Leveraging MIDI files for Music Information Retrieval

Colin Raffel

ISMIR 2017

Outline

- Preliminaries
- A brief history and overview of MIDI
- What kinds of information do MIDI files contain?
- Using `pretty_midi` to extract information from MIDI files
- Aligning MIDI files to audio recordings
- Matching MIDI files to audio corpora
- Exploring the Lakh MIDI Dataset
- How reliable is the LMD?
- Calls to action

If you want to follow along...

- Download "LMD mini":
  http://colinraffel.com/projects/lmd/lmd_mini.tar.gz
- Get a Python environment set up, with (all `pip` installable)
  - `pretty_midi`
  - `librosa`
  - `mir_eval`
  - `matplotlib`
  - `numpy`
  - `jupyter`
  - `tables`
  - `pyfluidsynth` (optional, requires `fluidsynth`)

Why am I doing this?
And what are you going to get out of it?


MIDI FILES
MIDI FILES EVERYWHERE
imgflip.com

What is MIDI?

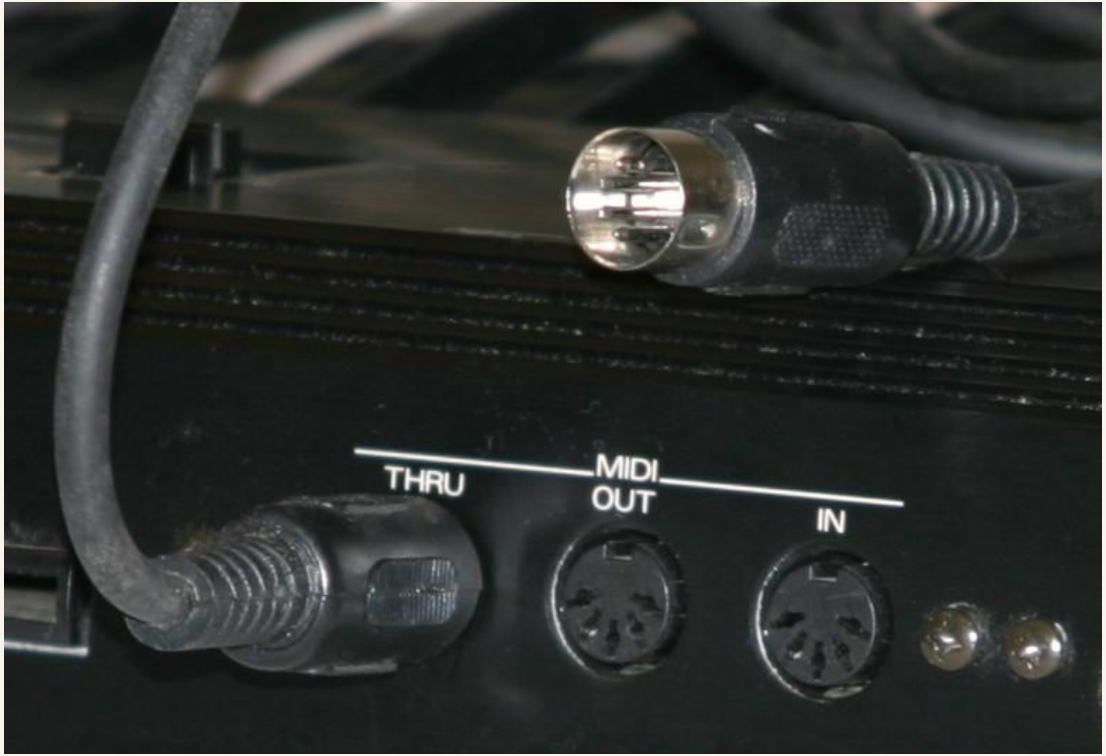


Photo by user Pretzelpaws, Wikipedia, CC-BY-SA

What is MIDI?

Photo by use[…]a, CC-BY-SA

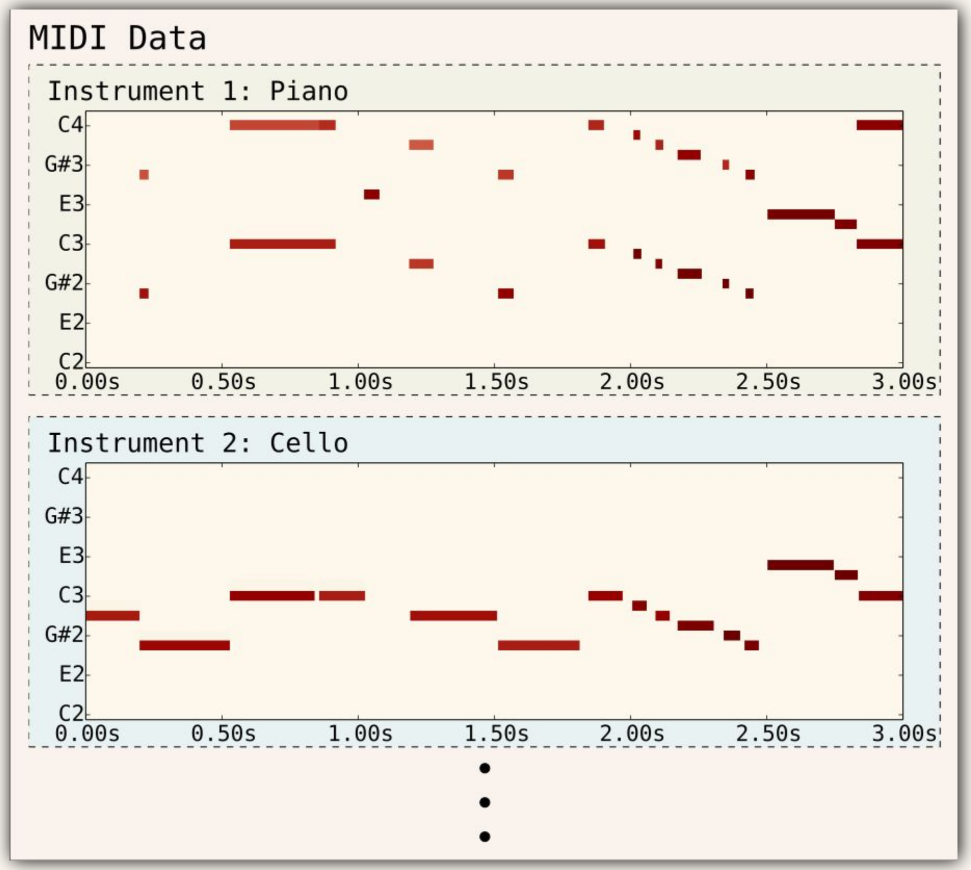
MIDI Files

```
0000000 4d 54 68 64 00 00 00 06 00 01 00 02 00 dc 4d 54
0000010 72 6b 00 00 00 13 00 ff 51 03 07 a1 20 00 ff 58
0000020 04 04 02 18 08 01 ff 2f 00 4d 54 72 6b 00 00 00
0000030 0e 00 c0 03 2c 90 3c 40 2c 3c 00 01 ff 2f 00
```

Change to
program number 3

Play note 60 (C4)
at velocity 64

MIDI Files

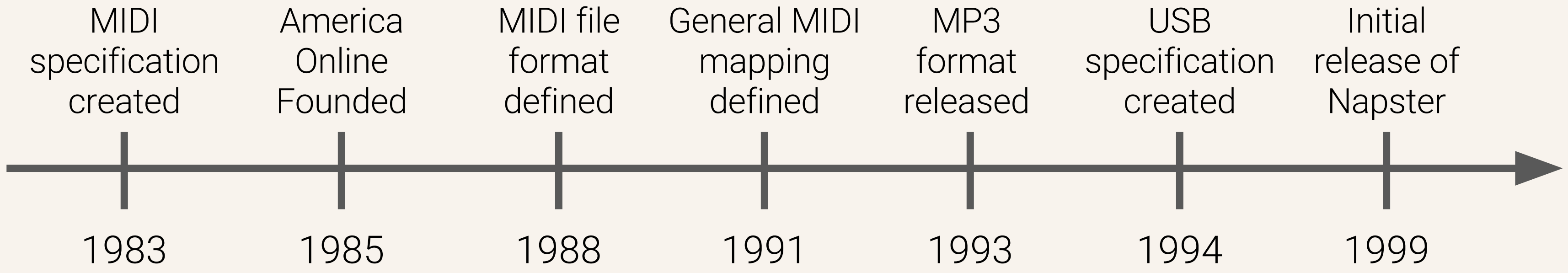


MIDI Data
Instrument 1: Piano
C4
G#3
E3
C3
G#2
E2
C2
0.00s
0.50s
1.00s
1.50s
2.00s
2.50s
3.00s
Instrument 2: Cello
C4
G#3
E3
C3
G#2
E2
C2
0.00s
0.50s
1.00s
1.50s
2.00s
2.50s
3.00s

General MIDI

Photo by user Darashinaikuma, Wikipedia, CC-BY-SA

MIDI Timeline



MIDI BBS

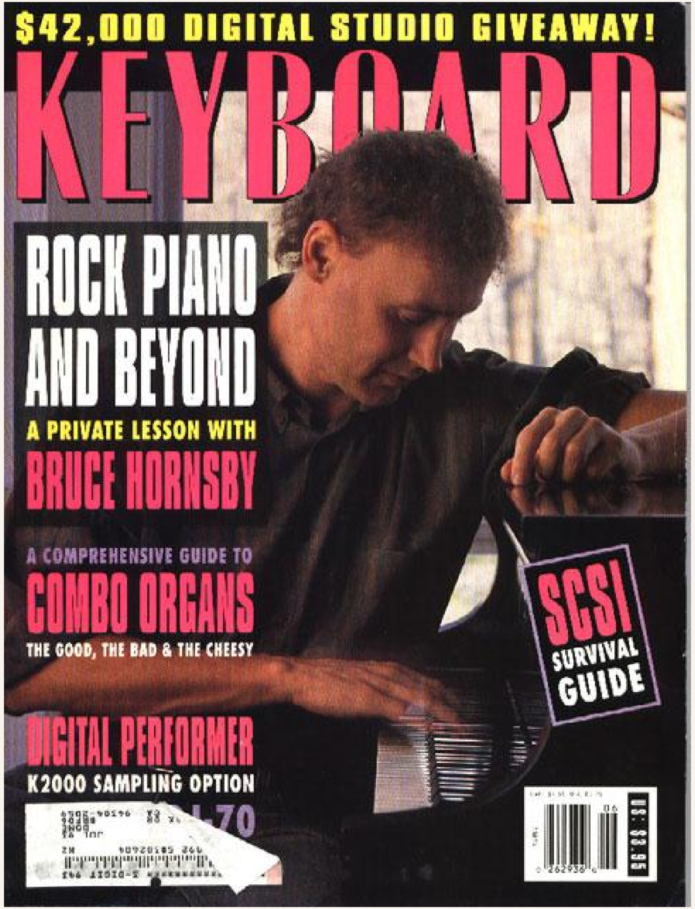
MystMagical MIDI	Omaha, NB	Pete Olsen	MIDINEB	402-293-0451
PGH-MIDI Music	Pittsburg, PA	Art Doud	MUSIC	412-882-3703
MIDI Thru	Aveiro, PORT	Fausto Carvalho	MIDITHRU	35-1-34915452
1st Dutch MIDI	Delft, HOLLAND	Johnan Corstjens		31-1-5138754
TBUS-BBS	Munich, GERMANY	Rudolf Stricker		49-8-9293881
1st Austria MIDI	Vienna, AUSTRIA	Erich Varga		43-1-7693132
Music Studio UK	UNITED KINGDOM	Paul Urmston		44-0+926403904
Rock & Jazz	Paris, FRANCE	Sam Przyswa		33+1-40548604
Slatch	Paris, FRANCE	Frank Gardes	SLATCH	33+1-48020814
Twilight Zone	Barrie, ONTARIO	Robin Wells	TWILIGHT	705-722-8184
Action Link	Bradenton, FL	Jim Davie	ACTION	813-747-9295

"O" after the telephone number means that BBS has a door for off-line message handling. Saves time on line, and $ long distance!

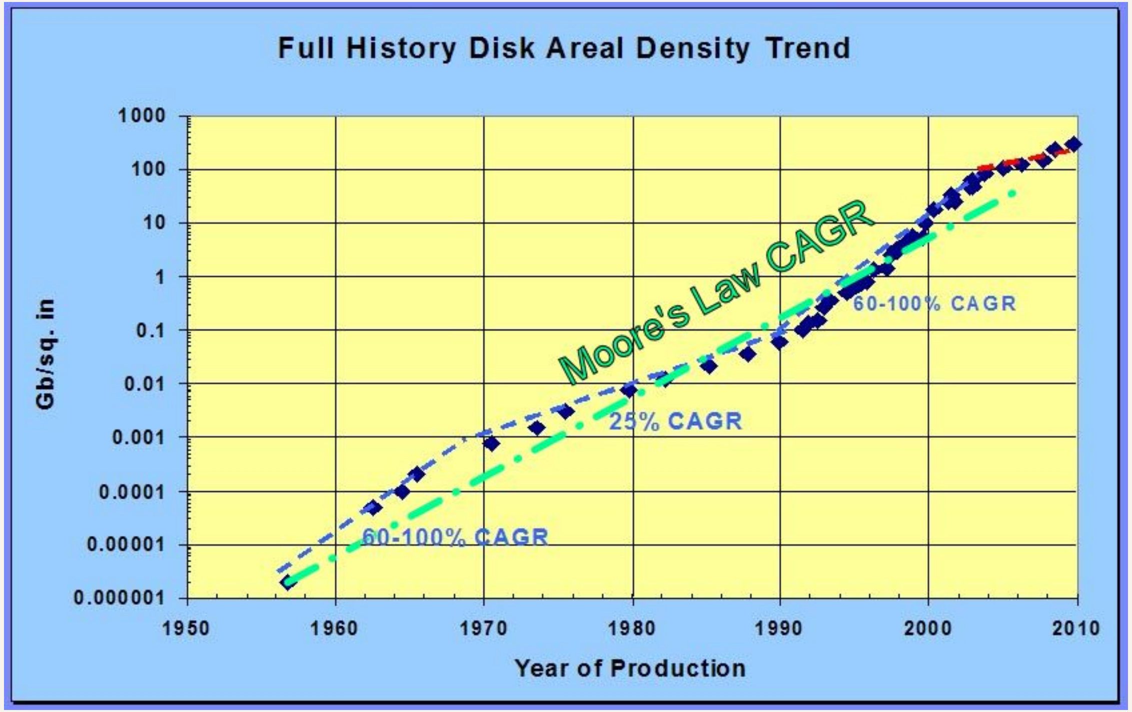
“Bulletin Board Systems”

Photo by user Liftarn, Wikipedia, CC-BY-SA

“You log on to your computer service (Compuserve, GEnie, Prodigy, America On Line, etc.), and you see that someone has made a MIDI Song File of your song, and has uploaded it so any Tom, Dick, or Harry can download it. You even download a copy out of curiosity. As they don't charge you any extra to download it, you assume that you'll get paid royalties out of the online flat-rate fee.
Wrong! The MIDI File of your song is being given away (yes - published electronically for free), and no royalties are being collected or distributed.”



Why are there so many of them?



"Prediction: The cost for 128 kilobytes of memory will fall below U$100 in the near future."

Creative Computing magazine
December 1981, page 6

Why are there so many of them?

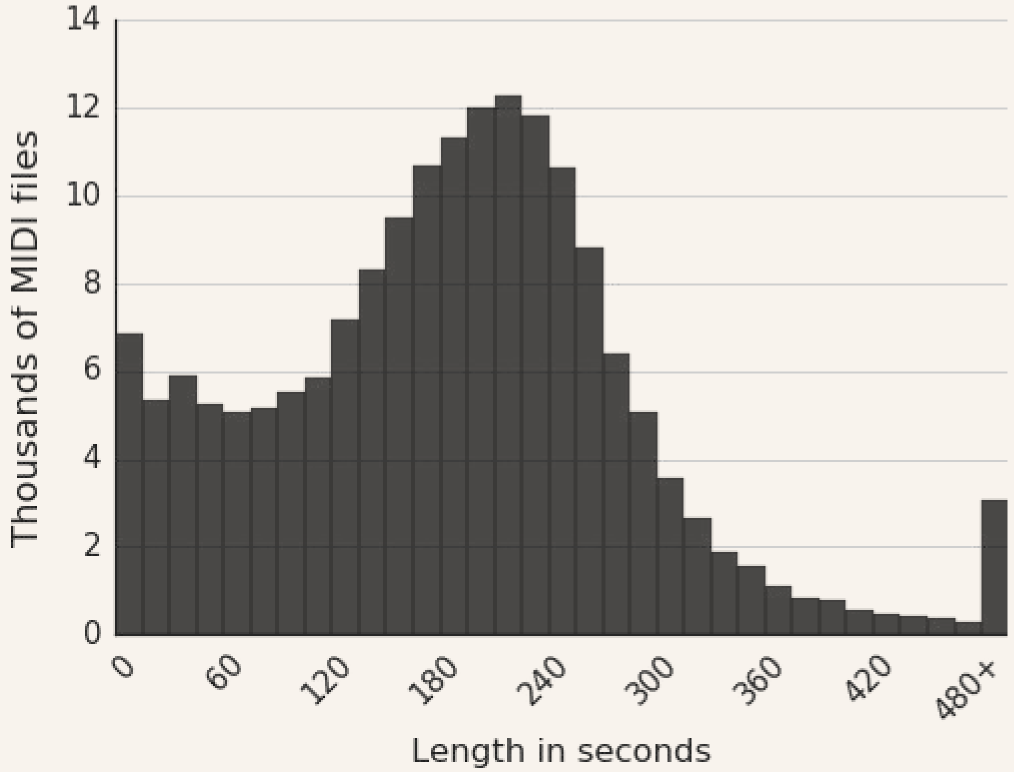
Photo by user Pokman817, Wikipedia, CC-BY-SA

Why are there so many of them?

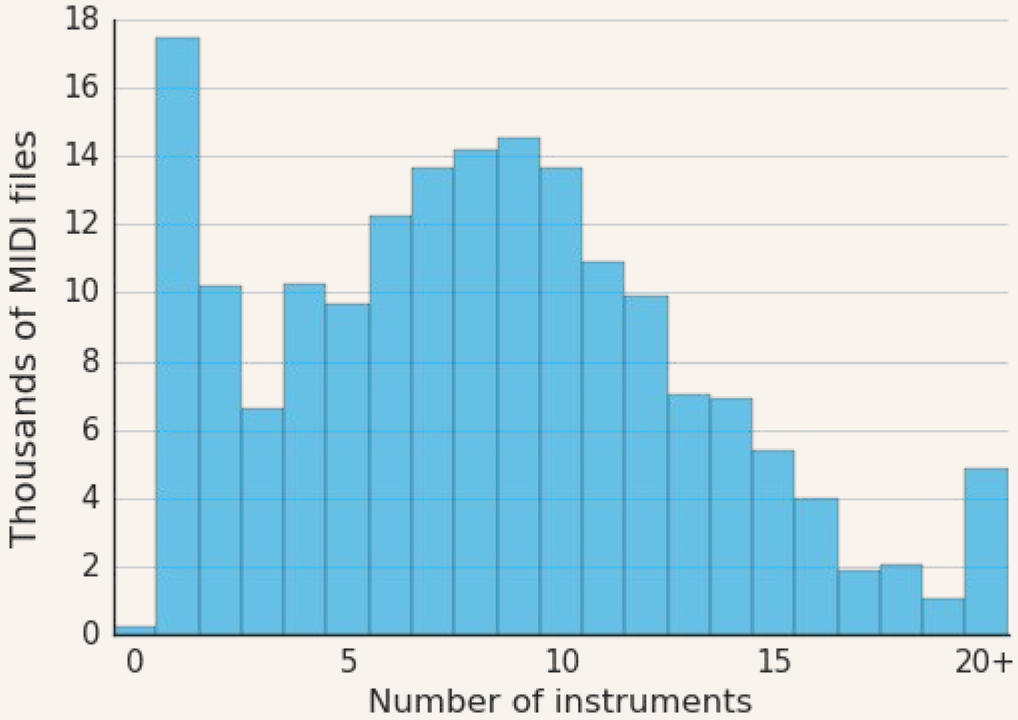
What information is typically in MIDI files?

178,561

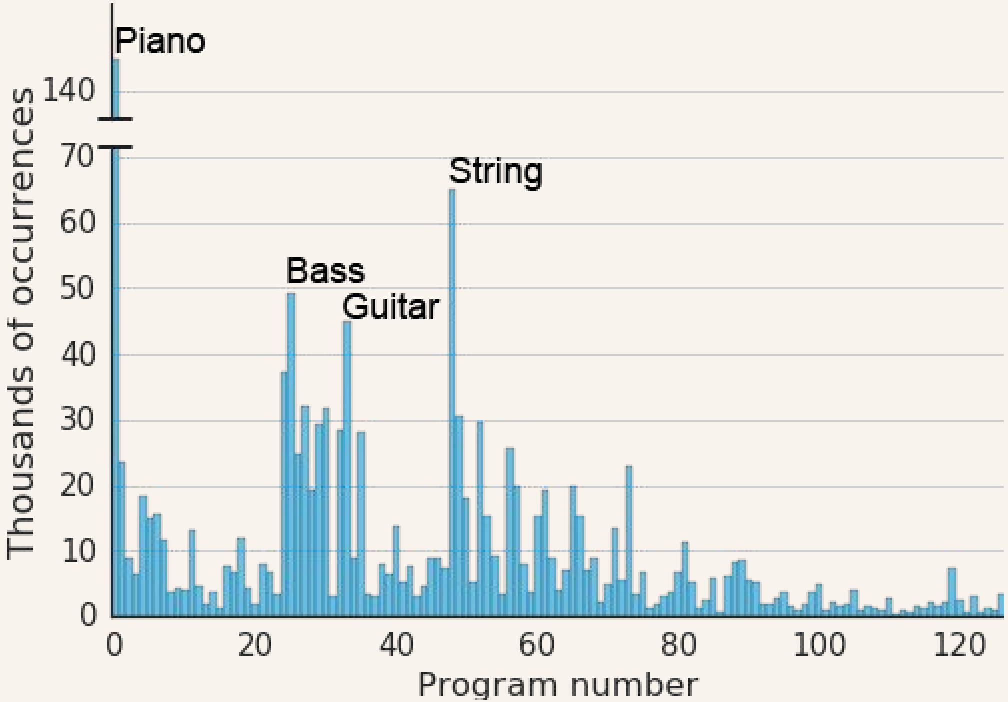
Lengths


Thousands of MIDI files
14
12
10
8
6
4
2
0
0
60
120
180
240
300
360
420
480+
Length in seconds

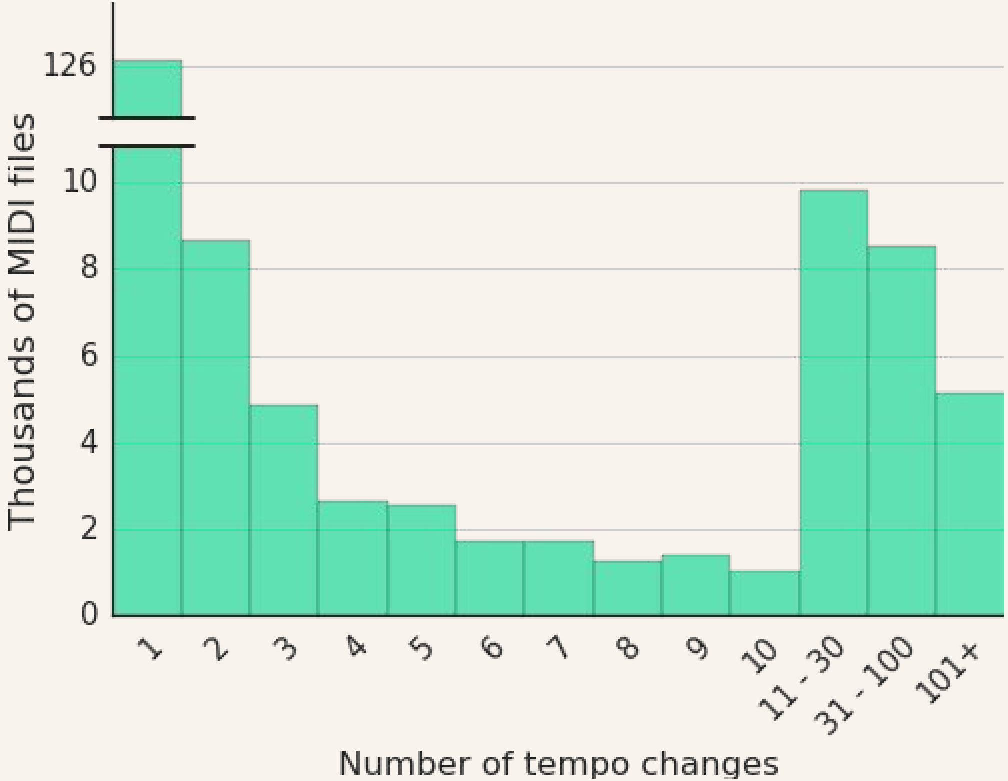
Transcription



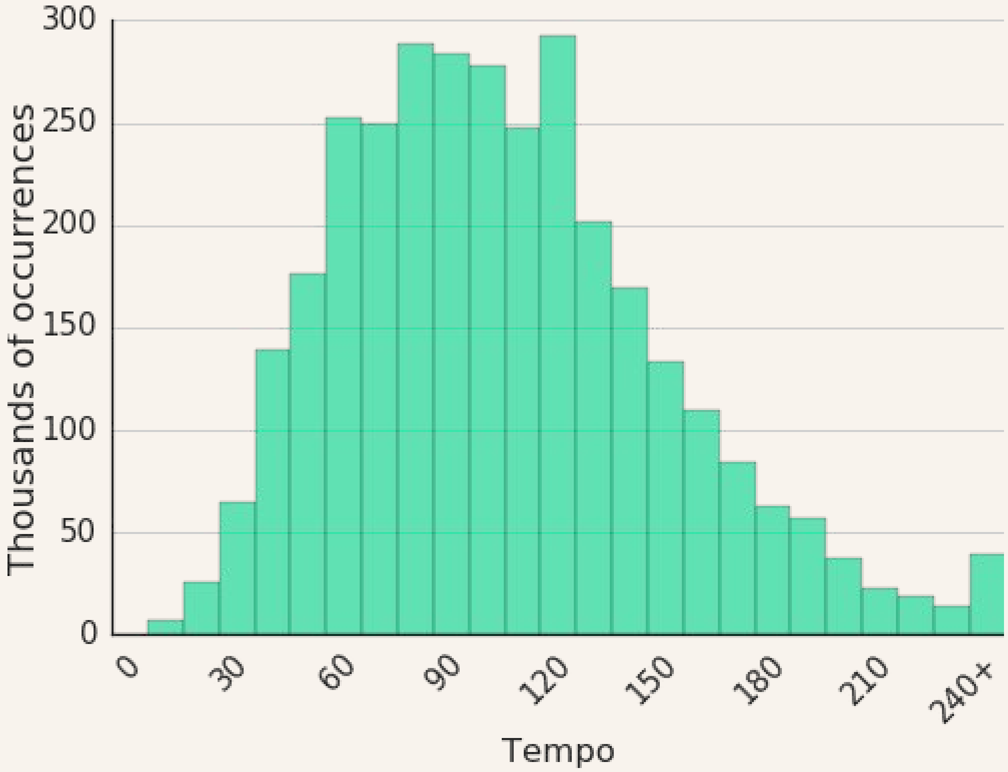
Instrumentation



Tempo

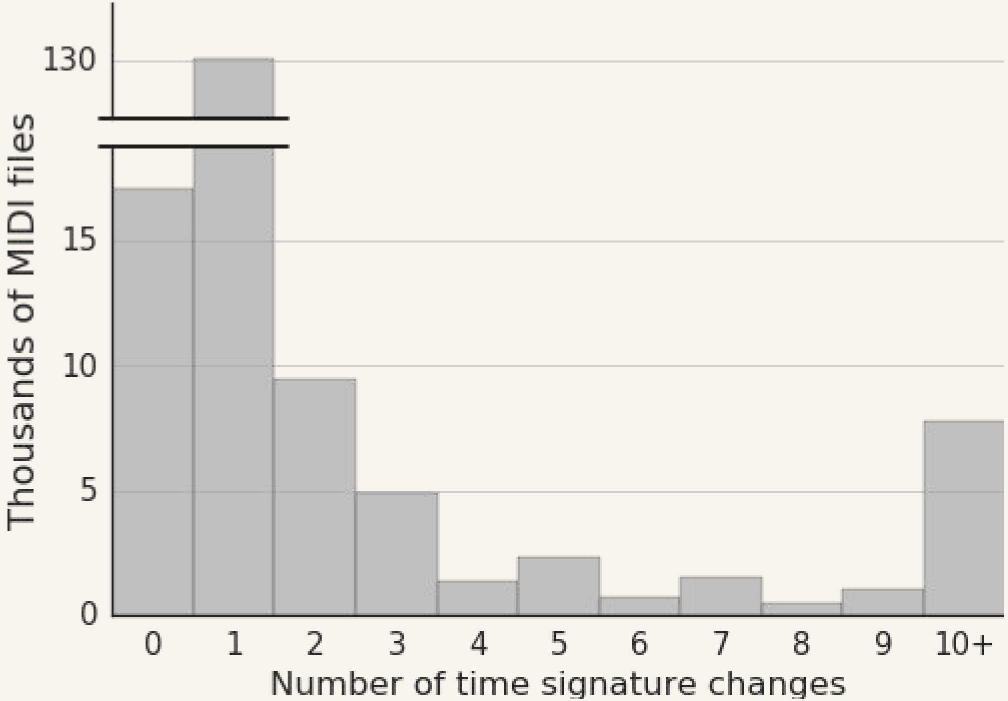

Thousands of MIDI files
126
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
11 - 30
31 - 100
101+
Number of tempo changes

Tempo

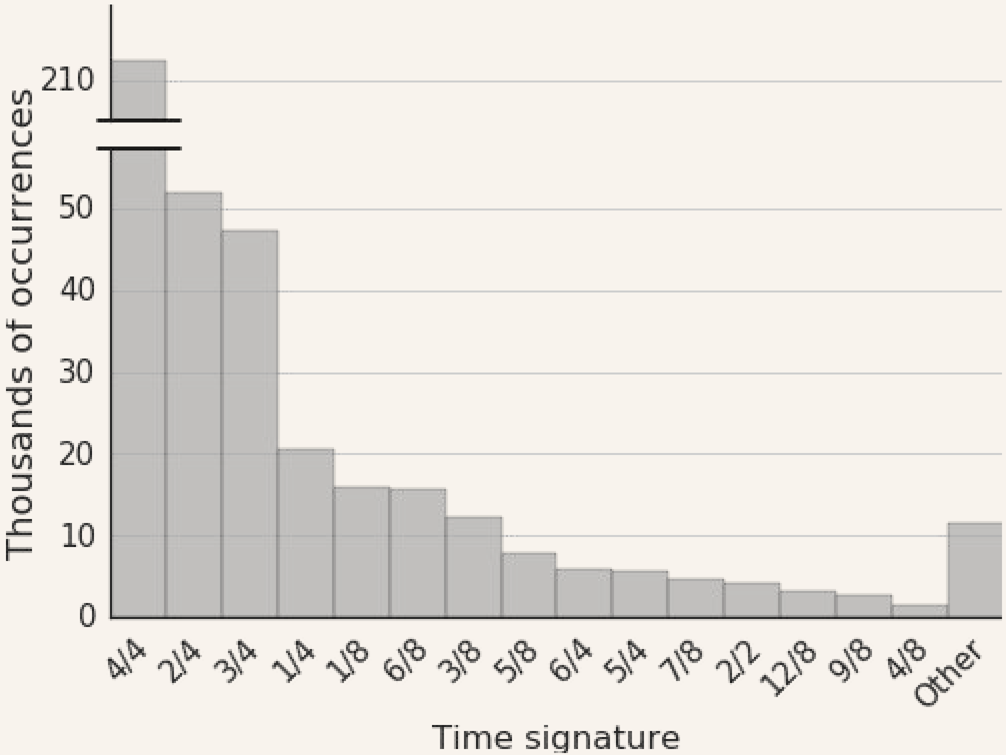


300
250
200
150
100
50
0
Thousands of occurrences
0
30
60
90
120
150
180
210
240+
Tempo

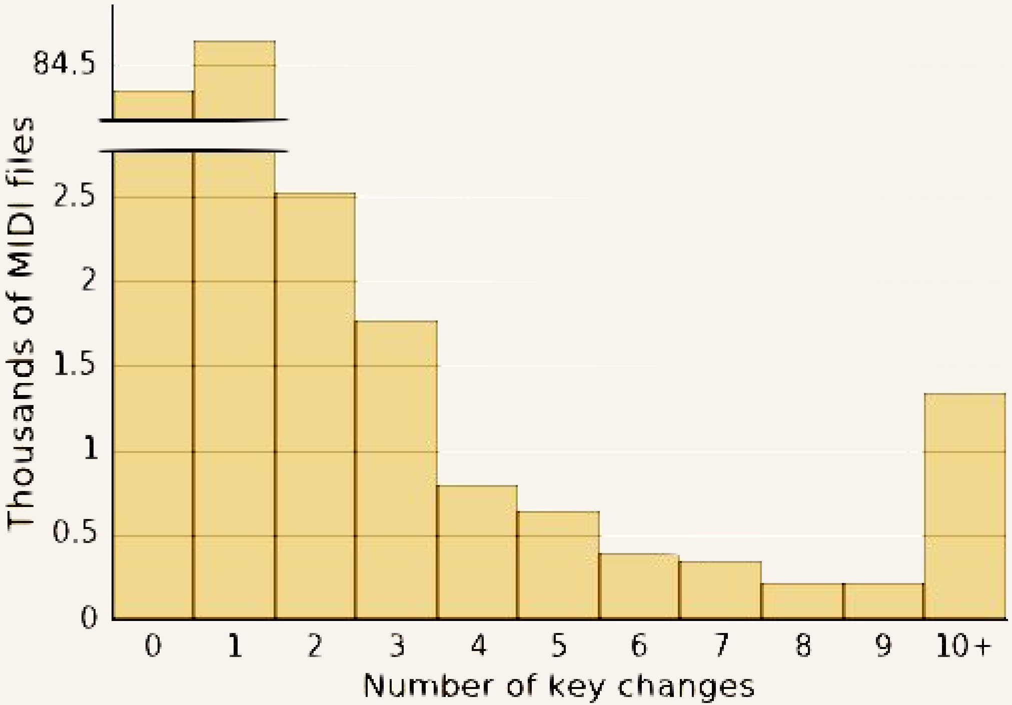
Time Signature



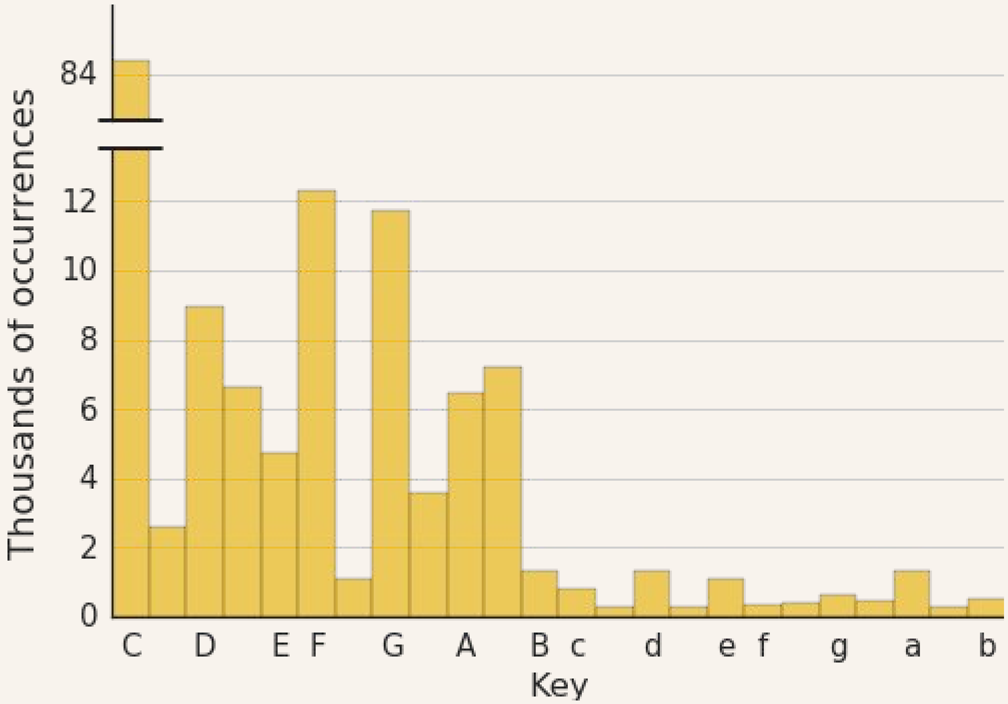
Time Signature



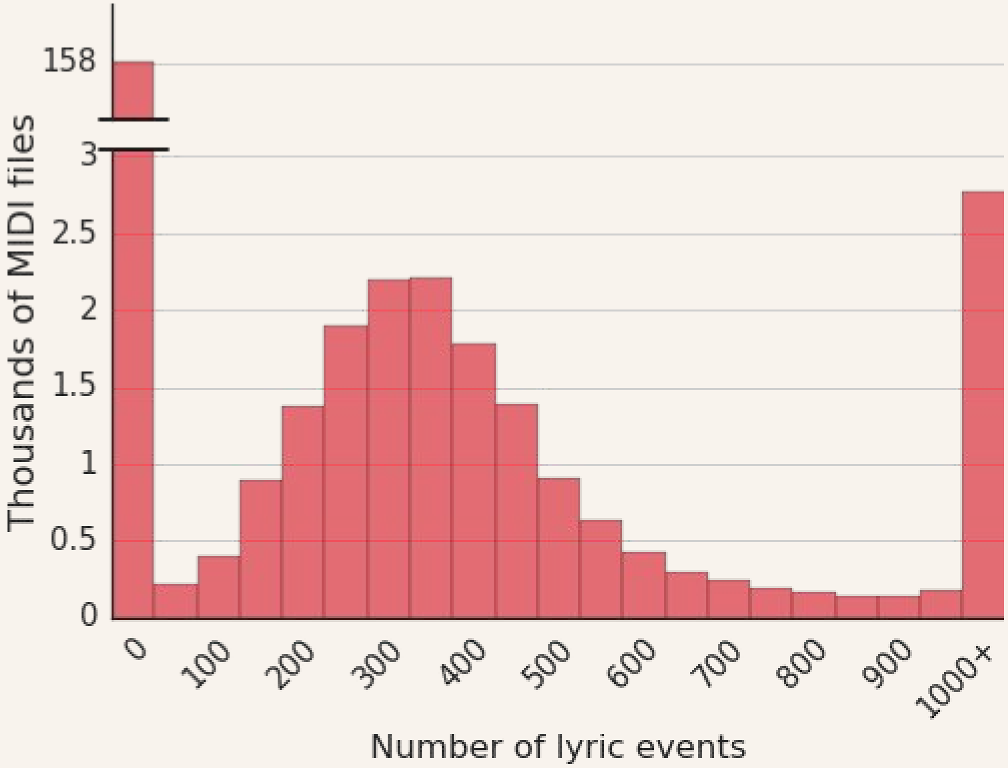
Key Signature



Key Signature

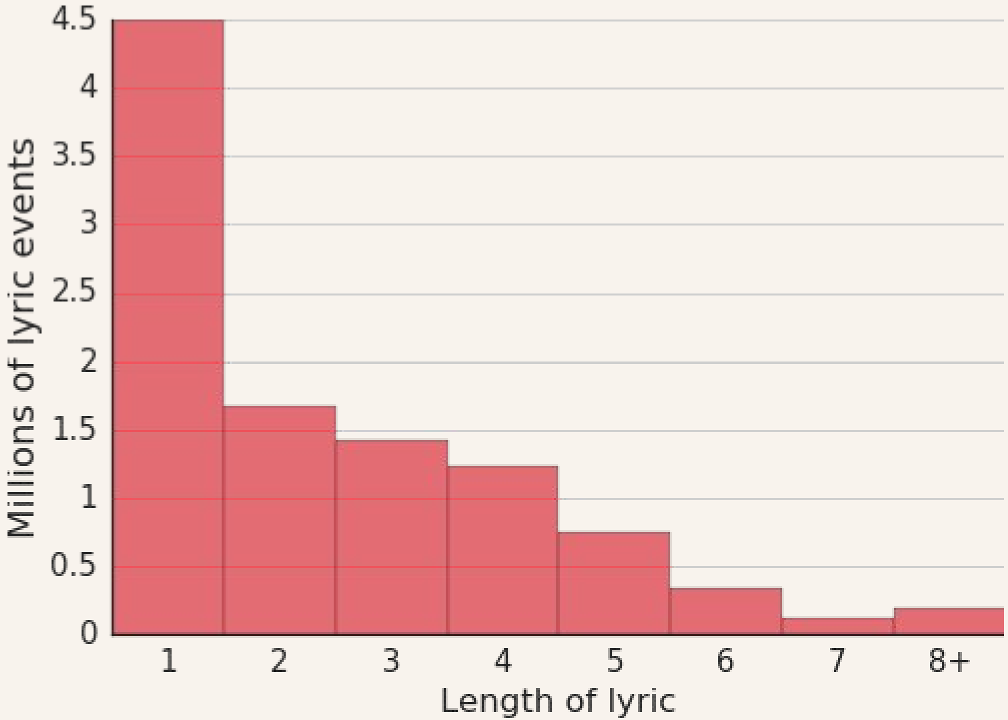


Lyrics



Thousands of MIDI files
158
3
2.5
2
1.5
1
0.5
0
0
100
200
300
400
500
600
700
800
900
1000+
Number of lyric events

Lyrics



Millions of lyric events
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7
8+
Length of lyric

Things you *can't* (reliably) get from MIDI

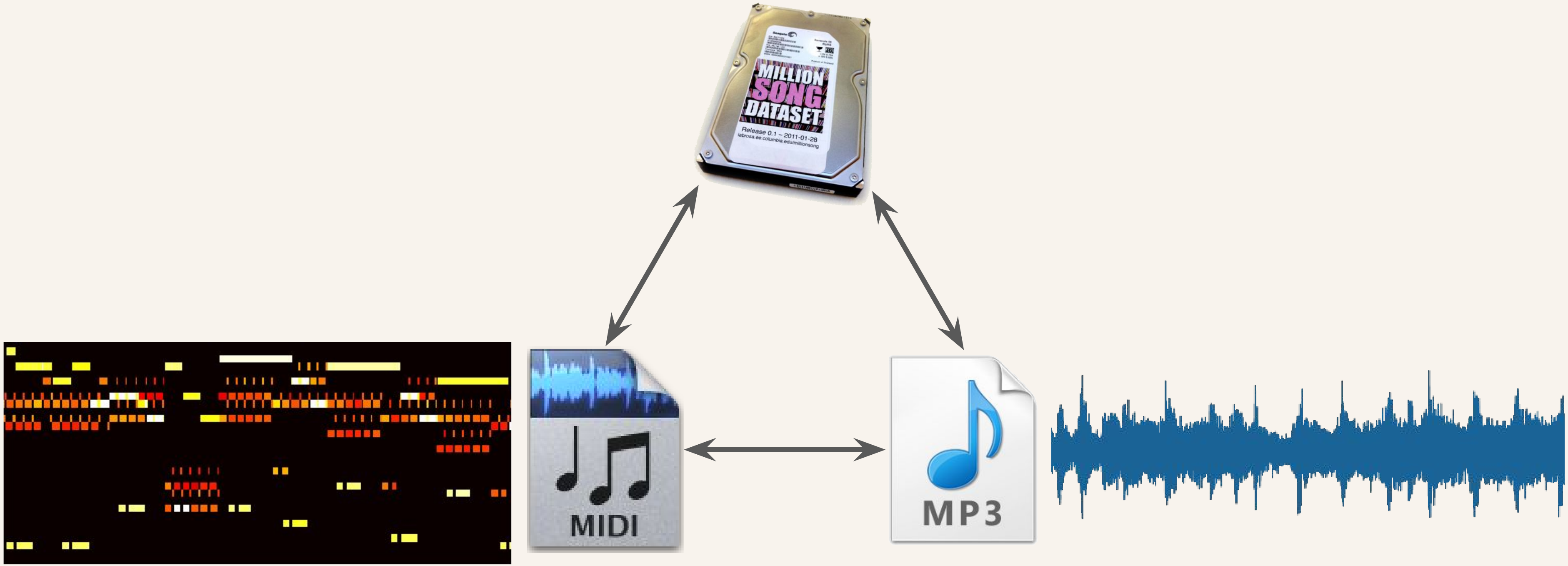
- Vocal track
- Melody track
- Instrument name
- Chord annotations
- Structural annotations
- Metadata

`pretty_midi` tutorial

`https://goo.gl/YL687S`

Matching

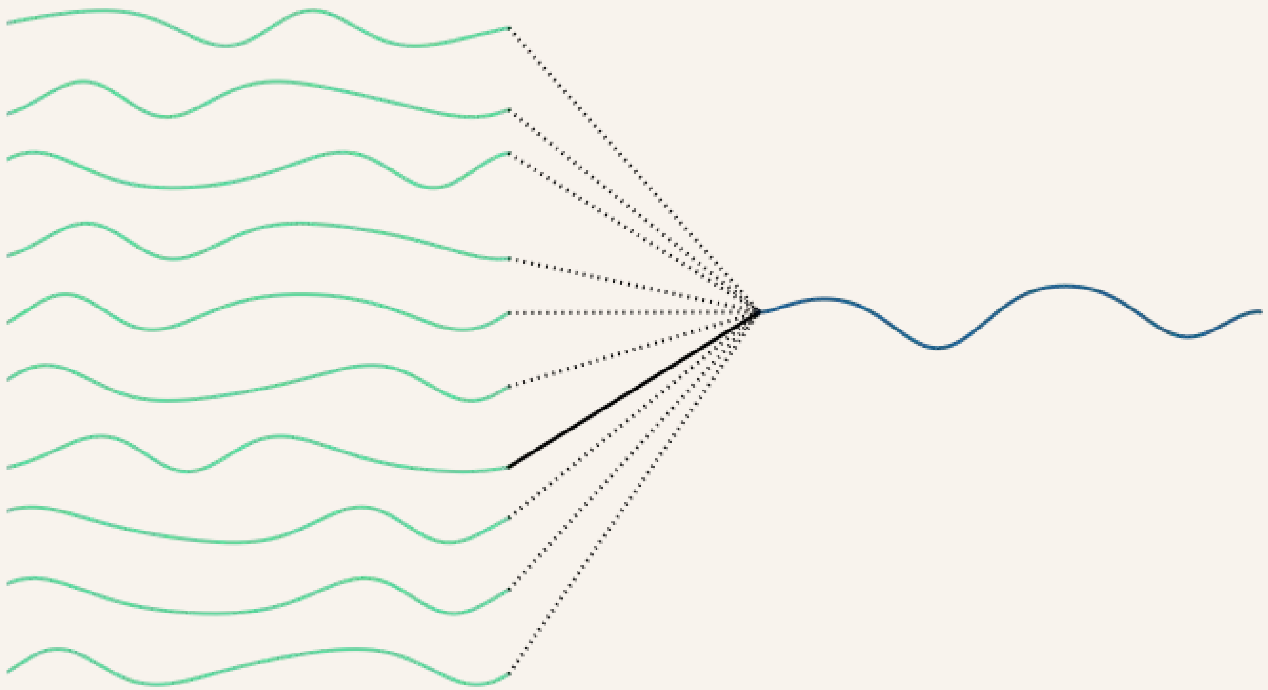
```
     artist: 'Tori Amos'
    release: 'LIVE AT MONTREUX'
      title: 'Smells Like Teen Spirit'
         id: 'TRKUYPW128F92E1FC0'
   duration: 216.4502
sample_rate: 22050
  audio_md5: '8'
 7digitalid: 5764727
       year: 1992
```



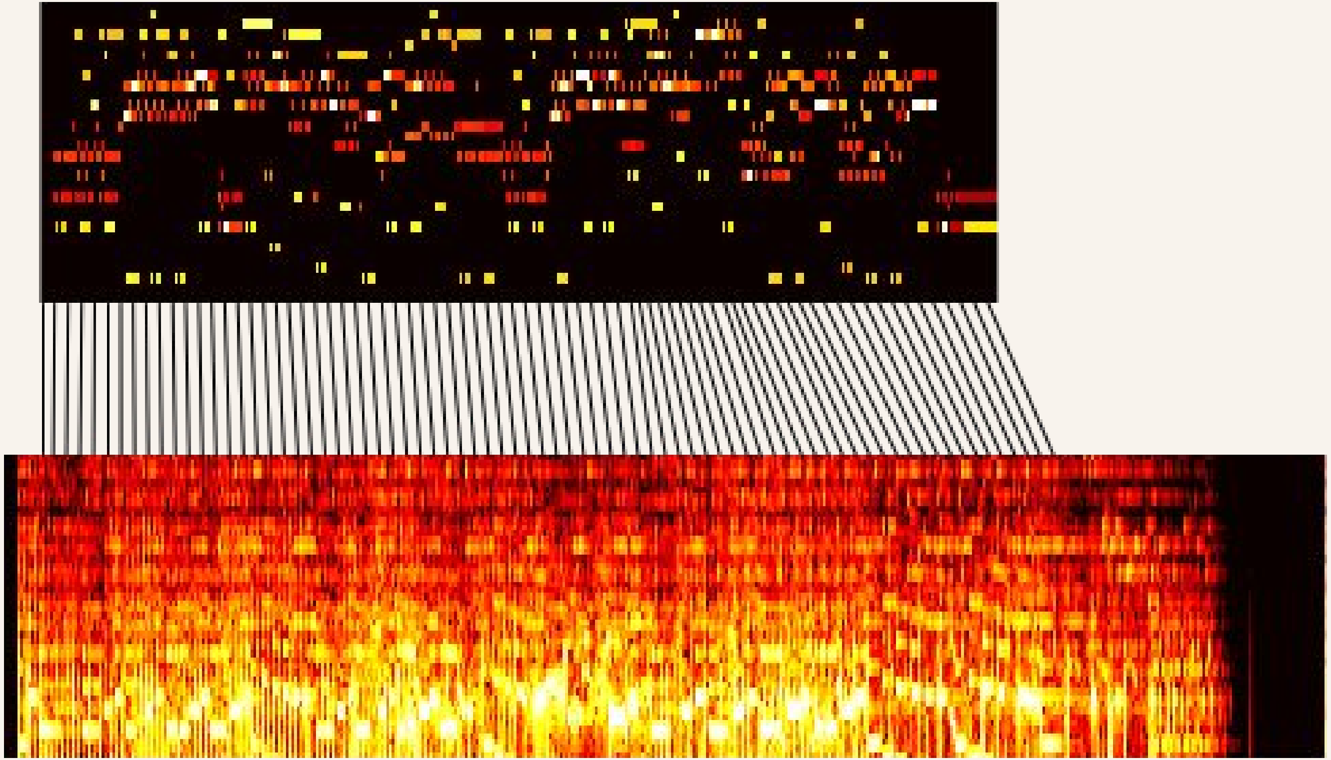
Matching by metadata won’t work

- J/Jerseygi.mid
- V/VARIA18O.MID
- Carpenters/WeveOnly.mid
- 2009 MIDI/handy_man1-D105.mid
- G/Garotos Modernos - Bailanta De Fronteira.mid
- Various Artists/REWINDNAS.MID
- GoldenEarring/Twilight_Zone.mid
- Sure.Polyphone.Midi/Poly 2268.mid
- d/danza3.mid
- 100%sure.polyphone.midi/Fresh.mid
- rogers_kenny/medley.mid
- 2009 MIDI/looking_out_my_backdoor3-Bb192.mid

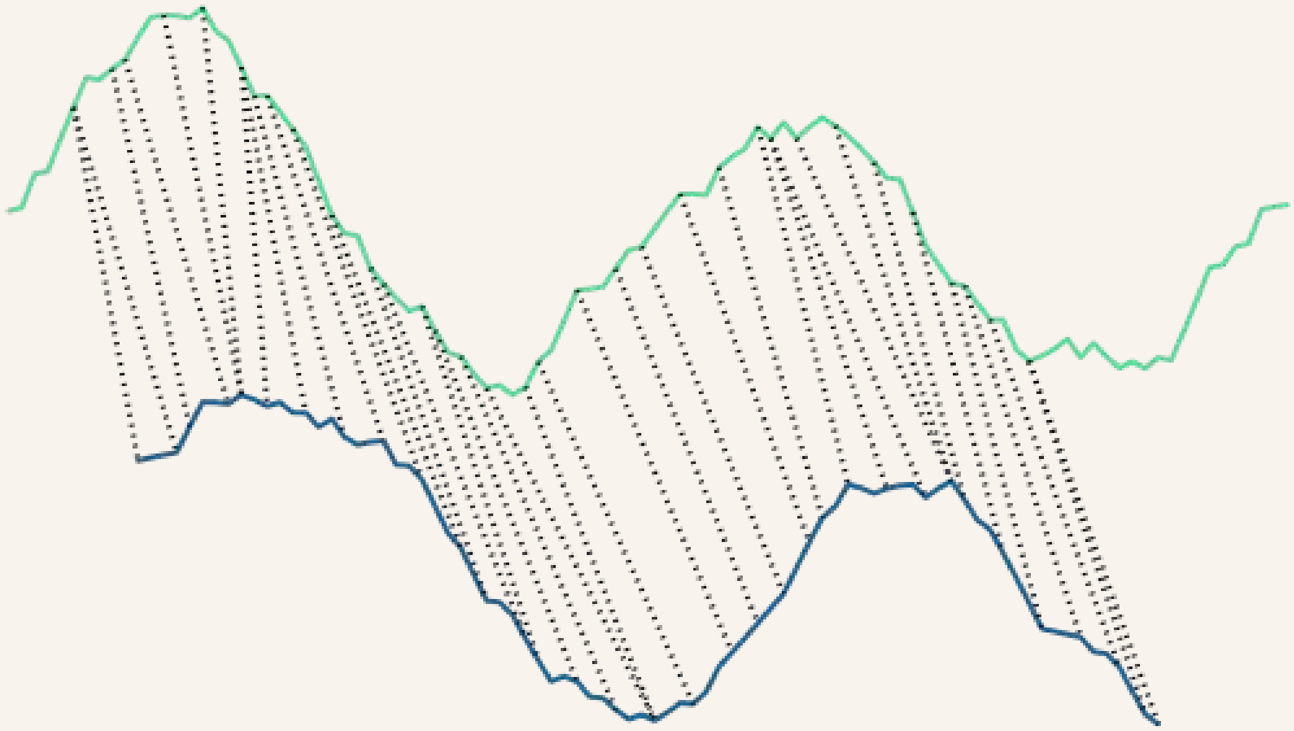
Sequence Retrieval



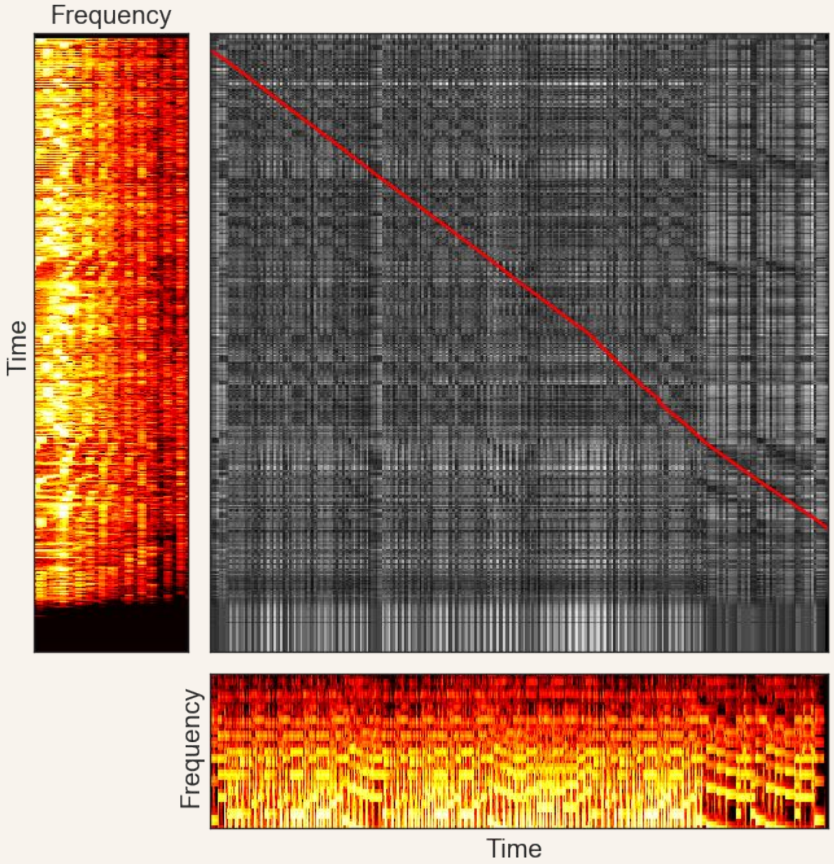
Aligning



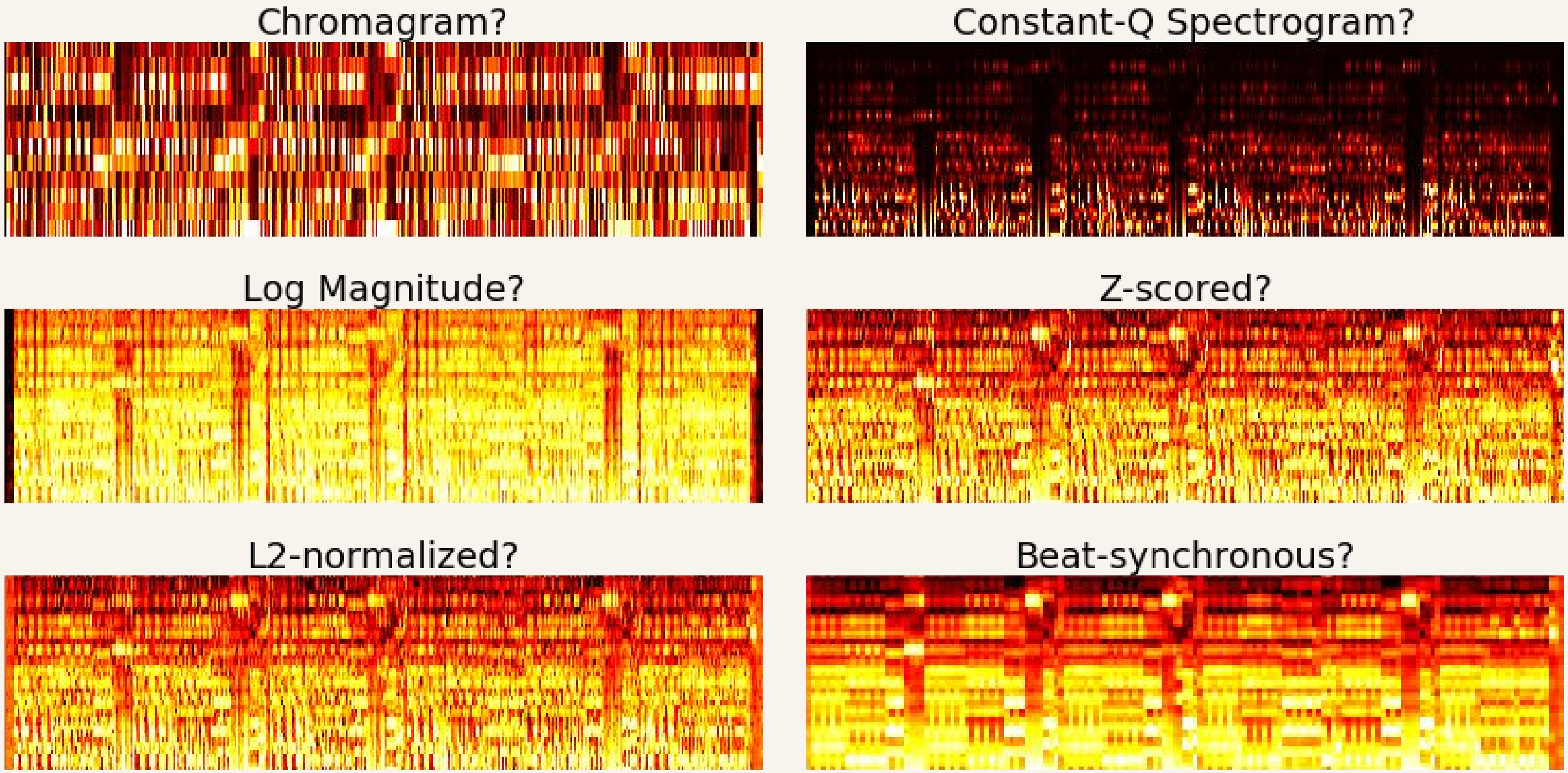
Dynamic Time Warping



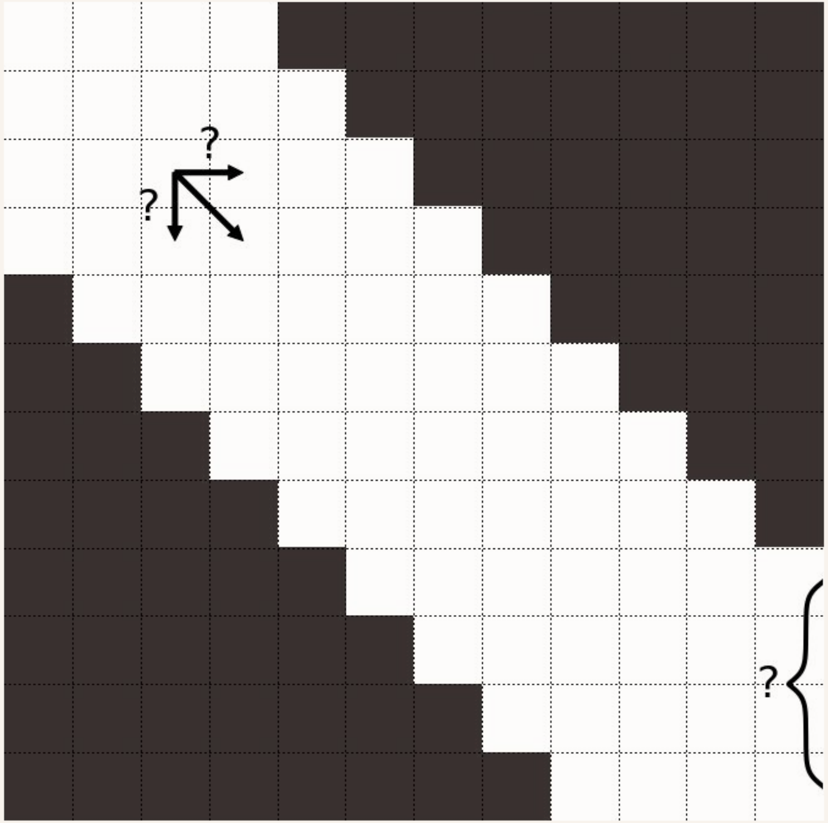
Dynamic Time Warping



Representation?



Path constraints?

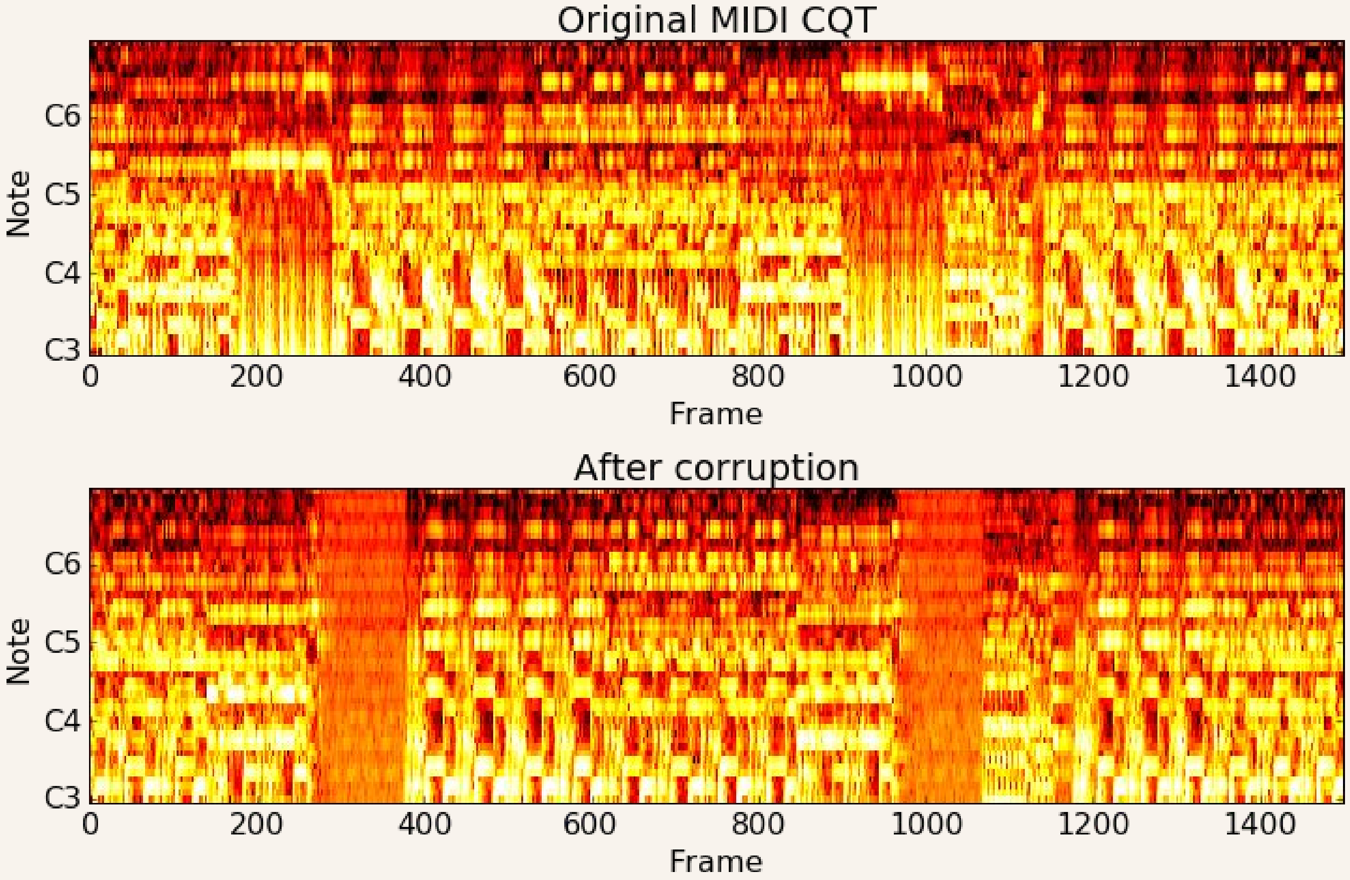


Score normalization?

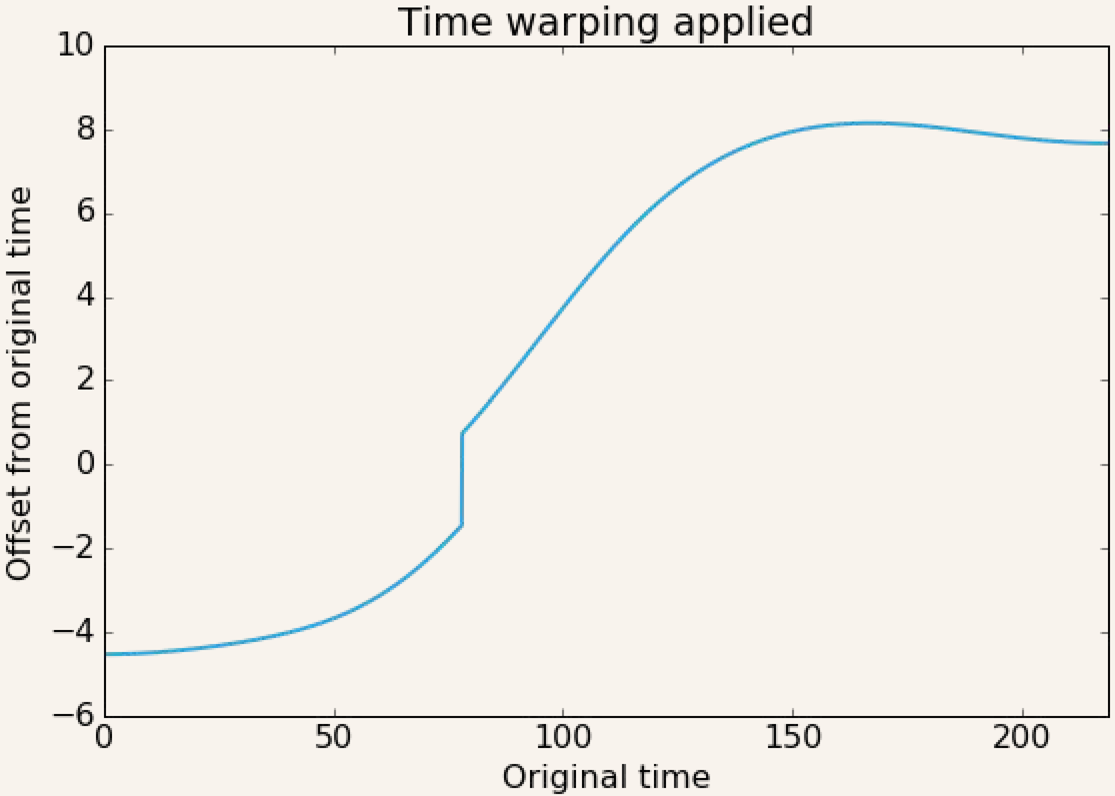
$$\text{score} = \frac{\sum_{i=1}^{|p_m|} D[p_m[i], p_a[i]] + \Phi(i)}{|p_m| \frac{\sum_{i=\min(p_m)}^{\max(p_m)} \sum_{j=\min(p_a)}^{\max(p_a)} D[i,j]}{|\max(p_m)-\min(p_m)||\max(p_a)-\min(p_a)|}}$$

Bayesian Optimization

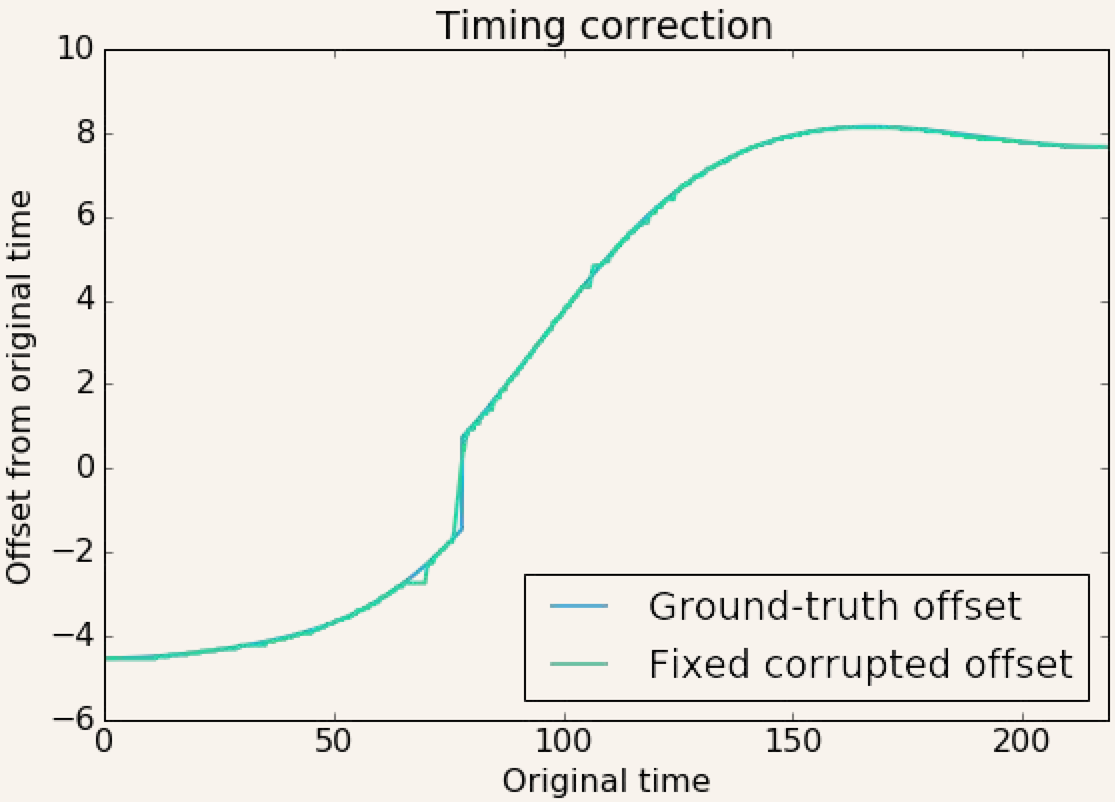
Idea: Synthetic Alignment Data



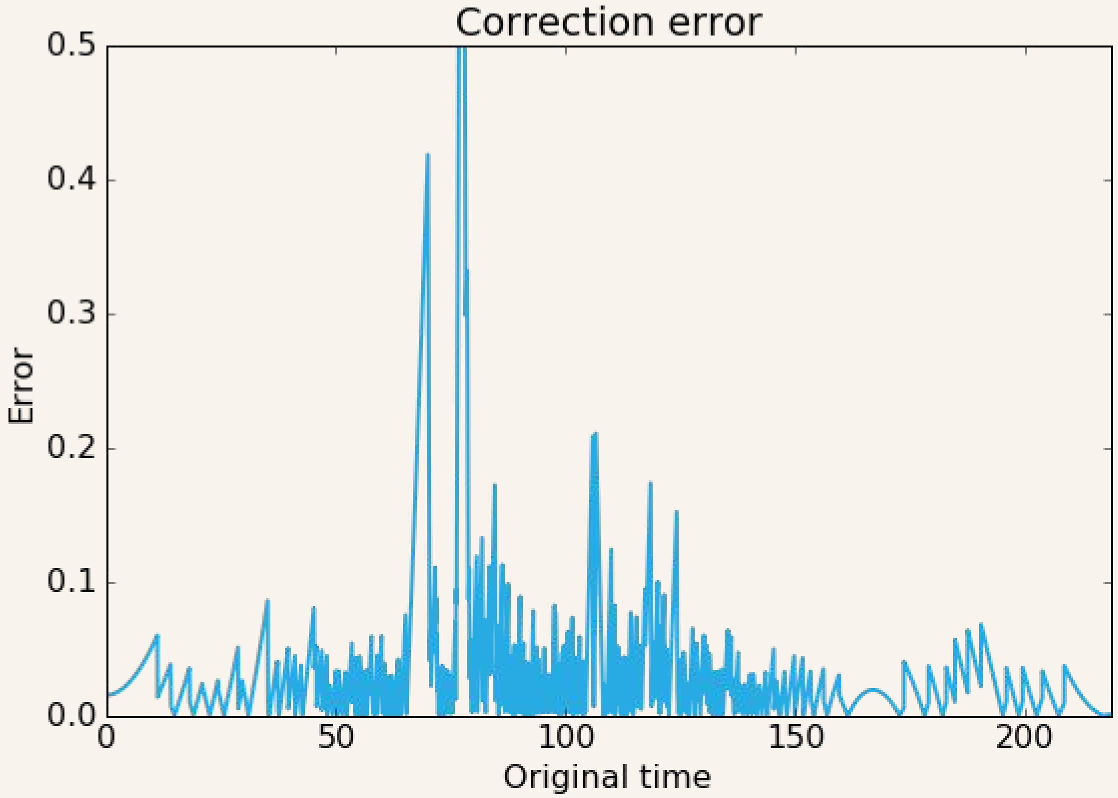
Artificial time warping



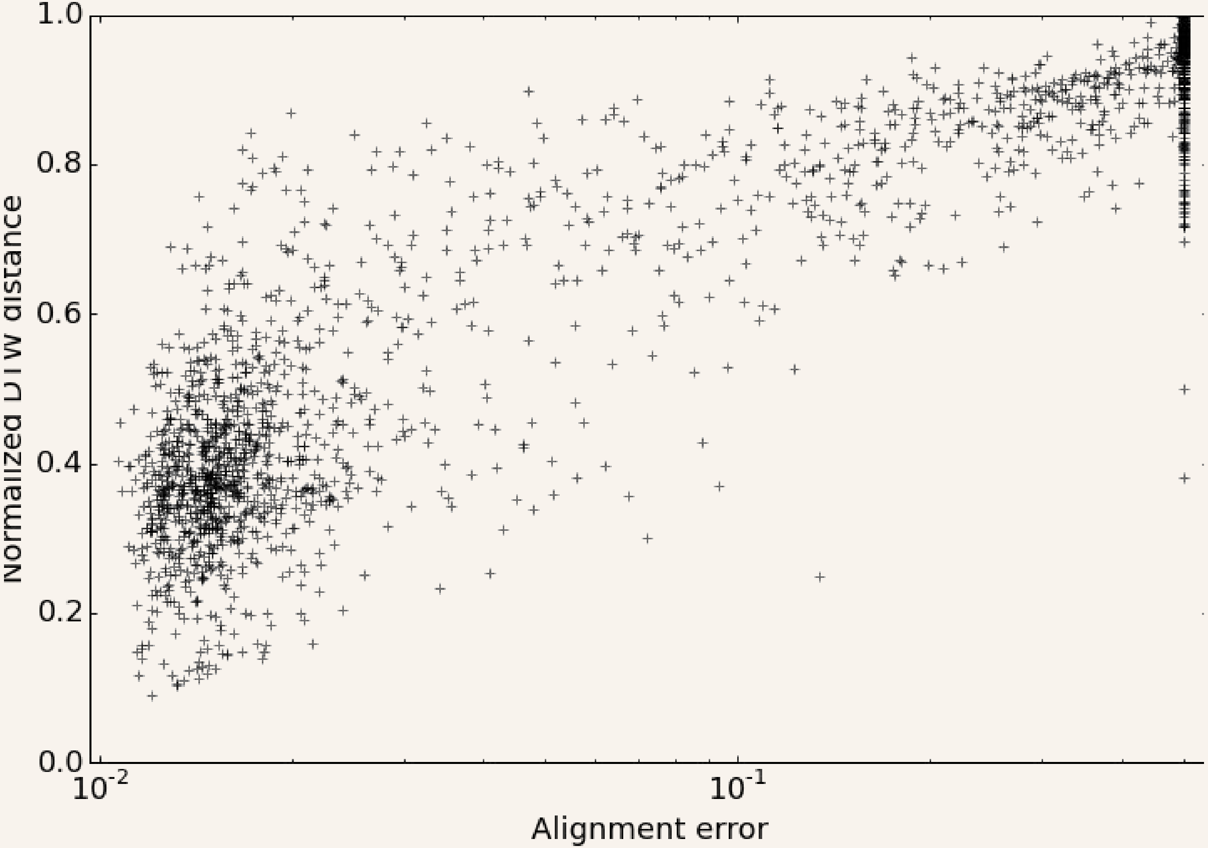
Correcting time warping



Measuring error



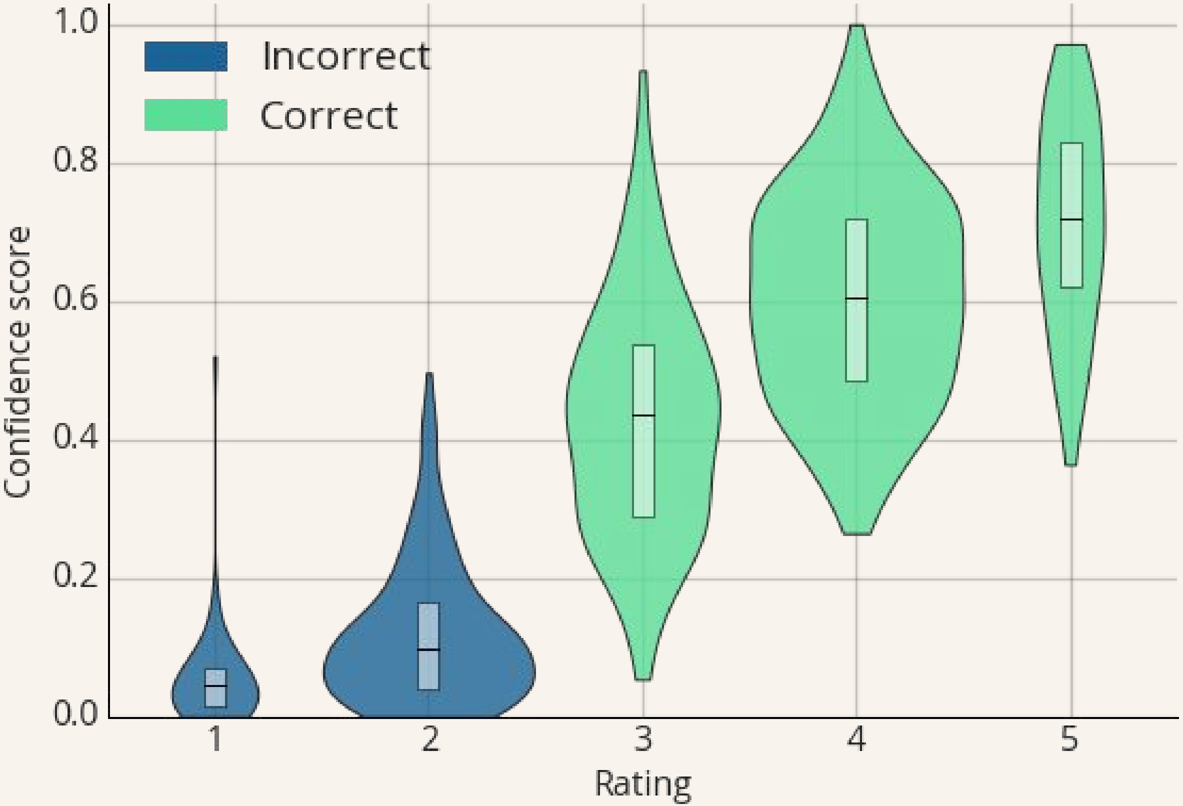
Score normalization search



Best system

- Use log-magnitude constant-Q spectrograms
- Don’t beat synchronize
- L2 normalize spectra (cosine distance)
- Don’t z-score spectrograms
- Use median distance as non-diagonal penalty
- Force sequences to match up to 96% of shorter
- Don’t use a band path constraint
- Include penalties in confidence score
- Normalize by path length and submatrix mean
- Example implementation in `pretty_midi` examples

Real-world test



DTW is too slow

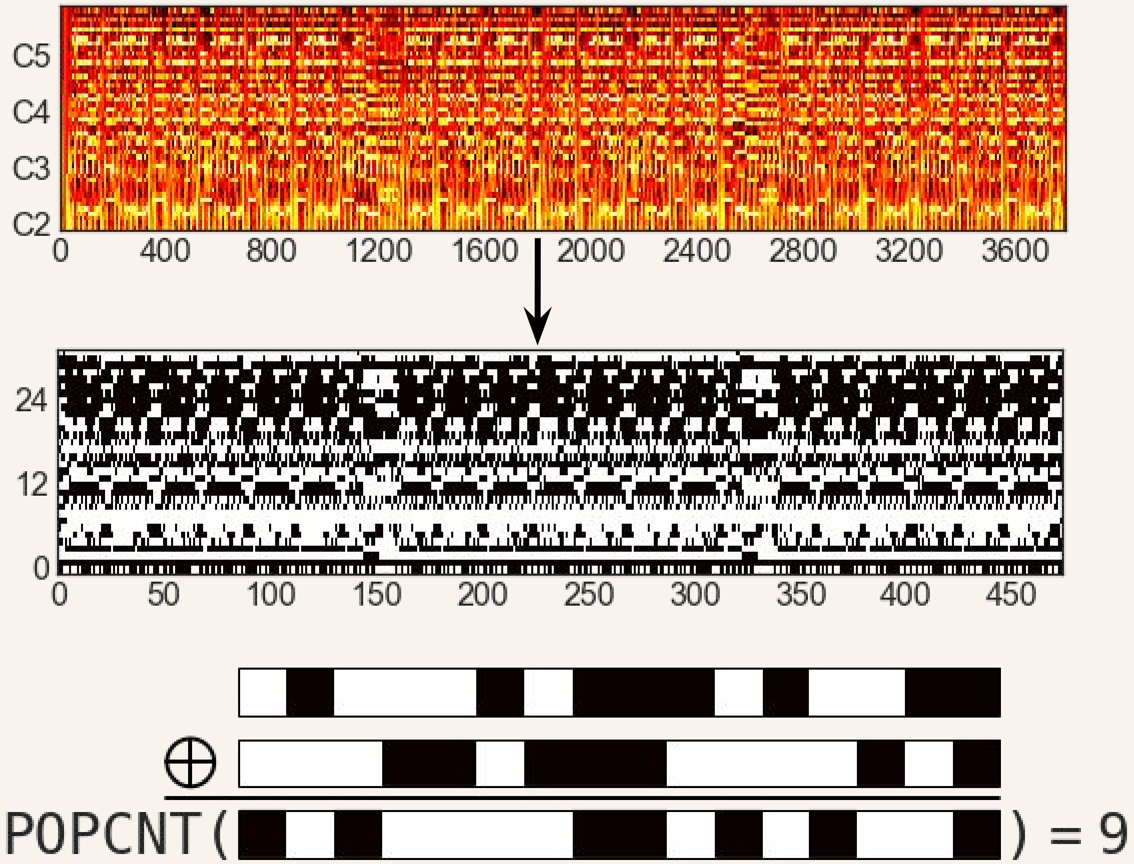
$$\frac{(.247 \text{ seconds})(178{,}561 \text{ MIDI files})(994{,}960 \text{ audio files})}{(60 \text{ seconds})(60 \text{ minutes})(24 \text{ hours})(365 \text{ days})}$$

DTW is too slow

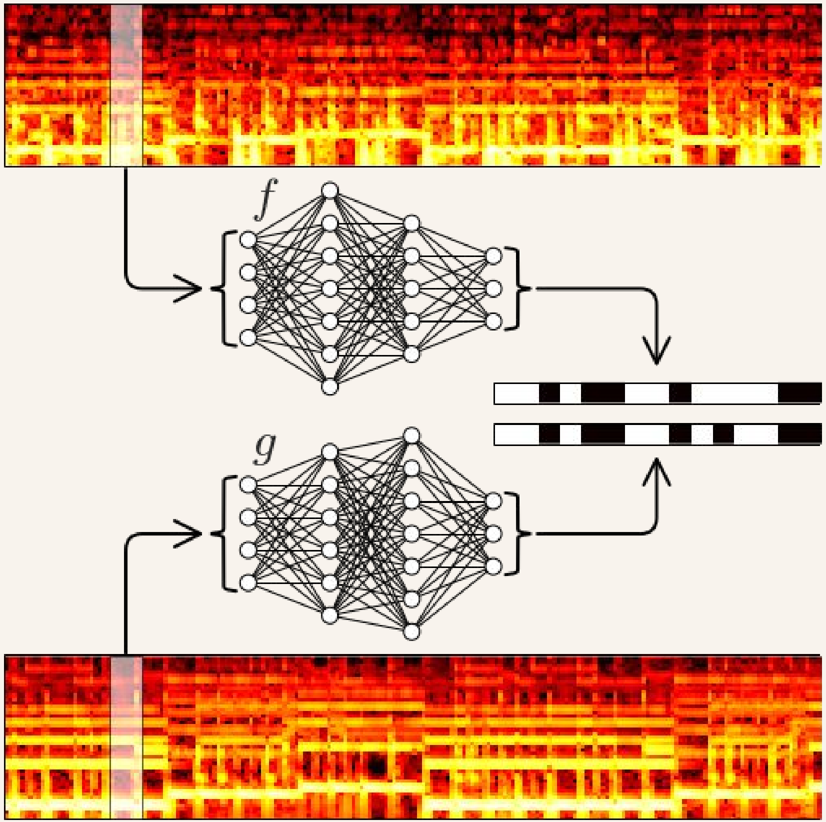
$$\frac{(.247 \text{ seconds})(178{,}561 \text{ MIDI files})(994{,}960 \text{ audio files})}{(60 \text{ seconds})(60 \text{ minutes})(24 \text{ hours})(365 \text{ days})}$$

$$\approx 1{,}391 \text{ years}$$

Downsampled Hash Sequences



Similarity-preserving hashing



Collecting data

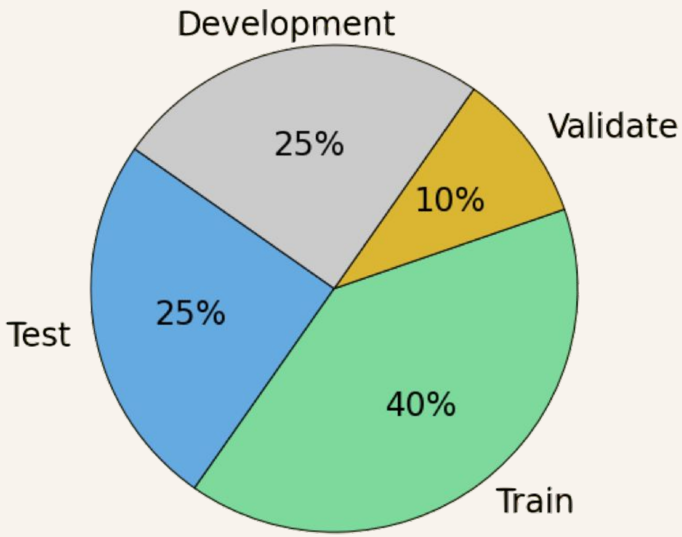
140,910

24,850

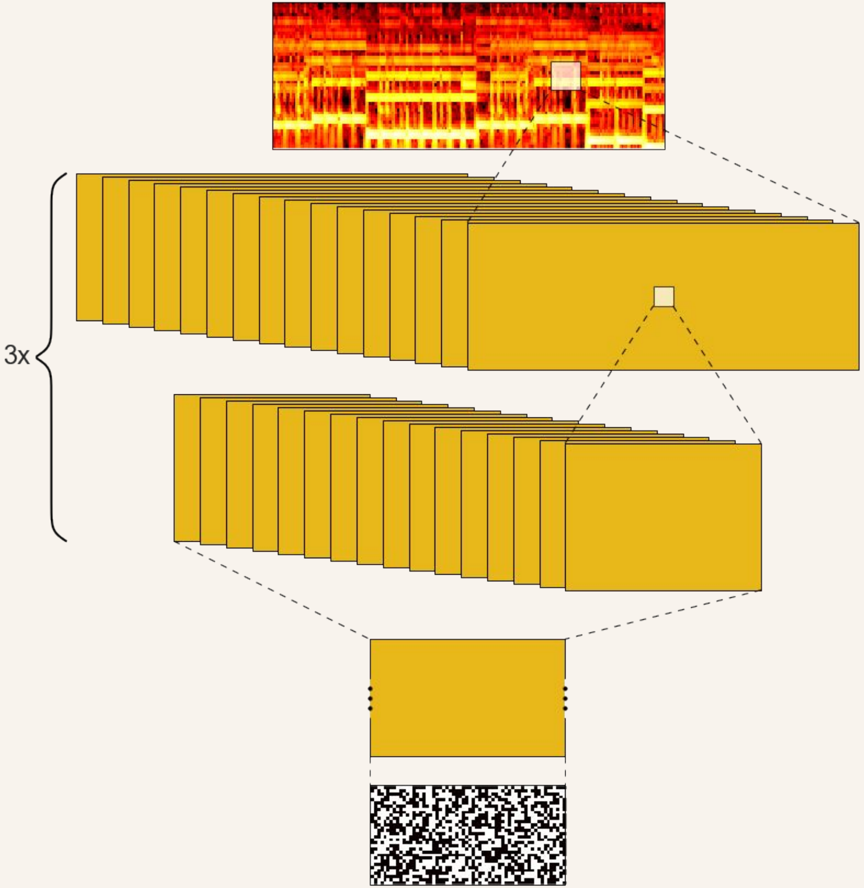
17,243

26,311

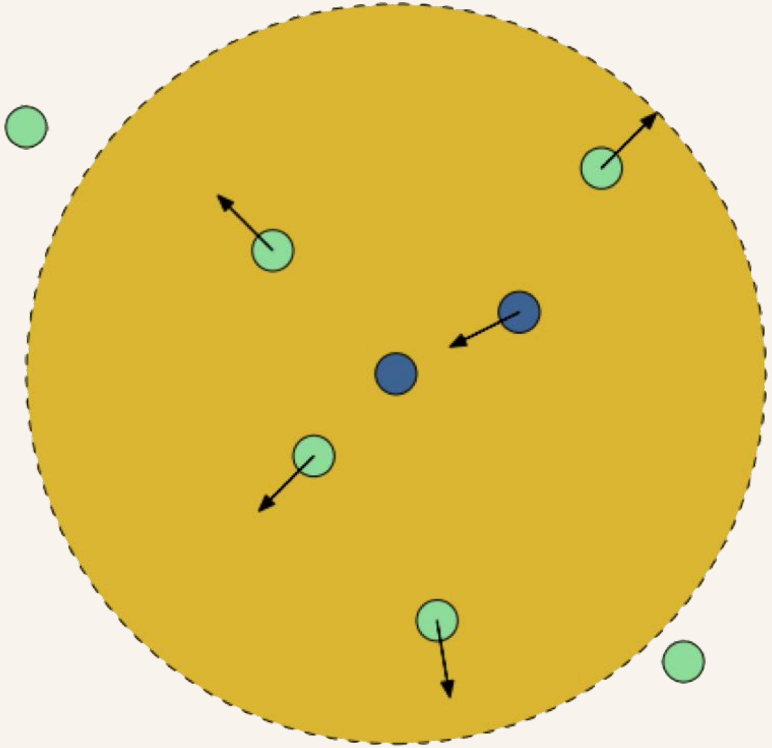
10,035



Network structure

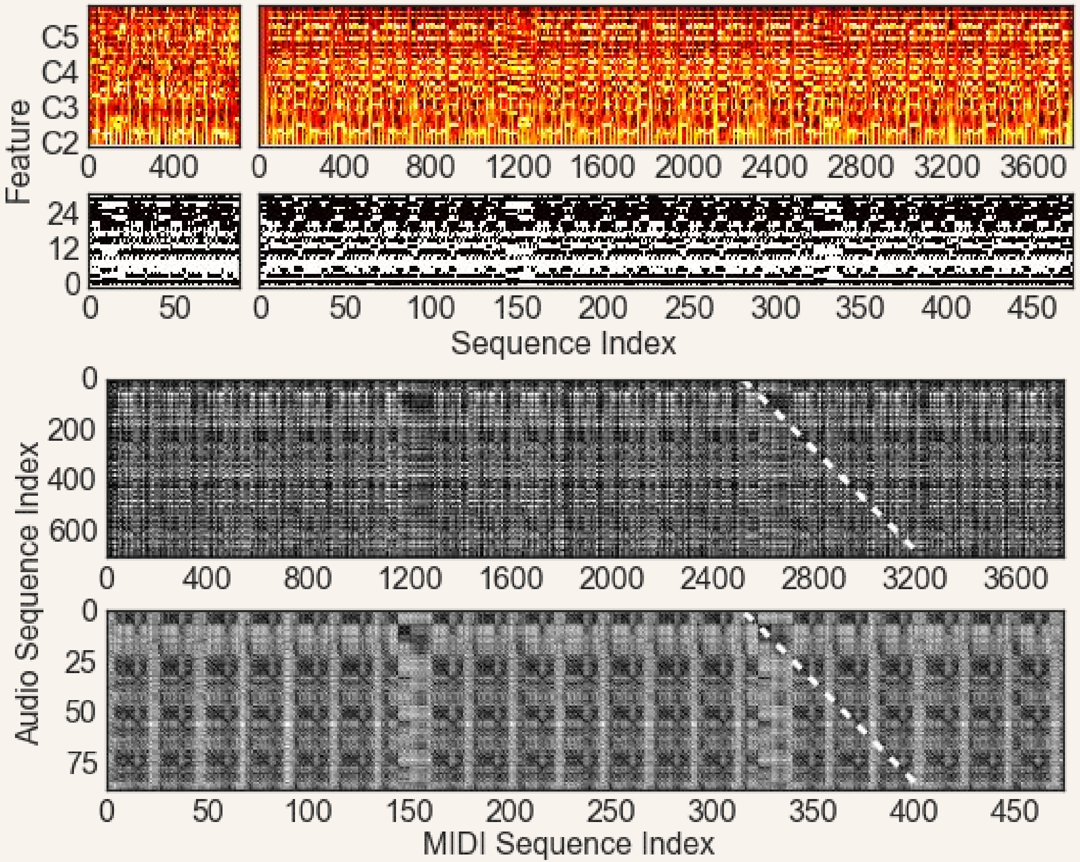


Loss function

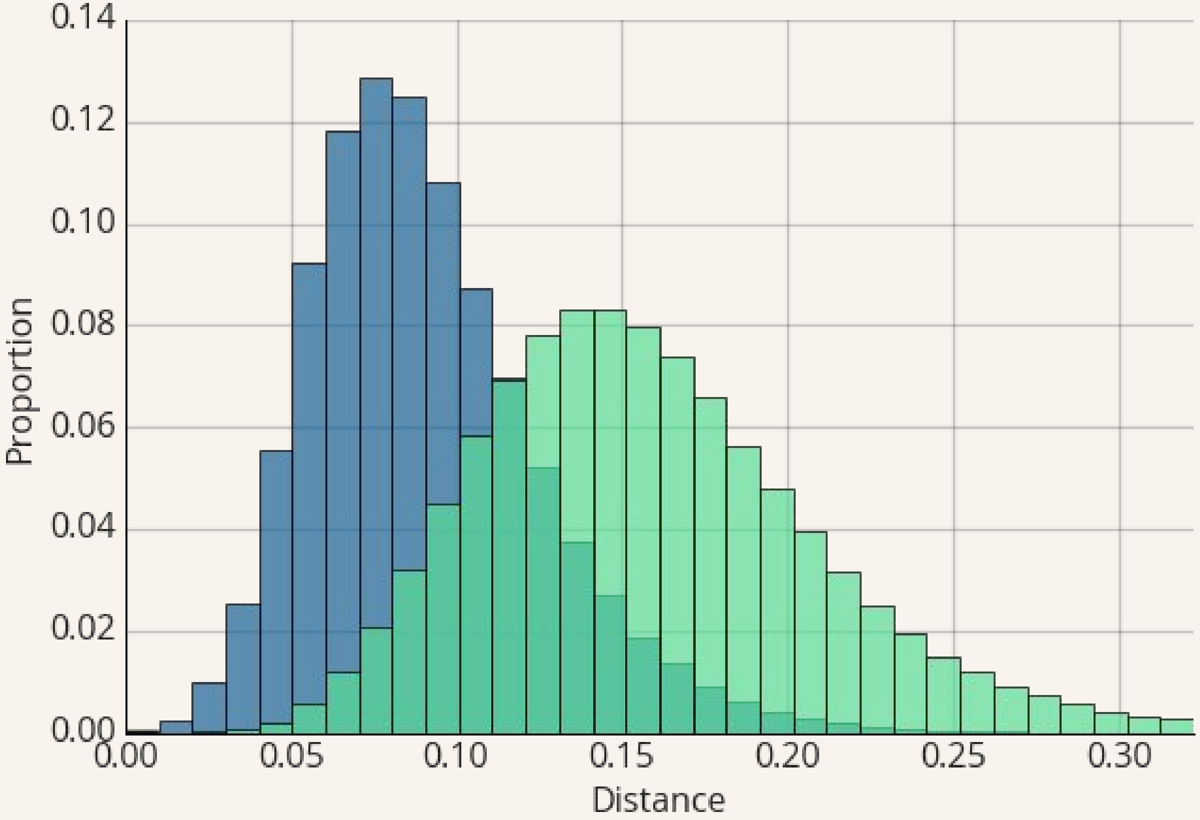


$$\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(x,y)\in\mathcal{P}} \|f(x) - g(y)\|_2^2 + \frac{\alpha}{|\mathcal{N}|} \sum_{(x,y)\in\mathcal{N}} \max(0, m - \|f(x) - g(y)\|_2)^2$$

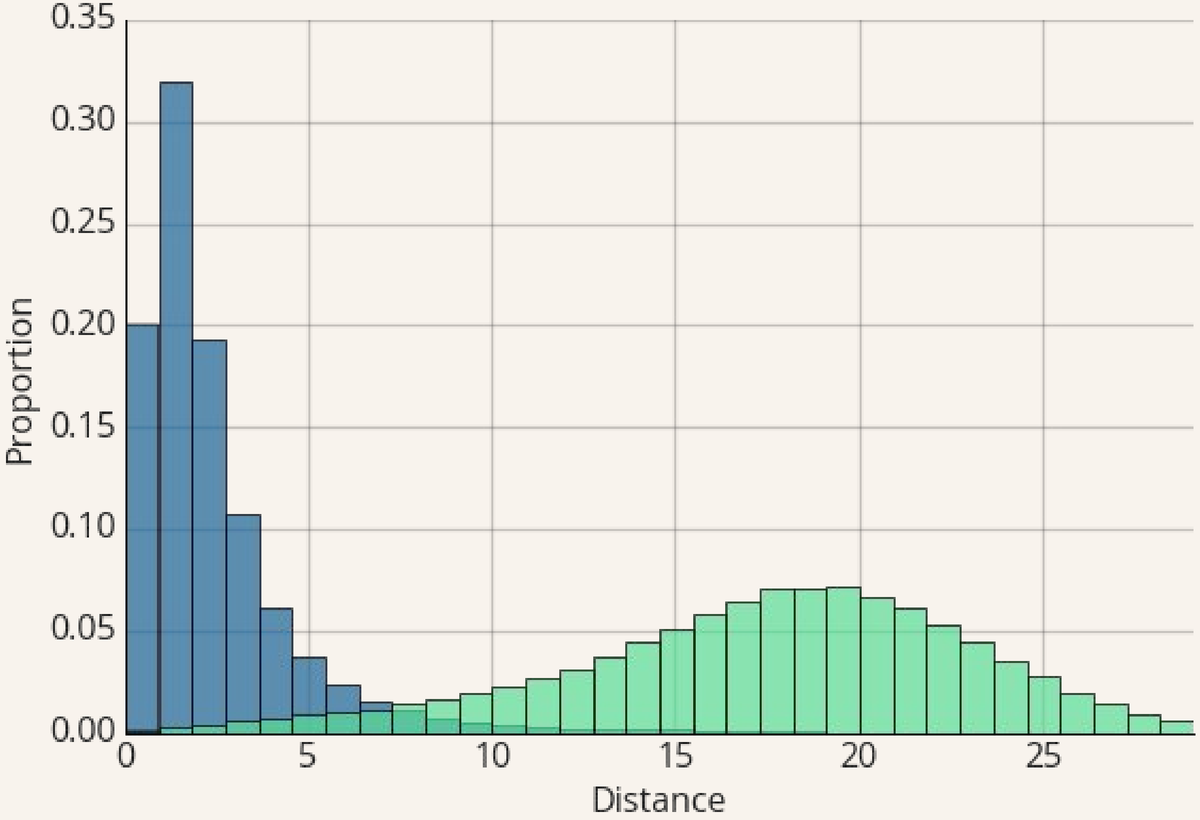
Example output



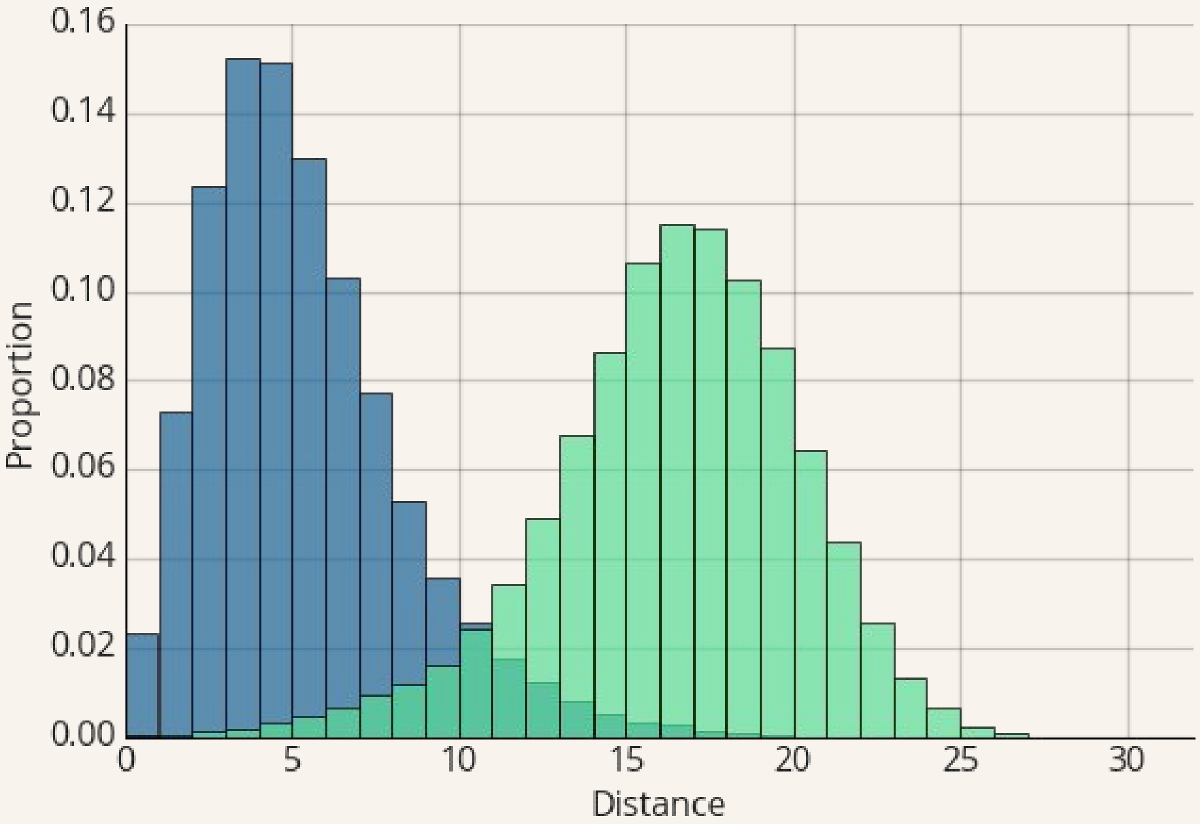
Raw distance distributions



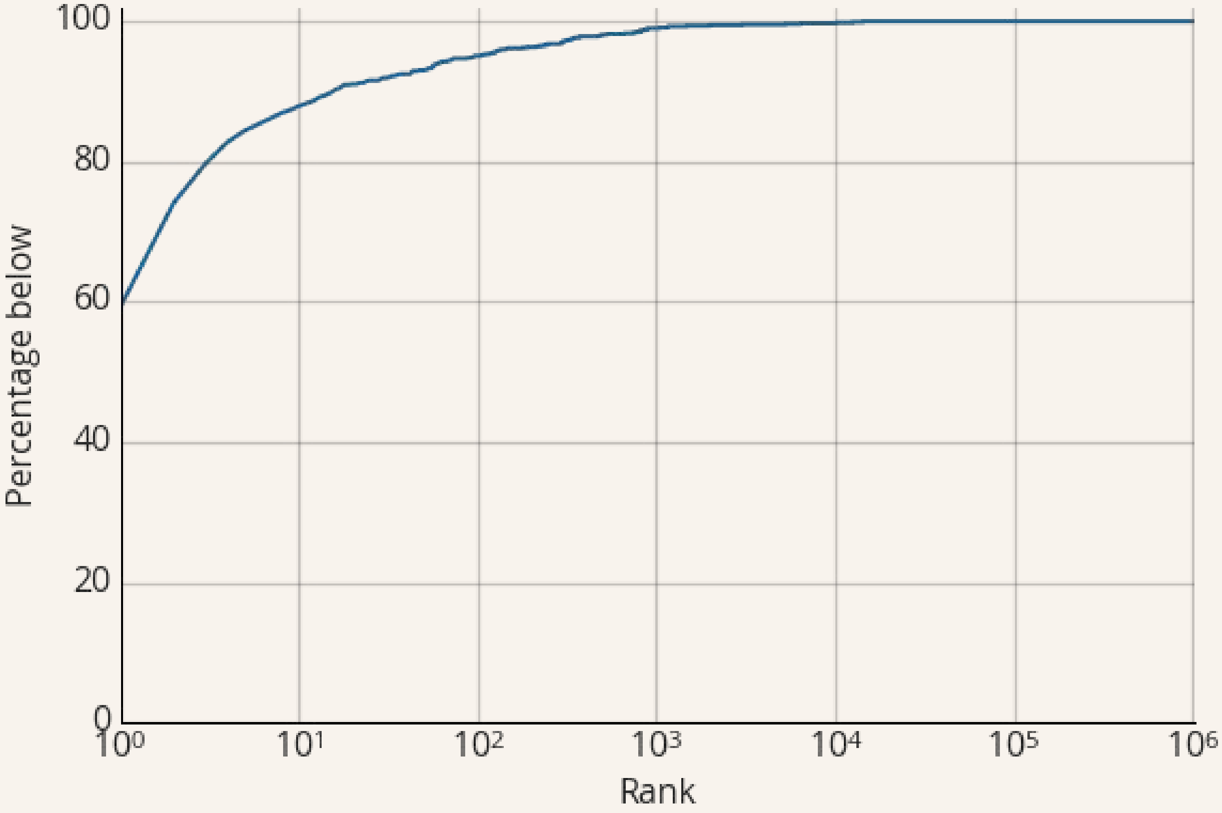
Network output distance distributions



Hash distance distributions

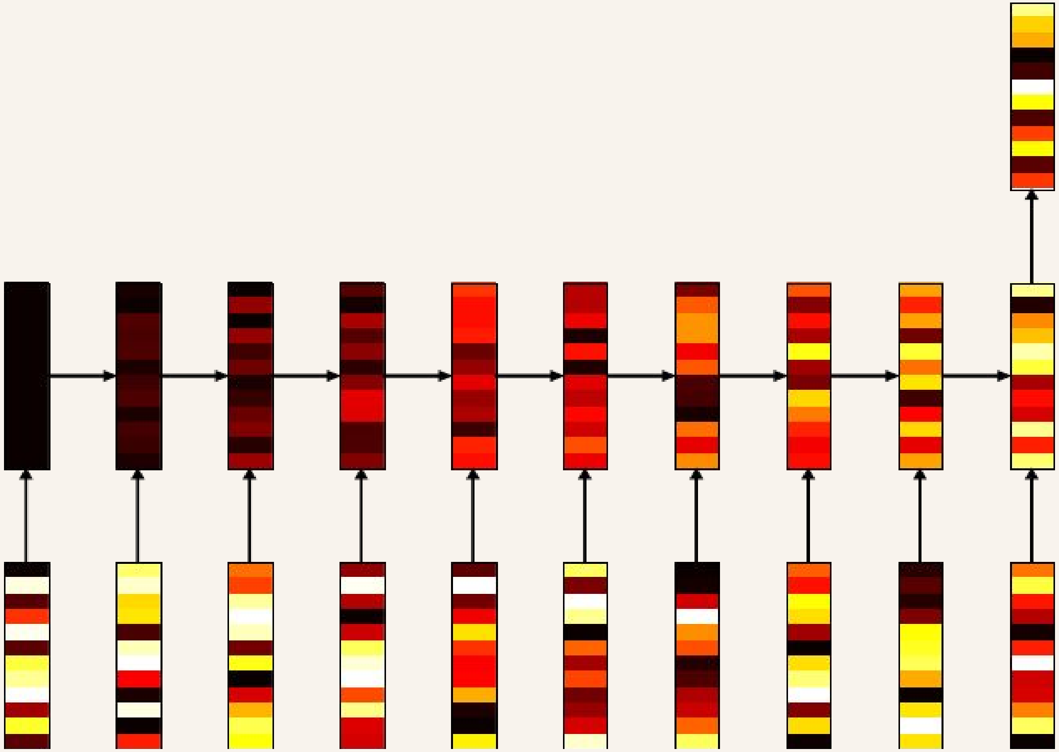


Match ranks

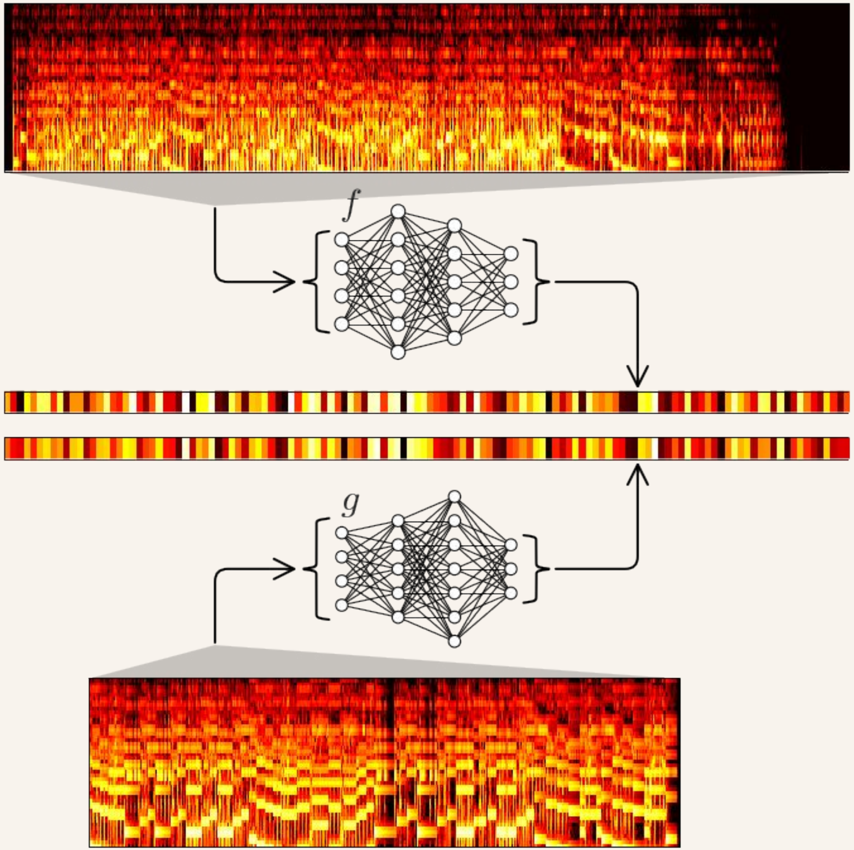


Percentage below
100
80
60
40
20
0
10^0
10^1
10^2
10^3
10^4
10^5
10^6
Rank

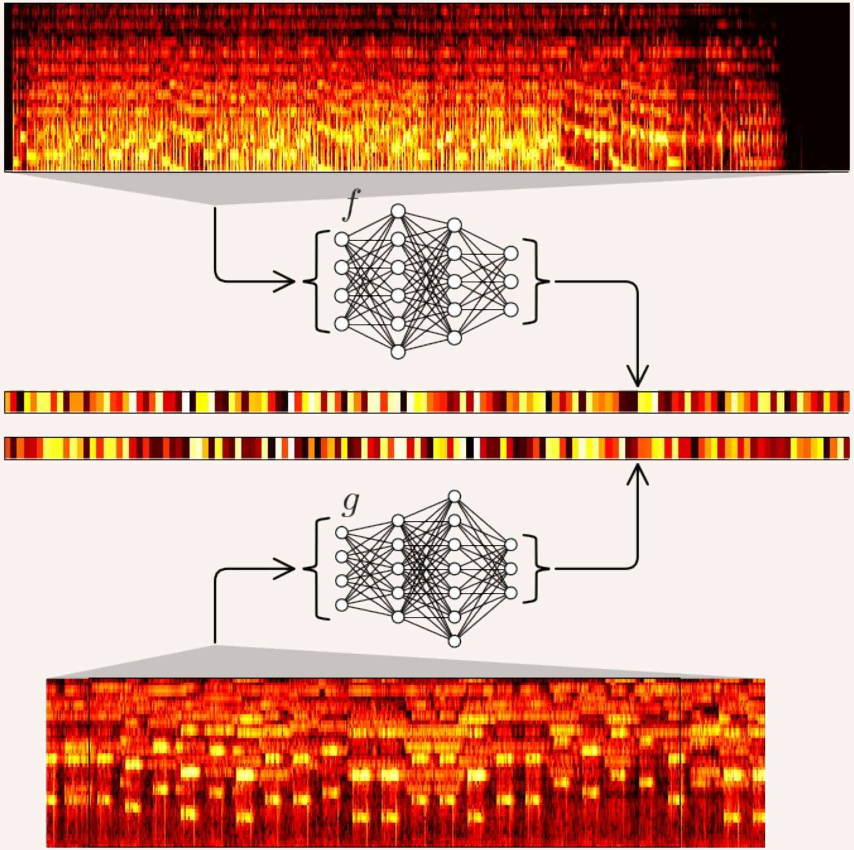
Sequence embedding



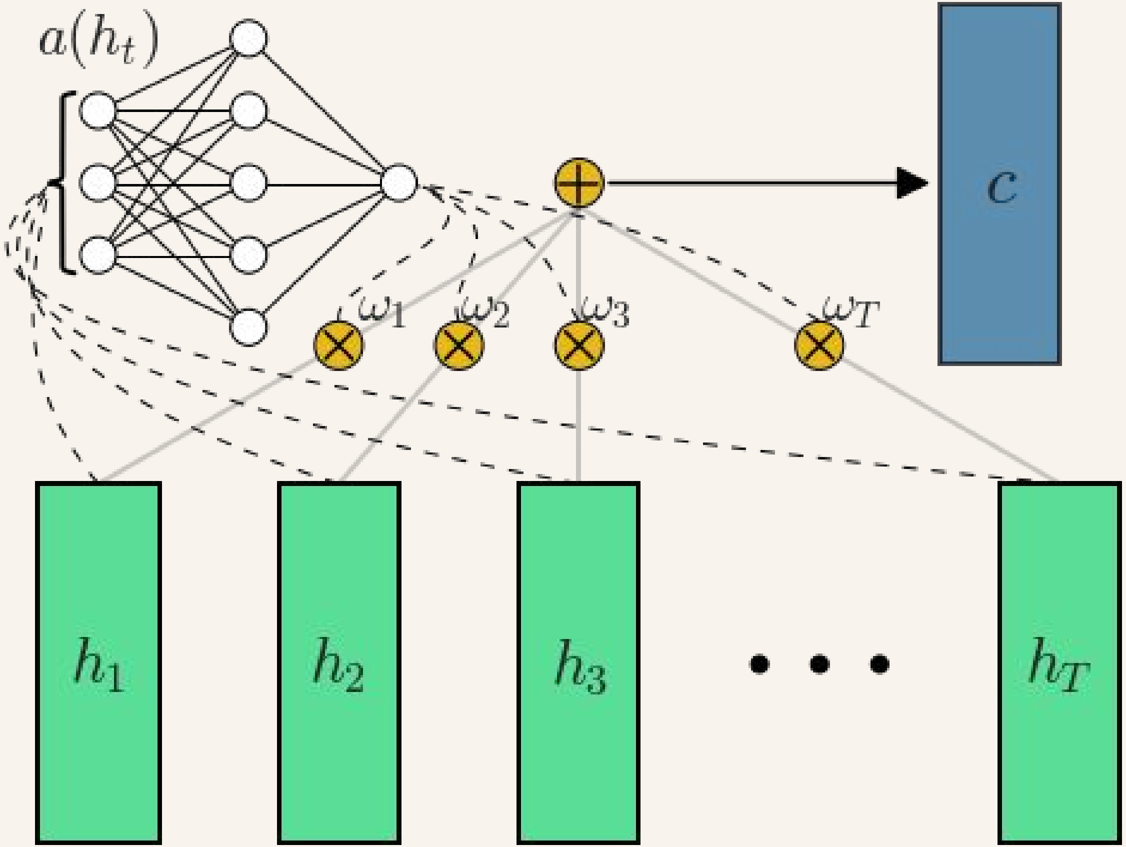
Pairwise sequence embedding



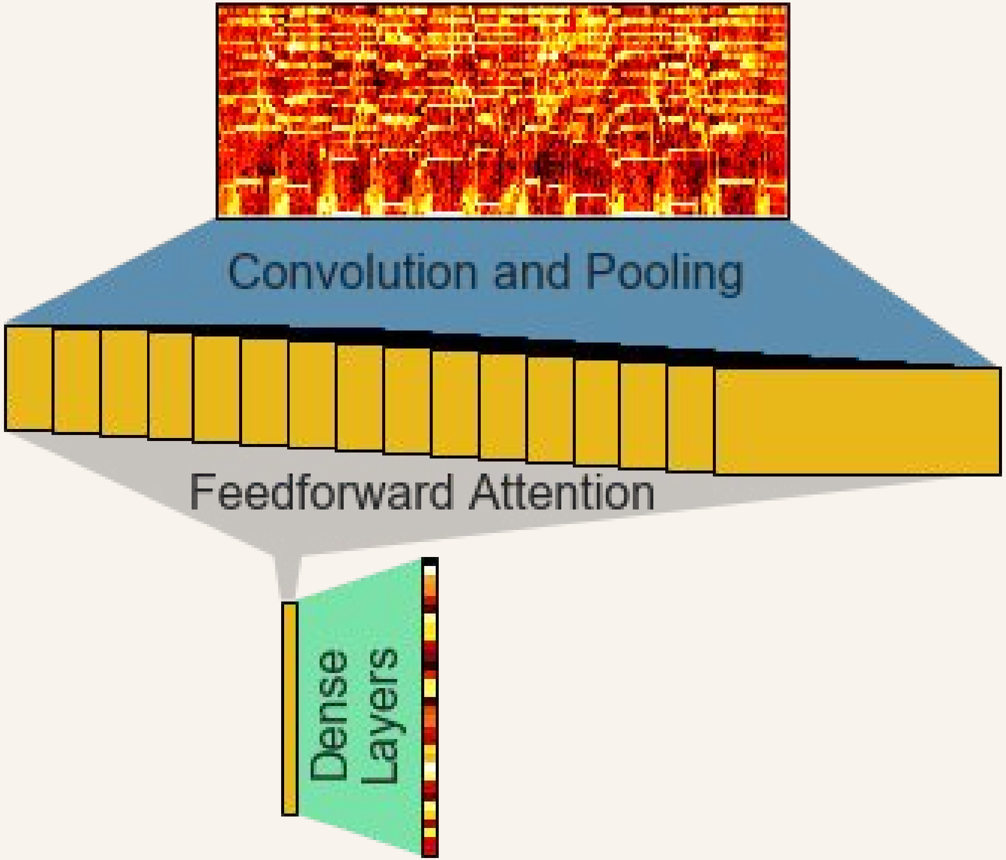
Pairwise sequence embedding



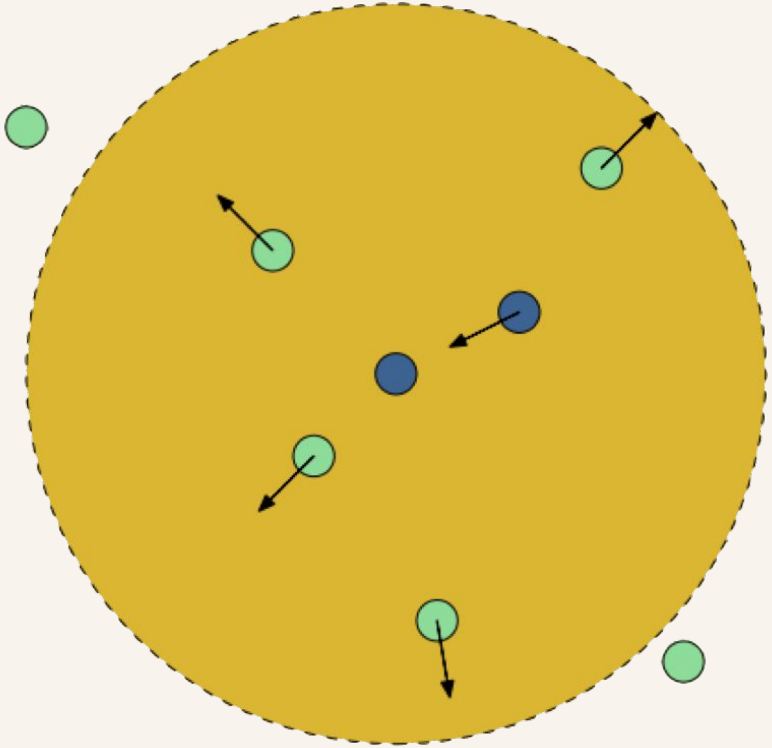
Feedforward attention



Embedding network

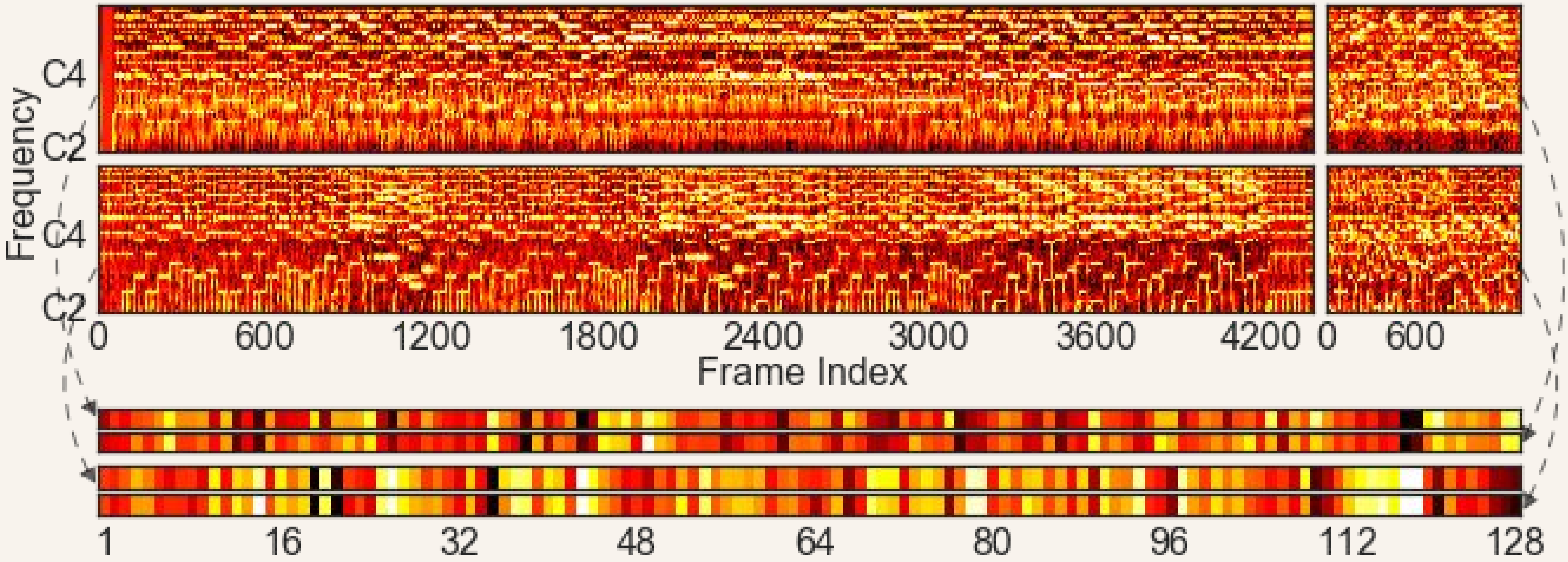


Loss function

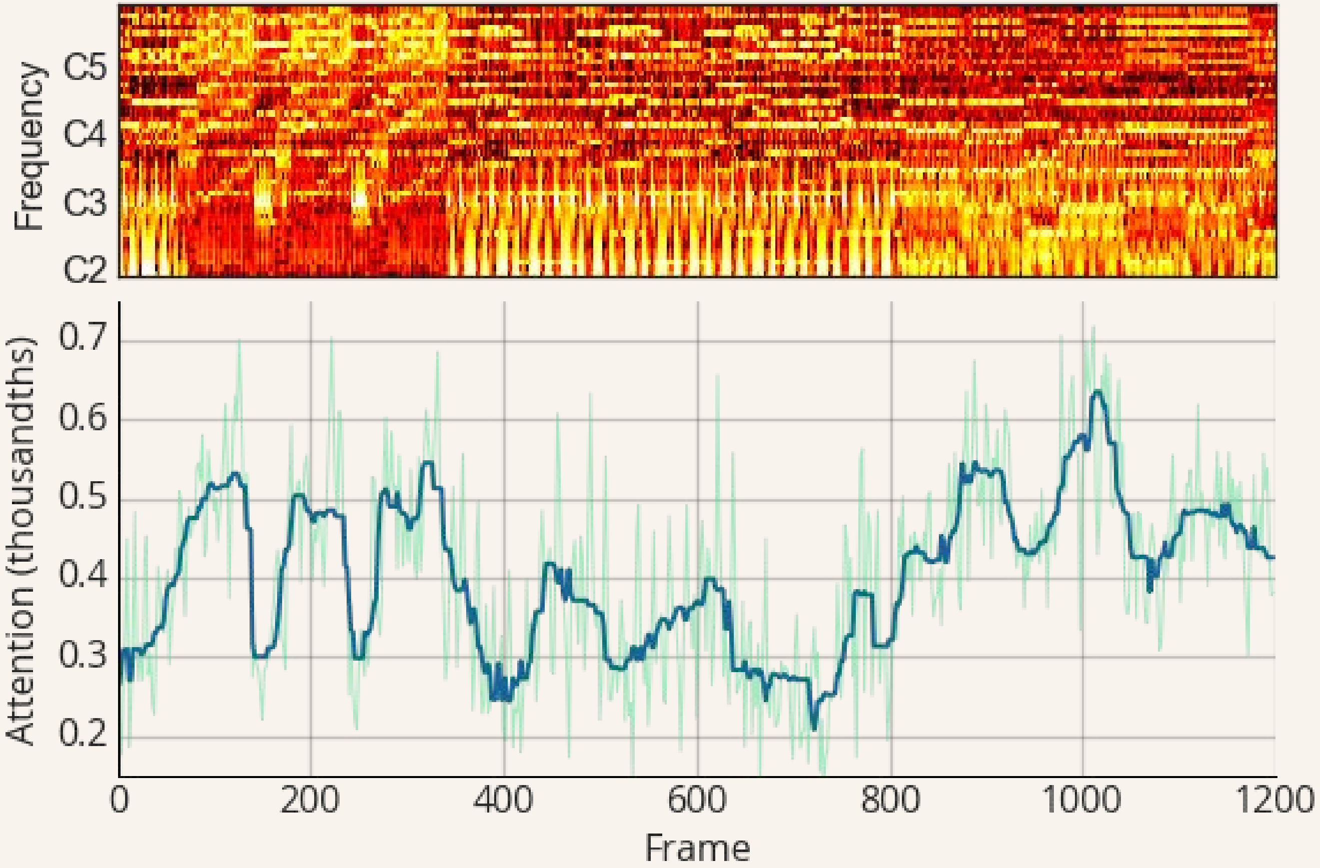


$$\mathcal{L} = \frac{1}{|\mathcal{P}|} \sum_{(x,y)\in\mathcal{P}} \|f(x) - g(y)\|_2^2 + \frac{\alpha}{|\mathcal{N}|} \sum_{(x,y)\in\mathcal{N}} \max(0, m - \|f(x) - g(y)\|_2)^2$$

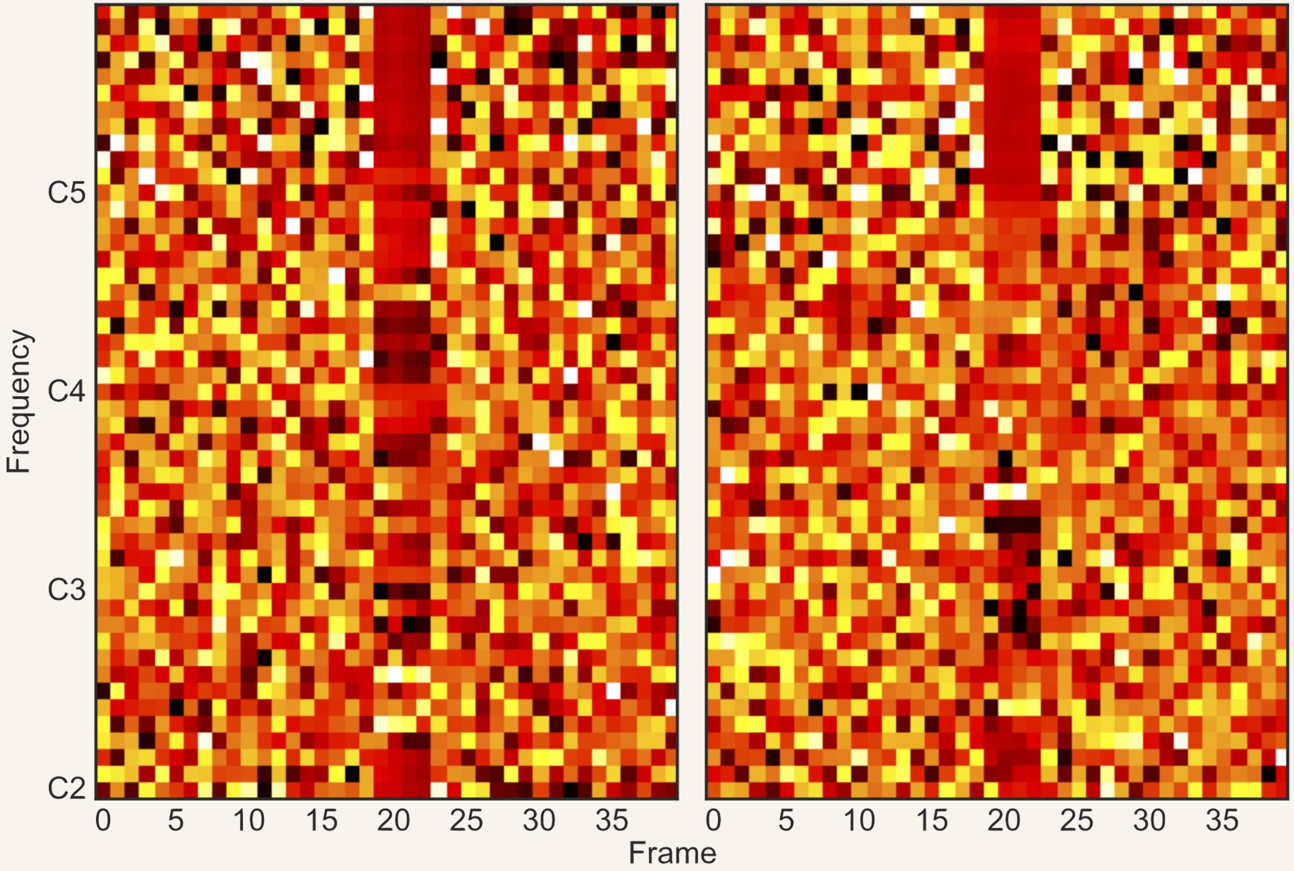
Example embeddings



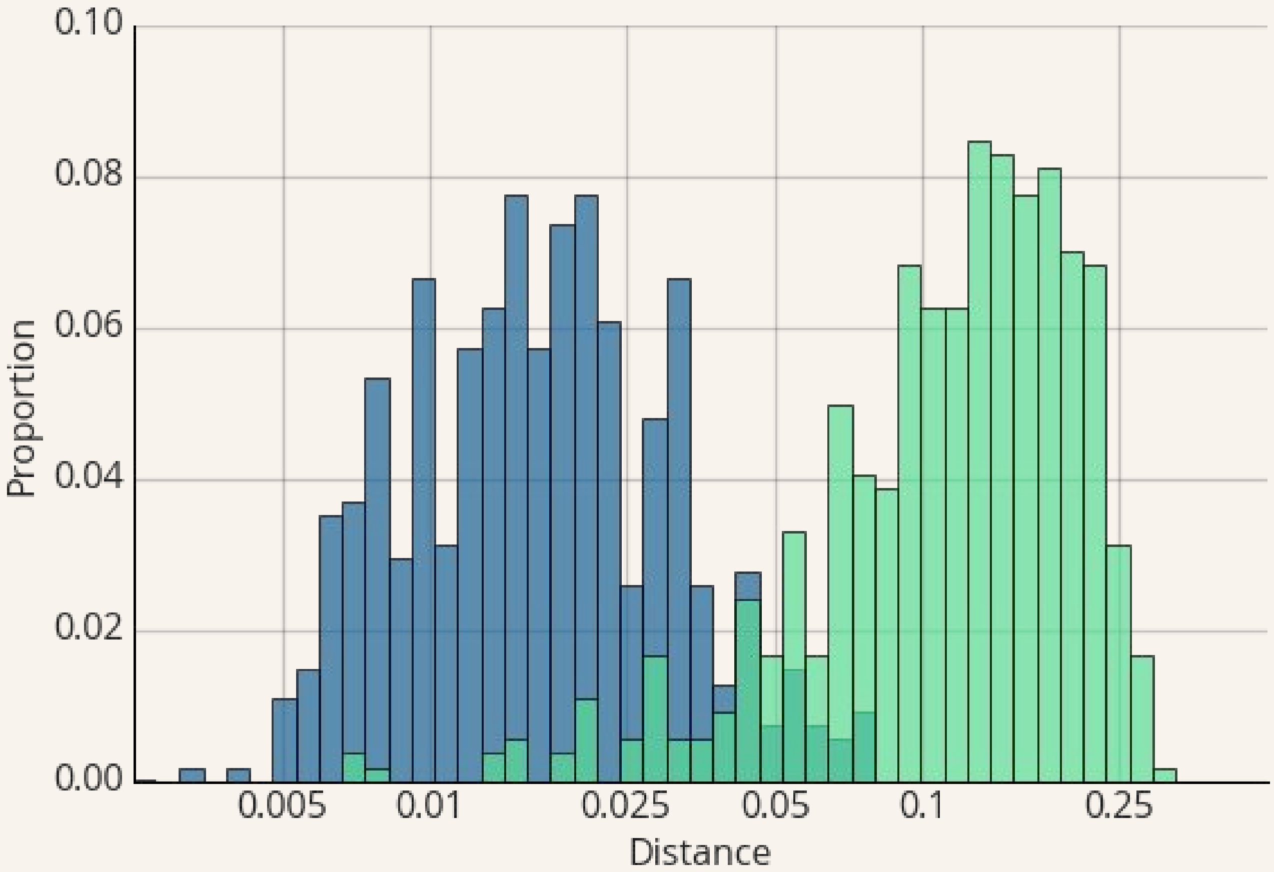
Example attention



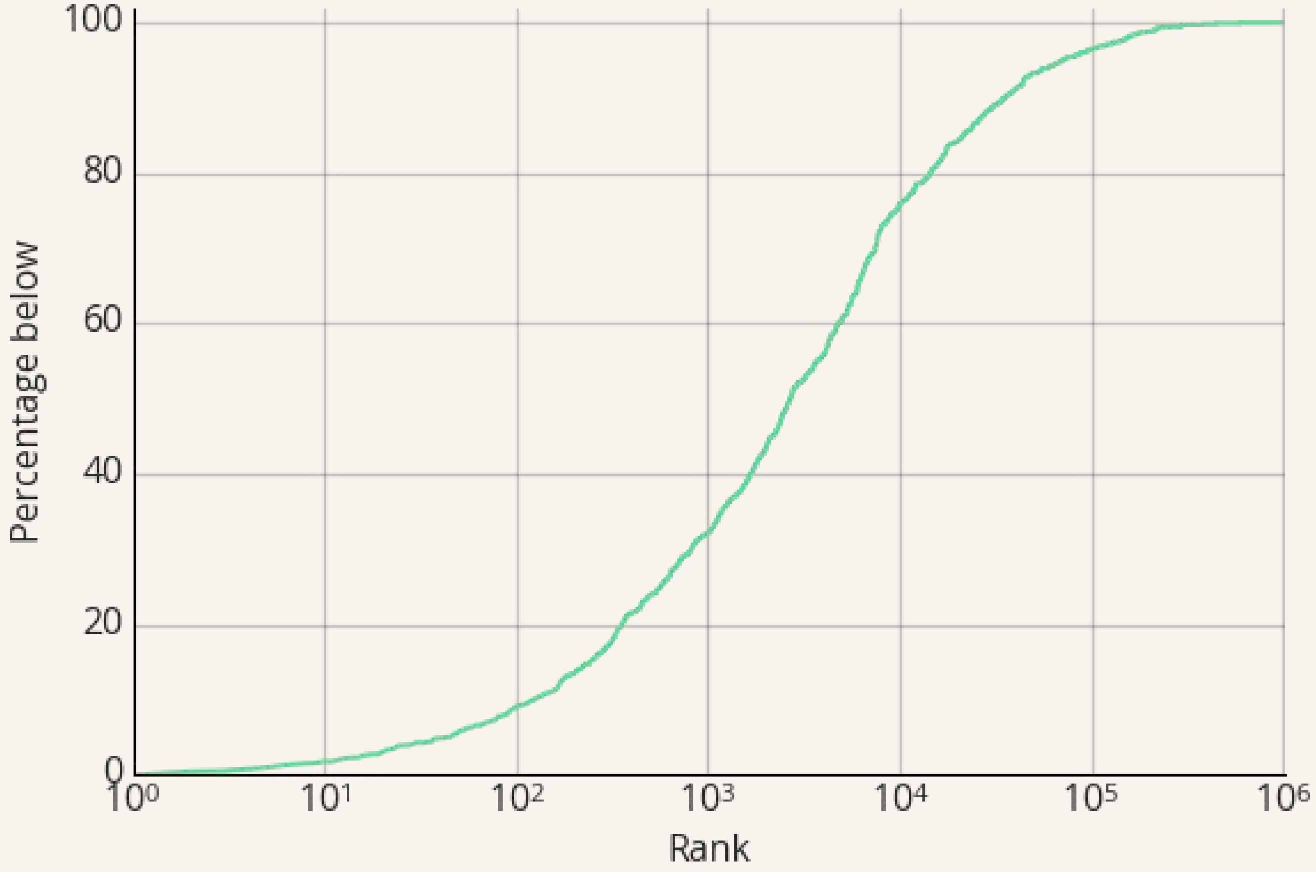
Attention dreaming



Embedding distances

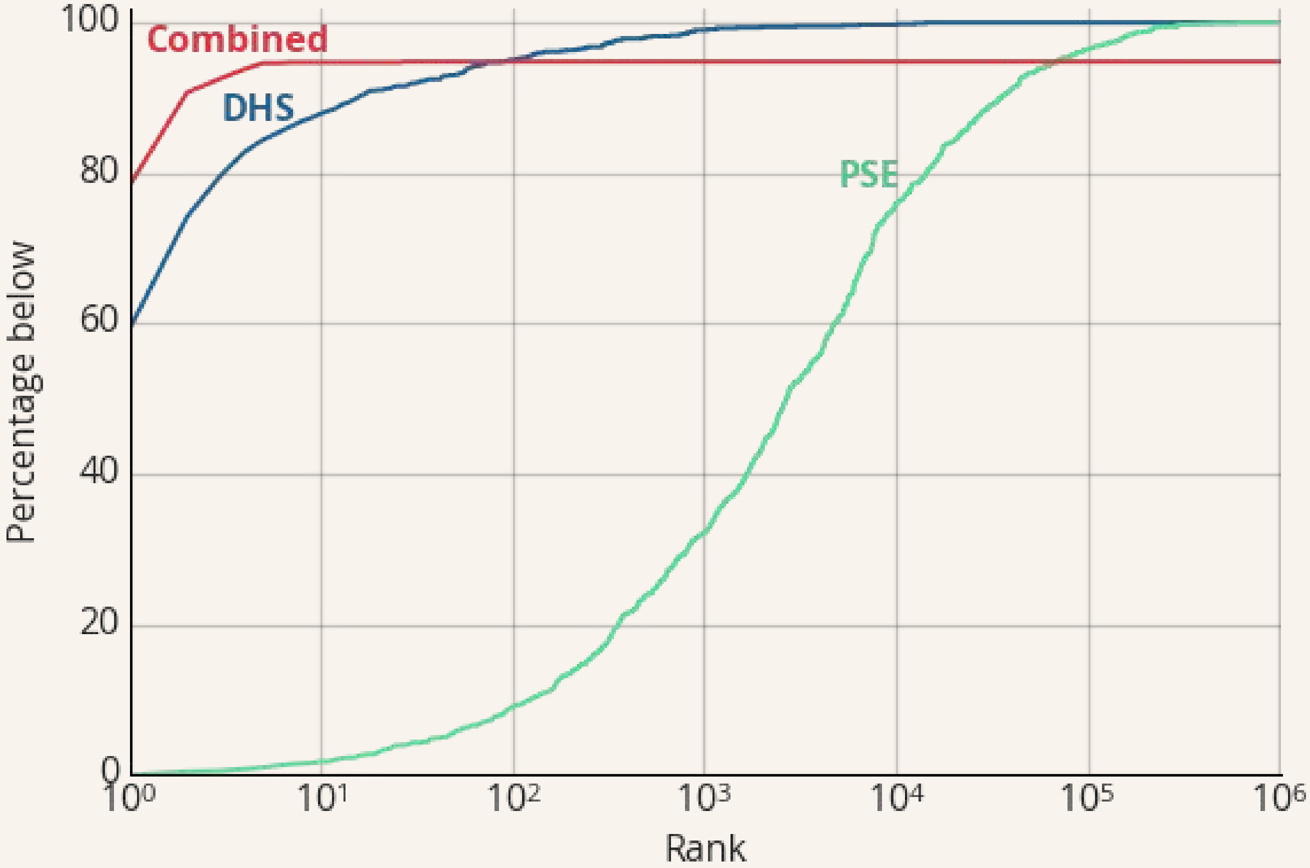


Match ranks

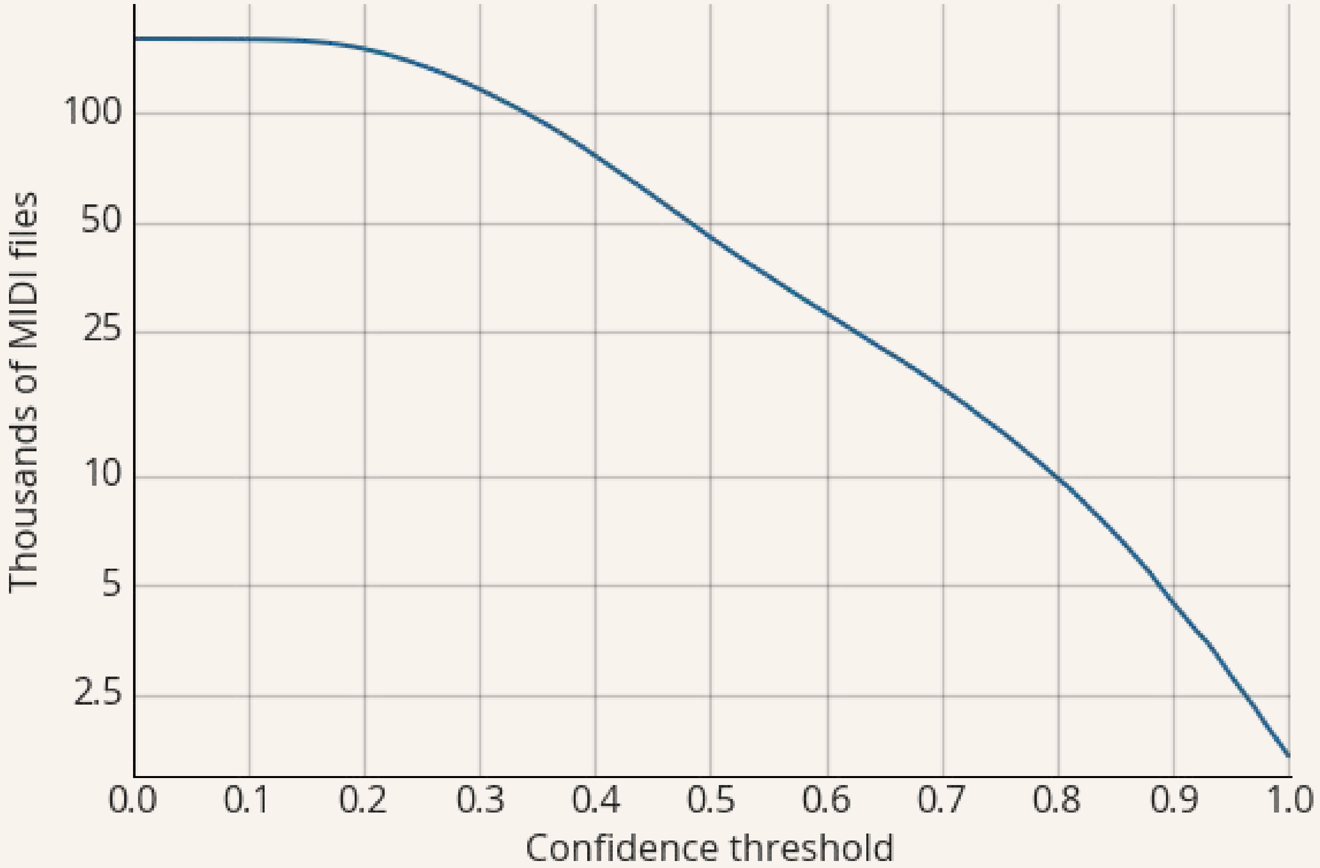


100
80
60
40
20
0
Percentage below
10^0
10^1
10^2
10^3
10^4
10^5
10^6
Rank

Combined ranks



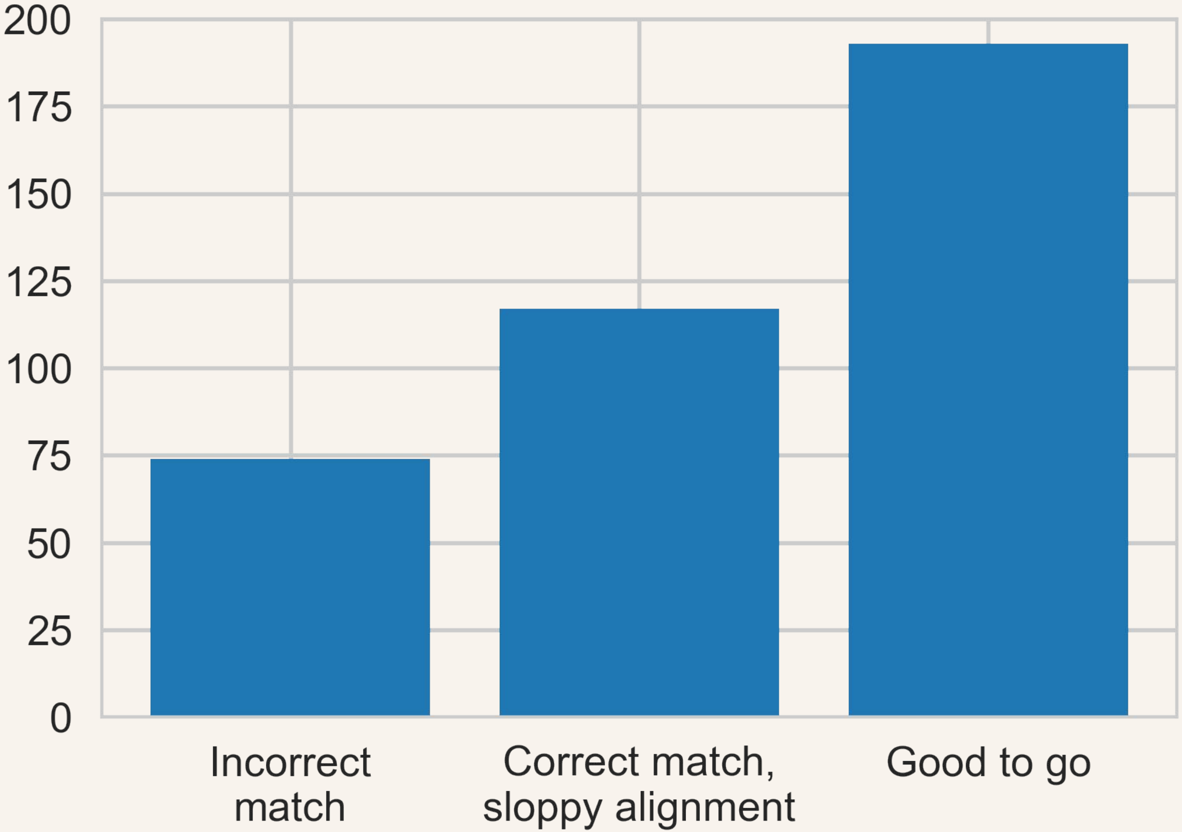
Number of matches



Lakh MIDI Dataset (mini) tutorial

`https://goo.gl/hU4GAK`

How well did we do?

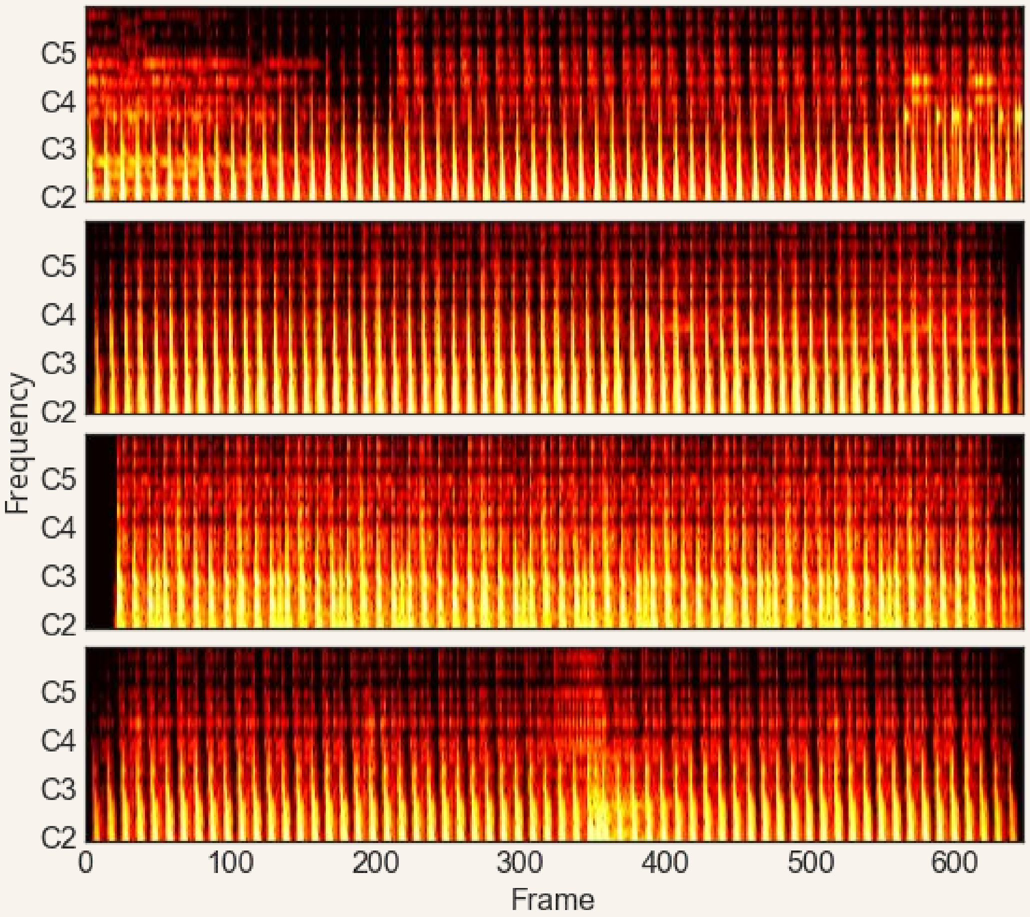


Decoys?

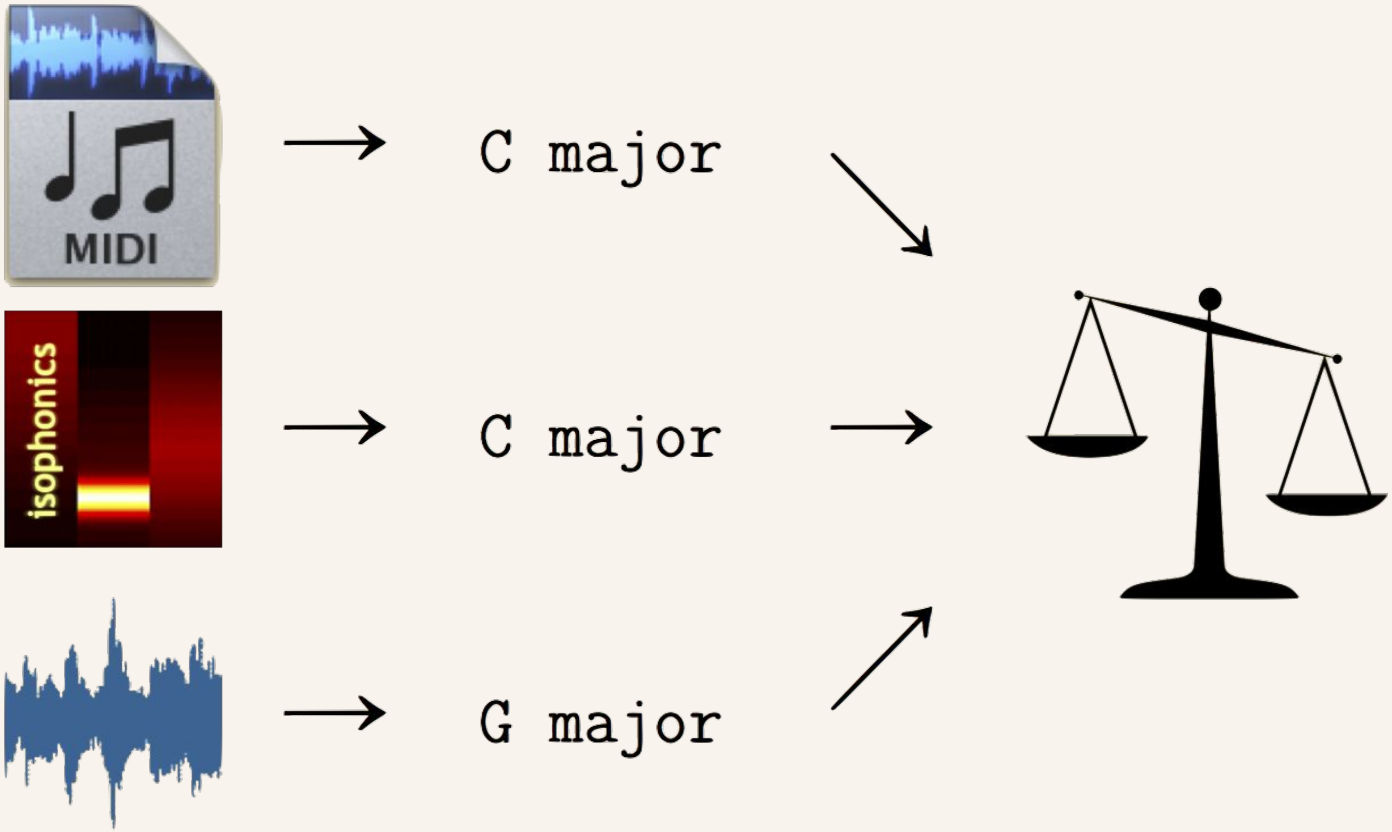
- Kaskade - In This Life (Mario Fabriani Remix)
- Marcos Hernandez - If You Were Mine
- D-Unity - Afrika
- Anane - Let’s Get High (Yves C Mix)
- DJ Silver - Wardance
- Ann Nesby - So Much Joy (Praise Party Beats)
- Kim English - My Destiny (Kobbe & Austin Leeds Club Mix)
- Gianluca Motta / Snap! / NG3 - Ooops Up
- Dave Clarke - Protective Custody
- Karizma - Ride (Original Mix)
- Logic - The Warning (Claude Monnet & Torre Bros Main Mix)
- Ultra Nat - Automatic (Shawn Q’s Soltribe Vocal Mix)

More examples

Decoys



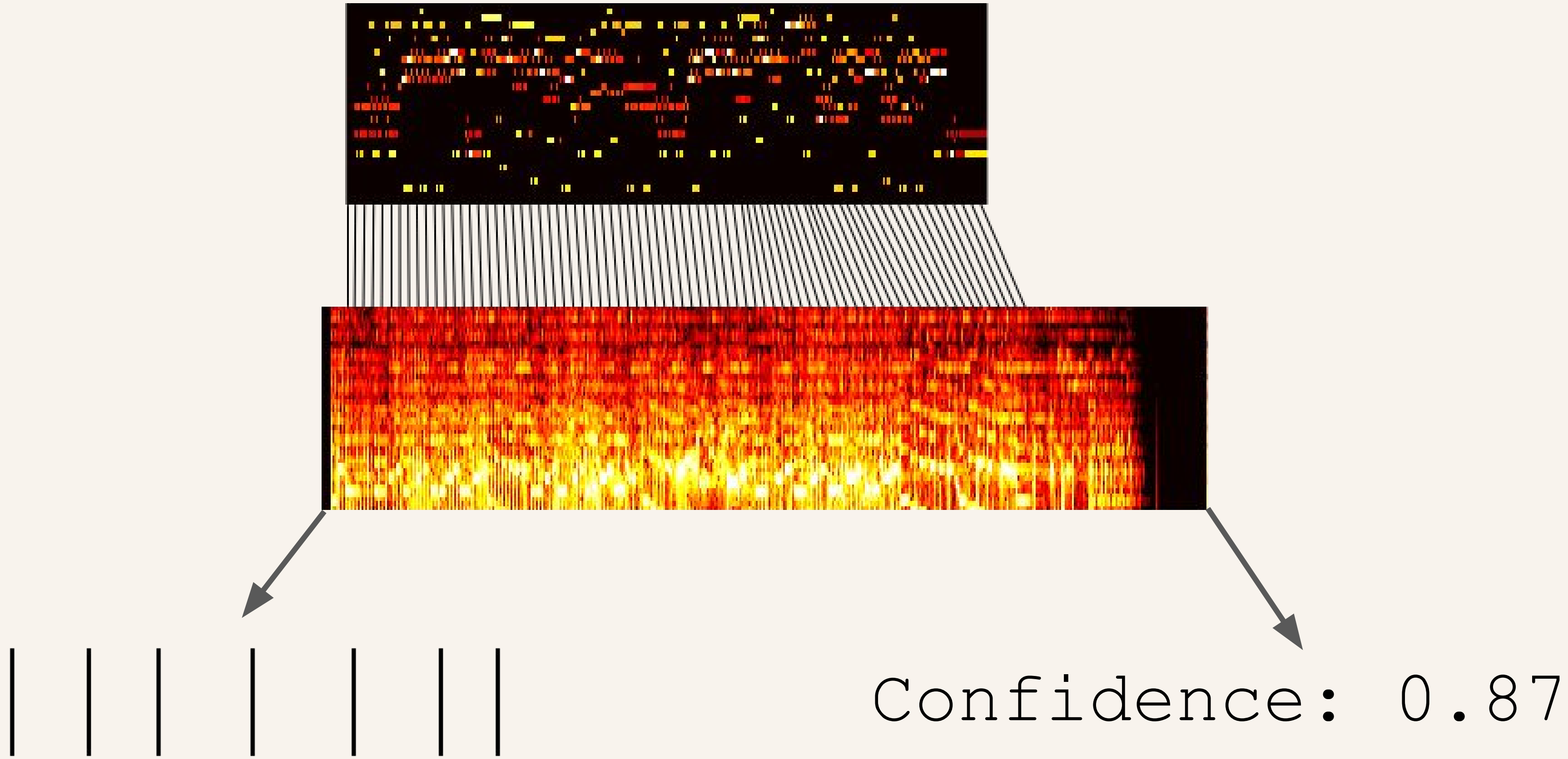
Testing key reliability



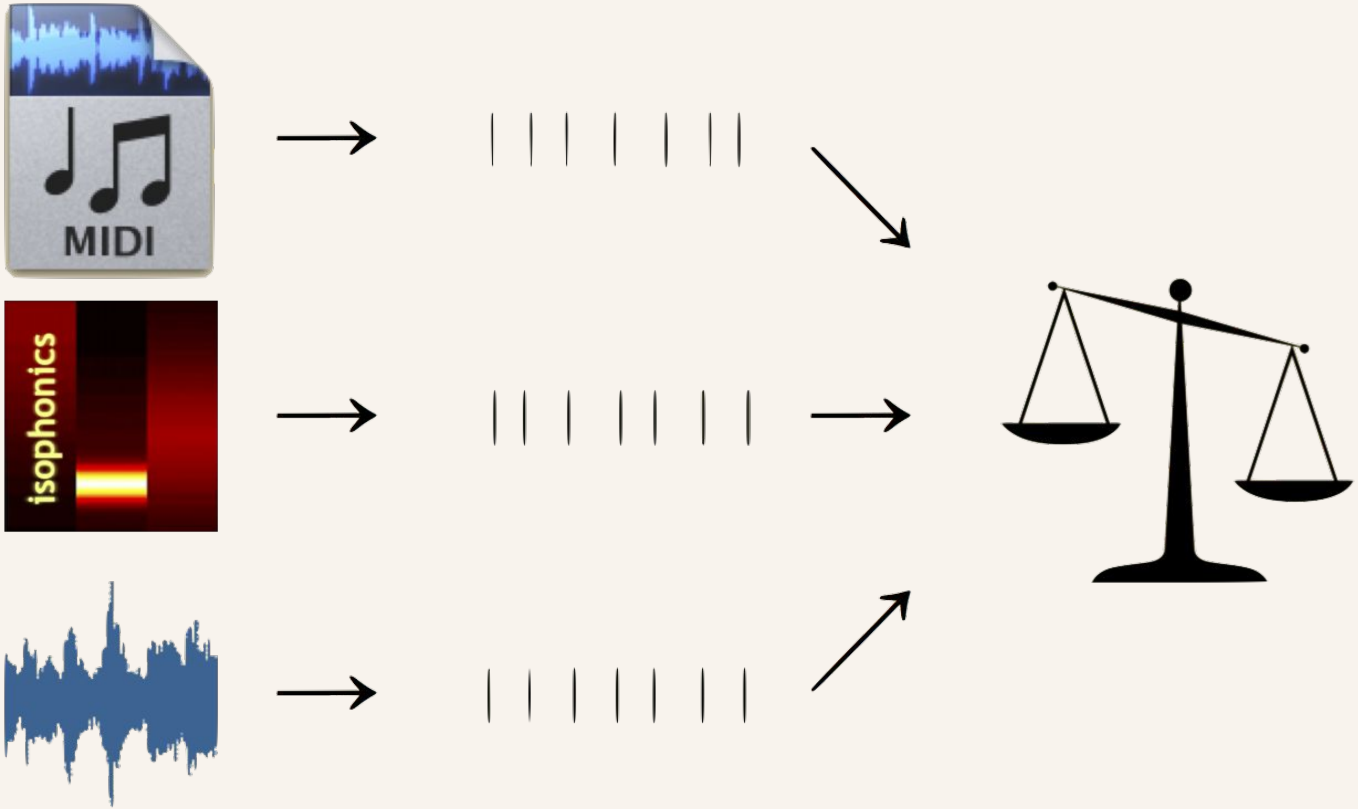
Key Results

Source	Score	Comparisons
MIDI, all keys	0.400	223
MIDI, C major only	0.167	146
MIDI, non-C major	0.842	77
Audio content-based	0.687	151
Human Agreement	0.857	145

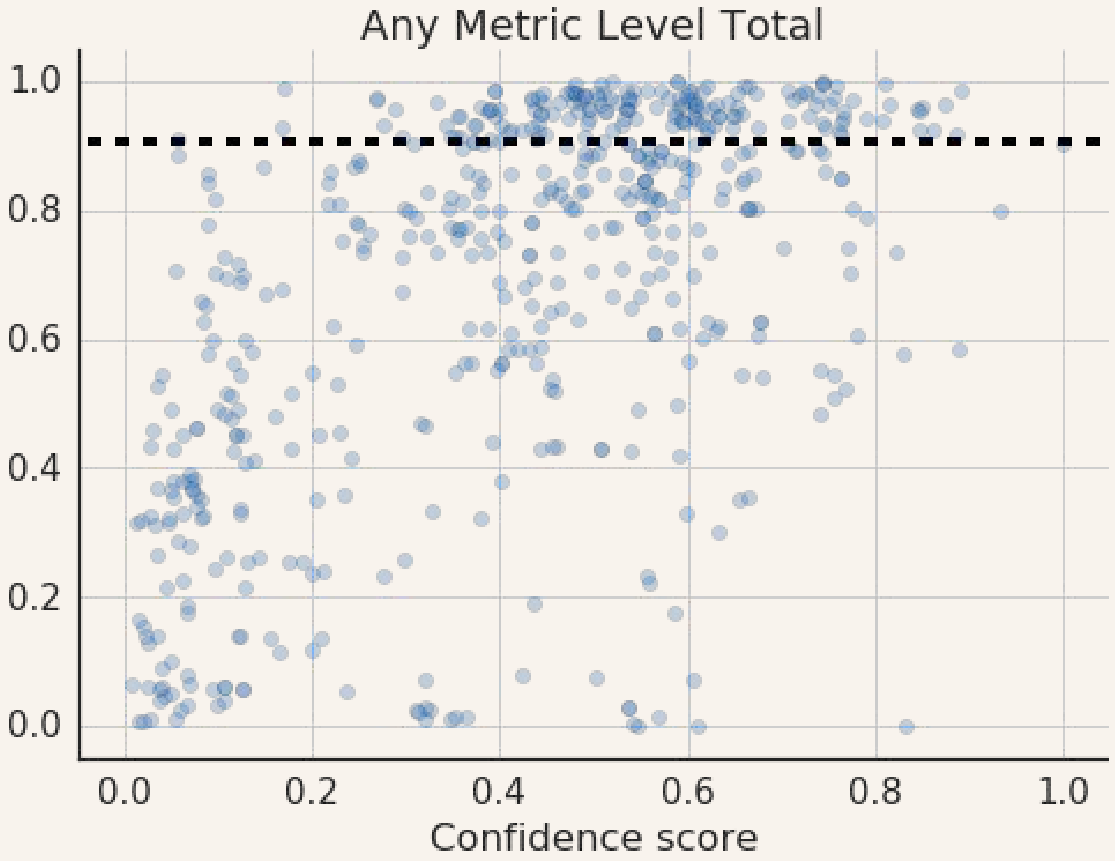
Testing beat reliability



Testing key reliability



Beat results



Idea: Improve DTW confidence score reporting

- Confidence score seems somewhat invariant to sloppiness
- Also is not reliable for covers, silence/garbage
- This is the final step for matching, so it better be good!
- Can we take a learning-based approach?
- Annotate (more) bad match/bad alignment/good to go
- Train a classifier on (features of) the alignment
- Preliminary results are promising

Idea: Improve/replace DTW

- DTW itself also has some failure modes...
- … but we did a pretty exhaustive search.
- Can we improve its fine-grained temporal alignment?
- What if the features are better? We saw that network output distance distributions were much nicer.
- Can we combine many existing alignment approaches?
- Can we learn the alignment algorithm?

Idea: Better pruning

- The pairwise sequence embedding approach is simplistic
- Feedforward attention is explicitly order invariant
- Recent results show using a position encoding is helpful
- The network architecture search preferred a small network
- Some recent advances for training recurrent models on long sequences without running out of memory
- Better loss functions for encouraging good embeddings

Idea: Getting chord annotations from MIDI

- MIDI files don't have chord annotations
- Chords are easier to estimate from MIDI chromagrams
- Can even separate out the guitar instruments!
- Can we transfer chord labels from aligned MP3s to the MIDI, and train as usual?
- What other things about MIDI can we exploit?
- Will the resulting estimated chords be ~human level?

Idea: Detecting whether C major key is true

- We saw that many MIDI files are erroneously given C key
- Simple idea: Train a C major vs. not C major model
- Also, MIDI is way easier to estimate key from!
- Can easily get pitch class histogram and transition matrix
- Do these benefits make the resulting key annotations as reliable as human labelling?

Idea: Does this MIDI file have vocals transcribed?

- Many MIDI files are karaoke files
- Even if we get a good match and alignment, it’s not a perfect transcription if an instrument is not transcribed
- Can we create a “vocal instrument” classifier?
- Simple heuristics will probably work (monophonic, in the vocal note range)
- Can also look at program number, instrument name
- Similar ideas for finding “melody” tracks

Idea: Vocal embellishment correction

- For MIDIs with vocals, there are often embellishments
- This also makes the transcription imperfect (but maybe who cares, it’s still gives us what we want)
- Can we correct this? E.g. given vocal instrument from MIDI, do fine-grained alignment
- May help to have rudimentary vocal extraction
- Can be viewed as an informed transcription problem

Idea: Do things with the dataset

- Some things don’t require matching or aligning (training generative models, corpus studies…)
- Matching can gives more metadata in both directions
- With good alignments: Transcription, instrument activation, key changes, (down)beat tracking, lyrics transcription, tempo estimation, ...

Thanks!