1154-68-1573 **Hassan Rafique*** (hassan-rafique@uiowa.edu), **Tong Wang** and **Qihang Lin**.
Model-Agnostic Linear Competitors - When Interpretable Models Compete and Collaborate with Black-box Models. Preliminary report.

Driven by the increasing need for model interpretability, interpretable models have become strong competitors for black-box models in many real applications. In this paper, we propose a novel type of model where interpretable models compete and collaborate with black-box models. We present the Model-Agnostic Linear Competitors (MALC) for partially interpretable classification. MALC is a hybrid model that uses linear models to locally substitute an (any) black-box model, capturing subspaces that are most likely to be in a class while leaving the rest of the data to the black box. MALC brings together the interpretable power of linear models and good predictive performance of a black-box model. We formulate the training of a MALC model as a convex optimization, where predictive accuracy and transparency (defined as the percentage of data captured by the linear models) balance through a carefully designed objective function, and solve it with the accelerated proximal gradient method. Experiments show that MALC can effectively trade prediction accuracy for transparency and provide an efficient frontier that spans the entire spectrum of transparency. (Received September 16, 2019)