

- Accelerate machine learning research for music information retrieval.
- Promote open evaluation and open data.
- Need both open benchmarking and standardized challenges.
- Available at <https://github.com/mdeff/fma>.

Key numbers

- 106,574 tracks from 16,341 artists and 14,854 albums.
- 917 GiB and 343 days of Creative Commons-licensed audio.
- Arranged in a hierarchical taxonomy of 161 genres.
- 518 pre-computed features per track.

Dataset

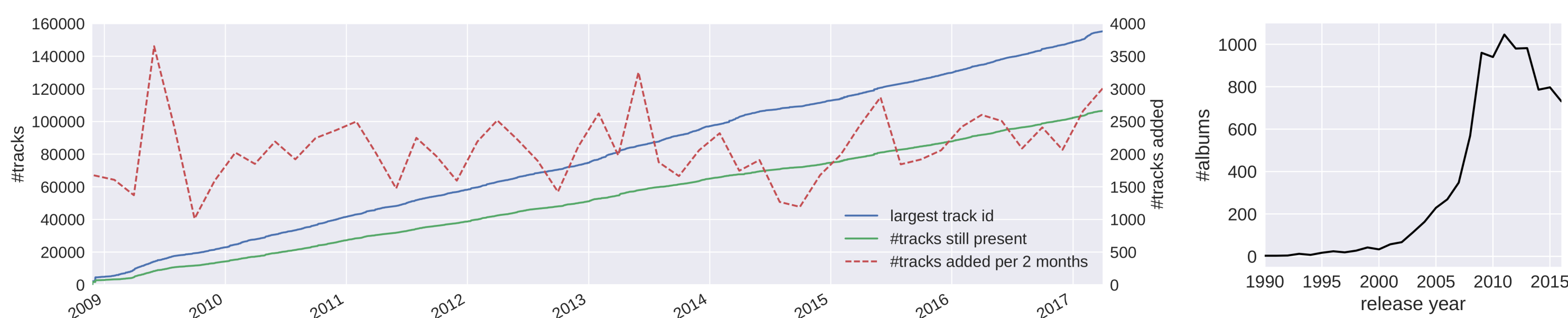


Figure: (left) Growth of the archive. (right) Number of albums released per year (1902 to 2017).

100% track_id	100% title	93% number
2% information	14% language_code	100% license
4% composer	1% publisher	1% lyricist
98% genres	98% genres_all	98% genres_top
100% duration	100% bit_rate	100% interest
100% #listens	2% #comments	61% #favorites
100% date_created	6% date_recorded	22% tags
100% album_id	100% title	
94% type	96% #tracks	
76% information	16% engineer	18% producer
97% #listens	12% #comments	38% #favorites
97% date_created	64% date_released	18% tags
100% artist_id	100% name	25% members
38% bio	5% associated_labels	
43% website	2% wikipedia_page	
	5% related_projects	
37% location	23% longitude	23% latitude
11% #comments	48% #favorites	10% tags
99% date_created	8% active_year_begin	
	2% active_year_end	

dataset	clips	genres	length	size	
			[s]	[GiB]	#days
small	8,000	8	30	7.4	2.8
medium	25,000	16	30	23	8.7
large	106,574	161	30	98	37
full	106,574	161	278	917	343

Table: Subsets of the FMA.

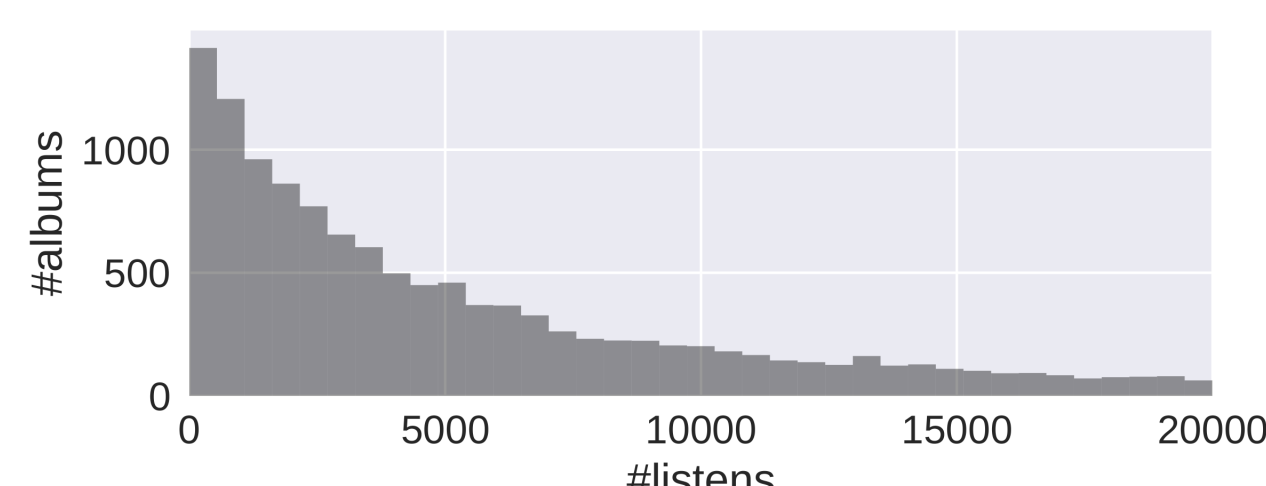


Figure: Album listens (max 3.6 millions).

Table: Available metadata. Percentages indicate coverage.

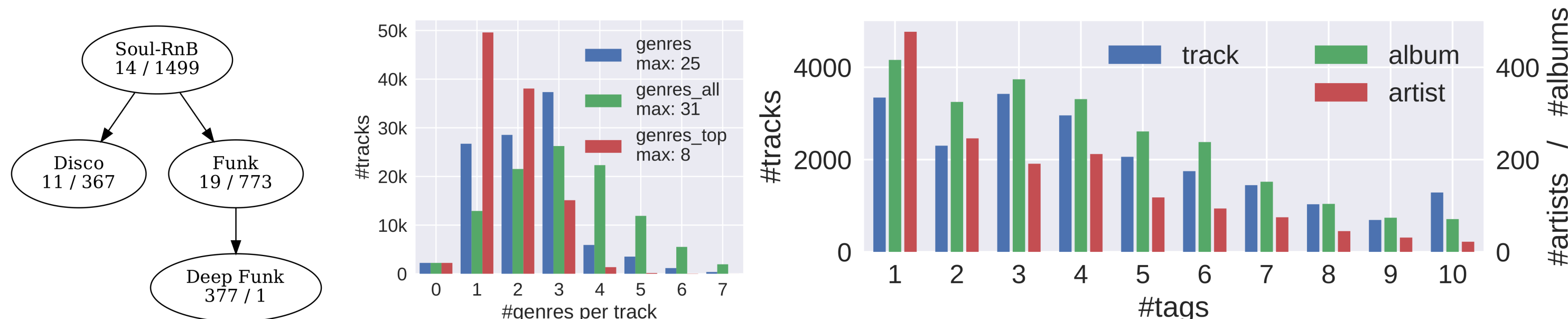


Figure: (left) Genre hierarchy. Numbers: genre_id / number of tracks per genre. (middle) Number of genres per track. (right) Per-track, album and artist tags (min 0, max 150).

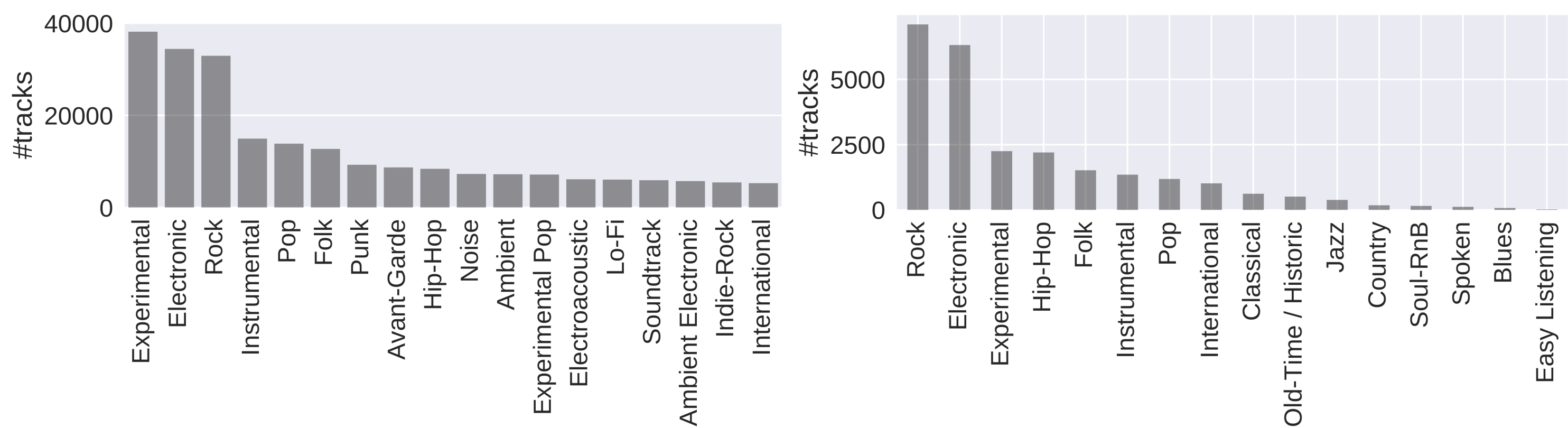
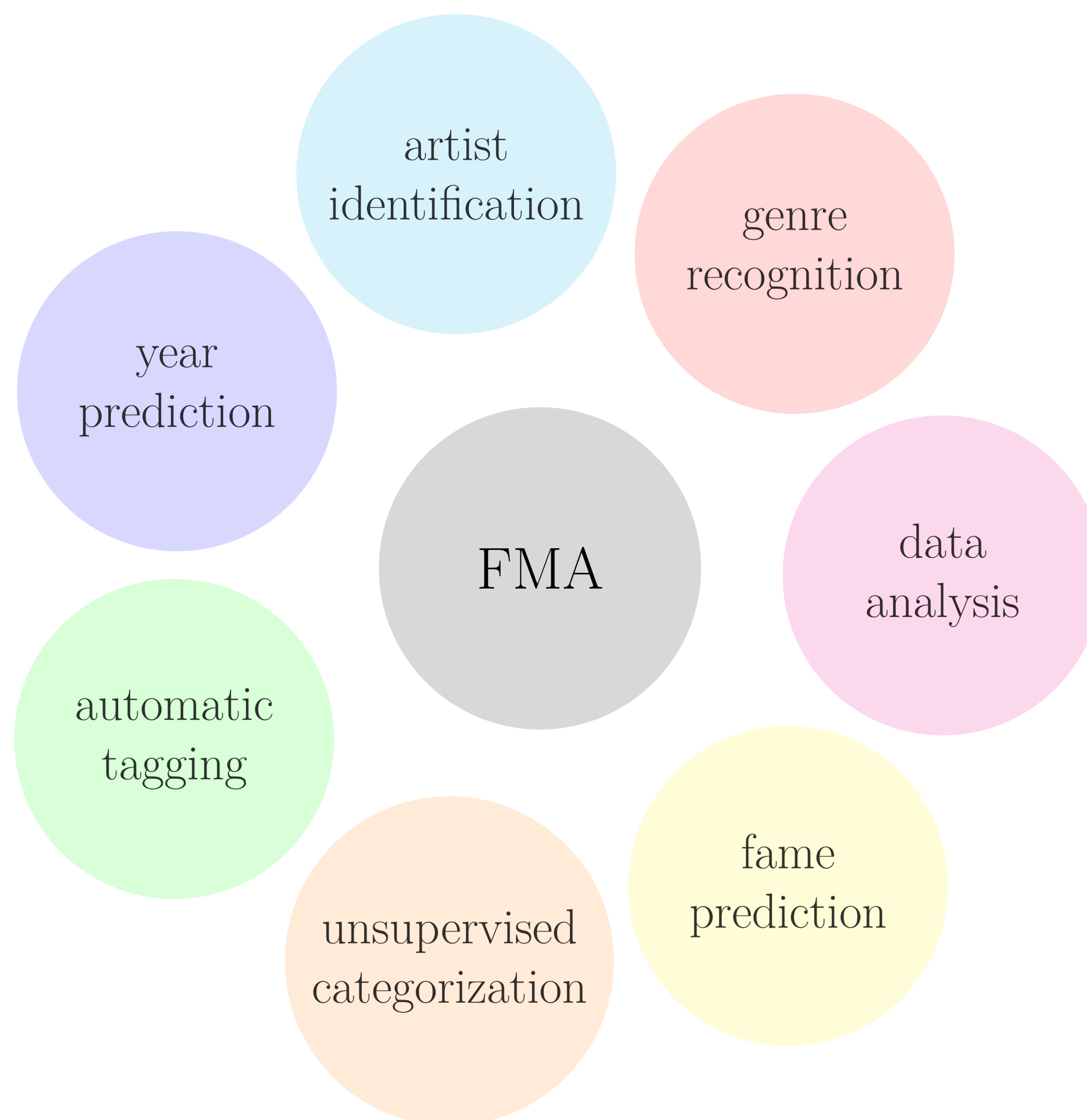


Figure: (left) Tracks per (sub-)genre on the full set (min 1, max 38,154). (right) Tracks per all 16 root genres on the medium subset (min 21, max 7,103).

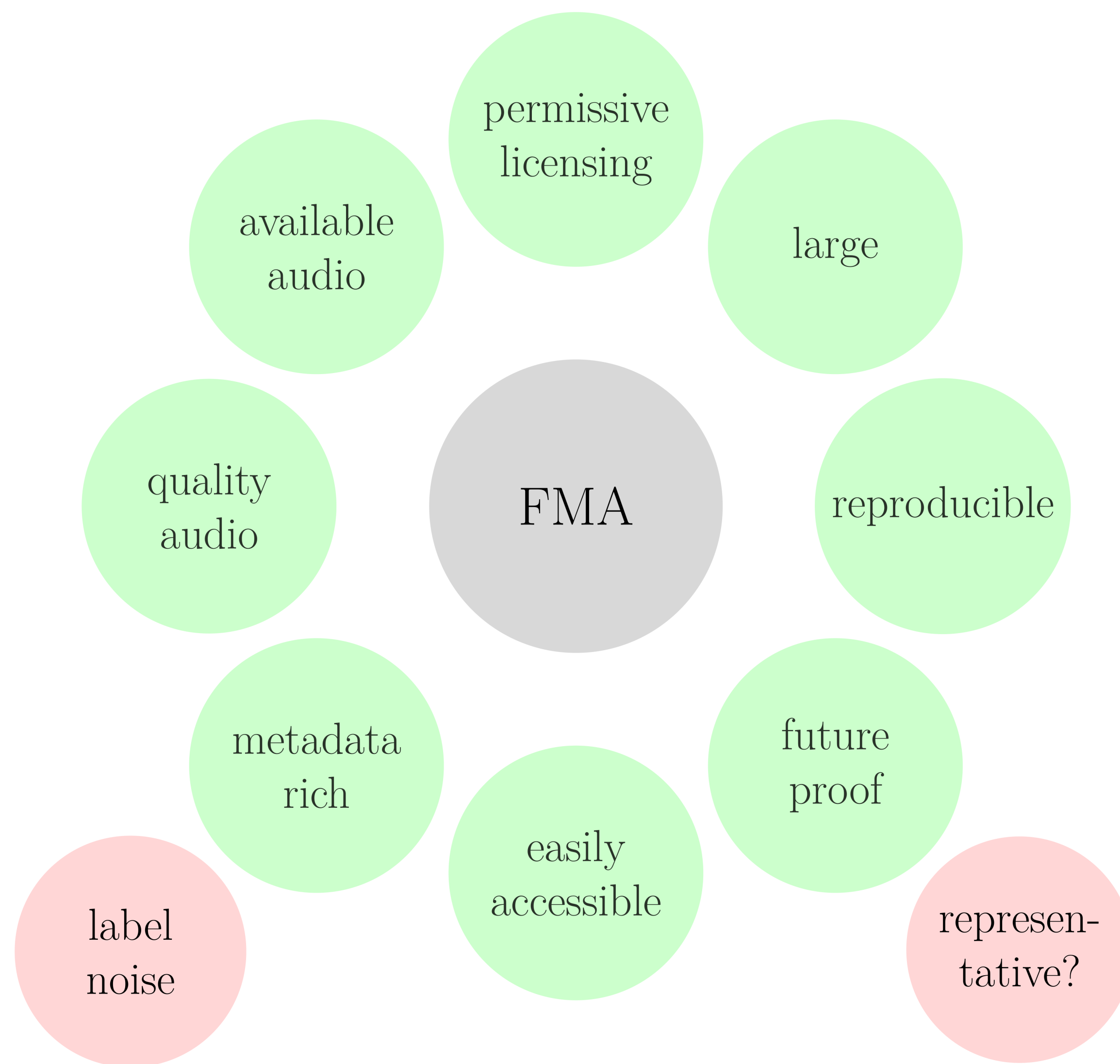
track id	title	genres	genres_all	genres_top	dur.	listens	album title	listens	tags	artist name	long.	lat.
76217	Jared C. Balogh...	[1, 4, 38...	[1, 322...	[4, 5, 38]	340	594	Resonant Tableaux	8018	[balogh, visual...	Charles Rice...	-94.6	39.1
67523	a	[32]	[32, 38]	[38]	75	358	Untitled	772	[noise, beats, exper...	A. P. Vague	-75.2	40.0
60161	Room of Our Own	[38, 71, 94]	[38, 71, 12...	[17, 12, 38]	244	413	Live at WFMU...	3727	[vocal experimental...	Muscles of Joy	-4.2	55.9
46779	Onwards...	[15, 66]	[66, 12, 15]	[12, 15]	195	455	Onwards	2354	[mge, maine]	Jeremy...	-119.3	37.3
66139	Little Respect...	[10, 15, 362]	[10, 362, 15]	[10, 15]	196	1397	Halo	36288	[electronic, indie-electro...	Blackbird...	-122.3	37.8
39897	Invisible Roots...	[1, 32, 38]	[32, 1, 38]	[38]	446	237	Violin Massage	10945	[improv, noise, fin...	Uton	-3.1	48.1
129505	Downhill Mader	[1, 38...	[1, 38, 41...	[20, 38]	354	524	The Singles	9107	[aufait, ungetreu...	suRRism	8.7	50.1

Table: Some rows and columns of the metadata table.

Usage



Qualities & Limitations



Encouraging reception



The screenshot shows the GitHub repository for 'mdeff / fma' and two Twitter posts. The first tweet is from Sander Dieleman (@sedielem) replying to @Kikohs, @m_deff, and @AtuAcharya, expressing his wish that the dataset had existed when he was doing his PhD. The second tweet is from Keunwoo Choi (@keunwoochoi) replying to the same users, stating that he used the dataset last month and encouraging others to stop using GTZAN and start using FMA.

What's next

- Data Jam Days, November 24–25, Lausanne.
- Challenge in collaboration with crowdAI (crowdai.org) and the Swiss Data Science Center (datascience.ch).
- Ground truths: cross-referencing and crowd-sourcing.
- What would you like to see?

