samsi
NSF·Duke·NCSU·UNC

# Deep Learning Program
# Triangle Machine Learning Day
# September 20, 2019

## SPEAKER TITLES/ABSTRACT

**Ted Enamorado**
Princeton University

"Active Learning for Probabilistic Record Linkage"

Integrating information from multiple sources plays a key role in social science research. However, when a unique identifier that unambiguously links records is not available, merging datasets can be a difficult and error-prone endeavor. Probabilistic record linkage (PRL) aims to solve this problem by providing a framework in which common variables between datasets are used as potential identifiers, with the goal of producing a probabilistic estimate for the unobserved matching status across records. In this paper, I propose an active learning algorithm for PRL, which efficiently incorporates human judgment into the process and significantly improves PRL's performance at the cost of manually labelling a small number of records. Using data from local politicians in Brazil, where a unique identifier is available for validation, I find that the proposed method bolsters the overall accuracy of the merging process. In addition, I examine data from a recent vote validation study conducted for the American National Election Studies (ANES), and I show that the proposed method can recover estimates that are indistinguishable from those obtained from a more extensive, expensive, and time-consuming clerical review.