METHODS AND RESULTS SECTION

MODEL

Our model builds on research across sociology, psychology, economics and philosophy on the idea of self-concept and self-presentation. This paper posits that individuals present different 'selves' in different situations (Mead's (1934), Goffman (1975)). Online platforms, such as dating sites, provide an avenue to strategically construct a different 'self' from what may be observed or perceived in real life.

The text section of these profiles provides considerable freedom in the process of this construction. It may prove critical in cases where other aspects of the profile- such as the photo and fixed-response answers may not project an attractive image as desired by the user. For the sake of this study, we assume that the variables where such a dichotomous allocation of profiles into 'attractive' and 'not attractive' would include height, education, weight and ethnicity. Past research (see literature review) suggests that female users accord much greater importance to this written component of men's profiles than men do to women's. Hence, we will limit ourselves to men's profiles.

Individuals who consider themselves to be 'deficient' or somehow lacking in the four listed dimensions may then rely on the use of words to reinforce their strengths or possibly even present a factually untrue version of themselves. A large volume of such essays could then serve as 'documents' for computational methods to discern out major topics and themes that may be seen to emerge across the consolidated corpus from all essays.

Thus, this research seeks to identify and analyze any substantial differences across the categories of the above variables in the proportion of men using these topics. For example, our models posits a different distribution for short and non-short men, or men with only high-school education versus those with more. For initial results, we will only be considering two of the four variables- height and education.

In a related vein, selectivity about mates has typically been observed more in women than men (Finkel et al, 2011). The setting of high expectations could act as a signal of high demand by other women, and the ability to choose. It would also reflect self-awareness. Alternatively, many responses may reflect the woman's preferences- that the man is on board provided that the woman is.

DATA

This study will leverage publicly open and anonymized user profile data for 59,946 users of OkCupid within a 25 mile radius of San Francisco that were extracted with permission (Kim & Escobedo-Land, 2012). To ensure that these were active users, profiles were accepted into dataset only if they had been members as of 06/26/2012, had been active in the previous year, and had at least one photo in their profile. The data set includes "typical user information, lifestyle variables, and text responses to 10 essay questions"

https://github.com/rudeboybert/JSE_OkCupid

Unlike several studies in this domain, this set does not contain details of interactions between users. In this sense, it limits our scope of testable hypotheses, since we cannot explore which profiles ‘triggered’ other users to interact with them, or sustain that interaction. Nonetheless, it does provide a critical window into the realm of self-concept and self-presentation, which is the focus of this paper.

The data were themselves submitted by users in the San Francisco Bay Area for the purposes of the dating app. In this sense, they suffer from self-reporting bias, especially in some of the areas where individuals are likely to veer towards dishonesty in their response so as to improve their dating prospects. This is an inherent limitation of the data. Data can be triangulated as a whole- say for the distribution of heights within the sample vis-a-vis census records. However, they cannot be verified for each individual. In subsequent analysis, we may make adjustments to account for actual height distribution. For now, we will trust the veracity of these claims.

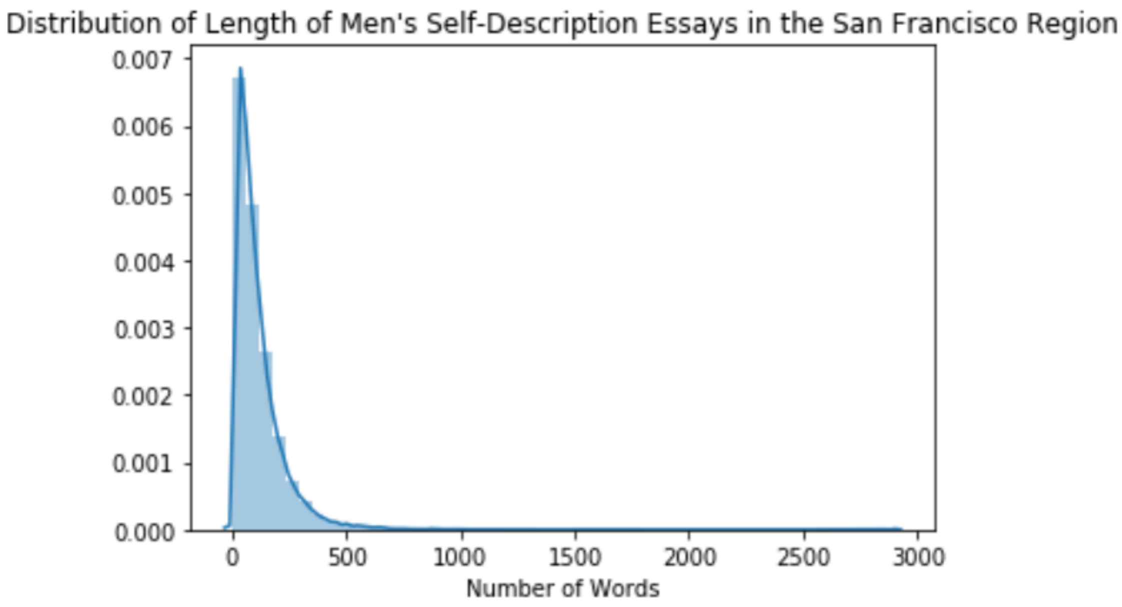
After their submission, the data were then scraped but with anonymization and removal of personally identifiable information for legal and privacy concerns. We dropped all female users, and male users with missing entries for any of the 4 key variables or essays. This reduced the total sample size to 25,281.

This data includes 9 optional essays available in this data. While all of them enquire into different aspects of self-concept, the first one- entitled ‘About Me’ most closely matches the kind of presentation this study is interested in. In real-life situations, it is the first to be displayed, and thus builds the crucial first impression. Subsequent essays may not even be browsed if this first one fails to hold a viewer’s attention. For these reasons, we will only rely on the first essay to link to the theoretical construct of self-concept. Later work will then explore aspects of selectivity through Essay 9’0 ‘You Should Message Me If’

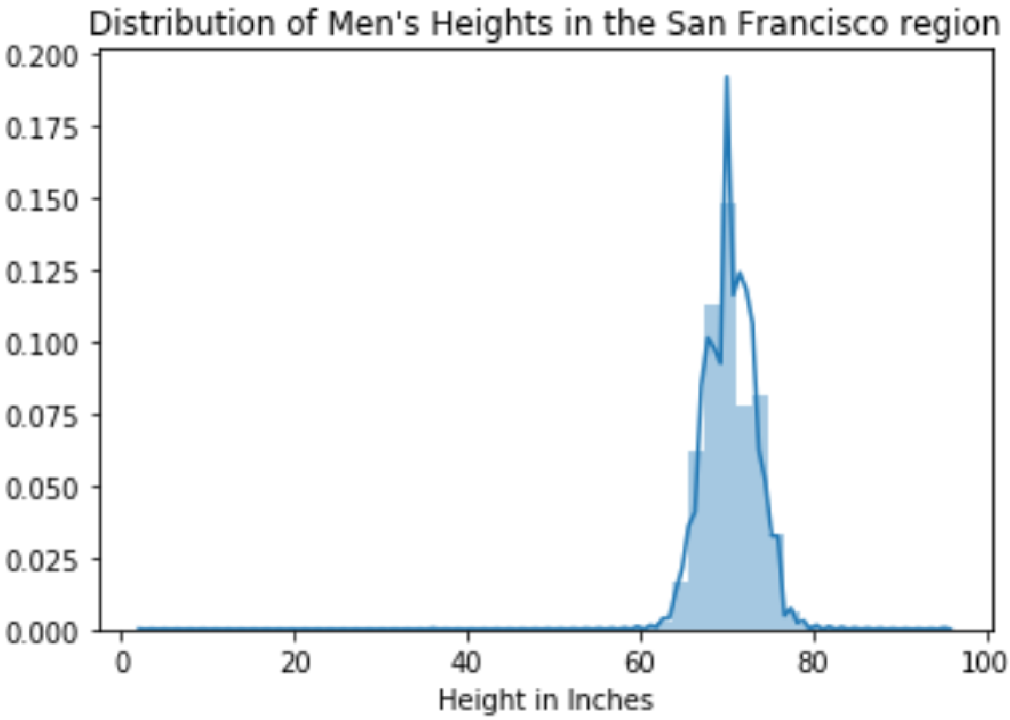
The remaining variables link directly to their stated role in the study- height and education. We will limit the number of categories under each to more of a dichotomy, as mentioned in the model introduction. This can be developed into a more fine-grained analysis at later stages.

SUMMARIZING KEY VARIABLES

We begin with an exploration of the two continuous variables in this model- the essay length.



Height



Key Statistics for Continuous Variables

Statistic	Essay Length (in words)	Height (in inches)
mean	109.251058	70.462165
std	113.561951	3.012987
min	1	3

25%	39	69
50%	78	70
75%	143	72
max	2894	95

This confirms our hunch that height is largely normally distributed, while essay size is heavily skewed.

The quartiles distribution also helps define the cutoff for what will qualify as ‘short’ in this study: the 1st quartile. Thus, male users under 69 inches will form the ‘short’ category, while everyone else will fall in the ‘not short’ category.

METHODS

The key methods harnessed here are topic modelling applied to different populations a tfidf matrix and use NMF to cluster the data points into 25 groups.

Vocabulary includes unigrams, bigrams, and trigrams without redundancies.

Machine learning tasks require numerical inputs. There are several ways to represent text as numerical feature vectors. Features typically correspond to distinct tokens or to sequences of adjacent tokens. A token is a series of characters, such as a word, that is treated as a distinct unit [Bir10].

We now have a corpus through this collection of text documents. This corpus may, in turn, be represented as a matrix of token counts which weighs each term by its absolute frequency. Extremely common terms- such as article (‘a’, the’) in English do not provide any new informations. Thus, instead of raw counts, tokens may be weighted through term frequency-inverse document frequency (tf-idf). These are well defined in Shishido (2016):

“Tf-idf is the product of the term frequency and the inverse document frequency. The term frequency refers to the relative frequency of term t in document d. The inverse document frequency is the log of the total number of documents N to the number of documents that contain term t.

INITIAL RESULTS

CONCLUSION

REFERENCES

Finkel, E. J., & Eastwick, P. W. (2009). Arbitrary social norms influence sex differences in romantic selectivity. *Psychological Science*, *20*(10), 1290-1295.

Goffman, Erving. (1959). The presentation of self in everyday life. Garden City, N.Y. :Doubleday,

Mead, G. H. (1934). *Mind, self and society* (Vol. 111). University of Chicago Press.: Chicago

Shishido, Juan, Jaya Narasimhan, and Matar Haller. "Tell Me Something I Don’t Know: Analyzing OkCupid Profiles." (2016).