Phase 1 – Team 3 – Team Proposal

[The problem that motivates the analysis we will conduct]

Members: Hanyu Chen, Yipeng Guo, Luke Hong, Ting-Ann Lu, Yuyan Ma

A) - The problem that motivates us is the possibility of launching updating to dating Apps in the following ways.

- Fitted lover recommendation with Unsupervised Learning (Clustering and Text Analysis)
- Exploration of dating profiles and preferences

The rationale behind this is that each person has their specific status and preference when choosing their date. Either one can be vegetarian hence wants a vegetarian as well, or one can be gay or lesbian who craves a same-sexual couple. What's more, maybe someone who suffers from depressed cannot appreciate more if there were someone shine like mid-noon sun. Sometimes the pairing & recommendation system will not do this kind of specific filter for people, which will share a higher possibility of misunderstanding and waste of time.

However, realistically speaking, the pairing system can be updated somehow. For instance, we can launch clustering system to distinguish gay or lesbian with straight guys, on-drug person with non-drug ones, so on and so for.

With tons of dating Apps in the market nowadays, the following data set has become outstanding for containing extremely specific characteristics of users, how would they define themselves and what kind of date are they looking for in details.

B) - The data set we plan to use and some specific description & what to do with it.

The dataset was explicitly granted and supported by **OkCupid** (the dating App itself). With thousands of new users joining in and completing their personal status together with preference, this dataset allows us to dig deeper than we would imagine. The dataset includes 31 attributes & characteristics of around 60k of users through years.

About the attributes, we some specific ones (columns in the dataset) triggered our interest as listed:

- Basic information: Age, Relationship status, Height, Sex, Sex orientation etc.
- Living habit: Diet, Drink or not, Drugs or not, smokes or not
- Education, Ethnicity and Language
- Income and Job
- Description: One sentence about yourself, one about your hobby, what can bring smile on your face, about your future etc.
- Personal preference: What kind of person are you looking for (text description)

With this detailed and sophisticated dataset, we plan to basically cluster them based on the basic information into separate pools. In this way once people fill in their basic info the pairing system can dive into a deeper and more specific way. Also, we can also conduct more detailed clustering target on their living habit or income. What’s more, **OkCupid** gives users questionnaire about themselves and wanted partners so we can apply language analysis – find significant features based on their descriptions and complete the pairing & recommendation.

C) - Our proposed analysis methodology

Generally, we plan to use the methodology acknowledged from class in the following ways:

- H-Clustering and K-Means Clustering – By applying these two methodologies, we expect to be able to cluster newbie users into separate main user pools. Also, segment them into more specific pools if needed.
- Principal Components Analysis (PCA) – By conducting PCA, we expect to shrink those attributes and waive the multi-collinearity to simplify our analysis.
- Text analysis including Sentiment Analysis, Text Classification and Clustering (maybe☺ we don’t know that exactly for now) – By doing this kind of text analysis, we hopefully expect to analyze the hidden features of existed users and promote the ability of pairing filter completing specific match on personality.