In this talk, I will highlight several recent advancements in reducing the computational cost of large-scale machine learning models. Notably, these models often require intensive matrix operations, such as in Transformers and graph neural networks. My focus will thus be on leveraging statistical structures to improve the efficiency in training and inference. I will present a novel modification of Nyström methods for attention approximation in transformers, based on the kernel structure of attention; also delve into a geometric perspective of graph coarsening, justifying the usage of graph coarsening in graph-level tasks and providing a simple yet effective algorithm to strengthen previous methods.

Yifan Chen is a fifth-year Ph.D. student in Statistics at the University of Illinois Urbana-Champaign (UIUC). He is broadly interested in developing efficient algorithms for machine learning, encompassing both statistical and deep learning models. He collaborates with Prof. Yun Yang, Prof. Ruoqing Zhu, Prof. Heng Ji, and Prof. Jingrui He during his Ph.D. studies. Before joining UIUC, Yifan earned his B.S. in Statistics from Fudan University in 2018, under the guidance of Prof. Juan Shen.