Siddhanth Sabharwal, ssabharwal@ucdavis.edu
Dang Dang, dvqdang@ucdavis.edu
John Nguyen, jolgnuyen@ucdavis.edu
Robin Hothi, rshothi@ucdavis.edu

# STA 141A Final Project: Proposal

## I. Project Description

Project will use a Twitter dataset about users and try to predict their gender. The dataset has the actual gender of users, so we will try to develop a model using a training dataset which is a subset of the entire dataset, and validate our model on the out of sample points and see if our model fits well. Our actual goal will be to try to predict the user's gender based on when they created their profile, their profile description, the number of favorite tweets they have, the color of their profiles, their usernames, their number of retweets, their profile sidebar color, their number of tweets, and their tweet locations. We hope to come up with interesting relationships between gender and these variables, and see which one of them have predictive power.

## II. Data Description

We are using a dataset from kaggle which contains 20,000 rows, each with a username, a random tweet, account profile and image, location, and a link to their tweet and also sidebar color. It comes nicely in a csv file from kaggle. There are 20 different columns that describe the profile. We are trying to predict the gender of the user by using these variables. The dataset also has a gender column which include male, female or just a brand. We are planning on using such column as our test dataset in order to estimate the classification error rate of our model.

## III. Key Questions

A. Did men or women adopt Twitter significantly before the other gender? Did specific regions adopt Twitter significantly before other regions?
B. Can certain words in a Twitter user's description of themselves be used to predict their gender accurately?
C. Does one gender have more retweets or favorites then other genders? Do genders prefer certain profile colors over others?

## IV. Methodology

The methodologies we will use in this project are some kind of discriminant analysis or clustering to able to classify gender. We will also use logistic regression to assign probabilities of each gender to each user. Throughout the project, we plan to use the ggplot2() package to come up with interesting visualizations to express the relationships between variables. Since the dataset has a lot of categorical variables, frequency distributions will also have to be computed.