**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 improvements to dating apps in the following ways.

* Fitted match/lover recommendation with unsupervised learning clustering and text analysis)
* Exploration of dating profiles and preferences

The rationale behind this project is that each person has their own specific status and preference when choosing their date. For example, either user can be vegetarian hence wants to date a vegetarian as well, or one can be gay or lesbian who craves a same-sexual couple. Even further, maybe someone who suffers from depression cannot appreciate their match more if there were someone outgoing. Sometimes the pairing & recommendation system will not do this kind of specific filter for people, which will lead to a higher possibility of misunderstanding and waste of time.

However, realistically speaking, the pairing system can be updated. 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 they would define themselves and what kind of date they are looking for in detail.

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 their preferences, this dataset allows us to dig deeper than we would imagine. The dataset includes 31 attributes & characteristics of around 60k users through years.

Regarding the attributes, 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 a 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 cluster them based on the user data 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 conduct more detailed clustering targets on their living habits or income. Furthermore, **OkCupid** gives users a questionnaire about themselves and potential so we can apply analysis through natural language processing – 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 certain attributes and reduce the multicollinearity 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 existing users and promote the ability of pairing filters completing specific matches on personality.