**Team 11 Project Proposal: Track A**

**Group Members**

Dutch Bultema, Rafael Cintron, Pranathi Manthri, and Murali Prateek Manthri

**Problems to Solve**

1. Find what are the strongest features that drive your music taste.
2. Find songs from another genre that share similar attributes.

**Potential Data Sources**

* Kaggle Spotify Database
  + The Kaggle spotify database has fantastic columns for metadata that include popularity (0-100), danceability (0.0-1.0), energy (0.0-1.0), loudness (in decibels) and more. The challenge with this dataset is that it is of relatively small size with only 125 genres and 114,000 tracks. We have worked with in the past, and it has proven to be fairly clean and functional.
* MusicBrainz
  + MusicBrainz has the most songs in any of the databases, but would be a challenge to work with since it has a unique way to connect due to the size. The metadata specified on the website may not be strong enough indicators.
* Spotify API
  + The Spotify API is an established Python library called Spotify. This database has been specifically designed for research and should have similar metadata to the Kaggle Spotify database. The library contains all of the media which resides on the platform, which would be large enough to pull highly applicable songs.
* Free Music Archive
  + This dataset, although small, only containing 106,574 tracks, is similarly structured to the Spotipy Library, where the metadata can be strong indicators in a model due to its research focused nature.
* Use Group Member’s Personal Spotify data
  + The personal data pulled from Spotify contains the following columns: track, album, and artist. This data would not be used alone, but rather a way to confirm that our model works with real world data and for demonstration.

**Method Procedure**

If we utilize Spotify API, our approach will be to extract audio features (tempo, energy, danceability) and genre information. We'll then use Random Forest to identify the most important features for preference, then implement a neural network with softmax to model songs distributions across genres. To develop recommendations, we’ll first identify songs with similar audio characteristics and then filter songs from different genres. We'll evaluate our system using precision/recall metrics on both the validation playlist data (personal and held out) and implement a simple feedback mechanism for continuous improvement.