# ITEC 657 Data Science

Project Proposal

## Finding the perfect match: A speed dating match classifier

Speed dating is a common way method used by psychologists to study different human behaviors when it comes to courtship between people. There are a lot of interesting facts that have come out of performing studies with this method and whilst some of them are fun factoids, every discovery brings us closer to understanding attraction and mateship.

## Project Goal:

* To predict whether the male and female will be a good match in a speed dating event.

The outcome of this project can use for a matching engine for a dating app or use in future speed dating events to improve the successful rate.

## Data Source and Background

The dataset came from a study by Columbia Business School professors Ray Fisman and Sheena Iyengar. The original study focused on the gender difference in mate selection and our project aim to extend the use of the data to create classification model.

The data was collected from participants from 21 experimental speed dating events from 2002 – 2004. The attendees would have a four minutes ‘first date’ with every other participant of the opposite sex. At the end of their four minutes, participants were asked if they would like to see their date again. And they will be asked to rate their date again. They were also asked to rate their date on six attributes: Attractiveness, Sincerity, Intelligence, Fun, Ambition, and Shared Interests.

The dataset also includes questionnaire data gathered from participants at different points in the process. These fields include: demographics, dating habits, self-perception across key attributes, beliefs on what others find valuable in a mate and lifestyle information.

## Format

The data is in csv file that consist of 8379 observations and 195 variables.

## What work need to do

### Data Cleaning

**Missing values**: There is a lot of missing values in the data due to some fields are not applicable for every person. For example, if a person never goes to college, his or her undergraduate study area will be missing. This created a lot of variables with missing values. We will need to make a decision to fill the missing slot or omit that record.

**Duplicated records**: The setting of the experiment is a participant will have 5-7 “first date” with different partners and the combination create a duplication of information. For example, Participant A will have “4 minutes date” with Participant B, and then move to next “4 minutes” with Participant C. Two records will be stored in the data. Although the result of matching of this two records may be different, the field for selection criteria and demographic of participant A will be duplicated in our data. And we need to tackle this issue in our model building.

### Data Manipulation

**Calculating differences**: We need to create a differences measure so we can calculate how difference of the two participants to contribute to the match or not. Areas need to be calculate includes:

* Differences between two profiles demographically
* Differences between the mating criteria in a survey and the partners end up being match
* Differences between two mating criteria

## Techniques expect to use in the project

We expect to use machine learning techniques that classify binary response such as logistic regression and K-nearest neighbour.

## A project plan

Week 10 Complete Data Cleaning and Manipulation

Week 11 Complete Model building with at least 3 algorithms

Week 12 Run tests and evaluate model