Nonnegative Tensor Factorization for Source Separation of Loops in Audio

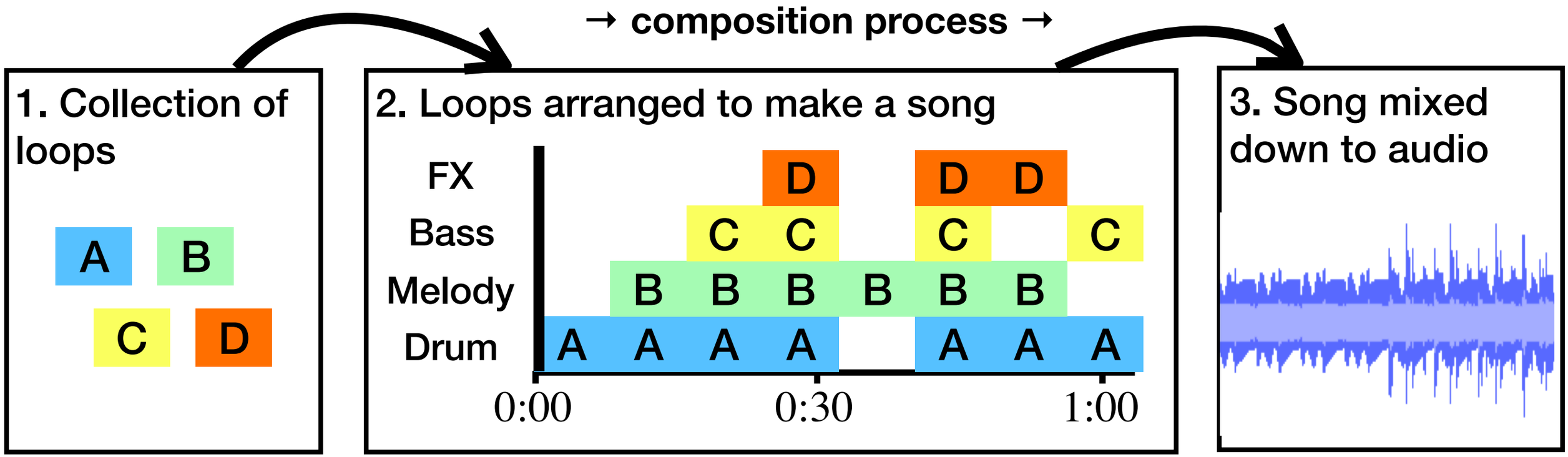
Jordan B. L. Smith
National Institute of Advanced Industrial Science and Technology (AIST), Japan

Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST), Japan

Introduction

Extracting loops from music

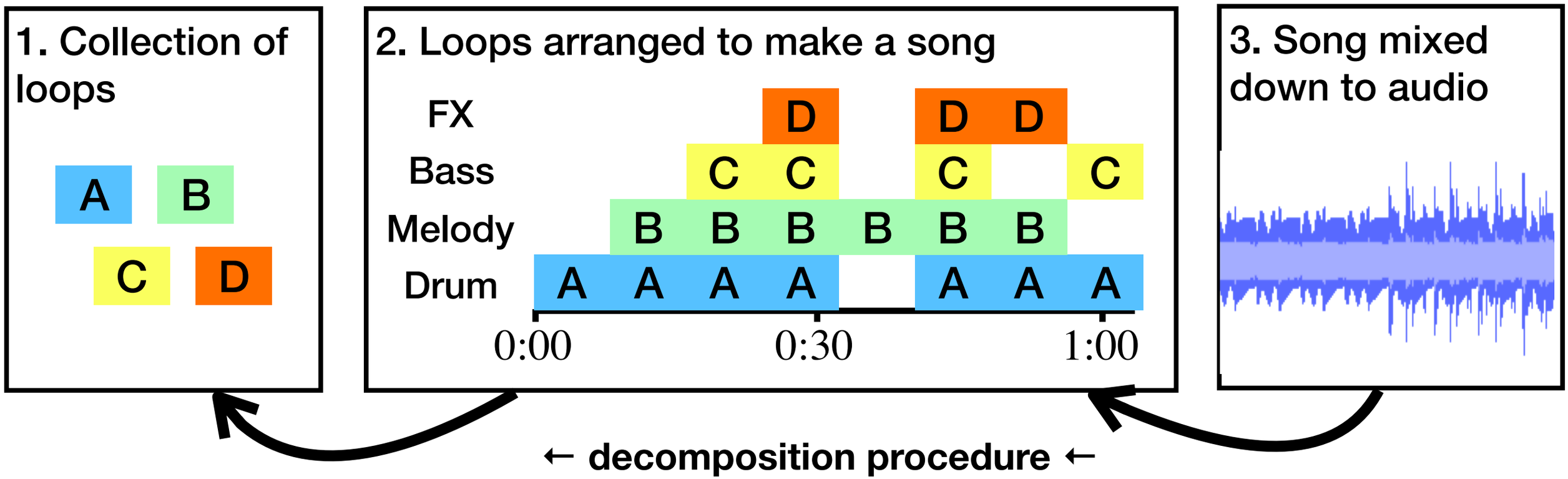
- In some musical styles, songs are built from loops. E.g.:



Audio examples (and test data) all borrowed from [López-Serrano et al. 2016]

Extracting loops from music

- In some musical styles, songs are built from loops. E.g.:

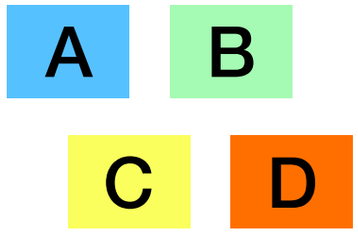


- Goal: decompose the audio signal to recover:
  - the layout of the song
  - the source-separated loops

Extracting loops from music

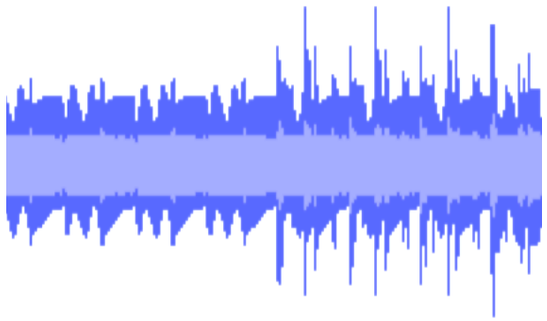
- Two previous approaches that inspired us:
  - Fingerprint-based loop detection [López-Serrano et al. 2016]

Inputs:



Original loops

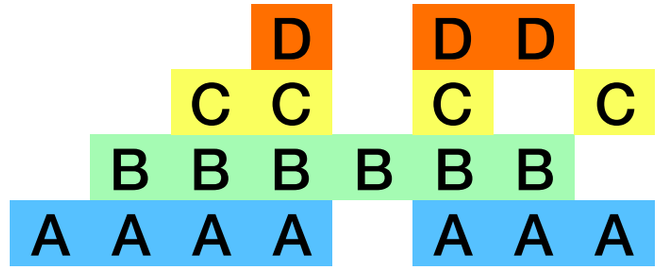
+



Mixed audio

→

Output:



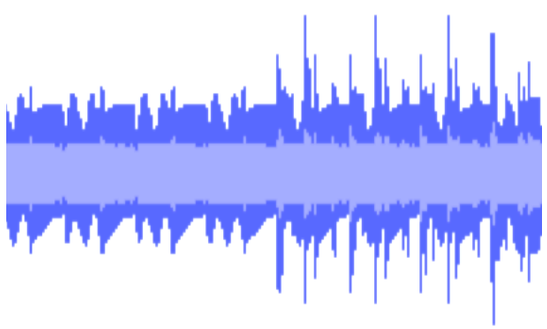
Map of loop activations

- Iterative NMF [Seetharaman & Pardo 2016]

Assumption that loops are introduced additively

Inputs:

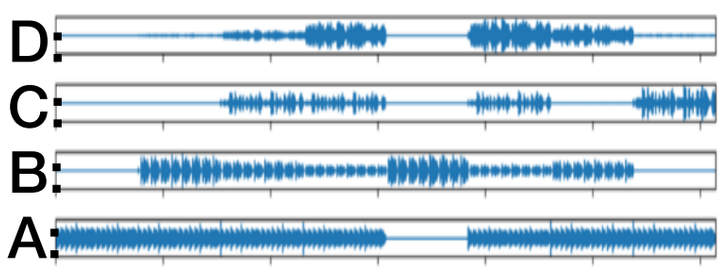
+



Mixed audio

→

Output:

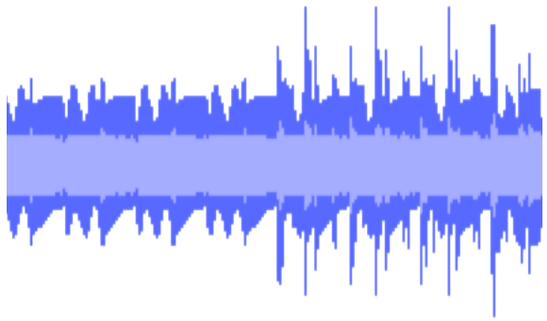


Separated tracks,
one per loop

Extracting loops from music

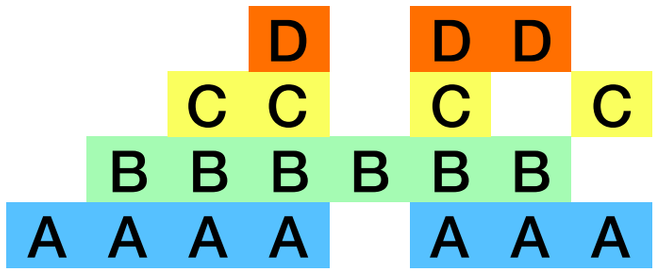
- Our proposed system:

Input:



Mixed audio

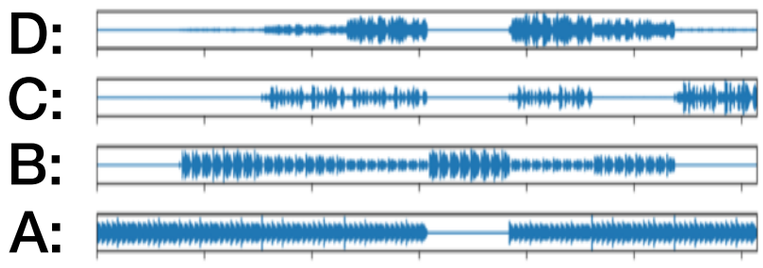
→



Map of loop activations

Outputs:

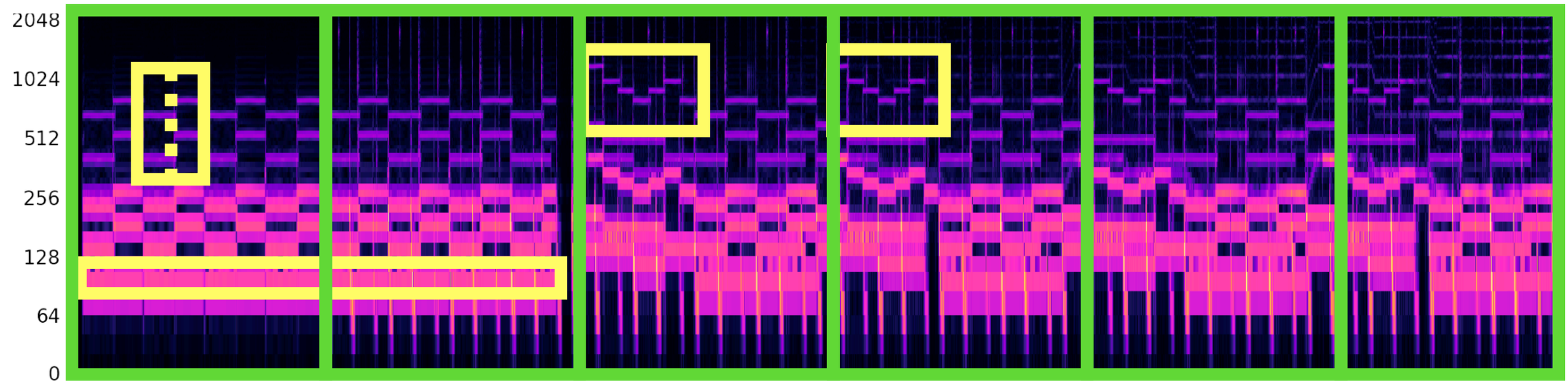
+



Separated tracks, one per loop

- We attempt to solve both problems in one step, without assumption of additive layout
- We do so by extending nonnegative matrix factorization (NMF) to handle periodicity

Source separation using NMF*



NMF can handle many types of repetition:

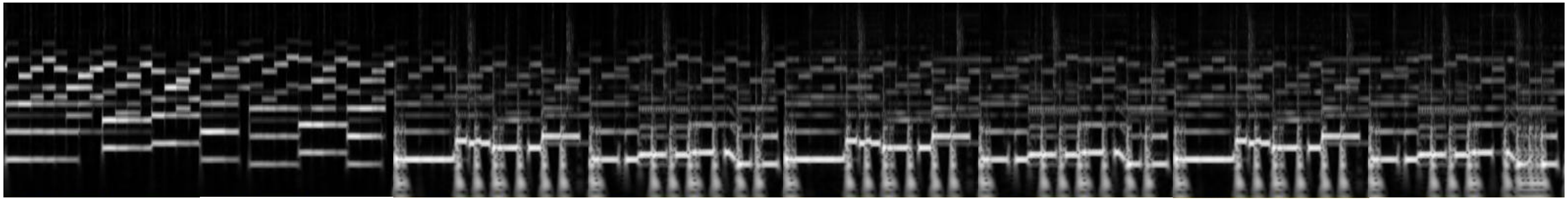
- Steady-state notes
- Note sequences repeated in time
- Transposed notes
- **Periodicity (especially at downbeats)**

- NMF with harmonic templates
- NMFD with time-evolving templates
  [Smaragdis 2004]
- NMF2D with transposed harmonic templates
  [e.g., FitzGerald, Cranitch & Coyle 2008]
- **...no nonnegative approach! NB: REPET, a median-filtering approach**
  [Rafii, Liutkus, & Pardo 2014]

Method

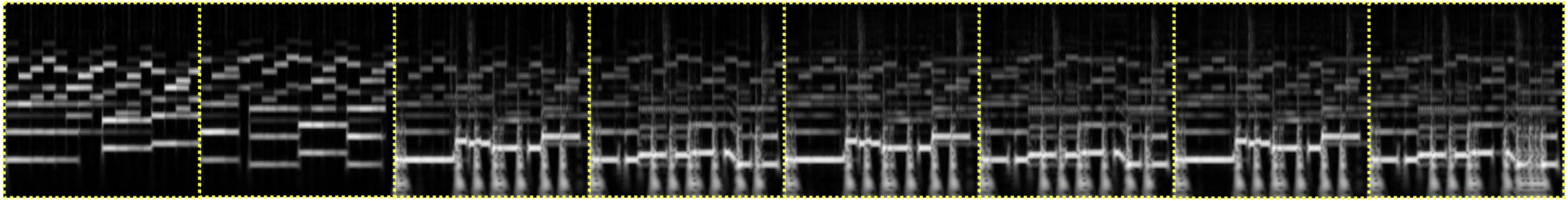
Nonnegative tensor factorization

- Step 1: estimate downbeats [madmom, Böck et al. 2016]



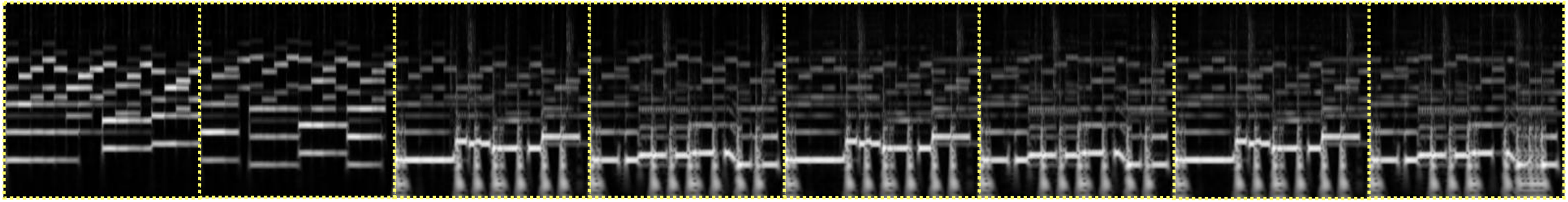
Nonnegative tensor factorization

- Step 1: estimate downbeats [madmom, Böck et al. 2016]



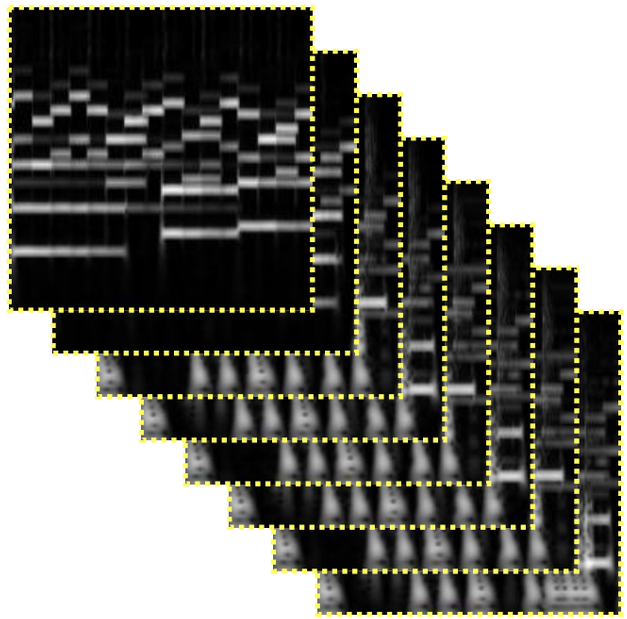
Nonnegative tensor factorization

- Step 1: estimate downbeats
- Step 2: stack the 2D spectrograms into a 3D volume (a “spectral cube”)



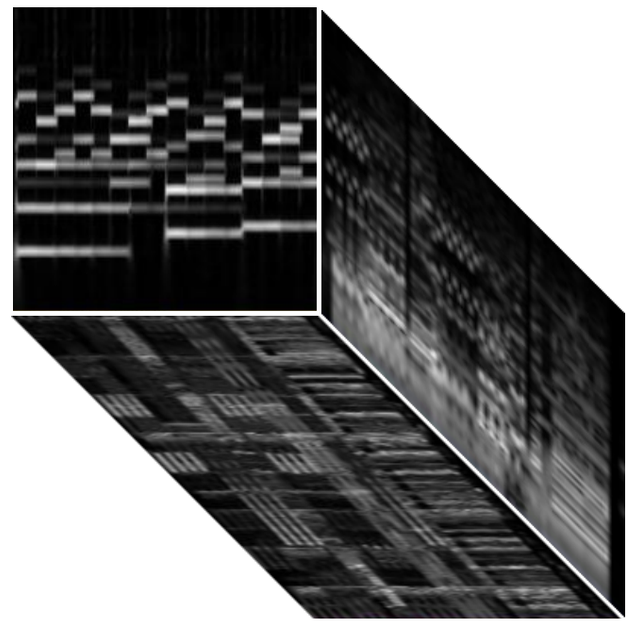
Nonnegative tensor factorization

- Step 1: estimate downbeats
- Step 2: stack the 2D spectrograms into a 3D volume (a “spectral cube”)

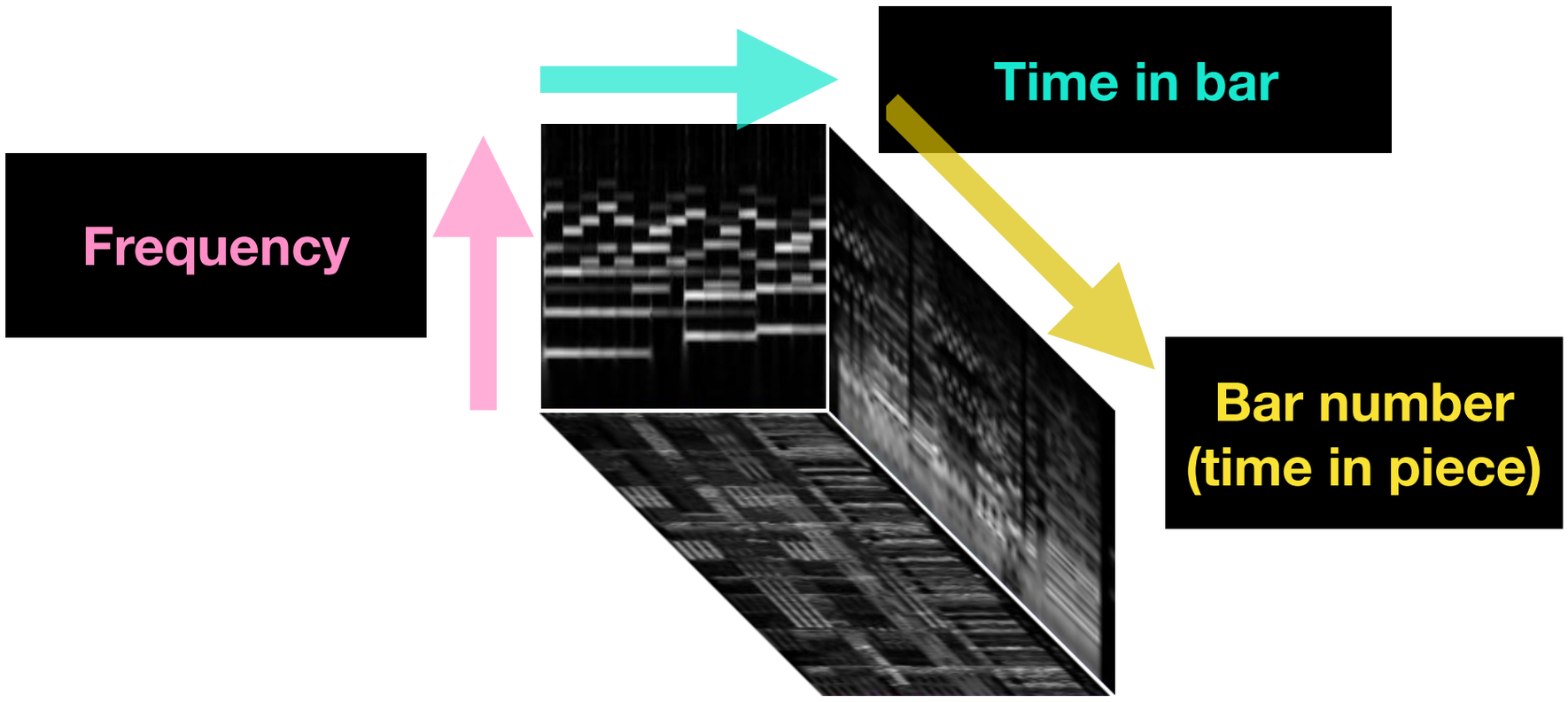


Nonnegative tensor factorization

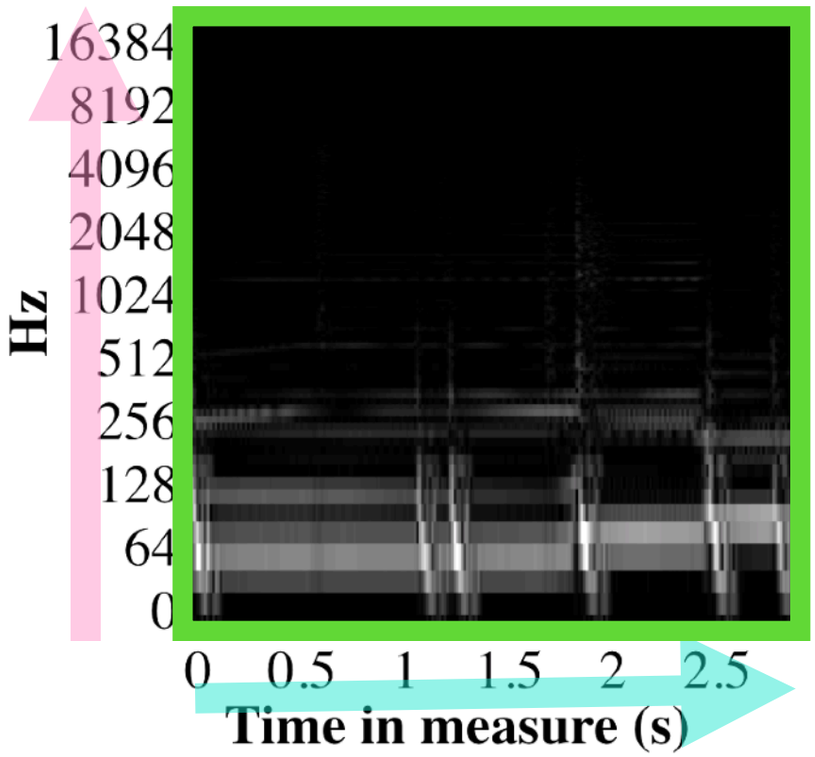
- Step 1: estimate downbeats
- Step 2: stack the 2D spectrograms into a 3D volume (a “spectral cube”)

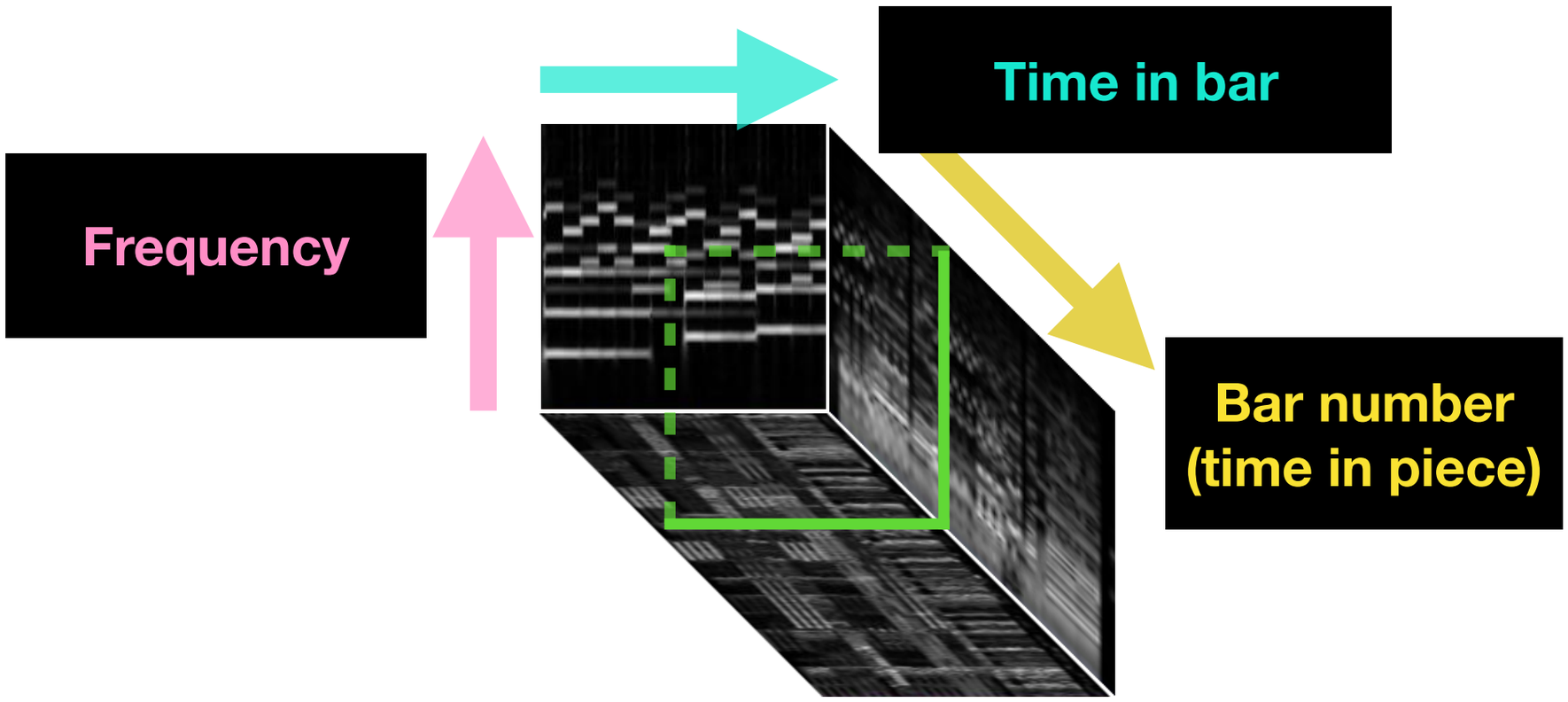


Detour: understanding the spectral cube

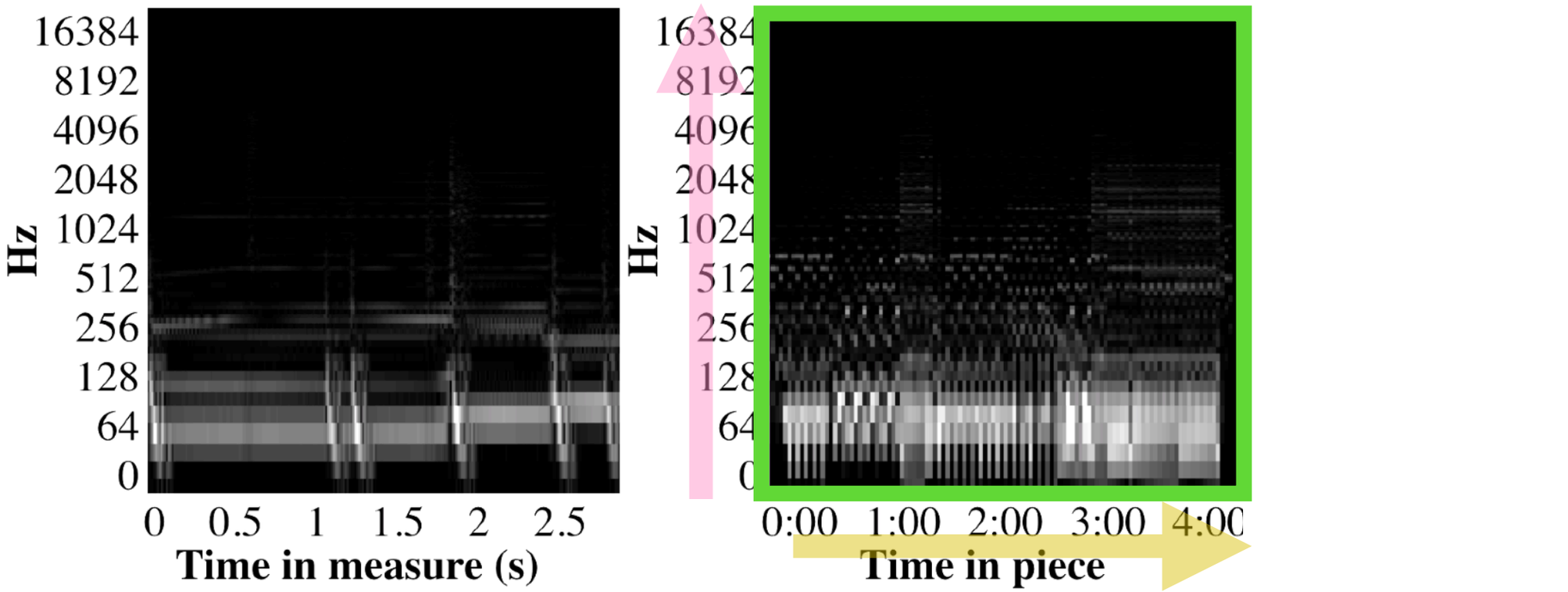


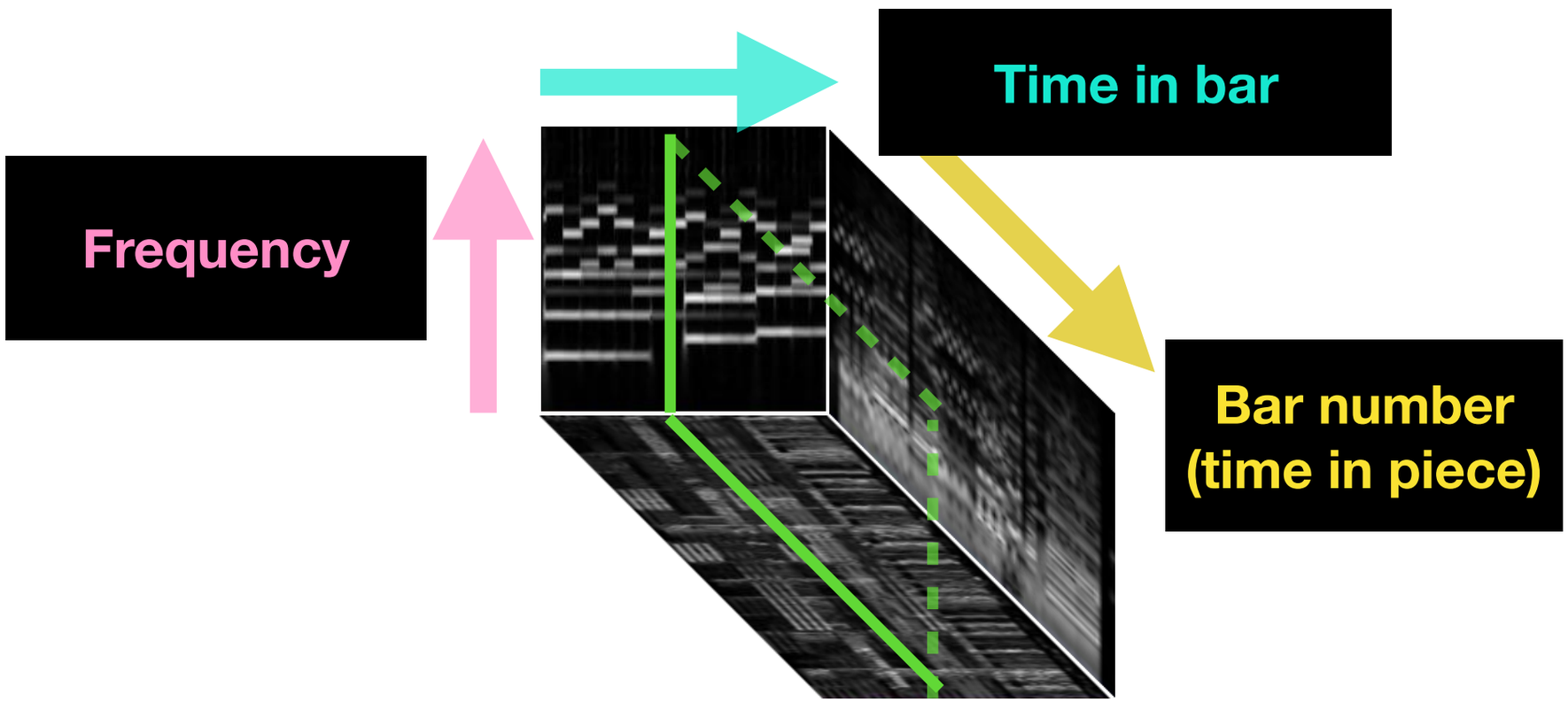
Detour: understanding the spectral cube



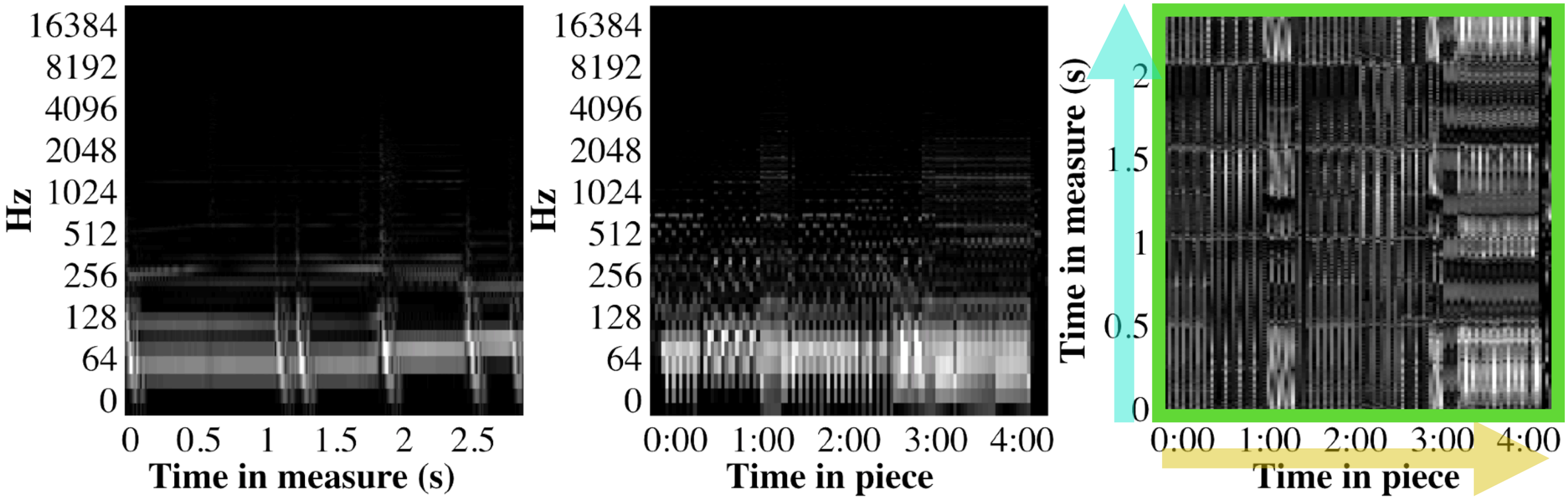


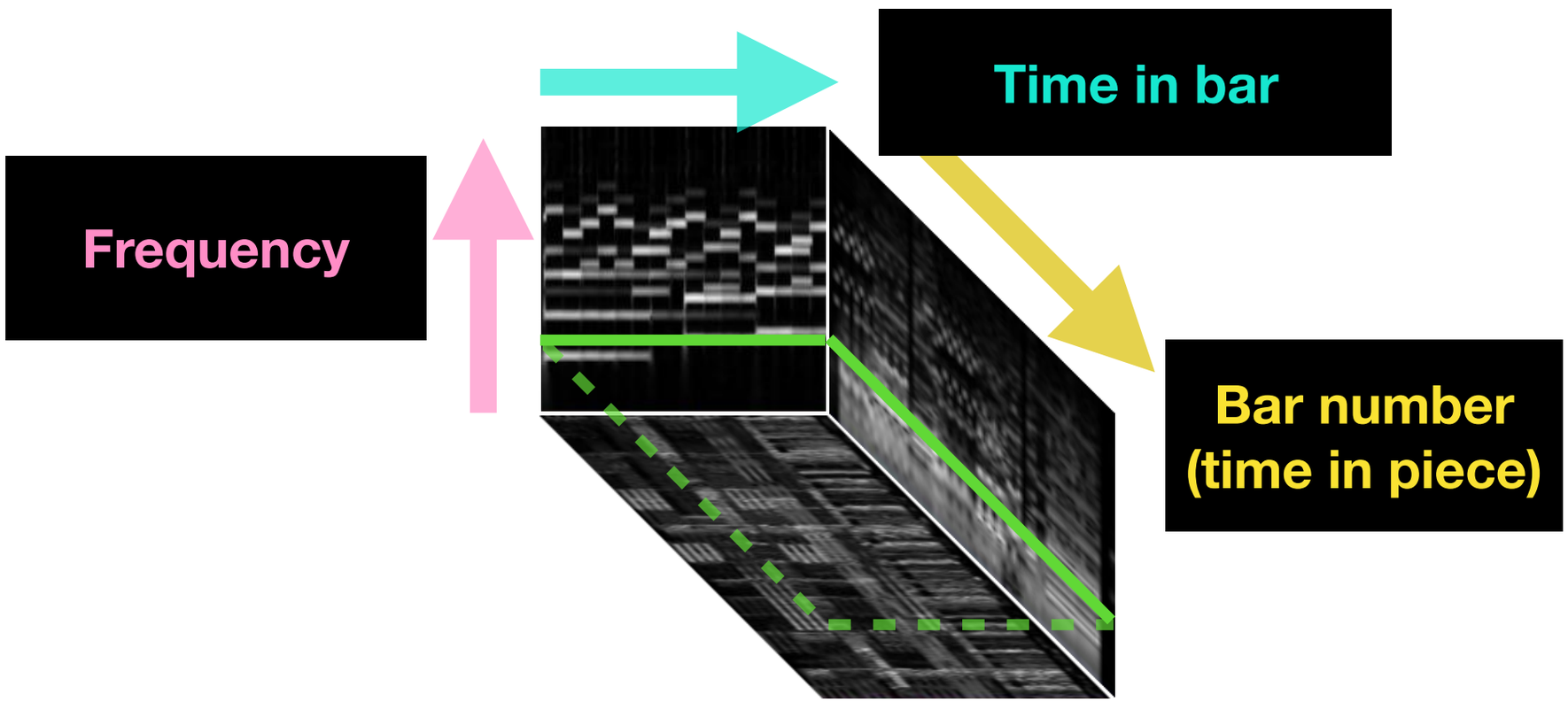
Detour: understanding the spectral cube



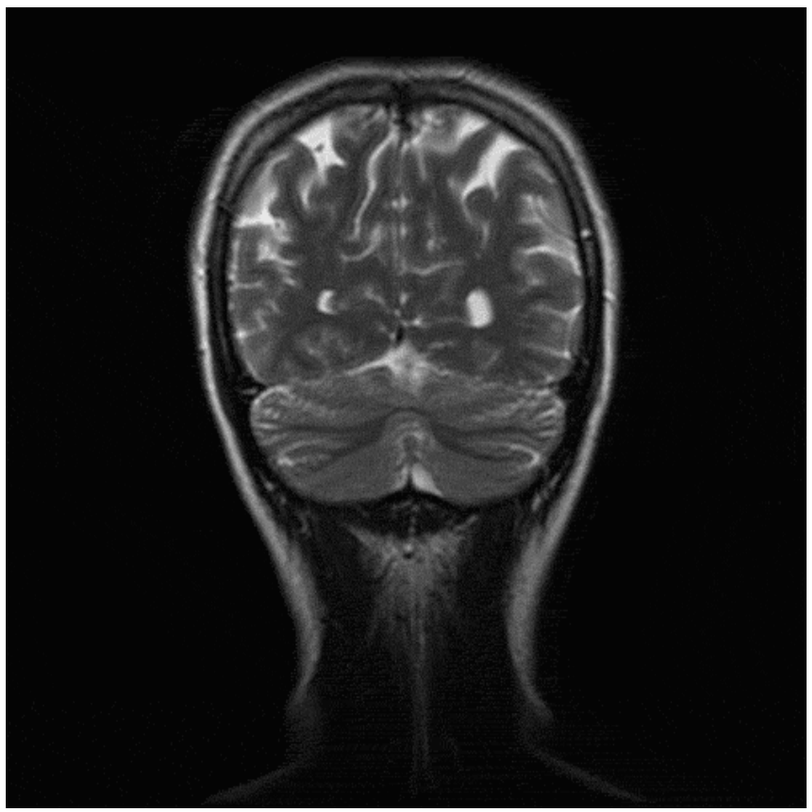


Detour: understanding the spectral cube

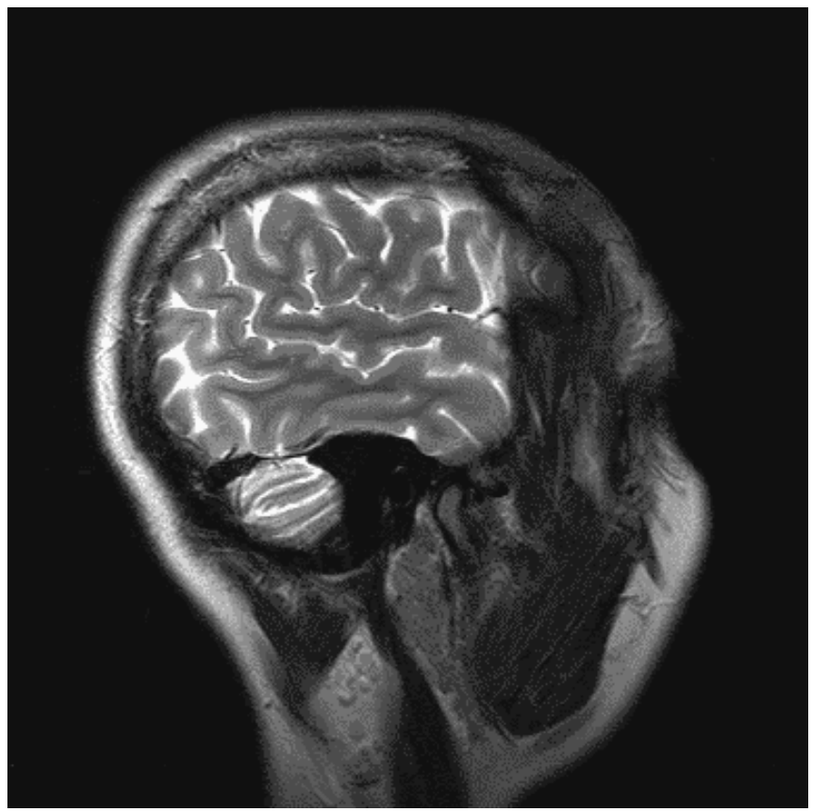




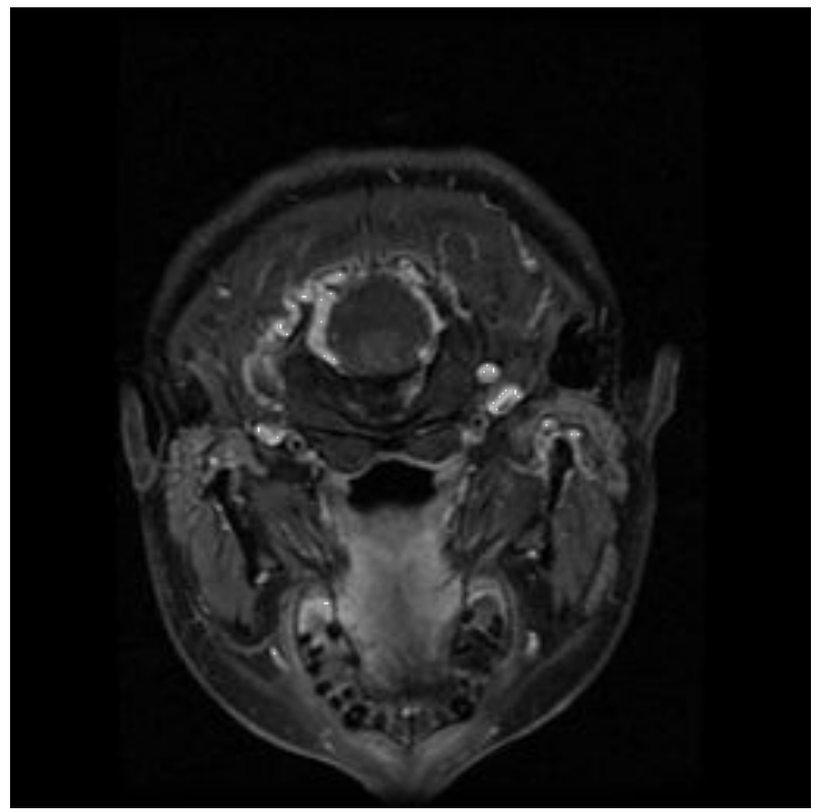
Visualizing a 3D volume: CT scan



Back to front

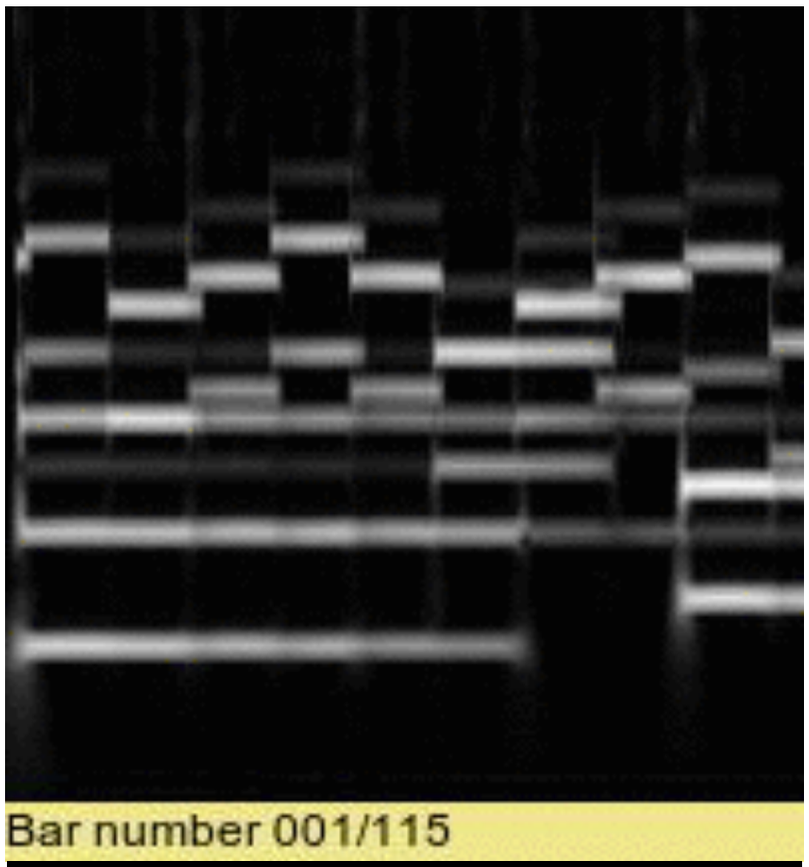


Left to right

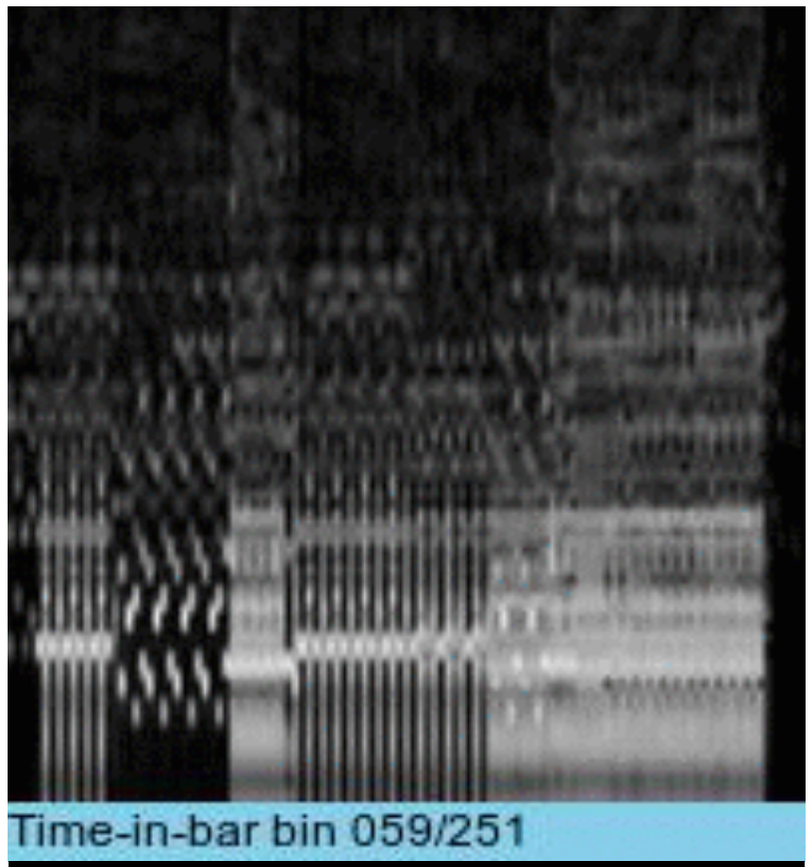


Bottom to top

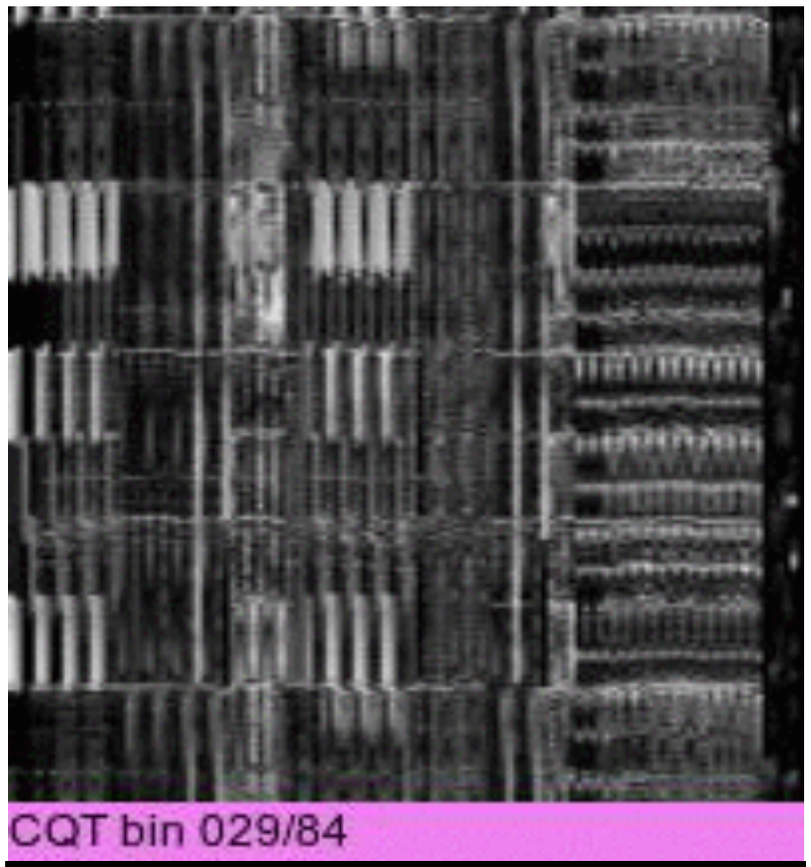
Visualizing a 3D volume: CT scan



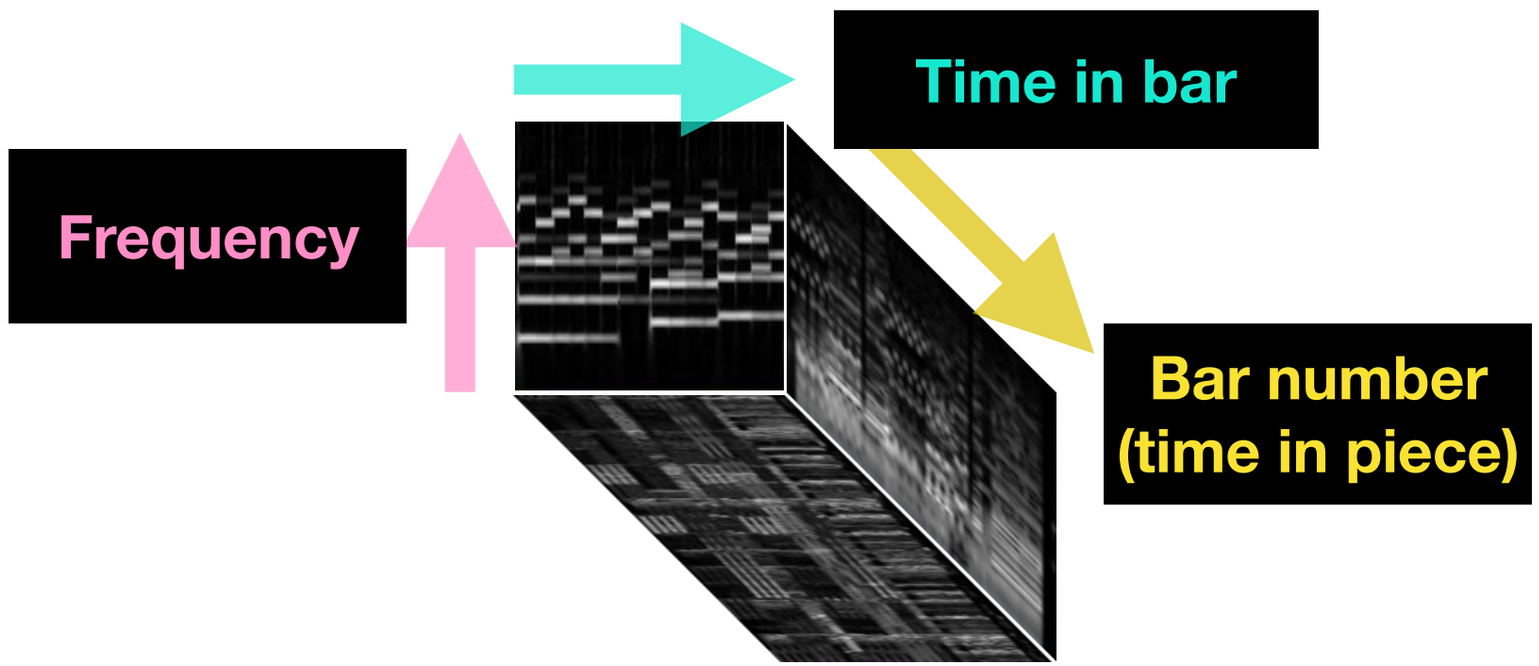
Beginning to end of piece



Beginning to end of a bar



Low frequency to high

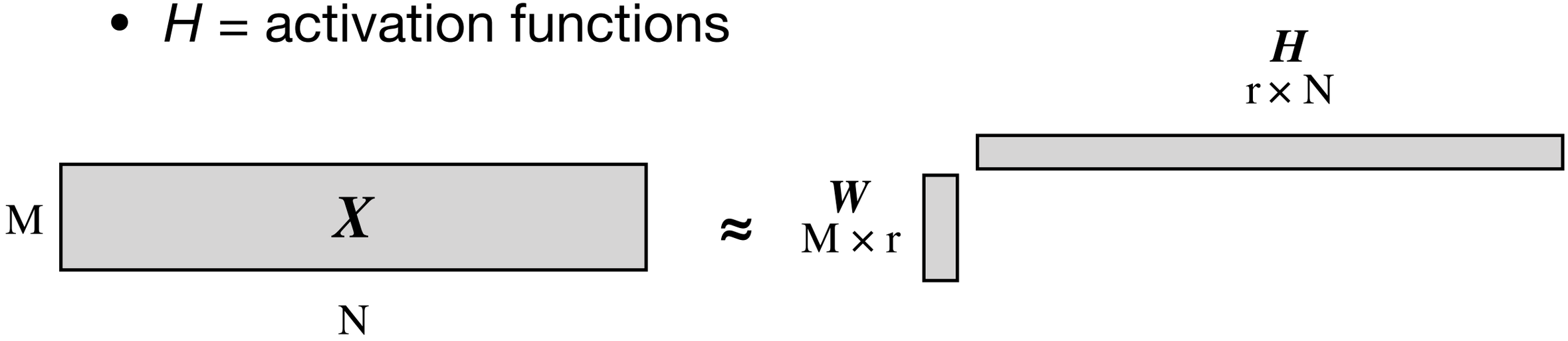


Nonnegative tensor factorization

- Step 1: estimate downbeats
- Step 2: stack the 2D spectrograms into a 3D volume (a “spectral cube”)
- Step 3: use nonnegative tensor factorization (NTF) to model the spectral cube

Nonnegative matrix factorization

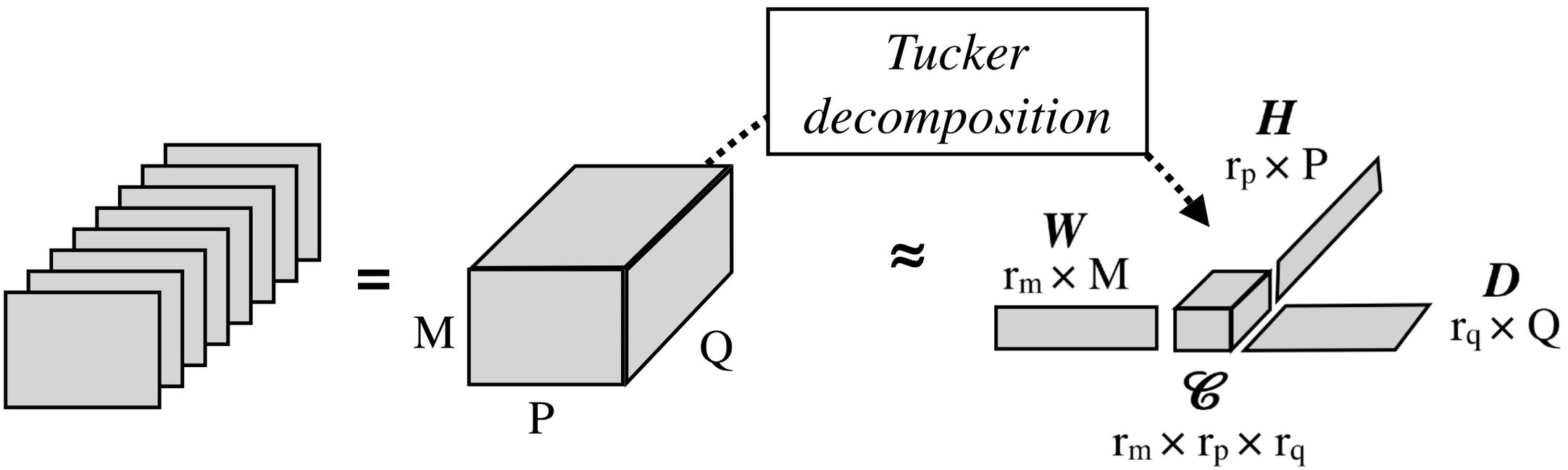
- NMF: $X \approx W \circ H$
  - W = note templates
  - H = activation functions



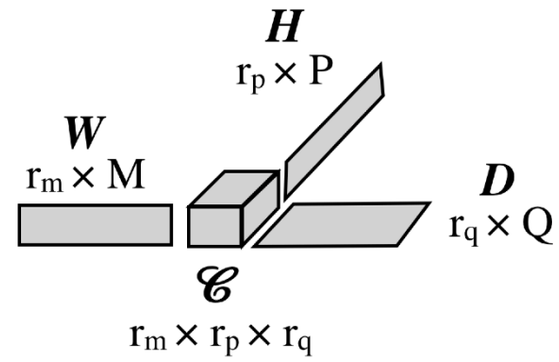
- Needs post-processing to separate sources:
  - which templates in W belong to the same source?
  - different sources could use the same harmonic components!

Nonnegative tensor factorization

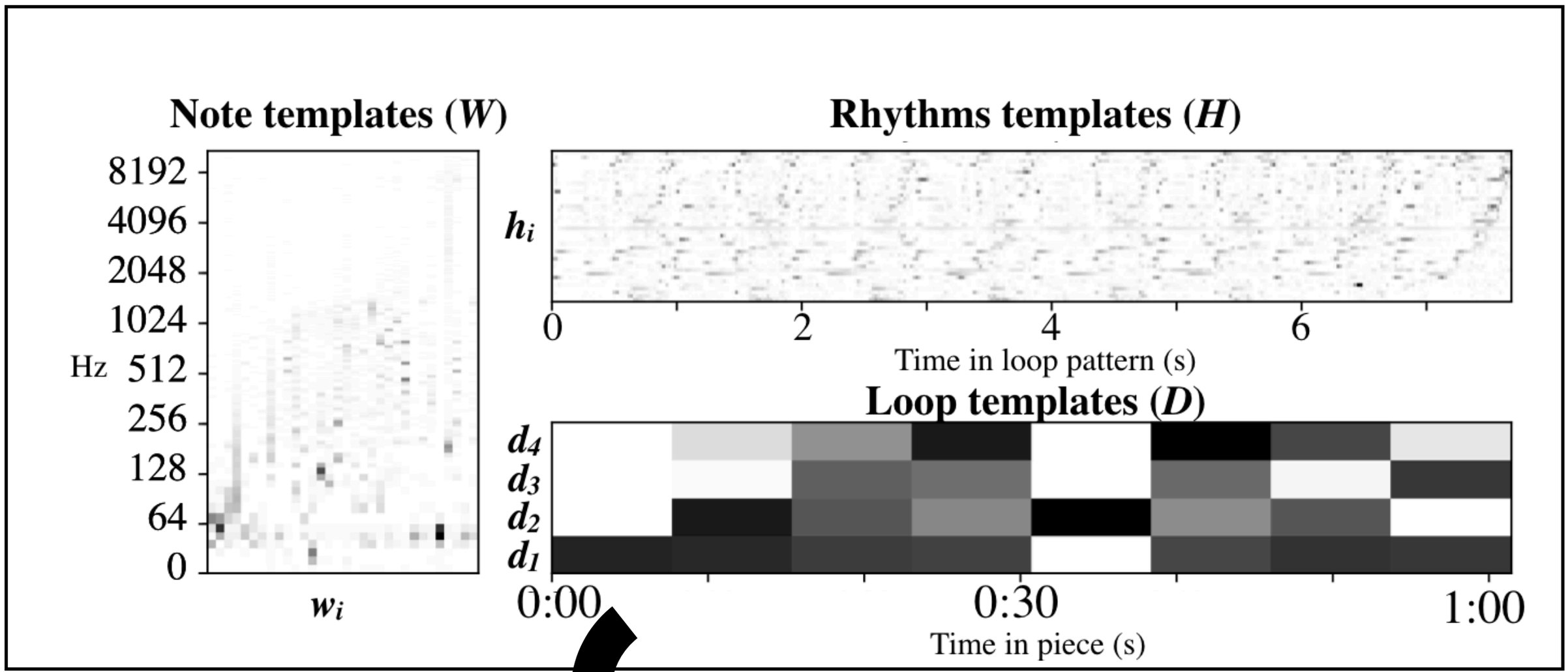
- Tucker Decomposition: $X \approx C \circ (W \circ H \circ D)$
  - W = note templates
  - H = activation functions (time-in-bar)
  - D = loop activation functions (time-in-piece)
  - C = core tensor = recipe for each loop type



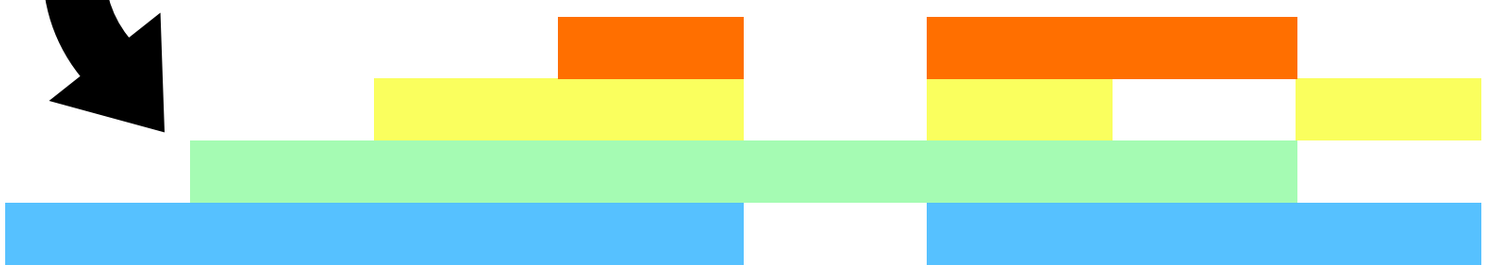
Interpreting the NTF model



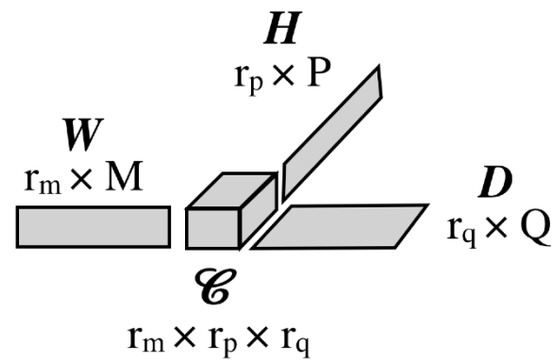
- *W, H*, and *D* all musically intuitive:



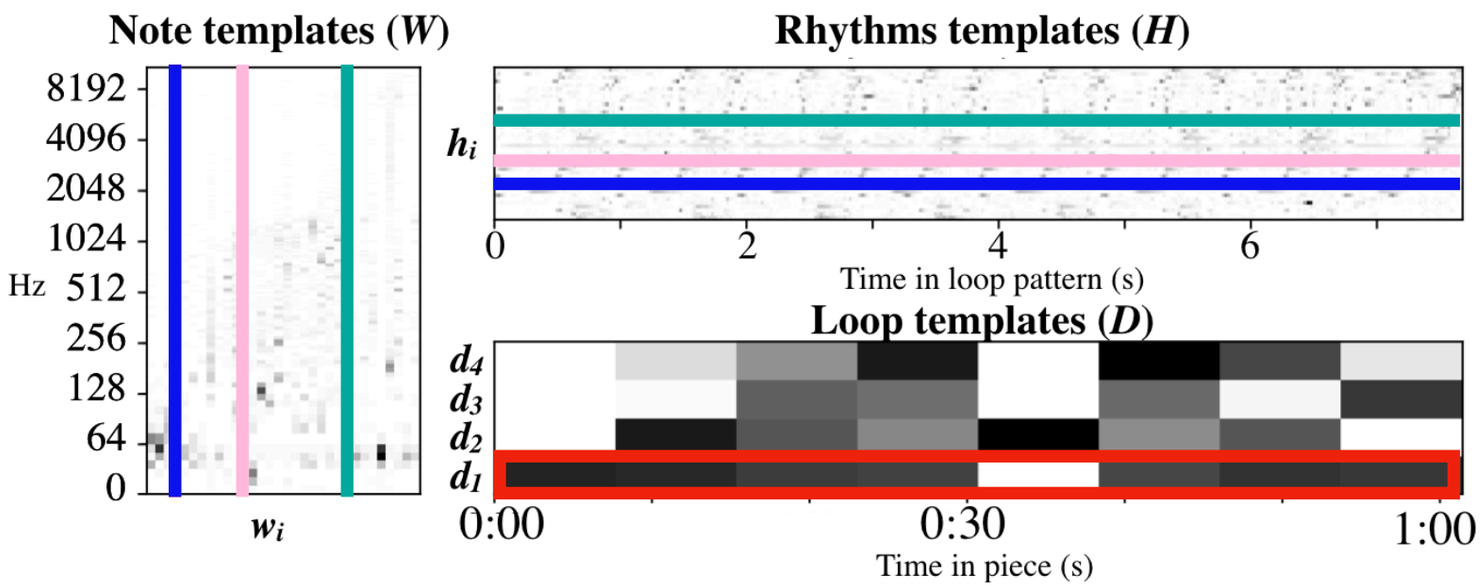
Loop template activations directly estimate layout of song

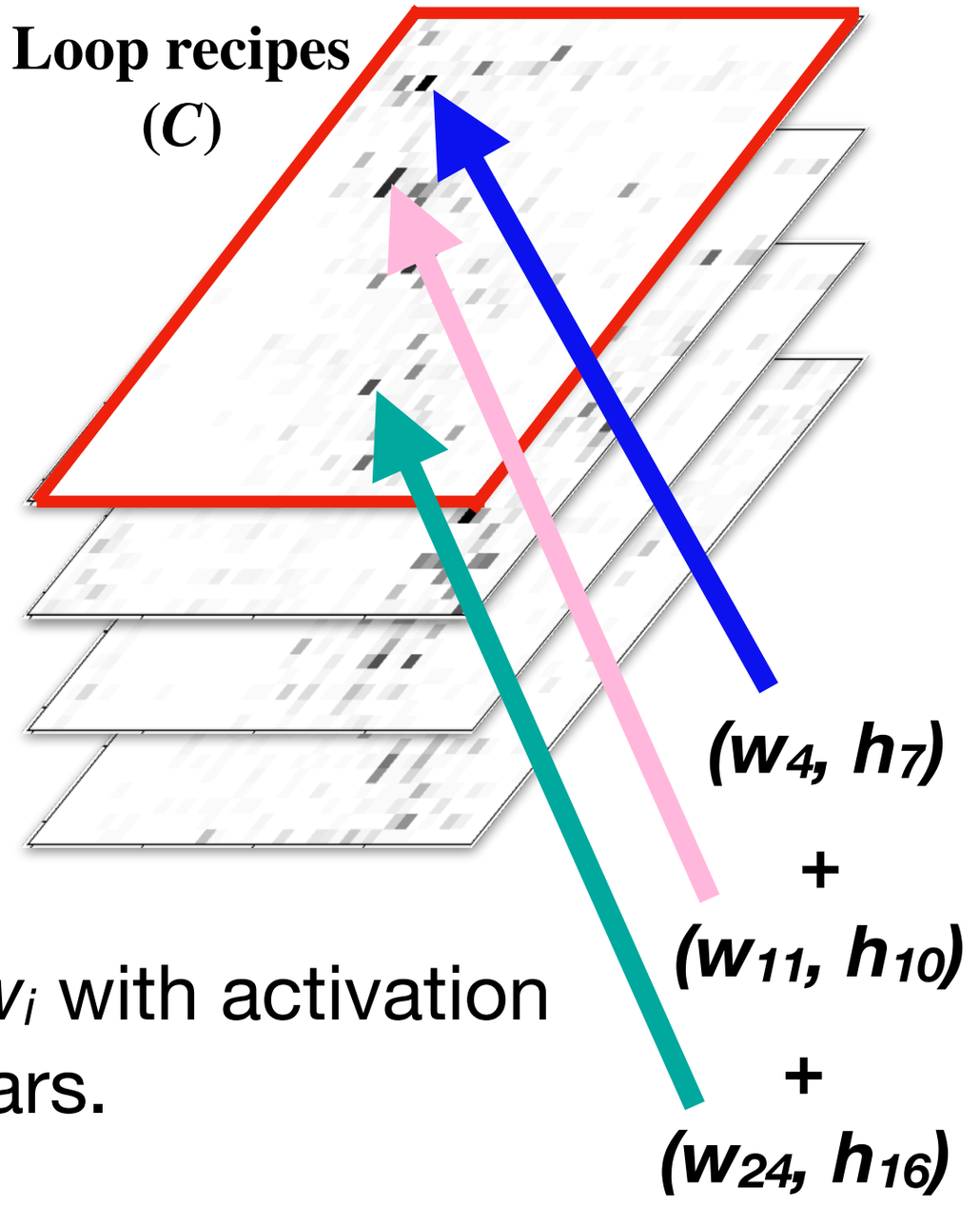


Interpreting the NTF model



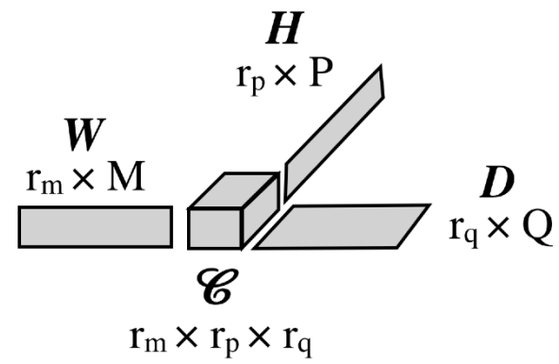
- Core tensor *C* = recipe for each loop type



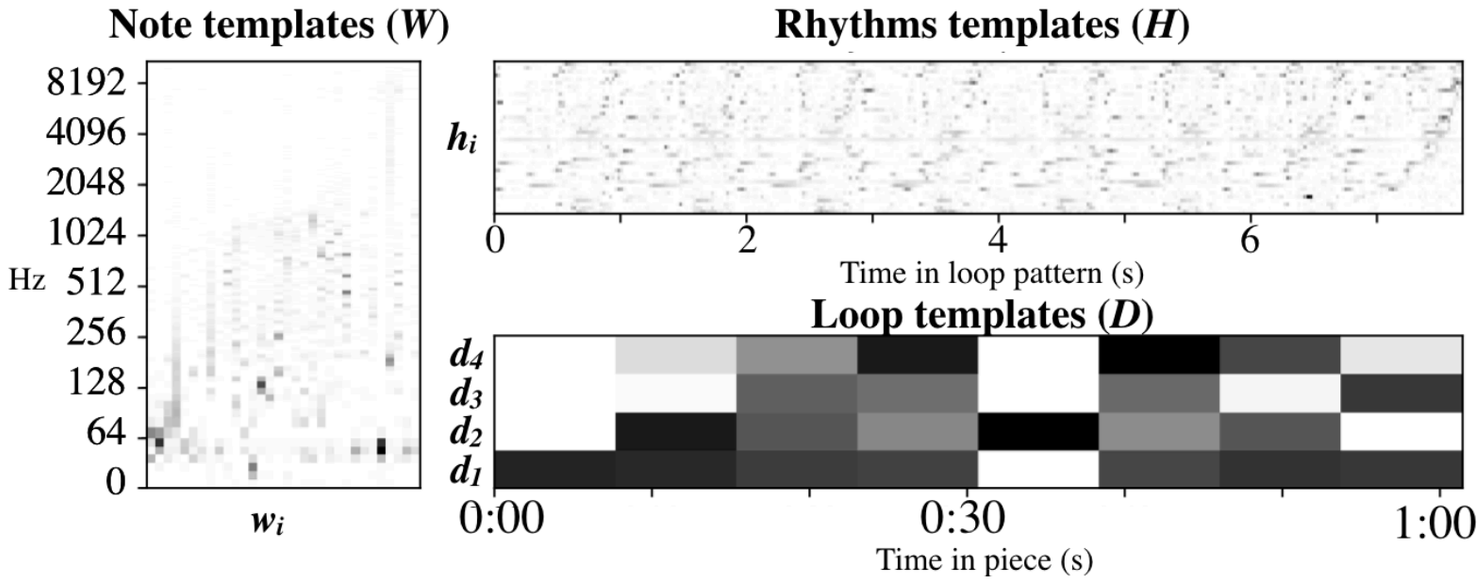


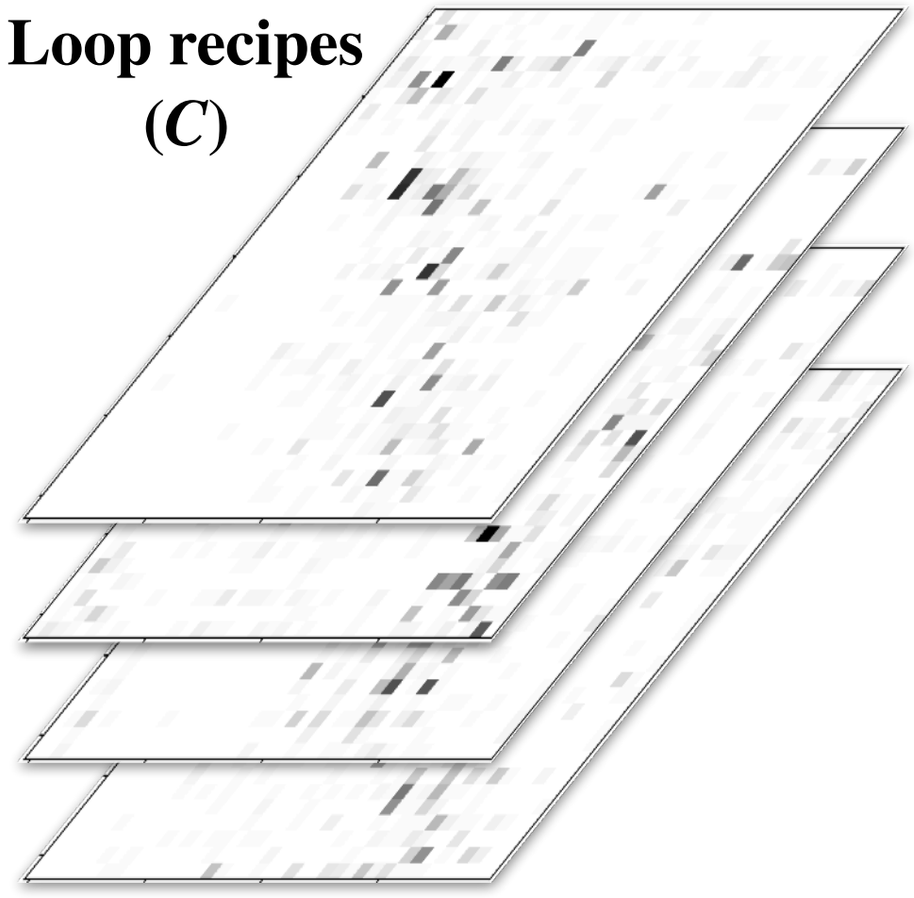
- Pixel *C(i, j, k)* tells us to play note w_i with activation function h_j whenever loop d_k appears.

Interpreting the NTF model



- Core tensor *C* = recipe for each loop type





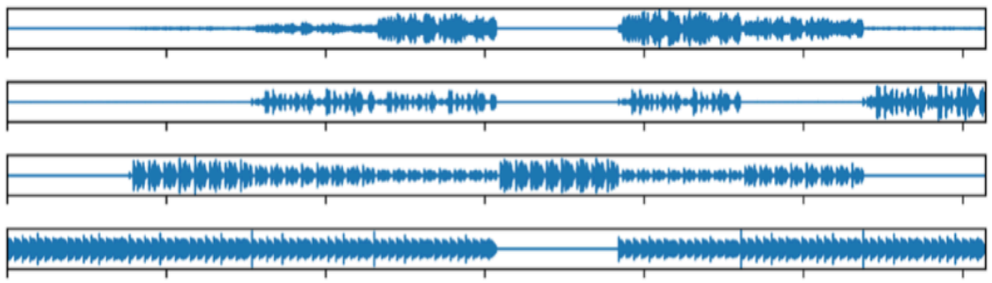
- To recover entire spectrogram: $C \circ (W \circ H \circ D)$
- To recover individual loop source: $C_{[:,:,k]} \circ (W \circ H \circ D_{[k,:]})$

Evaluation

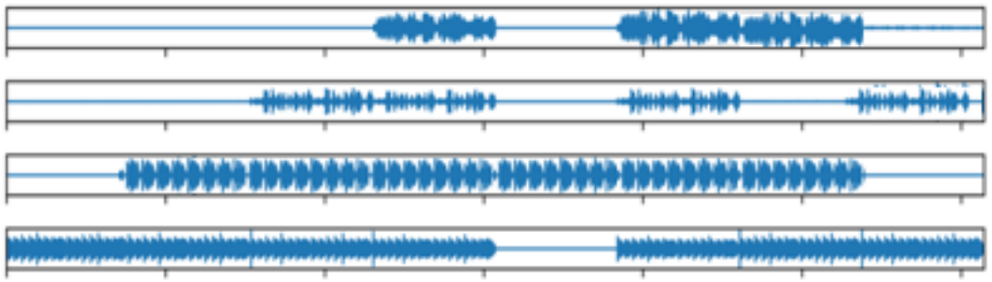
Evaluation

- We used synthetic data [López-Serrano et al. 2016]
  - 7 sets of loops x 3 different layouts (arrangements)
- Algorithm output 1: separated signals
  - Evaluate quality with SDR, SIR, SAR

estimated source tracks

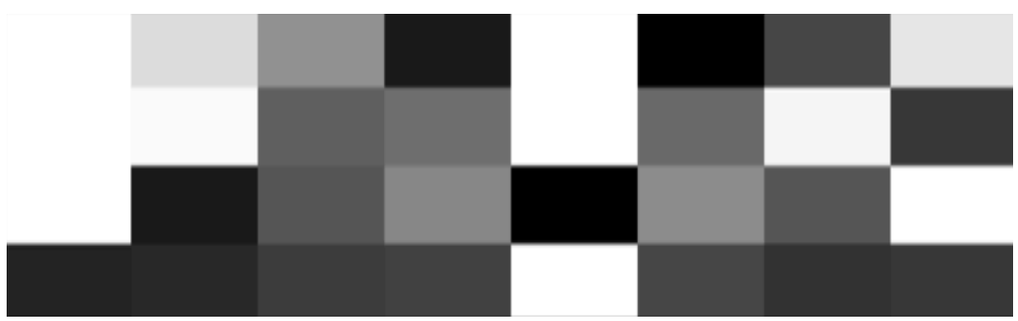


stem tracks

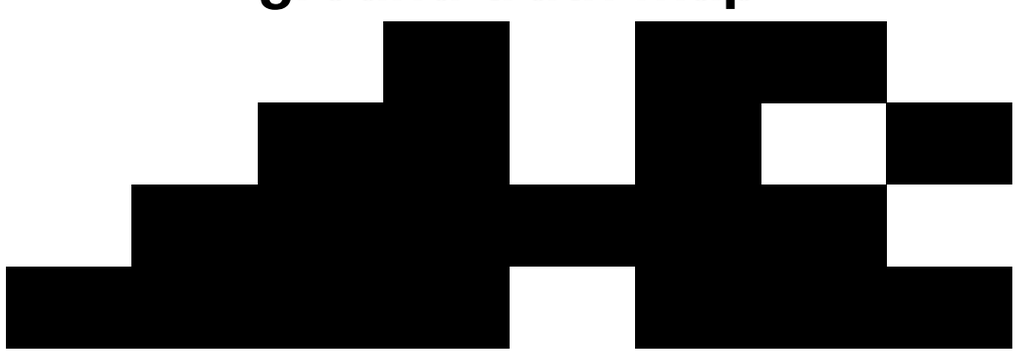


- Algorithm output 2: loop layout
  - Evaluate accuracy with correlation

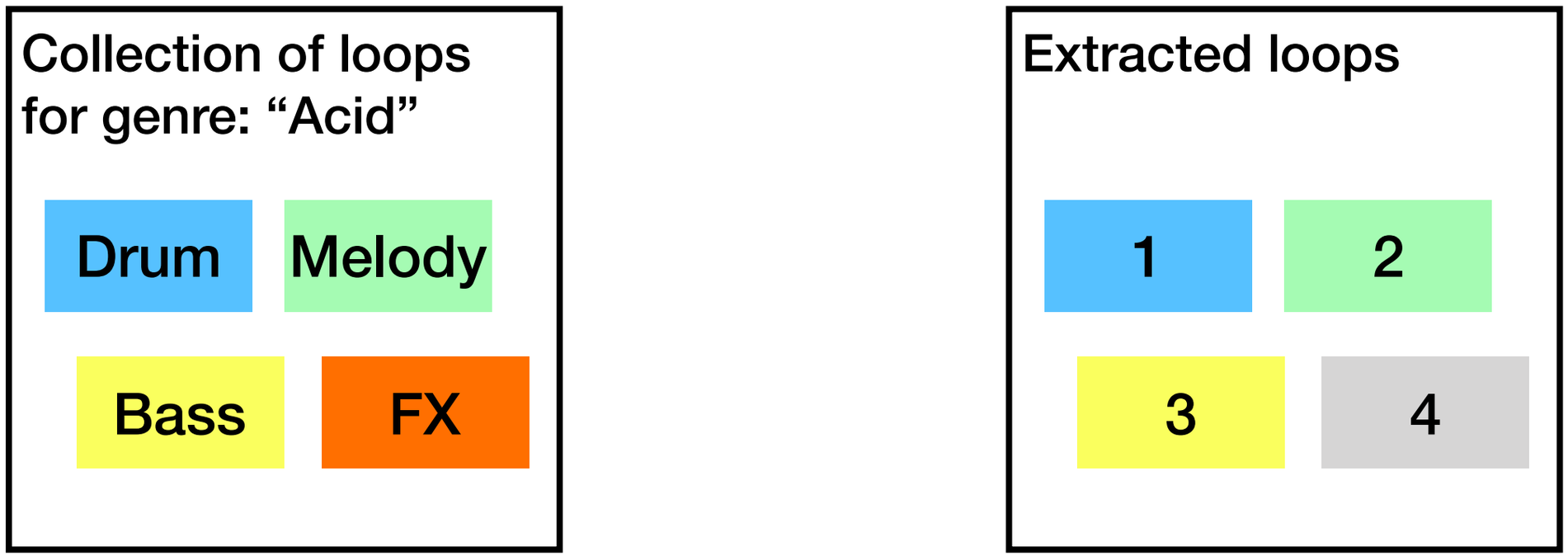
estimated map



ground truth map

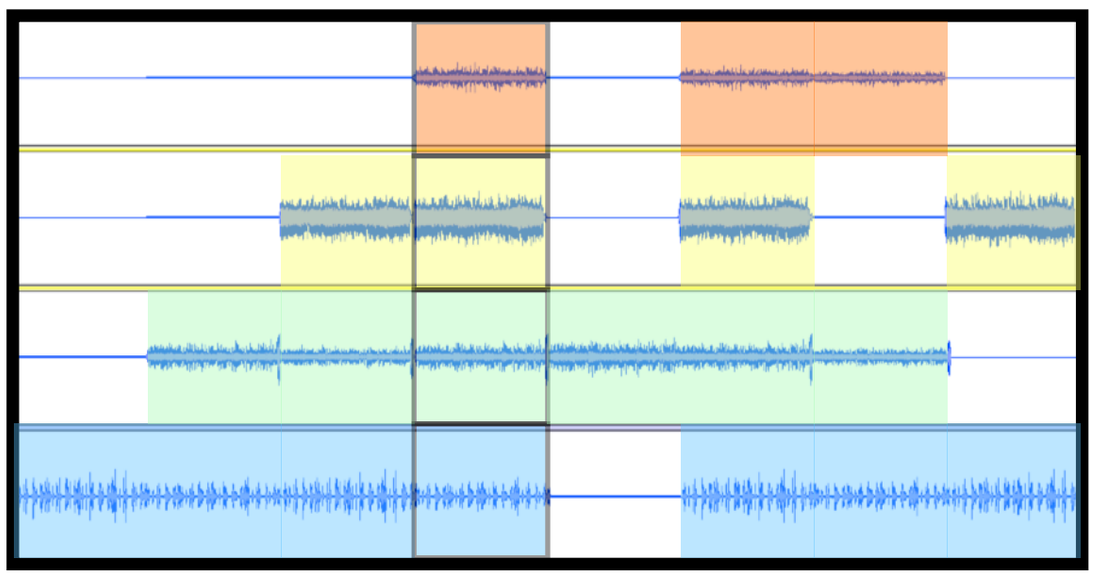


Good separation example

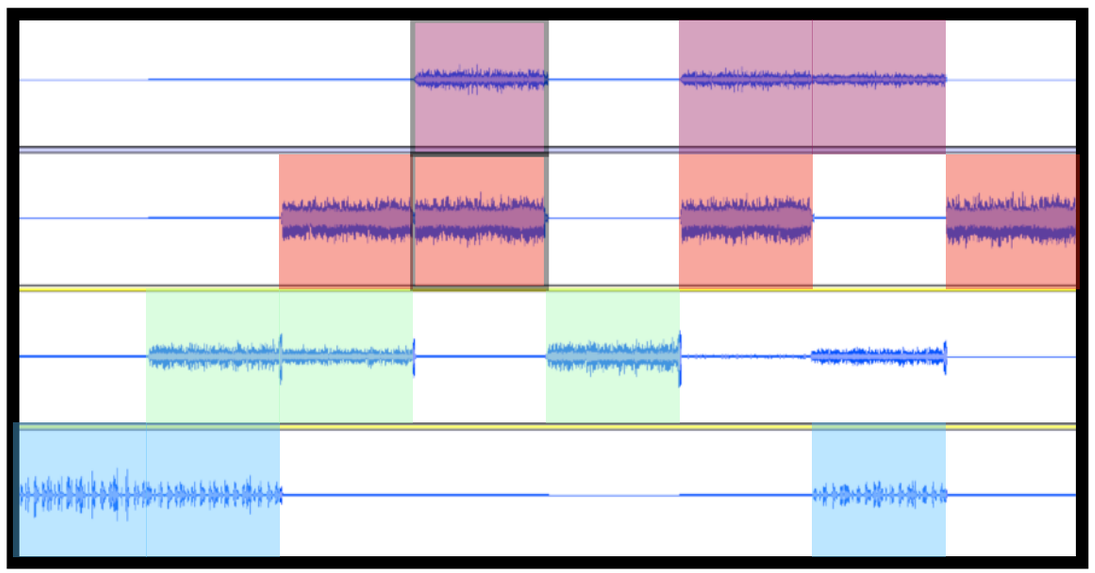


Flawed separation example

Original tracks for genre “Brezo”

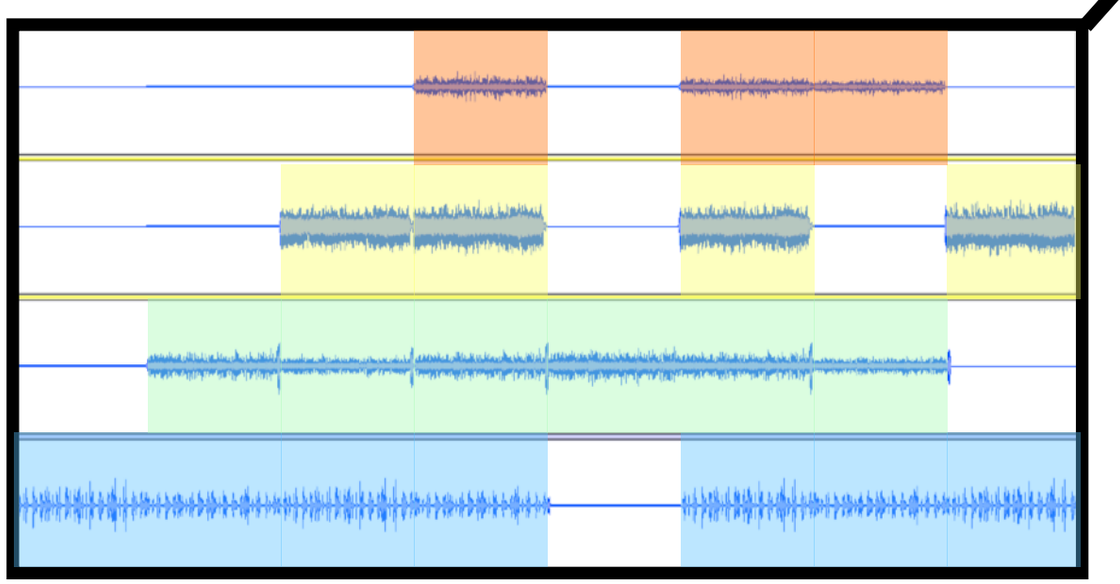


Source separated tracks



Flawed separation example

Original tracks for genre “Brezo”



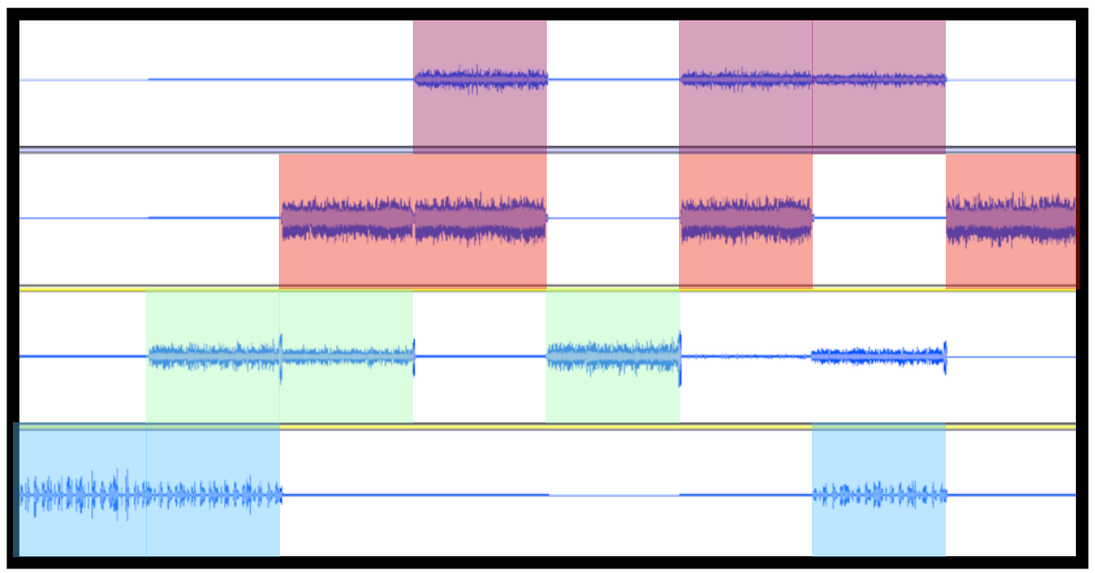
swap rows

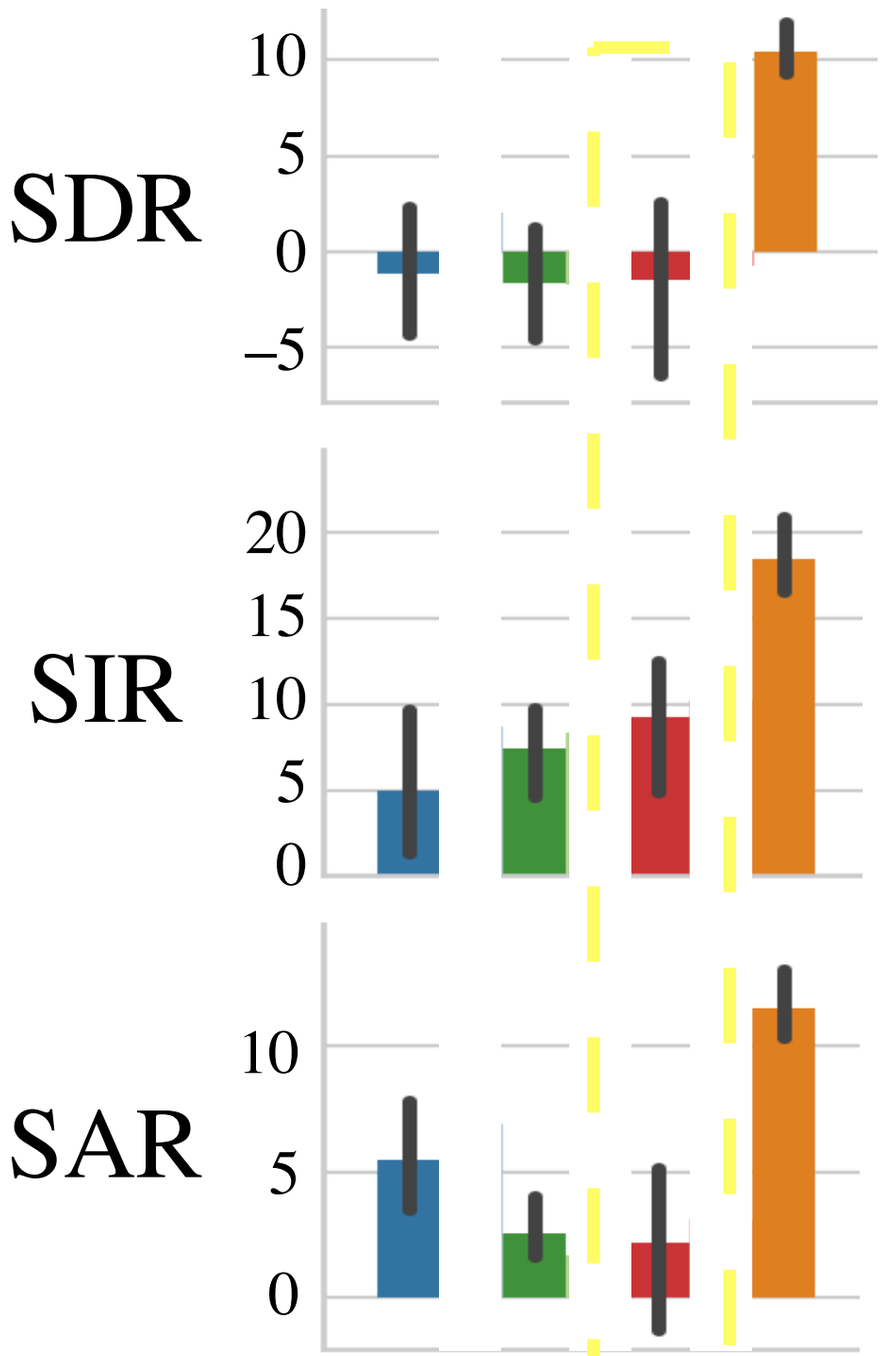
substitute =

substitute =

swap rows

Source separated tracks





Our reconstruction quality is average. :-|

We have less crosstalk than others! :-D

We have more noisy artifacts. :-(

We get very clean layouts! :-D

Algorithm

Iterative NMF

[Seetharaman & Pardo 2016]

NMFD

NTF **(proposed)**

Ideal Mask

(performance ceiling)

Conclusion

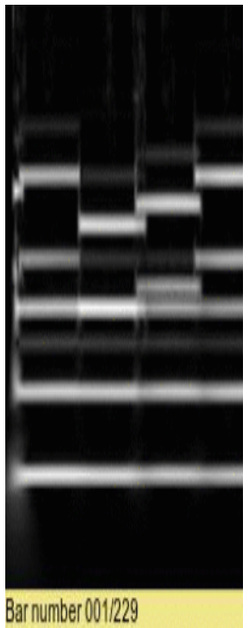
Conclusion

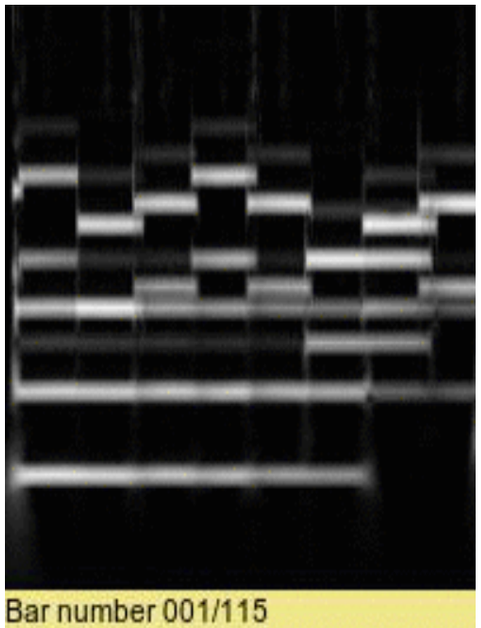
- Proposed method of decomposing audio into loops that:
    - Models periodicity using the spectral cube
    - Models source signals and song composition jointly
    - Tucker decomposition is musically intuitive
- Weaknesses include:
    - Very conservative reconstructions don’t model the whole signal
    - Like NMFD, we cannot distinguish between algebraically equivalent decompositions
- Future work: searching for repetitions at multiple hierarchical time scales

Future work: hierarchical analysis

- Different loops in the song have different lengths and periods

- Spectral cubes with different periods highlight different consistent repetitions



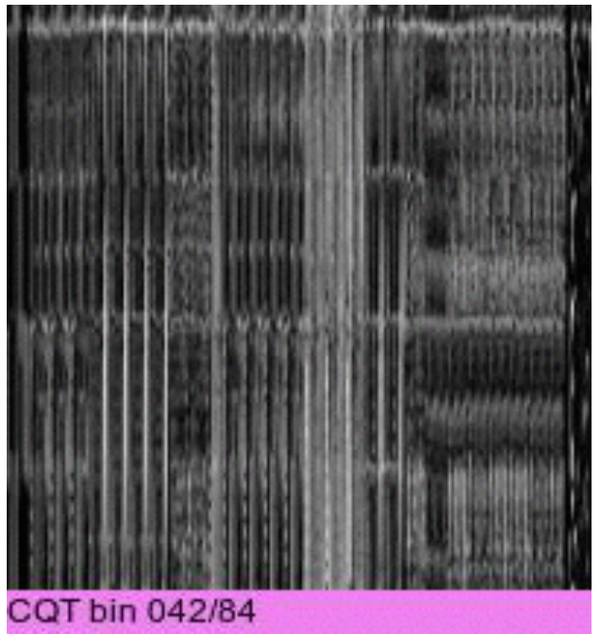


PERIOD: 2 beats 1 downbeat 4 downbeats

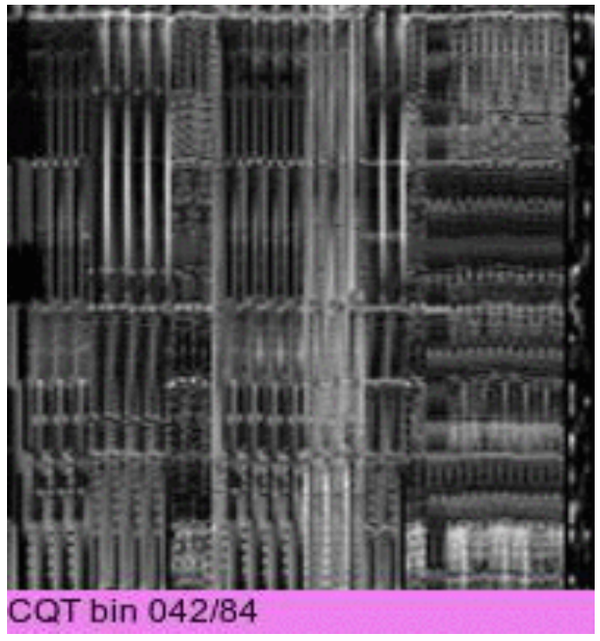
Future work: hierarchical analysis

- Different loops in the song have different lengths and periods

- Spectral cubes with different periods highlight different consistent repetitions

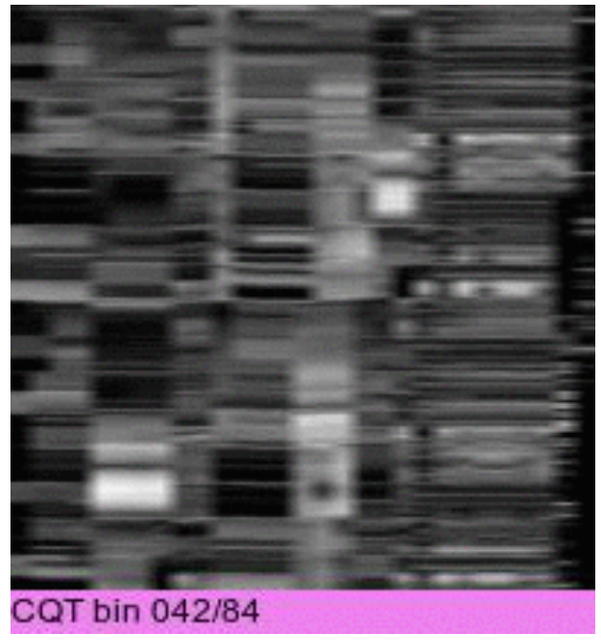


PERIOD: 2 beats



1 downbeat

2 downbeats



4 downbeats

Thank you!

PS. Jordan is now at:

ircam Centre Pompidou + L2S