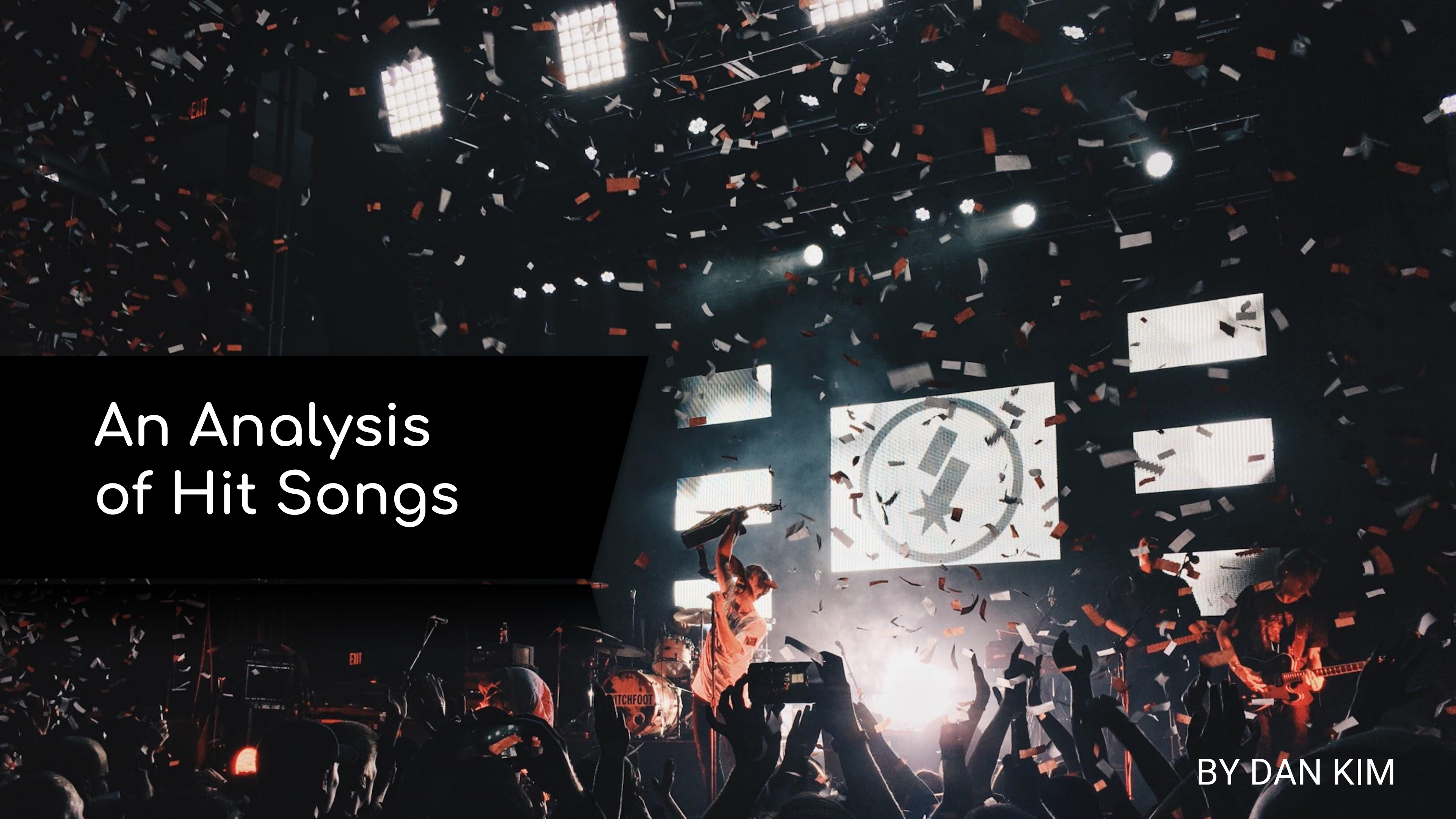
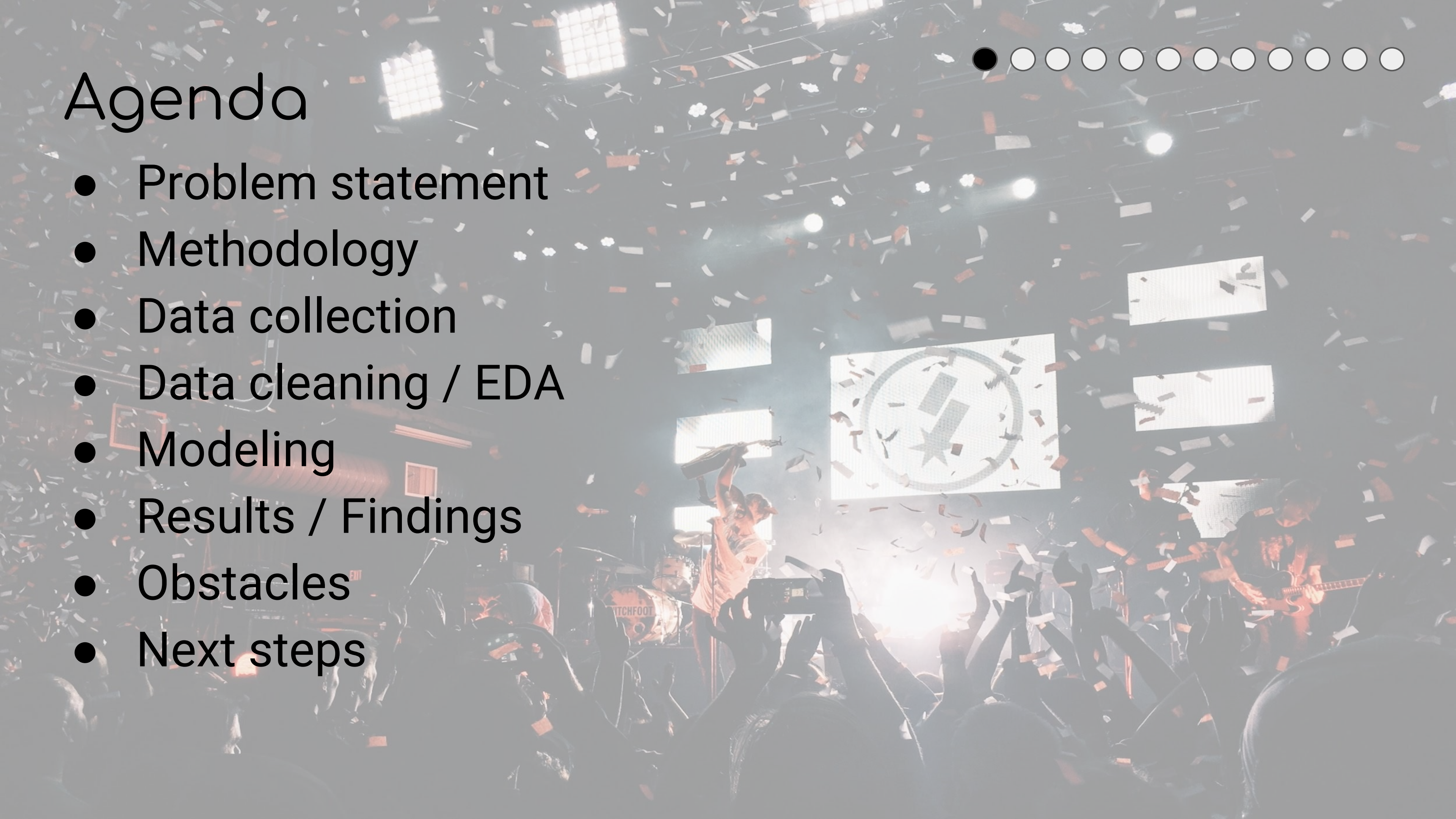


An Analysis of Hit Songs

BY DAN KIM



Agenda

- Problem statement
- Methodology
- Data collection
- Data cleaning / EDA
- Modeling
- Results / Findings
- Obstacles
- Next steps

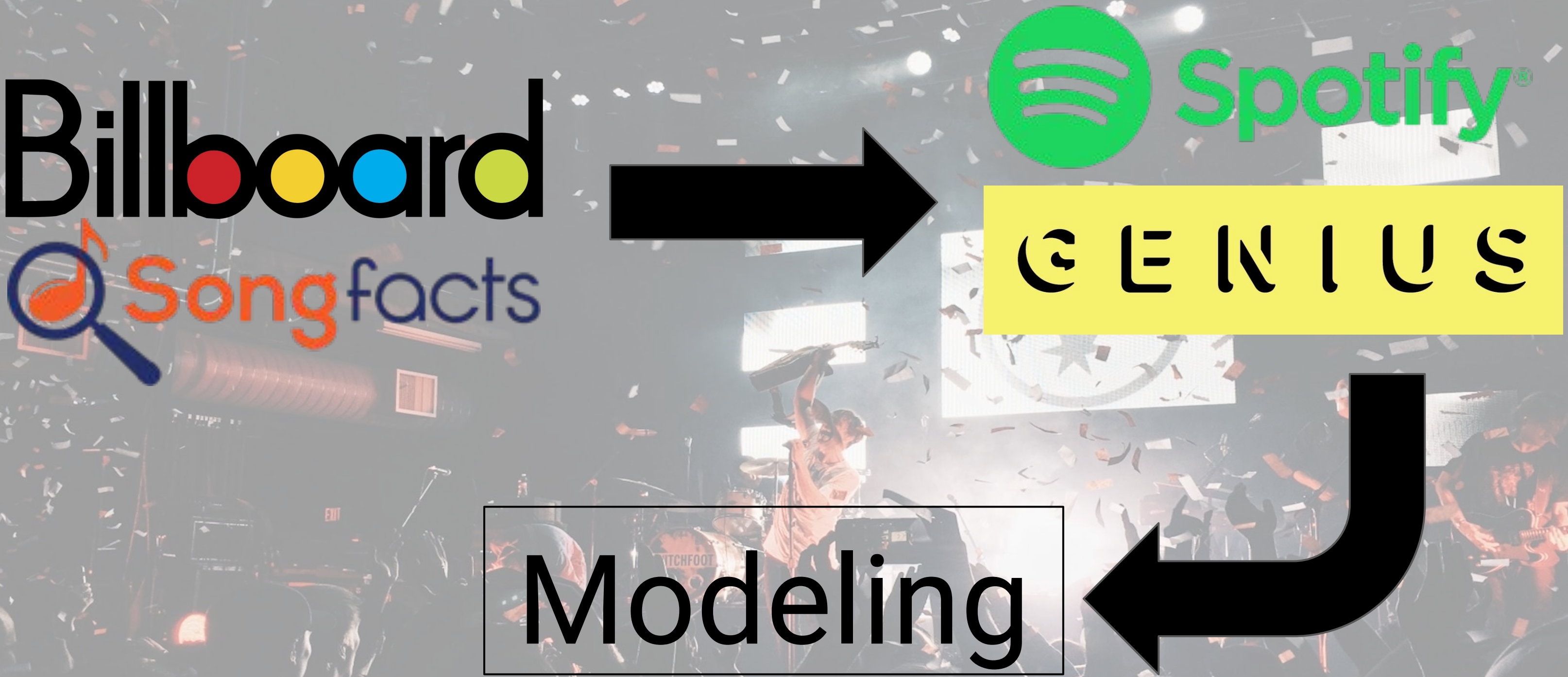
Problem Statement

Which audio features have the most influence on making a hit song?

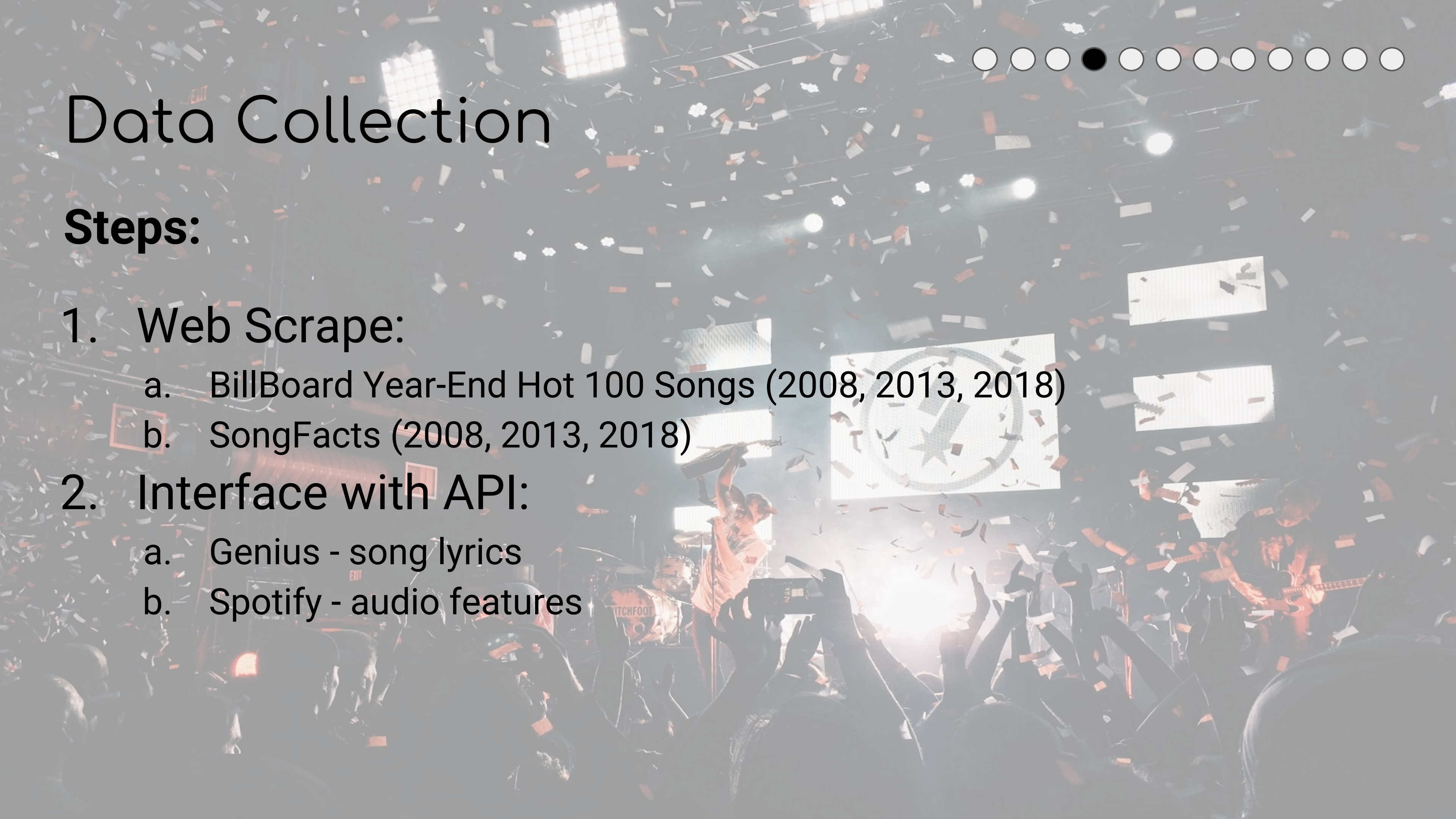
Who can this benefit?

Record Labels, independent artists

Methodology



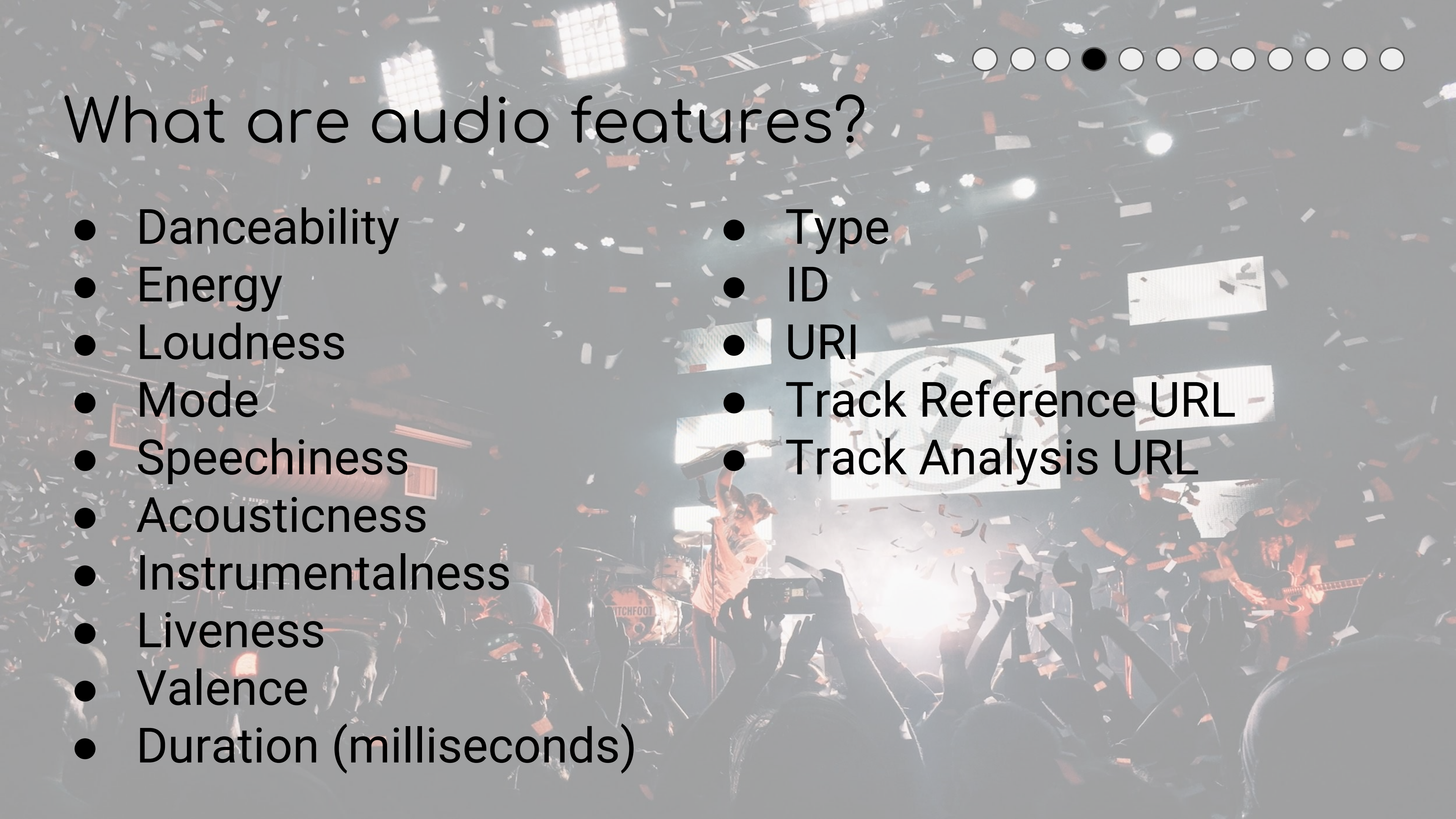
Modeling



Data Collection

Steps:

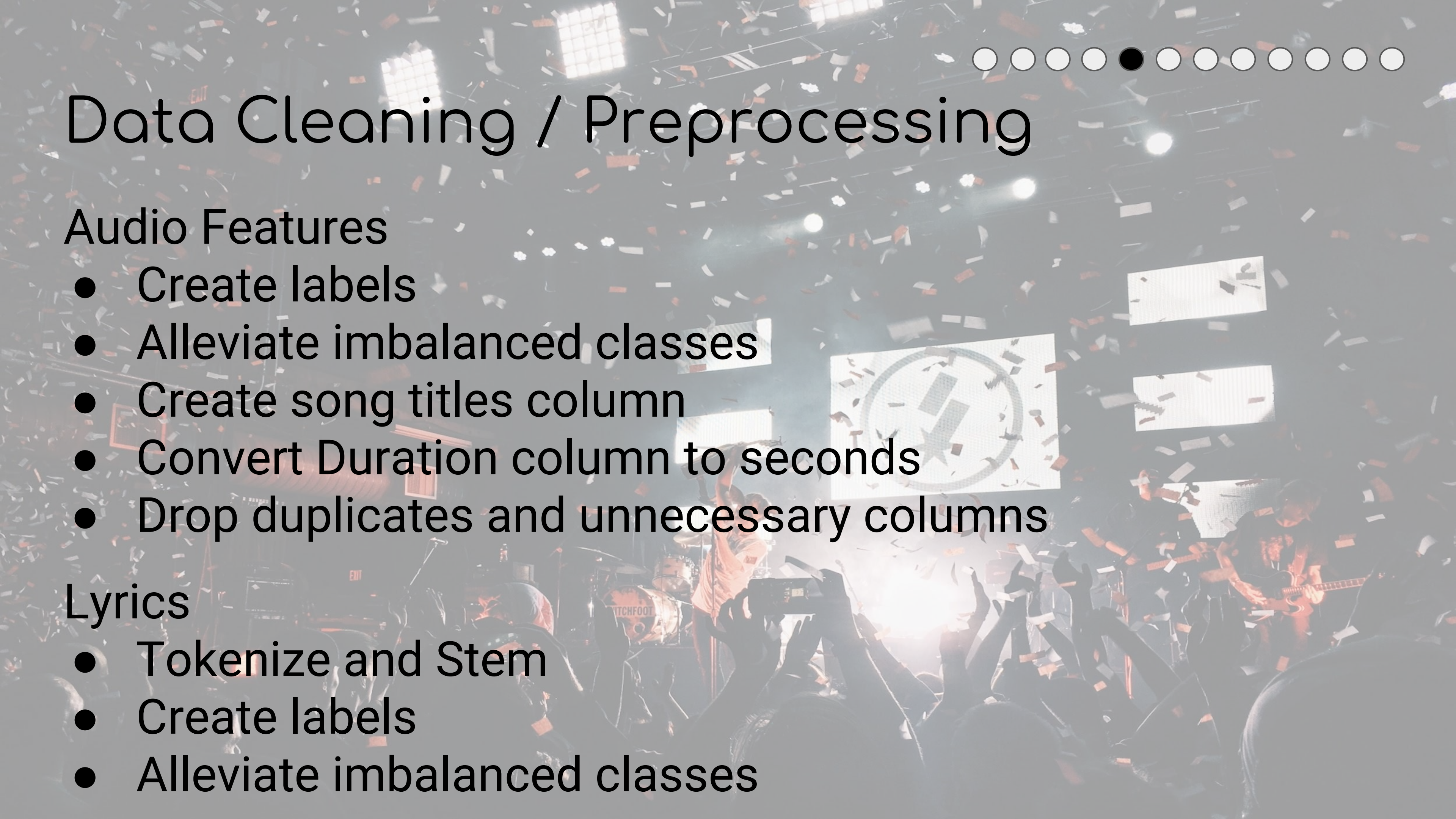
1. Web Scrape:
   a. BillBoard Year-End Hot 100 Songs (2008, 2013, 2018)
   b. SongFacts (2008, 2013, 2018)
2. Interface with API:
   a. Genius - song lyrics
   b. Spotify - audio features



What are audio features?

- Danceability
- Energy
- Loudness
- Mode
- Speechiness
- Acousticness
- Instrumentalness
- Liveness
- Valence
- Duration (milliseconds)

- Type
- ID
- URI
- Track Reference URL
- Track Analysis URL



Data Cleaning / Preprocessing

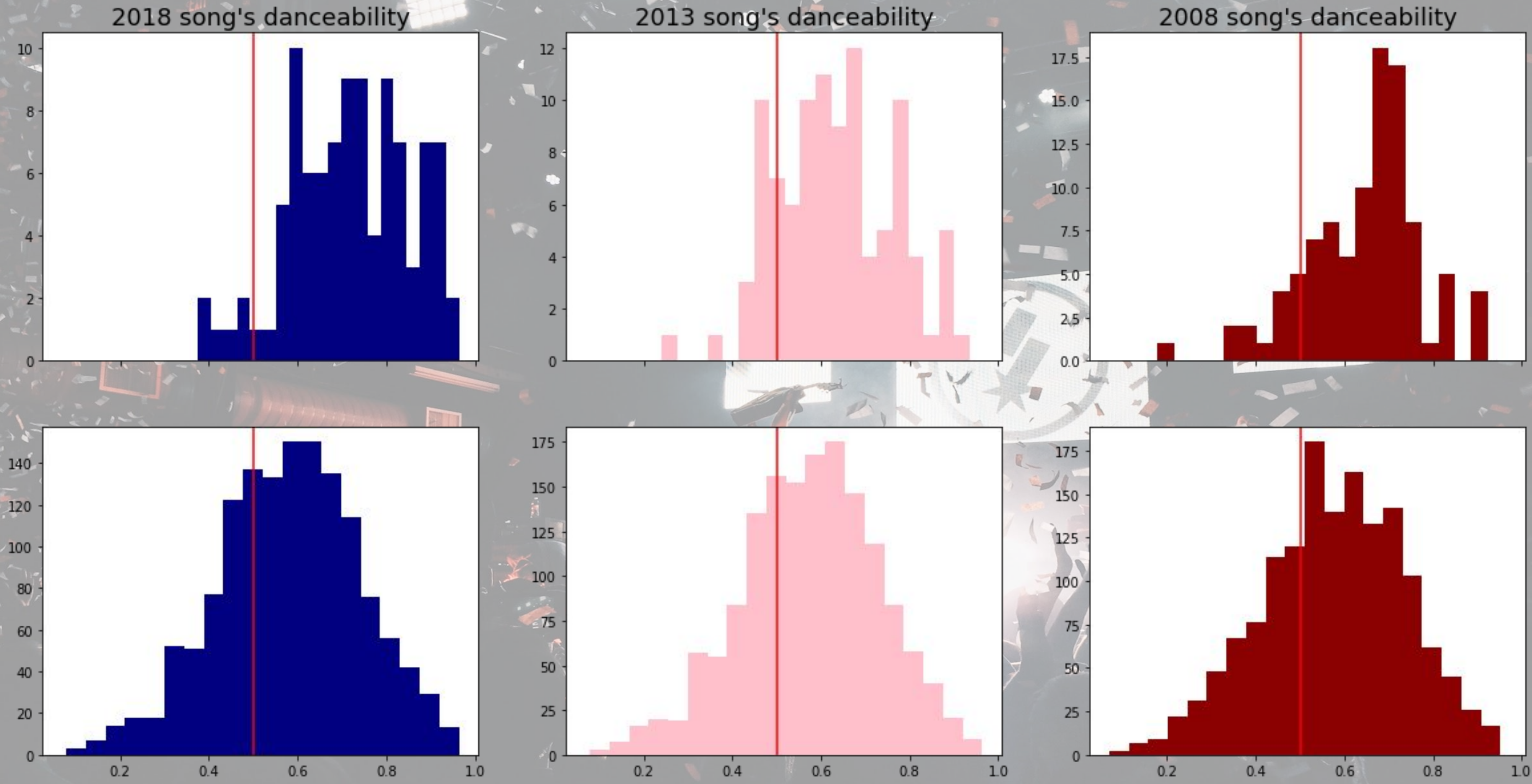
Audio Features

- Create labels
- Alleviate imbalanced classes
- Create song titles column
- Convert Duration column to seconds
- Drop duplicates and unnecessary columns

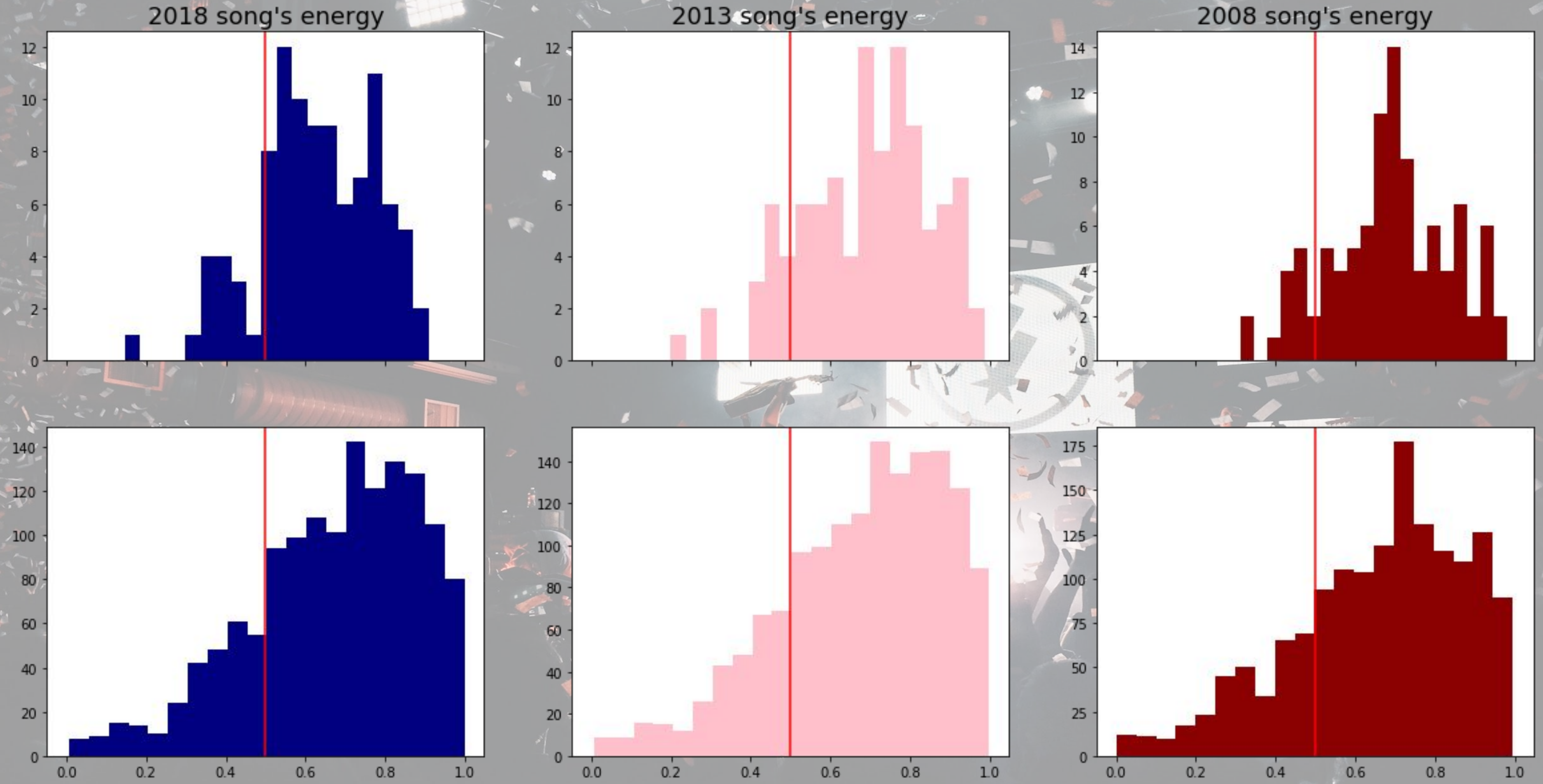
Lyrics

- Tokenize and Stem
- Create labels
- Alleviate imbalanced classes

EDA - danceability



EDA - energy

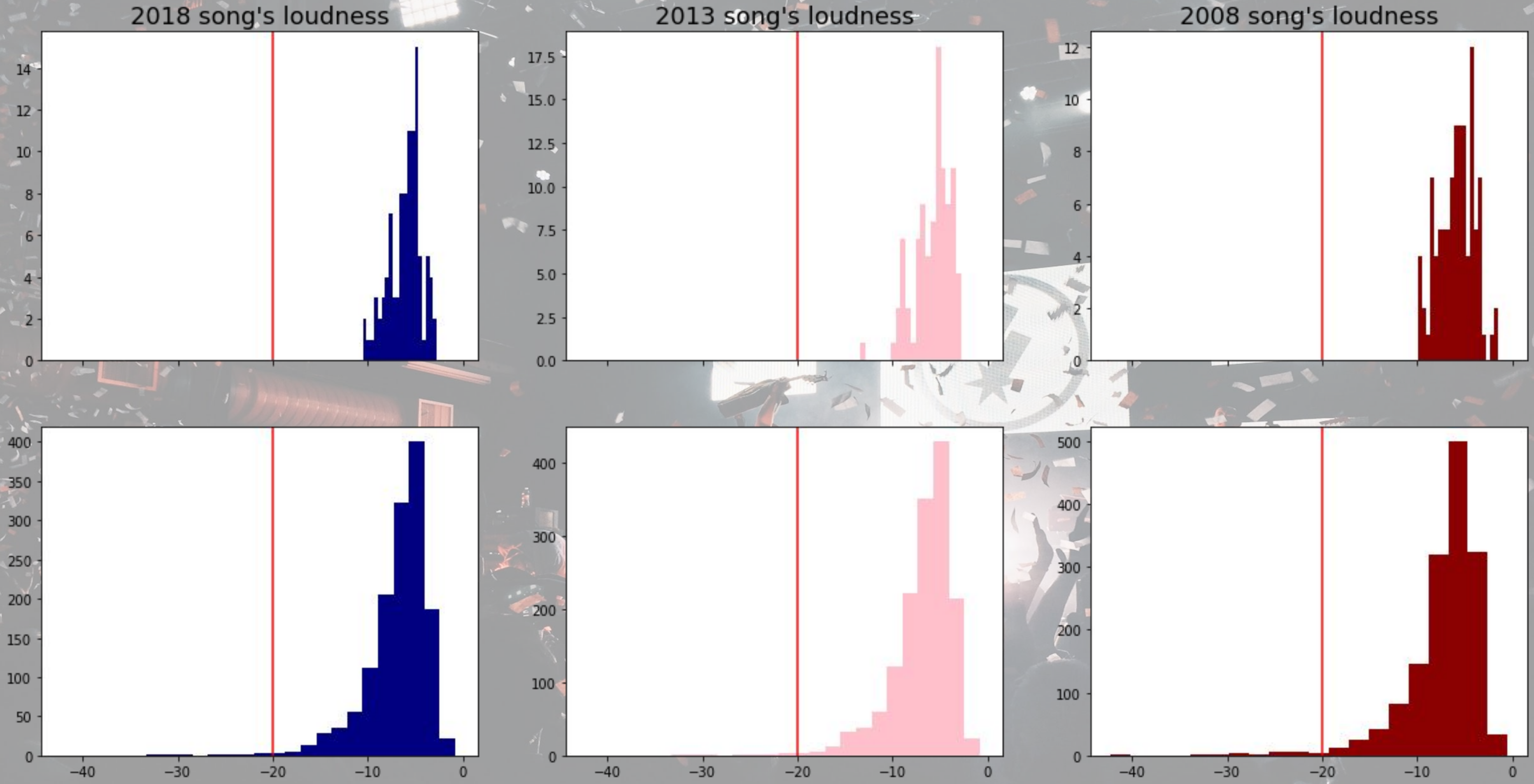


2018 song's energy

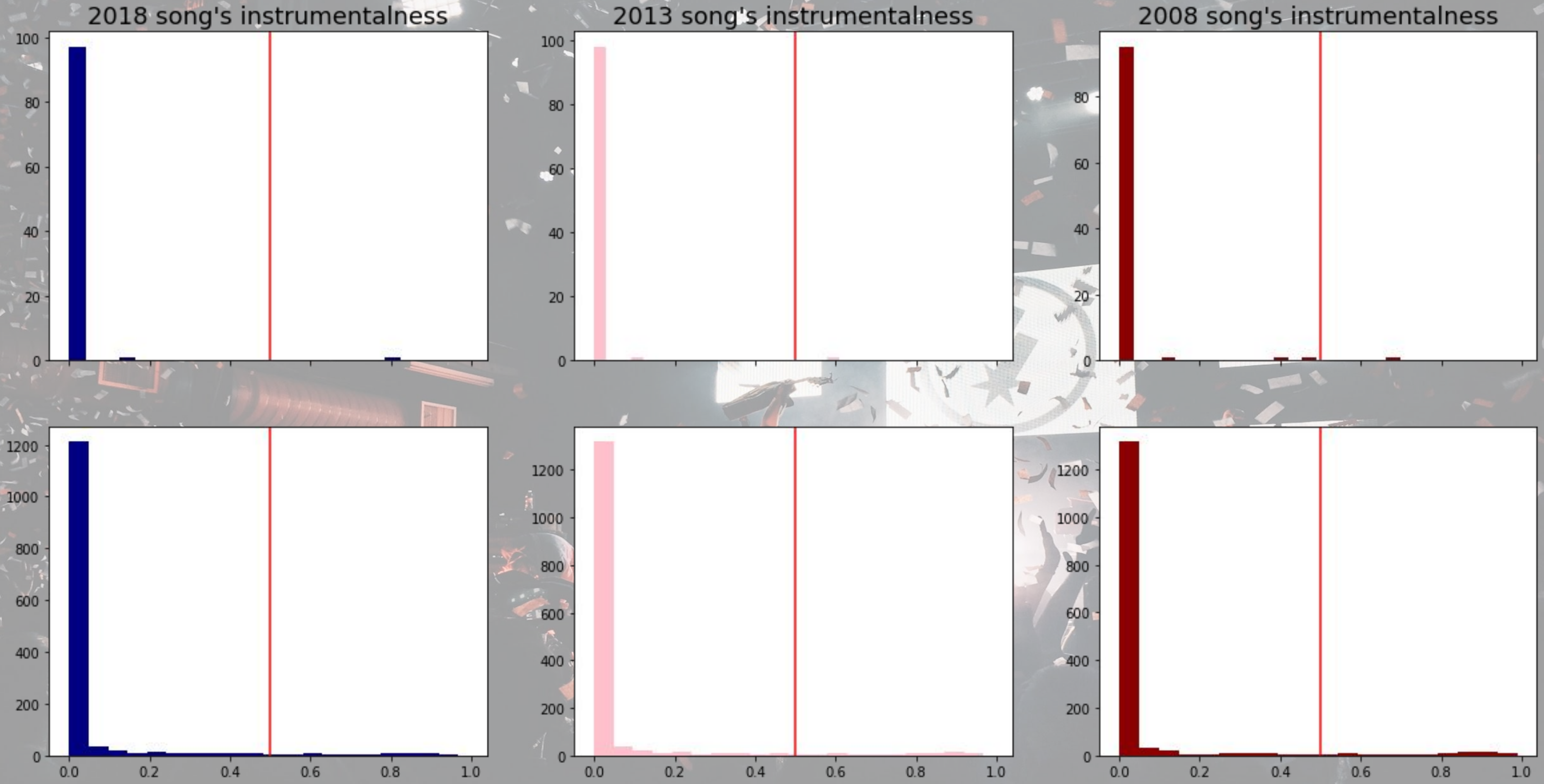
2013 song's energy

2008 song's energy

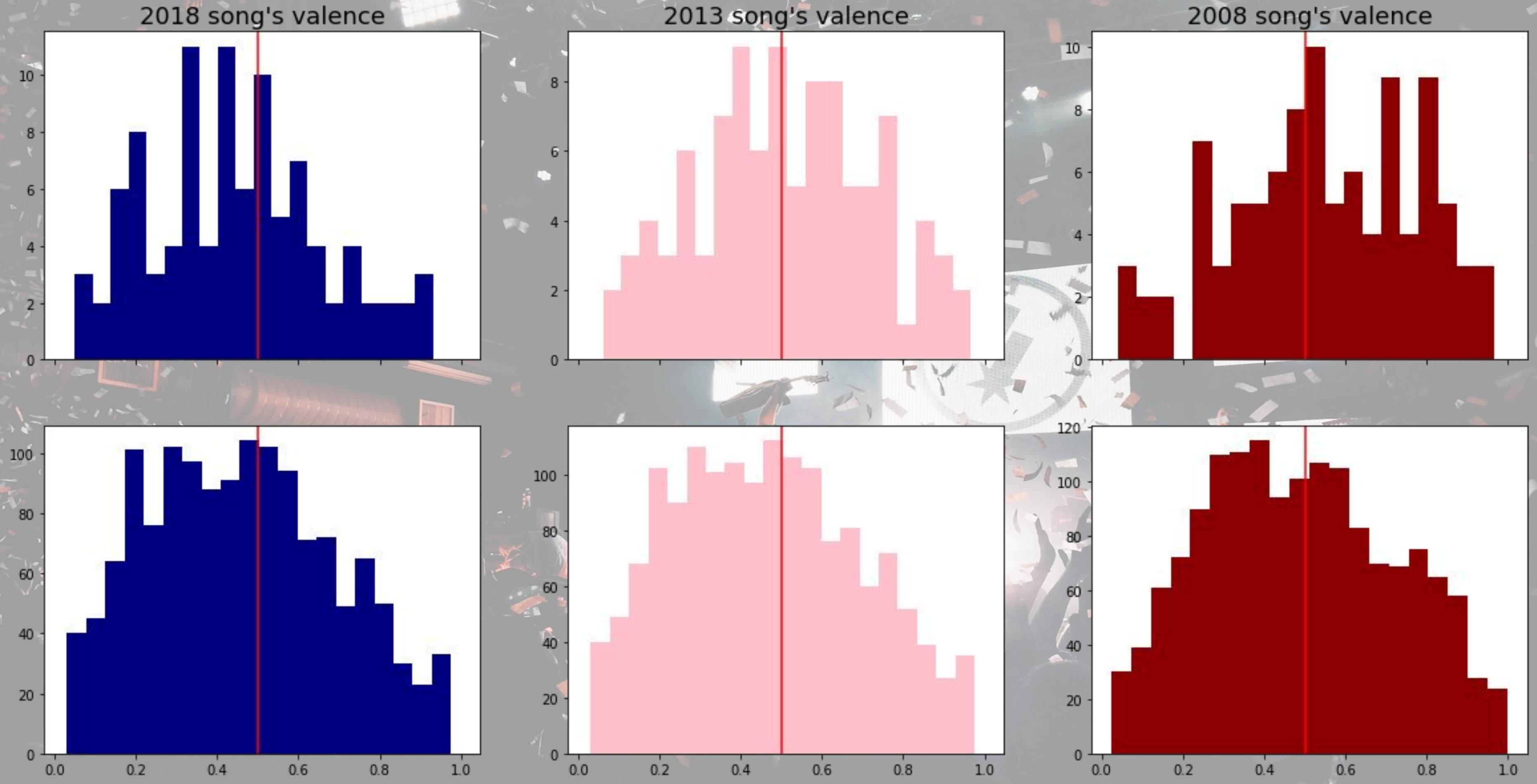
EDA - loudness

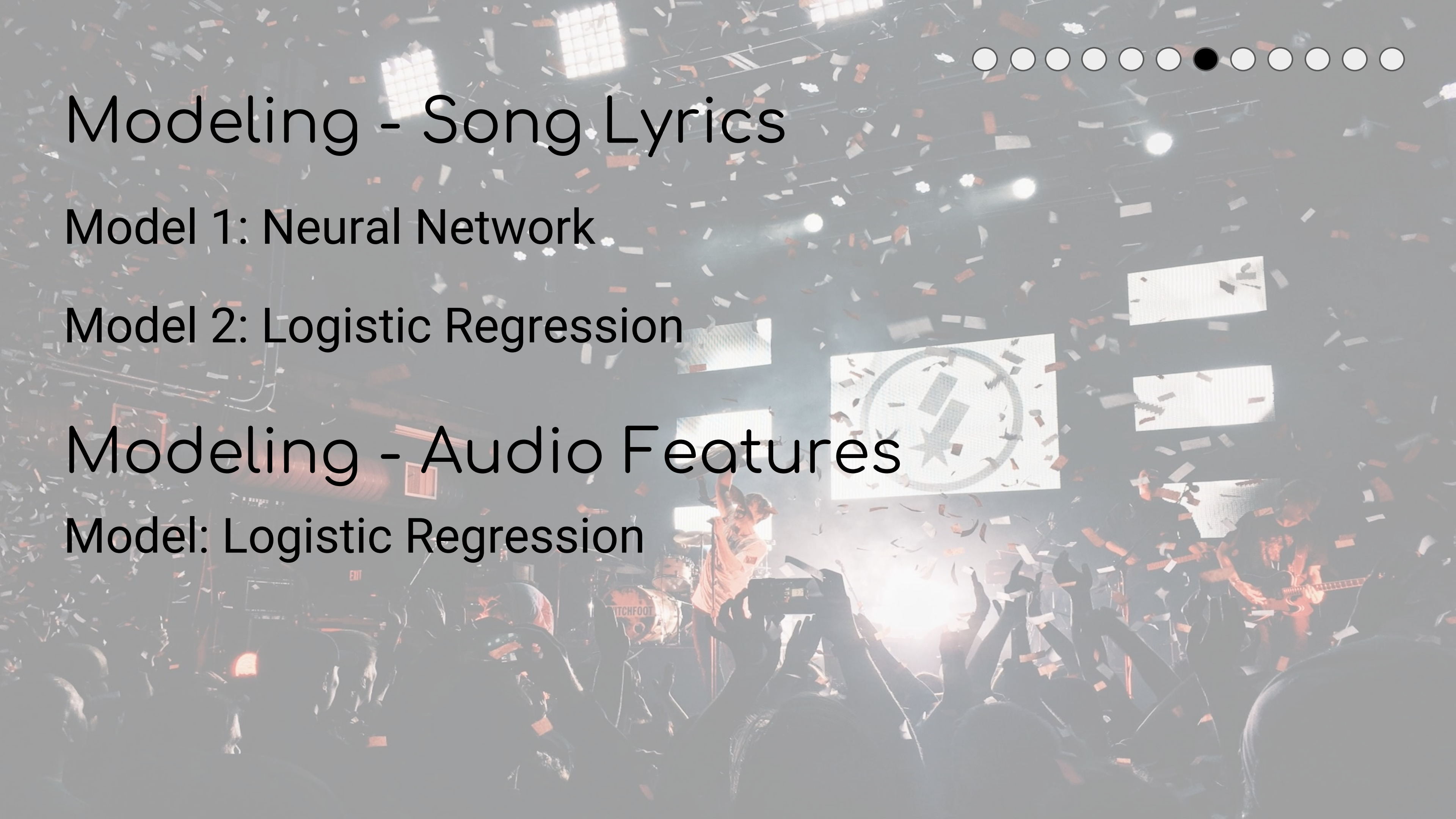


EDA - instrumentalness



EDA - valence


2018 song's valence
2013 song's valence
2008 song's valence
0
2
4
6
8
10
20
40
60
80
100
120
0.0
0.2
0.4
0.6
0.8
1.0



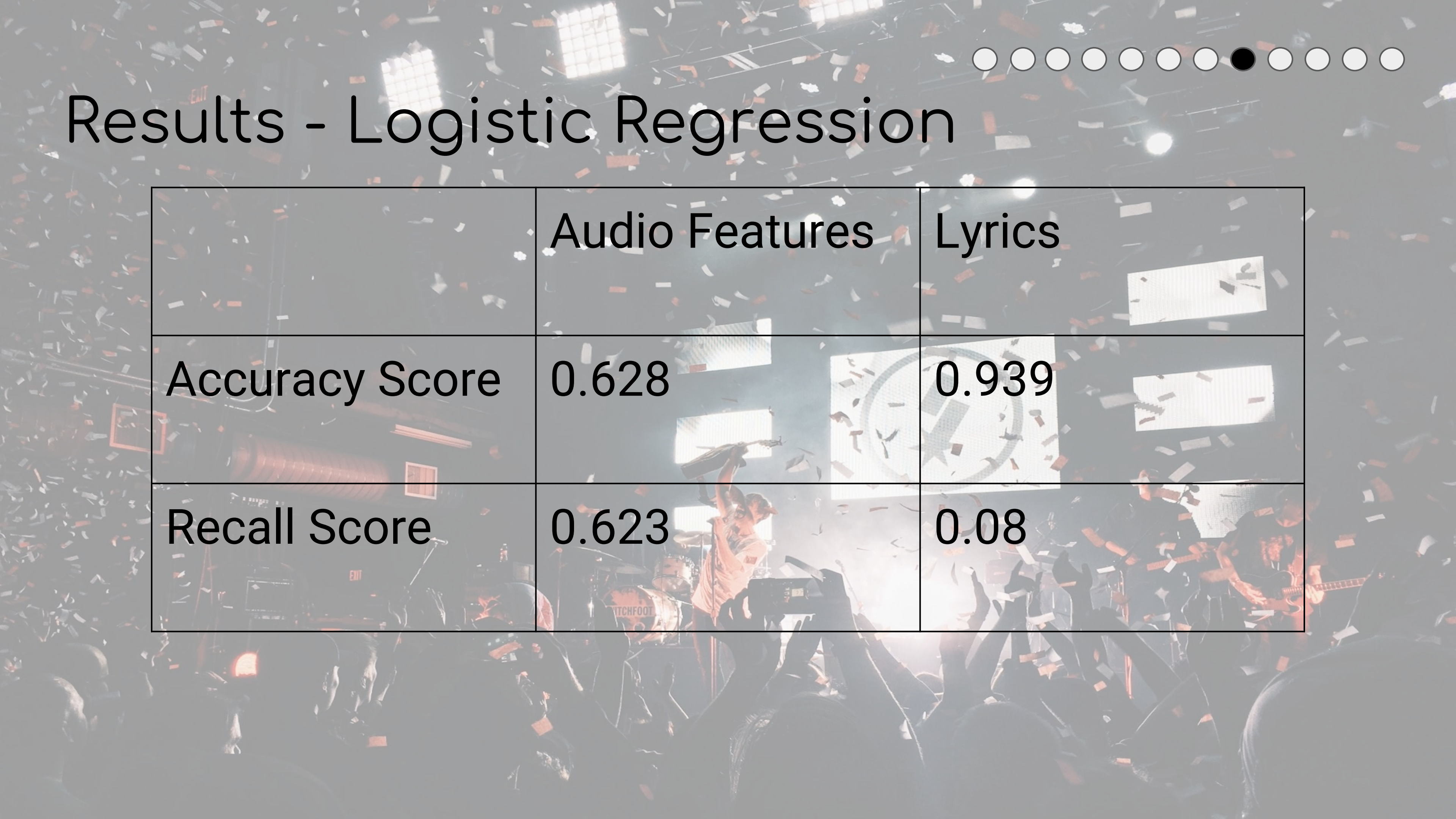
Modeling - Song Lyrics

Model 1: Neural Network

Model 2: Logistic Regression

Modeling - Audio Features

Model: Logistic Regression

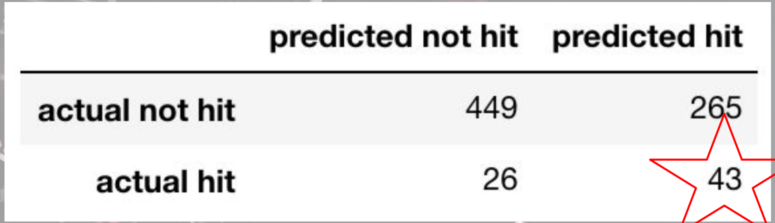


Results - Logistic Regression

	Audio Features	Lyrics
Accuracy Score	0.628	0.939
Recall Score	0.623	0.08

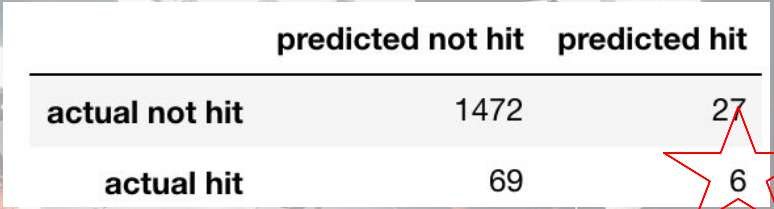
Results - Confusion Matrices

Audio Features



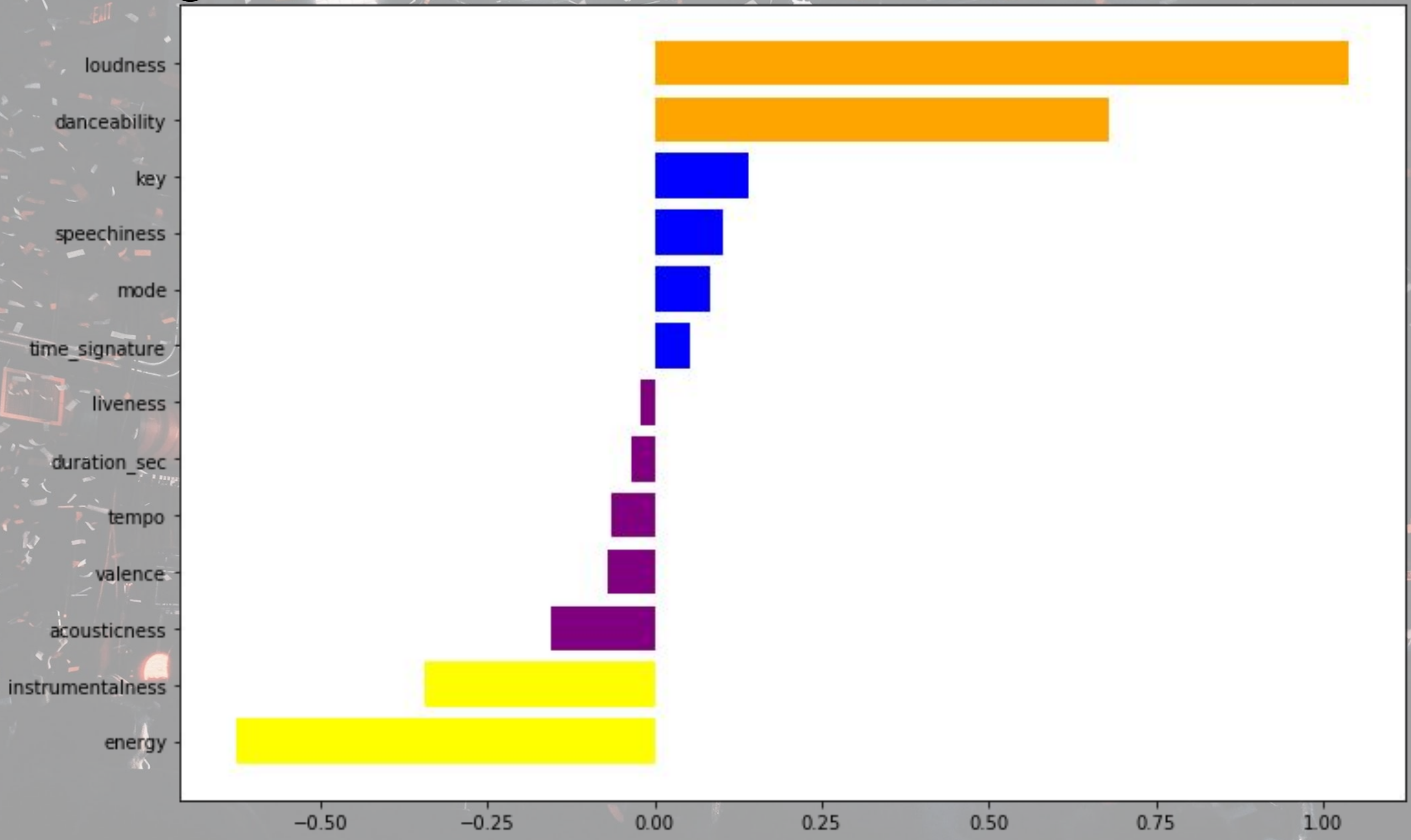
	predicted not hit	predicted hit
actual not hit	449	265
actual hit	26	43

Lyrics



	predicted not hit	predicted hit
actual not hit	1472	27
actual hit	69	6

Findings

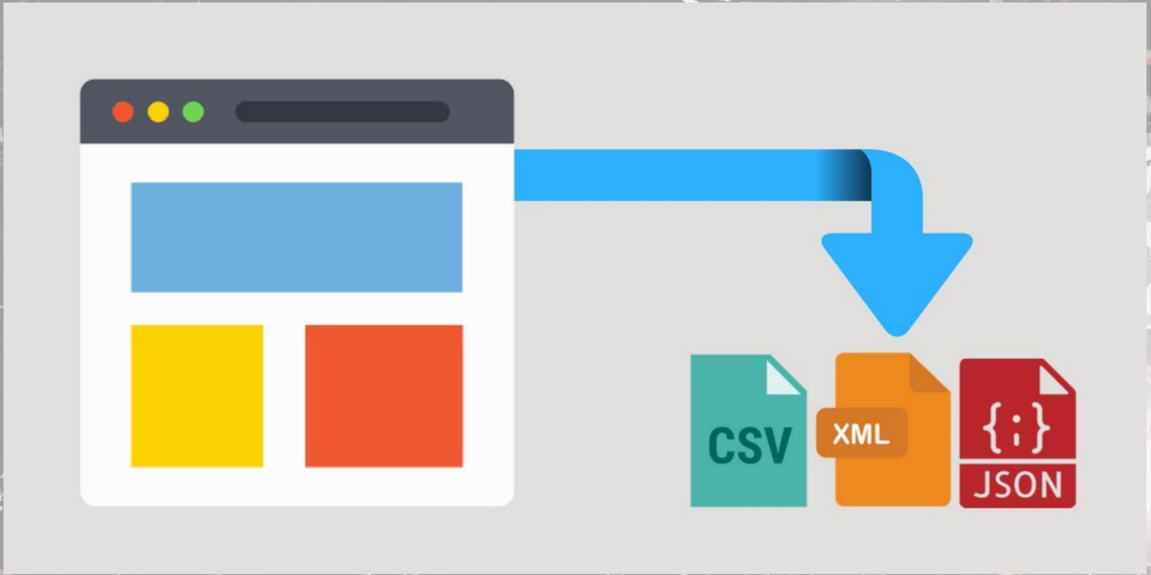


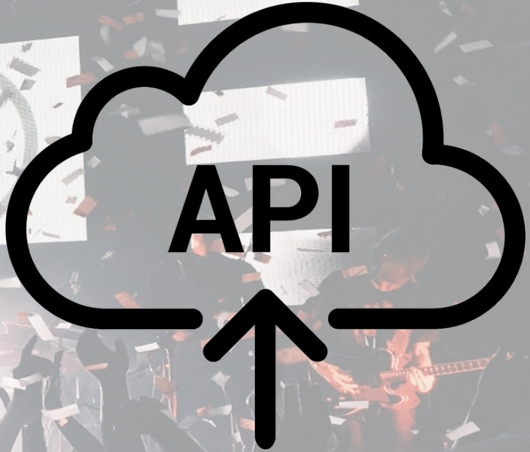
loudness
danceability
key
speechiness
mode
time_signature
liveness
duration_sec
tempo
valence
acousticness
instrumentalness
energy
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00

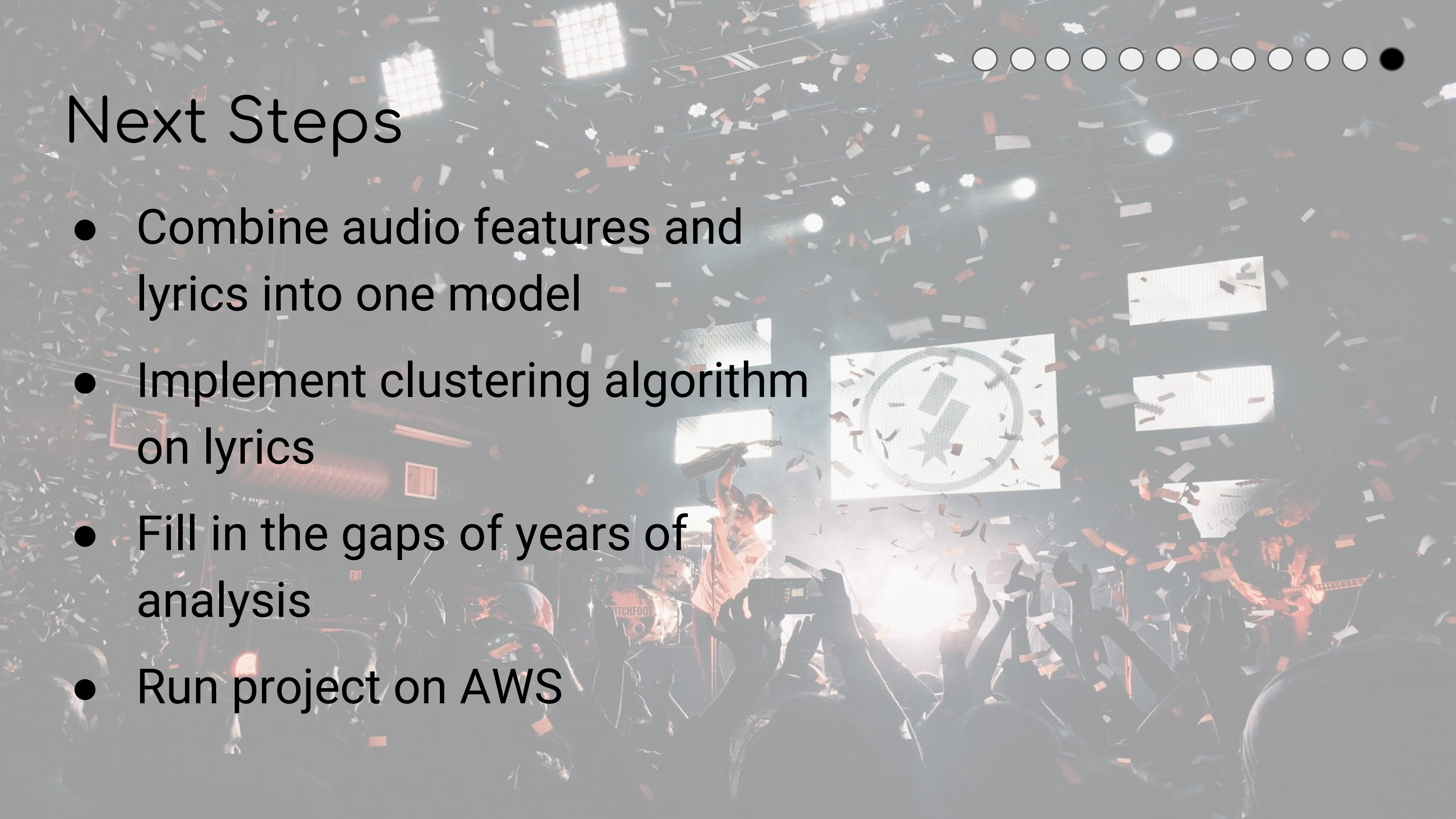
Recommendation

Invest in creating louder songs that have upbeat tempos and strong beats that are ideal for dancing!

Obstacles - Data Collection







Next Steps

- Combine audio features and lyrics into one model
- Implement clustering algorithm on lyrics
- Fill in the gaps of years of analysis
- Run project on AWS

Q & A

Dan Kim | GA Capstone