**Background & Data:**

Million Song Dataset is a collection of audio features and metadata for a million contemporary popular music track, collected by music intelligence platform the Echo Nest. The data set was created for research purposes under a grant from the National Science Foundation.

*Question of interest: Can we use song feature data to predict the year of a song?*

A subset of the Million Song Dataset including 515,345 songs with 91 attributes (1 year, 12 mean timbre features, 78 covariances) was utilized in the analysis. Since the columns were not labeled, it was nearly impossible to decipher between the means and covariances. Principal components analysis will help to alleviate this problem by finding patterns in the data and reducing the size of the data (i.e. reducing the number of variables).

(insert table of data distribution by decade)

**Machine Learning Method:**

WHY?

HOW?

1. Take smaller random subset of 10,000
2. Split into 7,500 training, 2,500 test set
3. Run PCA on correlation matrix of 90 predictors
4. Used first 8 species

**Findings:**

Assumptions & Limitations:

Scaling:

Sub-training set pulled out the same # of PCs as the bigger training set: 8 PCs of the correlation matrix predict 80%.

**References:**

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.