**Business understanding**

For our project we’re examining data from a company called RentHop as part of a Kaggle competition. RentHop’s business strategy relies on making apartment hunting easier because it sorts rental listings via quality. They achieve this by leveraging as much data as possible. However, as one can imagine in the world of big city renting, things like Baits and Switches, Craigslist scams, and inexperienced users happen all the time. So as these electronic postings funnel into the system, RentHop needs a fast sorting algorithm to separate and highlight those results most likely to have the most responses from the ones nobody wants to see. Speed is especially a factor here because in big cities rental properties turn over very quickly, and there are many agents/renters/companies in the real estate market all competing for the same business.

We’re going to try to classify the listing data to find groups of posters that are similar. Because different areas have different acceptable demands when it comes to price and # of rooms, there will probably be a decent amount of classification broken down by interest level and neighborhood. The variables we will be using are # of bathrooms, # of bedrooms, price, neighborhood (extrapolated from longitude and latitude), and interest level.

**Data Understanding**

In this competition, the training dataset included details of some 49,000 apartment listings in New York City, along with ratings of whether each listing generated a High, Medium, or Low Interest Level. The Interest Level is based on the number of inquiries for a particular listing.

The testing dataset included the details of some 75,000 apartment listings, but without the ratings for Interest Level. The goal is to use the information provided in the training dataset to predict the probabilities that a given listing in the testing dataset would receive a High, Medium, or Low Interest Level. This kind of predictive modeling would help RentHop predict the Interest Level of apartment listings.

**Data Preparation**

(I wrote this part based on the codes order from mysigma4)

* Import train and test json files as data frames
* Data Exploration and Take out outliers for bedrooms, bathrooms, price
* Select features based on frequency
* Make attributes from created and photos column
* Create new attributes from price
* Define attributes and dependent variable
* Concatenate latitude and longitude into one column and Import into csv with zip codes of unique latitude and longitude. And create id for unique zip codes and finally merge zip code and its id with train and test set
* Create index for unique building and manager ids, then merge with train and test set
* Define attributes and dependent variable