The Million Song Dataset

AUDIO FEATURES

The Million Song Dataset

"There is no data like more data" Bob Mercer of IBM (1985).

T. Bertin-Mahieux, D.P.W. Ellis, B. Whitman, P. Lamere, The Million Song Dataset, In Proceedings of the 12^{th} International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.

Introduction

- The Million Song Dataset (MSD) contains metadata and extracted audio features for a million songs from The Echo Nest.
- Licensing
  - GZTAN a smaller dataset
  - Magnatagatune
  - MSD Legally available

MSD Goals

- Scale MIR and related research to commercial sizes
- Provide reference dataset for research evaluation
- Alternative shortcut for The Echo Nest’s API
- Kick start new MIR researchers

MIR Datasets Critical Requirements

- Algorithms should be scalable
- Realistically sized datasets are necessary

dataset	# songs / samples	audio
RWC	465	Yes
CAL500	502	No
GZTAN genre	1,000	Yes
USPOP	8,752	No
Swat10K	10,870	No
Magnatagatune	25,863	Yes
OMRAS2	50,000?	No
MusiCLEF	200,000	Yes
MSD	1,000,000	No

MSD Creation

- The Echo Nest API with Python wrapper pyechonest.
- Echo Nest provides:
  - Metadata: artist, title, etc.
  - Audio Features: short time scale – global scale
  - Defined by Echo Nest Analyze API (per segment)
- Additional info from musicbrainz server
- 5 Threads during 10 days
- Code available

MSD Content

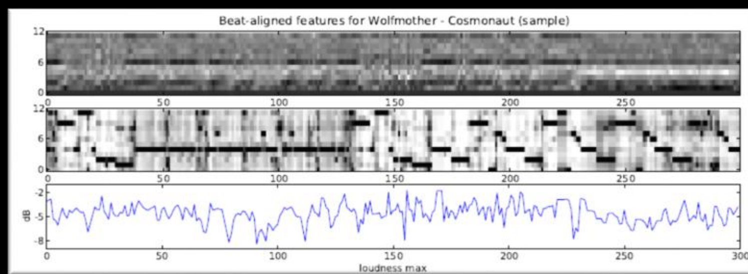
- 280 GB of data
- 1,000,000 songs/files
- 44,745 unique artists
- 7,643 unique terms (Echo Nest tags)
- 2,321 unique musicbrainz tags
- 43,943 artists with at least one term
- 2,201,916 asymmetric similarity relationships
- 515,576 dated tracks starting from 1922

MSD Content

- HDF5 format
- 55 fields per song
- Audio Features
  - Timbre
  - Pitches
  - Loudness max
  - Beats
  - Bars (~3 – 4 beats)
  - Note onsets/tatum

analysis_sample_rate	artist_7digitalid
artist_familiarity	artist_hotttnesss
artist_id	artist_latitude
artist_location	artist_longitude
artist_mbid	artist_mbtags
artist_mbtags_count	artist_name
artist_playmeid	artist_terms
artist_terms_freq	artist_terms_weight
audio_md5	bars_confidence
bars_start	beats_confidence
beats_start	danceability
duration	end_of_fade_in
energy	key
key_confidence	loudness
mode	mode_confidence
num_songs	release
release_7digitalid	sections_confidence
sections_start	segments_confidence
segments_loudness_max	segments_loudness_max_time
segments_loudness_start	segments_pitches
segments_start	segments_timbre
similar_artists	song_hotttnesss
song_id	start_of_fade_out
tatums_confidence	tatums_start
tempo	time_signature
time_signature_confidence	title
track_7digitalid	track_id
year	

MSD Audio Features



- Timbre, Pitches (both 12 elements per segment) and Loudness max for one song.

MSD Integration

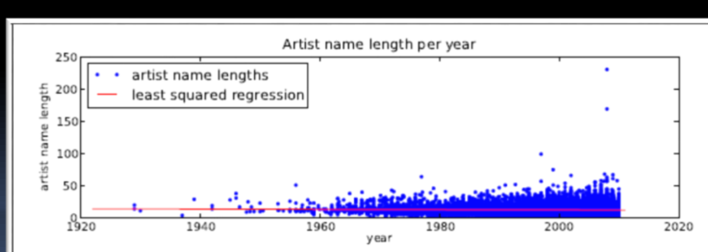
- Using Echo Nest identifiers (track, song, album, artist) the API can provide updates on dynamic values: popularity, familiarity, etc.
- Yahoo Music Ratings Datasets provides user ratings for 97 954 artists
  - 15 780 artists in MSD (91% overlap with the more popular artists in MSD)
  - One of the largest benchmarks for evaluating content-based music recommendation
- Identifiers
  - Artist, album, song names
  - Echo Nest id
  - Musicbrainz id
  - MusiXmatch id => lyrics
  - 7digital identifiers > 30sec samples

MSD Usage

- Metadata Analysis
- Artist Recognition
- Automatic Music Tagging
- Recommendation
- Cover Song Recognition
  - SecondHandSong Dataset 18 196 covers of 5 854 songs
  - Most methods based on chroma features
- Lyrics
  - Mood prediction
- Year Prediction

Metadata Analysis

- Are all good artist names already taken?
- Do newer bands have to use longer names?
  - Seems false, apart from outliers. See graph.
- Etc.



Artist Recognition

- 18 073 artists with at least 20 songs in MSD
- 2 standard training/test datasets
  - 20 songs/artist
  - 15 songs/artist
- Benchmark k-NN algorithm with accuracy of 4% provided => much room for improvement?

Automatic Music Tagging

- Core of MIR research for the last years
- 300 most popular terms in The Echo Nest
- Split all artists in training/test sets according to terms
- Lacking song tags
- Correlations between artist names and genre, or year and genre etc.

artist	EN terms	musicbrainz tags
Bon Jovi	adult contemporary arena rock 80s	hard rock glam metal american
Britney Spears	teen pop soft rock female	pop american dance

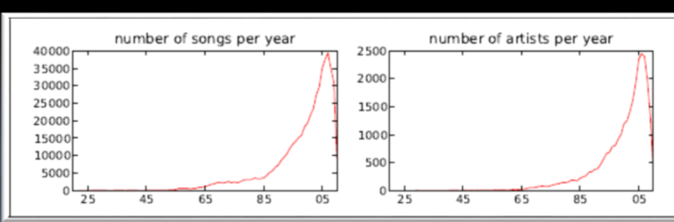
Music Recommendation

- Music recommendation and music similarity have high potential commercial value.
- Content based systems underperform when compared to collaborative filtering methods
  - Also novelty and serendipity are important.
- Integration with Yahoo Music Ratings
  - Enables large scale experiments
  - Clean ground truth
- Similar Artists according to Echo Nest:

Ricky Martin	Weezer
Enrique Iglesias	Death Cab for Cutie
Christina Aguilera	The Smashing Pumpkins
Shakira	Foo Fighters
Jennifer Lopez	Green Day

Year Prediction

- Little studied
- Practical applications in music recommendation
- Years-of-release field (1922 – 2011)
  - 515 576 tracks of 28 223 artists
  - Errors
  - Non-uniformity over the years



Year Prediction

- K-NN: the predicted year is the average of the k nearest training songs
- Vowpal & Wabbit (**VW**): regression by learning a linear transformation **T** of the features using gradient descent => predicted year is equal to the application of **T** on the features of the song
- Table shows
  - average absolute difference between predicted and actual yaer
  - the square root of the average squared difference between predicted and actual year.
- Benchmark average release year predicted from the training set. VW improves this baseline.

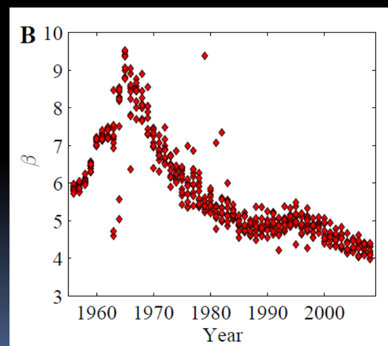
method	diff	sq. diff
constant pred.	8.13	10.80
1-NN	9.81	13.99
50-NN	7.58	10.20
vw	**6.14**	**8.76**

Evolution of Pop Music

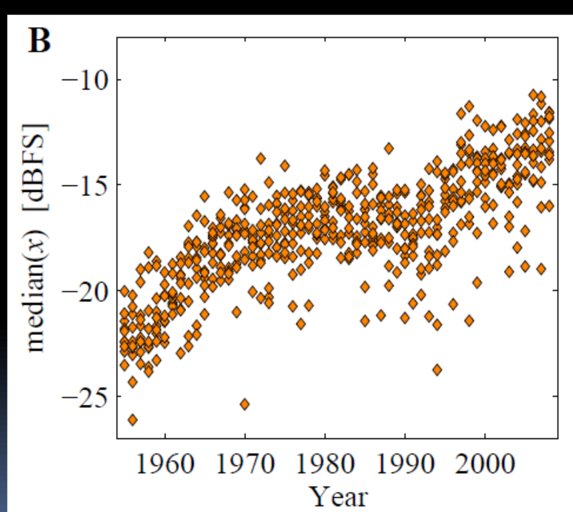
Measuring the evolution of contemporary western popular music, J. Serra, A. Corral, M. Boguna, M. Haro and J.L. Arcos, *2012*

Timbre of Pop Music

- The distributions of timbre codewords are fitted to a power-law distribution with parameter β.
- Lower β indicates less timbre variety, i.e., frequent code words become more frequent and infrequent ones less frequent.
- More homogeneity in timbre



Loudness of Pop Music



MSD Limitations

- No or limited access to original audio
  - Novel audio feature analysis and acoustic features
- Lack of album and song level meta data and tags
- Limited Diversity
  - World, ethnic, and classic music is not represented, or very limited
- Accurate time stamps problematic
  - No guarantee that audio features have been computed using the same audio track
  - As a result from many official releases, different ripping and encoding schemes, etc

the Million Song Dataset Challenge

B. McFee, et al., WWW 2012 Companion, April 16-20 2012, Lyon, France.

"... a large scale, personalized music recommendation challenge, where the goal is to predict the songs that a user will listen to, given both the user's listening history and full information (including meta-data and content analysis) for all songs. We explain the taste profile data, our goals and design choices in creating the challenge, and present baseline results using simple, off--the-shelf recommendation algorithms."

the Million Song Dataset Challenge

http://www.kaggle.com/c/msdchallenge

"What is the task in a few words? You have:
1) the full listening history for 1M users,
2) half of the listening history for 110K users (10K validation set, 100K test set), and
3) you must predict the missing half. .."

Winner: *aio* with a MAP@k score of 0.17910
(MAP@k = Mean average precision over k queries)

Future

- Very recent effort => Time will tell.
- Hopefully used as one of the default benchmarks
- Depends on efforts of research community
- Preserving commonality and comparability
- Important for visibility of MIR research
- Subsets on UCI Machine Learning Repository

ISMIR (http://www.ismir.net/)

- ISMIR 2014 Proceedings
  - http://dblp.uni-trier.de/db/conf/ismir/ismir2014.html

- Li Su, Li-Fan Yu, Yi-Hsuan Yang: Sparse Cepstral, Phase Codes for Guitar Playing Technique Classification. 9-14
- Antti Laaksonen: Automatic Melody Transcription based on Chord Transcription. 119-124
- Nikolay Glazyrin: Towards Automatic Content-Based Separation of DJ Mixes into Single Tracks. 149-154
- Dominique Fourer, Jean-Luc Rouas, Pierre Hanna, Matthias Robine: Automatic Instrument Classification of Ethnomusicological Audio Recordings. 295-300
- Po-Sen Huang, Minje Kim, Mark Hasegawa-Johnson, Paris Smaragdis: Singing-Voice Separation from Monaural Recordings using Deep Recurrent Neural Networks. 477-482
- Po-Kai Yang, Chung-Chien Hsu, Jen-Tzung Chien: Bayesian Singing-Voice Separation. 507-512

MIREX 2015
http://www.music-ir.org/mirex/wiki/MIREX_HOME

Challenges 2015

- Audio Classification (Train/Test) Tasks, incorporating:
  - Audio US Pop Genre Classification
  - Audio Latin Genre Classification
  - Audio Music Mood Classification
  - Audio Classical Composer Identification
- Singing Voice Separation
- Structural Segmentation
- Audio Cover Song Identification
- Audio Fingerprinting
- Audio Beat Tracking
- Etc.