# Airbnb Suggested Pricing Project

#### Milestone #2 Team: Stephen Camera-Murray, Himani Garg, and Vijay Thangella TF: Christine Hwang

## Literature Review

### Neighborhood and Price Prediction for San Francisco Airbnb Listings

#### Goal

The goal of this study is to predict price and the neighborhood for San Francisco rentals based on their Airbnb listings. The listings include numeric, text, and image data. Pricing is predicted in two discrete ranges.

#### Dataset

The listing dataset contains: 1) numeric and categorical descriptive information like the property type, number of bedrooms, bathrooms, etc., 2) free-form text description of the property, and 3) property images. Neighborhoods with fewer than 70 listings are dropped and the data is split into train, dev, and test. The study goes on to describe the approach for feature extraction, focusing on text and image analysis.

#### Feature Extraction

For text feature extraction, the study creates a bag of words by stemming words from all listing descriptions and choosing the most frequent 1000 stems. From this list, all stop words and neighborhood words are removed. Frequencies for these words within each property listing are created and the sum of the features is normalized to one. An additional set of word class features is then generated by categorizing each word into one of 9 classes, frequencies for each listing are created, and the sum of word class features is normalized to one. Sentiment features are also generated using the TextBlob package, but details on the method used in the study are light.

Similarly, the visual features follow the bag of words model, by creating a dictionary from the images, create 1000 clusters (the “words”) with K-means clustering, create visual feature vectors for each listing, and normalize.

#### Modeling Approach and Findings

A Support Vector Machine (SVM) with a linear kernel is used for both predictions. The sensitivity parameter, *C*, is tuned by comparing the performance between the train and dev dataset. Overfitting is then reduced using Recursive Feature Elimination (RFE). Once overfitting is corrected, both classifiers, for price and neighborhood, performed better than the baseline, suggesting that listing information may be used for classification.

### Predicting Airbnb Listing Prices with Scikit-Learn and Apache Spark

#### Goal

The goal of this study is to predict prices for San Francisco rentals based on the listing information provided to Airbnb.

#### Dataset

Only numeric and categorical descriptive information like the property type, number of bedrooms, bathrooms, etc. are used from the listing dataset. Data cleansing is performed by: 1) replacing NaNs in count fields with zero, 2) dropping listings with problem data (e.g. price of $0), 3) converting price to a float datatype, and 4) applying one-hot encoding to categorical variables. Finally, on observing that 70% of the dataset is for one bedroom listings, all other bedroom sizes were dropped from the dataset.

#### Modeling Approach and Findings

After checking for multicollinearity with a scatterplot matrix, an ensemble model is built using: linear regression, ridge regression, lasso regression, ElasticNet, Bayesian ridge, and Orthogonal Matching Pursuit. All parameter combinations are attempted and cross-validated with the GridSearchCV function in Scikit-learn to find the parameter combinations with the least error. This CPU-intensive activity was then sped up using the Spark implementation of sklearn. In conclusion, the study visualized the relative importance of each feature on predicting price.

## Team Project Definition

The goal of the team project is to provide guidance to New York City property owners listing their properties on Airbnb. Using the information they provide as predictors, our model will suggest a rental price range based on what we learned from past rentals, taking seasonality into account. The price range we choose depends on how much accuracy we get from our model and how “user-friendly” we can make it (e.g. $300-400 vs. $231-387).

As a stretch goal, we’ll analyze text descriptions of the properties to recommend words and/or phrases that result in higher rental prices. This data is available at the Inside Airbnb website: http://data.insideairbnb.com/united-states/ny/new-york-city/2016-07-02/visualisations/listings.csv

## Approach

Our approach will borrow from both studies. The Stanford study only predicts whether or not the price is above or below the median using an SVM model, but we plan to follow their approach for feature extraction from the text titles and description data. The MapR study takes an iterative ensemble approach only on 1 BRs using several modeling techniques and makes point estimates for price. We’ll take a similar approach for price, but bin the final outcome in a friendly manner.

For the stretch goal, we will, again, follow a similar approach to the Stanford study for text analysis to determine which words and phrases are associated with higher rental prices.