We plan on using the Million Song Dataset for our project. We have already downloaded and taken a look at a subset of the data, and believe it is large and interesting enough to warrant a project. The Million Song Dataset contains data on approximately one million songs, user play counts on these songs, and many attributes of each song and user. Ultimately, we would like to have a model that can predict what songs users tended to listen to the most.

The data will need a bit of preprocessing. The dataset is extremely large, so we will likely only consider a subset of the data. We have play count data, which is unbounded, unlike ratings are. Therefore we will need to normalize the play counts based on total user play counts and other things. One potential problem is interpreting play counts - e.g. how will we treat play counts of 0, or how will we interpret play counts for a user with many different songs played as opposed to only a few? The dataset is extremely sparse, so any algorithms we use will have to very strongly account for this. Also, due to the subjective nature of the dataset, in the sense that the data is deeply rooted in human actions, it is very difficult to accurately extract any underlying structure. Such problems will be challenging to address, but with proper attention to the dataset we should be able to produce interesting results. We plan on using collaborative filtering, while also exploring some other techniques (e.g. using RNNs alongside CF.)