MIPFrontiers

Universitat Pompeu Fabra Barcelona

MTG Music Technology Group

Exploration of Music Collections with Audio Embeddings

Philip Tovstogan [UPF1]
Supervisors: Dmitry Bogdanov, Xavier Serra
Music Technology Group, Universitat Pompeu Fabra
MIP-Frontiers Final Workshop, Online, Oct 14-15, 2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 765068

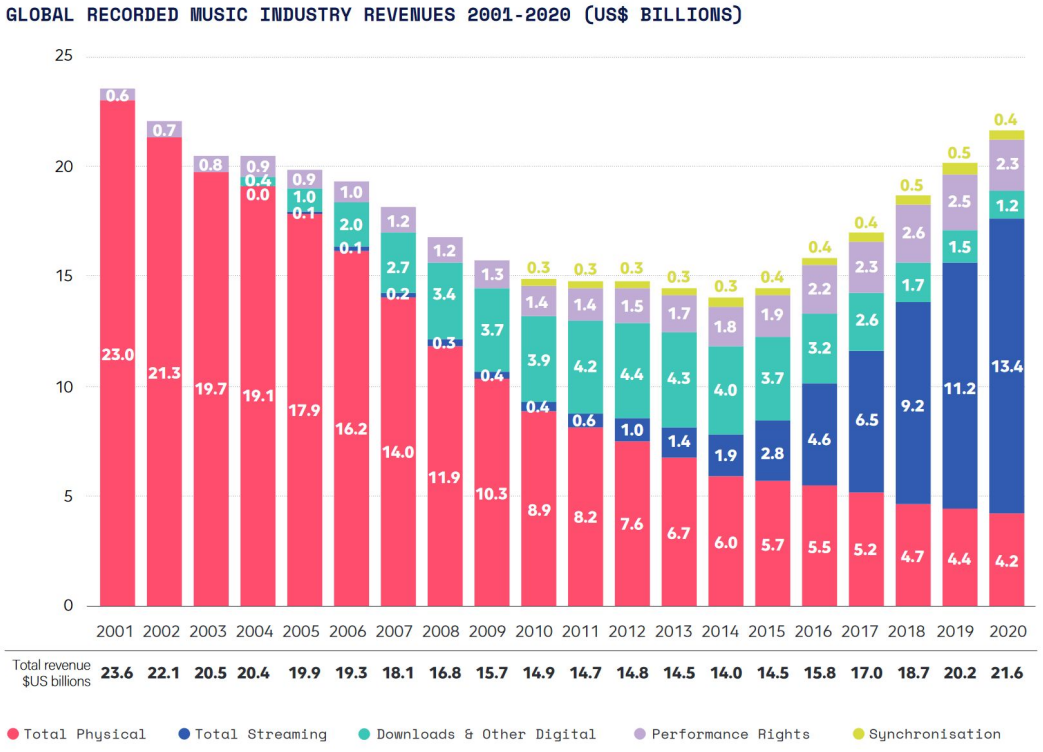
Project

UPF1: Exploration of Music Collections with Audio Embeddings

(original title: Tag propagation from structured to unstructured audio collections)

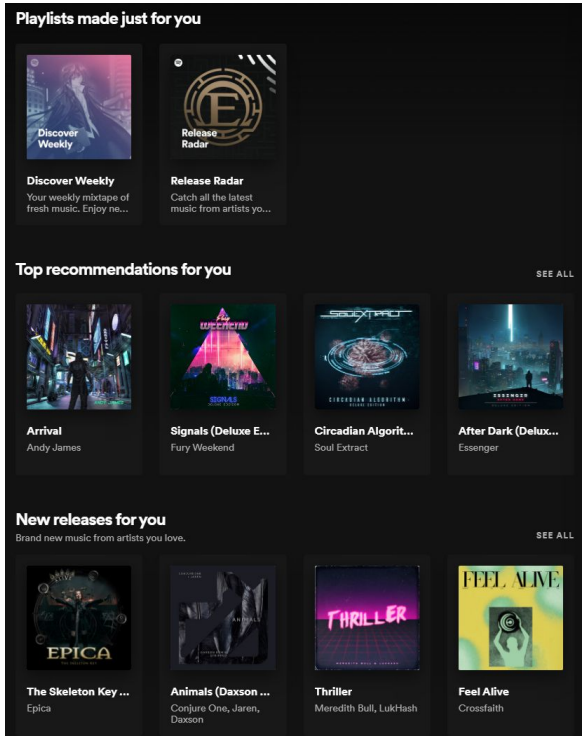
- Institution: Music Technology Group, UPF, Barcelona
- Secondment: Jamendo, Luxembourg
- Supervisor: Xavier Serra, Dmitry Bogdanov

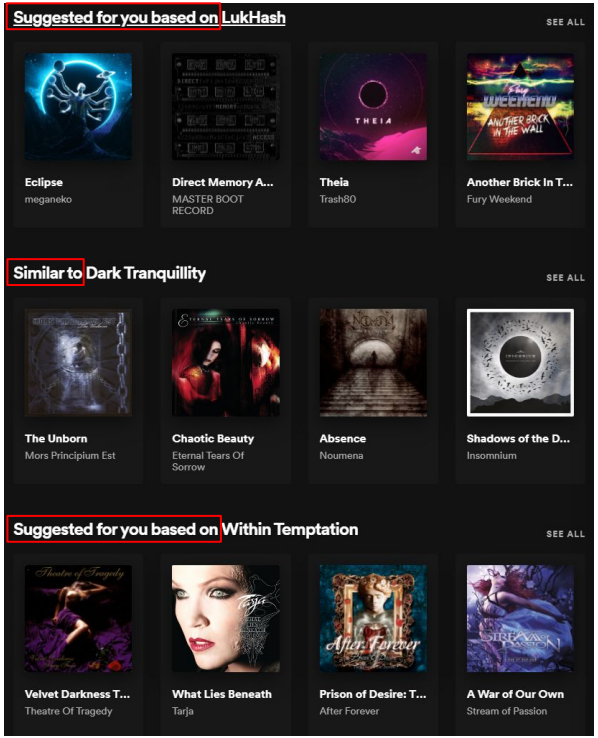
Music Industry and Streaming

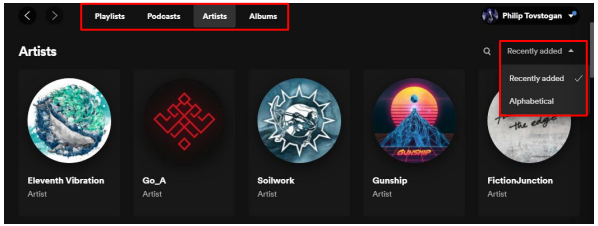


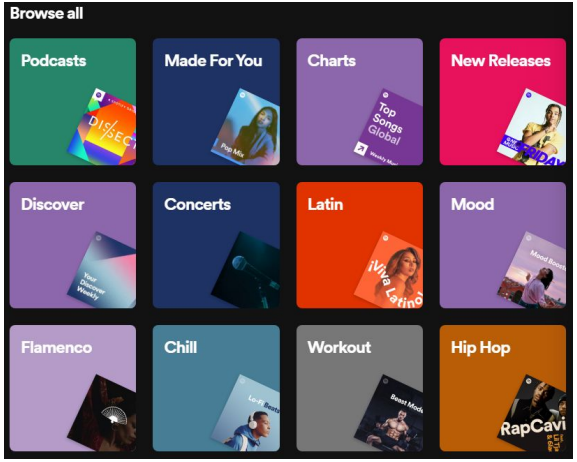
IFPI. (2021). Global music report 2021

Industry: RecSys, Interfaces, Browsing

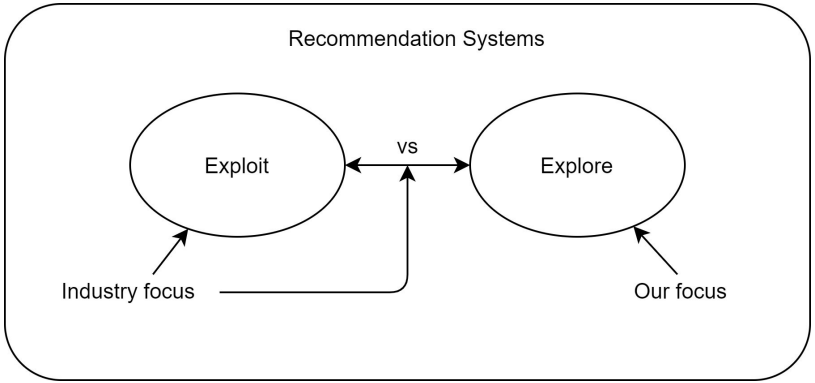








Recommendations: Explore vs Exploit



Exploit

- Predict what user would like (user retention, **passive**)
- **Short-term** reward, lean-out, safe

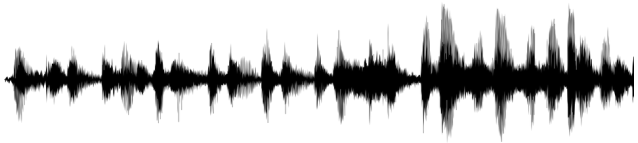
Explore

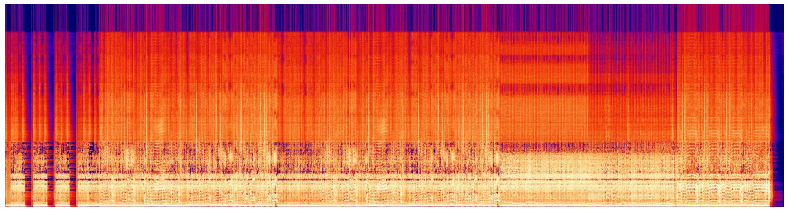
- Expose user to **something new**, guide him into unknown (**active**)
- **Long-term** reward, lean-in, risky

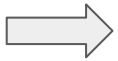
Barraza-Urbina, A. (2017). The exploration-exploitation trade-off in interactive recommender systems. RecSys, 431–435

Exploration and Auto-Tagging

Music exploration → using categories, thematic playlists, tags







Existing Open Auto-Tagging Datasets

Name	Tracks	Artists	Tags	Audio	Split
Million Song Dataset	505 216	-	522 366	N/A*	1*
MagnaTagATune	25 877*	230	188	Poor	No
Free Music Archive	106 574	16 341	161	Good	1
Music4All	109 269	16 269	19 541	Good	No
Melon	649 091	-	30 652	Spec.	No
MTG-Jamendo	55 609	3 565	195	Good	5

Non-uniform qlt. of audio

Need to request audio

Only specs. of 10-30s

Table 2.1: Auto-tagging datasets

Bogdanov, D., Won, M., Tovstogan, P., Porter, A., & Serra, X. (2019, June). The MTG-Jamendo dataset for automatic music tagging. ML4MD, ICML.

The MTG-Jamendo Dataset

- Creative Commons license
- Quality audio and labels (curated by Jamendo)
  - Also spectrograms and Essentia features
- Tag categories and subsets:
  - Genre, instrument, mood/theme, top50
- Tag pre-processing (*rockpop* → *poprock*)
- Standardized 5 splits without artist effect
- Reproducible pre-processing and baseline

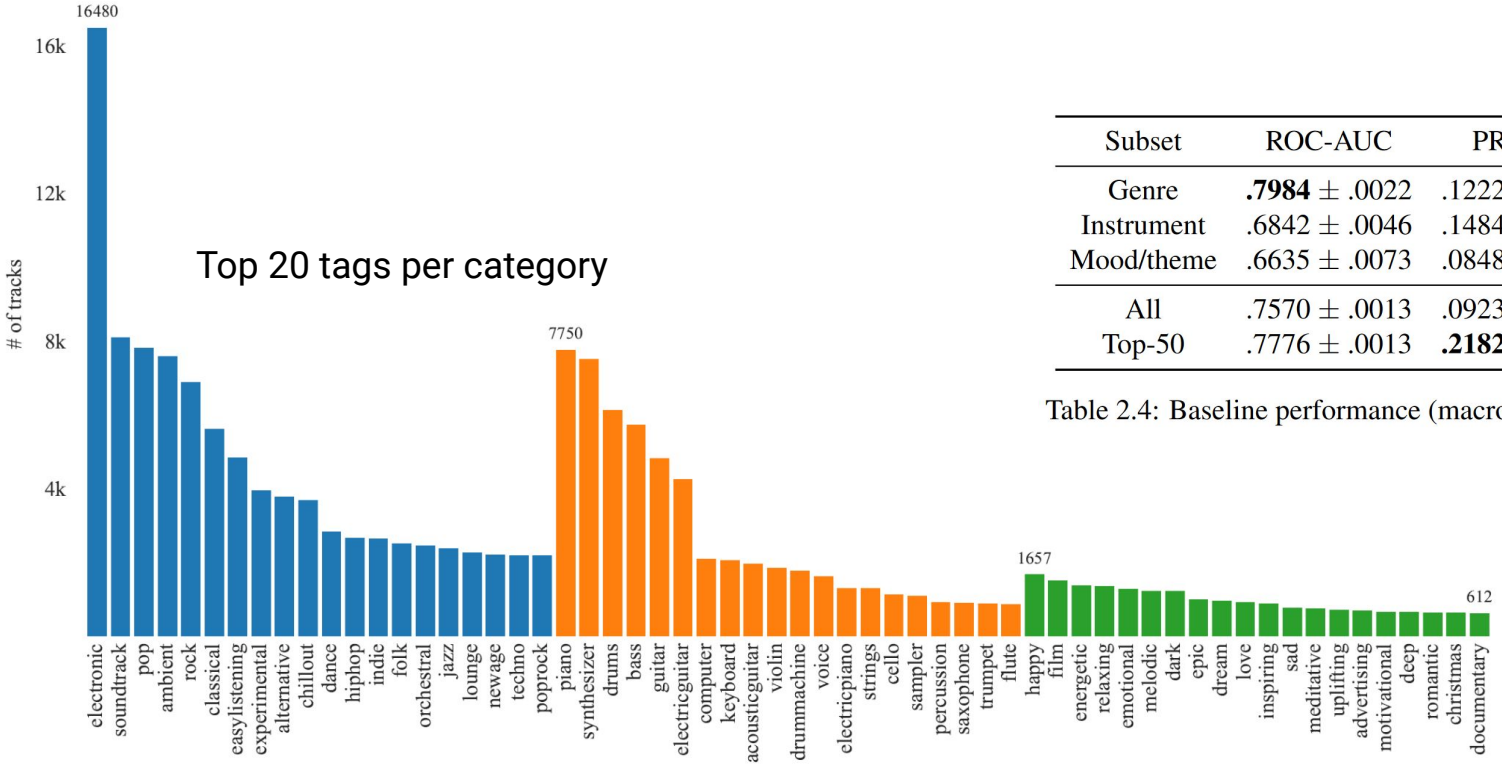
Category	Tags	Tracks	Albums	Artists
Genre	87	55 094	11 186	3 546
Instrument	40	24 976	5 672	2 003
Mood/theme	56	17 982	4 423	1 508
All	183	55 525	11 256	3 565
Top-50	50	54 380	11 107	3 517

Table 2.2: Statistics for category subsets

MTG/mtg-jamendo-dataset

Bogdanov, D., Won, M., Tovstogan, P., Porter, A., & Serra, X. (2019, June). The MTG-Jamendo dataset for automatic music tagging. ML4MD, ICML.

MTG-Jamendo: More Numbers



Subset	ROC-AUC	PR-AUC
Genre	**.7984** ± .0022	.1222 ± .0019
Instrument	.6842 ± .0046	.1484 ± .0030
Mood/theme	.6635 ± .0073	.0848 ± .0018
All	.7570 ± .0013	.0923 ± .0016
Top-50	.7776 ± .0013	**.2182** ± .0018

Table 2.4: Baseline performance (macro-averaged)

Bogdanov, D., Won, M., Tovstogan, P., Porter, A., & Serra, X. (2019, June). The MTG-Jamendo dataset for automatic music tagging. ML4MD, ICML.

Auto-Tagging Architectures: Background

- MusiCNN
  - musically-motivated CNN
  - vertical and horizontal filters
- VGG
  - computer vision
  - deep stack of 3×3 convolutional filters
  - adapted for audio
- VGGish
  - original implementation of VGG
  - 3087 output units

Methods	MTAT		MSD		MTG-Jamendo	
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
FCN [1]	0.9005	0.4295	0.8744	0.2970	0.8255	0.2801
FCN (with 128 Mel bins)	0.8994	0.4236	0.8742	0.2963	0.8245	0.2792
Musicnn [2]	0.9106	0.4493	0.8803	0.2983	0.8226	0.2713
Musicnn (with 128 Mel bins)	0.9092	0.4546	0.8788	3036	0.8275	0.2810
Sample-level [3]	0.9058	0.4422	0.8789	0.2959	0.8208	0.2742
Sample-level + SE [4]	0.9103	0.4520	0.8838	0.3109	0.8233	0.2784
CRNN [6]	0.8722	0.3625	0.8499	0.2469	0.7978	0.2358
CRNN (with 128 Mel bins)	0.8703	0.3601	0.8460	0.2330	0.7984	0.2378
Self-attention [7]	0.9077	0.4445	0.8810	0.3103	0.8261	0.2883
Harmonic CNN [9]	0.9127	0.4611	**0.8898**	**0.3298**	0.8322	0.2956
Short-chunk CNN	0.9126	0.4590	0.8883	0.3251	**0.8324**	**0.2976**
Short-chunk CNN + Res	**0.9129**	**0.4614**	**0.8898**	0.3280	0.8316	0.2951

Table 2: Performances of state-of-the-art models.

ESSENTIA

Architecture	Dataset	Classes	AUC-PR
MusiCNN	MSD [22]	50	88.01
MusiCNN	MTT [23]	50	90.69
VGG	MSD [22]	50	87.67
VGG	MTT [23]	50	90.26
VGG	Audioset [18]	3087	91.00

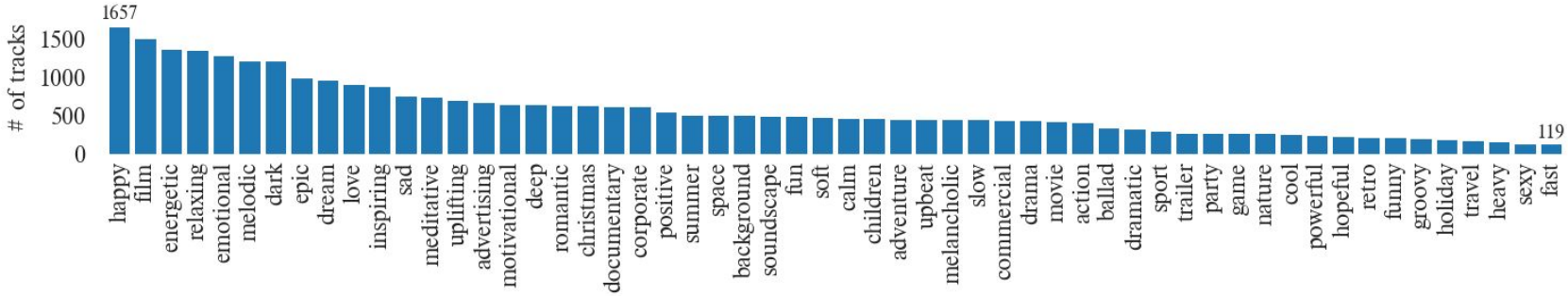
Table 1: State-of-the-art CNN models included in Essentia.

Won, M., Ferraro, A., Bogdanov, D., & Serra, X. (2020). Evaluation of CNN-based automatic music tagging models. SMC, 331–337.
Alonso-Jiménez, P., Bogdanov, D., Pons, J., & Serra, X. (2020). Tensorflow audio models in Essentia. ICASSP, 266–270.

MediaEval: Jamendo Moods & Themes

Mediaeval: the benchmarking initiative for multimedia evaluation, since 2010

- Goal: Invite teams to try their approaches to auto-tagging
- Data: MTG-Jamendo dataset mood/theme subset, **split-0**
  - Allowing usage of external data to pre-train
- Evaluation: PR-AUC on publicly available test set



Bogdanov, D., Porter, A., Tovstogan, P., & Won, M. (2020, December). MediaEval 2020: Emotion and theme recognition in music using Jamendo. MediaEval 2020

Mediaeval: Submissions

2020 (6/12)

- VGGish (MSD, Music4All), Mixup, different losses
- EfficientNet, WaveNet (NSynth), MobileNetV2, SpecAug
- CBAM, Self-Attention + RNNs
- ResNet, Self-Attention, Mixup, SpecAug
- VGGish, Self-Attention, AReLU, smaller nets
- CRNN, pre-processing, moods vs themes

2021 (?/10+)

2019 (6/14):

- Shake-FA-ResNet + FA-ResNet
- MobileNetV2, Self-attention
- Simple CNN
- CRNN models fusion (Audioset)
- CRNN, spectrograms + features
- Pre-trained VQ-VAE on MSD + CNN

Architectures, Data Augmentation

Mediaeval: Results from 2020 & 2019

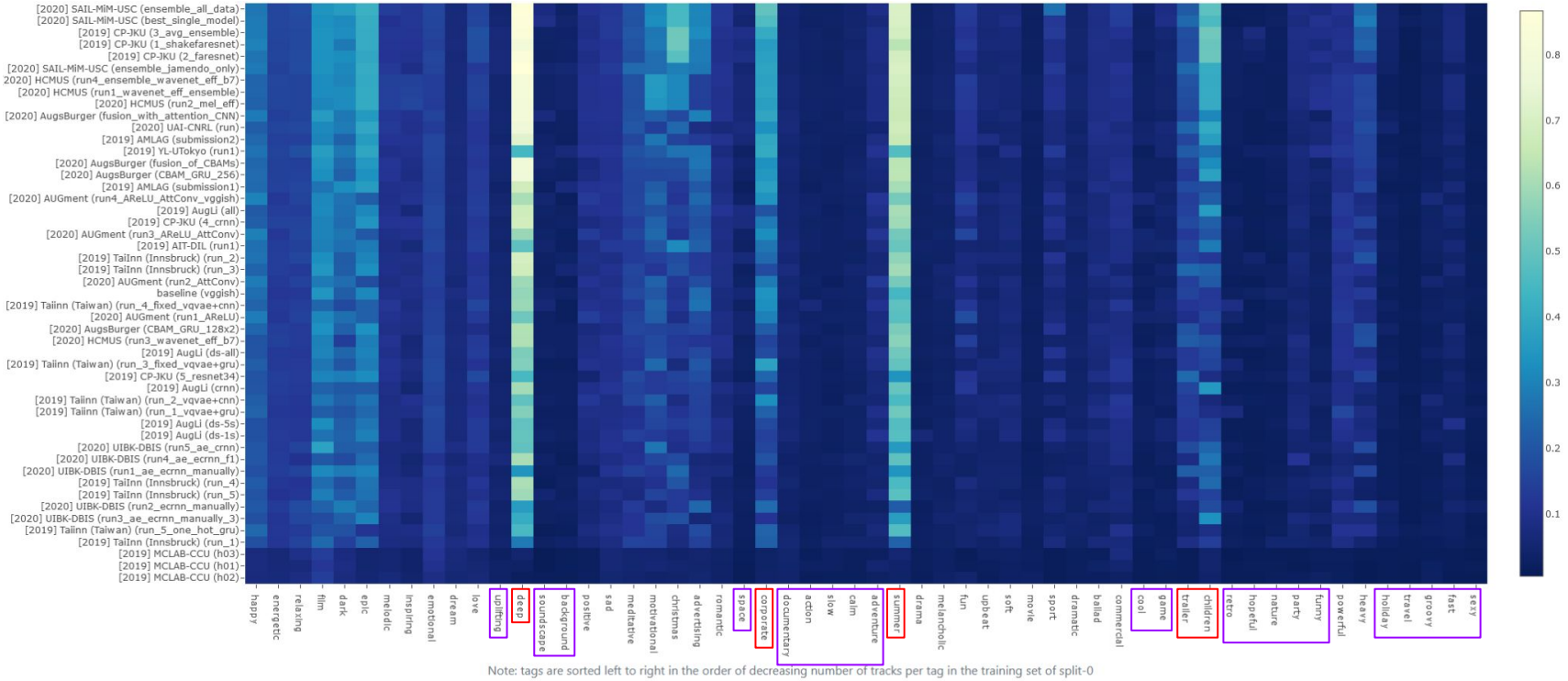
	PR-AUC	ROC-AUC	F-Score	Approach
Baseline VGG	.1077	.7258	.1656	VGG
	+ .0469	+ .0471	+ .0468	
Best 2019	**.1546**	.7729	.2124	Shake-FA-ResNet + FA-ResNet
	+ .0063	+ .0083	+ .0079	
Best 2020	**.1609**	.7812	.2203	VGGish (***MSD, Music4All***), Mixup, focal, CB, CD loss

2021 edition in progress!

multimediaeval/2020-Emotion-and-Theme-Recognition-in-Music-Task

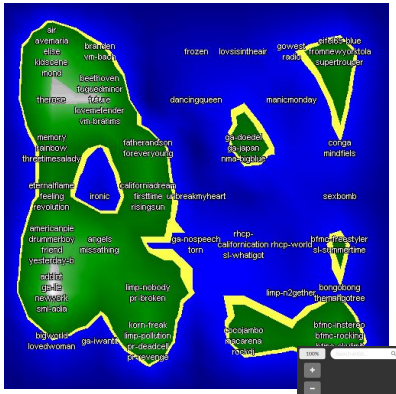
Knox, D., Greer, T., Ma, B., Kuo, E., Somandepalli, K., & Narayanan, S. (2020, December). MediaEval 2020 emotion and theme recognition in music task: Loss function approaches for multi-label music tagging. MediaEval 2020.
Koutini, K., Chowdhury, S., Haunschmid, V., Eghbal-Zadeh, H., & Widmer, G. (2019, October). Emotion and theme recognition in music with frequency-aware RF-regularized CNNs. MediaEval 2019.

Mediaeval: Per-Tag Performances

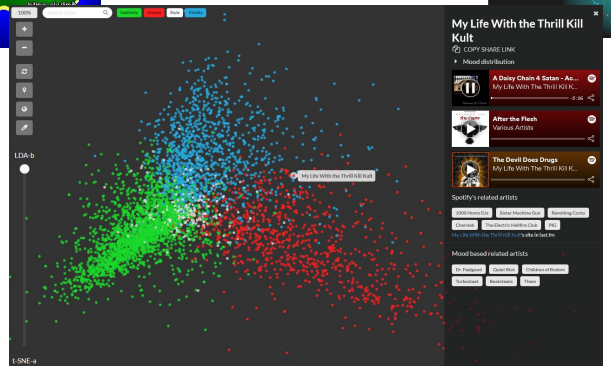


Note: tags are sorted left to right in the order of decreasing number of tracks per tag in the training set of split-0

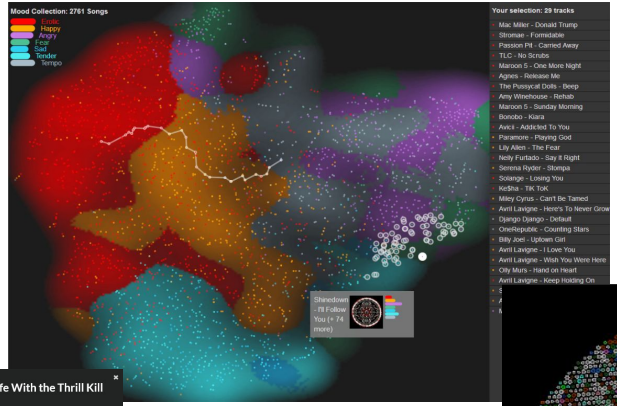
Interfaces: Background



Island of Music



MoodPlay



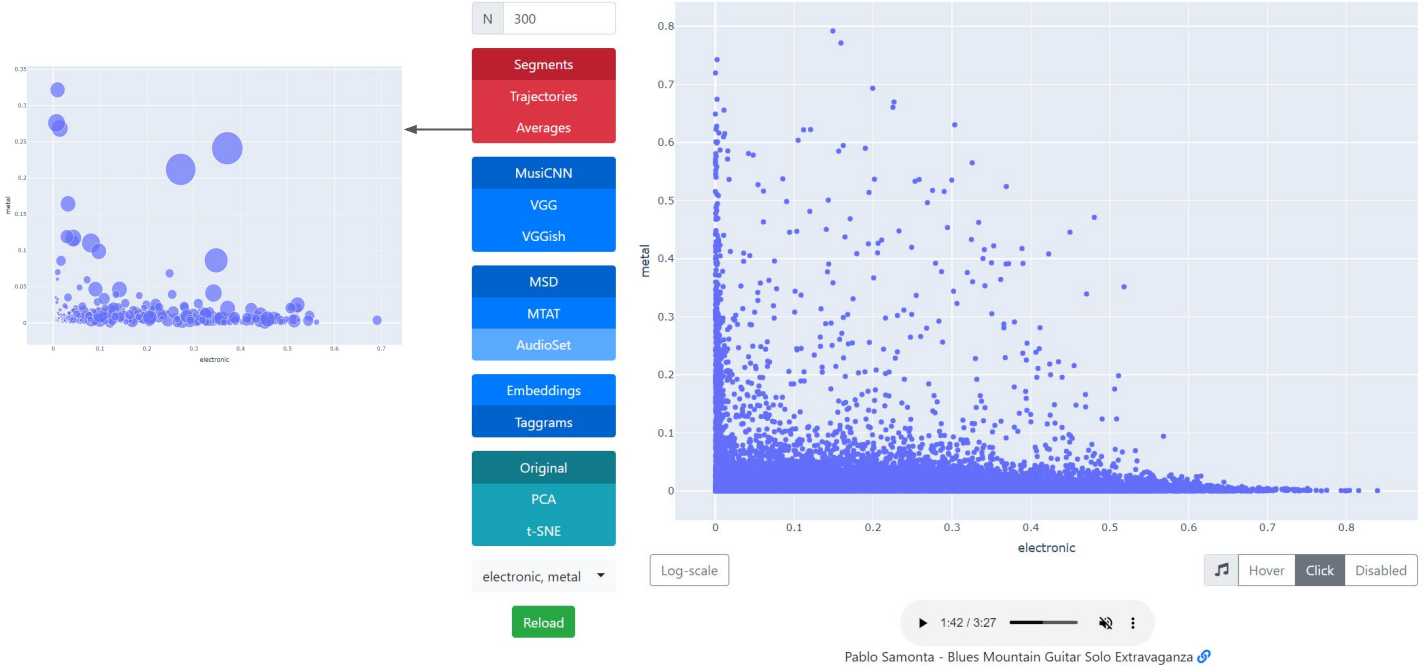
Vad et.al.

SongExplorer

Knees, P., Schedl, M., & Goto, M. (2020). Intelligent user interfaces for music discovery. Transactions of the International Society for Music Information Retrieval, 3(1), 165–179

Web Interface: Exploration & Evaluation

music-explore.upf.edu



MTG/music-explore

Tovstogan, P., Serra, X., & Bogdanov, D. (2020, July). Web interface for exploration of latent and tag spaces in music auto-tagging. ML4MD, ICML.

Web Interface: Second Iteration

Web Interface: Comparing Visualizations

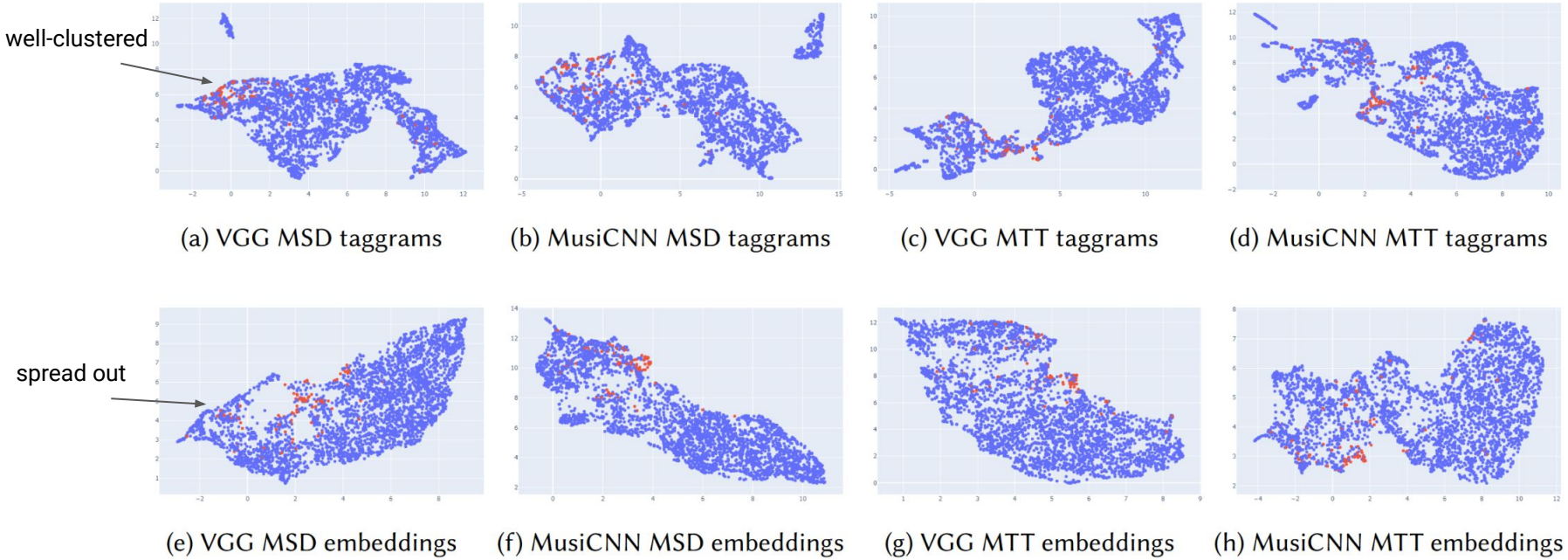
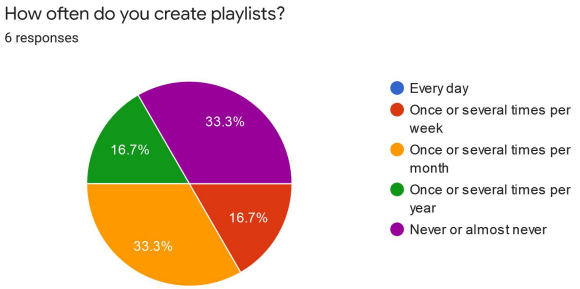


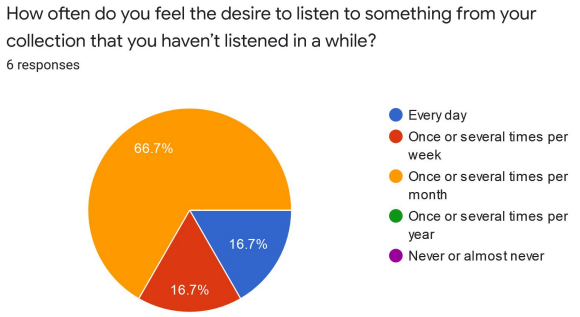
Fig. 3. UMAP visualizations of *new age* (in red) in mostly rock and metal collection (reduction of 20)

User Study: Participants' Background

User study: semi-structured interviews

- 6 participants so far
- Age: 27-36 years
- Music training: 1 to 20 years (mdn 6, avg 8)
- Listen to music: 0.5-8 hrs per day
  - Actively: less than 1 hr
  - To playlists: 0%-40% (avg 15%)





User Study: Interview and Feedback

- Personal music collections (600-1200 tracks)
- 10 min introduction of functionalities

Task: *Imagine that you want to listen to something from your library that you haven't listened in a while. Explore the system and make a playlist for yourself.*

Table 1. Summarized results from Likert scale questions

Question	Mean ± STD
Liked interacting with system	**4.8** ± 0.4
Had preference for particular model	3.5 ± **1.4**
Preferred over browsing	**4.5** ± 0.5
Preferred over random	4.3 ± 1.0
Liked big picture	3.5 ± 1.0
Liked segment groupings	4.3 ± 0.8
Discovered unexpected connections	**4.7** ± 0.5
Rediscovered something	**4.5** ± 1.2
Want to use for playlist creation	4.2 ± 1.2
Want to use for inspiration	**4.7** ± 0.6
Had rewarding experience	4.2 ± **1.3**
Had engaging experience	**4.7** ± 0.8

User Study: Preferences

- Preferred particular combination (6):
  - VGG-MSD (3)
  - VGG-MTT (3)
  - MusiCNN-MTT (2)
- Taggrams (4) > Embeddings (2)
- Projections:
  - PCA, STD-PCA → big picture
  - t-SNE and UMAP → zooming in

"I would never think to put these two artists together in a playlist, but it works quite well for these tracks"

"It seems that (MusiCNN-MTAT) can separate ambient from drums, while (VGG-MSD) gets the timbral aspect of sounds together well"

Summary

- New auto-tagging dataset: MTG-Jamendo
  - 500 GB of 55,000+ CC-licensed full audio tracks with 190+ tags split into categories
- Organizing “Emotions and Themes in Music” task in Mediaeval 2019-2021
  - Improving state-of-the-art performance on challenging mood/theme subset
- Web interface for exploration of latent and tag spaces for MTG-Jamendo
  - Allows for quick qualitative evaluation and sanity check of models, exploration with tags
- Visualization of embeddings for exploration of personal collections
  - Evaluated with semi-structured interview, strong positive feedback

Thanks for listening!
Q&A

philip.tovstogan@upf.edu