Personalized Recommender System for Conferences

## Background

Academic conferences provide researchers with unparalleled channels to demonstrate their new progress and exchange original ideas with colleagues. However, given that there are so many talks in a conference, researchers often have difficulty deciding which sessions to attend and might miss some valuable talks if they have not thoroughly searched through thousands of them on the website. To help resolve this issue, I and Prof. Jacob Bien developed a recommender system that leverages enormous (200M-record) databases on academic publication to identify the most relevant people of a given researcher based on citation relationships and outputs the corresponding talks given by this cohort. Furthermore, we embedded it into a user-friendly web application that streamlined users' experience of retrieving recommendations, customizing the results, and exporting the schedule to calendar. The work was done in this summer and the app we implemented for Joint Statistical Meetings (JSM) benefited 226 distinct attendees over the week of August 6-11.

## Key Problem

There is a major challenge that we have not rigorously tackled during the summer. Given that author names have different formats across multiple data sources, we need to determine whether a name on the conference website and a similar name in a data record refer to the same identity. For instance, one name in Semantic Scholars database is "R. Tibshirani," which we cannot apparently tell whether it is "Ryan Tibshirani" or "Robert Tibshirani" on the JSM website without other information. A wrong match means we incorporate some citation relations of another author and miss some relations that should be included, which will affect who are on the list of relevant authors and their order of relevance. In addition, there may be multiple records within a database that correspond to the same person, which we should merge into one single record for more accurate citation history. These two tasks are known as *record linkage*, which is an active area of research in the field of statistics.

## Methods

Now our naïve solution is manually designing a set of criteria to determine whether a pair of data records are compatible. If 1) the last names are exactly the same AND 2) the first names are either the same or have the same initial AND 3) they have at least one common co-author, then we identify the pair as a match. However, imposing a solid rule is often less accurate than the approach of "soft" alignment, which regards that the pair has certain probability of being a match and otherwise not a match with the remaining probability. The probabilistic framework is advantageous for its flexibility and higher complexity for fitting complicated real-world data. We will introduce a latent random variable that represents the ground-truth matching status and apply expectation-maximization to find the parameters that maximize the data likelihood. With the parameters, we can calculate the probability of being a match given name and other features and compare with a threshold (for example 0.5) to make a prediction.