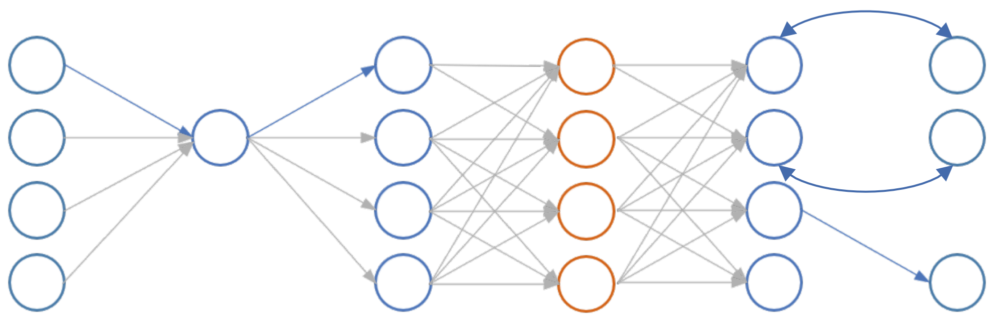
AUDIO LABS

Literature Overview

Deep Neural Networks in MIR



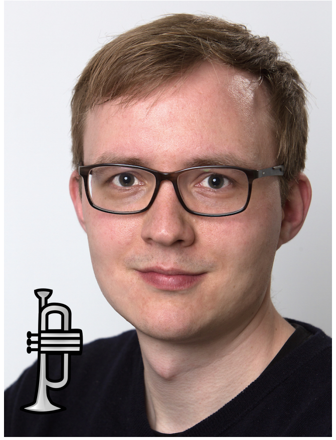
Stefan Balke and Meinard Müller

International Audio Laboratories Erlangen

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Fraunhofer IIS

Introduction
Stefan Balke



- 2008-2013: Electrical Engineering
  Leibniz Universität Hannover
- Since 2014: Working towards my PhD

- Research Interests:
  - Content-based audio retrieval
  - Deep learning and MIR
  - Web and multimedia
  - Jazz music

- Hobby: Trumpet playing!
- Further infos: https://www.audiolabs-erlangen.de/fau/assistant/balke

Motivation

- DNNs become a general method (almost easy to use).
- Lots of decisions involved in designing a DNN
  - Input representation, input preprocessing
  - #layers, #neurons, layer type, dropout, regularizers, cost function
  - Initialization, mini-batch size, #epochs, early stopping (patience)
  - Optimizer, learning rate…
- Provide a starting point for beginners.

Considered MIR Tasks

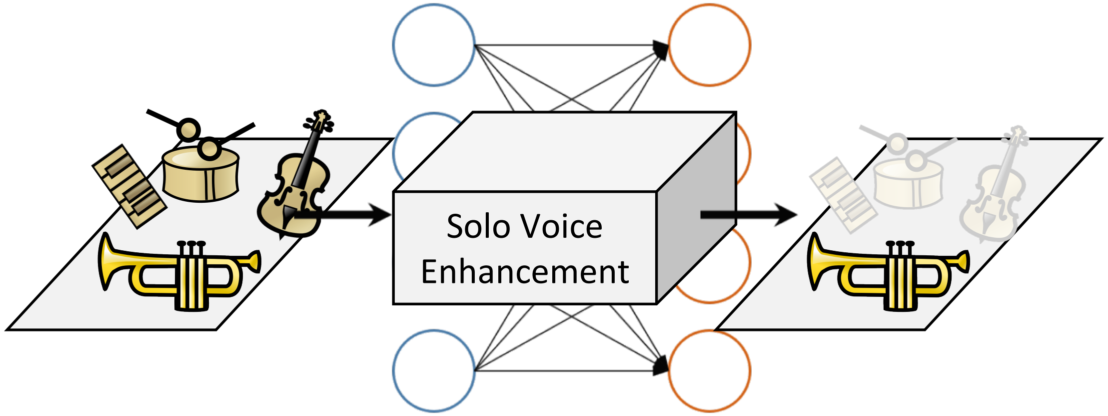
- 7 Categories
  - Feature Learning (FL)
  - F0-Estimation (F0)
  - Automatic Music Transcription (AMT)
  - Beat and Rhythm Analysis (BAR)
  - Music Structure Analysis (MSA)
  - Chord Recognition (CR)
  - Audio Source Separation (ASP)
  - Various (e.g., Singing Voice Detection, Tagging, …) (VAR)

- 76 publications, 149 authors

Overview

Philippe Halsman, "Louis Armstrong"

1. Feature Learning
2. Beat and Rhythm Analysis
3. Music Structure Analysis
4. Literature Overview



Feature Learning

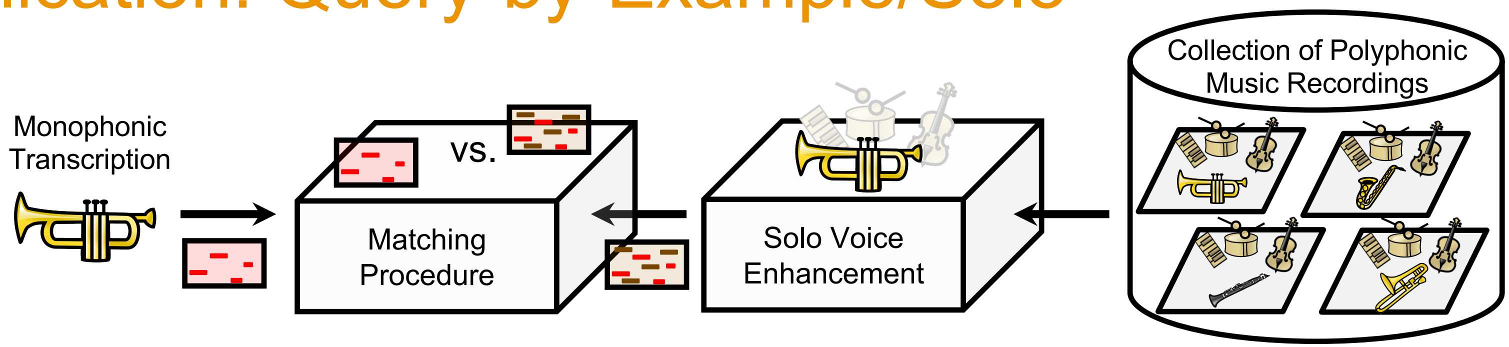
Feature Learning

…where it all began

- Core task for DNNs:
  Learn a representation from the data to solve a problem.
- Task is very hard to define!
  Often evaluated in tagging, chord recognition, or retrieval application.

Task	Year	Authors	Ref.	Type	Input	Pre-proc.
FL	2013	Schmidt and Kim	[67]	DBN	HC	—
FL	2010	Hamel and Eck	[30]	DBN	LinS	—
FL	2017	Dai et al.	[15]	CNN	Raw	—
FL	2012	Hamel et al.	[33]	FNN	LogMelS	PCA
FL	2016	Korzeniowski and Widmer	[43]	FNN	LogLogS	—
FL	2017	Balke et al.	[2]	FNN	LogS	—
FL	2011	Hamel et al.	[32]	FNN	MelS	PCA
FL	2014	Dieleman and Schrauwen	[17]	CNN	Raw	—

Application: Query-by-Example/Solo



Retrieval Scenario

Given a monophonic transcription of a jazz solo as query, find the corresponding document in a collection of polyphonic music recordings.

Solo Voice Enhancement

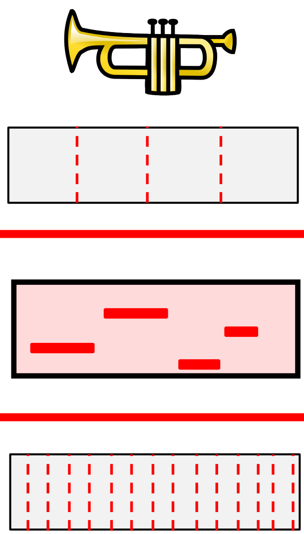
1. Model-based Approach [Salamon13]
2. Data-Driven Approach [Rigaud16, Bittner15]

Our Data-Driven Approach

Use a **DNN** to learn the mapping from a "polyphonic" TF representation to a "monophonic" TF representation.

Weimar Jazz Database (WJD)

[Pfleiderer17]



Transcription

Beats

| E^7 A^7 | D^7 G^7 | …

Chords

…

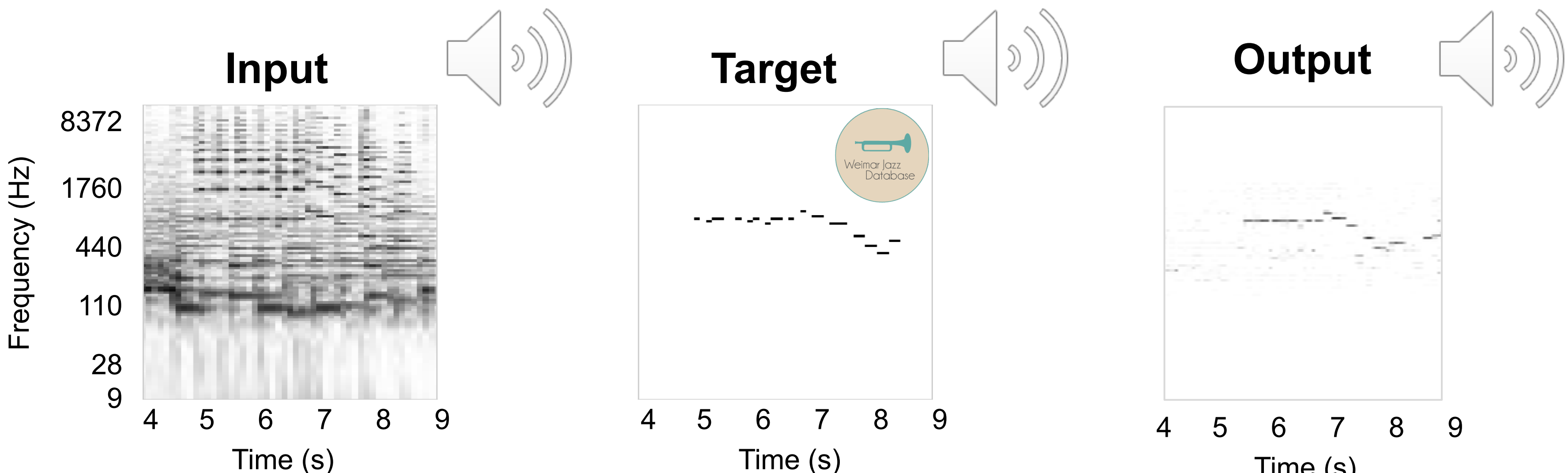
- 456 transcribed jazz solos of monophonic instruments.
- Transcriptions specify a musical pitch for physical time instances.
- 810 min. of audio recordings.

Thanks to the Jazzomat research team: M. Pfleiderer, K. Frieler, J. Abeßer, W.-G. Zaddach

DNN Training

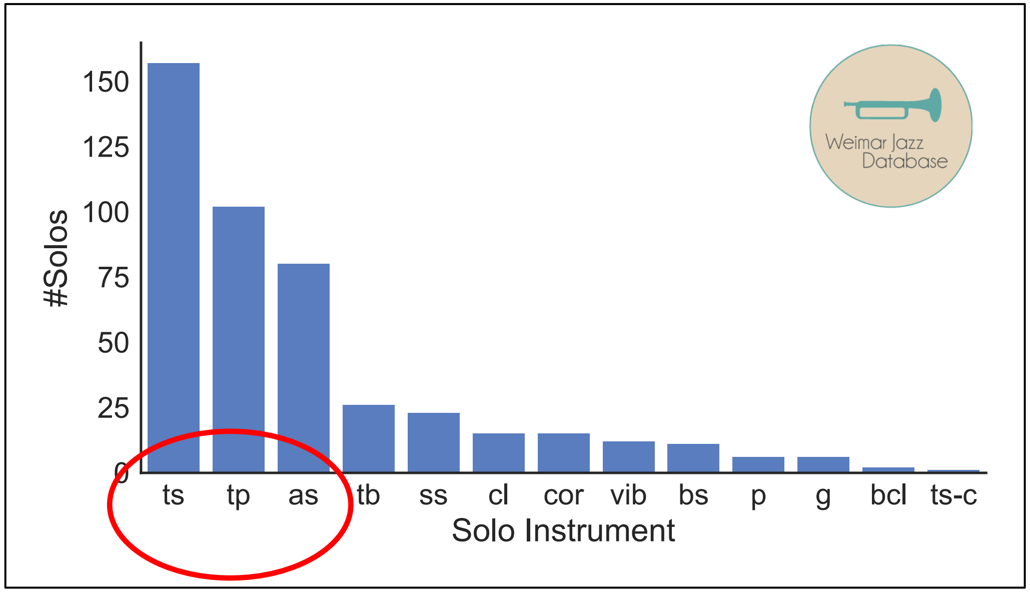
Stefan Balke, Christian Dittmar, Jakob Abeßer, Meinard Müller, ICASSP 17

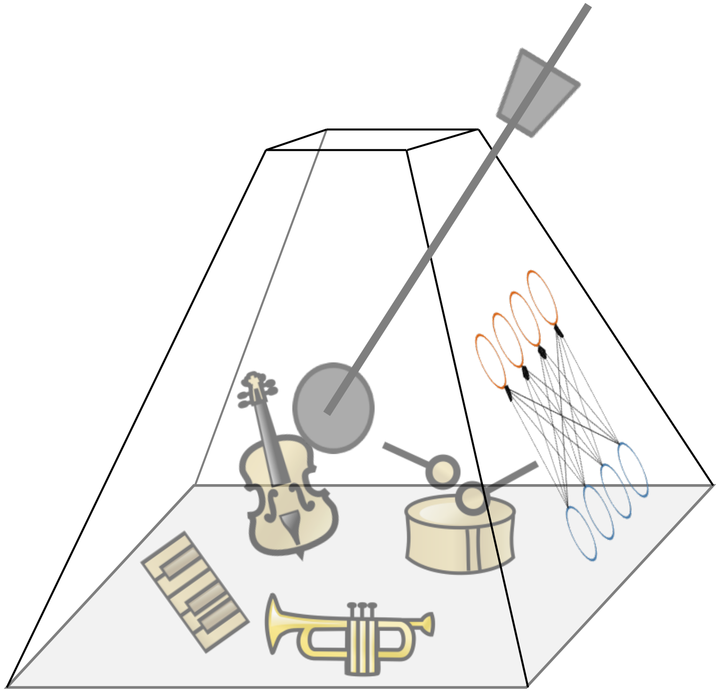
- **Input:** Log-freq. Spectrogram (120 semitones, 10 Hz feature rate)
- **Target:** Solo instrument’s pitch activations
- **Output:** Pitch activations (120 semitones, 10 Hz feature rate)
- **Architecture:** FNN, 5 hidden layers, ReLU, Loss: MSE, layer-wise training



Feature Learning

- Less domain knowledge needed to learn working features.
- Know your task/data.
  Accuracy is not everything!





Beat and Rhythm Analysis

Beat and Rhythm Analysis

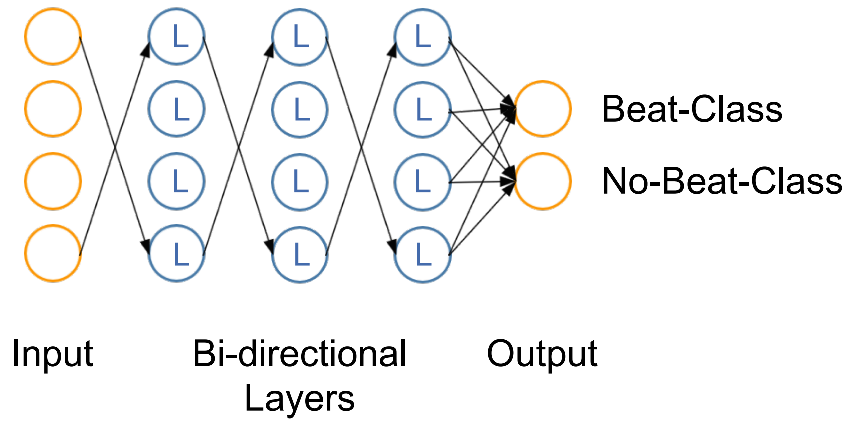
Task	Year	Authors	Ref.	Type	Input	Pre-proc.
BRA	2010	Eyben et al.	[25]	RNN-BLSTM	LogMelS	DERIV
BRA	2011	Böck and Schedl	[5]	RNN-BLSTM	LogMelS	DERIV
BRA	2012	Battenberg and Wessel	[3]	DBN	—	—
BRA	2014	Böck et al.	[7]	RNN-BLSTM	LogS	—
BRA	2016	Böck et al.	[9]	RNN-BLSTM	LogS	DERIV
BRA	2016	Elowsson	[23]	FNN	HC	—
BRA	2016	Holzapfel and Grill	[35]	CNN	LogLogS	STDF
BRA	2016	Krebs et al.	[46]	RNN-BGRU	HC	—
BRA	2016	Durand and Essid	[21]	CNN	HC	—
BRA	2017	Durand et al.	[22]	CNN	HC	—
BRA	2015	Böck et al.	[8]	RNN-BLSTM	LogMelS	DERIV

- **Beat Tracking:**
  Find the pulse in the music which you would tap/clap to.

Beat and Rhythm Analysis

Sebastian Böck, Florian Krebs, and Gerhard Widmer, DAFx 2011

- **Input:** 3 LogMel spectrograms (varying win-length) + derivatives
- **Target:** Beat annotations
- **Output:** Beat activation function $\in [0, 1]$
- **Post-processing:** Peak picking on beat activation function
- **Architecture:** RNN, 3 bidirectional layers, 25 LSTM per layer/direction



AUDIO LABS

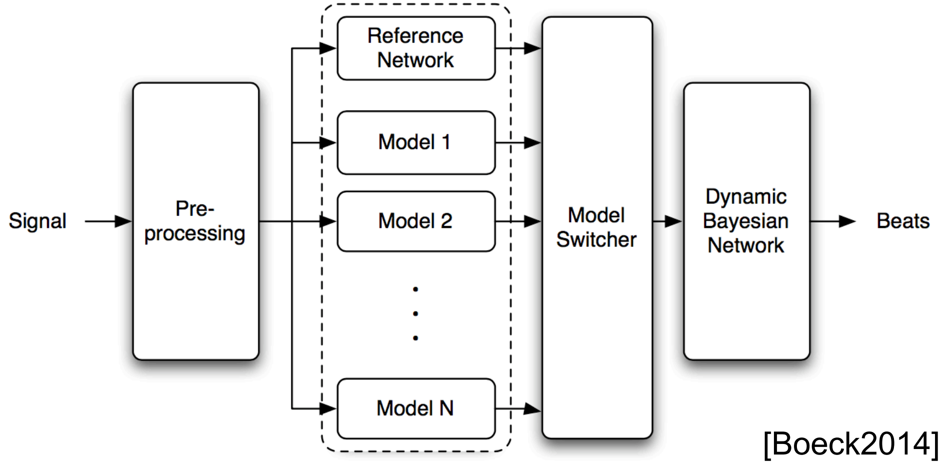
Beat Tracking
Examples

	Borodin String Quartet 2, III. 65 bpm	Carlos Gardel Por una Cabeza 114 bpm	Sidney Bechet Summertime 87 bpm	Wynton Marsalis Caravan 195 bpm	Wynton Marsalis Cherokee 327 bpm
Original					
Ellis (librosa) Init = 120 bpm					
Böck2015 (madmom)					

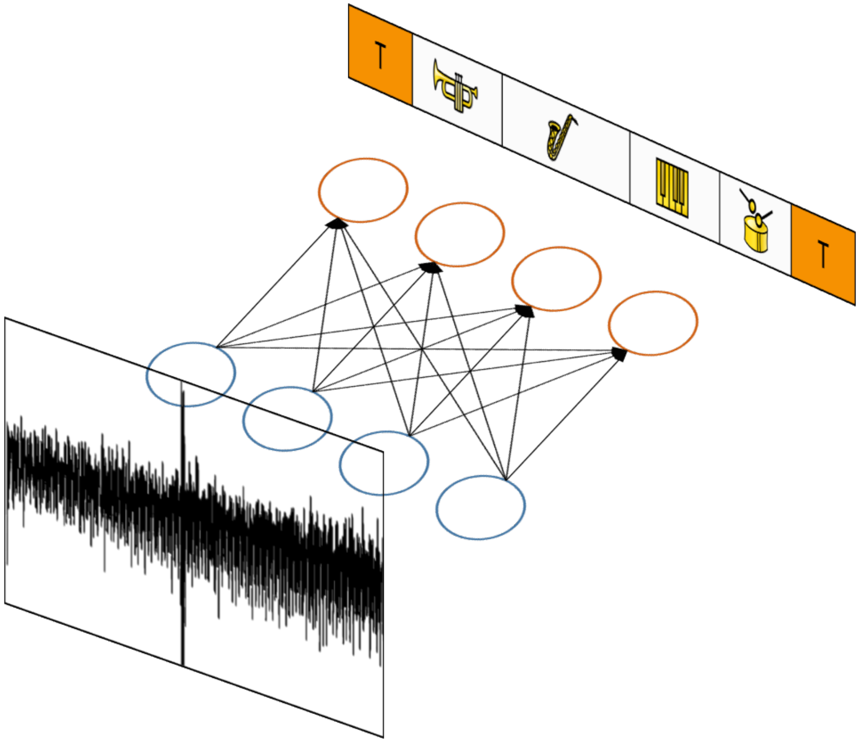
Beat Tracking

- DNN-based methods need less task-specific initialization (e.g., tempo).
- Closer to a “universal” onset detector.

- Task-specific knowledge is introduced as post-processing step:



[Boeck2014]

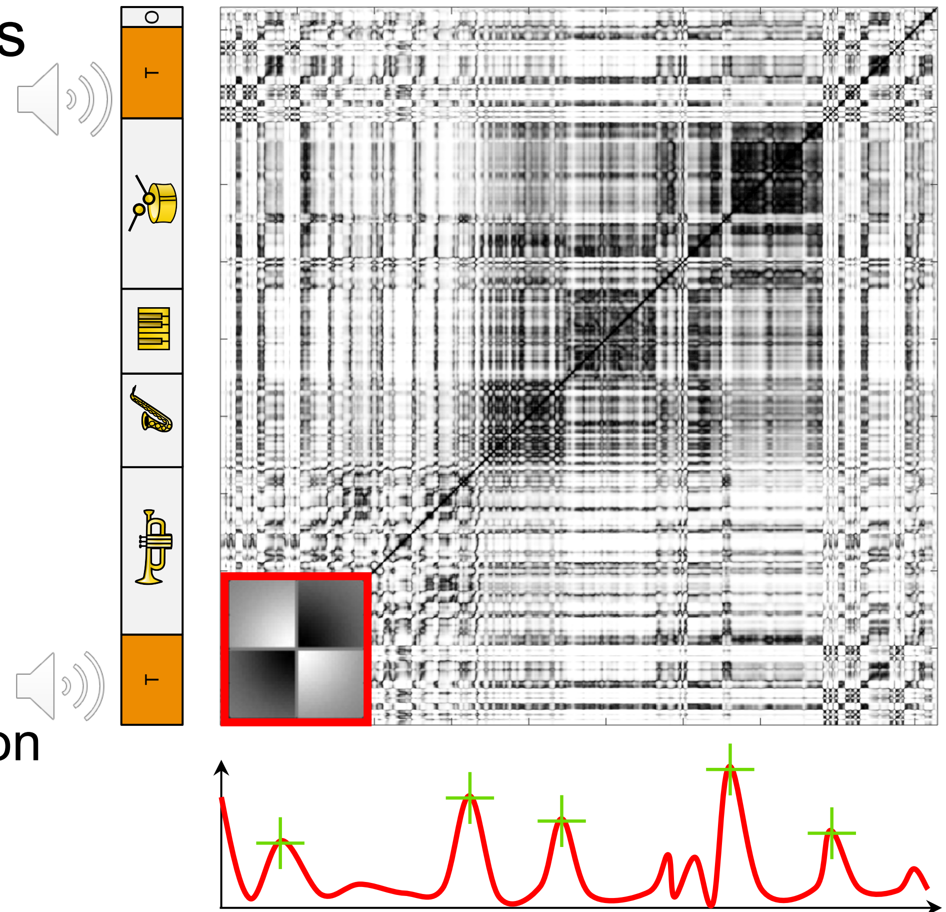


Music Structure Analysis

Music Structure Analysis

Task	Year	Authors	Ref.	Type	Input	Pre-proc.
MSA	2017	Cohen-Hadria and Peeters	[14]	CNN	LogMelS, SSM	—
MSA	2014	Ullrich et al.	[75]	CNN	LogMelS	—
MSA	2015	Grill and Schlüter	[28]	CNN	LogMelS	—
MSA	2015	Grill and Schlüter	[29]	CNN	LogMelS	HPSS

- Find boundaries/repetitions in music
- Classic approaches:
  - Repetition-based
  - Homogeneity-based
  - Novelty-based
- Main challenges:
  - What is structure?
  - Model assumptions based on musical rules (e.g., sonata).

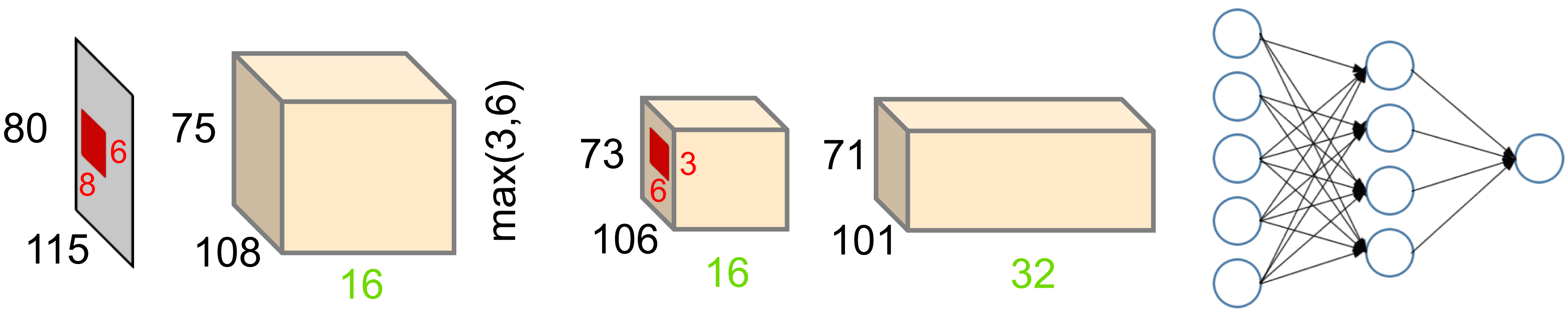


[Foote]

Music Structure Analysis

Karen Ullrich, Jan Schlüter, and Thomas Grill, ISMIR 2014

- **Input:** LogMel spectrogram
- **Target:** Boundary annotations
- **Output:** Novelty function $\in [0, 1]$
- **Post-processing:** Peak picking on novelty function



71 * 101 * 32 = 229'472 128 1

8 * 6 * 16
= 768

6 * 3 * 16 * 32
= 9216

229'472 * 128
= 29'372'416

128 * 1
= 128

* ignoring bias

Music Structure Analysis

Results

Tolerance	SALAMI 1.3 Ullrich et al. (2014)

0.5 s:

Algorithm	F-measure	Precision	Recall
Upper bound (est.)	0.68		
16s_std_1.5s	**0.4646**	0.5553	0.4583
MP2 (2013)	0.3280	0.3001	0.4108
MP1 (2013)	0.3149	0.3043	0.3605
OYZS1 (2012)	0.2899	0.4561	0.2583

3.0 s:

Algorithm	F-measure	Precision	Recall
Upper bound (est.)	0.76		
32s_low_6s	**0.6164**	0.5944	0.7059
16s_std_1.5s	0.5726	0.5648	0.6675
MP2 (2013)	0.5213	0.4793	0.6443
MP1 (2013)	0.5188	0.5040	0.5849

SALAMI 2.0

Grill et al. (2015)

Algorithm	**F_1**	**$F_{.58}$**	**Rec.**	**Prec.**
Upper bound (est.)	.74	.74		
All features, multi+fine ann.	**.508**	.529	.502	.572
MLS+SSLM-near, multi+fine	.496	.506	.509	.536
MLS+SSLM-near, single ann.	.469	.466	.504	.475
SUG1 (2014)	.422	.442	.422	.490
MP2 (2013)	.294	.280	.362	.271
MP1 (2013)	.276	.270	.311	.269
NB1 (2014)	.270	.246	.374	.229
KSP2 (2012)	.263	.231	.422	.209
Baseline (est.)	.15	.21		

- Added features (SSLM)
- Trained on 2 levels of annotations
- SUG1 is similar to [Ullrich2014]

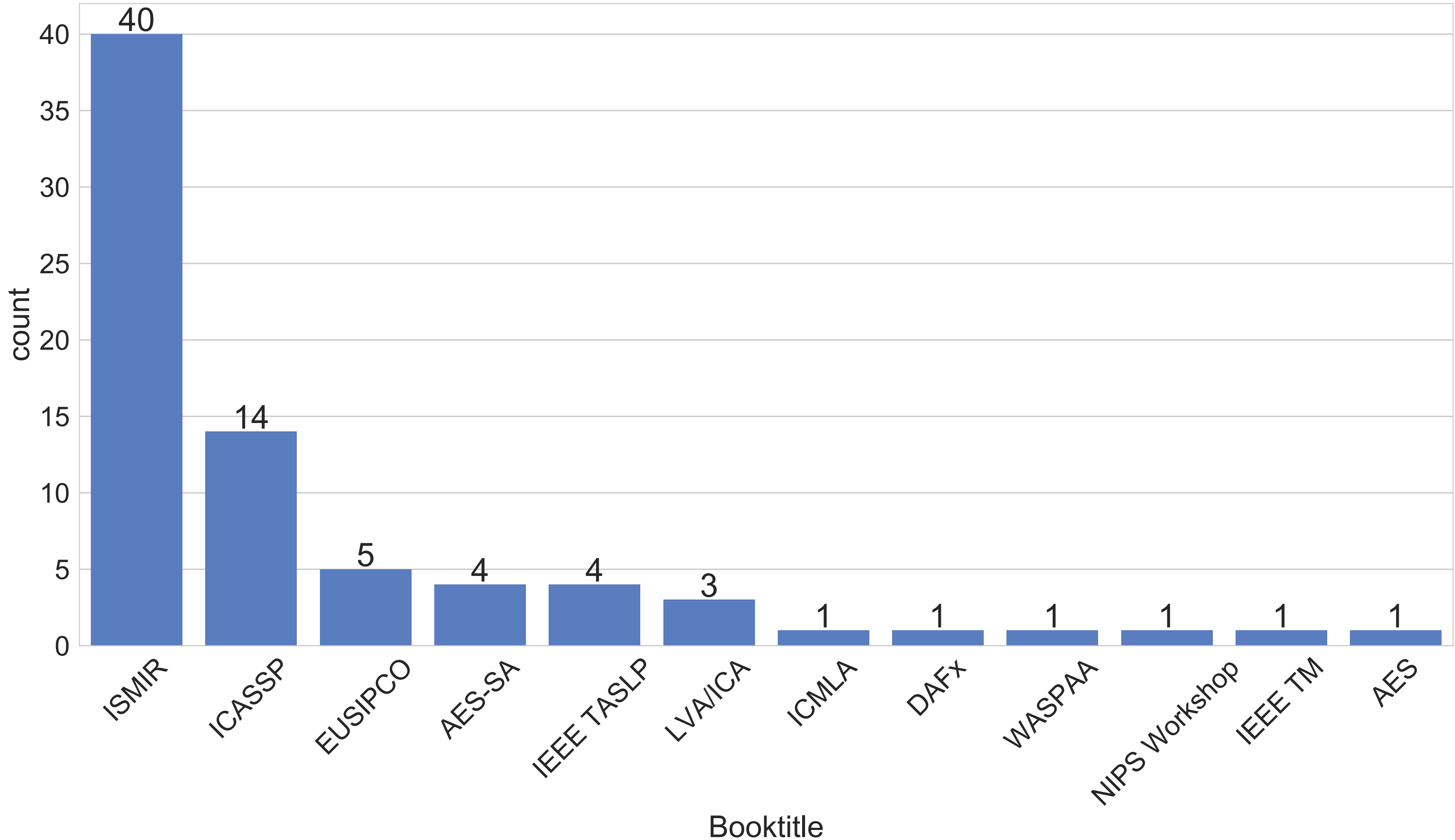
Music Structure Analysis

Task	Year	Authors	Ref.	Type	Input	Pre-proc.
MSA	2017	Cohen-Hadria and Peeters	[14]	CNN	LogMelS, SSM	—
MSA	2014	Ullrich et al.	[75]	CNN	LogMelS	—
MSA	2015	Grill and Schlüter	[28]	CNN	LogMelS	—
MSA	2015	Grill and Schlüter	[29]	CNN	LogMelS	HPSS

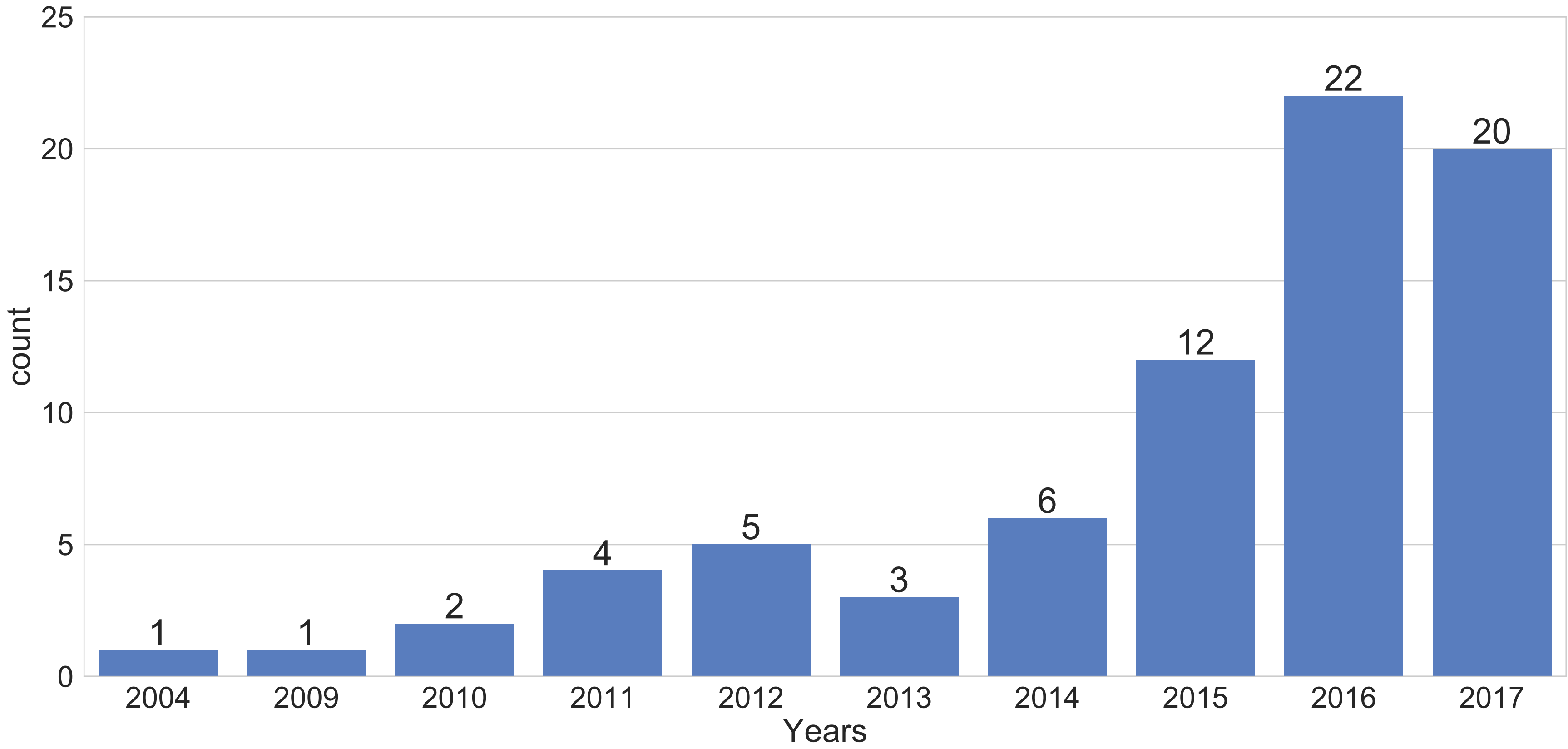
- Re-implementation by *Cohen-Hadria and Peeters* did not reach reported results.
- Possible reasons:
  - Data identical?
  - Different kind of convolution? What was the stride?
  - Didn’t ask?
  - Availability of pre-trained model would be awesome!

Literature Overview

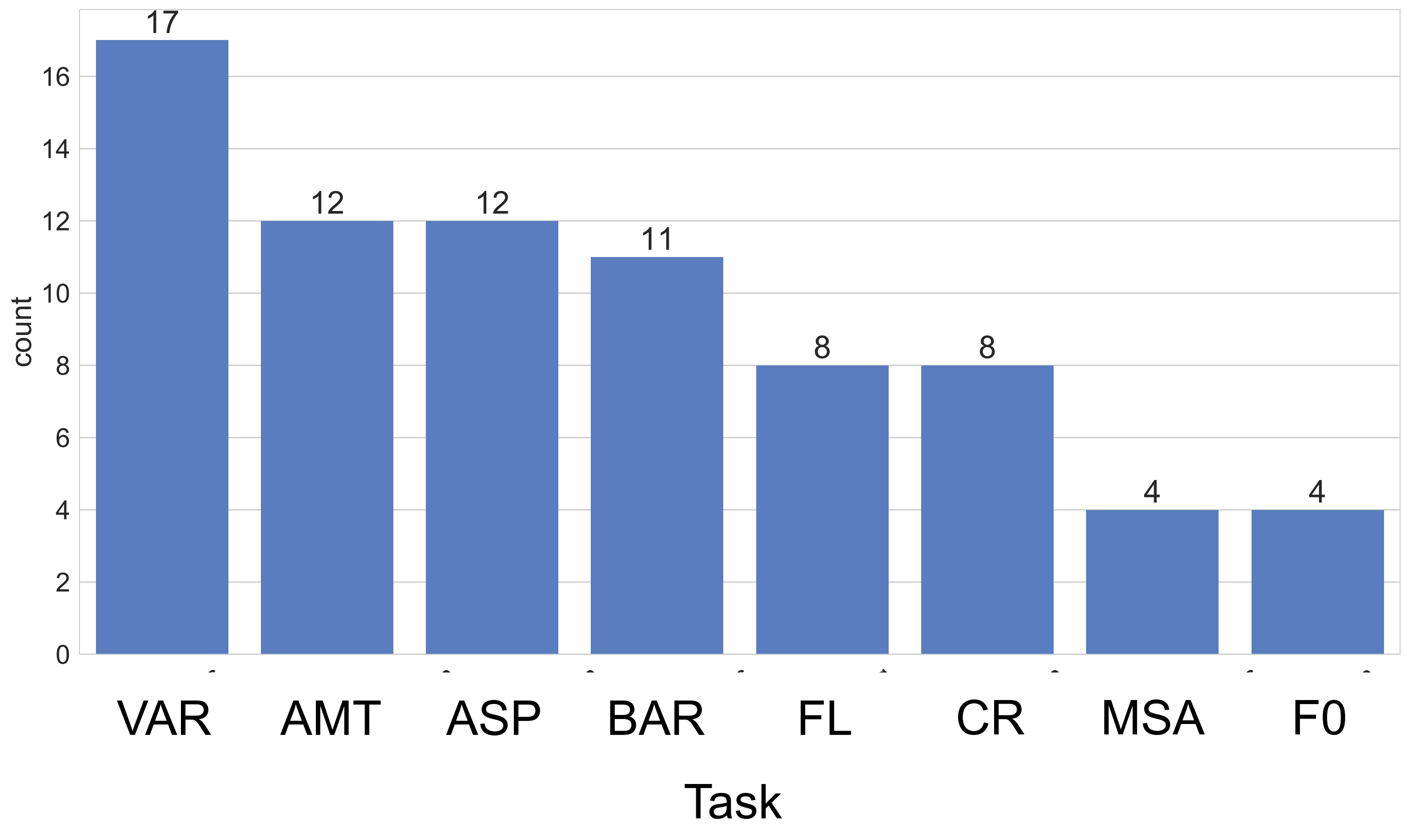
Publications by Conference



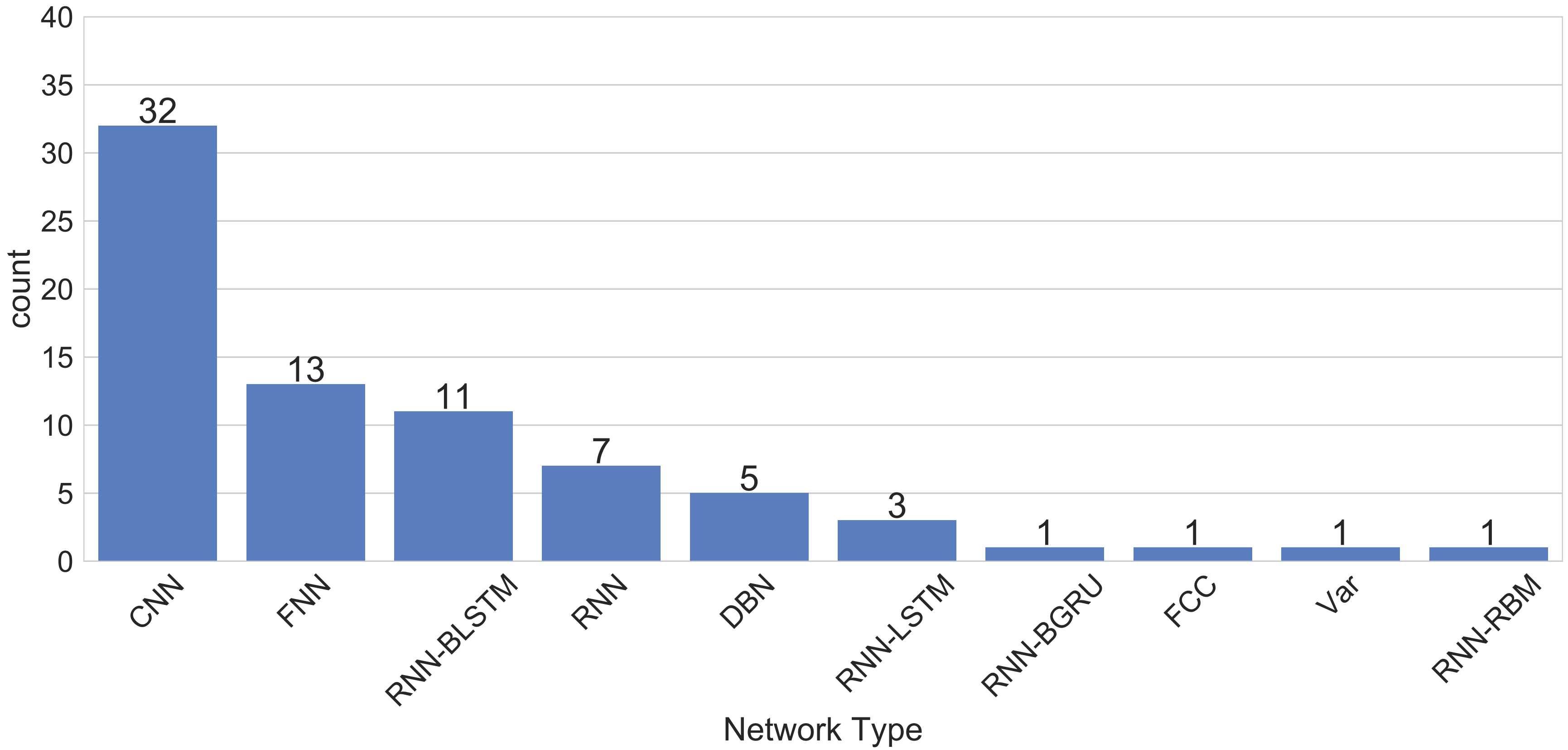
Publications by Year



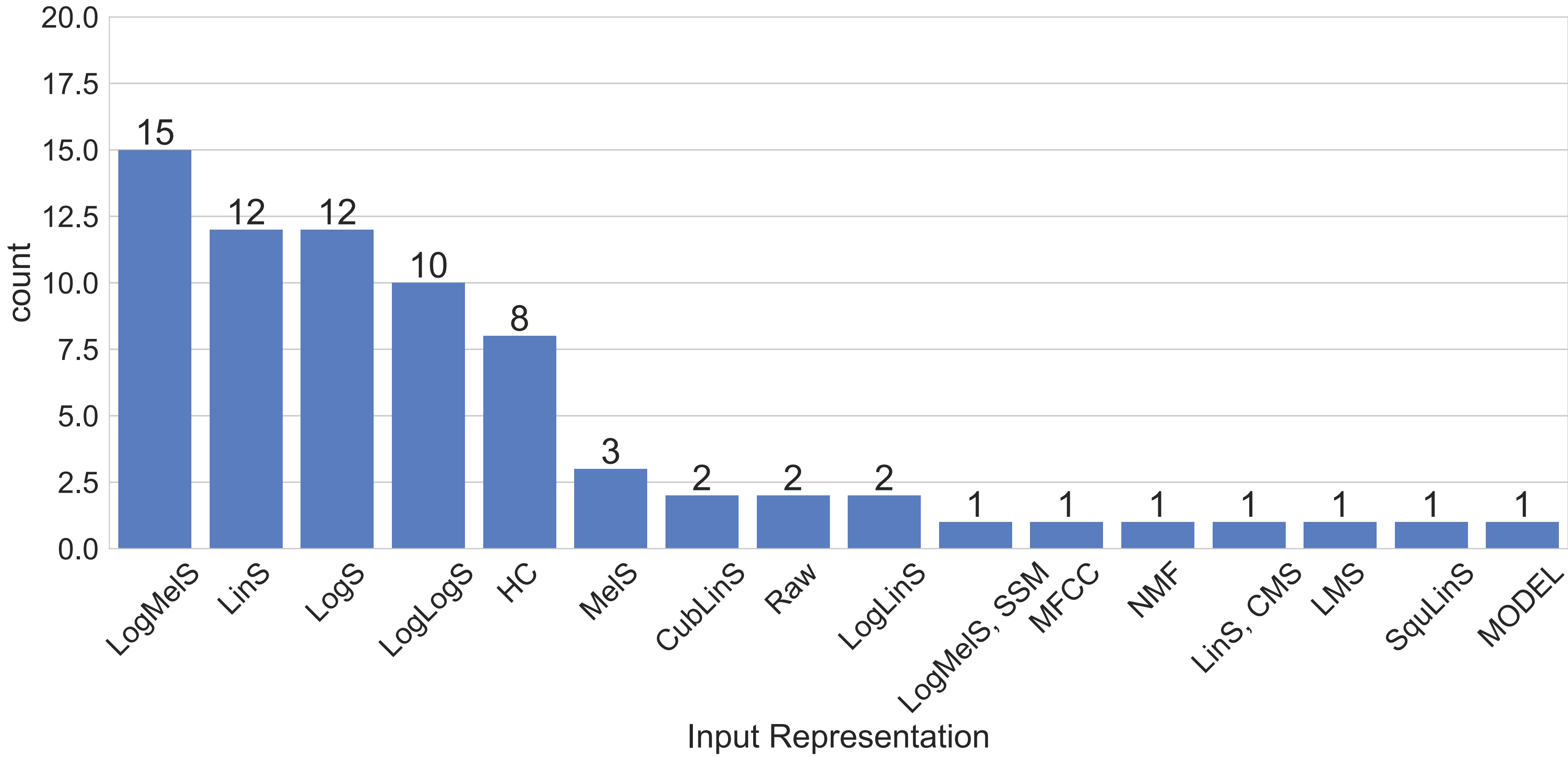
Publications by Task



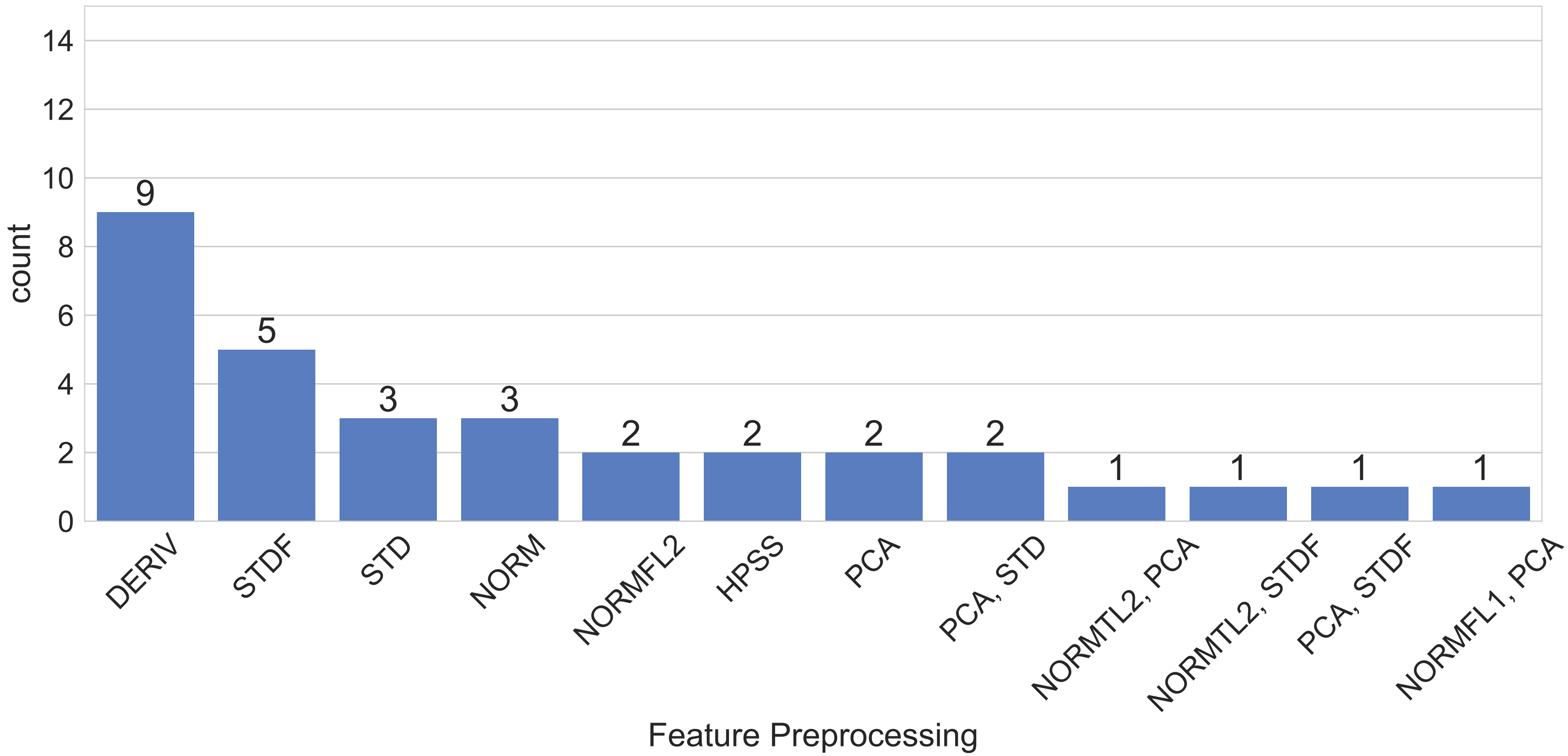
Publications by Network



Input Representations



Feature Preprocessing



Deep Neural Networks in MIR

- Other resources:
  - Jordi Pons
    http://jordipons.me/wiki/index.php/MIRDL
  - Keunwoo Choi
    https://docs.google.com/spreadsheets/d/1cIR7sp-HFDs7UI72CA-98yFc5fimQxMrq13e4fj3iA4
  - Yann Bayle
    https://github.com/ybayle/awesome-deep-learning-music

- Work in progress…

Conclusion

- How can we contribute to the progress of DNN research?
  - Provide well-/ill-defined tasks and labeled data.
  - Much existing experience for sanity-checks (e.g., network inspection, feature sonification).
  - Explore generalization with different genres.
  - Tweak architectures for a given task (e.g., use musical knowledge).

- Interested in the “report“?

- Interested in jazz music? Happy to collaborate!