THE MILLION SONG DATASET

Thierry Bertin-Mahieux, Daniel P.W. Ellis
Columbia University
LabROSA, EE Dept.
{thierry, dpwe}@ee.columbia.edu

Brian Whitman, Paul Lamere
The Echo Nest
Somerville, MA, USA
{brian, paul}@echonest.com

ABSTRACT

We introduce the Million Song Dataset, a freely-available collection of audio features and metadata for a million contemporary popular music tracks. We describe its creation process, its content, and its possible uses. Attractive features of the Million Song Database include the range of existing resources to which it is linked, and the fact that it is the largest current research dataset in our field. As an illustration, we present year prediction as an example application, a task that has, until now, been difficult to study owing to the absence of a large set of suitable data. We show positive results on year prediction, and discuss more generally the future development of the dataset.

1. INTRODUCTION

"There is no data like more data" said Bob Mercer of IBM in 1985 [7], highlighting a problem common to many fields based on statistical analysis. This problem is aggravated in Music Information Retrieval (MIR) by the delicate question of licensing. Smaller datasets have ignored the issue (e.g. GZTAN [11]) while larger ones have resorted to solutions such as using songs released under Creative Commons (Magnatagatune [9]).

The Million Song Dataset (MSD) is our attempt to help researchers by providing a large-scale dataset. The MSD contains metadata and audio analysis for a million songs that were legally available to The Echo Nest. The songs are representative of recent western commercial music. The main purposes of the dataset are:

- to encourage research on algorithms that scale to commercial sizes;
- to provide a reference dataset for evaluating research;
- as a shortcut alternative to creating a large dataset with The Echo Nest's API;
- to help new researchers get started in the MIR field.

Some have questioned the ability of conferences like ISMIR to transfer technologies into the commercial world, with scalability a common concern. Giving researchers a chance to apply their algorithms to a dataset of a million songs is a step in the right direction.

2. THE DATASET

2.1 Why?

The idea for the Million Song Dataset arose a couple of years ago while discussing ideas for a proposal to the US National Science Foundation's GOALI (Grant Opportunities for Academic Liaison with Industry) program. We wanted an idea that would not be possible without academic-industrial collaboration, and that would appeal to the NSF as contributing to scientific progress.

One of the long-standing criticisms of academic music information research from our colleagues in the commercial sphere is that the ideas and techniques we develop are simply not practical for real services, which must offer hundreds of thousands of tracks at a minimum. But, as academics, how can we develop scalable algorithms without the large-scale datasets to try them on? The idea of a "million song dataset" started as a flippant suggestion of what it would take to solve this problem. But the idea stuck – not only in the form of developing a very large, common dataset, but even in the specific scale of one million tracks.

There are a several possible reasons why the community does not already have a dataset of this scale:

- We all already have our favorite, personal datasets of hundreds or thousands of tracks, and to a large extent we are happy with the results we get from them.
- Collecting the actual music for a dataset of more than a few hundred CDs (i.e. the kind of thing you can do by asking all your colleagues to lend you their collections) becomes something of a challenge.

- The well-known antagonistic stance of the recording industry to the digital sharing of their data seems to doom any effort to share large music collections.
- It is simply a lot of work to manage all the details for this amount of data.

On the other hand, there are some obvious advantages to creating a large dataset:

- A large dataset helps reveal problems with algorithm scaling that may not be so obvious or pressing when tested on small sets, but which are critical to real-world deployment.
- Certain kinds of relatively-rare phenomena or patterns may not be discernable in small datasets, but may lead to exciting, novel discoveries from large collections.
- A large dataset can be relatively comprehensive, encompassing various more specialized subsets. By having all subsets within a single universe, we can have standardized data fields, features, etc.
- A single, multipurpose, freely-available dataset greatly promotes direct comparisons and interchange of ideas and results.

A quick look at other sources in Table 1 confirms that there have been many attempts at providing larger and more diverse datasets. The MSD stands out as the largest currently available for researchers.

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

Table 1. Size comparison with some other datasets.

2.2 Creation

The core of the dataset comes from The Echo Nest API [5]. This online resource provides metadata and audio analysis for millions of tracks and powers many music applications on the web, smart phones, etc. We had unlimited access to the API and used the python wrapper pyechonest [1] . We captured most of the information provided, ranging from timbre analysis on a short time-scale, to global artist similarity. From a practical point of view, it took us 5 threads running non-stop for 10 days to gather the dataset. All the code we used is available, which would allow data on additional tracks to be gathered in the same format. Some additional information was derived from a local musicbrainz server [2].

2.3 Content

The MSD contains audio features and metadata for a million contemporary popular music tracks. It contains:

- 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

The data is stored using HDF5 format [2] to efficiently handle the heterogeneous types of information such as audio features in variable array lengths, names as strings, longitude/latitude, similar artists, etc. Each song is described by a single file, whose contents are listed in Table 2.

The main acoustic features are *pitches*, *timbre* and *loudness*, as defined by the Echo Nest Analyze API. The API provides these for every "segment", which are generally delimited by note onsets, or other discontinuities in the signal. The API also estimates the tatums, beats, bars (usually groups of 3 or 4 beats) and sections. Figure 1 shows beat-aligned timbre and pitch vectors, which both consist of 12 elements per segment. Peak loudness is also shown.

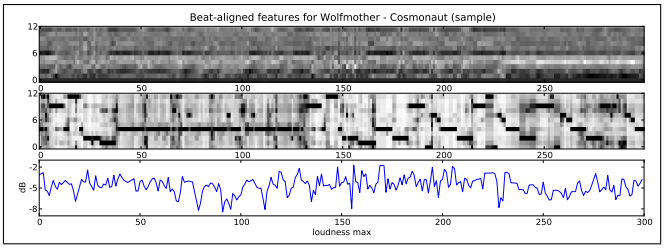

Figure 1. Example of audio features (*timbre*, *pitches* and *loudness max*) for one song.

[1] http://code.google.com/p/pyechonest/

[2] http://www.hdfgroup.org/HDF5/

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	

Table 2. List of the 55 fields provided in each per-song HDF5 file in the MSD.

The website [1] is a core component of the dataset. It contains tutorials, code samples [3], an FAQ, and the pointers to the actual data, generously hosted by Infochimps [4].

2.4 Links to other resources

The Echo Nest API can be used alongside the Million Song Dataset since we provide all The Echo Nest identifiers (track, song, album, artist) for each track. The API can give updated values for temporally-changing attributes (song hotttnesss, artist familiarity, ...) and also provides some data not included in the MSD, such as links to album cover art, artist-provided audio urls (where available), etc.

Another very large dataset is the recently-released Yahoo Music Ratings Datasets [5]. Part of this links user ratings to $97,954$ artists; $15,780$ of these also appear in the MSD. Fortunately, the overlap constitutes the more popular artists, and accounts for 91% of the ratings. The combination of the two datasets is, to our knowledge, the largest benchmark for evaluating content-based music recommendation.

The Echo Nest has partnered with 7digital [6] to provide the 7digital identifier for all tracks in the MSD. A free 7digital account lets you fetch 30 seconds samples of songs (up to some cap), which is enough for sanity checks, games, or user experiments on tagging. It might be feasible to compute some additional audio features on these samples, but only for a small portion of the dataset.

To support further linking to other sources of data, we provide as many identifiers as available, including The Echo Nest identifiers, the musicbrainz artist identifier, the 7digital and playme [7] identifiers, plus the artist, album and song names. For instance, one can use MusiXmatch [8] to fetch lyrics for many of the songs. Their API takes Echo Nest identifiers, and will also perform searches on artist and song title. We will return to musiXmatch in the next section.

3. PROPOSED USAGE

A wide range of MIR tasks could be performed or measured on the MSD. Here, we give a somewhat random sample of possible uses based on the community's current interests, which serves to illustrate the breadth of data available in the dataset.

3.1 Metadata analysis

The original intention of the dataset was to release a large volume of audio features for machine learning algorithms. That said, analyzing metadata from a million song is also extremely interesting. For instance, one could address questions like: Are all the "good" artist names already taken? Do newer bands have to use longer names to be original? This turns out to be false according to the MSD: The average length might even be reducing, although some recent outliers use uncommonly long names. The Figure 2 summarizes this. The least squared regression has parameters: gradient = -0.022 characters/year and intercept = 55.4 characters (the extrapolated length of a band name at year 0!).

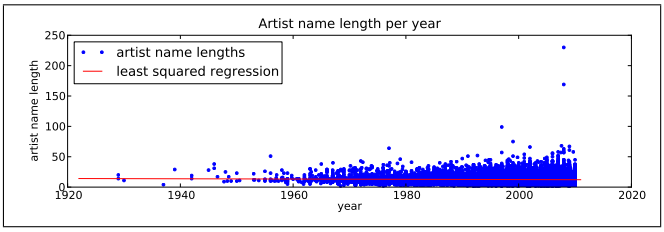

Figure 2. Artist name length as a function of year.

3.2 Artist recognition

Recognizing the artist from the audio is a straightforward task that provides a nice showcase of both audio features and machine learning. In the MSD, a reasonable target is

[3] https://github.com/tb2332/MSongsDB
[4] http://www.infochimps.com/
[5] http://webscope.sandbox.yahoo.com/
[6] http://www.7digital.com
[7] http://www.playme.com
[8] http://www.musixmatch.com

the $18,073$ artists that have at least 20 songs in the dataset (in contrast to the 5 artists reported a decade ago in [12]). We provide two standard training/test splits, the more difficult of which contains just 15 songs from each artist in the training set. This prevents the use of artist popularity. Our benchmark k-NN algorithm has an accuracy of 4% (code provided), which leaves plenty of room for improvement.

3.3 Automatic music tagging

Automatic tagging [4] has been a core MIR tasks for the last few years. The Echo Nest provides tags (called "terms") at the artist level, and we also retrieved the few terms provided by musicbrainz. A sample is shown in Table 3. We split all artists between train and test based on the 300 most popular terms from The Echo Nest. This makes it the largest available dataset for tagging evaluation, as compared to Magnatagatune [9], Swat10K [10] and the Last.FM corpus in [3]. That said, the MSD currently lacks any tags at the song, rather than the artist, level. We would welcome the contribution of such tags.

Although less studied, the correlation between tags and metadata could be of great interest in a commercial system. Certain "genre tags", such as "disco", usually apply to songs released in the 70s. There are also correlations between artist names and genres; you can probably guess the kind of music the band *Disembowelment* plays (if you are not already a fan).

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

Table 3. Example of tags for two artists, as provided by The Echo Nest and musicbrainz.

3.4 Recommendation

Music recommendation and music similarity are perhaps the best-studied areas in MIR. One reason is the potential commercial value of a working system. So far, content-based system have fallen short at predicting user ratings when compared to collaborative filtering methods. One can argue that ratings are only one facet of recommendation (since listeners also value novelty and serendipity [6]), but they are essential to a commercial system.

The Yahoo Music Ratings Datasets, mentioned above, opens the possibility of a large scale experiment on predicting ratings based on audio features with a clean ground truth. This is unlikely to settle the debate on the merit of content-based music recommendation once and for all, but it should support the discussion with better numbers.

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

Table 4. Some similar artists according to The Echo Nest.

3.5 Cover song recognition

Cover song recognition has generated many publications in the past few years. One motivation behind this task is the belief that finding covers relies on understanding something deeper about the structure of a piece. We have partnered with Second Hand Songs, a community-driven database of cover songs, to provide the SecondHandSong dataset [9]. It contains $18,196$ cover songs grouped into $5,854$ works (or *cliques*). For comparison, the MIREX 2010 Cover Song evaluation used 869 queries. Since most of the work on cover recognition has used variants of the chroma features which are included in the MSD (*pitches*), it is now the largest evaluation set for this task.

3.6 Lyrics

In partnership with musiXmatch (whose API was mentioned above), we have released the musiXmatch dataset [10], a collection of lyrics from $237,662$ tracks of the MSD. The lyrics come in a bag-of-words format and are stemmed, partly for copyright reasons. Through this dataset, the MSD links audio features, tags, artist similarity, etc., to lyrics. As an example, mood prediction from lyrics (a recently-popular topic) could be investigated with this data.

3.7 Limitations

To state the obvious, there are many tasks not suited for the MSD. Without access to the original audio, the scope for novel acoustic representations is limited to those that can be derived from the Echo Nest features. Also, the dataset is currently lacking album and song-level metadata and tags. Diversity is another issue: there is little or no world, ethnic, and classical music.

[9] SecondHandSongs dataset, the official list of cover songs within the Million Song Dataset, available at: `http://labrosa.ee.columbia.edu/millionsong/secondhand`

[10] musiXmatch dataset, the official lyrics collection for the Million Song Dataset, available at: `http://labrosa.ee.columbia.edu/millionsong/musixmatch`

Tasks that require very accurate time stamps can be problematic. Even if you have the audio for a song that appears in the MSD, there is little guarantee that the features will have been computed on the same audio track. This is a common problem when distributing audio features, originating from the numerous official releases of any given song as well as the variety of ripping and encoding schemes in use. We hope to address the problem in two ways. First, if you upload audio to The Echo Nest API, you will get a time-accurate audio analysis that can be formatted to match the rest of the MSD (code provided). Secondly, we plan to provide a fingerprinter that can be use to resolve and align local audio with the MSD audio features.

4. YEAR PREDICTION

As shown in the previous section, many tasks can be addressed using the MSD. We present year prediction as a case study for two reasons: (1) it has been little studied, and (2) it has practical applications in music recommendation.

We define year prediction as estimating the year in which a song was released based on its audio features. (Although metadata features such as artist name or similar artist tags would certainly be informative, we leave this for future work). Listeners often have particular affection for music from certain periods of their lives (such as high school), thus the predicted year could be a useful basis for recommendation. Furthermore, a successful model of the variation in music audio characteristics through the years could throw light on the long-term evolution of popular music.

It is hard to find prior work specifically addressing year prediction. One reasons is surely the lack of a large music collection spanning both a wide range of genres (at least within western pop) and a long period of time. Note, however, that many music genres are more or less explicitly associated with specific years, so this problem is clearly related to genre recognition and automatic tagging [4].

4.1 Data

The "year" information was inferred by matching the MSD songs against the musicbrainz database, which includes a year-of-release field. This resulted in values for $515,576$ tracks representing $28,223$ artists. Errors could creep into this data from two main sources: incorrect matching, and incorrect information in musicbrainz. Informal inspection suggests the data is mostly clean; instead, the main issue is the highly nonuniform distribution of data per year, as shown in Figure 3. A baseline, uniform prediction at the mode or mean year would give reasonable accuracy figures because of the narrow peak in the distribution around 2007. However, we have enough data to be able to show that even small improvements in average accuracy are statistically significant: With $2,822$ test artists and using a z-test with a 95% confidence level, an improvement of 1.8 years is significant. Allowing some independence between the songs from a single artist reduces that number still more.

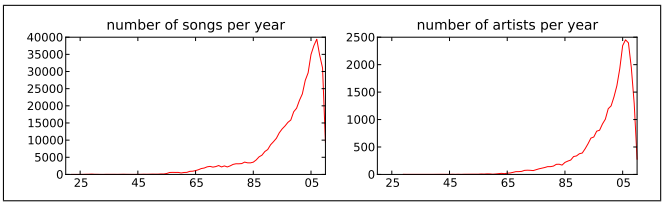

Figure 3. Distribution of MSD tracks for which release year is available, from 1922 to 2011. An artist's "year" value is the average of their songs.

Again, we define and publish a split between train and test artists so future results can be directly comparable. The split is among artists and not songs in order to avoid problems such as the "producer effect". The features we use are the average and covariance of the timbre vectors for each song. No further processing is performed. Using only the nonredundant values from the covariance matrix gives us a feature vector of 90 elements per track.

4.2 Methods

Our first benchmark method is k nearest neighbors (k-NN), which is easy to parallelize and requires only a single pass over the training set, given enough memory. Prediction can efficiently performed thanks to libraries such as ANN [11]. The predicted year of a test item is the average year of the k nearest training songs.

A more powerful algorithm, specifically designed for large-scale learning, is Vowpal Wabbit [8] (VW). It performs regression by learning a linear transformation w of the features x using gradient descent, so that the predicted value $\hat{y}^i$ for item i is:

$$\hat{y}^i = \sum_j w_j x_j^i$$

Year values are linearly mapped onto $[0, 1]$ using 1922 as 0 and 2011 as 1. Once the data is cached, VW can do many passes over the training set in a few minutes. VW has many parameters; we performed an exhaustive set of experiments using a range of parameters on a validation set. We report results using the best parameters from this search according to the average difference measure. The final model is trained on the whole training set.

4.3 Evaluation and results

Table 5 presents both average absolute difference and square root of the average squared difference between the predicted release year and the actual year.

[11] http://www.cs.umd.edu/~mount/ANN/

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**

Table 5. Results on year prediction on the test songs.

The benchmark is the "constant prediction" method, where we always predict the average release year from the training set (1998.4). With VW [12] we can make a significant improvement on this baseline.

5. THE FUTURE OF THE DATASET

Time will tell how useful the MSD proves to be, but here are our thoughts regarding what will become of this data. We have assemble a dataset which we designed to be comprehensive and detailed enough to support a very wide range of music information research tasks for at least the near future. Our hope is that the Million Song Dataset becomes the natural choice for researchers wanting to try out ideas and algorithms on data that is standardized, easily obtained, and relevant to both academia and industry. If we succeed, our field can be greatly strengthened through the use of a common, relevant dataset.

But for this to come true, we need lots of people to use the data. Naturally, we want our investment in developing the MSD to have as much positive impact as possible. Although the effort so far has been limited to the authors, we hope that it will become a true community effort as more and more researchers start using and supporting the MSD. Our vision is of many different individuals and groups developing and contributing additional data, all referenced to the same underlying dataset. Sharing this augmented data will further improve its usefulness, while preserving as far as possible the commonality and comparability of a single collection.

5.1 Visibility for MIR

The MSD has good potential to enhance the visibility of the MIR community in the wider research world. There have been numerous discussions and comments on how our field seems to take more that it gives back from other areas such as machine learning and vision. One reason could be the absence of a well-known common data set that could allow our results to be reported in conferences not explicitly focused on music and audio. We hope that the scale of the MSD will attract the interest of other fields, thus making MIR research a source of ideas and relevant practice. To that end, subsets of the dataset will be made available on the UCI Machine Learning Repository [13] . We consider such dissemination of MIR data essential to the future health of our field.

[12] The parameters to VW were –passes 100 –loss_function squared -l 100 –initial_t 100000 –decay_learning_rate 0.707106781187.

6. ACKNOWLEDGEMENTS

This work is supported by NSF grant IIS-0713334 and by a gift from Google, Inc. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reect the views of the sponsors. TBM is supported in part by a NSERC scholarship.

7. REFERENCES

[1] Million Song Dataset, official website by Thierry Bertin-Mahieux, available at: `http://labrosa.ee.columbia.edu/millionsong/`.

[2] Musicbrainz: a community music metadatabase, Feb. 2011. MusicBrainz is a project of The MetaBrainz Foundation, http://metabrainz.org/.

[3] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere. Autotagger: a model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research, special issue: "From genres to tags: Music Information Retrieval in the era of folksonomies."*, 37(2), June 2008.

[4] T. Bertin-Mahieux, D. Eck, and M. Mandel. Automatic tagging of audio: The state-of-the-art. In Wenwu Wang, editor, *Machine Audition: Principles, Algorithms and Systems*, pages 334–352. IGI Publishing, 2010.

[5] The Echo Nest Analyze, API, `http://developer.echonest.com`.

[6] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[7] F. Jelinek, 2004. `http://www.lrec-conf.org/lrec2004/doc/jelinek.pdf`.

[8] J. Langford, L. Li, and A. L. Strehl. Vowpal wabbit (fast online learning), 2007. `http://hunch.net/vw/`.

[9] E. Law and L. von Ahn. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the 27th international conference on Human factors in computing systems*, pages 1197–1206. ACM, 2009.

[10] D. Tingle, Y.E. Kim, and D. Turnbull. Exploring automatic music annotation with acoustically-objective tags. In *Proceedings of the international conference on Multimedia information retrieval*, pages 55–62. ACM, 2010.

[11] G. Tzanetakis and P. Cook. Musical genre classification of audio signals. *IEEE Trans. on Speech and Audio Processing*, 10(5):293–302, 2002.

[12] B. Whitman, G. Flake, and S. Lawrence. Artist detection in music with minnowmatch. In *Neural Networks for Signal Processing XI, 2001. Proceedings of the 2001 IEEE Signal Processing Society Workshop*, pages 559–568. IEEE, 2002.

[13] `http://archive.ics.uci.edu/ml/`