# Milestone 1

Joo-Wang John Lee, Andrew Patterson, Matthew Schlegel

November 17, 2016

A majority of the time we have spent on this project has been in the way of data cleaning, brainstorming, research, and data management. For this milestone we will talk about several steps we have taken/will take to make running many algorithms on the data easier and several methods we will employ to give our algorithms the best chance in predicting the desired values. Each set has a unique set of problems with which we need to handle to make prediction viable for the given set.

The housing data is relatively manageable in size, containing only 1460 samples. The major problem with this set is the features given for prediction. Many of these features are words and classes which have a less obvious impact on the price of the house compared to values such as the size of the house or lot. What needs to be done is a reduction in the number and complexity of the given features. The major features that seem to be the most useful from our intuition is the LotArea, LandSlope, Neighborhood, BldgType, OverallQual, OverallCond, YearBuilt, YearRemodAdd, BsmtQual, BsmtCond, Functional, YearSold, SaleType, CondSale. These seem to encompass a broad sense of the houses quality, size and location which all could be useful to predicting the price. One other major concern is overfitting to the training data. Because of the large space of features available it would be much better to create a representation that is as simple as possible.

The Outbrain set shows its obstacles upfront with its size and complexity. The data is separated into a number of tables all containing information about a different piece of the overall picture. Some examples of tables is the table describing a users page view, each row having a user id, document id and identifying attributes such as time and location of view. The document table has information on the documents content, publisher and types. And there are several other tables needed to get a better picture of the data. Our first goal with this data is to create a table describing which documents have been viewed by each user. This will give us a good baseline of information. Another idea we may pursue is the use of temporal information to create a hidden Markov model, from this we could hopefully predict trajectories of users and give a likelihood of a user clicking an ad based on their trajectory.

As a team we have made strides in understanding the type of data presented in these contests. We have spent the majority of time looking through the data and brainstorming on ideas on how to creatively predict the values for which we are looking. We are making further efforts in cleaning the data to make the implementation of various prediction algorithms much simpler. Our main ideas focus around using simple techniques on cleaned and clustered data. The housing pricing data has many features that makes cleaning and reduction a priority. We want to use a method such as Ridge regression to pare down what seems to be important features. After this we want to cluster the data by certain aspects of the information and then train smaller models to try and create a more specialized predictor. For the Outbrain set we are planning to employ matrix factorization of the users cross pageviews. From this factorization, we can calculate the angle between users and find similar candidates. We can also use this to cluster users based on pageviews, and find the probability that this group of user has clicked this data point. Something we also want to try is a streaming type algorithm to better handle the size of the data in the database. These are just simple techniques that we plan to start our exploration into modeling and predicting the given datasets.