**Million Song Project**

Group 3, Milestone 2: EDA

## Project statement

*Based on the project description, state a well-defined question that you’ll address in the project.*

Predict the ‘song hotness’ – or the popularity of a song - based on its features (tempo, danceability etc) and metadata (artist information, year of release etc).

## Project background

This project is based on the Million Song Dataset provided by labROSA at Columbia University.

The majority of record company profits are made from smash hits, and a great deal of money and effort is spent on seeking out the next success. Thus, predicting song hotness would be a useful way to test new songs and predict their commercial potential.

This project will explore the theory that there are certain features shared by popular songs that can make them appealing to a large number of people.

The ‘song hotness’ is one of the features available in the Million Song Dataset, ranging from 0-1, and indicates the popularity of a song. It is assigned algorithmically based on various metrics, including news mentions, play counts, music reviews, radio airtime and Billboard rankings.

## Literature review

*Please provide basic background and relevant references on your project topic. A paragraph description should suffice.*

## Preliminary EDA

<Rahul to provide code>

*It is important that at this stage you are already working with the data for your project, and should demonstrate some basic exploratory results working with your data.*

*Explain your preliminary data exploration. You may include simple visualizations or just a verbal description.*

The full dataset of a million songs is around 270 GB. There is also a subset of 10,000 songs (around 2.5 GB) that we will use for the purposes of developing the prediction approach, as indicated in the project guidelines. Both datasets are provided as compressed HDF5 files.

In the main dataset are audio analysis features for each track, including the target variable of song hotness. There are also an additional 53 variables that include audio features, algorithmic estimations and various tags from Echo Nest and Music Brainz.

The data is accompanied by additional files that provide the following information:

* List of track Echo Nest ID
* List of all artist IDs
* List of all unique artist terms
* List of all unique musicbrainz tags
* List of the tracks for which there is year information
* List of artist latitude and longitude, where available
* Summary file of the whole dataset containing metadata but no audio analysis
* SQLite database containing track metadata
* SQLite database containing links from artist ID to tags
* SQLite database containing similarity among artists

Suggested points to include in EDA write-up:

* Discussion of what attributes are in the data
* Discussion of data cleaning requirements – is the data already clean, how much missing data is there, what steps should be taken to fill gaps?
* A few visualisations showing distribution of data
* A few visualisations showing correlations with song hotness and identification of which features may be particularly important/interesting
* How will success of the project be measured?