Anya Kalogerakos

Dr. Haverals

HUM 346

21 February 2024

Data Biography: The Million Song Dataset

Compromised of nearly 300 GB of data, the Million Song Dataset includes metadata for precisely “one million contemporary popular music tracks.”[[1]](#footnote-1) A collaborative project funded by the NSF in 2011, commercial entity Echo Nest and Columbia University research group LabROSA partnered to create this database in the interest of helping researchers in the field of Music Information Retrieval (MIR). The goal was to develop a reference dataset for evaluating MIR research and, as often MIR researchers work with smaller amounts of data to develop their algorithms (since gaining the rights and data from over a hundred thousand songs is quite difficult), the creators of the Million Song Dataset also hoped to give researchers the capability to develop and test algorithms with commercial-sized data applications in mind (as many commercial music databases include millions of songs).[[2]](#footnote-2)

Accessing this data is a slightly complex process due to its large size, which cannot be stored or manipulated on a typical personal computer. Instead, the Million Song Dataset website provides a link to an Amazon Public Dataset snapshot, which users can manipulate with Amazon’s EC2 virtual machine[[3]](#footnote-3) for a relatively small cost of about ten dollars per week. An important clarification here is that while the data is free and open, the cost of storing and computing on such a large dataset is not. Unfortunately, at the time of writing this paper, the link to the Amazon Public Dataset snapshot holding the Million Song Dataset is not active.

The procedure of collecting and processing the data for the Million Song Dataset constituted of largely three parts: choosing the one million songs to include in the dataset, analyzing the million songs with The Echo Nest API, and creating the structured dataset based on analysis results and related API methods to access data in the dataset efficiently. For the initial step of choosing one million songs that would create a worthwhile MIR standard, the creators added as many songs as possible from what The Echo Nest determined to be the most popular artists (coined as “most familiar” by Echo Nest), then they gathered the top 200 terms from The Echo Nest and used each term to find 100 artists, adding as many of their songs as possible. Then they added the songs of artists from a popular MIR dataset called the “CAL500” dataset which hosts 500 songs from 500 different artists.[[4]](#footnote-4) Next, they added songs that The Echo Nest classified as search parameter extremes (e.g. highest energy song, fastest tempo song, lowest energy song, etc.). Finally, they executed a random walk beginning at the 100 most popular artists according to The Echo Nest and explored artists classified as “similar artists” by The Echo Nest.[[5]](#footnote-5) It is important to note that the collection of data for each of these songs did *not* include any audio of the songs, but instead a series of data points generated by The Echo Nest API. While the list of data points collected with each song is quite long, examples points include artist name, number of beats, “danceability,” estimation of what key the song is in, the loudness of the track, album name, number of segments in a song, and much more.[[6]](#footnote-6)

With such vast amounts of information, it is a bit difficult to fathom what projects might be accomplishable with the Million Song Dataset. Fortunately, the Million Song Dataset website offers some suggestions for users getting into data analysis that can inform what other uses this data may have. Year prediction for when a song was released, song title analysis, artist recognition, and cover song recognition are all examples of projects that can be attempted with this dataset. However, due to the nature of the data collected, there are limitations present in the research that can be pursued with the Million Song Dataset. One obvious constraint is the year in which all this data was collected. While the Million Song Dataset contains songs ranging in year from the 1920s to the 2010s, its data does not extend to songs released past 2011. Therefore, any conclusions drawn from this dataset may not apply to music released in the past decade. Furthermore, although some level of randomness was included in the process of determining which million songs to include, much of the collection process relied on song popularity and similarity, which may have led to non-traditional songs and genres being excluded from this dataset. While this exclusion is valuable, as it allows this dataset to make conclusions about popular music trends, it does disallow conclusions about *all* music or music belonging to non-popular genres or artists to be made.

As for the process of analyzing the million songs chosen, the methods used to do so are relatively mysterious, and no raw song data is provided. The dataset relied heavily on The Echo Nest API to create its data points; however, little information about The Echo Nest API and the algorithms it uses is provided. Additionally, in the years since the creation of the dataset, The Echo Nest API has been purchased by Spotify and privatized, such that there is little API open-source code still available. While The Echo Nest API is likely reliable, the Million Song Dataset does put a lot of faith in the API to correctly generate song information such as key estimation, popularity, similar artists, and much more. The inability to verify the algorithms behind these outputs may be a concern for those planning to use the dataset for research sensitive to algorithmic bias.

Overall, since this data was collected in 2011, the lack of upkeep is this dataset's biggest weakness. As just addressed, now that the API of The Echo Nest is privatized, the lack of record on how this API was applied, as well as the lack of raw data before The Echo Nest manipulation, makes it difficult to accurately assess how this data can be properly used. Furthermore, many of the links presented with the description of this dataset are now inactive—most importantly, the link to the dataset itself has expired. While the database’s potential for advancing MIR research and the effort put forward by The Echo Nest and LabROSA to create this database is certainly impressive, its current state of inaccessibility, coupled with outdated data and limited transparency in the collection process, hinders its utility for contemporary research endeavors.

Sources

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The Million Song Dataset. In Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011.

Million Song Dataset, official website by Thierry Bertin-Mahieux, available at: <http://millionsongdataset.com/>

1. [The Million Song Dataset](http://millionsongdataset.com/) [↑](#footnote-ref-1)
2. For instance, Spotify has data on over 100 million songs on its platform (See [Spotify.com](https://newsroom.spotify.com/company-info/))) [↑](#footnote-ref-2)
3. Though this dataset was created a while ago, it is still possible to run the Amazon EC2 VM [here](https://aws.amazon.com/getting-started/launch-a-virtual-machine-B-0/) [↑](#footnote-ref-3)
4. Information on the CAL500 dataset can be found [here](http://calab1.ucsd.edu/~datasets/cal500/) [↑](#footnote-ref-4)
5. Once a similar artist is explored, the algorithm then walks through the artists similar to that similar artist, thus generating a wide variety of artists via random walk [↑](#footnote-ref-5)
6. See all fields included in the dataset [here](http://millionsongdataset.com/pages/example-track-description/) [↑](#footnote-ref-6)