**Capstone Project Milestone Report**

Gaurav Sureka

Foundations of Data Science

**Introduction**

Recommender systems are backbones of some of the biggest companies like Amazon and Netflix . My aim is to compare different algorithms that existing recommender systems use and benchmark them in recommending the top products for a particular user based on their historical buying patterns. I will also try to combine both content-based and collaborative filtering methods into a hybrid recommender system for predictions on my dataset.

**Dataset Information**

What important fields and information does the data set have?

The important fields are :-

* UserId – This is the ID of the users who had purchased the product from the Amazon site and having this information will help us find similar users in their neighborhood who have similar product preferences based on their score by using User-User collaborative filtering.
* ProductId – This ID will help us predict products to the user based on the pair of items that are closest to each other in terms of score by using Item-Item collaborative filtering.
* Score – This is the metric which is the most important in determining both of the above metrics.
* Summary – This is a summary of the whole review that a customer gave based on which we can find the sentiment of the users for a collection of products.

What are its limitations i.e. what are some questions that you cannot answer with this data set?

* Since it is not given what categories the products belong to we cannot recommend products that are in the same category. Eg :- If someone bought dog food we cannot recommend another brand of dog food.
* Since there are many users who have just rated one product the recommendations might not be at par with users who have rated multiple products.
* We cannot answer which brand in a particular category is doing the best or worst and so which brand should we recommend to the customers more and which should be totally removed from the results.

**Preliminary exploration and initial findings.**

* The number of counts of reviews with a score of 4 and 5 is the highest (78%) which tells me that users are pretty satisfied with the products.
* A lot of users have only rated one product which will pose a challenge when using the recommender algorithms as the matrix will be very sparse.
* We might need to employ dimensionality reduction which is a way to reduce the data and get rid of the noise. This gets rid of the data below a certain threshold which we specify as that data does not add much value to our recommendations.