Deep learning and feature learning for MIR

Sander Dieleman – July 23rd, 2014

PhD student at Ghent University

Working on audio-based music classification, recommendation, …

Graduating in December 2014

Currently interning at Spotify in NYC

http://benanne.github.io

http://github.com/benanne

http://reslab.elis.ugent.be

sanderdieleman@gmail.com

I. Multiscale music audio feature learning

II. Deep content-based music recommendation

III. End-to-end learning for music audio

IV. Transfer learning by supervised pre-training

V. More music recommendation + demo

I. Multiscale music audio feature learning

Feature learning is receiving more attention from the MIR community

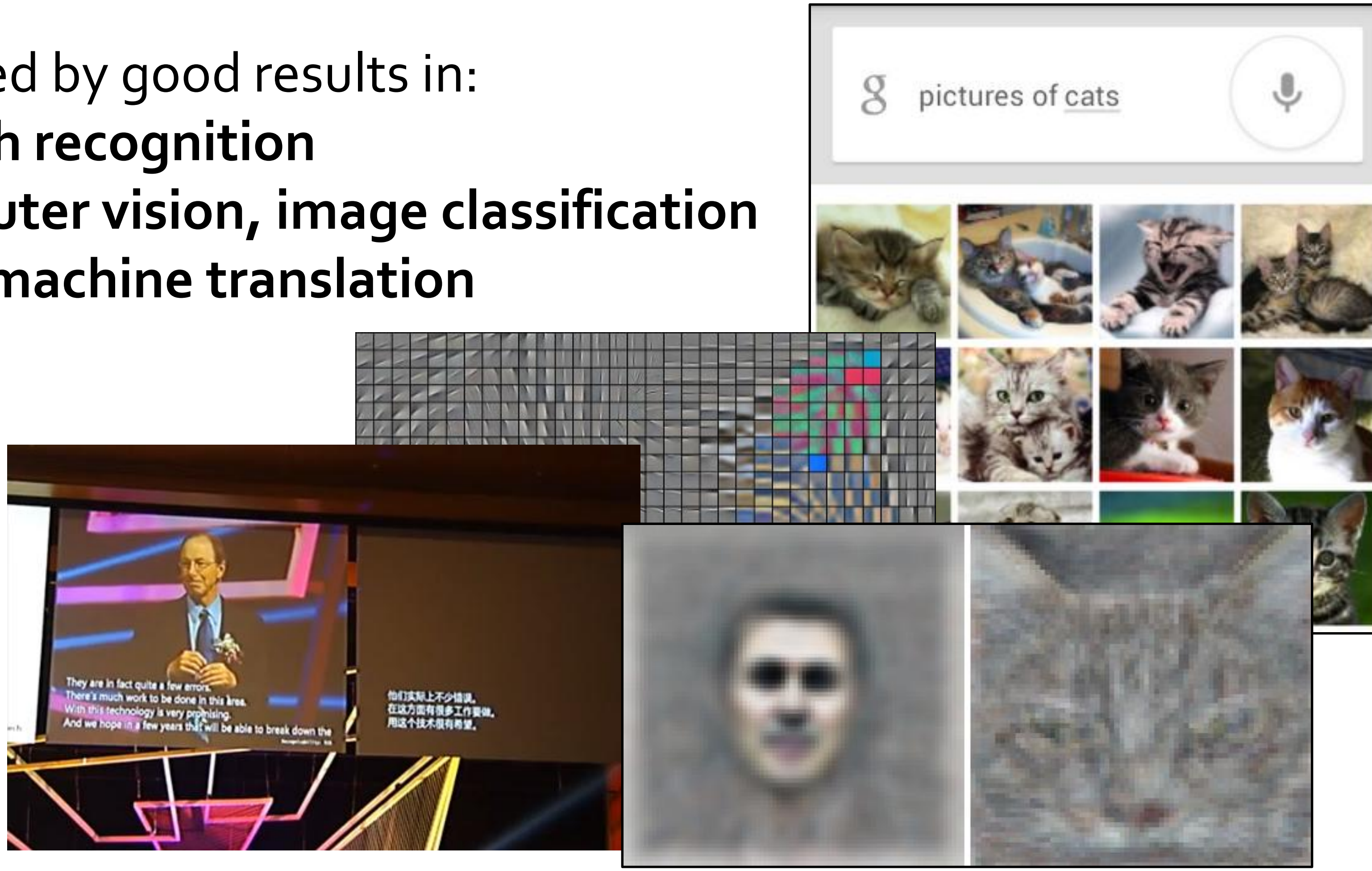
Inspired by good results in:

speech recognition

computer vision, image classification

NLP, machine translation

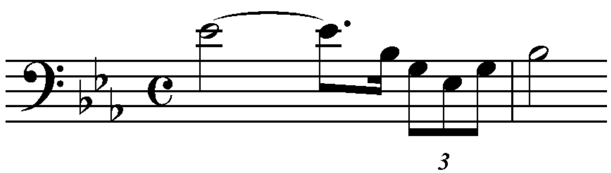
...



Music exhibits structure on **many different timescales**

Musical form

Themes



Motifs

Periodic waveforms

K-means for feature learning: cluster centers are features

Spherical K-means:
means lie on the unit sphere, have a **unit L2 norm**

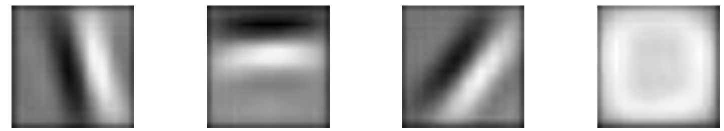
+ conceptually very **simple**

+ **only one parameter** to tune: number of means

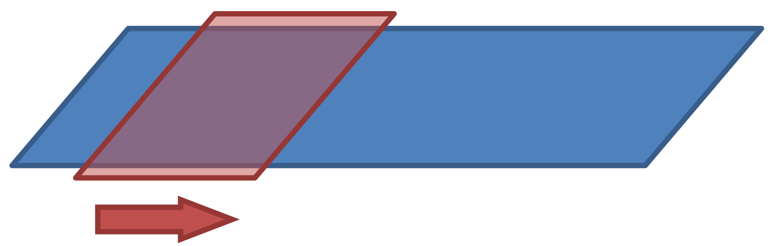
+ **orders of magnitude faster** than RBMs, autoencoders, sparse coding

(Coates and Ng, 2012)

Spherical K-means features work well with **linear feature encoding**



During training:	0	0	1.7	0	**One-of-K**
During feature extraction:	-0.2	2.3	1.7	0.7	**Linear**

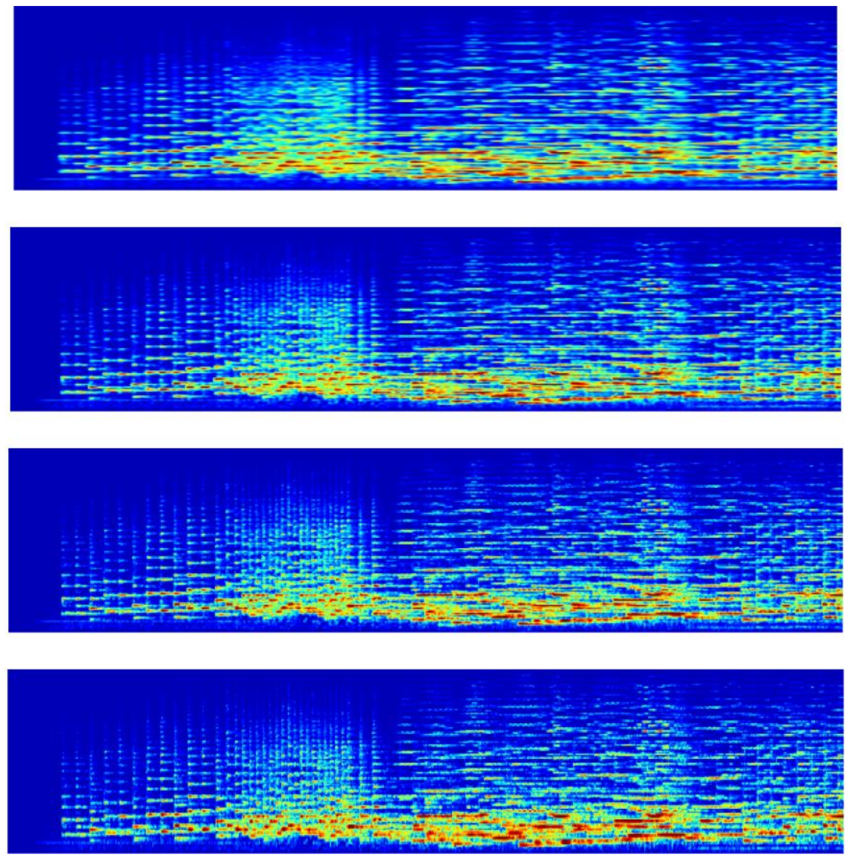


Feature extraction is a **convolution** operation

(Coates and Ng, 2012)

Multiresolution spectrograms: different window sizes

Coarse — 8192 samples



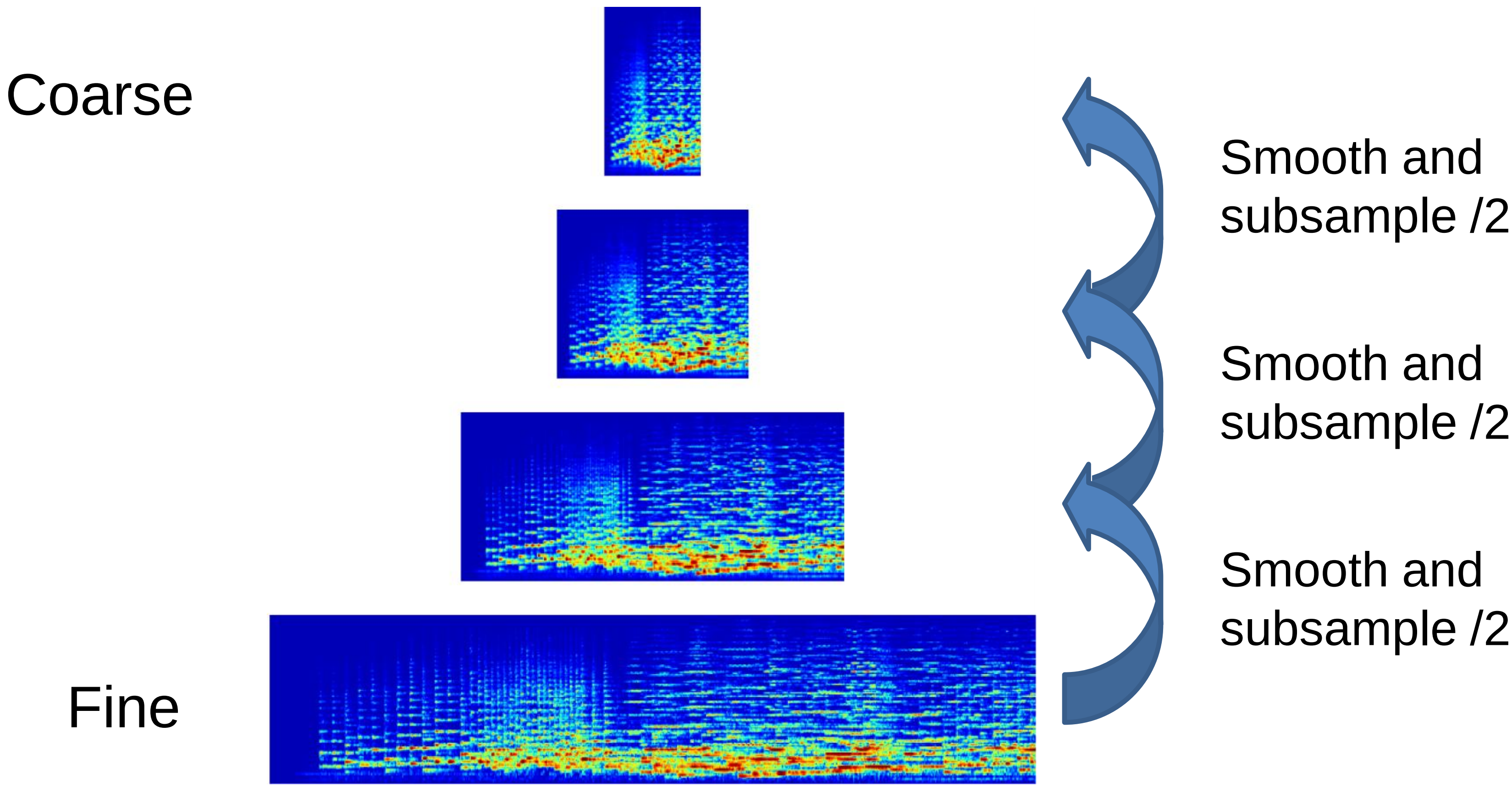
4096 samples

2048 samples

Fine — 1024 samples

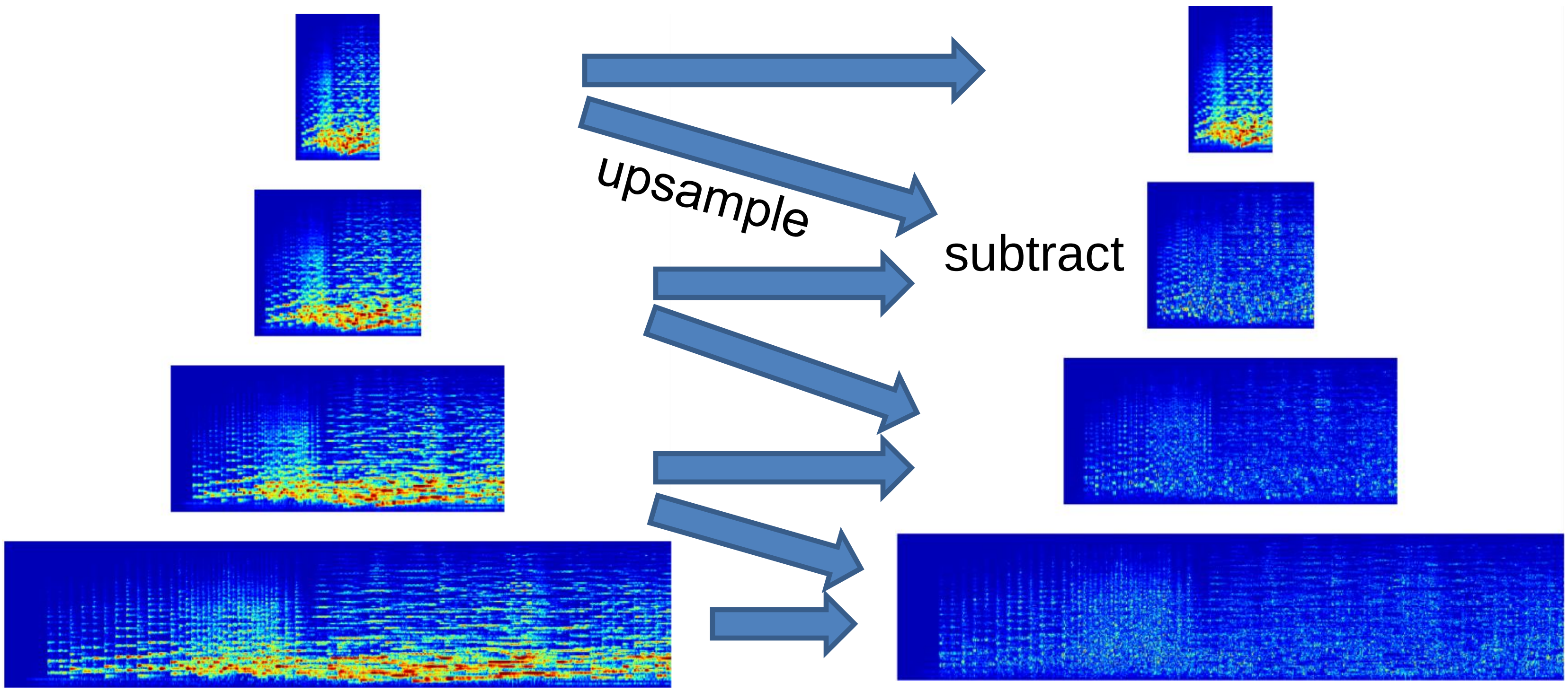
(Hamel et al., 2012)

Gaussian pyramid: repeated smoothing and subsampling



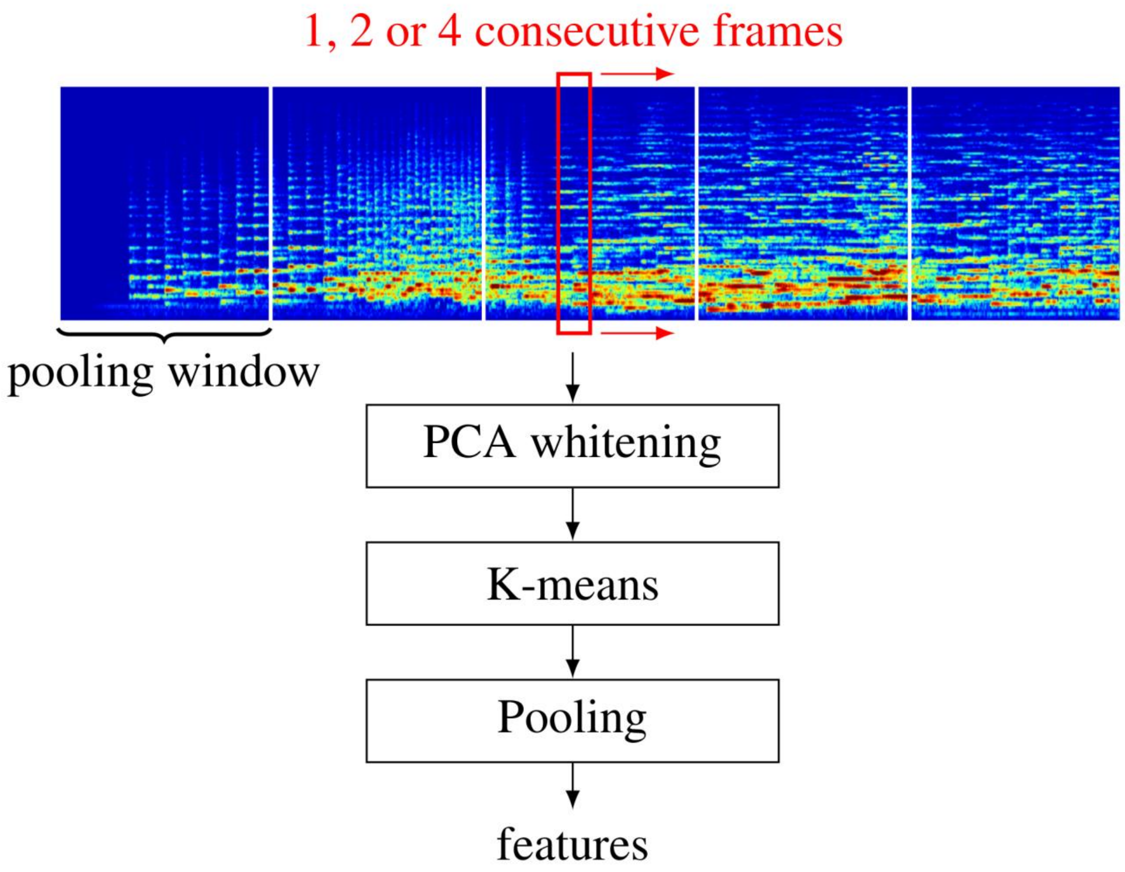
(Burt and Adelson, 1983)

Laplacian pyramid: difference between levels of the Gaussian pyramid



(Burt and Adelson, 1983)

Our approach: feature learning on multiple timescales



Task: **tag prediction** on the Magnatagatune dataset

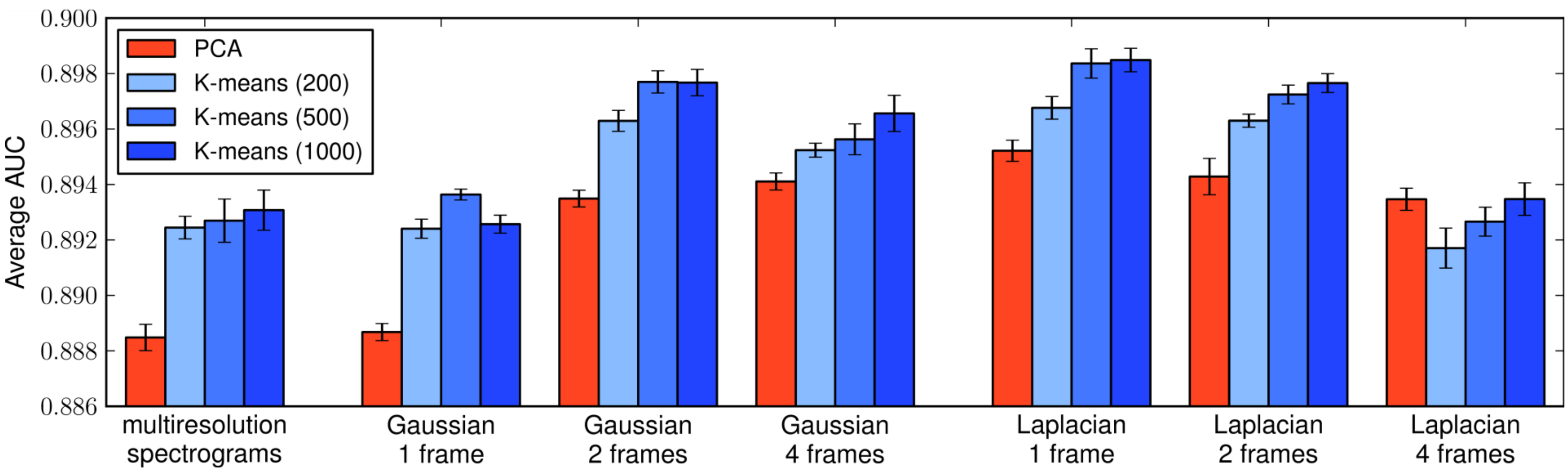
25863 clips of **29** seconds, annotated with **188** tags

Tags are **versatile**: genre, tempo, instrumentation, dynamics, .

We trained a **multilayer perceptron** (MLP):
- 1000 **rectified linear** hidden units
- **cross-entropy** objective
- predict 50 most common tags

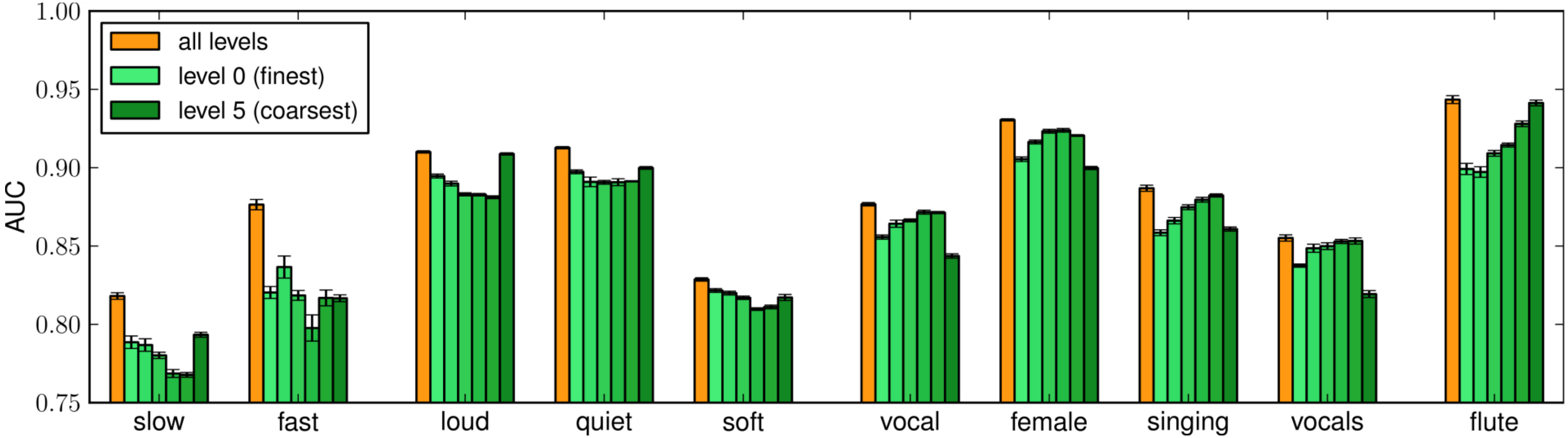
(Law and von Ahn, 2009)

Results: **tag prediction** on the Magnatagatune dataset



UNIVERSITEIT GENT

Results: importance of each timescale for different types of tags



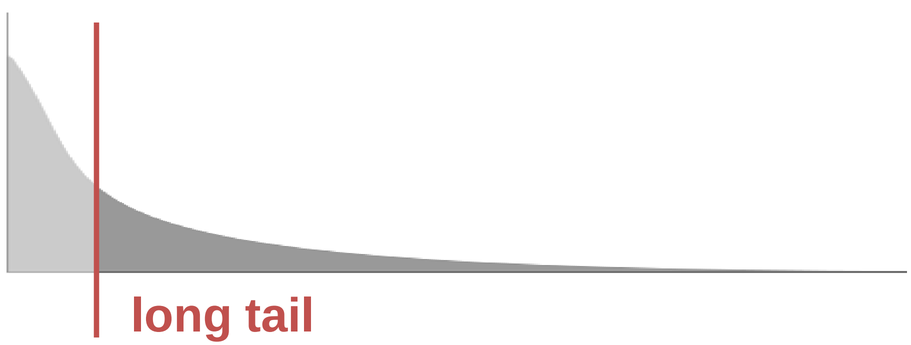
Learning features at **multiple timescales** improves performance over single-timescale approaches

Spherical K-means features consistently improve performance

II. Deep content-based music recommendation

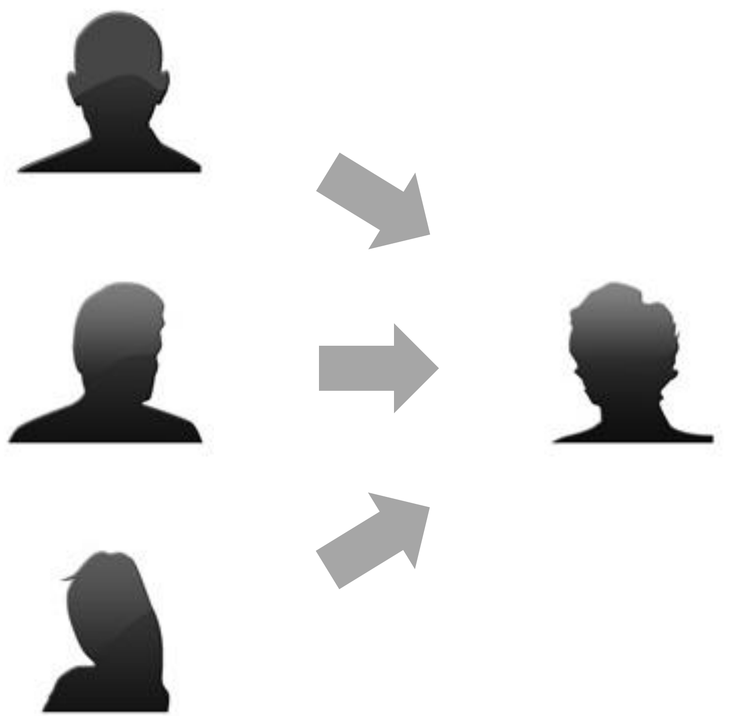
Music recommendation is becoming an increasingly relevant problem

Shift to **digital distribution**



The **long tail** is particularly long for music

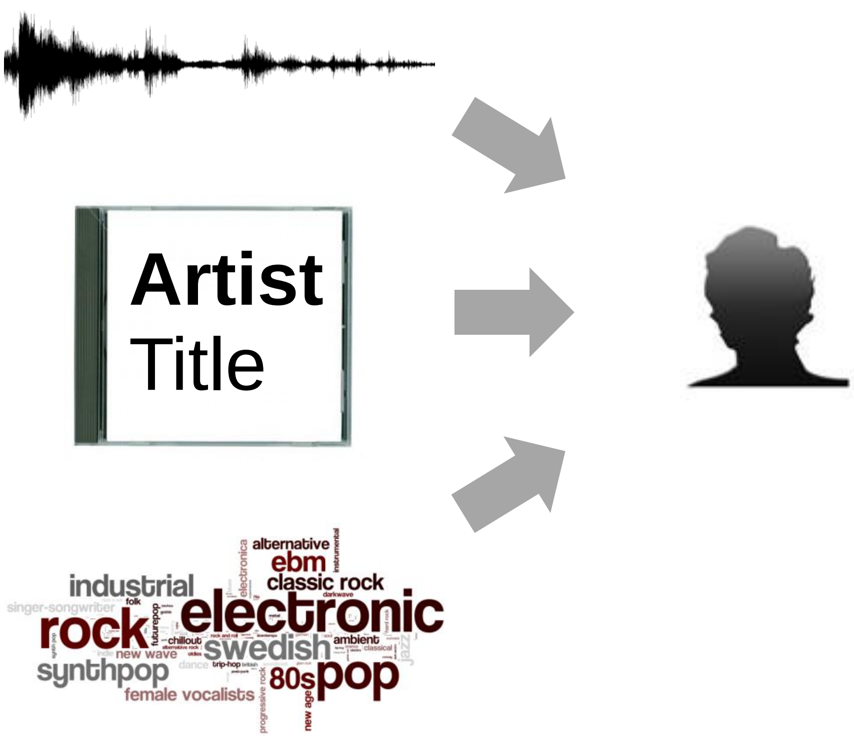
Collaborative filtering: use listening patterns for recommendation



+ good performance
- cold start problem

many **niche items** that only appeal to a small audience

Content-based: use audio content and/or metadata for recommendation



- worse performance

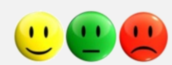
+ no usage data required

allows for all items to be recommended regardless of popularity

There is a large **semantic gap** between audio signals and listener preference

audio signals

genre



mood

popularity

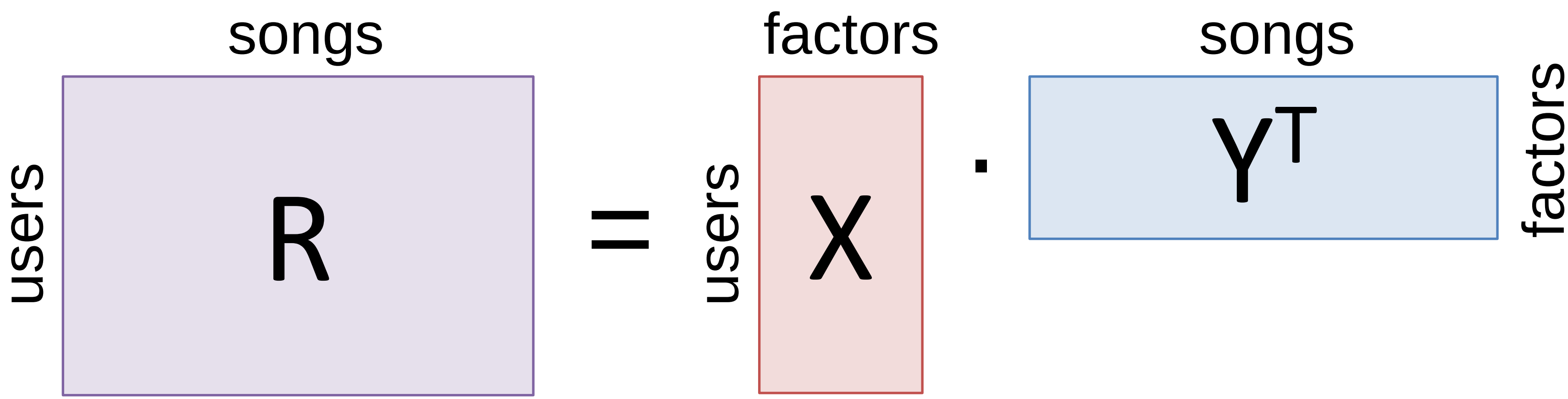
time

lyrical themes

location

instrumentation

Matrix Factorization: model listening data as a product of latent factors



listening data
play counts

user profiles
latent factors

song profiles
latent factors

Weighted Matrix Factorization: latent factor model for implicit feedback data

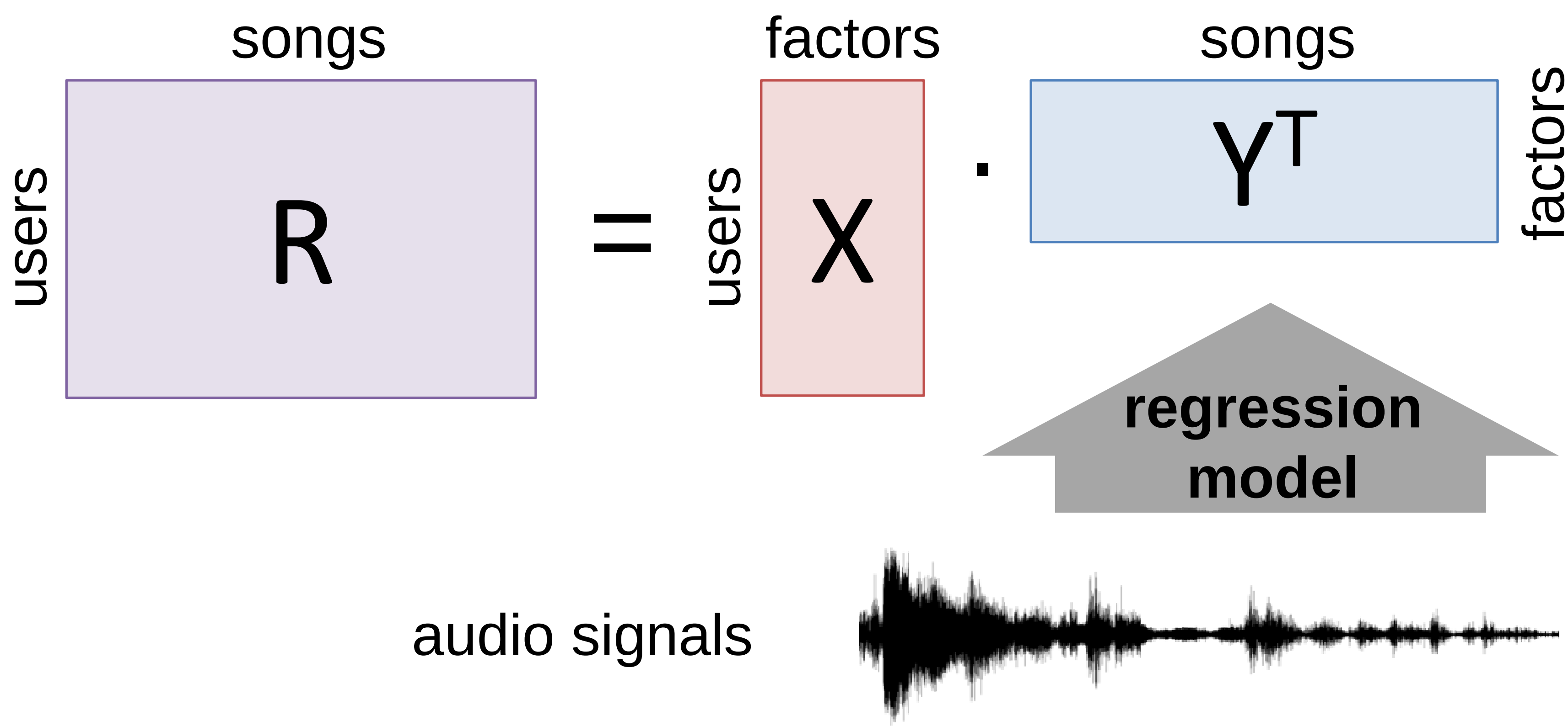
Play count > 0 is a **strong positive signal**
Play count = 0 is a **weak negative signal**

WMF uses a **confidence matrix** to emphasize positive signals

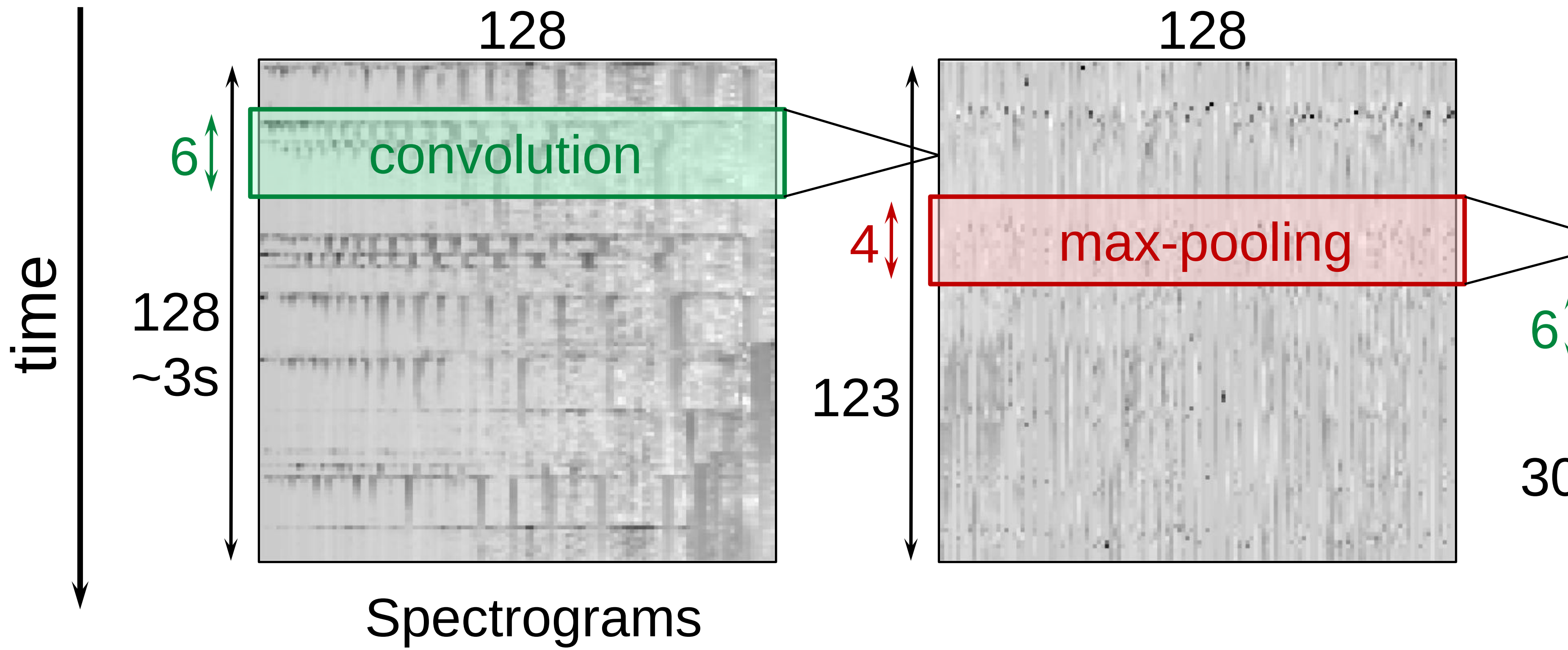
$$\min_{x_*, y_*} \frac{1}{2} \sum_{u,i} c_{ui} \left(p_{ui} - x_u^T y_i \right)^2$$

Hu et al., ICDM 2008

We **predict latent factors** from music audio signals



Deep learning approach: **convolutional neural network**



The **Million Song Dataset** provides metadata for 1,000,000 songs

+ Echo Nest Taste profile subset
Listening data from **1.1m users** for **380k songs**

+ 7digital
Raw audio clips (over 99% of dataset)

Bertin-Mahieux et al., ISMIR 2011

Quantitative evaluation: music recommendation performance

Subset (9330 songs, 20000 users)

Model	mAP@500	AUC
Metric learning to rank	0.01801	0.60608
Linear regression	0.02389	0.63518
Multilayer perceptron	0.02536	0.64611
CNN with MSE	0.05016	0.70987
CNN with WPE	0.04323	0.70101

Quantitative evaluation: music recommendation performance

Full dataset (382,410 songs, 1m users)

Model	mAP@500	AUC
Random	*0.00015*	*0.49935*
Linear regression	0.00101	0.64522
CNN with MSE	0.00672	0.77192
Upper bound	*0.23278*	*0.96070*

Qualitative evaluation: some queries and their closest matches

Query	Most similar tracks (WMF)	Most similar tracks (predicted)
Jonas Brothers Hold On 	**Jonas Brothers** Games	**Jonas Brothers** Video Girl
	Miley Cyrus G.N.O. (Girl's Night Out)	**Jonas Brothers** Games
	Miley Cyrus Girls Just Wanna Have Fun	**New Found Glory** My Friends Over You
	Jonas Brothers Year 3000	**My Chemical Romance** Thank You For The Venom
	Jonas Brothers BB Good	**My Chemical Romance** Teenagers

Qualitative evaluation: some queries and their closest matches

Query	Most similar tracks (WMF)	Most similar tracks (predicted)
Coldplay I Ran Away 	**Coldplay** Careful Where You Stand	**Arcade Fire** Keep The Car Running
	Coldplay The Goldrush	**M83** You Appearing
	Coldplay X & Y	**Angus & Julia Stone** Hollywood
	Coldplay Square One	**Bon Iver** Creature Fear
	Jonas Brothers BB Good	**Coldplay** The Goldrush

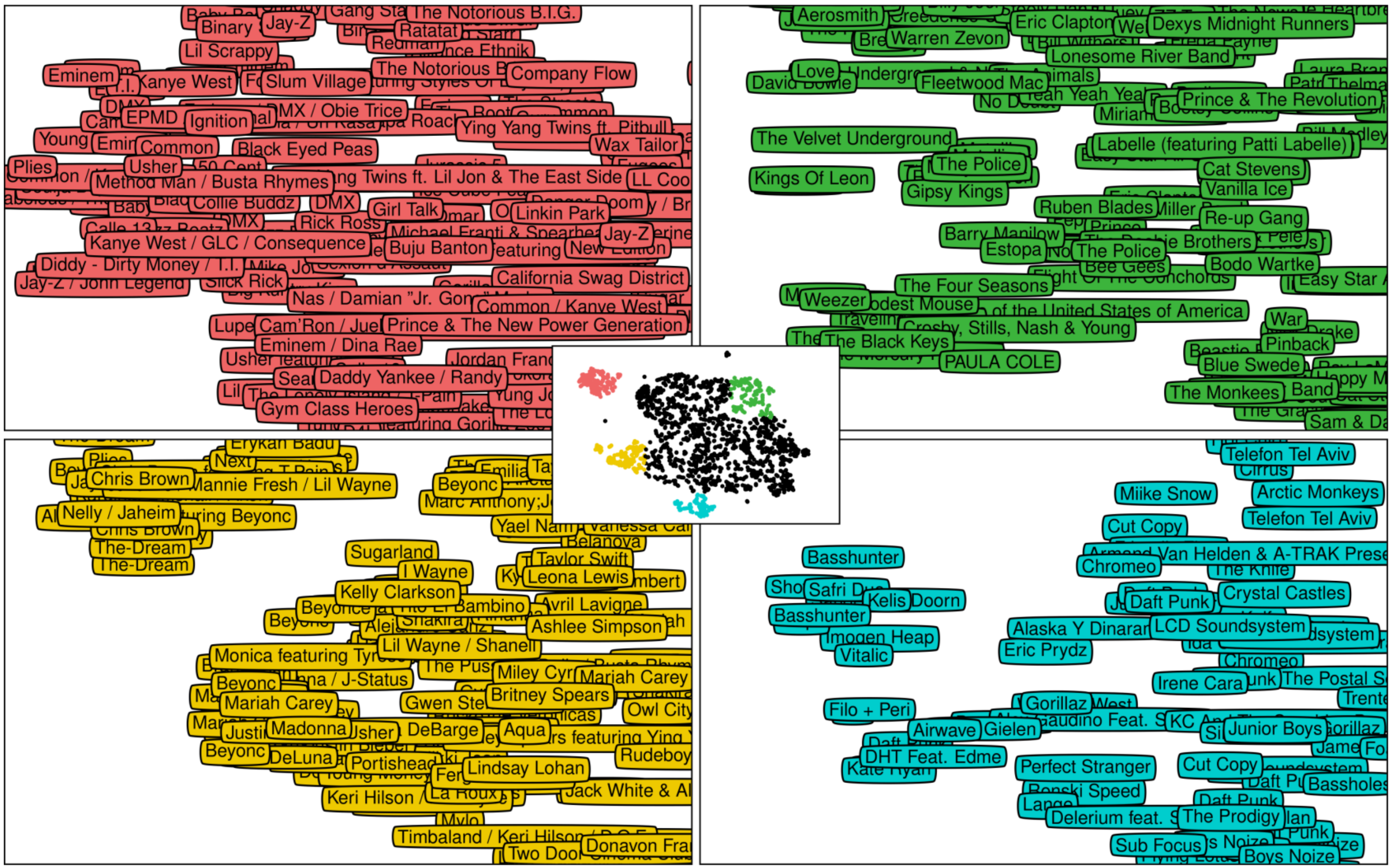
Qualitative evaluation: some queries and their closest matches

Query	Most similar tracks (WMF)	Most similar tracks (predicted)
Beyonce Speechless 	**Beyonce** Gift From Virgo	**Daniel Bedingfield** If You're Not The One
	Beyonce Daddy	**Rihanna** Haunted
	Rihanna / J-Status Crazy Little Thing Called ...	**Alejandro Sanz** Siempre Es De Noche
	Beyonce Dangerously In Love	**Madonna** Miles Away
	Rihanna Haunted	**Lil Wayne / Shanell** American Star

Qualitative evaluation: some queries and their closest matches

Query	**Most similar tracks (WMF)**	**Most similar tracks (predicted)**
Daft Punk Rock'n Roll	**Daft Punk** Short Circuit	**Boys Noize** Shine Shine
	Daft Punk Nightvision	**Boys Noize** Lava Lava
	**Daft Punk** Too Long	**Flying Lotus** Pet Monster Shotglass
	Daft Punk Aerodynamite	**LCD Soundsystem** One Touch
	Daft Punk One More Time	**Justice** One Minute To Midnight

Qualitative evaluation: visualisation of predicted usage patterns (t-SNE)

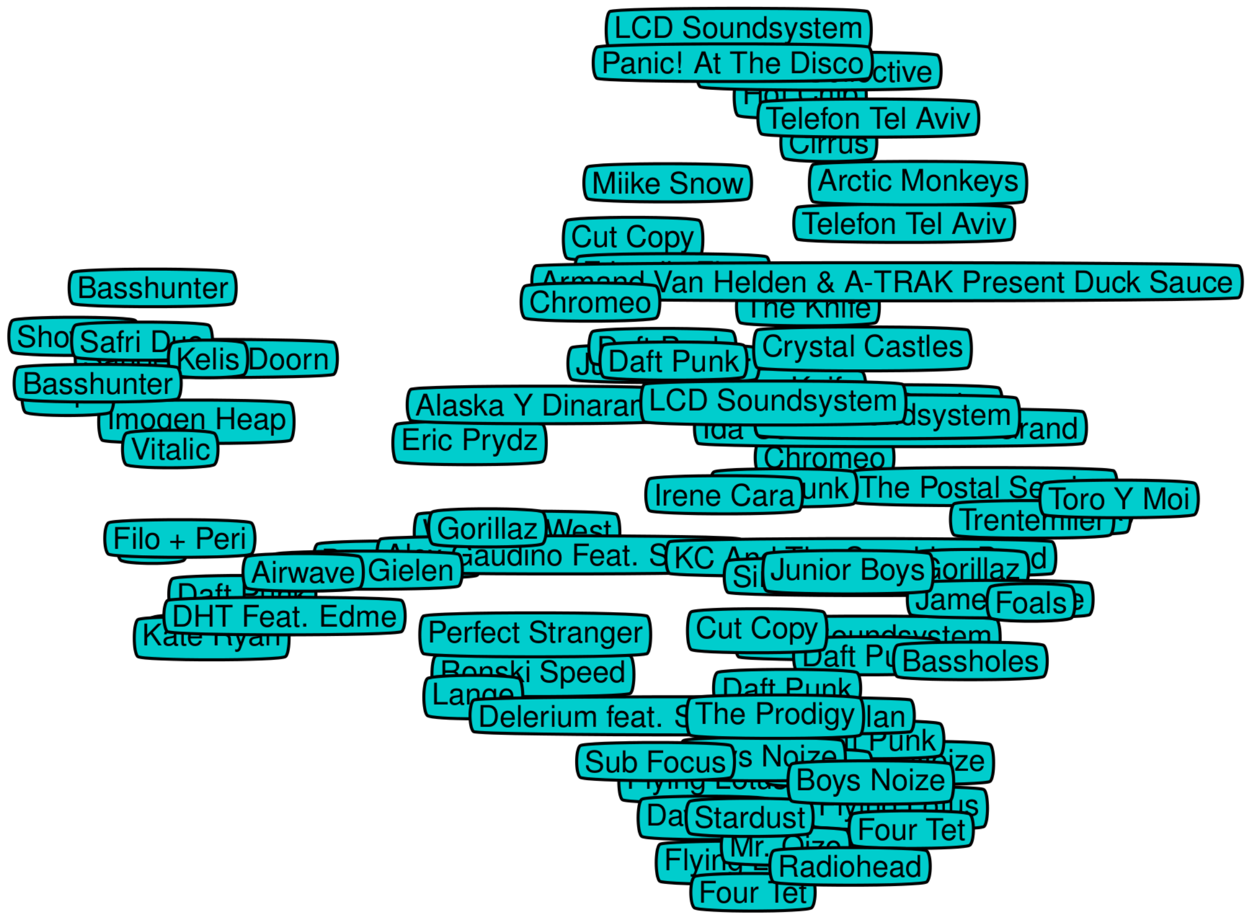


McFee et al., TASLP 2012

Qualitative evaluation: visualisation of predicted usage patterns (t-SNE)

Qualitative evaluation: visualisation of predicted usage patterns (t-SNE)

Qualitative evaluation: visualisation of predicted usage patterns (t-SNE)

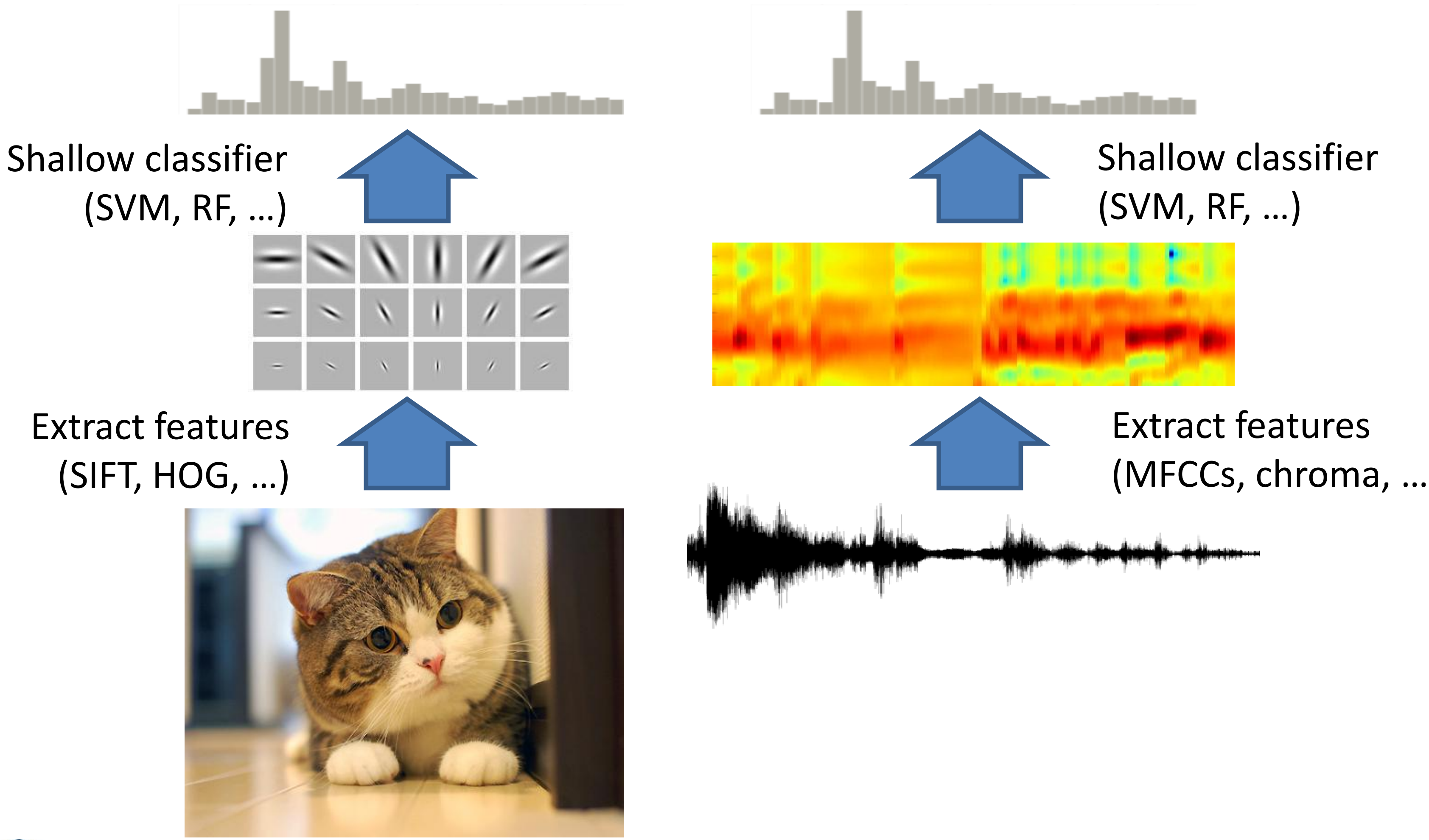


Qualitative evaluation: visualisation of predicted usage patterns (t-SNE)

Predicting latent factors is a viable method for music recommendation

III. End-to-end learning for music audio

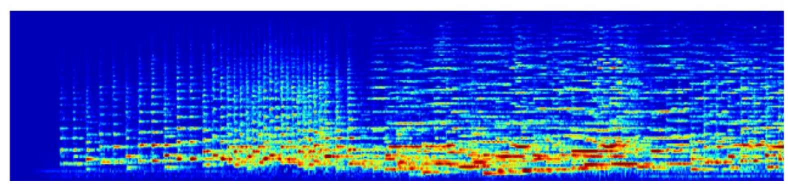
The traditional **two-stage approach**: feature extraction + shallow classifier



Integrated approach: learn both the features and the classifier

Learn features + classifier

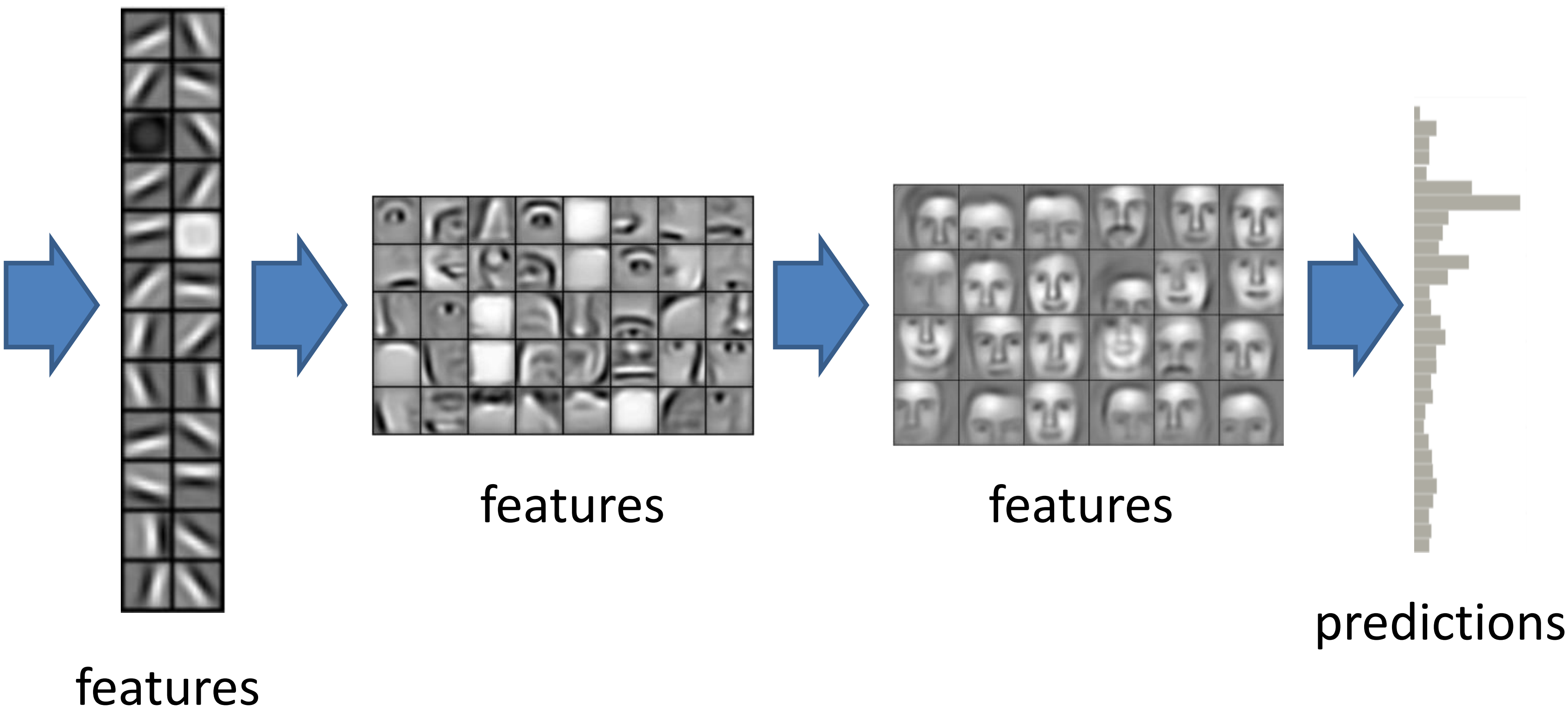
Learn features + classifier



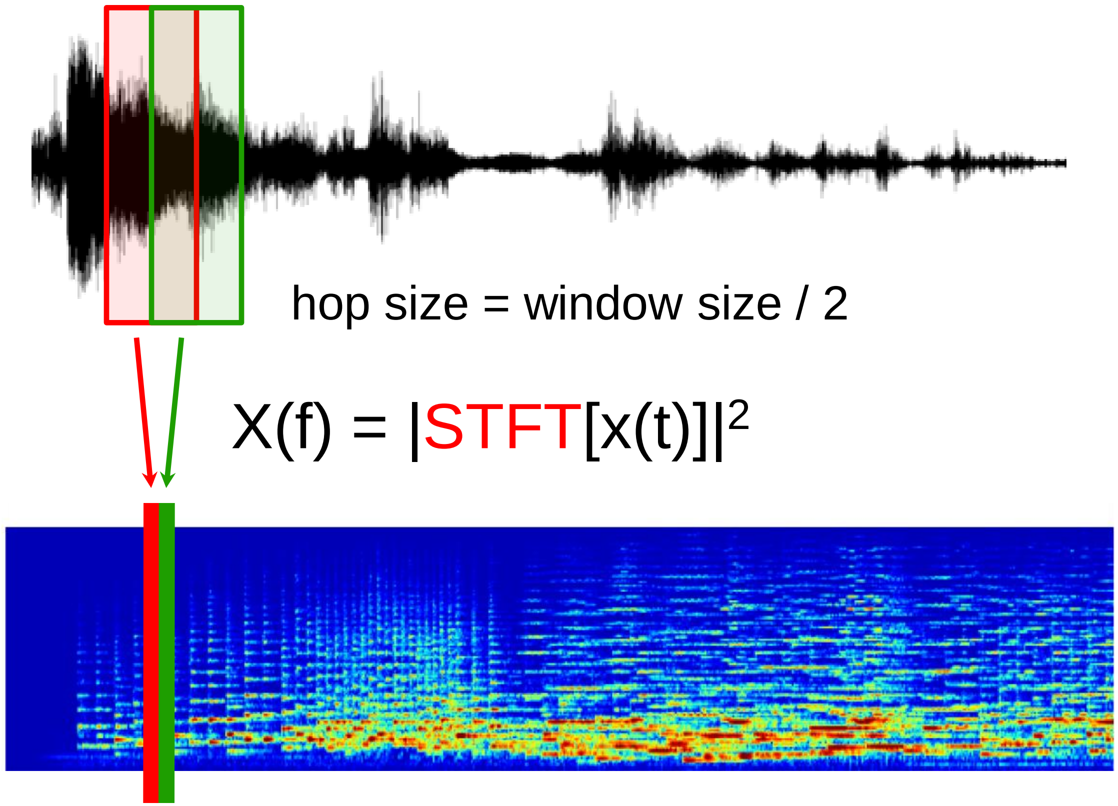
Extract mid-level representation (spectrograms, constant-Q)

End-to-end learning: try to remove the mid-level representation

Convnets can learn the features and the classifier simultaneously



We use log-scaled mel spectrograms as a **mid-level representation**



$$X(f) = |\mathrm{STFT}[x(t)]|^2$$

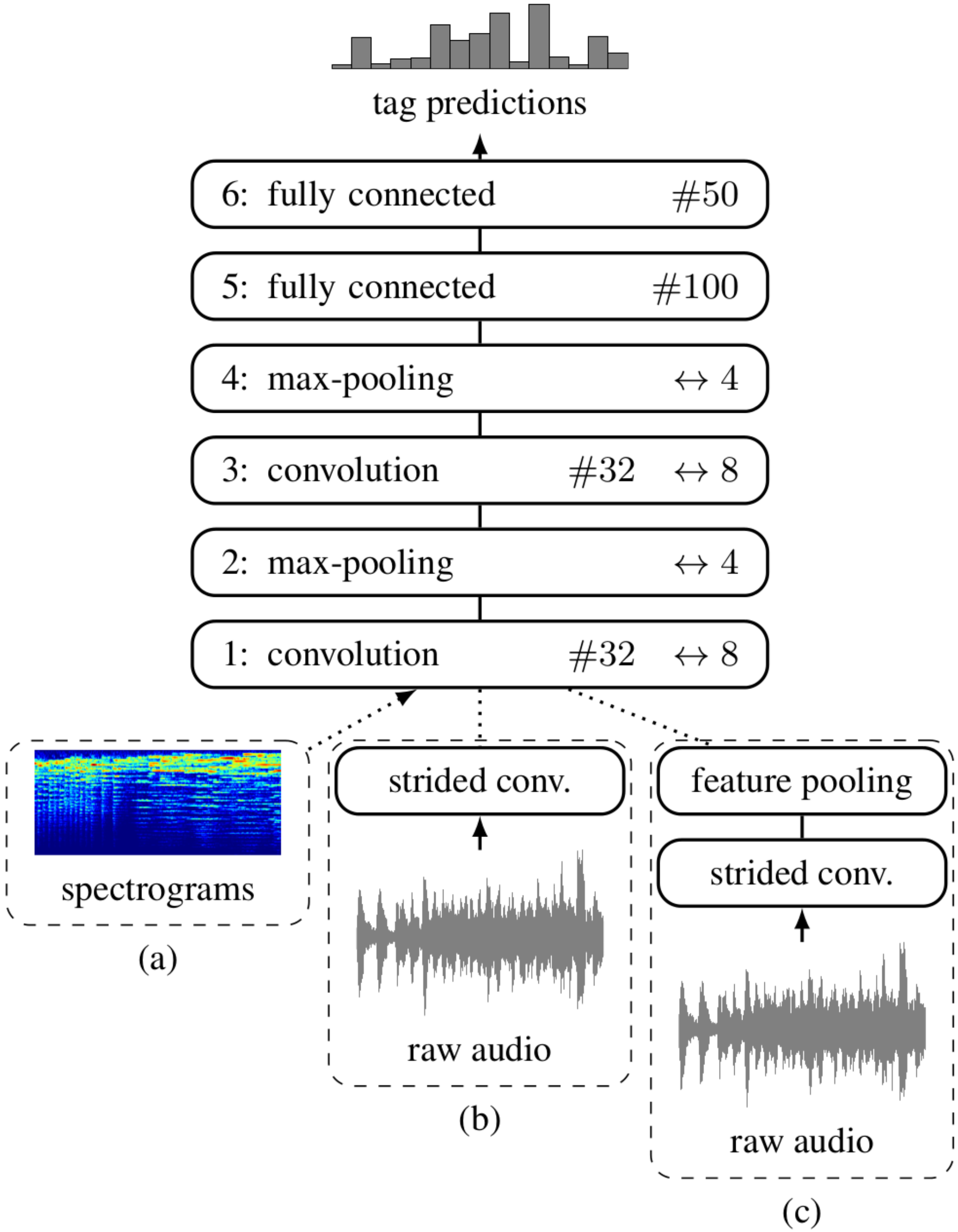
mel binning: $X'(f) = M\,X(f)$

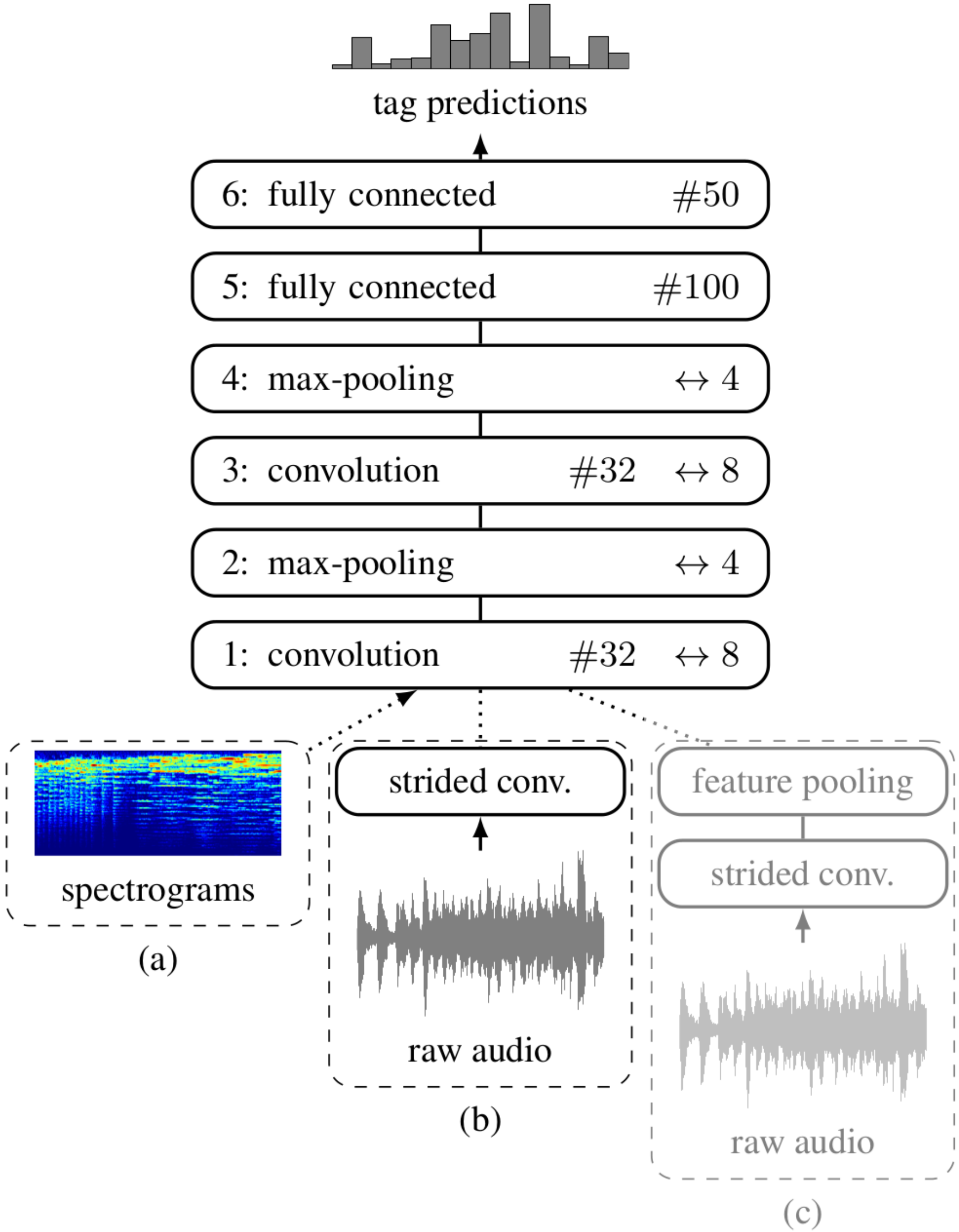
logarithmic loudness (DRC): $X''(f) = \log(1 + C\,X'(f))$

Evaluation: **tag prediction** on Magnatagatune

25863 clips of **29** seconds, annotated with **188** tags

Tags are **versatile**: genre, tempo, instrumentation, dynamics, …


tag predictions
6: fully connected #50
5: fully connected #100
4: max-pooling ↔ 4
3: convolution #32 ↔ 8
2: max-pooling ↔ 4
1: convolution #32 ↔ 8
spectrograms
(a)
strided conv.
raw audio
(b)
feature pooling
strided conv.
raw audio
(c)

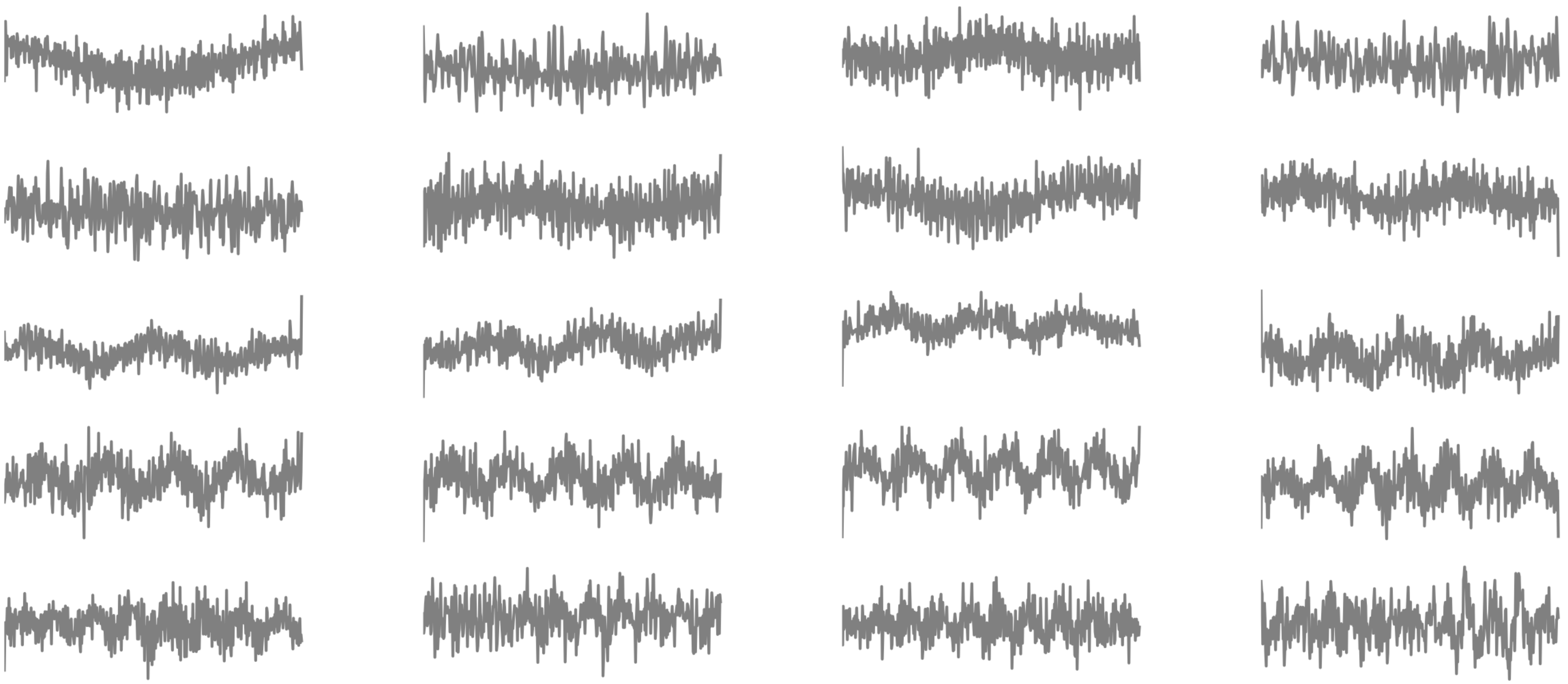


UNIVERSITEIT GENT

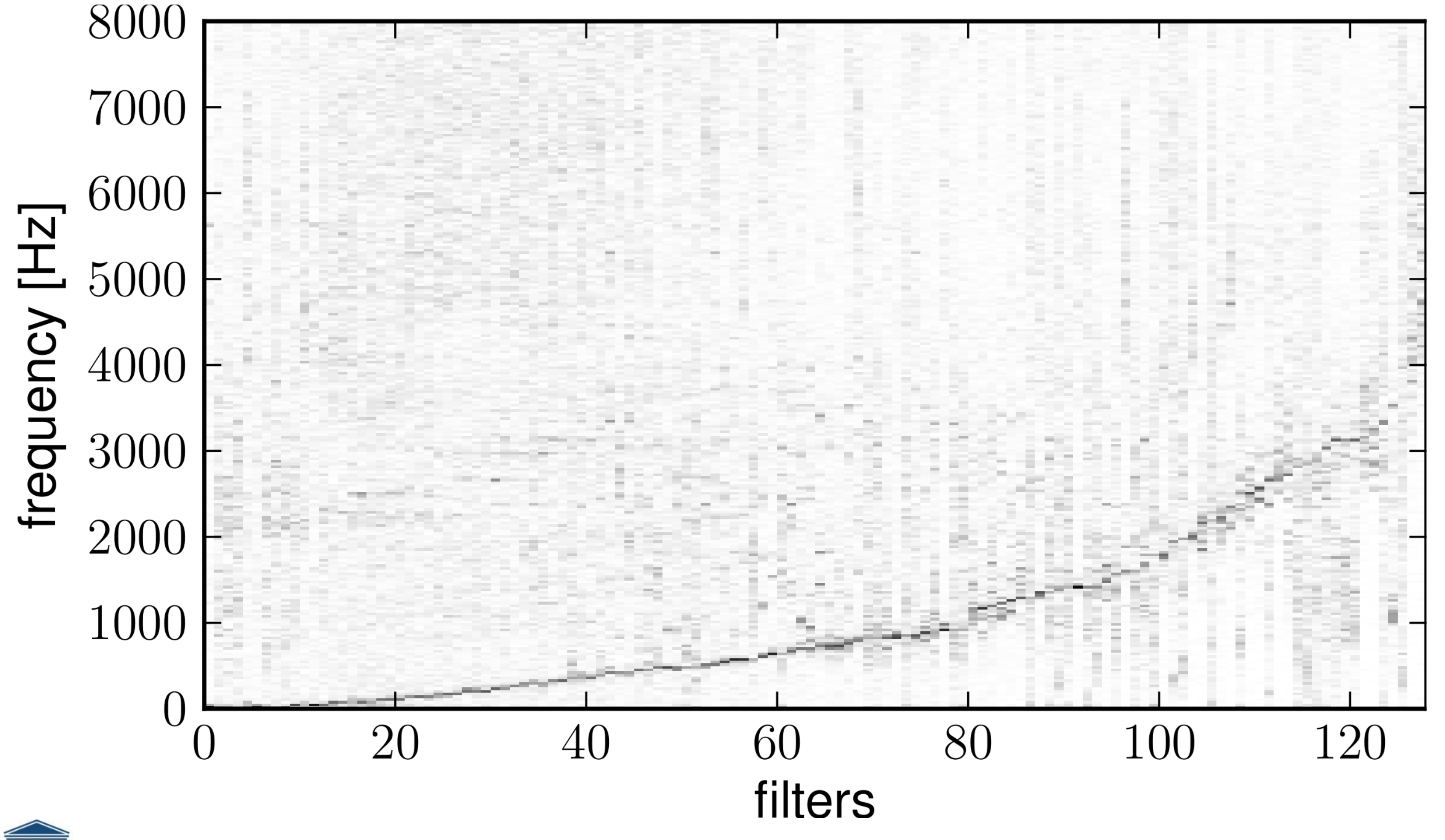
Spectrograms vs. raw audio signals

Length	Stride	AUC (spectrograms)	AUC (raw audio)
1024	**1024**	0.8690	0.8366
1024	**512**	0.8726	0.8365
512	**512**	0.8793	0.8386
512	**256**	0.8793	0.8408
256	**256**	0.8815	0.8487

The learned filters are mostly **frequency-selective** (and noisy)

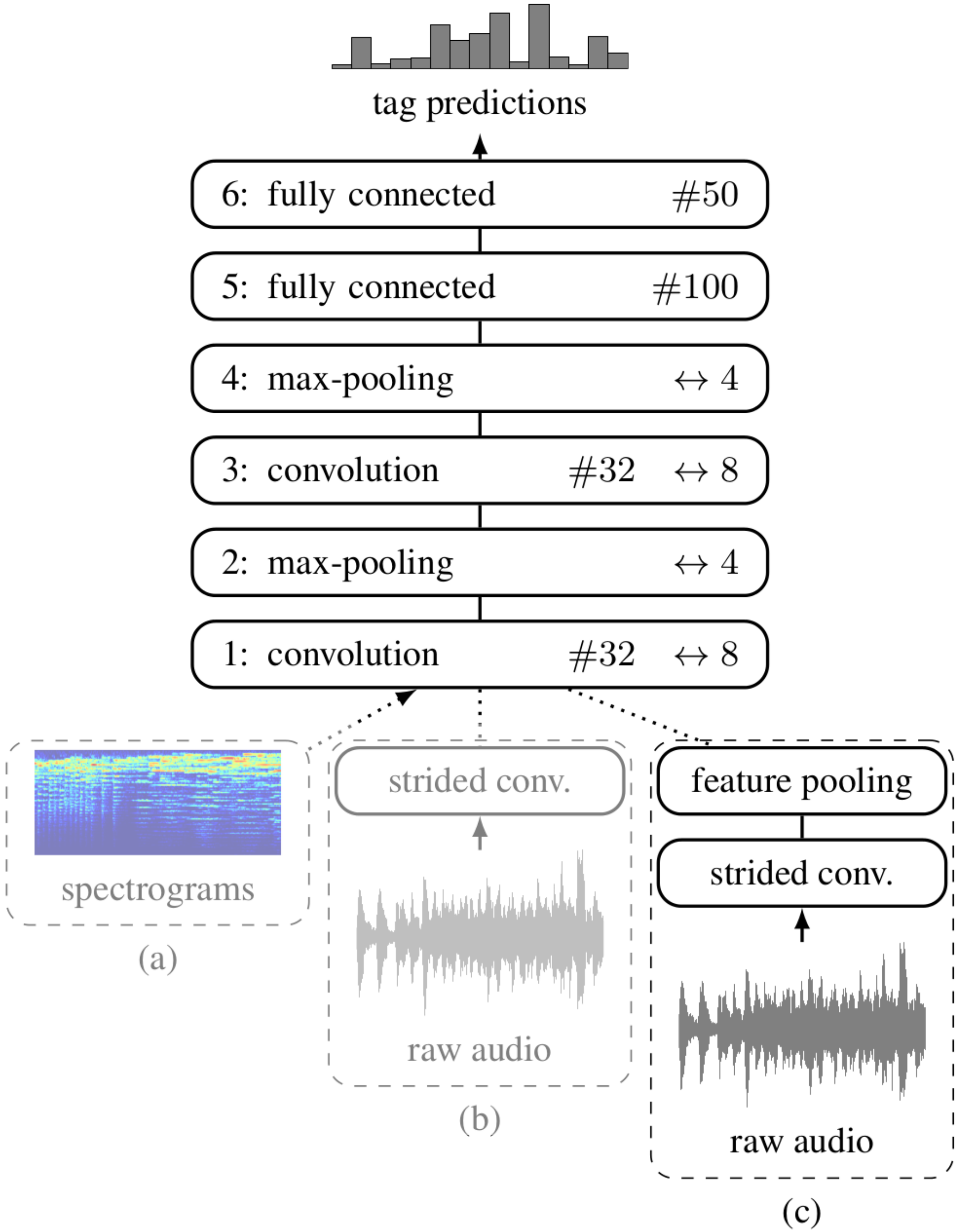


Their dominant frequencies resemble the **mel scale**



Changing the nonlinearity to introduce compression does not help

Nonlinearity	AUC (raw audio)
Rectified linear, **max(0, x)**	0.8366
Logarithmic, **log(1 + C x^2)**	0.7508
Logarithmic, **log(1 + C |x|)**	0.7487

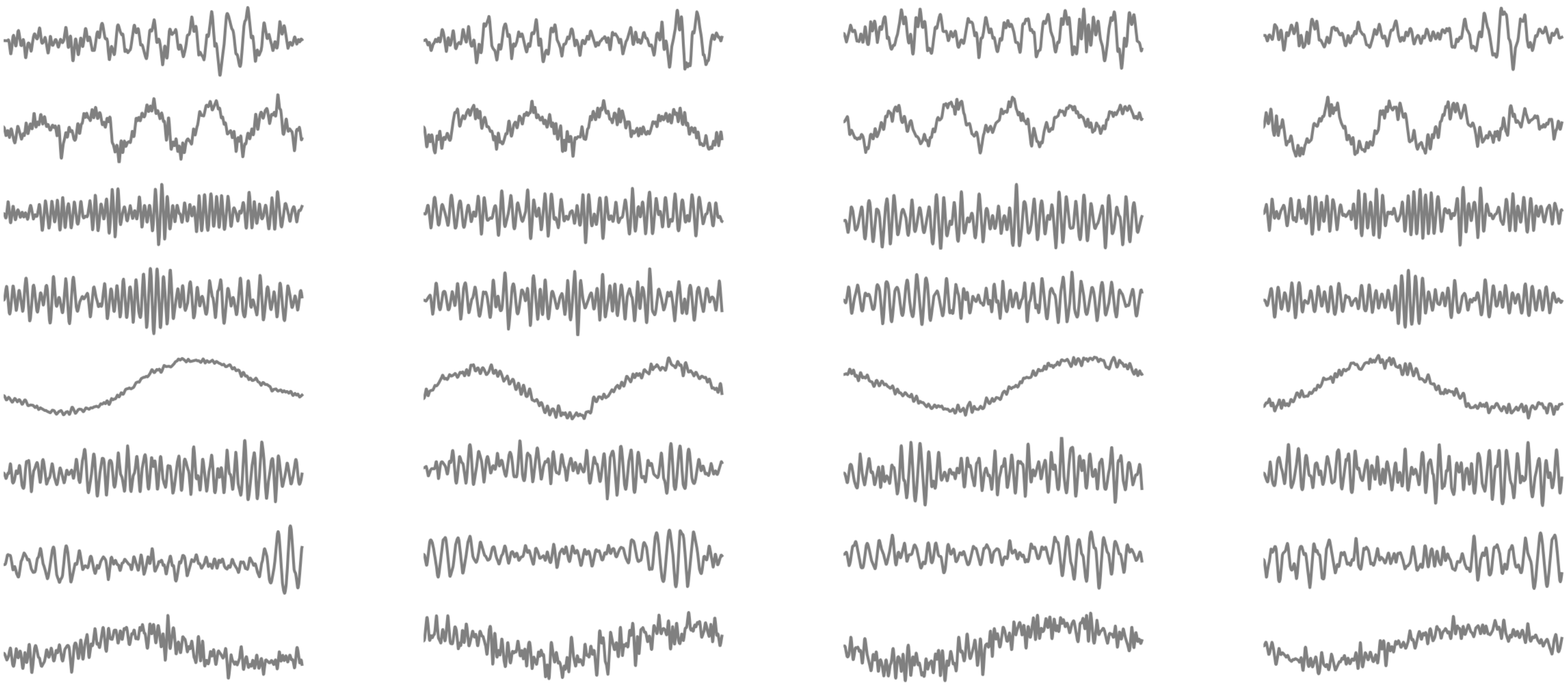


UNIVERSITEIT GENT

Adding a **feature pooling** layer lets the network learn invariances

Pooling method	Pool size	AUC (raw audio)
No pooling	**1**	0.8366
L2 pooling	**2**	0.8387
L2 pooling	**4**	0.8387
Max pooling	**2**	0.8183
Max pooling	**4**	0.8280

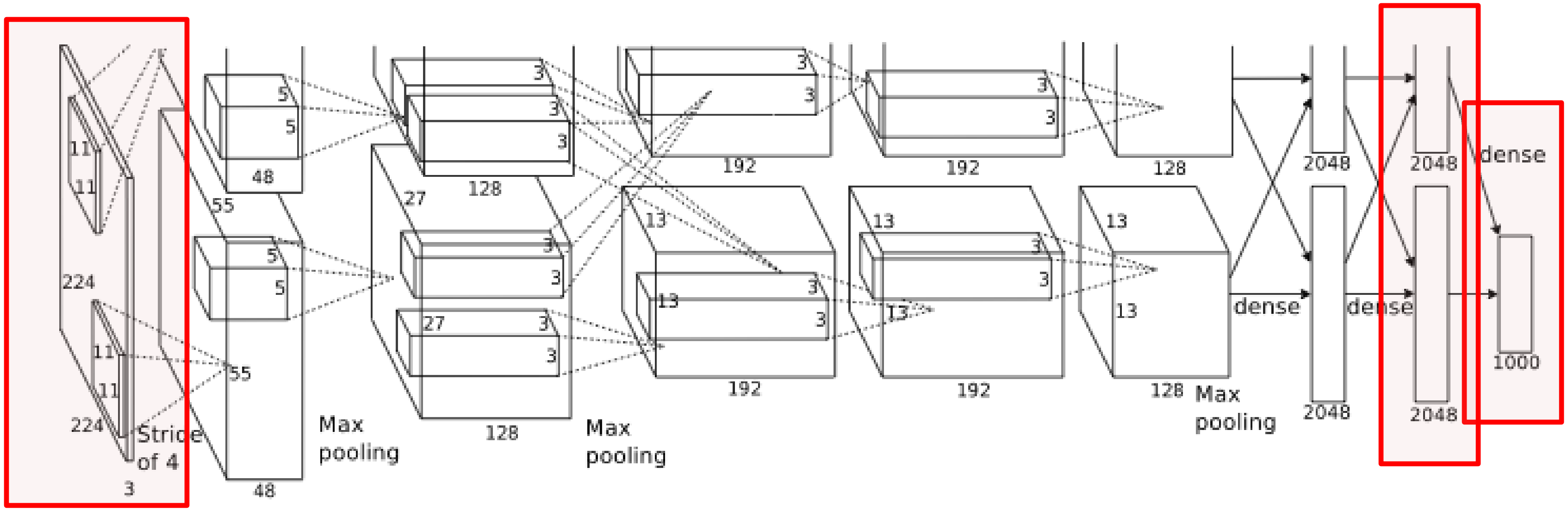
The pools consist of filters that are shifted versions of each other



Learning features from raw audio is possible, but this doesn’t work as well as using spectrograms (yet).

IV. Transfer learning by supervised pre-training

Supervised feature learning



features!

input

output

…
dog
cat
rabbit
penguin
car
table
…

Supervised feature learning for MIR tasks

lots of training data for:
- automatic tagging
- user listening preference prediction (i.e. recommendation)

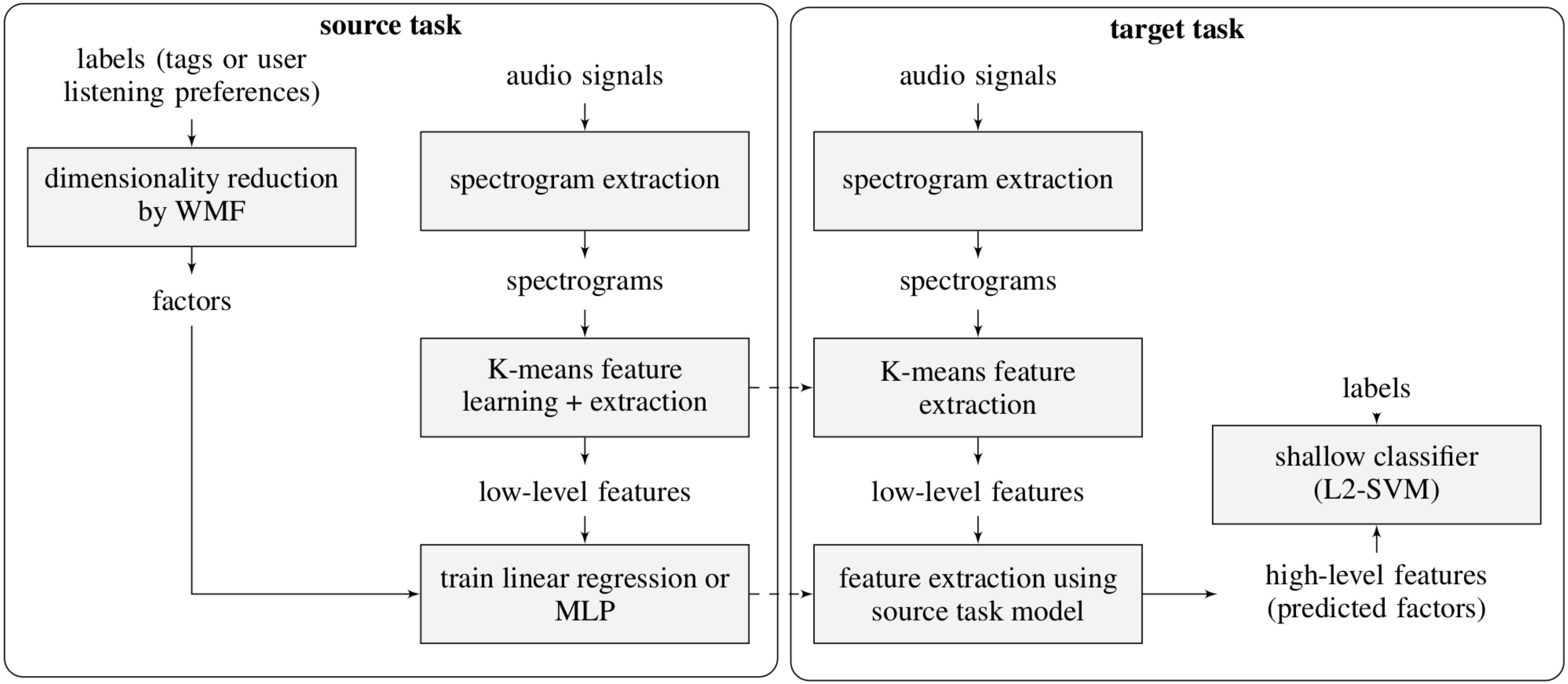
GTZAN	genre classification	**10** genres
Unique	genre classification	**14** genres
1517-artists	genre classification	**19** genres
Magnatagatune	automatic tagging	**188** tags

Tag and listening prediction differ from typical classification tasks

- multi-label classification
- large number of classes (tags, users)
- weak labeling
- redundancy
- sparsity

➡ use WMF for label space dimensionality reduction

Schematic overview



Source task results

User listening preference prediction

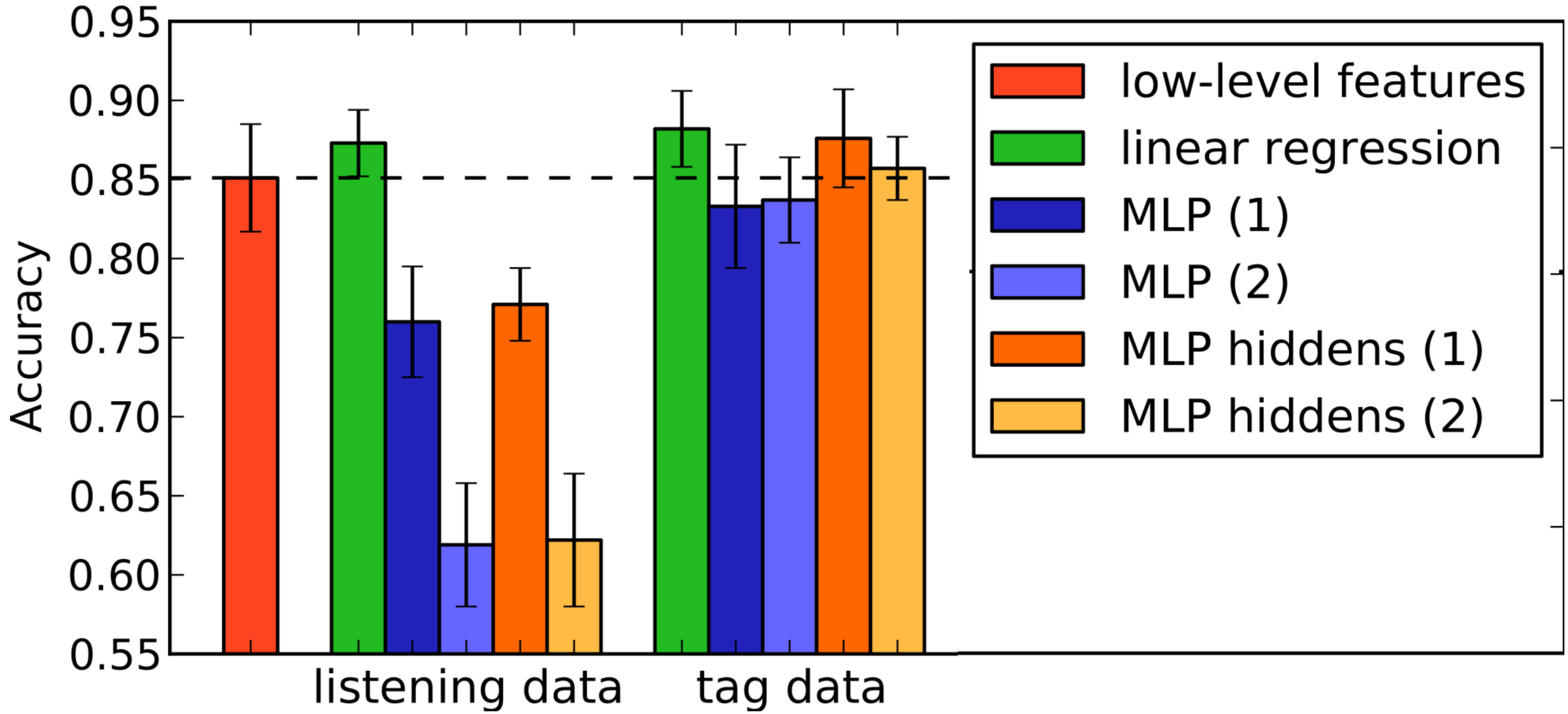
Model	NMSE	AUC	mAP
Linear regression	0.986	0.75	0.0076
MLP (1 hidden layer)	0.971	0.76	0.0149
MLP (2 hidden layers)	0.961	0.746	0.0186

Source task results

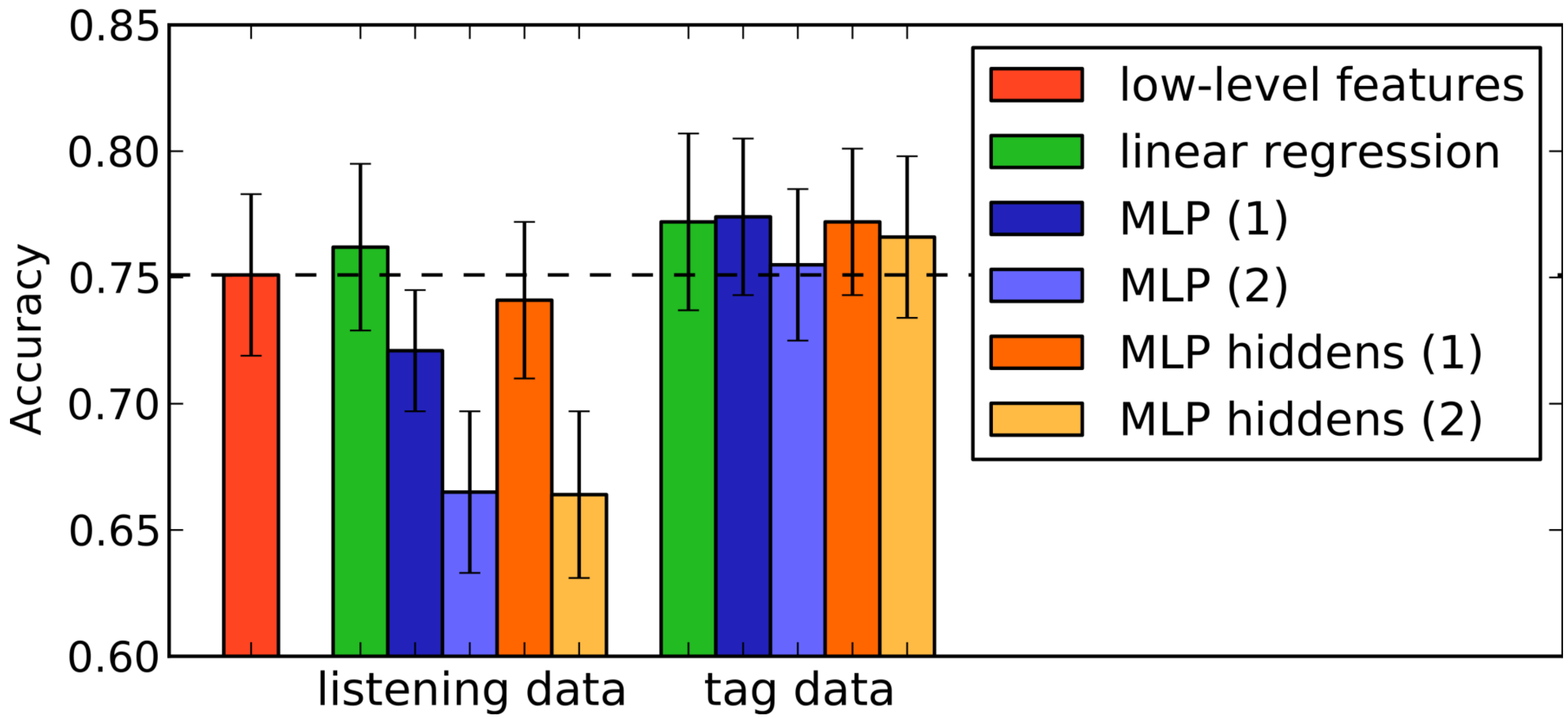
Tag prediction

Model	NMSE	AUC	mAP
Linear regression	0.965	0.823	0.0099
MLP (1 hidden layer)	0.939	0.841	0.0179
MLP (2 hidden layers)	0.924	0.837	0.0179

Target task results: GTZAN genre classification

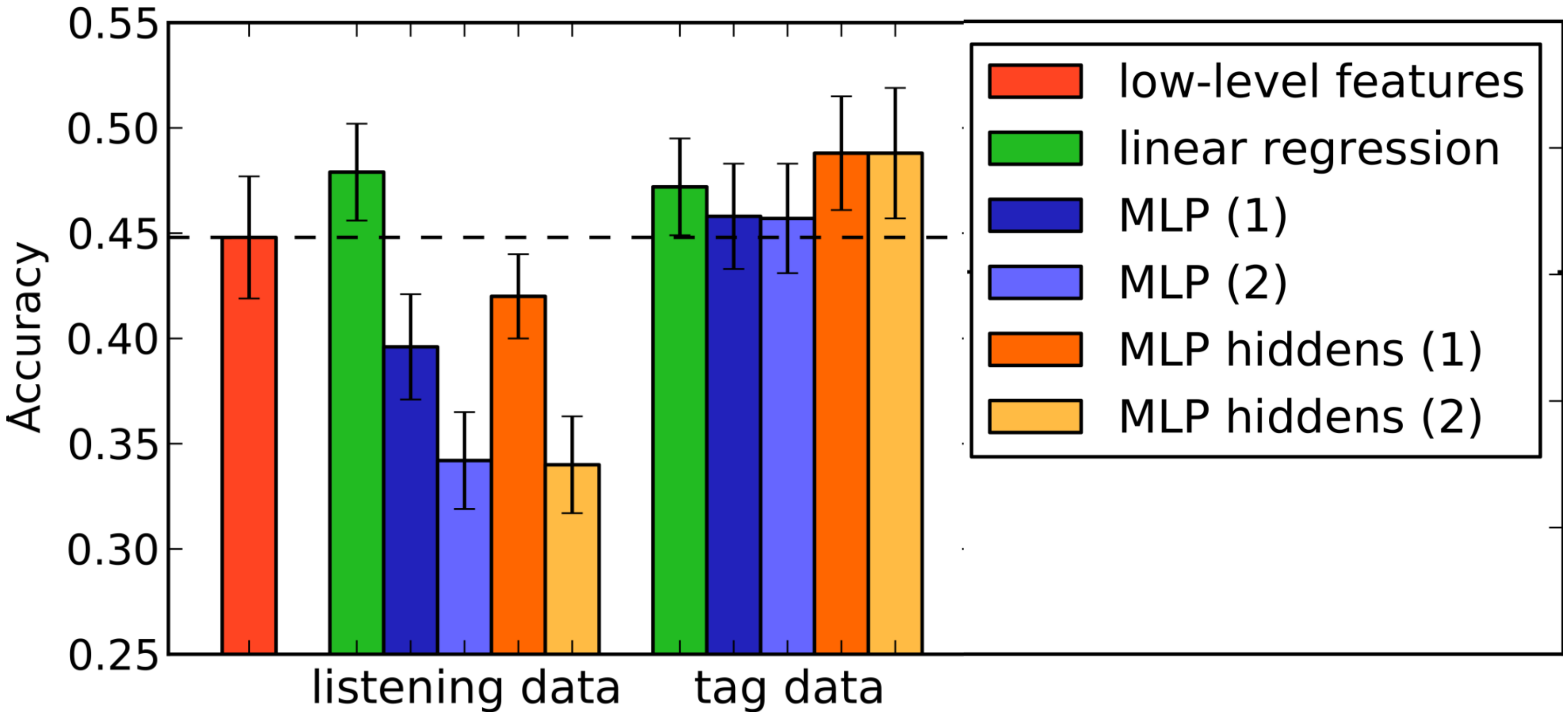


Target task results: Unique genre classification

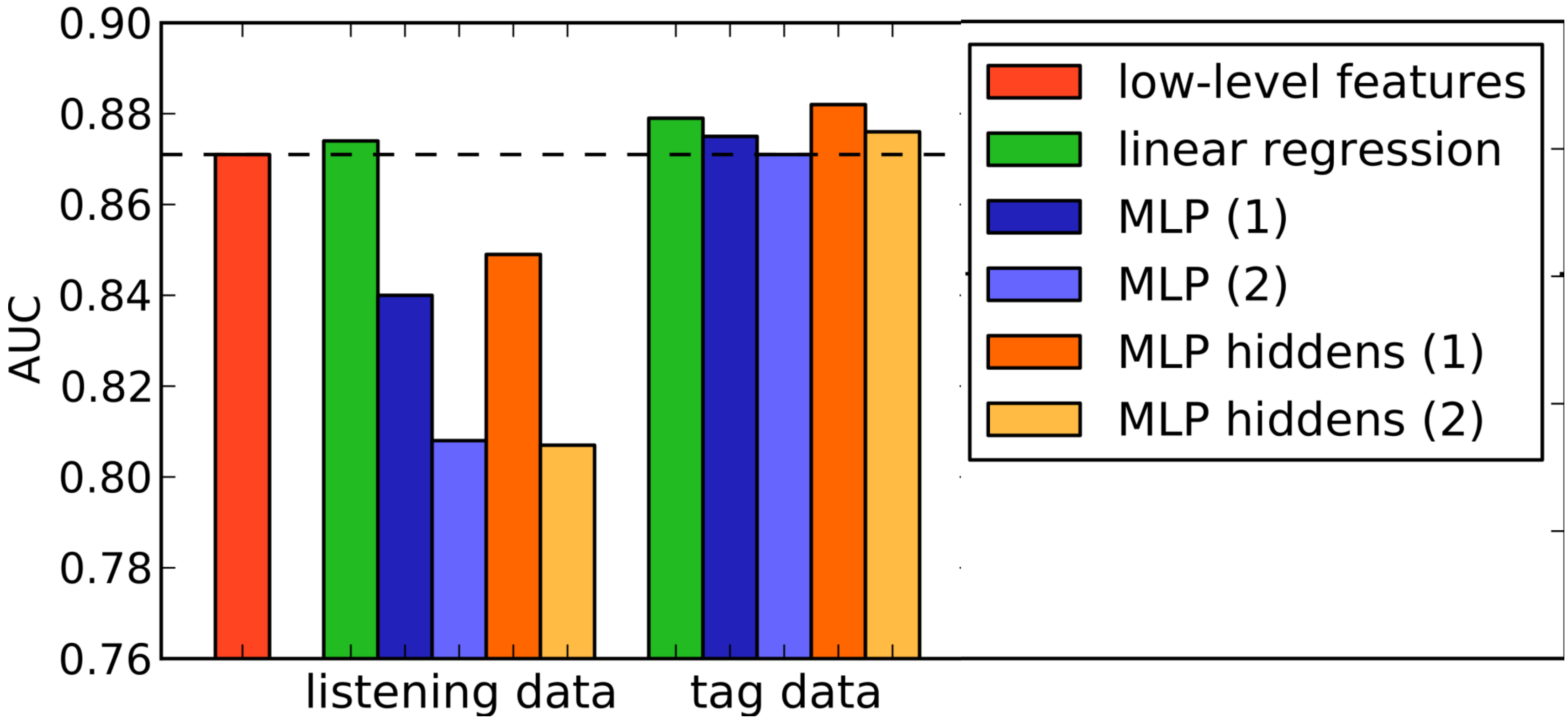


UNIVERSITEIT GENT

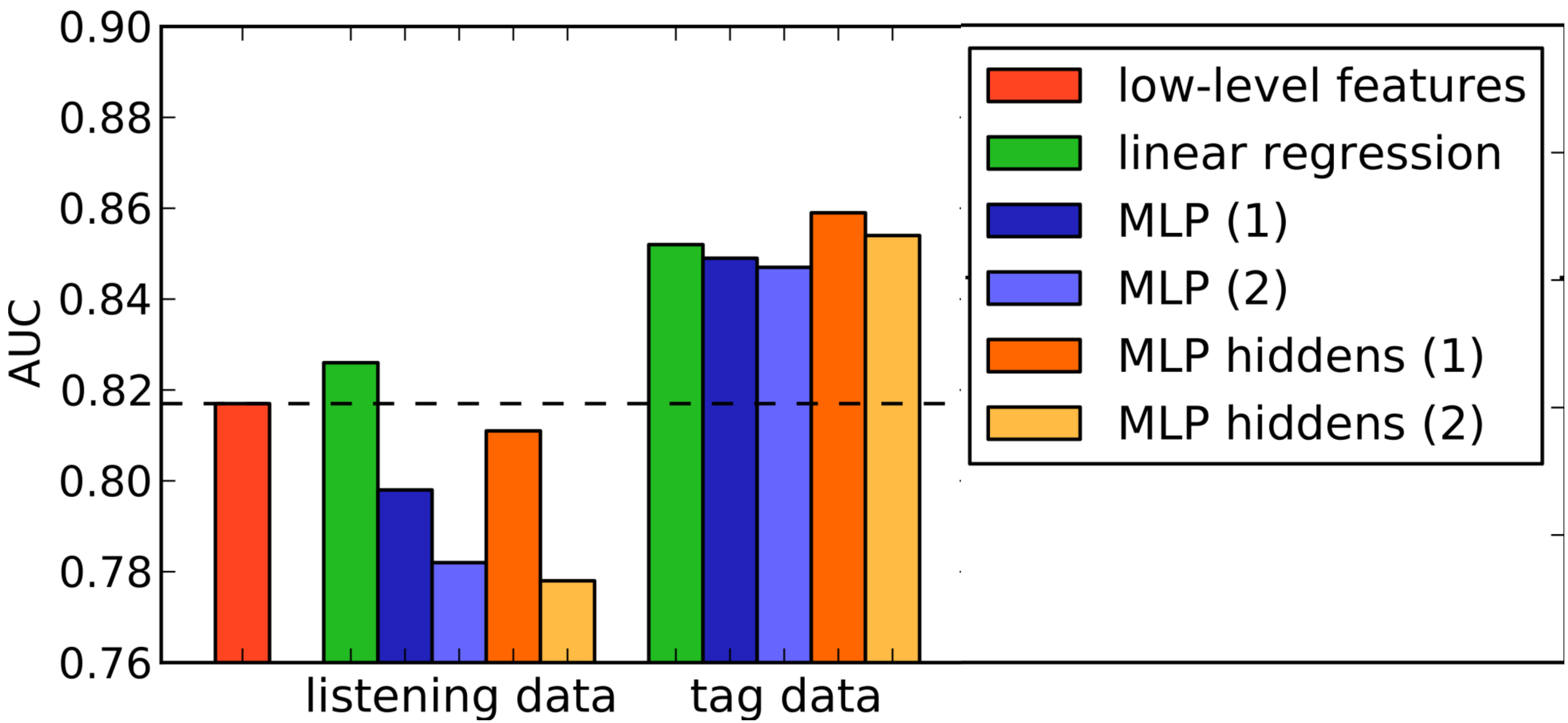
Target task results: 1517-artists genre classification



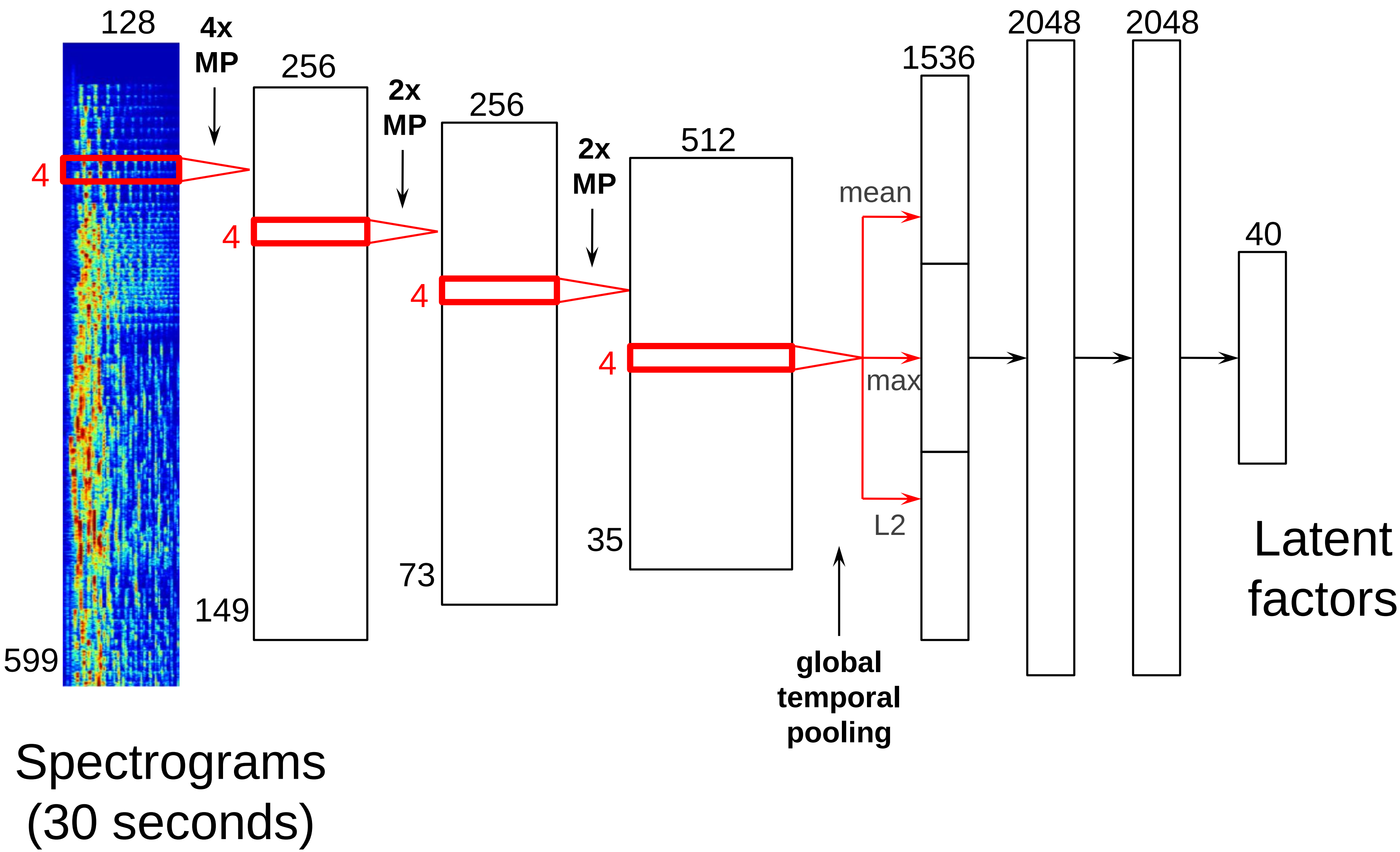
Target task results: Magnatagatune auto-tagging (50)

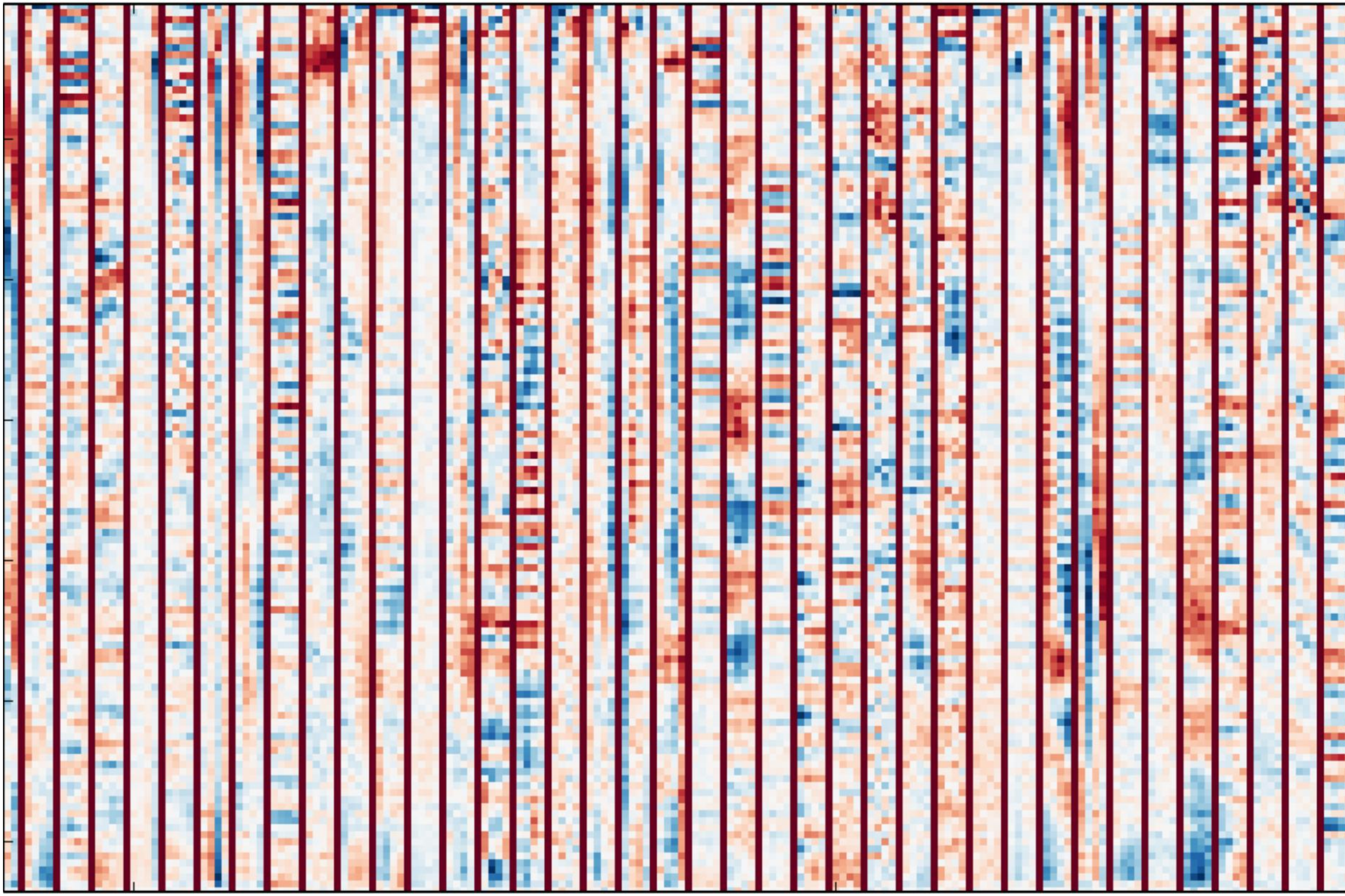


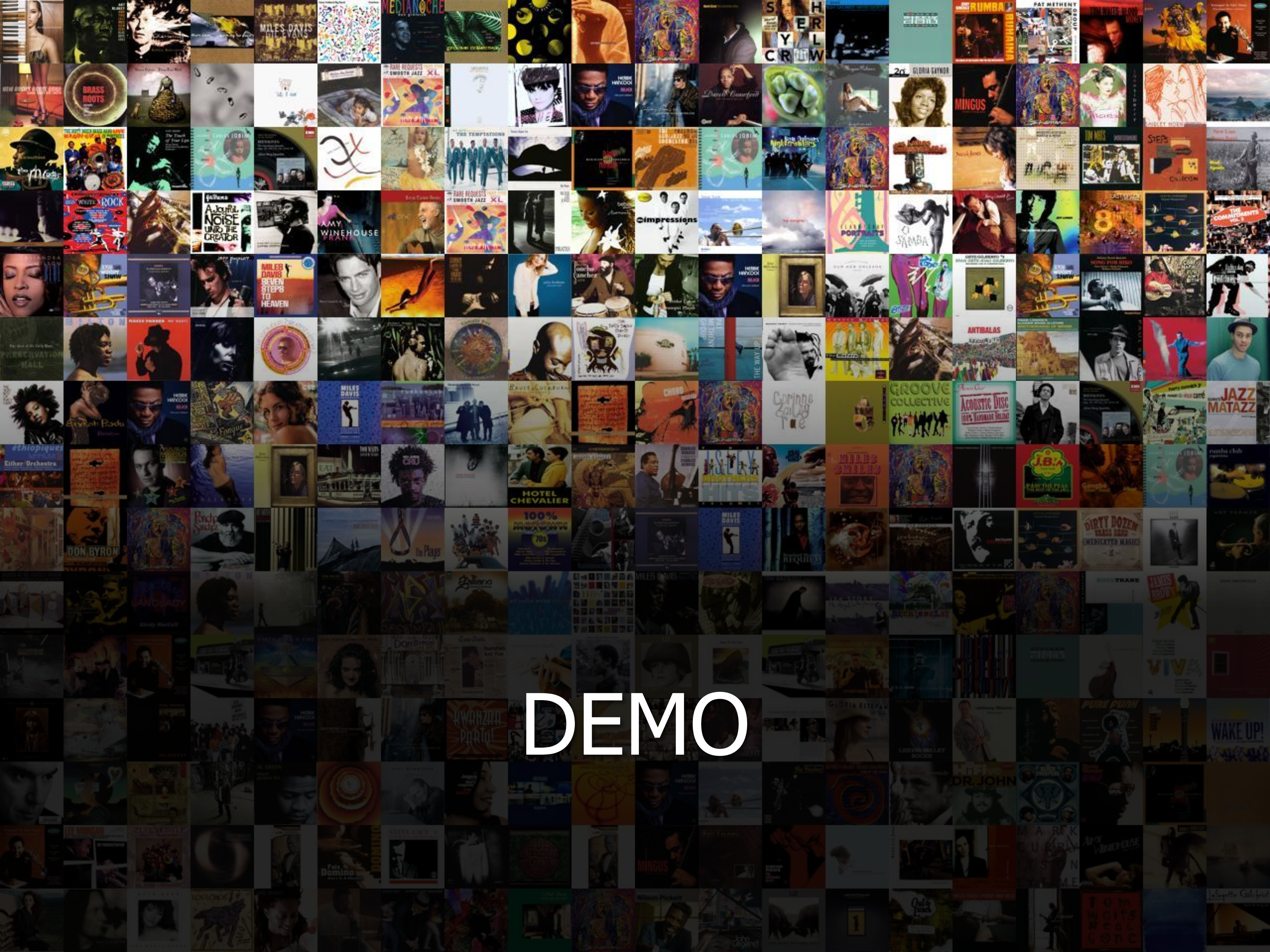
Target task results: Magnatagatune auto-tagging (188)



V. More music recommendation


128
4x
MP
256
2x
MP
256
2x
MP
512
1536
2048
2048
40
4
4
4
4
mean
max
L2
599
149
73
35
global
temporal
pooling
Latent
factors
Spectrograms
(30 seconds)




DEMO

Papers

Multiscale approaches to music audio feature learning

Sander Dieleman, Benjamin Schrauwen, ISMIR 2013

Deep content-based music recommendation

Aäron van den Oord, Sander Dieleman, Benjamin Schrauwen, NIPS 2013

End-to-end learning for music audio

Sander Dieleman, Benjamin Schrauwen, ICASSP 2014

Transfer learning by supervised pre-training for audio-based music classification

Aäron van den Oord, Sander Dieleman, Benjamin Schrauwen, ISMIR 2014