upf. Universitat Pompeu Fabra Barcelona
MTG Music Technology Group

Deep learning for music data processing

A personal (re)view of the state-of-the-art

Jordi Pons

www.jordipons.me

Music Technology Group, DTIC,
Universitat Pompeu Fabra, Barcelona.

31st January 2017

What **problems** do we care about in **music technology research**?

- (Automatically) cataloging large-scale music collections.
- Music recommendation.
- Similarity – *ie.* Shanzam.
- Synthesis: instruments, singing voice.
- ...

Some of them can be approached with **deep learning**.

Why deep learning might be useful for music data processing?

- Music is **hierarchic** in frequency (note, chord) and time (onset, rhythm) and deep learning naturally allows this representation.
- **Contextual analysis**
  - Short time-scale features: CNNs - *ie.* note, chords.
  - Long time-scale features: RNNs - *ie.* structure.
- **Unsupervised learning**: potential of learning from any audio!
- **Time/frequency invariant** operations: max-pool.
- **Any input**: spectrogram, MFCCs, self-similarity matrices, video, text.

Acronyms:

- **MLP**: multi layer perceptron $\equiv$ feed-forward neural network.
- **RNN**: recurrent neural network.
- **LSTM**: long-short term memory.
- **CNN**: convolutional neural network.

Assumed notion of *deep learning*:

- It is *deep* when **several non-linearities** are applied to the input.
- The parameters of the network are *learnt*:
  $\rightarrow$ typically by using **back-propagation**.

Chronology: the big picture

Some papers as examples for discussion

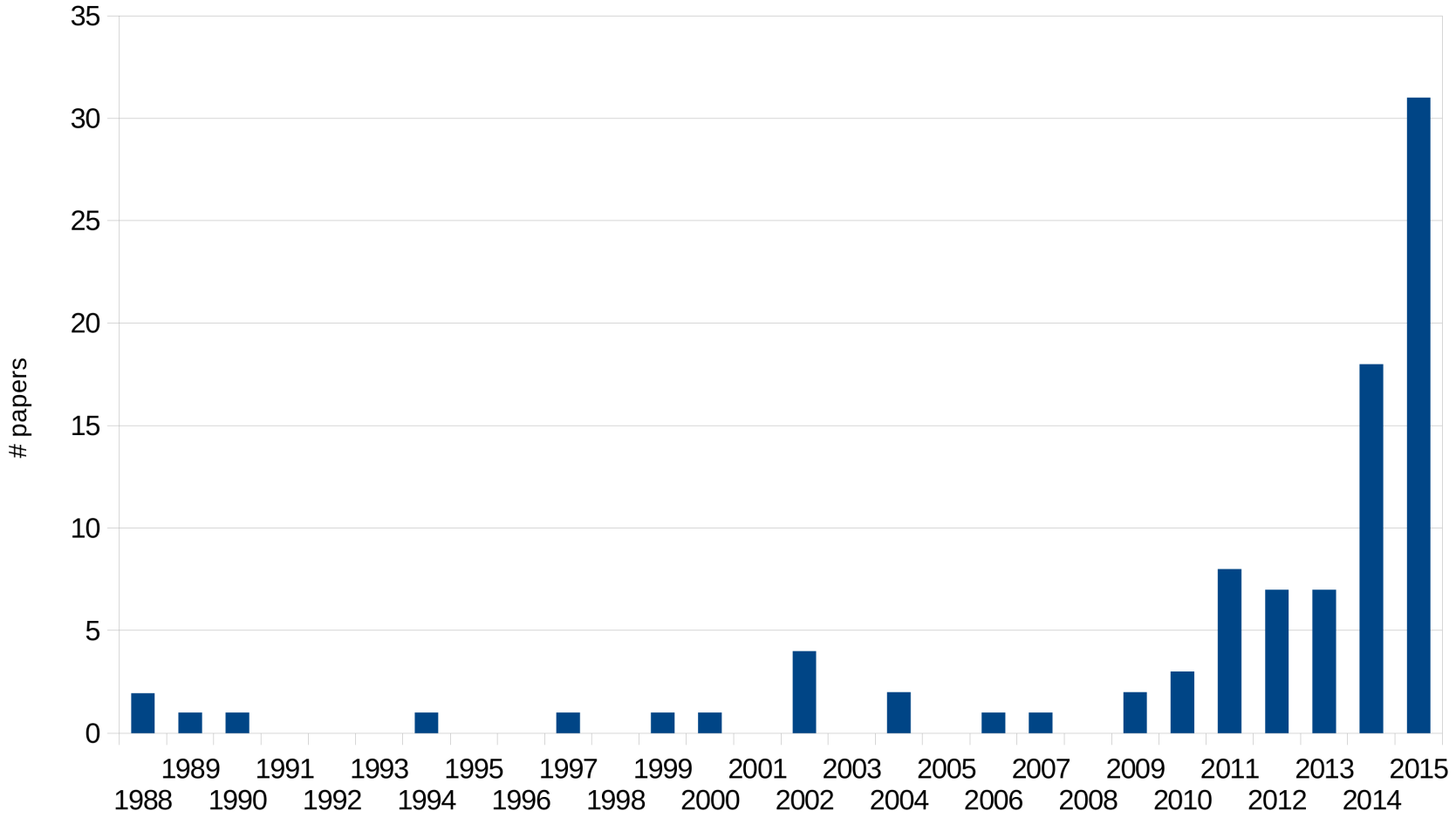
Current trends and research directions

Chronology: the big picture

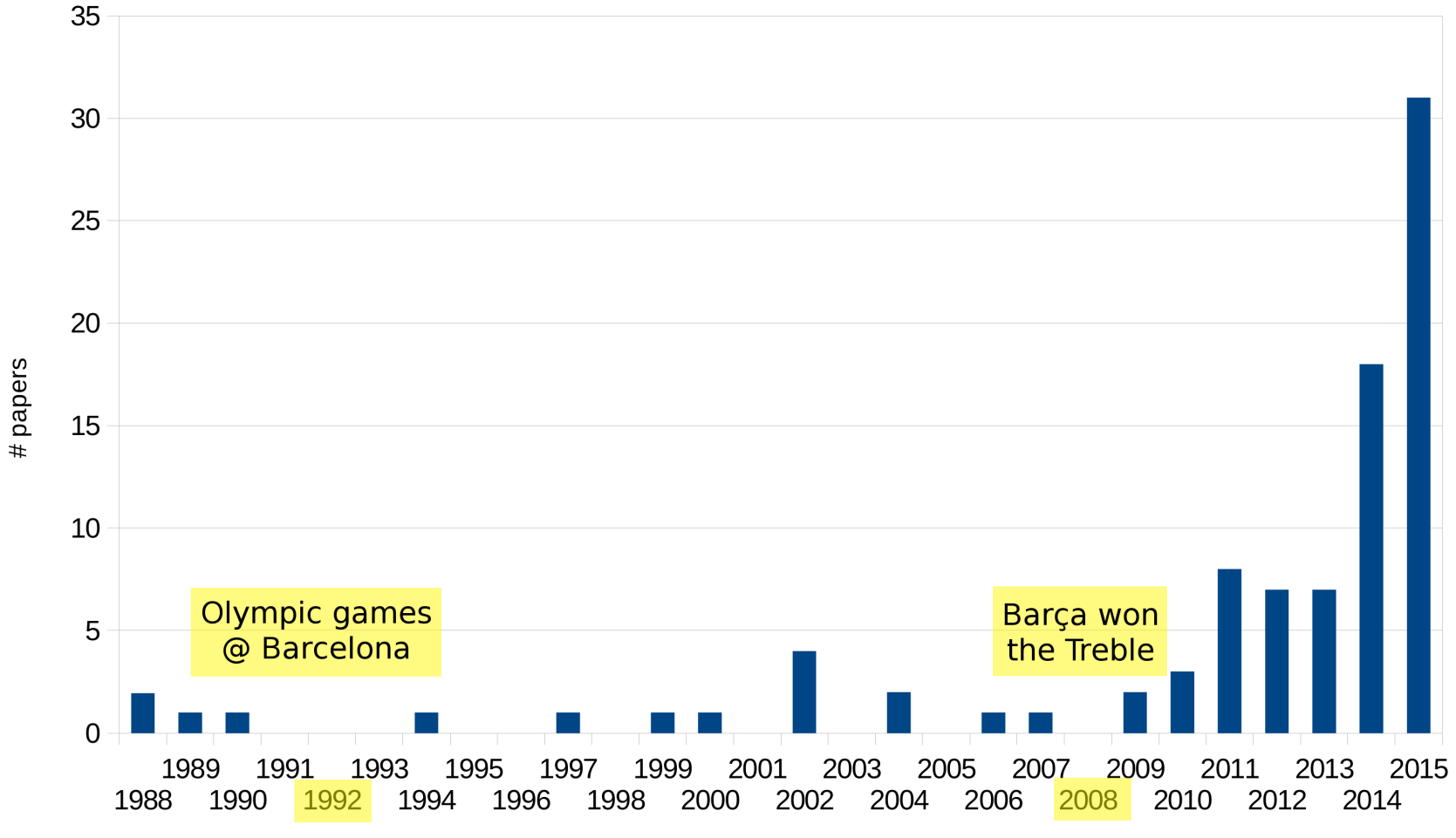
Some papers as examples for discussion

Current trends and research directions

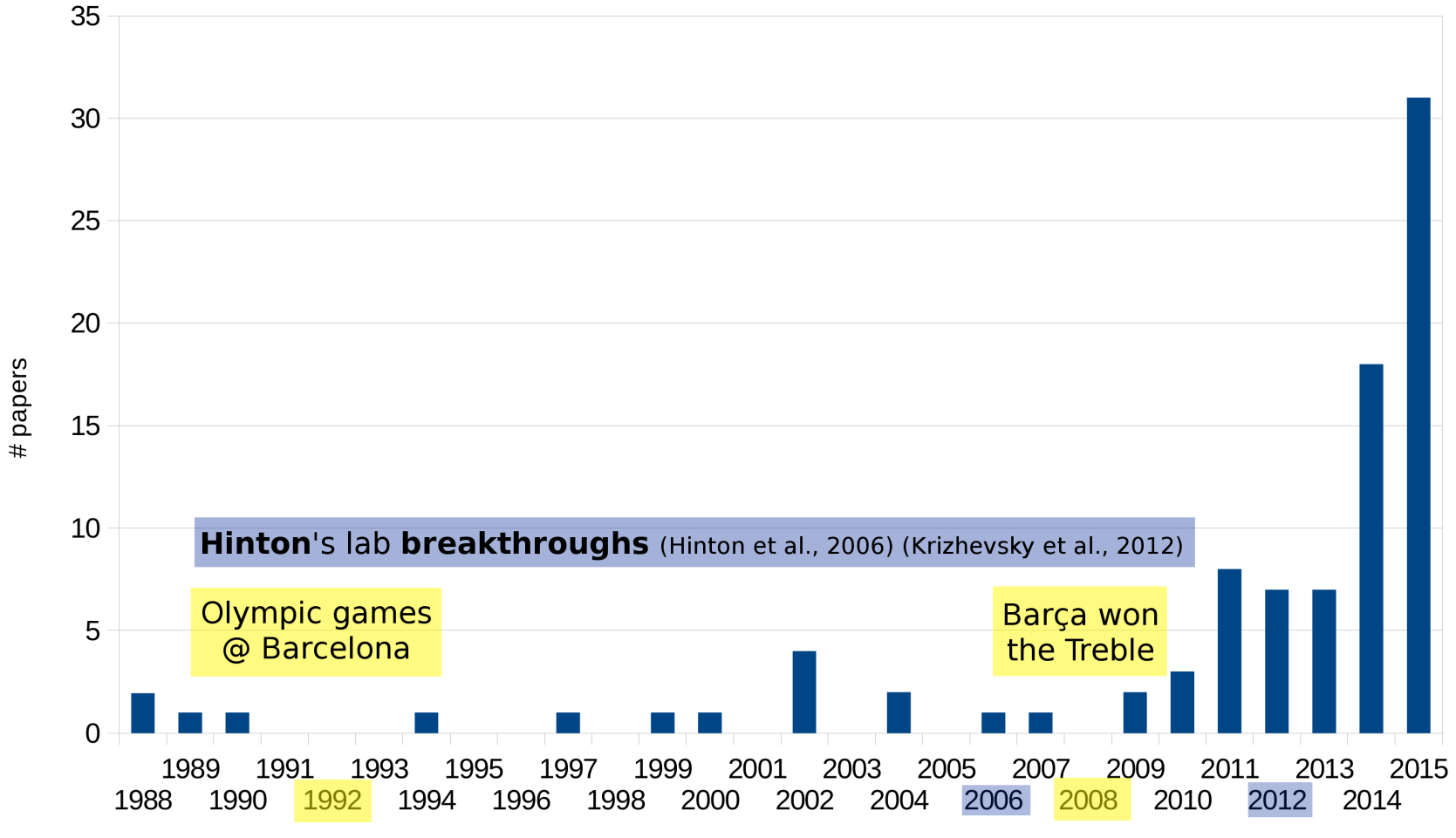
Approximate distribution: **deep learning** papers for **music** data processing over the years

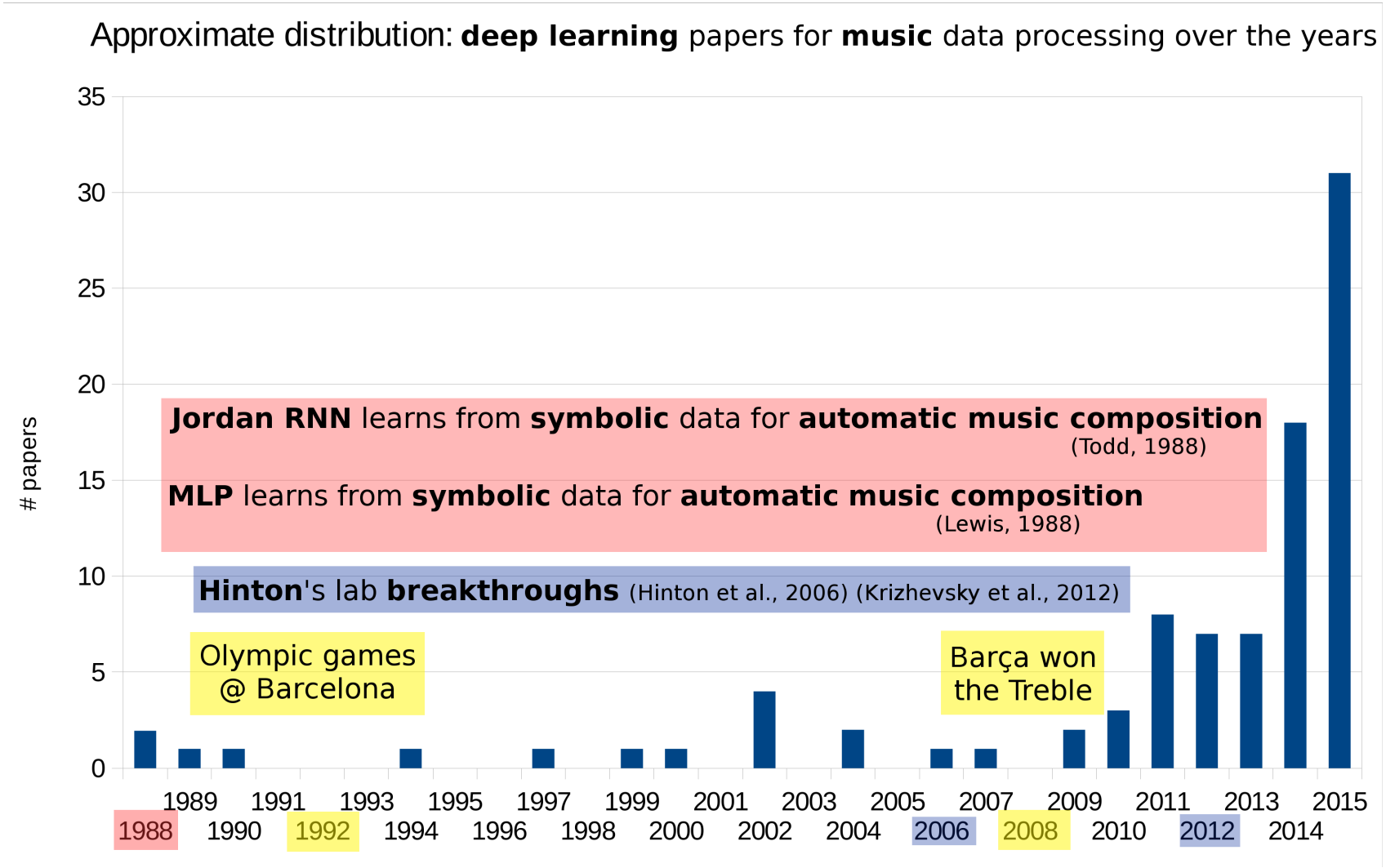


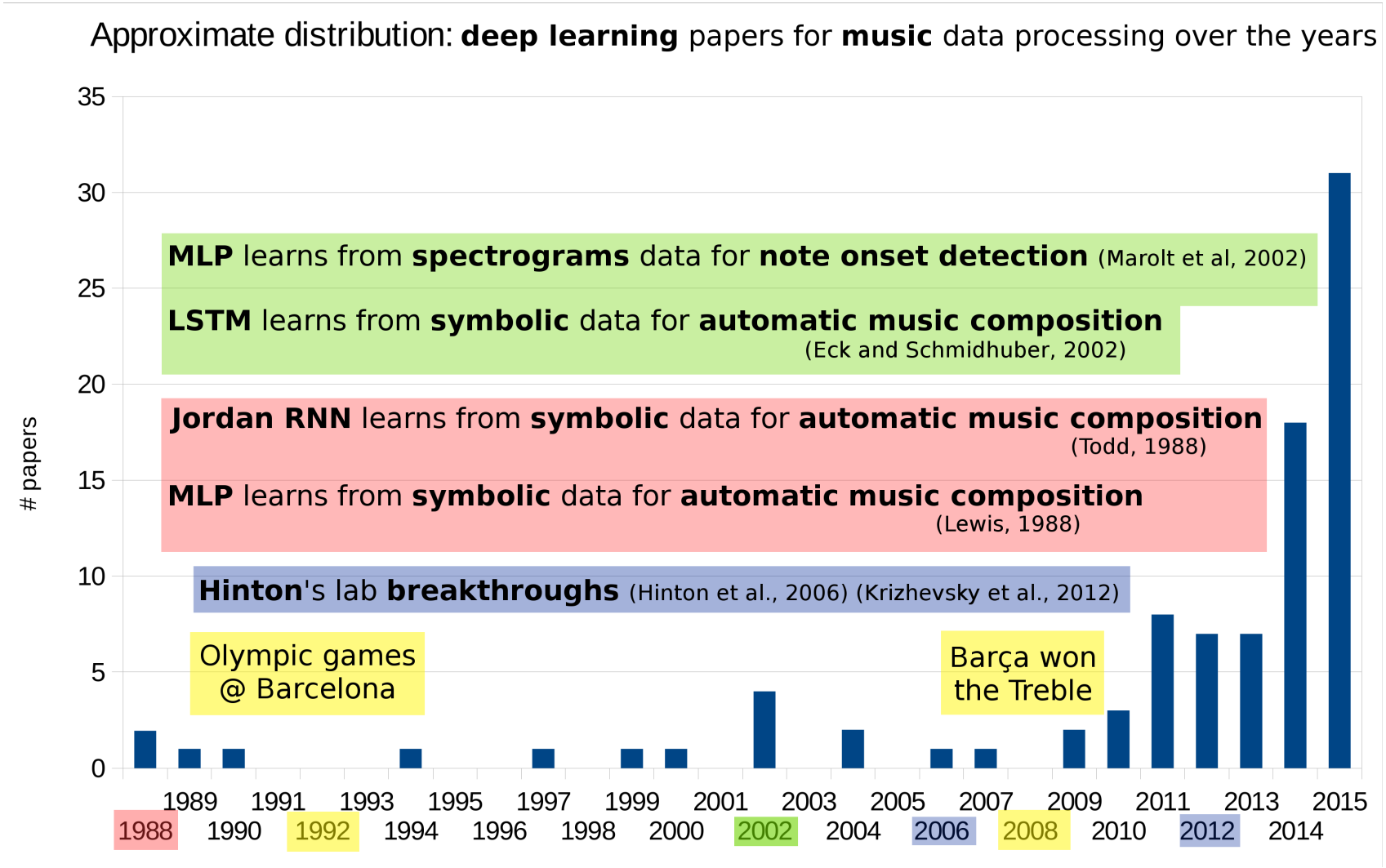
Approximate distribution: **deep learning** papers for **music** data processing over the years



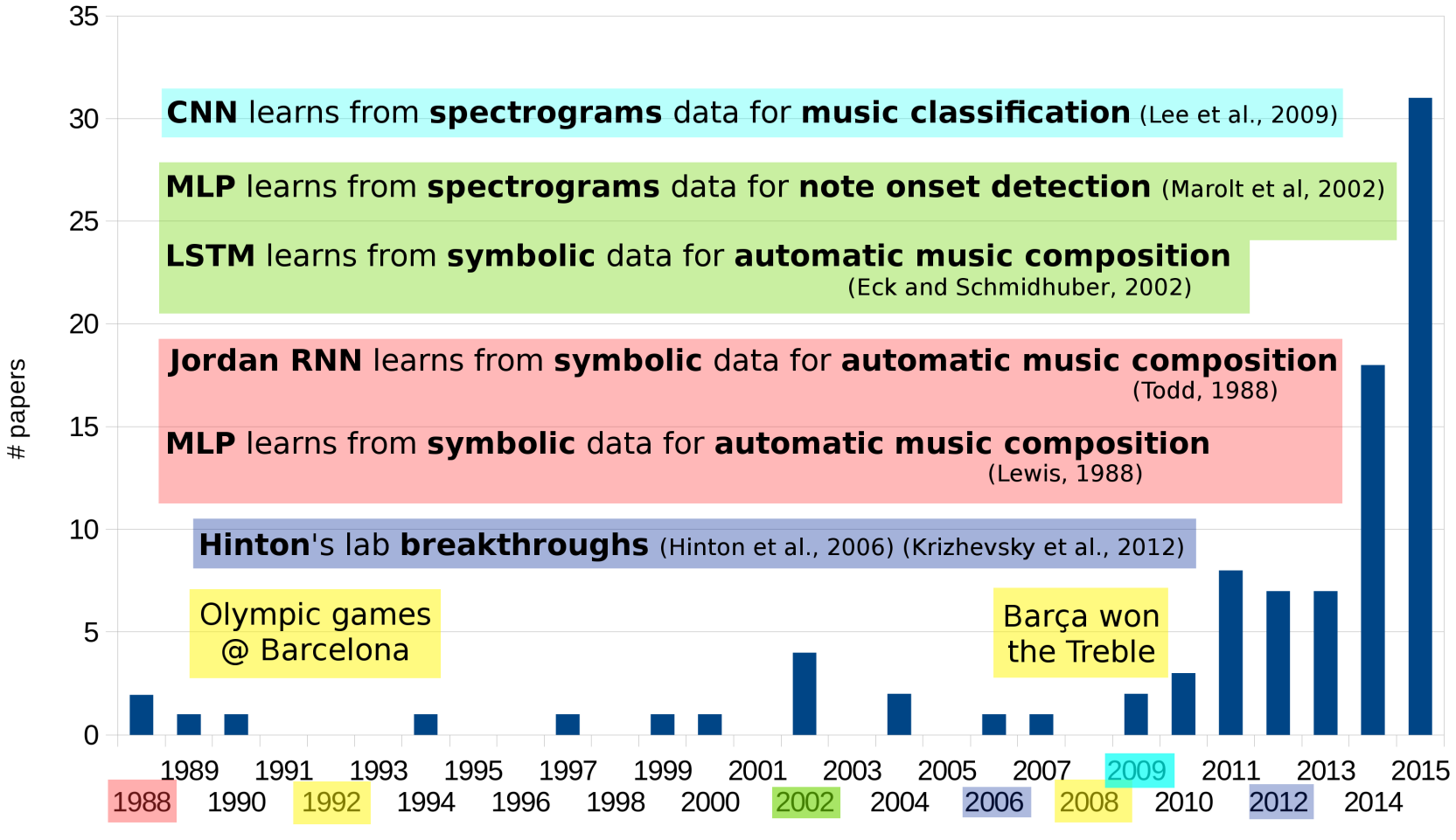
Approximate distribution: **deep learning** papers for **music** data processing over the years

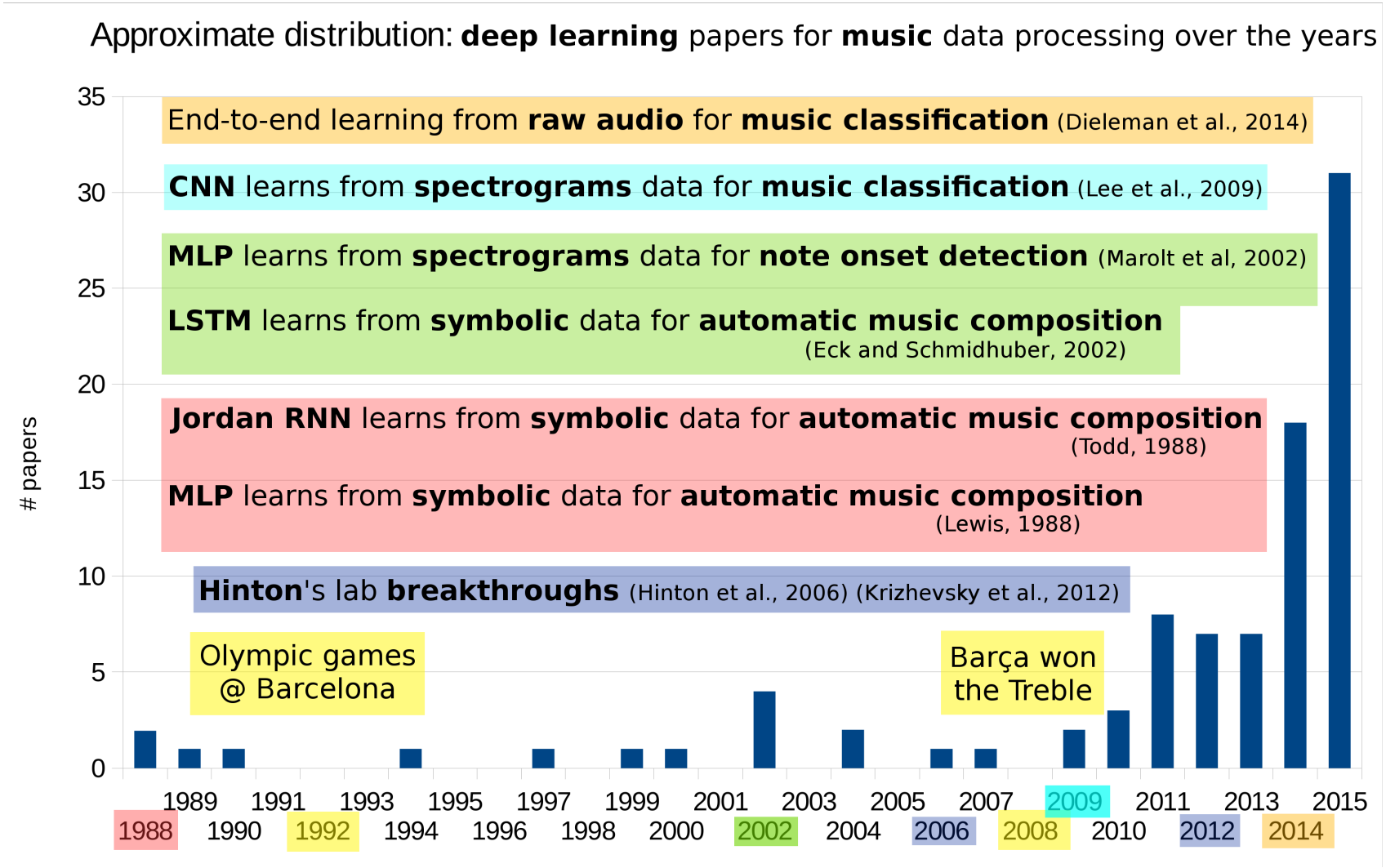



Approximate distribution: deep learning papers for music data processing over the years
papers
35
30
25
20
15
10
5
0
Jordan RNN learns from symbolic data for automatic music composition
(Todd, 1988)
MLP learns from symbolic data for automatic music composition
(Lewis, 1988)
Hinton's lab breakthroughs (Hinton et al., 2006) (Krizhevsky et al., 2012)
Olympic games @ Barcelona
Barça won the Treble
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

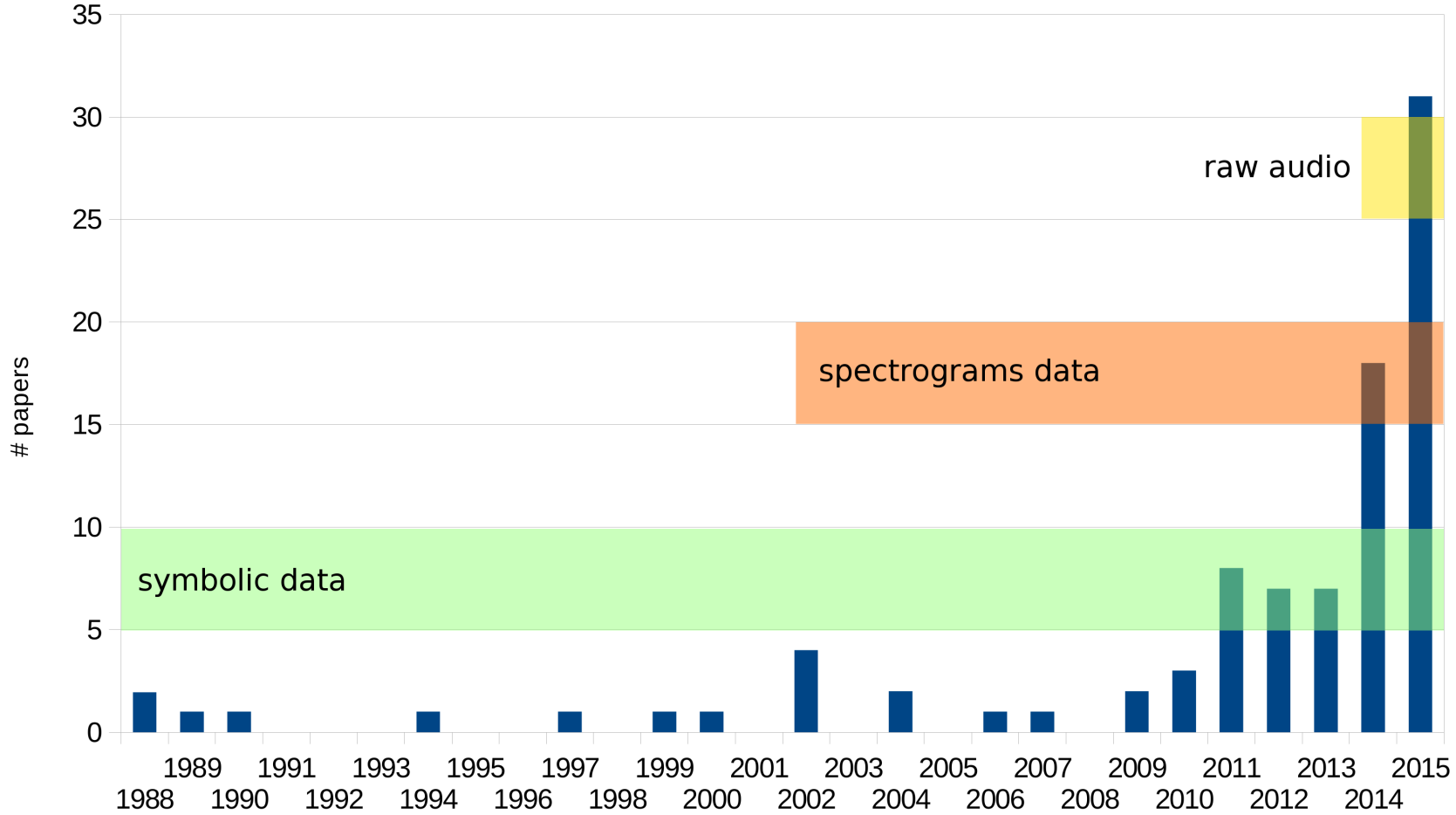

Approximate distribution: deep learning papers for music data processing over the years
papers
35
30
25
20
15
10
5
0
MLP learns from spectrograms data for note onset detection (Marolt et al, 2002)
LSTM learns from symbolic data for automatic music composition
(Eck and Schmidhuber, 2002)
Jordan RNN learns from symbolic data for automatic music composition
(Todd, 1988)
MLP learns from symbolic data for automatic music composition
(Lewis, 1988)
Hinton's lab breakthroughs (Hinton et al., 2006) (Krizhevsky et al., 2012)
Olympic games @ Barcelona
Barça won the Treble
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

Approximate distribution: **deep learning** papers for **music** data processing over the years




Approximate distribution: deep learning papers for music data processing over the years
papers
End-to-end learning from raw audio for music classification (Dieleman et al., 2014)
CNN learns from spectrograms data for music classification (Lee et al., 2009)
MLP learns from spectrograms data for note onset detection (Marolt et al, 2002)
LSTM learns from symbolic data for automatic music composition
(Eck and Schmidhuber, 2002)
Jordan RNN learns from symbolic data for automatic music composition
(Todd, 1988)
MLP learns from symbolic data for automatic music composition
(Lewis, 1988)
Hinton's lab breakthroughs (Hinton et al., 2006) (Krizhevsky et al., 2012)
Olympic games @ Barcelona
Barça won the Treble
0
5
10
15
20
25
30
35
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

Approximate distribution: **deep learning** papers for **music** data processing over the years



Used for:

- **Classification**: genre, artist, singing-voice detection, music-speech. Pons et al., Lidy et al.
- **Auto-tagging**. Dieleman et al., Choi et al.
- **Key** estimation. Humphrey et al., Korzeniowski et al.
- **Feature extraction** (unsupervised). Hamel et al., Lee et al.
- Music **similarity** estimation. Schlüter et al.
- Music **recommendation**. Aäron van den Oord et al.
- **Onset**/boundary detection. Böck et al., Durand et al.
- **Source separation**. Huang et al., Miron et al.
- Singing voice **synthesis**. Blaauw et al.

Chronology: the big picture

Some papers as examples for discussion

Current trends and research directions

LSTMs for automatic music composition with symbolic data

Eck and Schmidhuber. Learning The Long-Term Structure of the Blues. ICANN'02

"..compositions are quite pleasant"

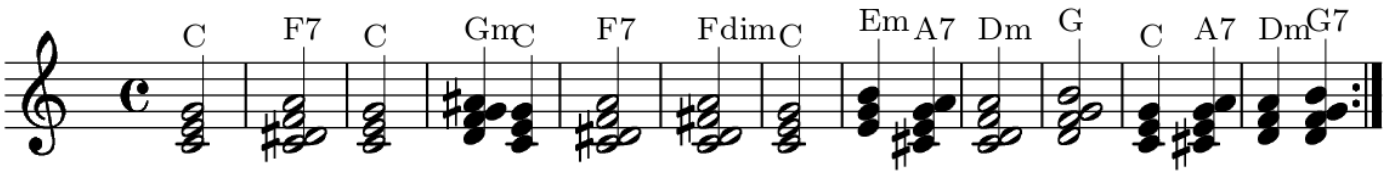


Fig. 1. Bebop-style blues chords used for training data (transposed up one octave).

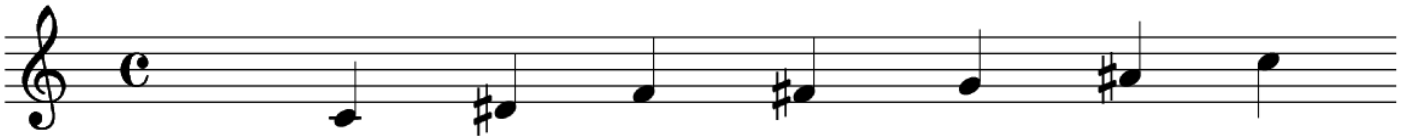


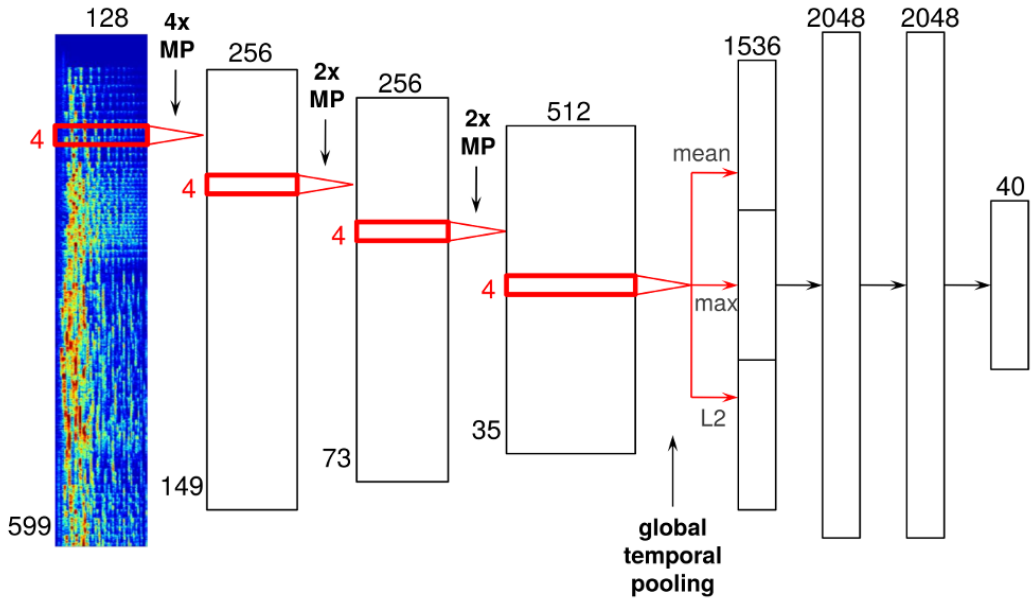
Fig. 2. Pentatonic scale used for training data melodies.

Some **examples** of **music composed** by LSTMs:

1. Bob Sturm plays: The Mal's Copporim.
2. LSTMetallica: Drums from Metallica. *Choi et al.*
3. LSTM Realbook: Generation of Jazz chord progressions.

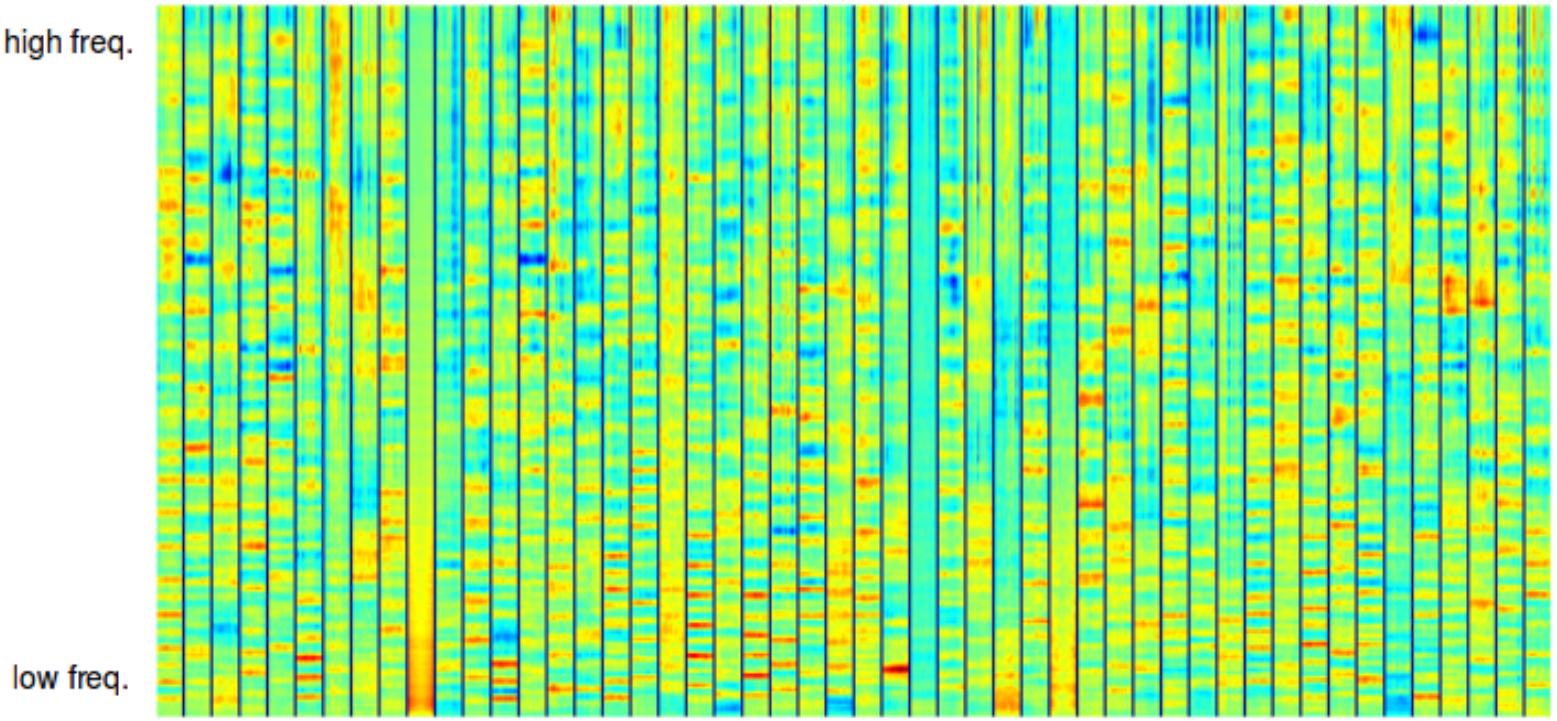
CNNs interpretation and filter shapes discussion

S. Dieleman. *http://benanne.github.io/2014/08/05/spotify-cnns.html*



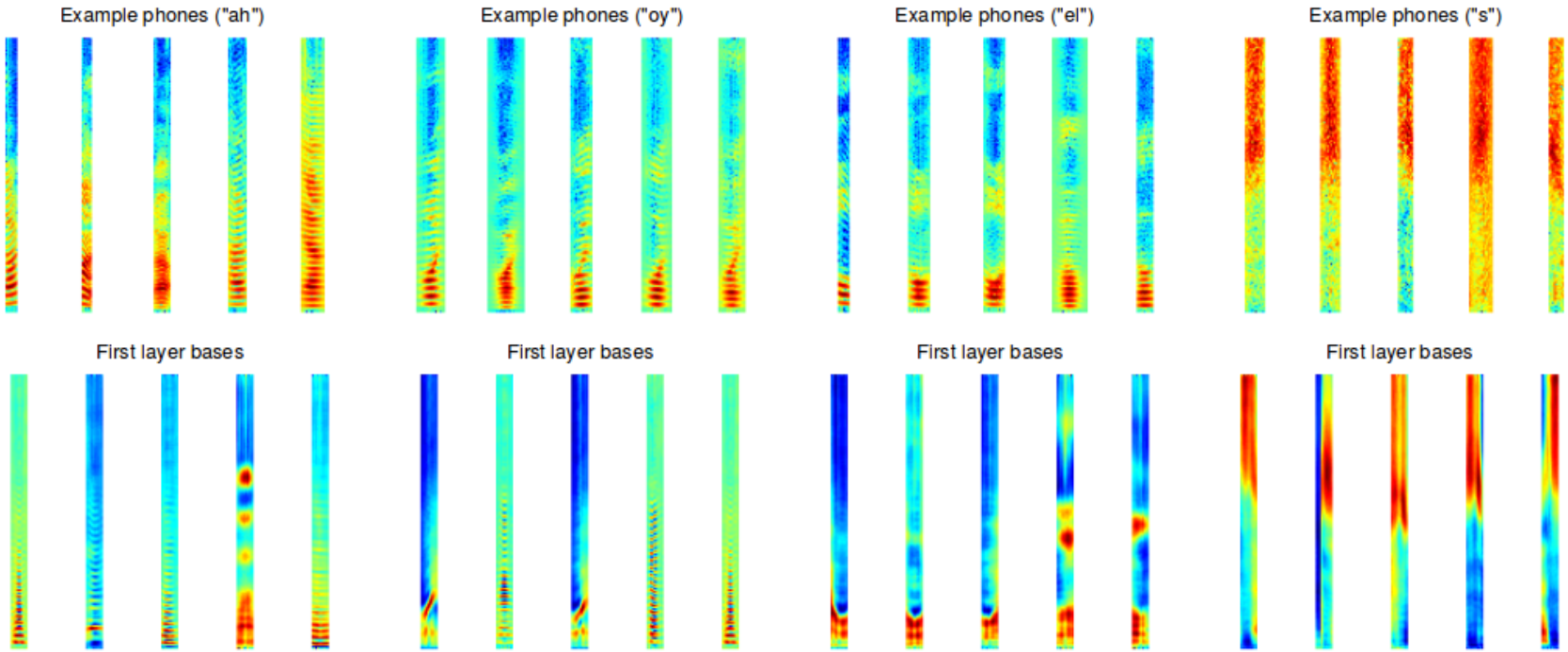
- **Content-based music recommendation** @ Spotify.
- CNN is learning (music) hierarchical features:
  - L1 Vibrato, vocal thirds, bass drums, A/Bb pitch, A/Am chord.
  - L3 Christian rock, Chinese pop, 8-bit, multimodal.

Lee et al. Unsupervised feature learning for audio classification using convolutional deep belief networks. NIPS'09



Visualization of some randomly selected first-layer convolutional filters trained with **music**.

Lee et al. Unsupervised feature learning for audio classification using convolutional deep belief networks. NIPS'09



Visualization of the four different phonemes and their corresponding first-layer convolutional filters trained with **speech**.

Choi et al. Explaining Deep CNNs on Music Classification. arXiv:1607.02444

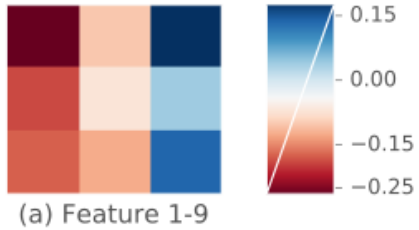


Figure : Filters of the first CNN layer trained for genre classification

Layer 1 : onsets.

Layer 2 : onsets, bass, harmonics, melody.

Layer 3 : onsets, melody, kick, percussion.

Layer 4 : harmonic structures, notes, vertical-horizontal lines.

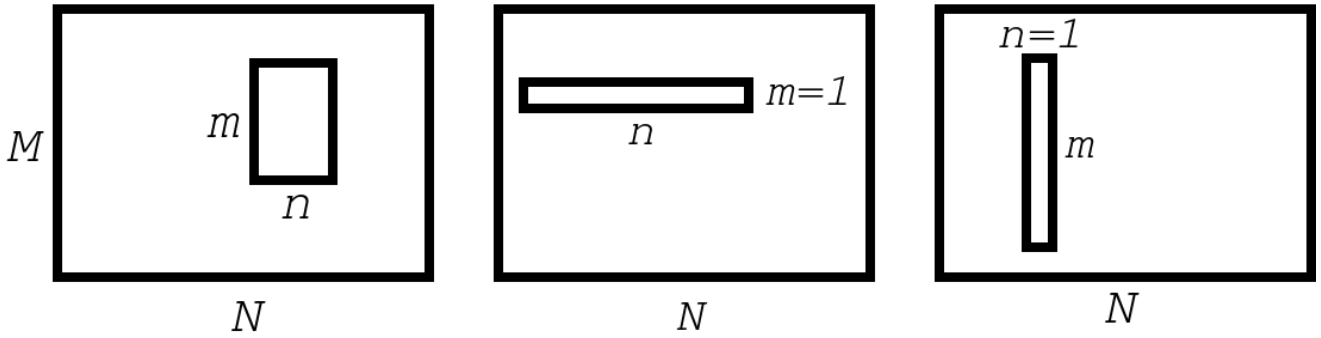
Layer 5 : textures, harmo-rhythmic patterns.

3x3 filters are **limiting** the representational power of the 1st layer!

Does it make sense then to use **computer vision** architectures?

as in: *Hershey et al. CNN architectures for large-scale audio classification. ICASSP'17*

Pons et al. Experimenting with musically motivated CNNs. CBMI'16



Squared/rectangular filters (m-by-n):

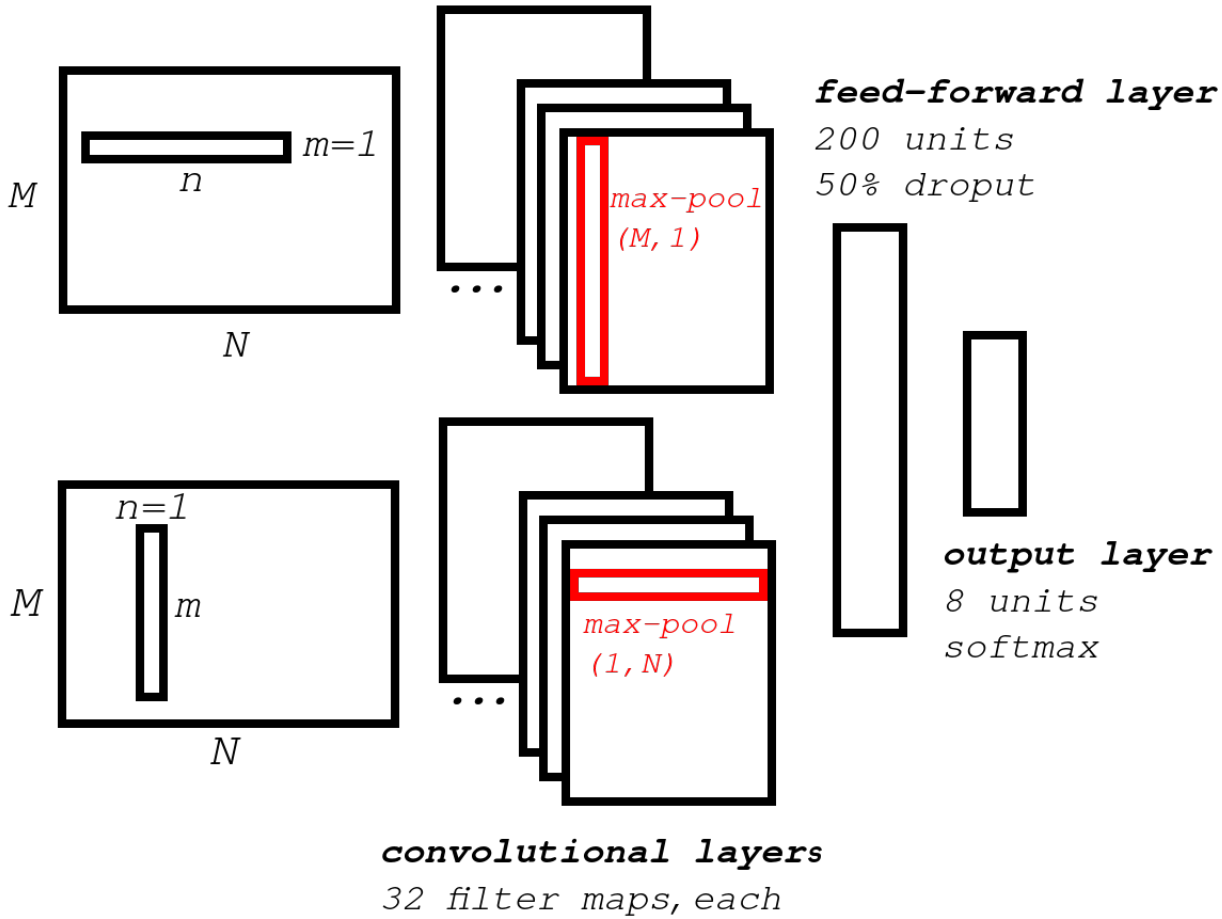
- kick, notes: $m \ll M$ and $n \ll N$

Temporal filters (1-by-n):

- onsets, patterns. *...very efficient!*

Frequency filters (m-by-1):

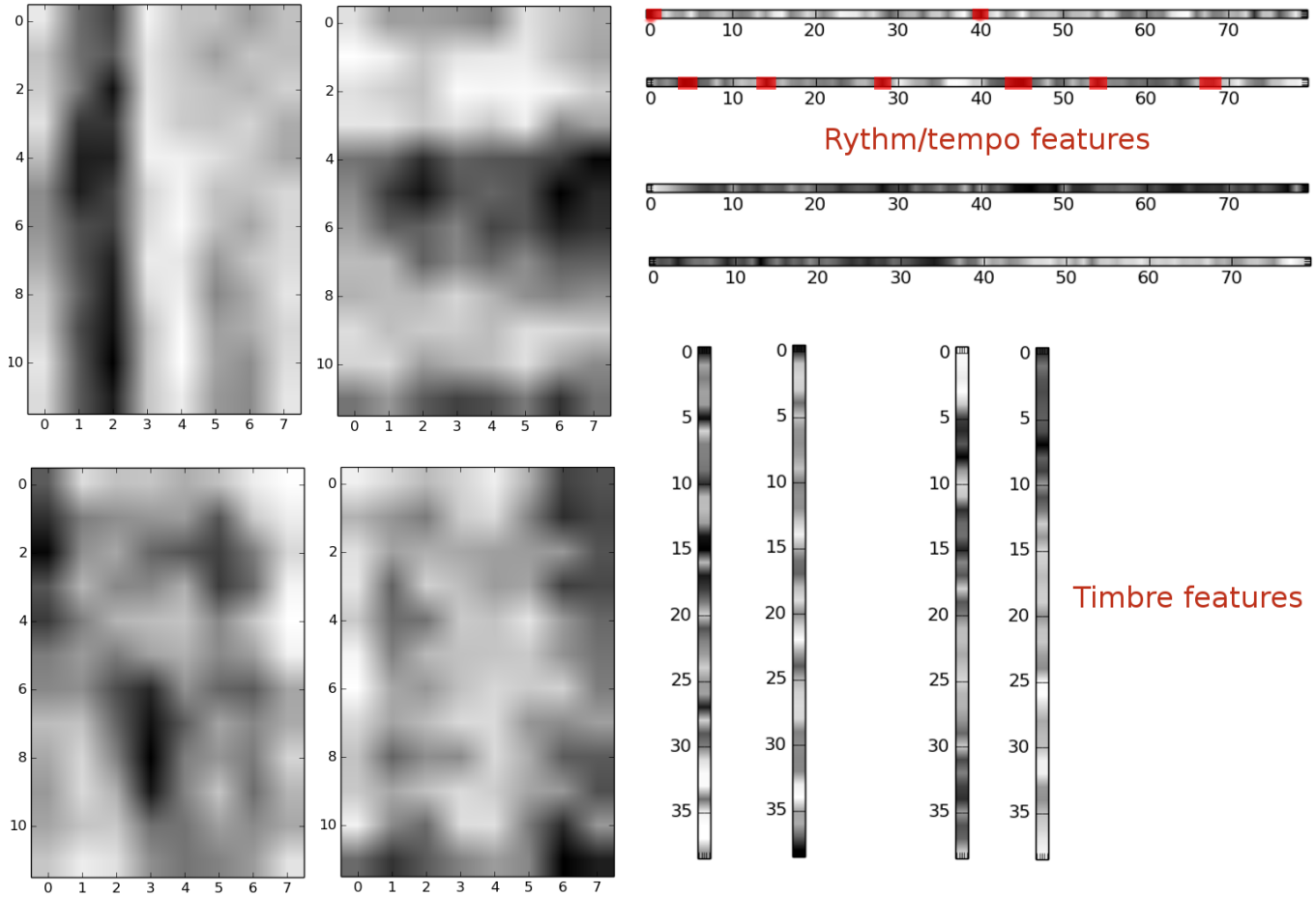
- timbre, chords. *...interpretable!*



Pons et al. Experimenting with musically motivated CNNs. CBMI'16

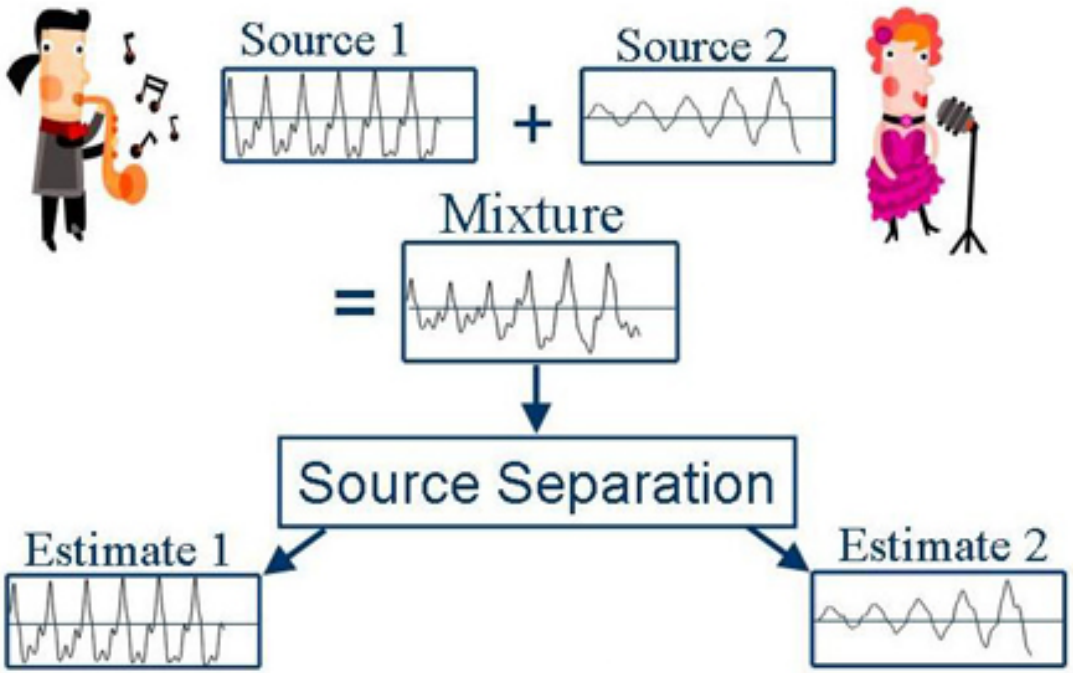
Pons & Serra. Designing efficient architectures for modeling temporal features with CNNs. ICASSP'17

in collaboration with Thomas Lidy: CNNs (12x8, 1x80, 40x1)

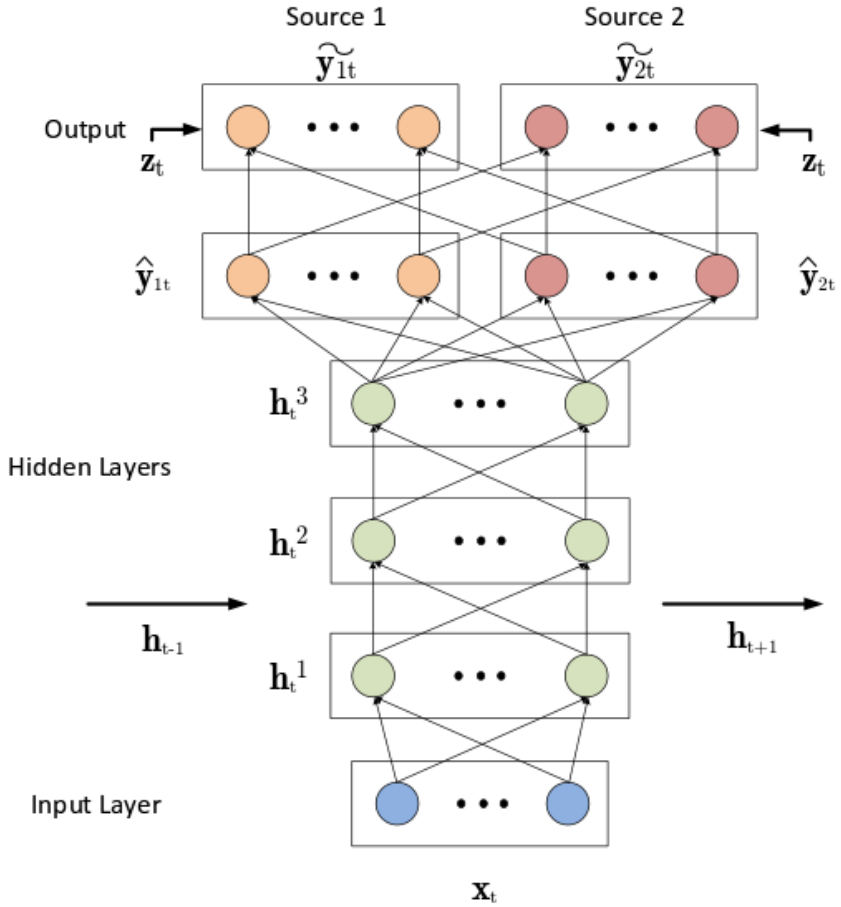


white > black

Source Separation

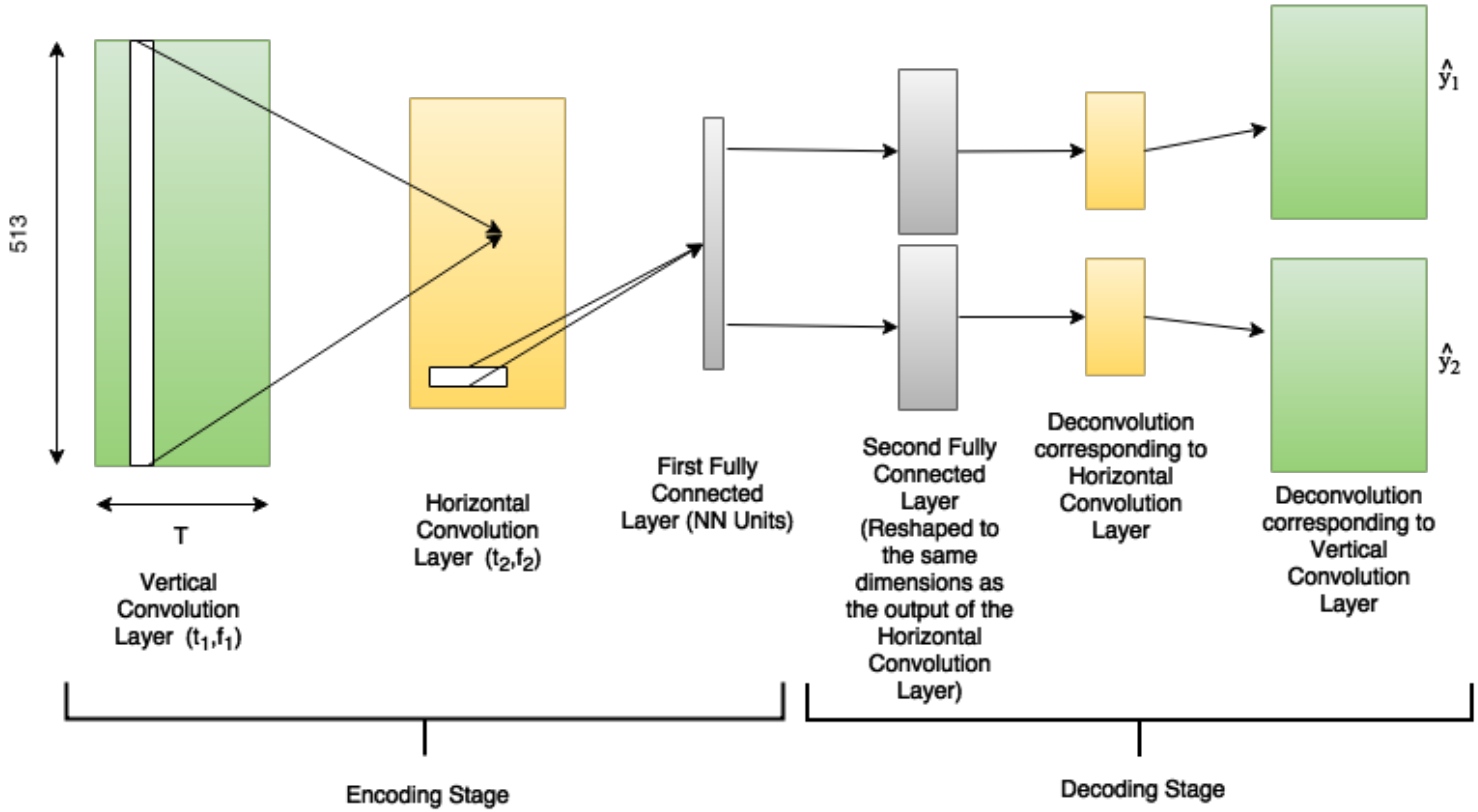


Po-Sen Huang et al. **Singing-Voice Separation** *from Monaural Recordings using Deep Recurrent Neural Networks ISMIR'14*



3 deep layers (*2nd recurrent*) estimating 2 sources simultaneously.
Joint modelling of **DRNN + mask** *with a discriminative cost.*

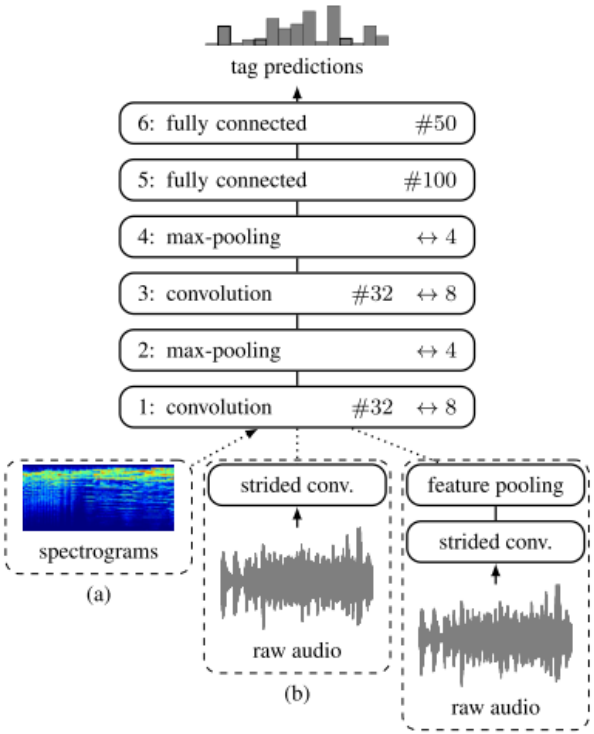
Chandna et al. Monoaural audio source separation using deep convolutional neural networks. LVA-ICA'17

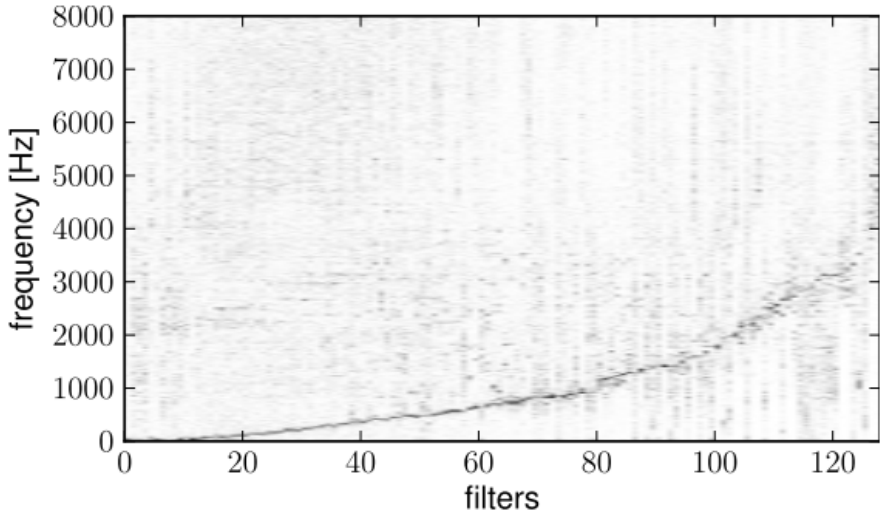


Presented to Signal Separation Evaluation Campaign 2017.

End-to-end learning

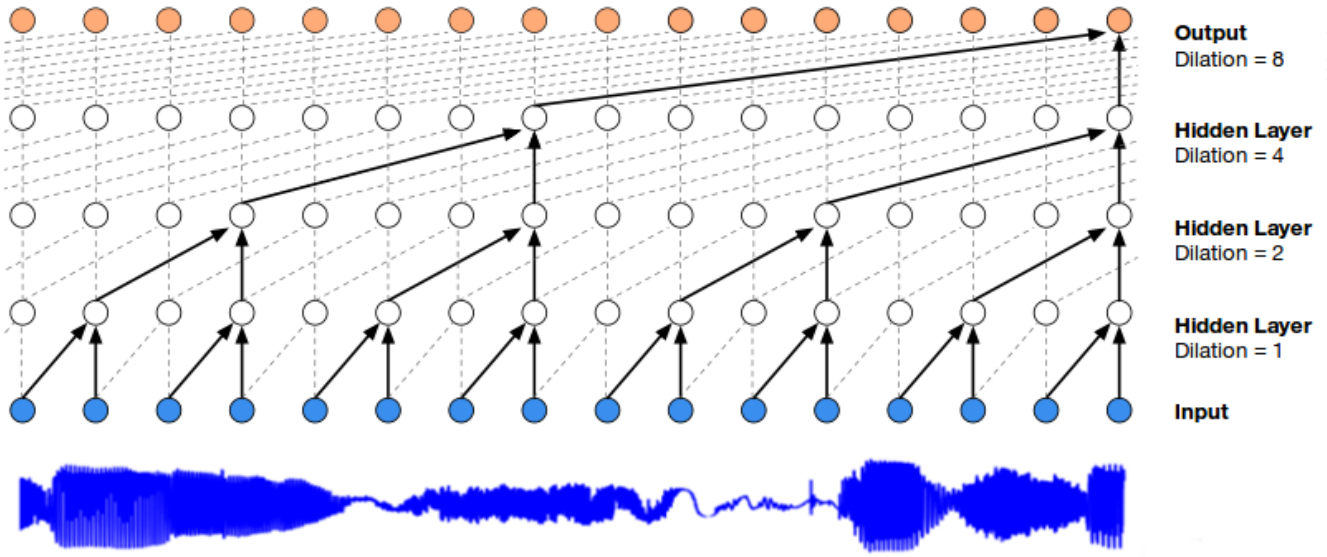
S. Dieleman and B. Schrauwen. End-to-end learning for music audio. ICASSP'14





Learning frequency selective filters similar to MEL filter bank.

Aäron van den Oord et al. **Wavenet**: *A generative model for raw audio. arXiv:1609.03499 (2016)*



Generative model for speech and music audio.

Chronology: the big picture

Some papers as examples for discussion

Current trends and research directions

Limitations the academic music technology community is facing when approaching their problems with deep learning:

- Lack of annotated data.
- Lack of hardware (GPUs) $\rightarrow$ Expertise goes to the industry.

Limitations the academic music technology community is facing when approaching their problems with deep learning:

- Lack of annotated data.
- Lack of hardware (GPUs) $\rightarrow$ Expertise goes to the industry.

Trends for solving the issue of annotated data:

- Collaborative effort for **jointly annotating music data**.
- Artificial **augmentation** of the annotated **data**.

Limitations the academic music technology community is facing when approaching their problems with deep learning:

- Lack of annotated data.
- Lack of hardware (GPUs) $\rightarrow$ Expertise goes to the industry.

Trends for solving the issue of annotated data:

- Collaborative effort for **jointly annotating music data**.
- Artificial **augmentation** of the annotated **data**.

Trends for solving hardware limitations:

- Researchers **avoid end-to-end learning** approaches:
  - Inputting hand-crafted features to deep networks.
  - Using non deep learning classifiers/models stacked on top of deep learning feature extractors.
- Constraining the solution space **considering prior information**: music nature or human audio perception.

Limitations the academic music technology community is facing when approaching their problems with deep learning:

- Lack of annotated data.
- Lack of hardware (GPUs) $\rightarrow$ Expertise goes to the industry.

Trends for solving the issue of annotated data:

- Collaborative effort for **jointly annotating music data**.
- Artificial **augmentation** of the annotated **data**.

Trends for solving hardware limitations:

- Researchers **avoid end-to-end learning** approaches:
  - Inputting hand-crafted features to deep networks.
  - Using non deep learning classifiers/models stacked on top of deep learning feature extractors.
- Constraining the solution space **considering prior information**: music nature or human audio perception.

References @ jordipons.me/lack-of-annotated-music-data-restrict-the-solution-space/

Imaginable **research directions**?

- End-to-end learning from raw audio.

  Aytar et al. **SoundNet***: Learning Sound Representations from Unlabeled Video.*

  @ NIPS'16

- Multimodal deep processing.

  Slizovskaia et al. Automatic musical instrument recognition in audiovisual recordings by **combining image and audio** *classification strategies.*

  @ SMC'16

- Unsupervised learning such as generative models.

  Aaron van den Oord et al. **Wavenet***: A generative model for raw audio.*

  @ arXiv:1609.03499 (2016)

- Efficient learning long-term dependencies.

  Eck and Schmidhuber. Learning The Long-Term Structure of the Blues.

  @ICANN02

- Understanding which features are learnt.

  Pons et al. Experimenting with **musically motivated convolutional NNs***.*

  @ CBMI'16

Thanks! :)

Deep learning for music data processing

A personal (re)view of the state-of-the-art

Jordi Pons

www.jordipons.me

Music Technology Group, DTIC,
Universitat Pompeu Fabra, Barcelona.

31st January 2017