Abstract:
Understanding how two datasets differ can help us determine whether one dataset under-represents certain sub-populations, and provides insights into how well models will generalize across datasets. Representative points selected by a maximum mean discrepancy (MMD) coreset can provide interpretable summaries of a single dataset, but are not easily compared across datasets. In this talk, I will introduce dependent MMD coresets, a data summarization method for collections of datasets that facilitates comparison of distributions. I will show that dependent MMD coresets are useful for understanding multiple related datasets and understanding model generalization between such datasets.

This is joint work with Jette Henderson (CognitiveScale)