# Data Science Seminar

Hosted by the Department of Mathematics and Statistics

- Date: Tuesday, November 18, 2025
- Time: 12:15pm – 1:15pm
- Room: Whitney Hall 100E
- Speaker: Dr. Bingxin Zhao (University of Pennsylvania)
- Title: Resampling-based pseudo-training in genomic predictions.


***Abstract***

In this talk, I will present a resampling-based pseudo-training framework for genomic prediction that enables model development using only summary-level data. We show that generating pseudo-training and validation statistics from summary results achieves asymptotic equivalence to conventional training while avoiding the need for individual-level datasets. Simulations and real data applications suggest that pseudo-training performs comparably to standard approaches with large datasets and substantially better when tuning data are limited. We highlight two platforms built on this framework: PennPRS (https://pennprs.org/), a cloud-based computing infrastructure supporting large-scale, no-code polygenic risk score training with purely summary data resources, and GCB-Hub (https://www.gcbhub.org/), which applies pseudo-training to proteome-wide association studies for protein-disease mapping and drug discovery. Together, these advances demonstrate how resampling-based pseudo-training methods can broaden accessibility, scalability, and impact of genomic prediction across diverse biomedical research settings.


Biography of the speaker: Dr. Bingxin Zhao is an Assistant Professor in the Department of Statistics and Data Science at the Wharton School, University of Pennsylvania, with a secondary appointment in Department of Medicine, Perelman School of Medicine. His research focuses broadly on statistics, AI in science and medicine, and inter-organ connections such as heart-brain and eye-brain links (https://www.bingxinzhao.com/).