generate data-driven solutions. The reliability of ML models primarily relies on diverse attributes of these datasets, such as diversity, high-quality annotations, balance, etc. Classical datasets in computer vision, such as *COCO* and *ImageNet*, exemplify these attributes. However, countless datasets may not always be accessible, causing challenges like data collection, labeling, data augmentation, and enhancement. These challenges signal the urgent demand for data engineering research and drive me to conduct advanced research by applying for an MS program in Data Engineering at UW–Madison.

In my previous projects, I have gained a lot of experience and insights in coding and data engineering. 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 data engineering through the use of ML algorithms.

While joining Dr. Peng Zhang's cybersecurity research group at the Chinese Academy of Sciences (CAS) as an intern and investigating URL-based phishing website detection, I realized how important a robust and balanced dataset is for building a *Deep Learning* (DL) model with satisfactory performance. During the experiment, I found that most contemporary datasets of phishing website detection are not sufficiently robust, as most phishing sites have been blocked and inactive. Using such datasets for model training may lead to poor robustness and low generalization. Therefore, I developed a real-time updating dataset using web crawlers and necessary data processing, including 564,434 latest benign and phishing URLs. Specifically, I leveraged a web crawling program to crawl the latest 556,305 phishing URLs from PhishTank, a public, community-driven database of phishing websites, and then continuously employed the web crawler and filtered out inaccessible or erroneous URLs, ultimately obtaining 276,239 phishing URLs that met the experimental requirements. Meanwhile, the benign URLs were obtained from Common Crawl, which collects and provides web datasets on a global scale. After data cleaning and processing using PySpark, I retained 288,195 benign URLs. Through my ingenuity, I built an efficient model based on my dataset, with the advantages of parallel CNN, 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%. There is no doubt that without the real-time updating dataset for training, the proposed model can hardly reach high accuracy on the latest test set. Data engineering for robust and stable datasets is the essential prerequisite for AI research.

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.

In the future, I hope to further my studies with a master's degree in Data Engineering and eventually attain a Ph.D. degree. I hope to contribute to innovative solutions, drive data-driven decision-making, and make meaningful impacts in industries. The MS program in Data Engineering at UW–Madison offers exciting courses like Big Data Systems, Advanced Deep Learning, Data Visualization, and Statistical Inference for Data Science. The comprehensive coursework, internships, and projects at this MS program will prepare me to become a data scientist professionally. Finally, the exceptional research traditions and ecosystem at this MS program provide a perfect platform to fulfill my professional objectives.