Stanford University
Department of Statistics

Departmental Seminar

*** Extra Seminar ***

4:30pm, Wednesday, January 16, 2019
Sloan Mathematics Center Room 380Y

Refreshments served at 4pm in Sequoia Lounge.

Speaker: Andee Kaplan, *Duke University*

Title: Life After Record Linkage: Tackling the Downstream Task with Error Propagation

Abstract:

Record linkage (entity resolution or de-duplication) is the process of merging noisy databases to remove duplicate entities that often lack a unique identifier. Linking data from multiple databases increases both the size and scope of a dataset, enabling post-processing tasks such as linear regression or capture-recapture to be performed. Any inferential or predictive task performed after linkage can be considered as the *downstream task*. While recent advances have been made to improve flexibility and accuracy of record linkage, there are limitations in the downstream task due to the passage of errors through this two-step process. In this talk, I present a generalized framework for creating a representative dataset post-record linkage for the downstream task, called prototyping. Given the information about the representative records, I explore two downstream tasks: linear regression and binary classification via logistic regression. In addition, I discuss how error propagation occurs in both of these settings. I provide thorough empirical studies for the proposed methodology, and conclude with a discussion of practical insights into my work.