Proyecto Final: Big Data

Seminario de Ciencias de la Computación

## Miembros del Equipo:

Báez Camacho Jesús Armando

Leyva Contreras José Alberto

Monge Tinoco Manuel Crisólogo

# Enunciado del Problema

A partir de un conjunto de datos de canciones, predecir el año de lanzamiento de una canción en base a las características de audio extraídas. El conjunto de datos se llama YearPredictionMSD, un subconjunto del Million Songs Dataset.

# Descripción del Problema

La predicción del año de lanzamiento se define como la estimación del año en que se lanzó una canción en base a las características de audio que posea. El Dataset Million Songs es un conjunto de datos famoso disponible libre en la web el cual contiene una colección de características de audio y metadata de un millón de canciones populares de la edad actual. Para predecir el año de lanzamiento de la canción nos enfocaremos principalmente en las características de *Timbre Average* y *Timbre Covariance*.

# Descripción del Dataset

El conjunto de datos usado es el Repositorio UCI de Machine Learning YearPredictionMSD, el cual como se mencionó anteriormente, es un subconjunto del Million Song Dataset. La mayoría de las canciones son del género western, con canciones comerciales desde el año 1922 al 2011, con un crecimiento exponencial a partir del año 2000.

Chart, histogram

Description automatically generated

# Características de los Datos

Algunas características son:

* Tamaño: 460 MB
* Formato: Archivo .txt
* Número de Canciones: 515345
* Atributos (Características): 90 (12 -> timbre average, 78 -> timbre covariance)

El conjunto de datos contiene características y etiquetas para cada canción, en donde la etiqueta es el año en el que la canción fue lanzada. Hay 12 timbre-average y 78 timbre-covariance características y se usó el Echo Nest para la segmentación. Se tomaron la media y la covarianza sobre todos los ‘segmentos’, cada segmento siendo descrito en un vector timbre de 12-dimensiones. El Echo Nest API divide la canción en múltiples segmentos de tiempo. Estos segmentos fueron delimitados por comienzo o discontinuidad en la señal del ruido.

# Implementación

Se usará Apache Spark con la librería de Machine Learning (MLlib), donde se aplicarán diversos algoritmos de aprendizaje automático para hacer la predicción como la regresión lineal, random forest y gradient boosted trees.

Se hace también un preprocesado de los datos donde se incluye el rescalamiento y normalización del conjunto de datos

# Referencias

* Conjunto de Datos: <https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD>