Mariana O. S. Silva
mariana.santos@dcc.ufmg.br

Laís M. Rocha
laismota@dcc.ufmg.br

Mirella M. Moro
mirella@dcc.ufmg.br

MusicOSet

An Enhanced Open Dataset for
Music Data Mining

UFMG CNPq

SBBD|2019

PUBLIC DATASETS

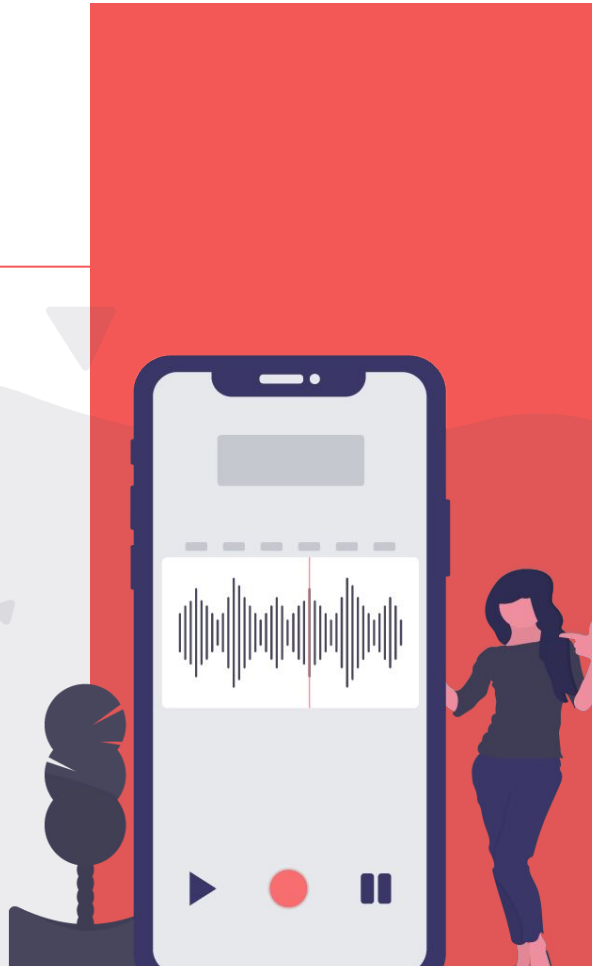
- Machine Learning
  - Iris Flower, Wine Quality
- Computer Vision
  - ImageNet
- Complex Networks
  - SNAP

PUBLIC DATASETS

- Machine Learning
  - Iris Flower, Wine Quality
- Computer Vision
  - ImageNet
- Complex Networks
  - SNAP
- **Music Retrieval Information (MIR)**

MUSIC RETRIEVAL INFORMATION

- Emerging research area dedicated to retrieving information from music
- Relevant musical content refers to **audio files** with:
  - lyrics
  - metadata
  - semantic information
- **Issue** ➜ apply musical multifaceted information for **predicting hit songs**
- New research area called *Hit Song Science* (HSS)



PROBLEM

HIT SONG SCIENCE

Predict *whether a song offers the potential to* ***become popular*** *and* ***commercially successful****, thus reaching the top of the charts*

COMPARISON EXISTING DATASETS

	size	metadata	acoustics	lyrics	popularity	year
MSD	1,000,000	✓	✓	☐	☐	2011
AudioSet	2,084,320	✓	✓	☐	☐	2017
FMA	106,574	✓	✓	☐	☐	2017
HSPD	1,000,000	☐	✓	☐	✓	2019

- Provide music information from different perspectives, focusing on a particular purpose
- Must provide a wide range of information in a centralized and easily accessible way

COMPARISON EXISTING DATASETS

	size	metadata	acoustics	lyrics	popularity	year
MSD	1,000,000	✔	✔	☐	☐	2011
AudioSet	2,084,320	✔	✔	☐	☐	2017
FMA	106,574	✔	✔	☐	☐	2017
HSPD	1,000,000	☐	✔	☐	✔	2019
MusicOSet	**20,405**	✔	✔	✔	✔	**2019**

FEATURES

EASY ACCESSIBLE

Unrestricted access in two formats (SQL database and compressed .csv files).

ENRICHED METADATA

Enriched metadata for music, artists, and albums from the US popular music industry.

CENTRALIZED

Integration and centralization of different musical data sources.

ACOUSTIC RESOURCES

Availability of acoustic fingerprints collected directly from *Spotify*.

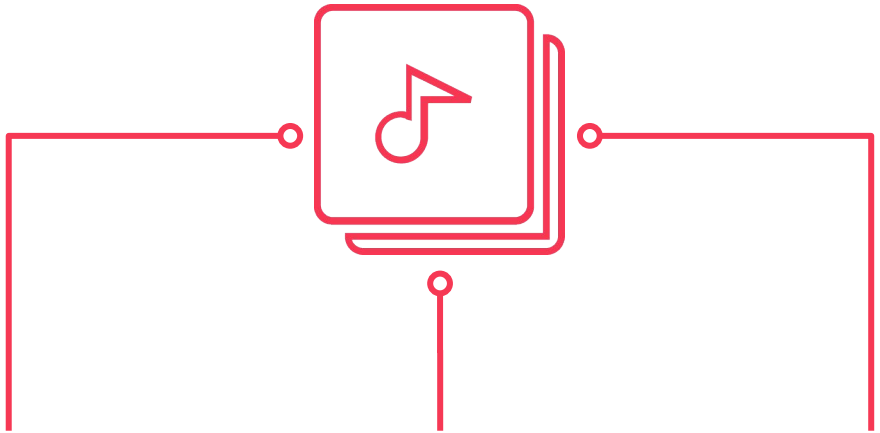
POPULARITY INFORMATION

Popularity scores and classification of hits and non-hits musical elements.

LYRICAL RESOURCES

Availability of lyrics resources collected using the *LyricsGenius* library.

CREATION PROCESS



Music Popularity Sources

Collected the last 56 years of the Hot 100 and Billboard 200 charts

(1962 to 2018)

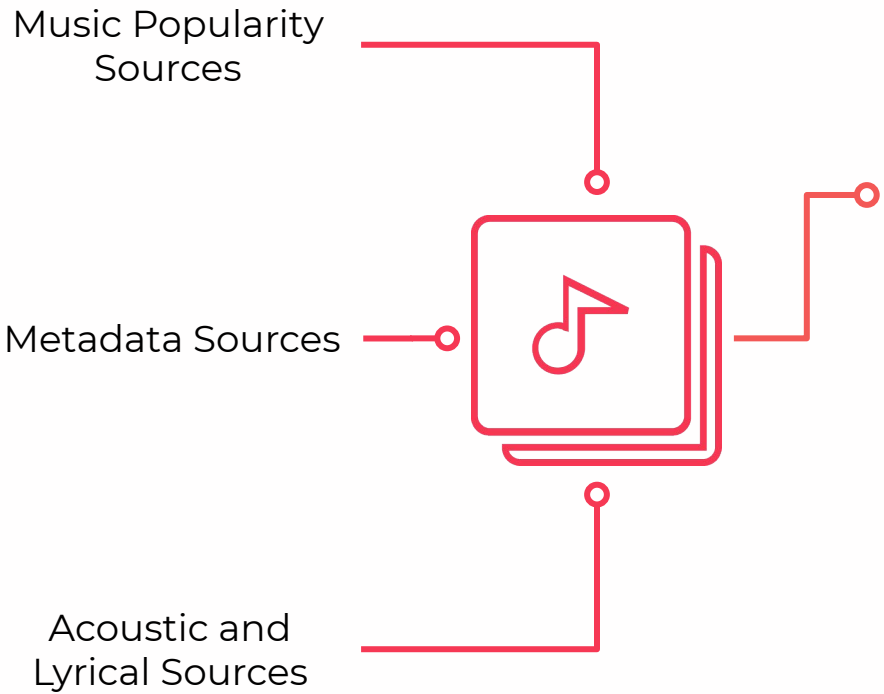
Metadata Sources

Used the *Spotify*, *Genius*, and *Wikipedia* platforms as sources of metadata and content

Acoustic and Lyrical Sources

Spotify ➜ collect acoustic fingerprints of the songs

Genius ➜ lyrics information

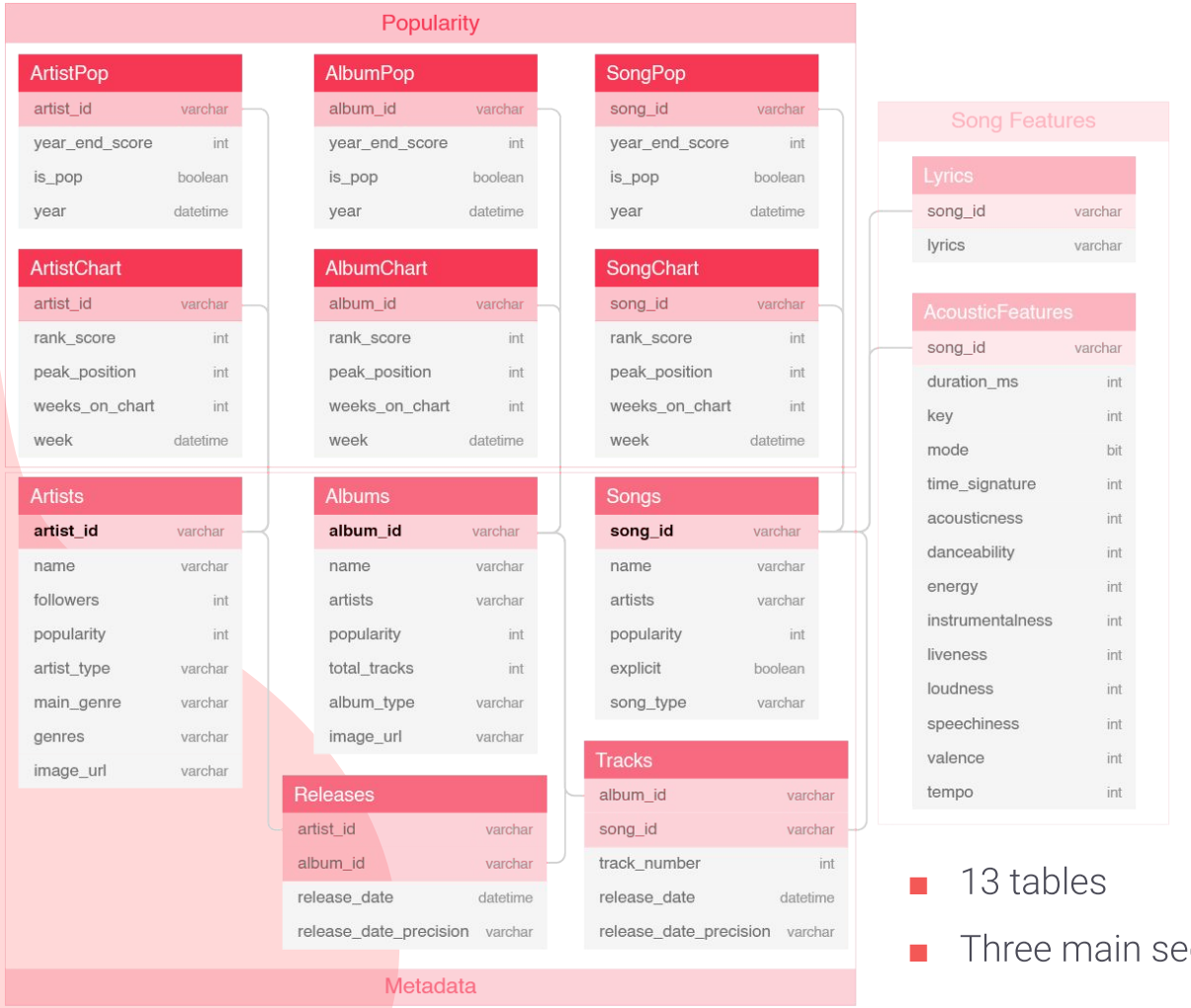


SOURCES INTEGRATION

- *SequenceMatcher*
- *Jaro-Winkler*
- Similarity ratio = 0.9

Mapped data **83%**

DATA CONTENT

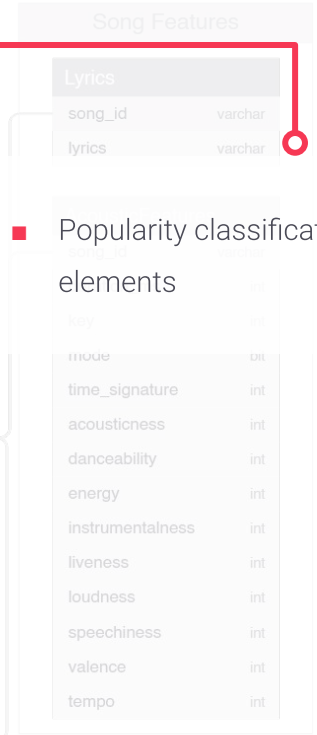


- 13 tables
- Three main segments

DATA CONTENT

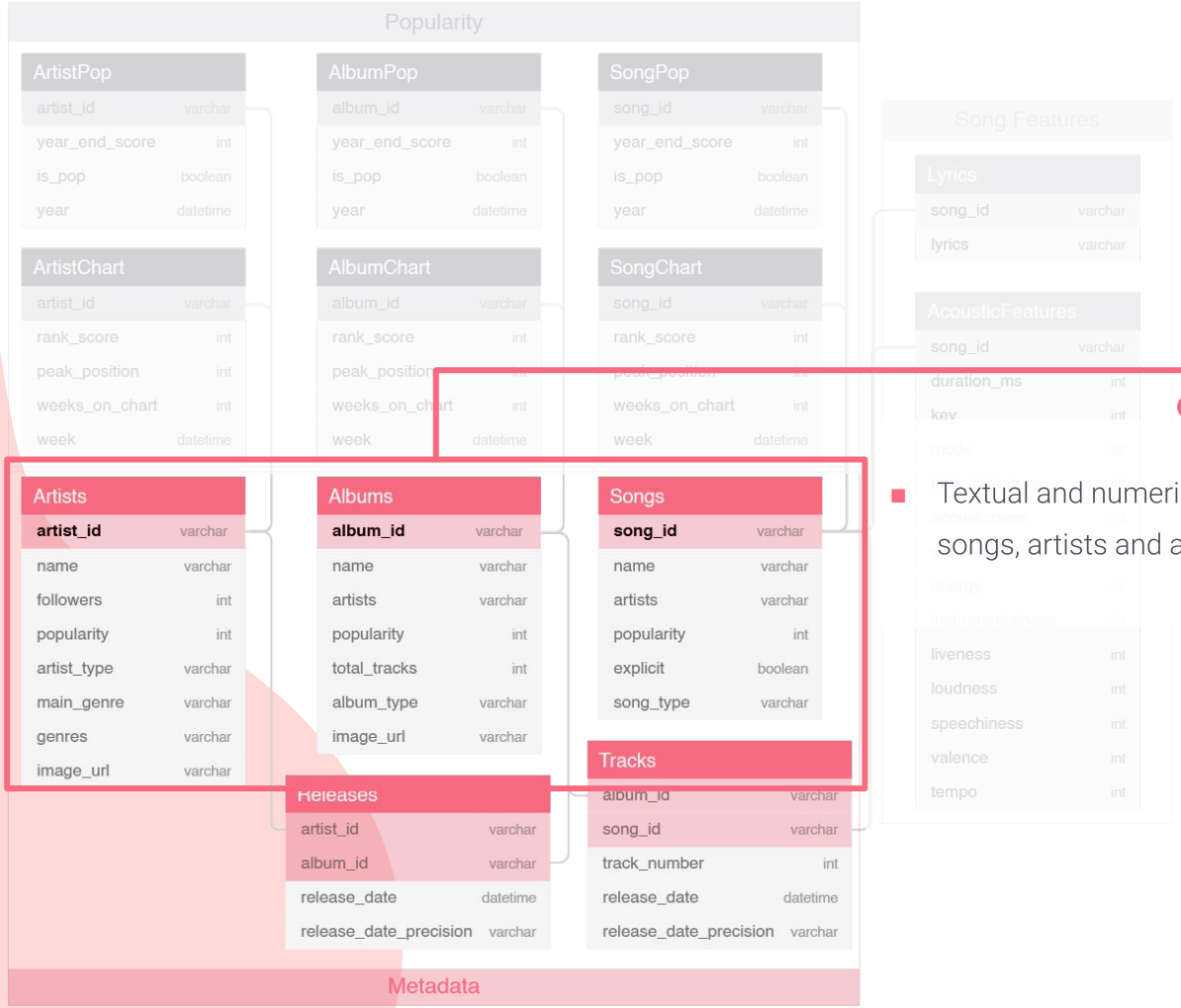
- Ranking information collected from Billboard charts

DATA CONTENT



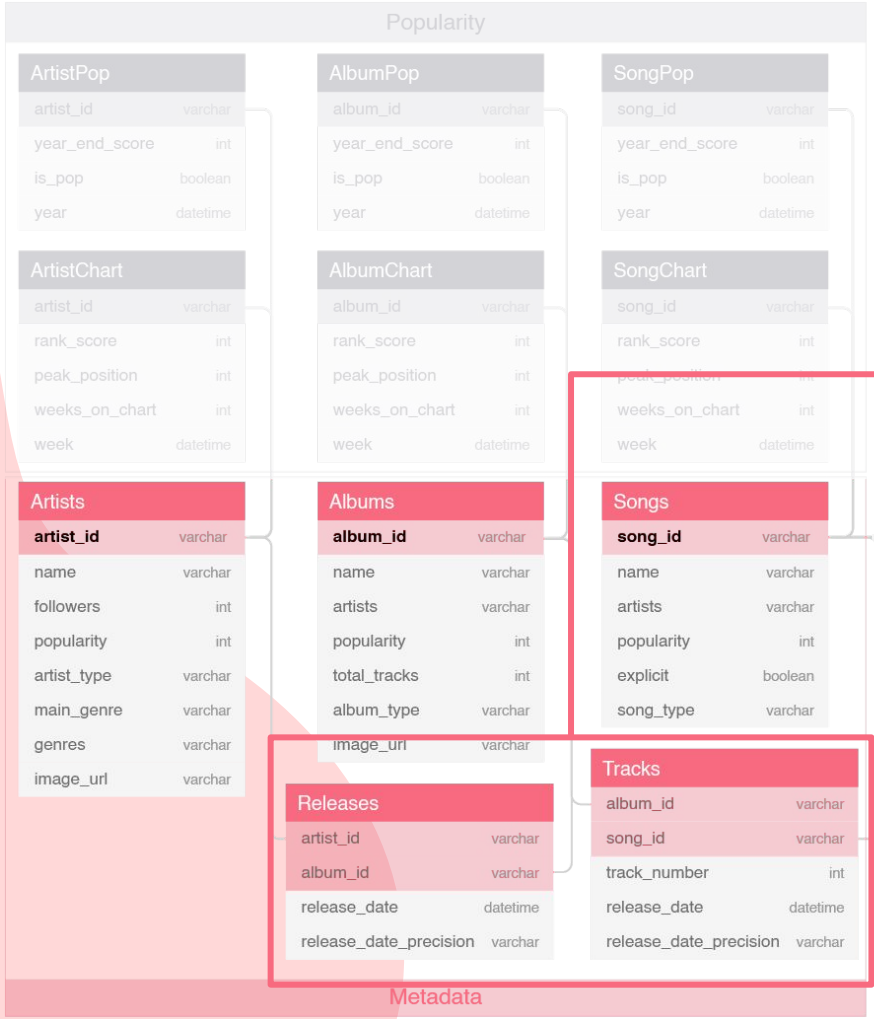
- Popularity classification of the musical elements

DATA CONTENT



- Textual and numeric information about songs, artists and albums

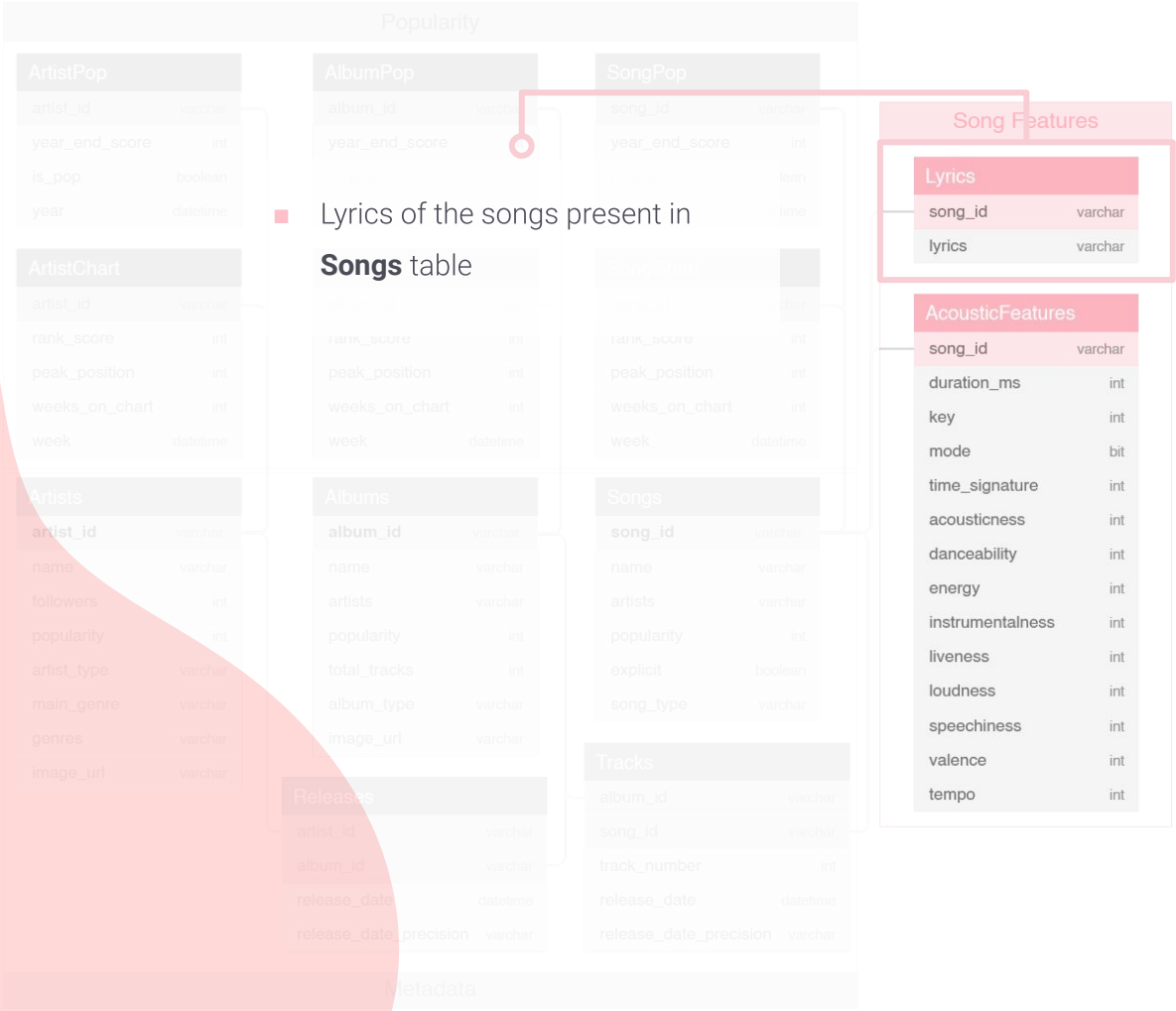
DATA CONTENT



- Store album release information
- Store song track information

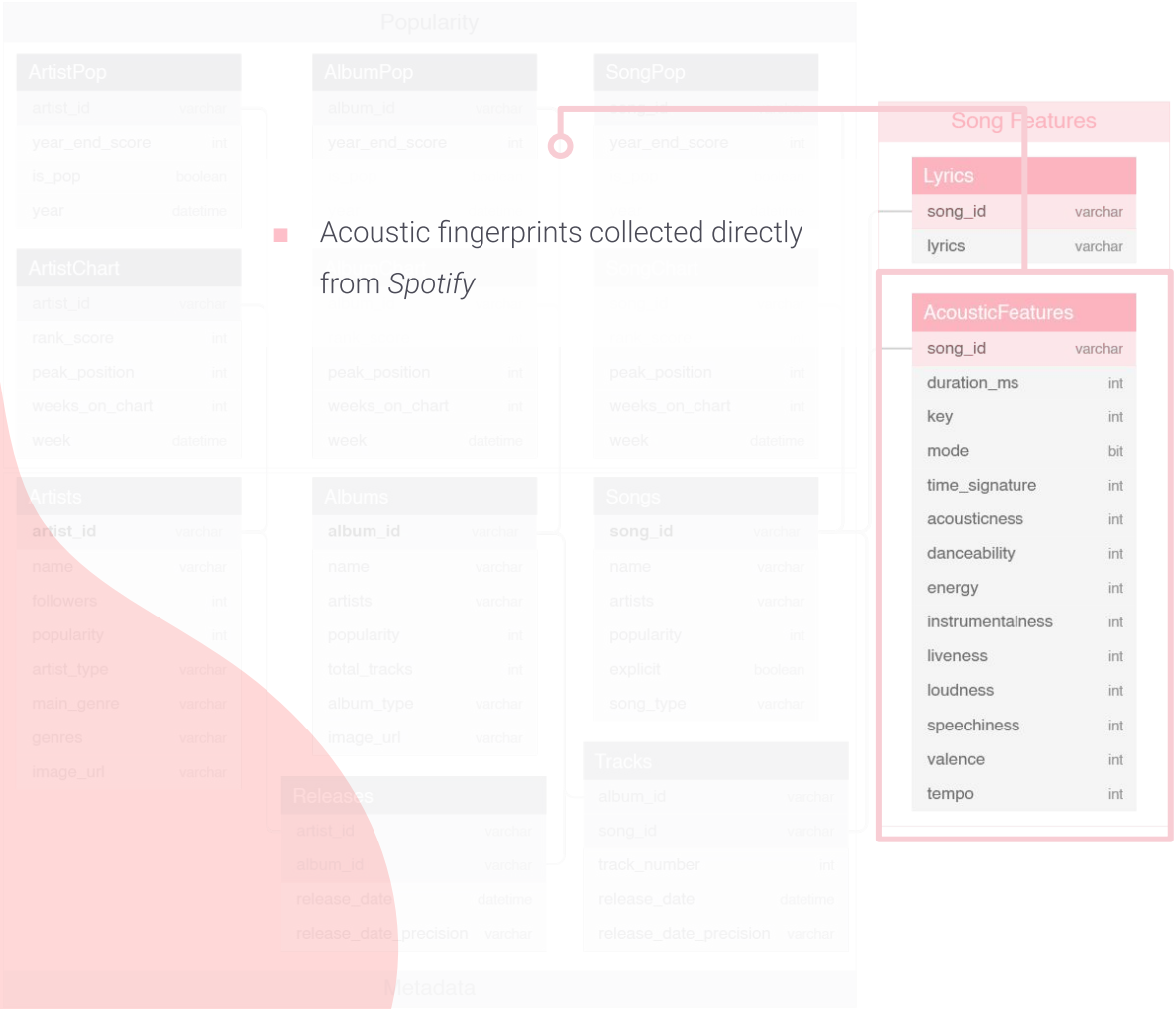
DATA CONTENT

- Lyrics of the songs present in **Songs** table

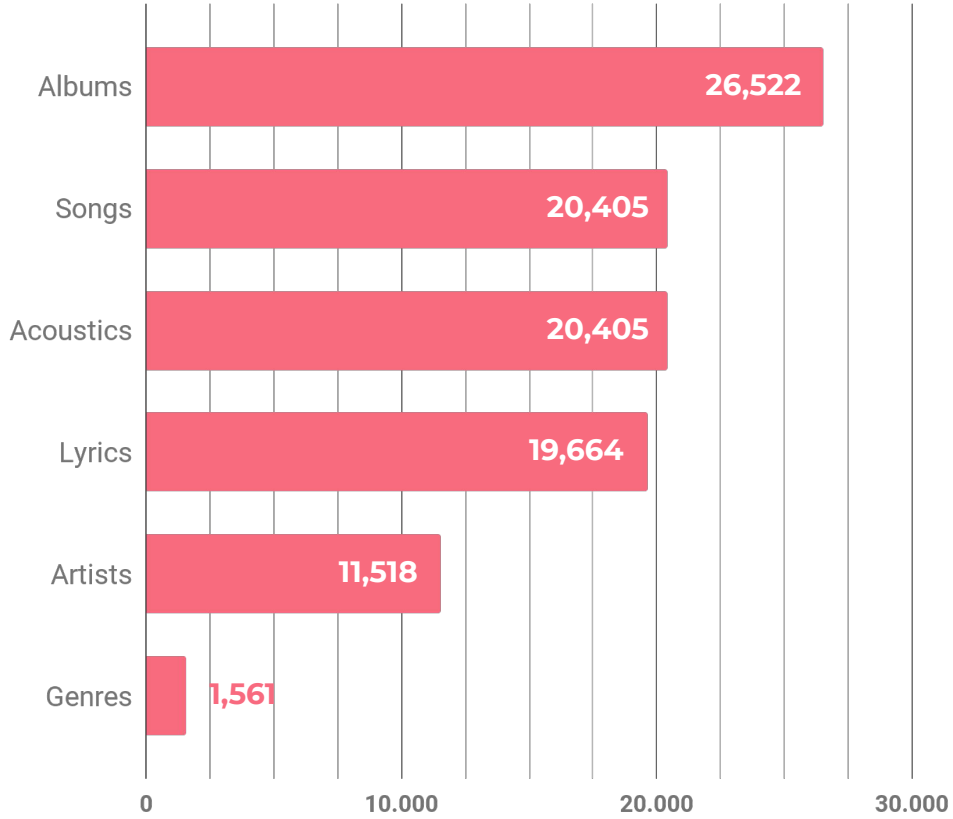


DATA CONTENT

- Acoustic fingerprints collected directly from *Spotify*



DATA STATISTICS



APPLICABILITY

Metadata Analysis

May involve: music visualization, association mining and clustering

Hit Song Science

The main goal is to predict the success of songs before they are released

Music Information Retrieval

Musical recommendation and musical similarity are well explored issues

APPLICABILITY

Metadata Analysis

May involve: music visualization, association mining and clustering

- Mariana O. Silva, Laís M. Rocha, and Mirella M. Moro. ***"Collaboration profiles and their impact on musical success."***, ACM SAC, 2019
- **Topological metrics + clustering algorithm ➜** identified three well-defined communities with distinct collaboration patterns
- **Successful artists** are more likely to have profiles with a **high degree of interaction and high diversification**

Hit Song Science

Music Information Retrieval

APPLICABILITY

Hit Song Science

The main goal is to predict the success of songs before they are released

- Mariana O. Silva and Mirella M. Moro. ***"Causality Analysis Between Collaboration Profilesand Musical Success."***, WebMedia, 2019
- Granger Causality
- Assess whether there is a causal relationship between collaboration profiles and artist popularity

Metadata Analysis

Music Information Retrieval

LIMITATIONS

MISSING DATA

Due to the different identification systems, not all sources provide information about all the data gathered

GENERALIZATION

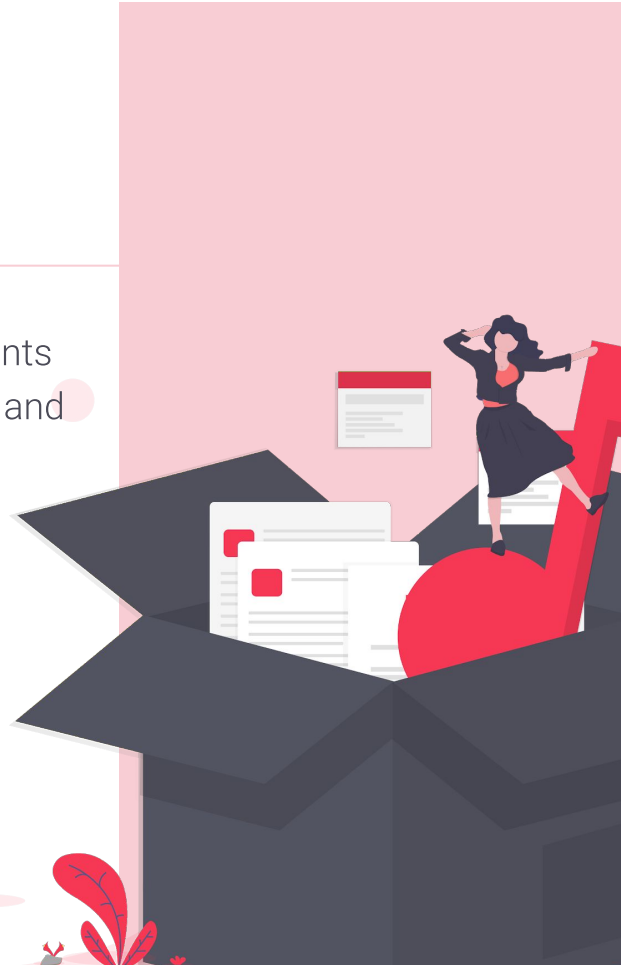
The data sources consider only the mainstream and popular music

DIVERSITY

Monopolization of US musical industry elements, as well as of pop and rock genres

MusicOSet

- A cured, open and enhanced dataset of musical elements
- **Contribution ➜** integrating metadata, audio resources and musical popularity information
- Can be used for many music data mining tasks
  - Recommendation
  - Clustering
  - Prediction of successful songs (HSS)



MusicOSet

- New data sources, increasing the scope of potential applications (e.g., Grammy Awards and Last.fm)
- Additional features
  - structure and content of the songs
  - listener information
  - extras metadata, etc

- **Ultimate Goal:** unified framework for predicting musical success by using machine learning methods

mariana.santos@dcc.ufmg.br laismota@dcc.ufmg.br mirella@dcc.ufmg.br

Dataset available here

MusicOSet

An Enhanced Open Dataset for Music Data Mining

UFMG CNPq SBBD|2019