**Background & Data:**

The 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?*

This is fundamentally a prediction problem: we want to predict the year of release of new songs based on information we already have about other songs. 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 originally prepared by T. Bertin-Mahieux. Since the columns were not labeled, it was impossible to differentiate between the means and covariances. Principal Components Analysis (PCA) 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).

[Netflix challenge]

**Machine Learning Method:**

Principal Components Analysis (PCA):

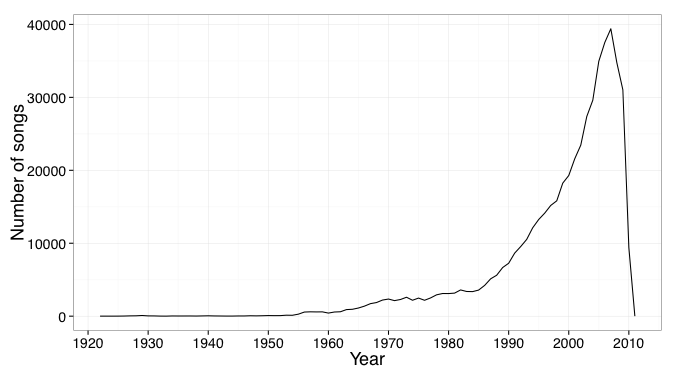
PCA develops linear combinations of variables that explain a desired percent of the overall variation in those variables. The ultimate goal of PCA is dimension reduction. The Pareto Principle, or ‘80/20’ rule was used as a threshold to identify the principle components to be used. The ‘80/20’ rule states that 80% of the variation will be explained by 20% of the variables.

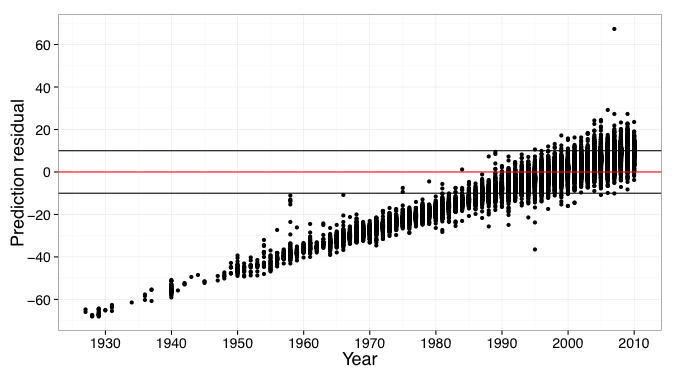
Using the principal components for prediction:

We created a training set (463,715 songs) and test set (51,630 songs), which were recommended by the subset creator. We performed PCA to reduce the dimensionality of our training data from 90 predictors to 8 principal components, because we found that ~80% of the variation was explained by the first 8 principal components. We then performed multiple linear regression on these 8 principal components, treating the year of the song as a continuous response. We also considered grouping the years into 3 categories: pre-1960, 1960-1989 and post-1990, on which we performed multinomial multiple linear regression to try to predict the era of the song.

**Findings:**

The data was heavily skewed in favor of the 1990s and 2000s, which made accurate predictions outside of these decades nearly impossible. The model over predicted the earlier years between 1920 and 1980. As expected, 1990s and 2000s had the best year predictions likely due to the large number of songs for those years.





**Assumptions & Limitations:**

We chose to consider year as a continuous response because we want the best approximation of the year as possible: if we “miss” a 2004 song, we'd like to end up somewhere in the same decade. One of the main concerns with PCA—that it becomes difficult to interpret results—was of little concern as our task was purely prediction, and because we have very little information about the content of our dataset, rendering interpretation difficult.

**Scaling**:

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

PCA scales well to very large datasets, although dimensional reduction may become computationally intensive. Alternative methods: [Tim’s paper here?]

**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.