**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 Software Engineering student at *Xi'an Jiaotong University* (SJTU), 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 like COCO and *ImageNet* exemplify these attributes. However, countless datasets may not always be accessible, adequately labeled, and balanced. These challenges drive me to conduct advanced research by applying for an MS program in Data Engineering at the University of *Wisconsin–Madison* (UW–Madison).

My coding skills were significantly refined in my research projects. For instance, I joined the group of Prof. *Wang Weiping—*a famous expert in Cybersecurity at Central South University—for summer research to create a model for generating a knowledge graph from countless unstructured Cyber Threat Intelligence (CTI) reports to handle malicious traffic intelligently. It implies extracting Structured Threat Information eXpression (STIX) entities from each report and specifying the correlations between various STIX entities, such as STIX Relationship Objects (SROs). The primary challenge was accurately determining SRO between distant entities, like one being in 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.

While joining Dr*. Peng Zhang's* cybersecurity research group at the Chinese Academy of Sciences (CAS) as an intern, I built datasets for Deep Learning (DL) models and investigated URL-based phishing website detection. Based on the research outcomes, my article has been accepted at CONF-SEML 2024. Under Dr. *Zhang's* guidance, I found that most contemporary datasets of phishing website detection are not sufficiently robust, as most 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 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. I 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, 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. 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.

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. I hope to contribute to innovative solutions and enable data-driven decision-making in industries. Finally, the exceptional research traditions and ecosystem at this MS program provide a perfect platform to fulfill my professional objectives.