Open Science in Practice

Summer School September 29, 2017

# Opening a Large Audio Dataset

Michaël Defferrard

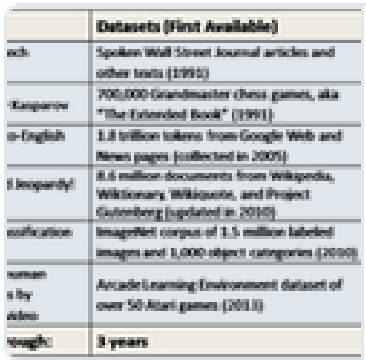
École Polytechnique Fédérale de Lausanne (EPFL)

# Why

**Fei-Fei Li**
@drfeifei

Following

Datasets play crucial roles in advancing AI. #ImageNet helped to enable the latest deep learning advances.




**Datasets Over Algorithms**

Content without method leads to fantasy; method without content to empty sophistry. — Johann Wolfgang von Goethe ("Maxims and Reflections", 1892) "Perhaps the most important news of ...

spacemachine.net

5:12 AM - 22 Apr 2016

**103** Retweets **174** Likes

2 103 174

# FMA: A Dataset For Music Analysis

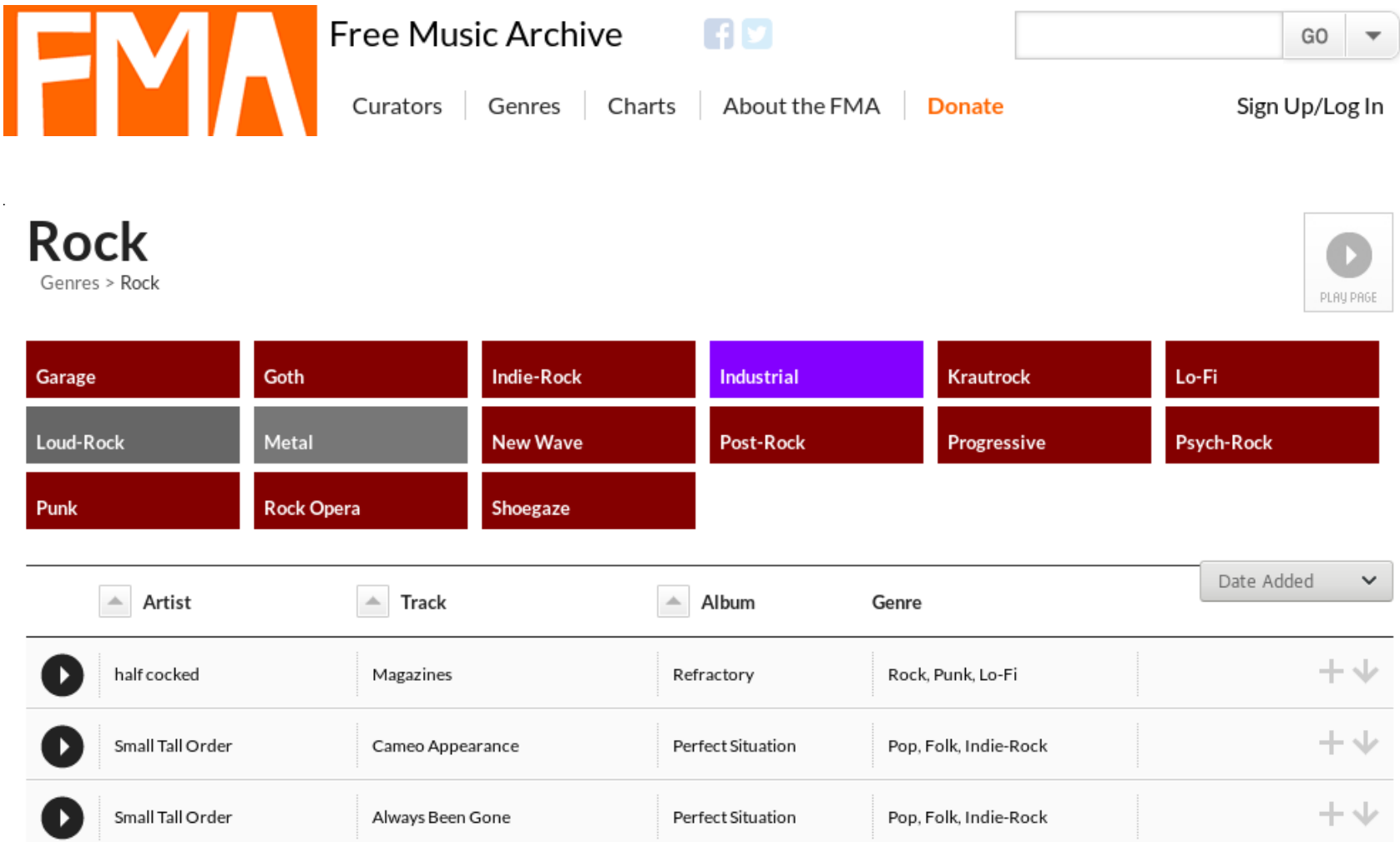
Defferrard, Benzi, Vandergheynst, and Bresson 2017

Goal: open dataset for Music Information Retrieval (MIR)
mostly for Machine Learning (ML)

- 106,574 tracks from 16,341 artists and 14,854 albums
- 917 GiB, 343 days of audio
- hierarchical taxonomy of 161 genres

# The Free Music Archive

http://freemusicarchive.org



# How

1. Collect
2. Clean
3. Package
4. Distribute

# Challenges

**Cleaning** Where to stop?

**Metadata** How to best describe my data?

**Data formats** What are the trade-offs?

**Publishing** Where? How?

| track | | | | | | album | | |
|---|---|---|---|---|---|---|---|---|
| track_id | title | genres_all | genre_top | dur. | listens | title | listens | tags |
| 150073 | Welcome to Asia | [2, 79] | International | 81 | 683 | Reprise | 4091 | [world music, dubtronica, fusion] |
| 140943 | Sleepless Nights | [322, 5] | Classical | 246 | 1777 | Creative Commons Vol. 7 | 28900 | [classical, alternate, soundtrack, piano, ... |
| 64604 | i dont want to die alone | [32, 38, 456] | Experimental | 138 | 830 | Summer Gut String | 7408 | [improvised, minimalist, noise, ... |
| 23500 | A Life In A Day | [236, 286, 15] | Electronic | 264 | 1149 | A Life in a Day | 6691 | [idm, candlestick, romanian, candle, ... |
| 131150 | Yeti-Bo-Betty | [25, 12, 85] | Rock | 124 | 183 | No Life After Crypts | 3594 | [richmond, fredericksburg, trash rock, ... |

# Challenges

## Licenses

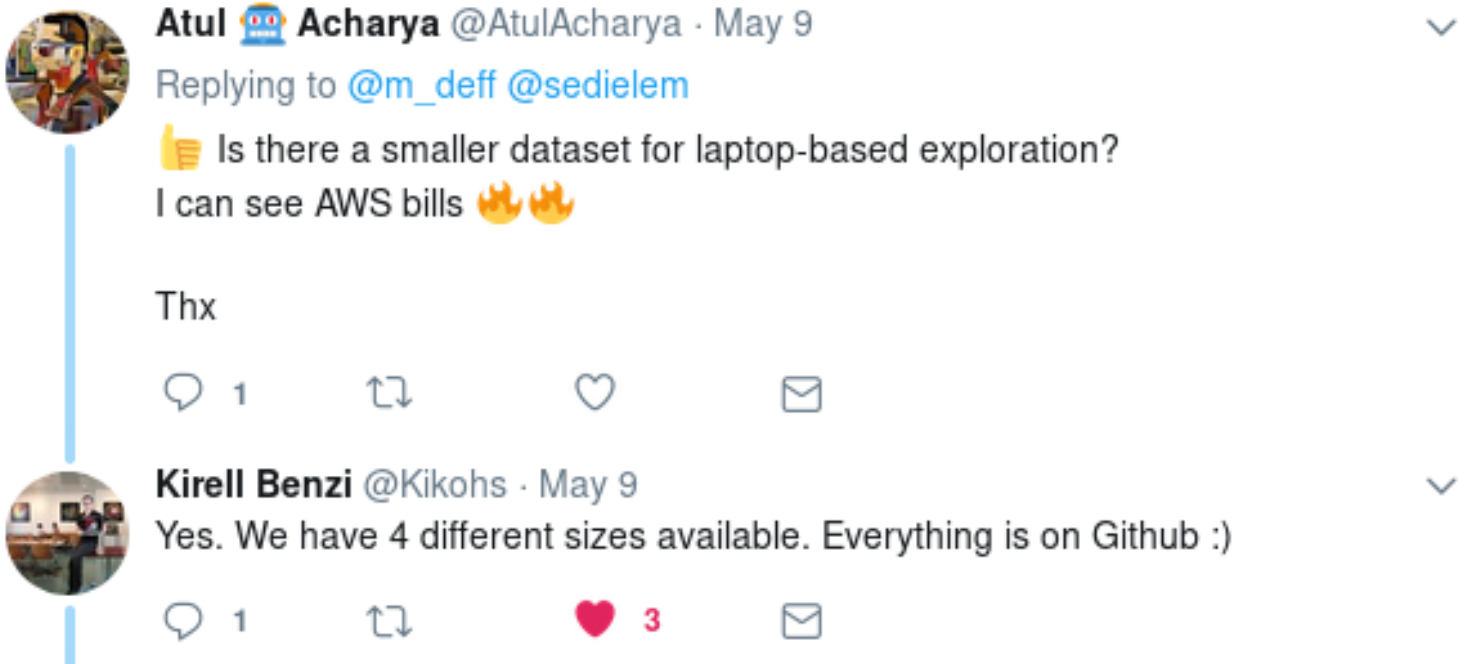
| | |
|---|---|
| Audio | various Creative Commons and similar, by the artists |
| Metadata | CC BY 4.0 |
| Code | MIT |

## Data formats

| | |
|---|---|
| Audio | mp3, by the artists |
| Metadata | csv |
| Code | Python |

# Challenges

Anticipate users



# How does it look like?

https://github.com/mdeff/fma

## FMA: A Dataset For Music Analysis

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, Xavier Bresson, EPFL LTS2.

### Data

All metadata and features for all tracks are distributed in **fma_metadata.zip** (342 MiB). The below tables can be used with pandas or any other data analysis tool. See the paper or the usage notebook for a description.

- `tracks.csv` : per track metadata such as ID, title, artist, genres, tags and play counts, for all 106,574 tracks.
- `genres.csv` : all 163 genre IDs with their name and parent (used to infer the genre hierarchy and top-level genres).
- `features.csv` : common features extracted with librosa.
- `echonest.csv` : audio features provided by Echonest (now Spotify) for a subset of 13,129 tracks.

Then, you got various sizes of MP3-encoded audio data:

1. **fma_small.zip**: 8,000 tracks of 30s, 8 balanced genres (GTZAN-like) (7.2 GiB)
2. **fma_medium.zip**: 25,000 tracks of 30s, 16 unbalanced genres (22 GiB)
3. **fma_large.zip**: 106,574 tracks of 30s, 161 unbalanced genres (93 GiB)
4. **fma_full.zip**: 106,574 untrimmed tracks, 161 unbalanced genres (879 GiB)

### Code

The following notebooks and scripts, stored in this repository, have been developed for the dataset.

1. usage: shows how to load the datasets and develop, train and test your own models with it.

### Usage

1. Download some data, verify its integrity, and uncompress the archives.

### History

### Contributing

Please open an issue or a pull request if you want to contribute. Let's try to keep this repository the central place around the dataset! Links to resources related to the dataset are welcome. I hope the community will like it and that we can keep it lively by evolving it toward people's needs.

### License & co

# Why code for a dataset?

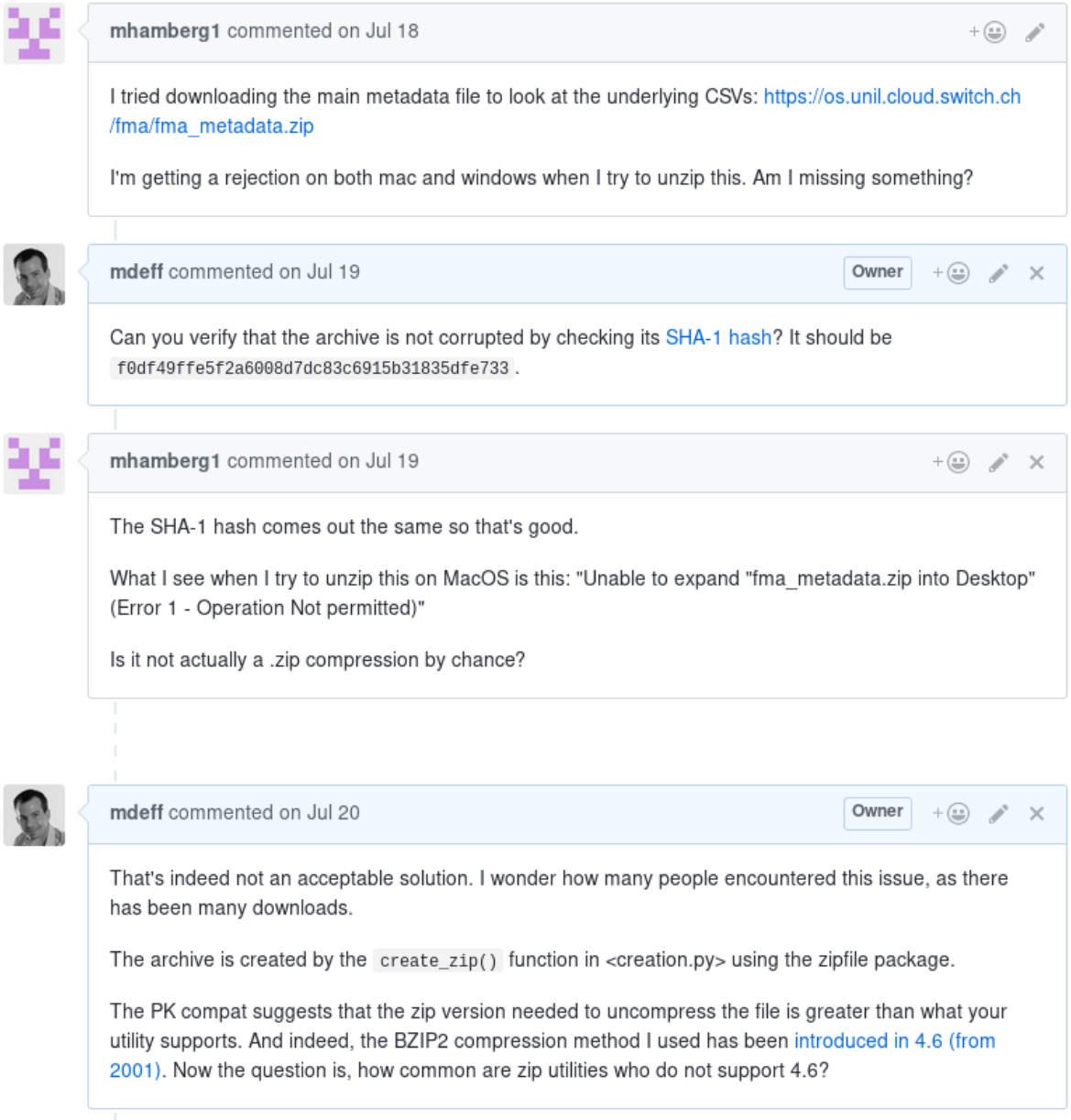
- Show how it was collected and cleaned. Reproducibility!
- Data exploration. Figures for the paper!
- Baselines for some tasks.
- A guide to import and use the data.

# Pros of openness

- Forces you to make it clean.
- Makes the work more useful.
- Be transparent!
- Feedback.

# Pros of openness

https://github.com/mdeff/fma



**mhamberg1** commented on Jul 18

I tried downloading the main metadata file to look at the underlying CSVs: https://os.unil.cloud.switch.ch/fma/fma_metadata.zip

I'm getting a rejection on both mac and windows when I try to unzip this. Am I missing something?

**mdeff** commented on Jul 19 — Owner

Can you verify that the archive is not corrupted by checking its SHA-1 hash? It should be `f0df49ffe5f2a6008d7dc83c6915b31835dfe733`.

**mhamberg1** commented on Jul 19

The SHA-1 hash comes out the same so that's good.

What I see when I try to unzip this on MacOS is this: "Unable to expand "fma_metadata.zip into Desktop" (Error 1 - Operation Not permitted)"

Is it not actually a .zip compression by chance?

**mdeff** commented on Jul 20 — Owner

That's indeed not an acceptable solution. I wonder how many people encountered this issue, as there has been many downloads.

The archive is created by the `create_zip()` function in <creation.py> using the zipfile package.

The PK compat suggests that the zip version needed to uncompress the file is greater than what your utility supports. And indeed, the BZIP2 compression method I used has been introduced in 4.6 (from 2001). Now the question is, how common are zip utilities who do not support 4.6?

# Pros of openness

https://twitter.com/m_deff/status/861985446116589569

mdeff / fma

Unwatch 28 | Unstar 377 | Fork 59

Code | Issues 3 | Pull requests 0 | Projects 0 | Wiki | Settings | Insights

FMA: A Dataset For Music Analysis https://arxiv.org/abs/1612.01840 Edit

dataset | music-analysis | music-information-retrieval | deep-learning | open-data | open-science | reproducible-research | Manage topics

**Michaël Defferrard**
@m_deff

Our FMA dataset is online! 106,574 songs, 161 hierarchical genres, 917 GiB, 343 days of audio under #creativecommons
github.com/mdeff/fma

| | | | |
|---|---|---|---|
| track_id | title | number | |
| information | language_code | license | |
| composer | publisher | lyricist | |
| genres | genres_all | genre_top | |
| duration | bit_rate | interest | |
| #listens | #comments | #favorites | tags |
| date_created | date_recorded | | |
| album_id | title | type | #tracks |
| information | engineer | producer | |
| #listens | #comments | #favorites | tags |
| date_created | date_released | | |
| artist_id | name | | |
| bio | members | associated_labels | |
| website | wikipedia_page | related_projects | |
| location | longitude | latitude | |
| | #comments | #favorites | tags |
| date_created | active_year_end | active_year_begin | |

Table 4. List of available per-track, per-album and per-artist metadata, i.e. the columns of `tracks.csv`.

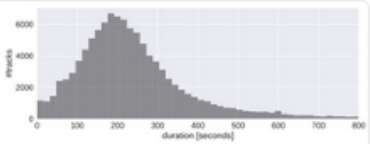



Figure 2. Track duration, from 0 to 3 hours.

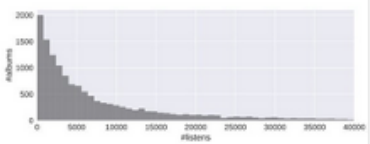



Figure 3. Album listens, from 0 to 3.6 millions.

RETWEETS 102 LIKES 178

6:45 PM - 9 May 2017

4 102 178

**Sander Dieleman**
@sedielem

Following

Replying to @Kikohs @m_deff @AtulAcharya

I wish this had existed when I was doing my PhD! Data drives research, so I think the impact of this on MIR could be massive. Amazing work!

RETWEET 1 LIKES 8

9:46 PM - 9 May 2017

1 8

cheyenne_h on 05/22/2017 at 06:15AM

## A Music Information Retrieval Dataset, Made With FMA

You may recall some news we shared last summer about a music dataset that was in progress - now it's complete!

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst & Xavier Bresson, a team of researchers interested in MIR (music information retrieval), have put together

# Cons of openness

- Maintenance.
- Support. Even more due to exposure!
- What about my competitive advantage?

# Cons of openness

https://github.com/mdeff/fma

**l3lackwood** commented 4 days ago

i have tried different ways to implement this codes but every time i found a new error
can you help me how can i use these codes
i had installed python 3.5 and all of packages in requirements after that i run the creation and i found this error :
C:\Users\l3lackwood\Downloads\WinPython-64bit-3.5.3.1Qt5\python-3.5.3.amd64\lib\site-packages\dotenv\main.py:24: UserWarning: Not loading - it doesn't exist.
warnings.warn("Not loading %s - it doesn't exist." % dotenv_path)
Traceback (most recent call last):
File "F:\farideh\python\WinPython-32bit-3.5.3.1Qt5\notebooks\genre\fma-master\creation.py", line 232, in
if sys.argv[1] == 'metadata':
IndexError: list index out of range

what should i do ?

**mdeff** commented 3 days ago — Owner

You need to add the path to the audio files, i.e. where you decompressed one of the `fma_*.zip` files containing `mp3`, in a `.env` file, like `AUDIO_DIR=/path/to/audio` (point 5 of the usage instructions). That file should be in the root directory, i.e. where the `creation.py` file resides. Without this, the `creation.py` script does not know where to search for those audio files.

👍 1

Slides https://doi.org/10.5281/zenodo.999353

Paper Defferrard, Benzi, Vandergheynst, Bresson,
FMA: A Dataset For Music Analysis, ISMIR, 2017.
https://arxiv.org/abs/1612.01840

Code & Data https://github.com/mdeff/fma

# Thanks Questions?