# CIS 545: Project Proposal

## Ben Demers, David Feng

1. Responsibilities
   1. David: EDA, Modeling, Next Steps (writeup)
   2. Ben: Cleaning/Wrangling, Introduction/Background, Modeling, Obstacles Faced (writeup)
2. <https://www.kaggle.com/zaheenhamidani/ultimate-spotify-tracks-db> - this is a dataset of different songs in Spotify and [features](https://developer.spotify.com/documentation/web-api/reference/#/operations/get-several-audio-features) associated with the song (instrumentalness, artist, acousticness, etc.).
3. Project Plan
   1. We intend to build a model to determine the danceability of a song. The current dataset assigns a danceability between 0 and 1, and we intend to classify songs by breaking the danceability into 5 classification ranges.
   2. The ultimate objective is to use the given features (genre, tempo, liveliness, etc.) to predict the danceability of a given song.
   3. We will try multiple multiclass classification algorithms such as Gradient Boosting, RandomForest, Logistic Regression, and Support Vector Machine models. We could compare the results of our model with unsupervised k-means clustering of the dataset based on danceability.
4. Project Interestingness
   1. This project is interesting because (we learned literally today that) there are many hidden Spotify metrics that users don’t often see, danceability being one of them. We’d like to know which features in the dataset contribute most to the danceability of a song. We as individuals (or even as a collective) know which songs to which we can dance, but it is difficult to pinpoint what aspects of a song make it so danceable. This project and dataset can help formalize and provide insight into what makes our body wanna move and dance to a song.
5. We anticipate that we will run into some challenges with feature selection to determine which are most important to the model. We also anticipate that we will need to use some sort of regression or sample balancing to avoid overfitting on larger genres, like pop. Finally, choosing and tuning hyperparameters to fit the models to the data could take a lot of time depending on how many models/combinations of parameters we use. To avoid this we will narrow down our choices using the principles we have learned in class and only do testing on those that we think will perform the best.
6. Calvin Hu