Tutorial
Cross-Modal Music Retrieval and Applications

Part I: Classical Approaches

Meinard Müller
International Audio Laboratories Erlangen
meinard.mueller@audiolabs-erlangen.de

Andreas Arzt, Stefan Balke
Johannes Kepler University
andreas.arzt@jku.at, stefan.balke@jku.at

AUDIO LABS

FAU FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG

Fraunhofer IIS

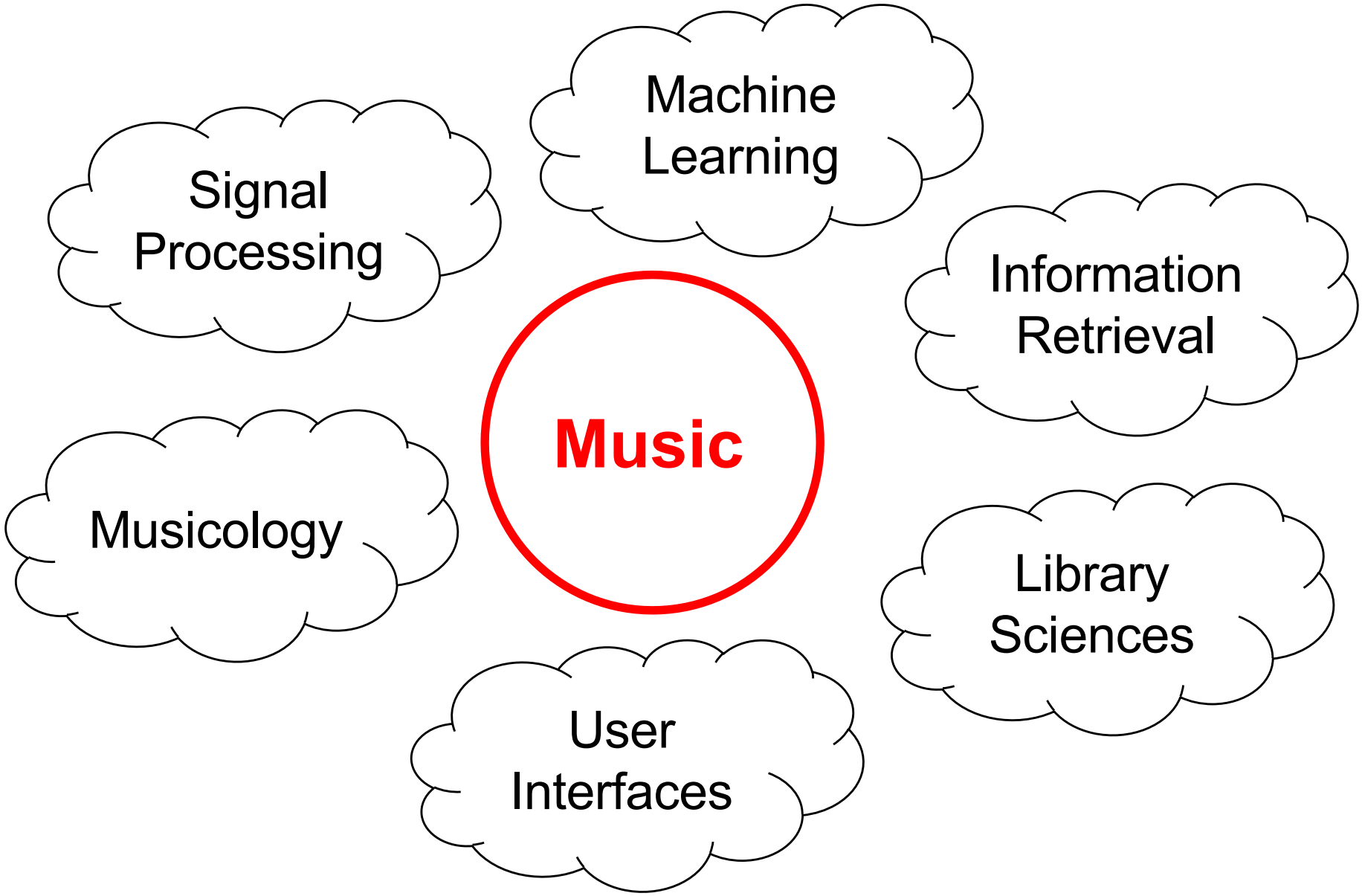
JKU JOHANNES KEPLER UNIVERSITÄT LINZ

Overview (Part I)

- Music representations
- Retrieval scenarios (modality , specificity, granularity)
- Music synchronization (chroma features, dynamic time warping)
- Audio matching (subsequence DTW)
- Cover song retrieval
- Shingle-based retrieval (embedding techniques, PCA, deep learning)
- Cross-modal retrieval (challenges, enhanced representations)
- …

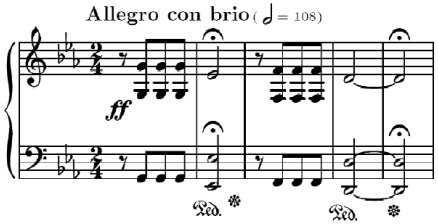

Music

Music Information Retrieval (MIR)

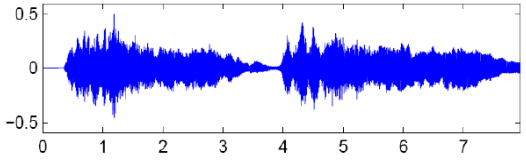


Music Information Retrieval (MIR)

Sheet Music (Image)



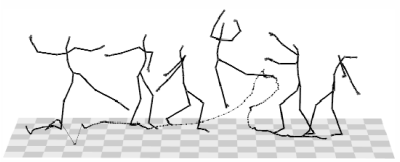
CD / MP3 (Audio)



MusicXML (Text)

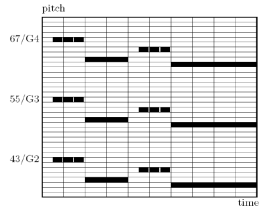
```
<note>
  <pitch>
    <step>E</step>
    <alter>-1</alter>
    <octave>4</octave>
  </pitch>
  <duration>2</duration>
  <type>half</type>
</note>
```

Dance / Motion (Mocap)



Music

MIDI

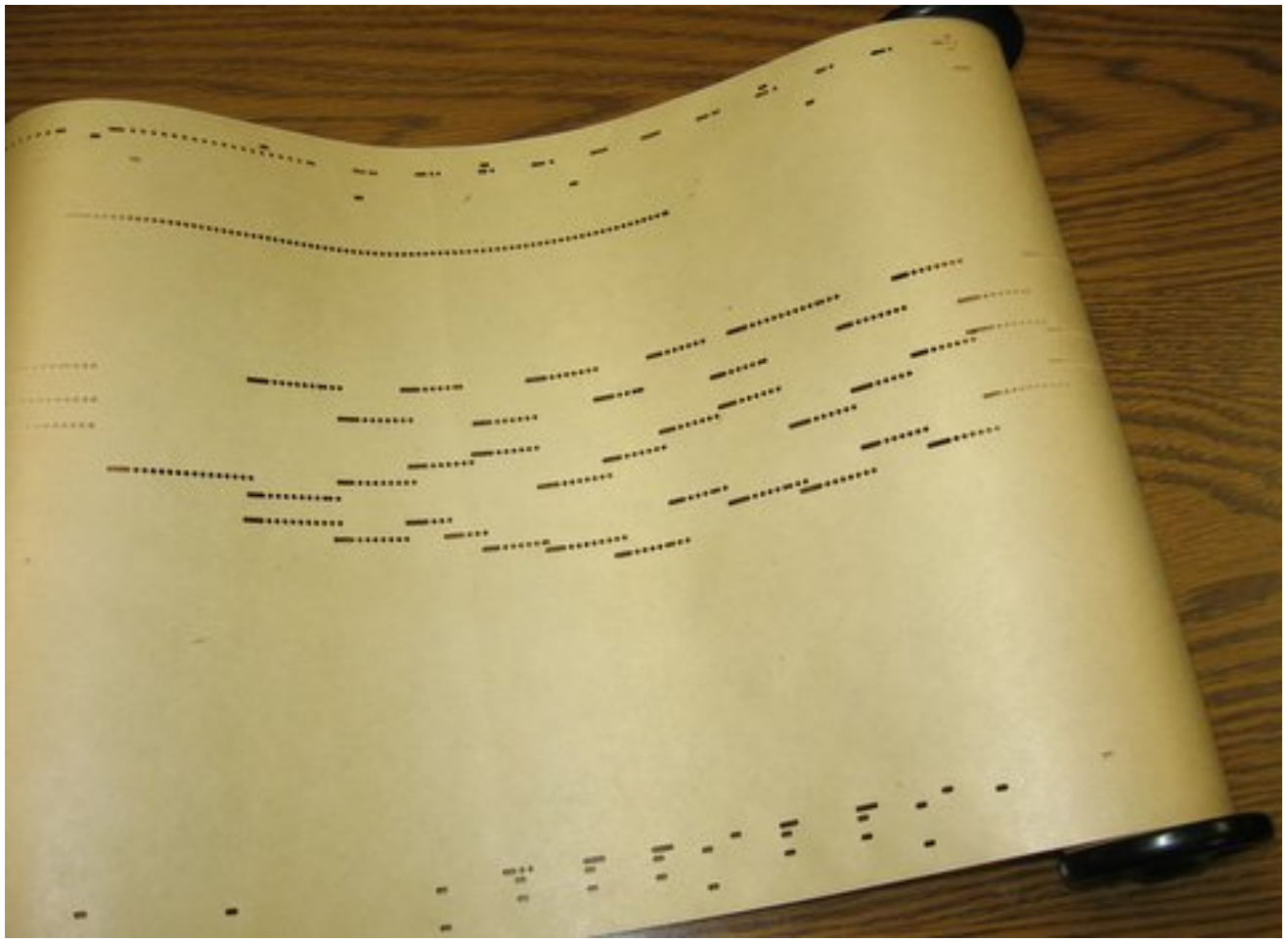


Singing / Voice (Audio)

Music Film (Video)

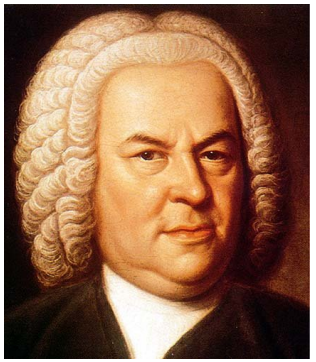
Music Literature (Text)

Piano Roll Representation



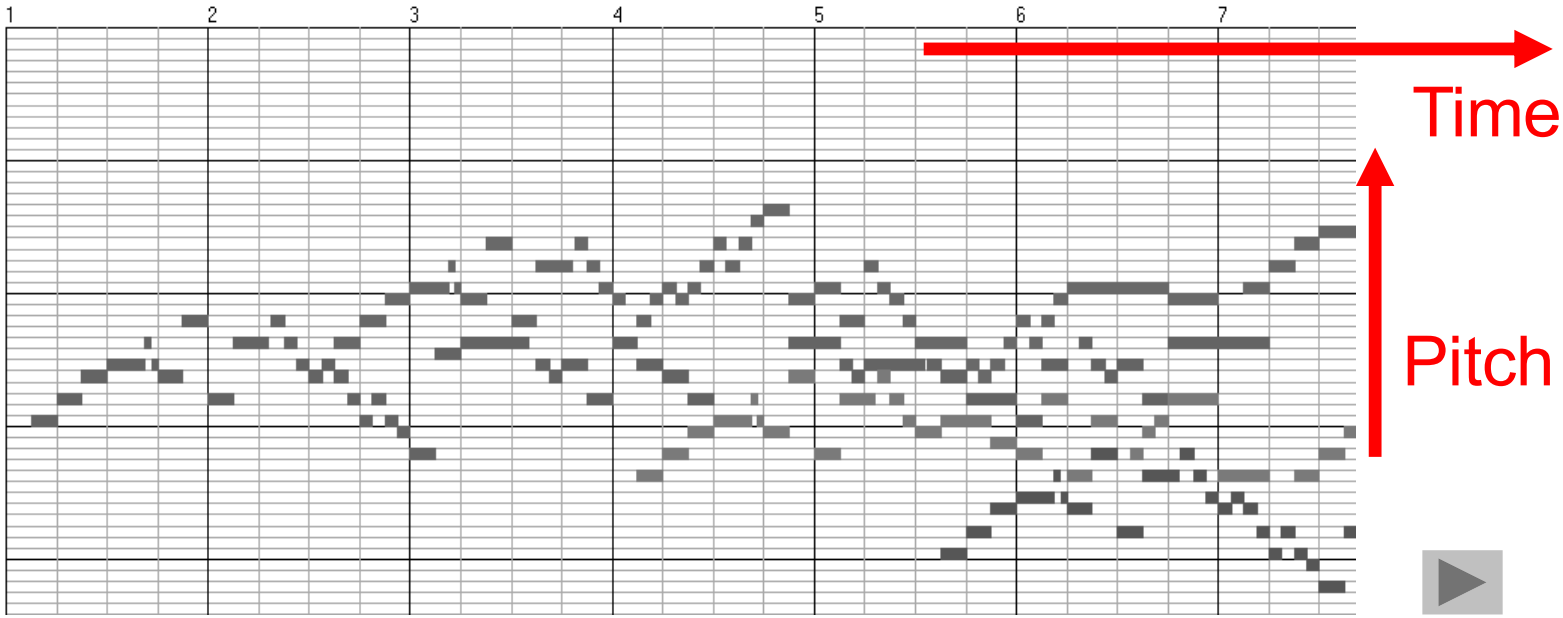
Player Piano (1900)

Piano Roll Representation (MIDI)

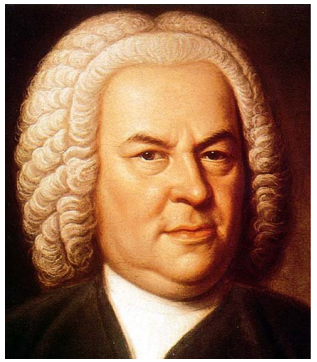


J.S. Bach, C-Major Fuge

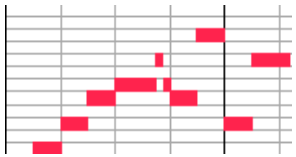
(Well Tempered Piano, BWV 846)



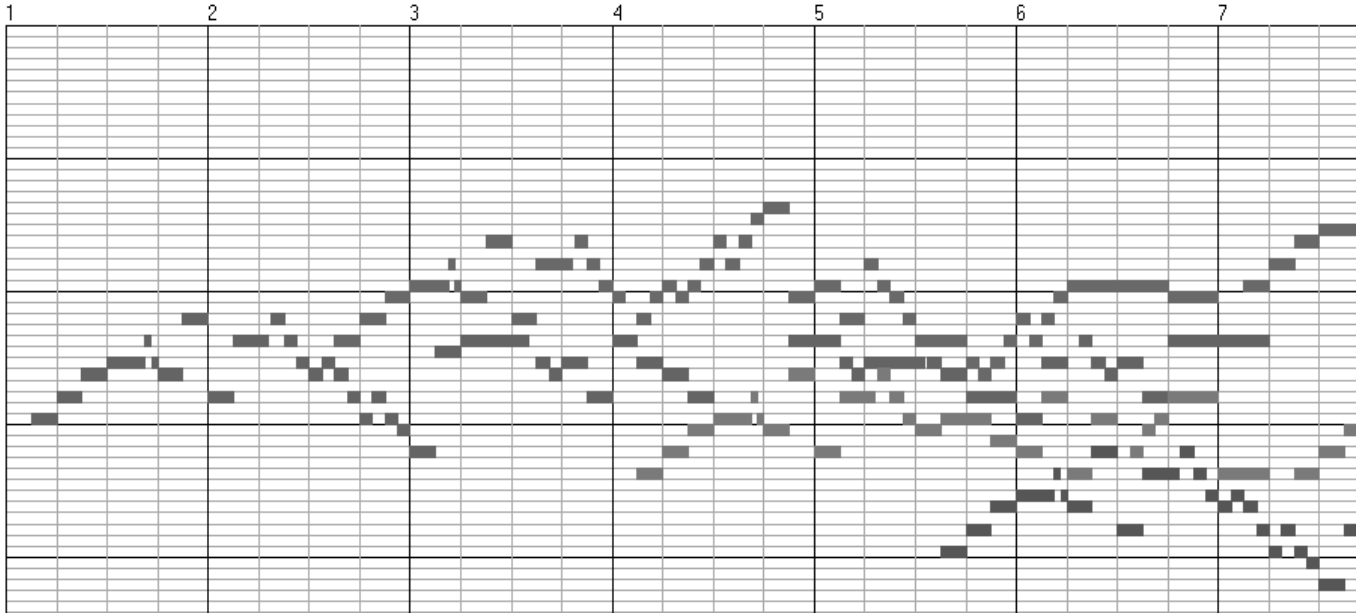
Piano Roll Representation (MIDI)



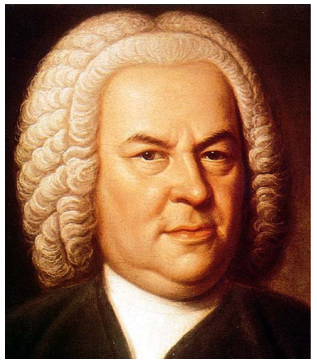
Query:



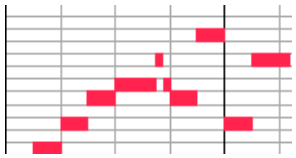
Goal: Find all occurrences of the query



Piano Roll Representation (MIDI)

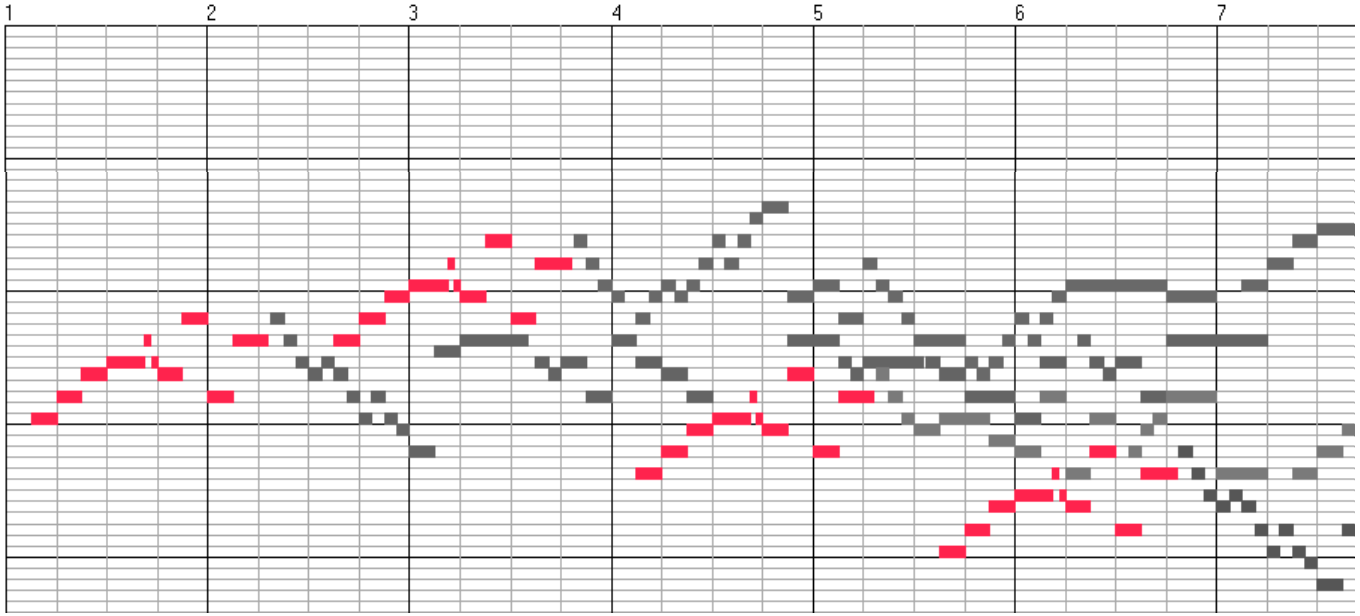


Query:

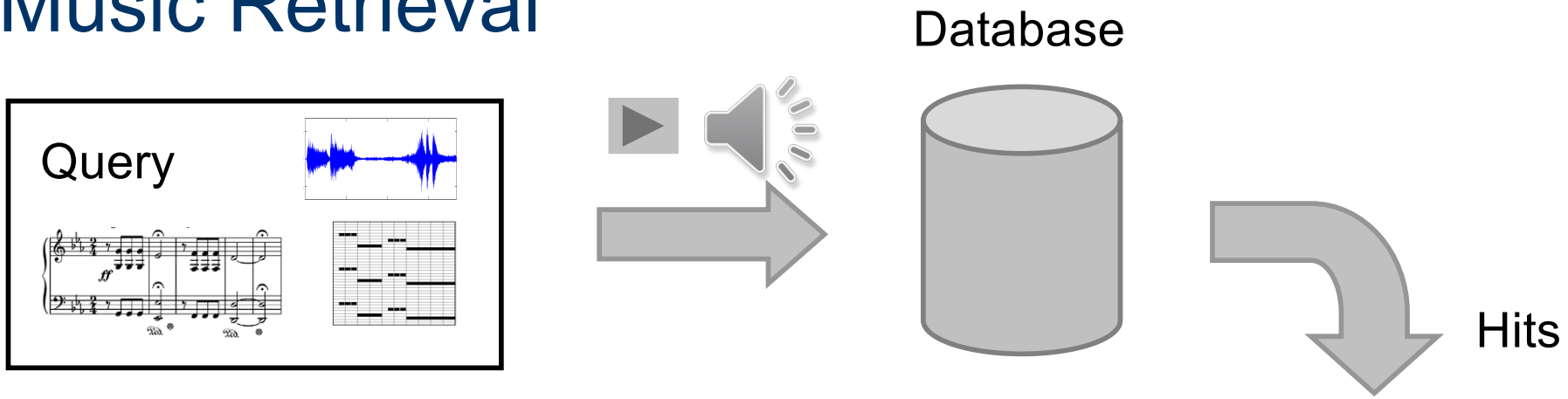


Goal: Find all occurrences of the query

Matches:



Music Retrieval



Retrieval tasks:

Audio identification

Bernstein (1962)
Beethoven, Symphony No. 5

Audio matching

Beethoven, Symphony No. 5:
- Bernstein (1962)
- Karajan (1982)
- Gould (1992)

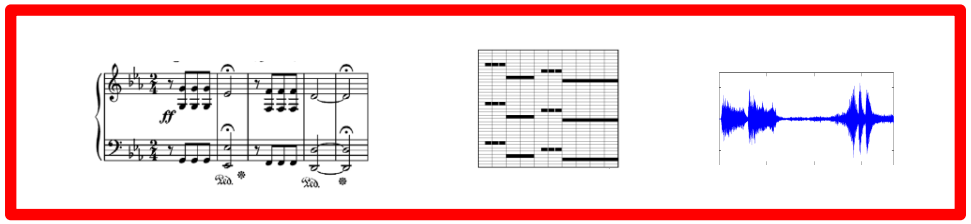
Version identification

Category-based music retrieval

- Beethoven, Symphony No. 9
- Beethoven, Symphony No. 3
- Haydn Symphony No. 94

Music Retrieval

Modalities



Retrieval tasks:

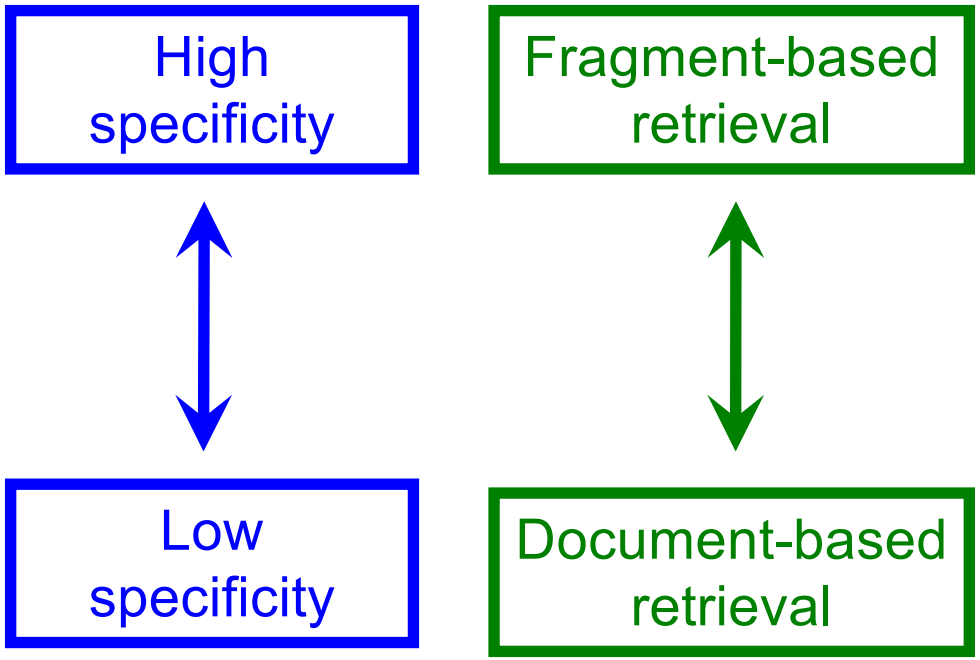
Audio identification

Audio matching

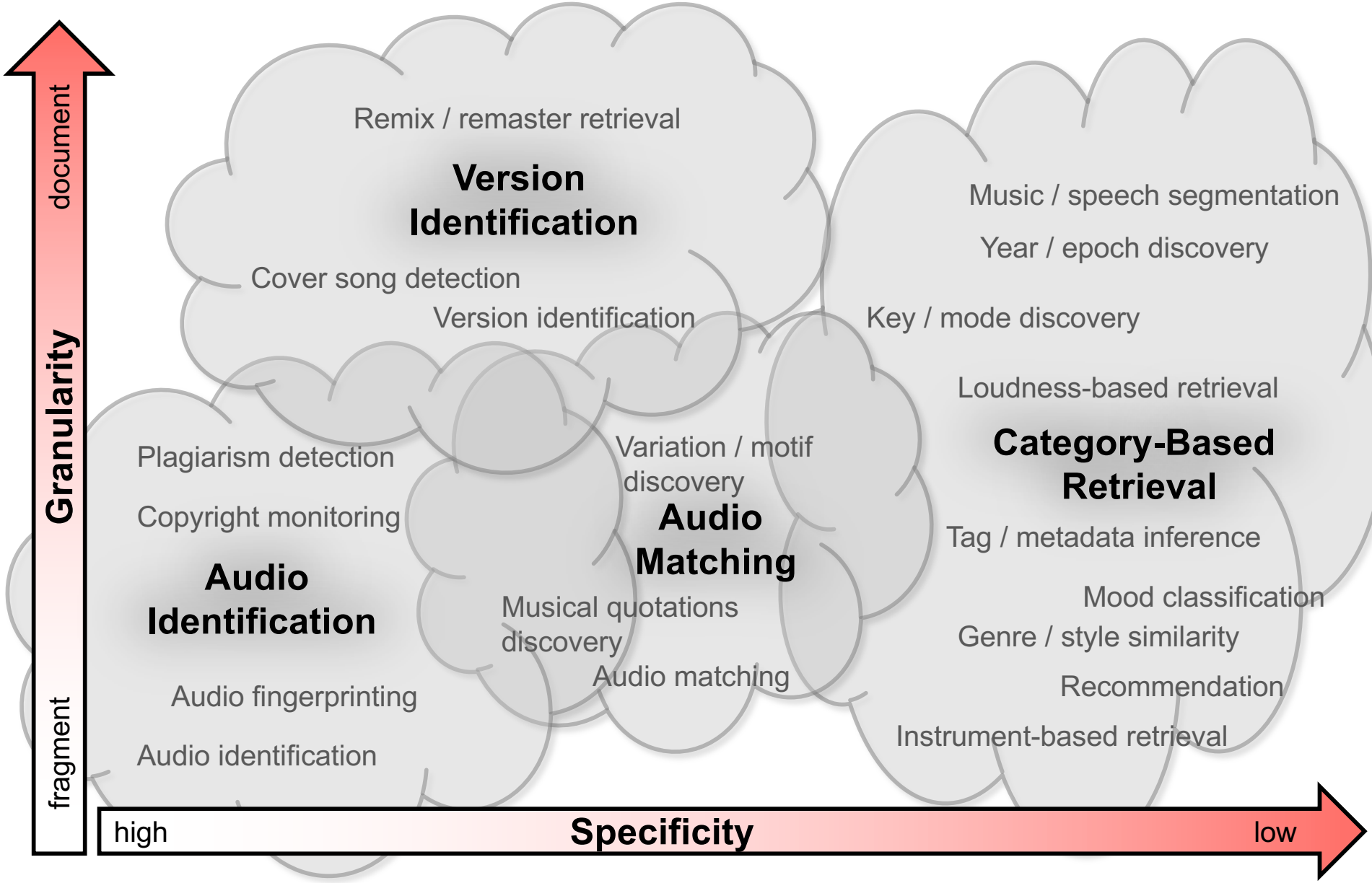
Version identification

Category-based music retrieval

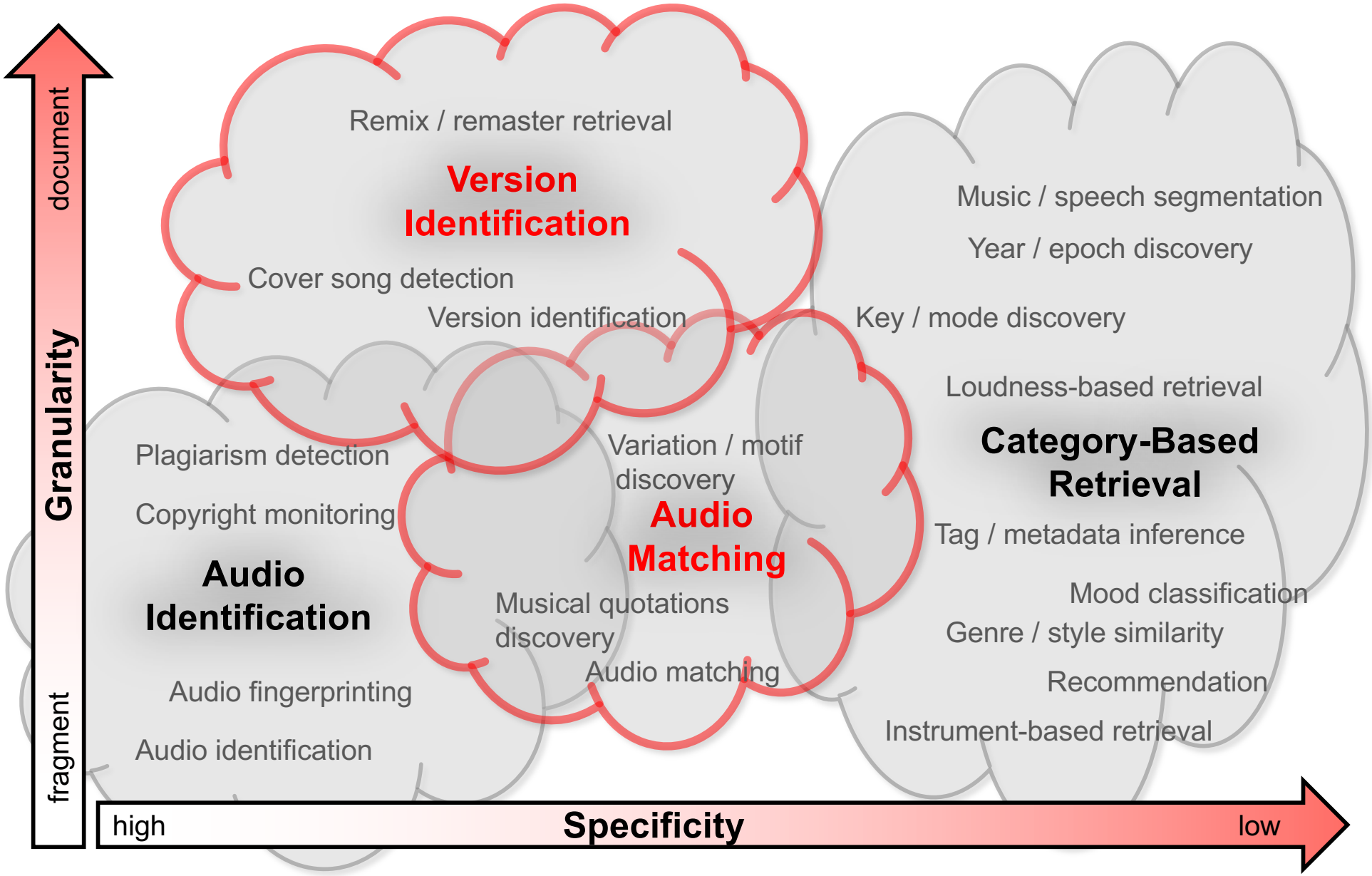
Specificity **Granularity**



Music Retrieval



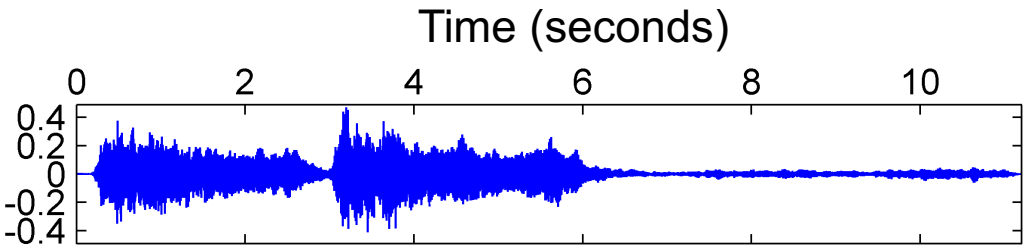
Music Retrieval



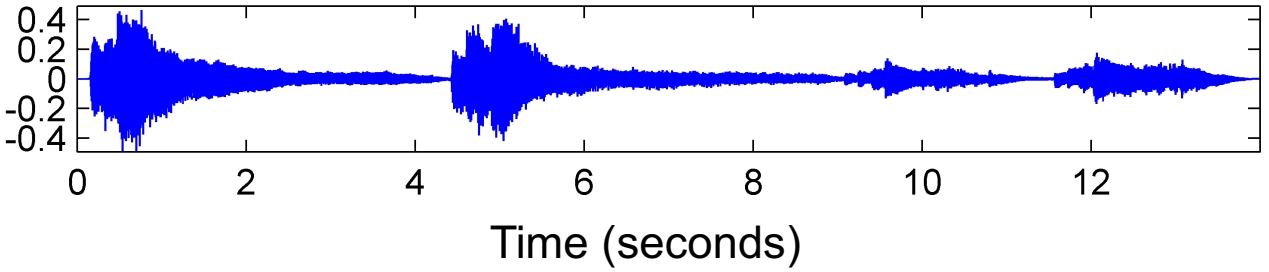
Music Synchronization: Audio-Audio

Beethoven's Fifth

Karajan



Gould

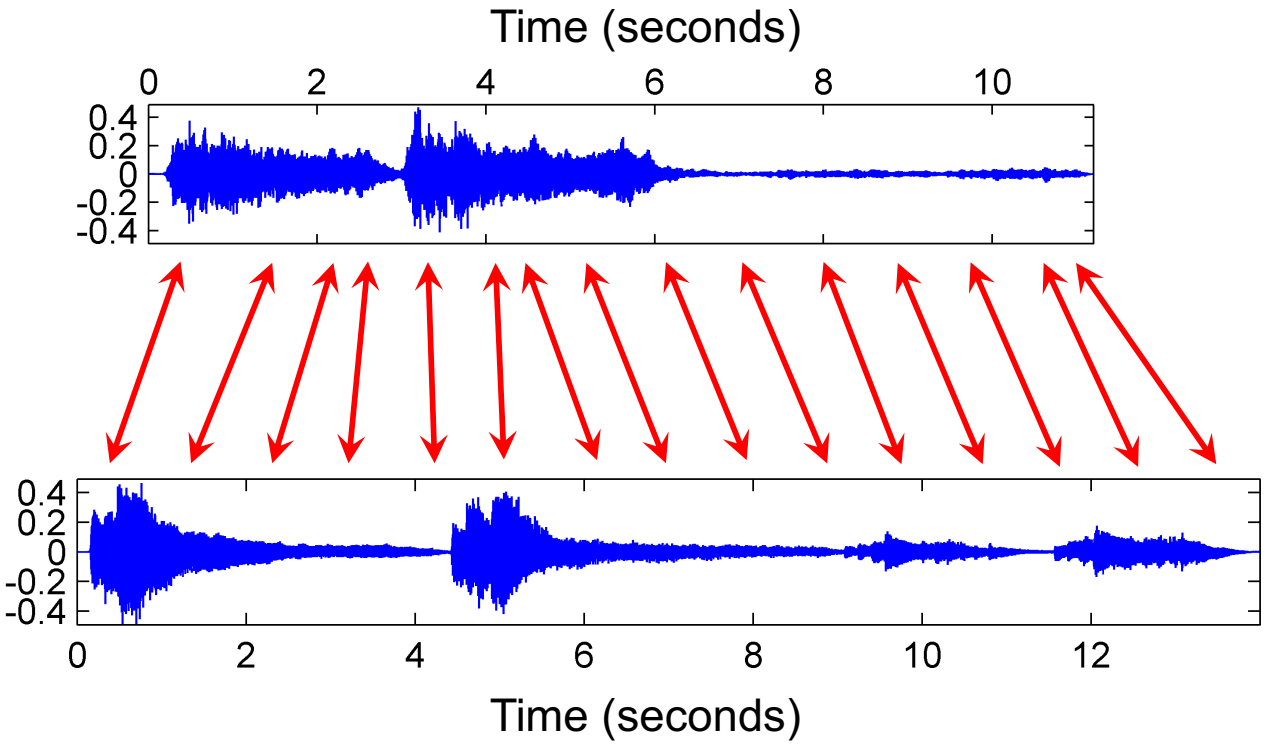


Music Synchronization: Audio-Audio

Beethoven's Fifth

Karajan

Gould



Music Synchronization: Audio-Audio

Task

Given: Two different audio recordings (two versions) of the same underlying piece of music.

Goal: Find for each position in one audio recording the musically corresponding position in the other audio recording.

Music Synchronization: Audio-Audio

Two main steps:

1.) Feature extraction

- Robust to variations (e.g., instrumentation, timbre, dynamics)
- Discriminative (e.g., capturing harmonic, melodic, tonal aspects)

➡ **Chroma features**

2.) Temporal alignment

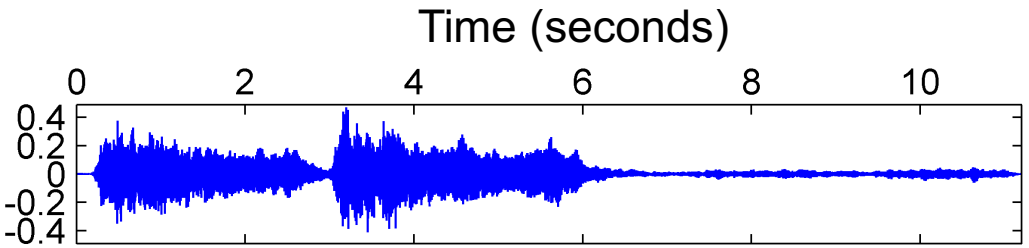
- Capturing local and global tempo variations
- Trade-off: Robustness vs. accuracy
- Efficiency

➡ **Dynamic time warping (DTW)**

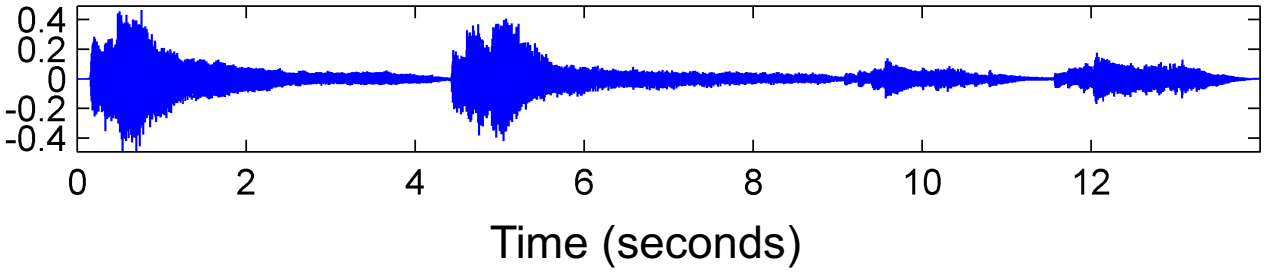
Music Synchronization: Audio-Audio

Beethoven's Fifth

Karajan



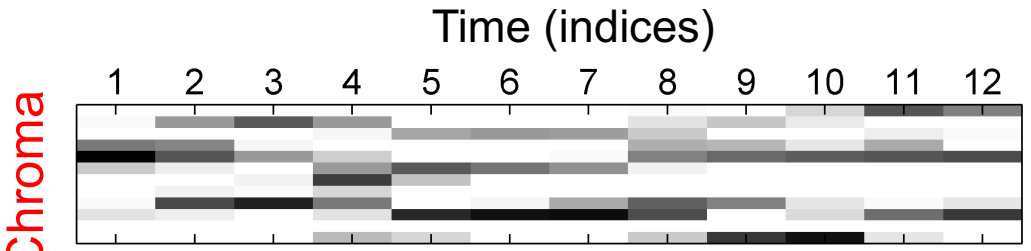
Gould



Music Synchronization: Audio-Audio

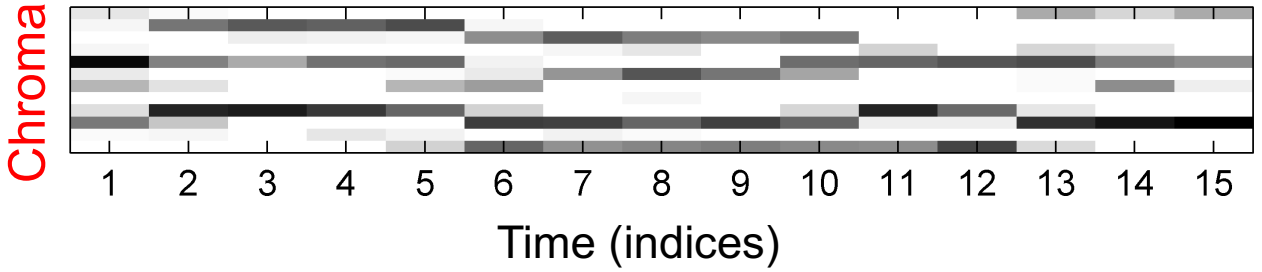
Beethoven's Fifth

Karajan



Time–chroma representations

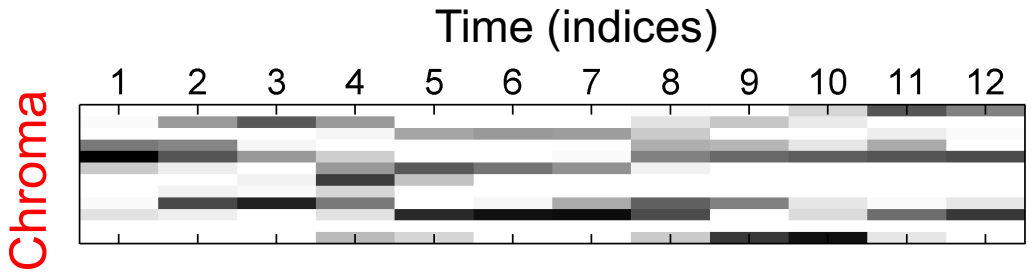
Gould



Music Synchronization: Audio-Audio

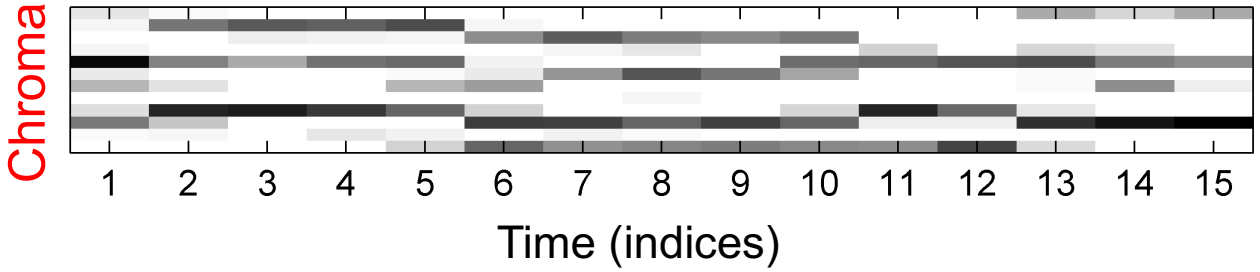
Beethoven's Fifth

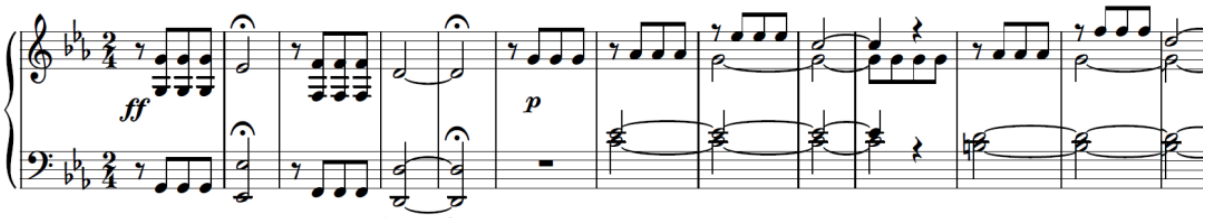
Karajan



Time–chroma representations

Gould

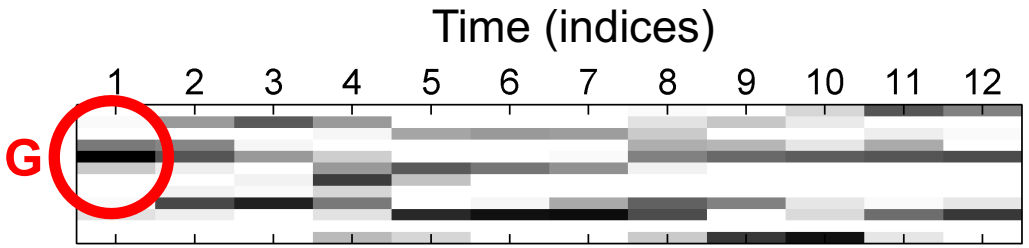




Music Synchronization: Audio-Audio

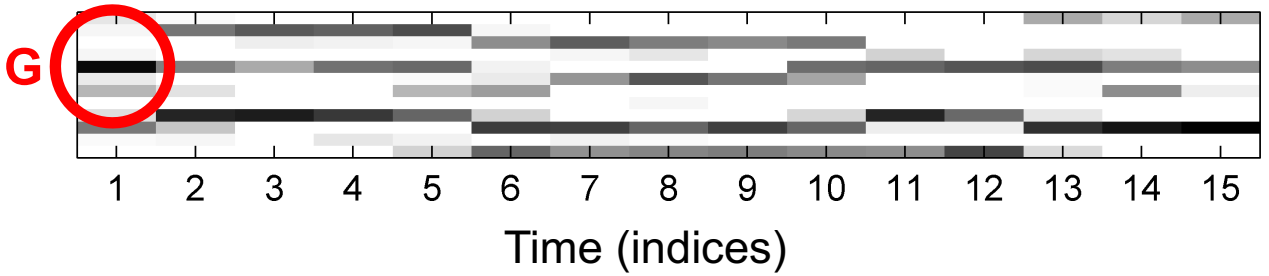
Beethoven's Fifth

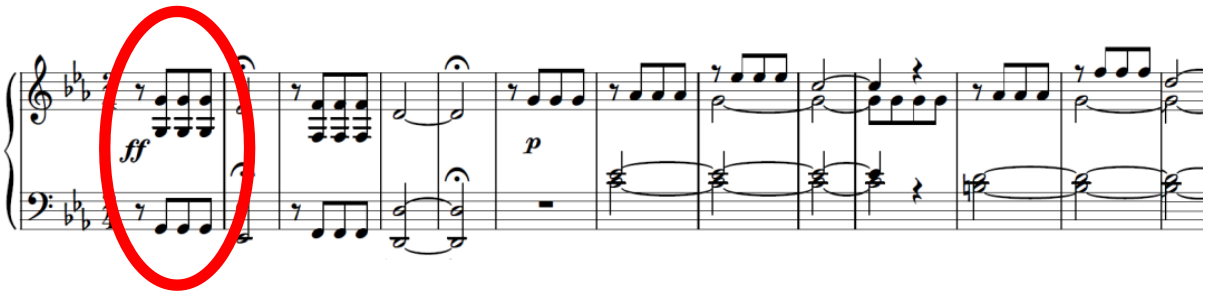
Karajan



Time–chroma representations

Gould

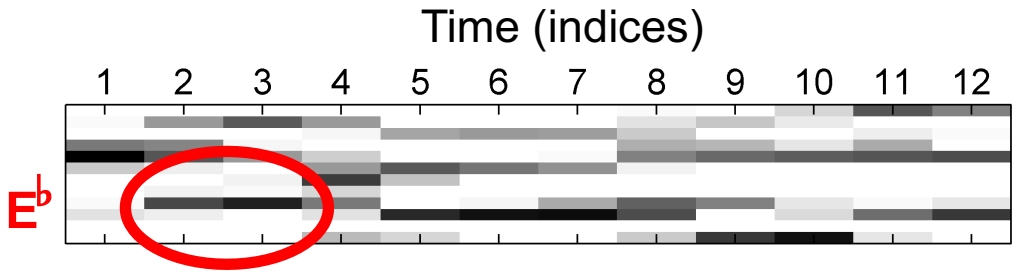




Music Synchronization: Audio-Audio

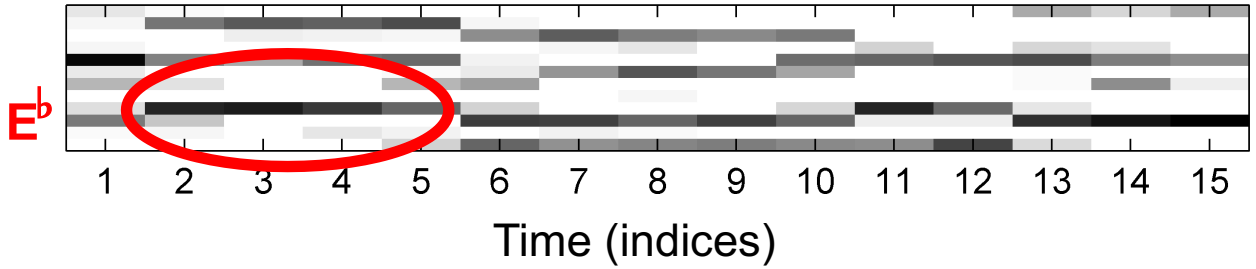
Beethoven's Fifth

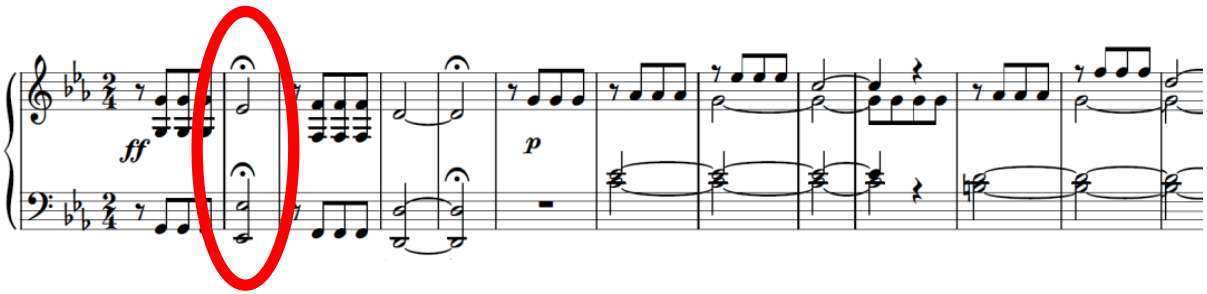
Karajan



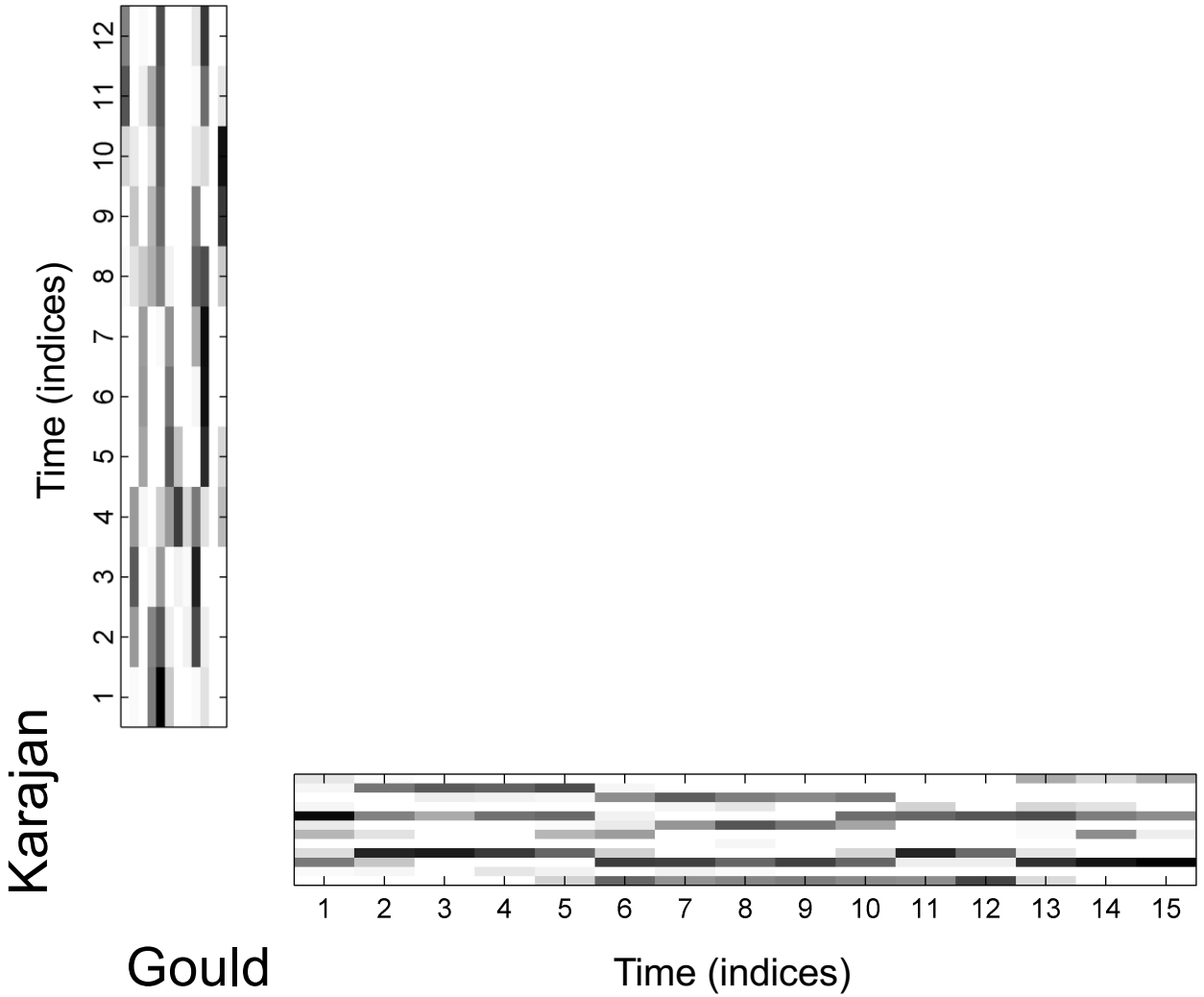
Time–chroma representations

Gould



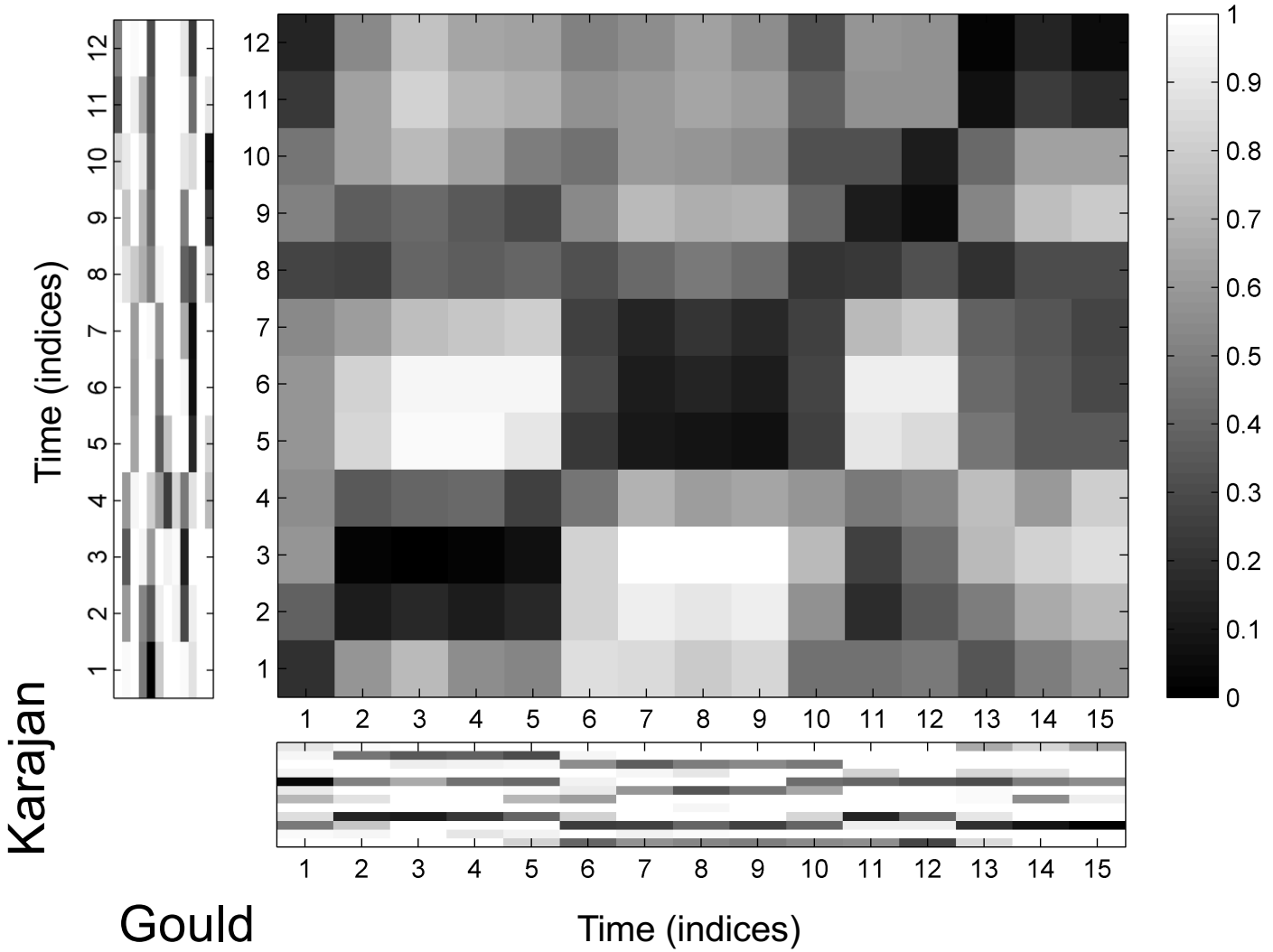


Music Synchronization: Audio-Audio



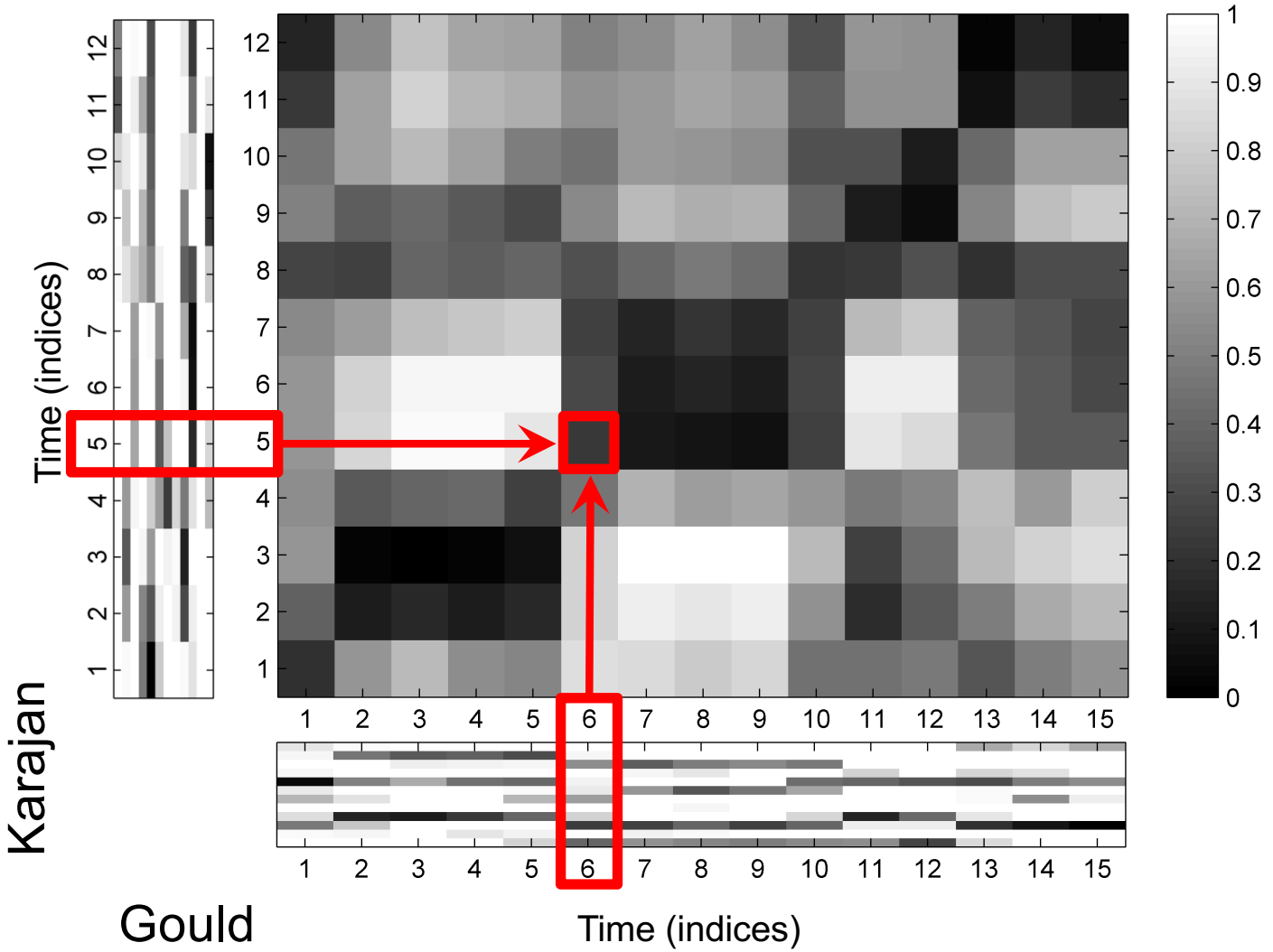
Music Synchronization: Audio-Audio

Cost matrix



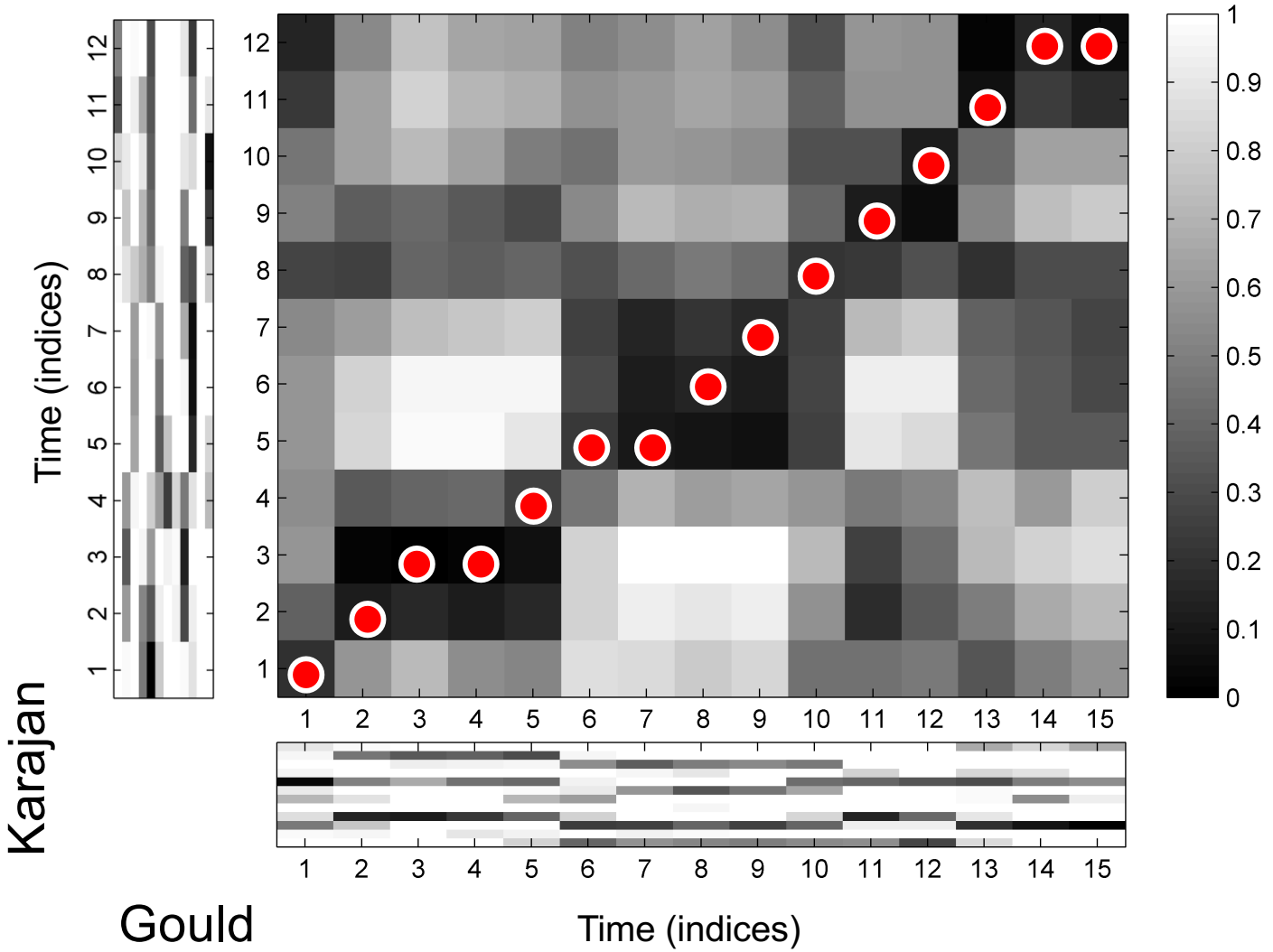
Music Synchronization: Audio-Audio

Cost matrix



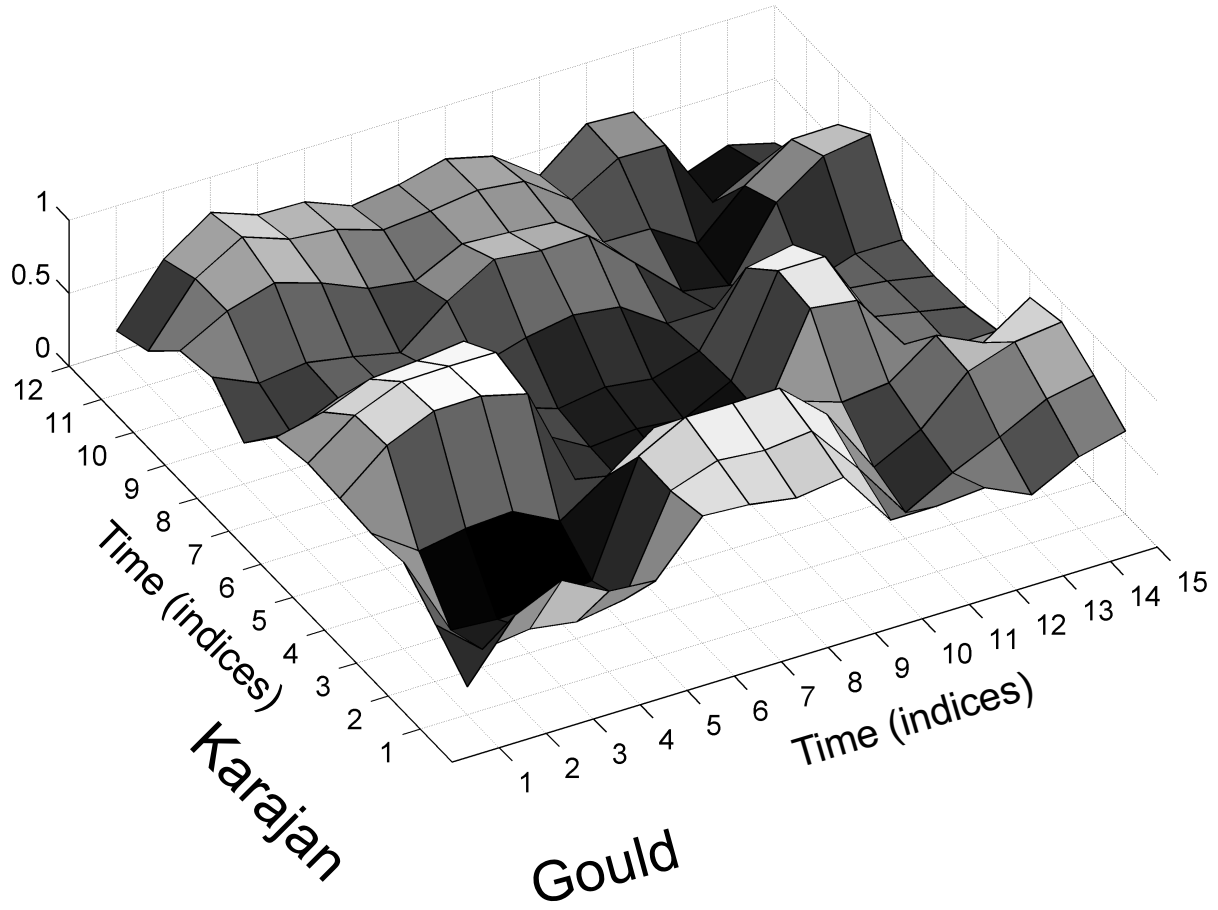
Music Synchronization: Audio-Audio

Cost-minimizing warping path



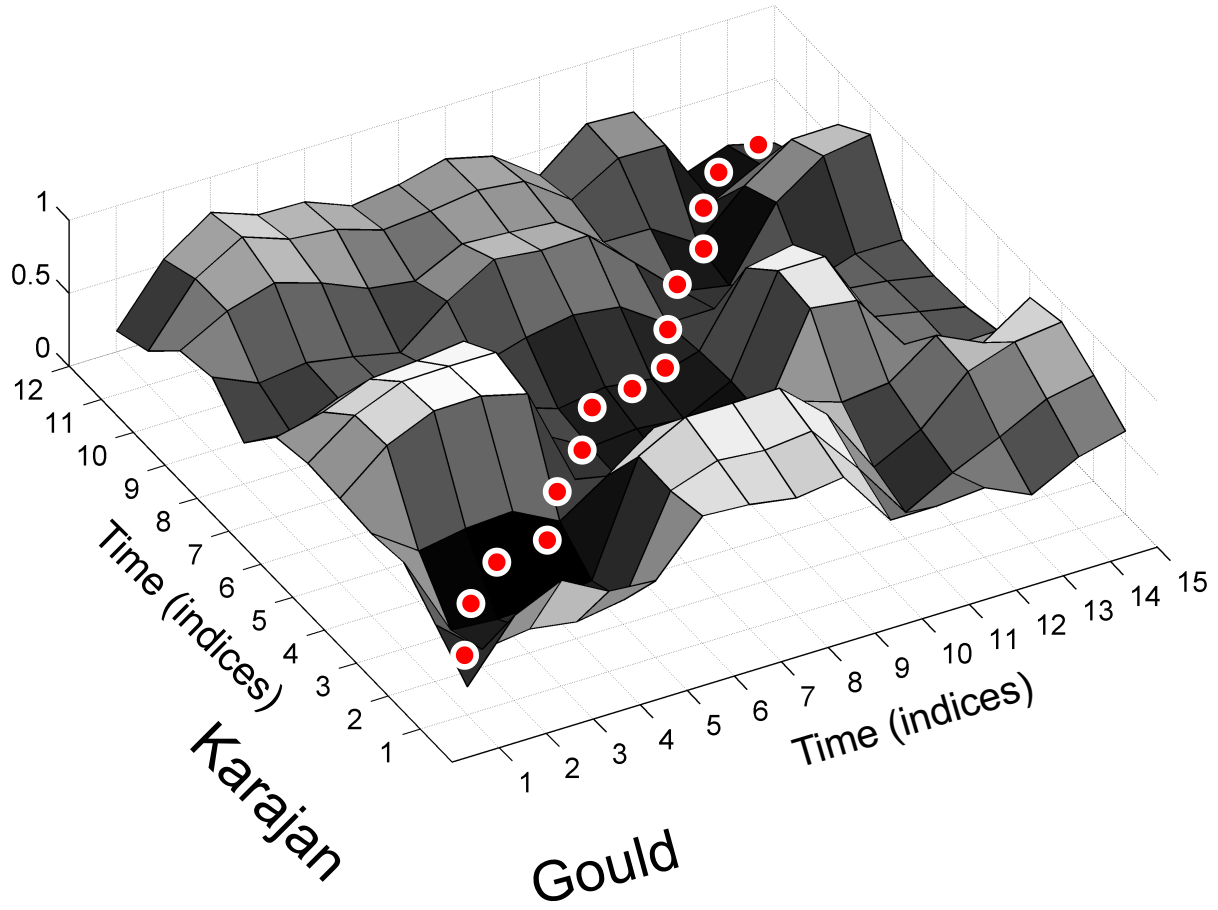
Music Synchronization: Audio-Audio

Cost matrix



Music Synchronization: Audio-Audio

Cost-minimizing warping path

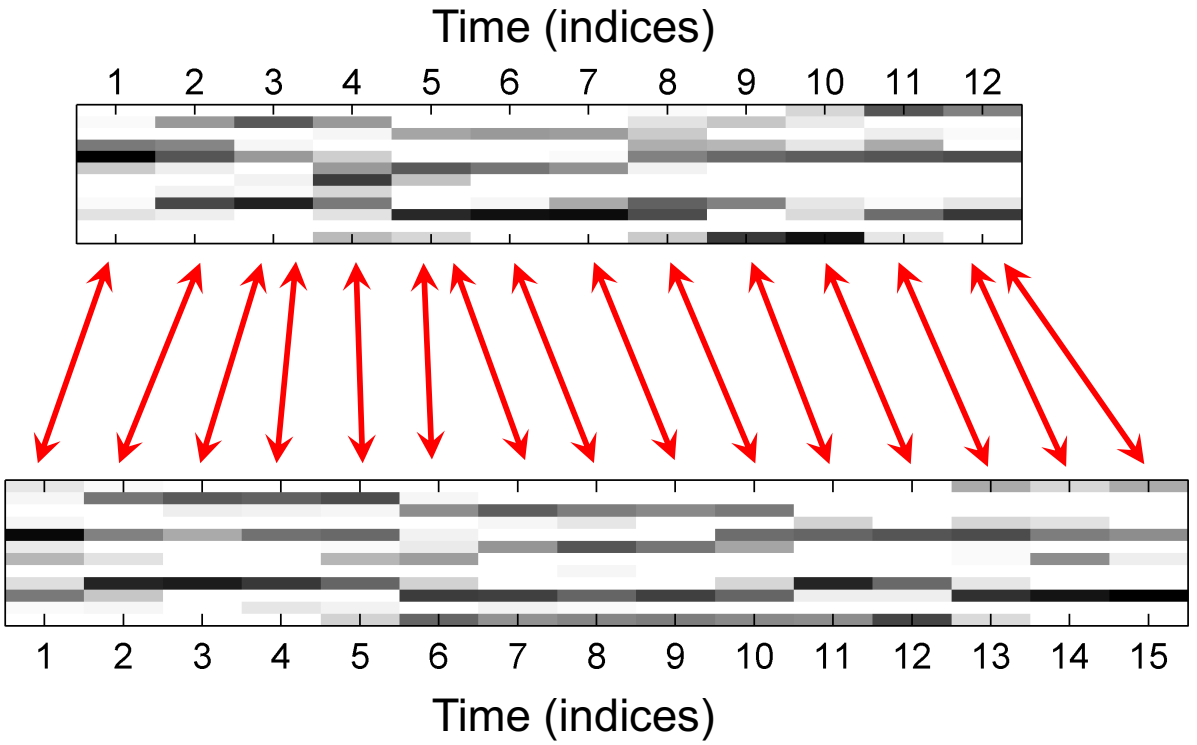


Music Synchronization: Audio-Audio

Optimal alignment (cost-minimizing warping path)

Karajan

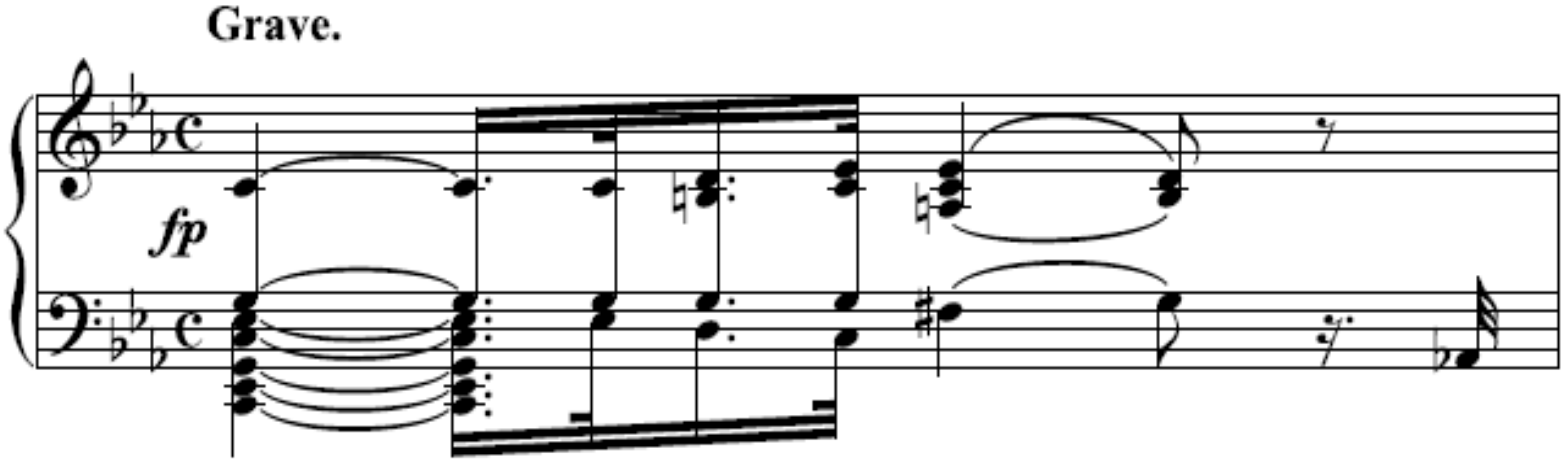
Gould



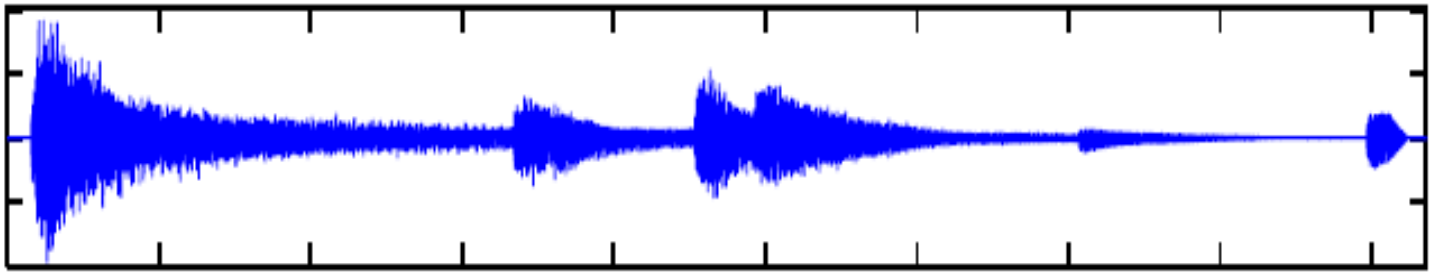
Application: Interpretation Switcher

Music Synchronization: Image-Audio

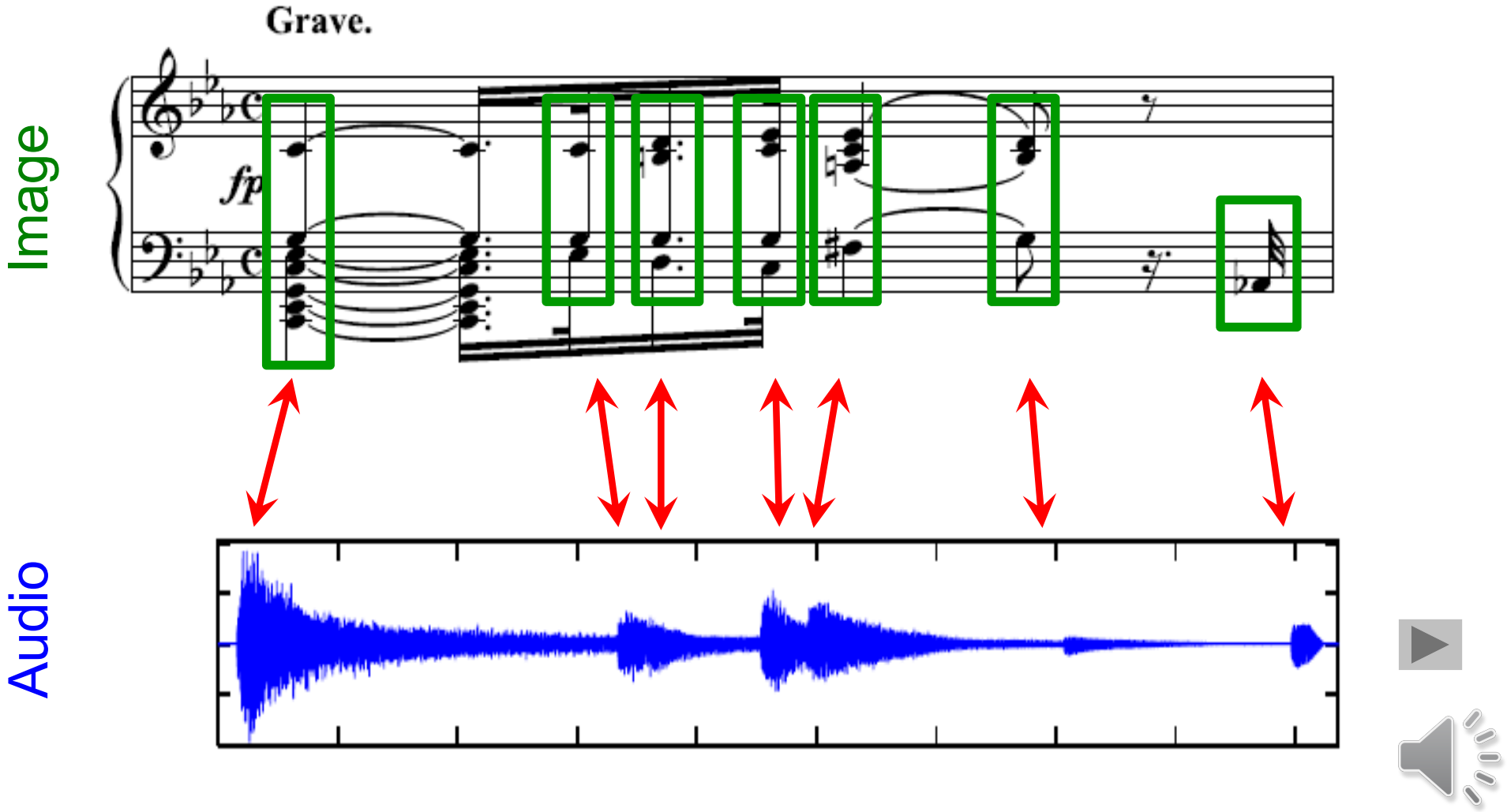
Image



Audio

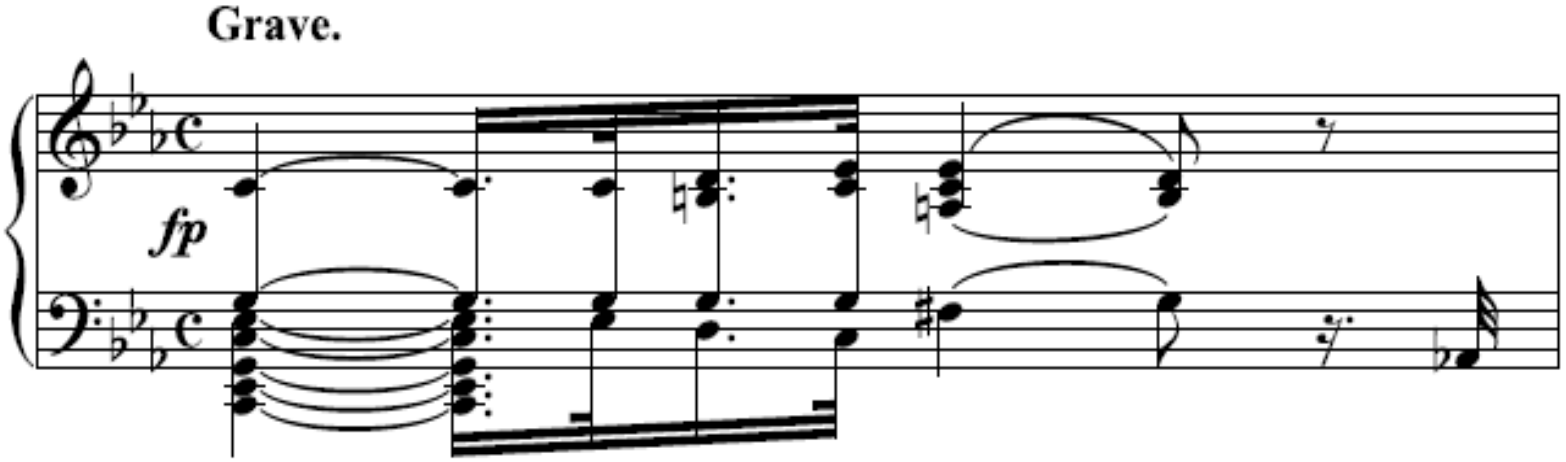


Music Synchronization: Image-Audio

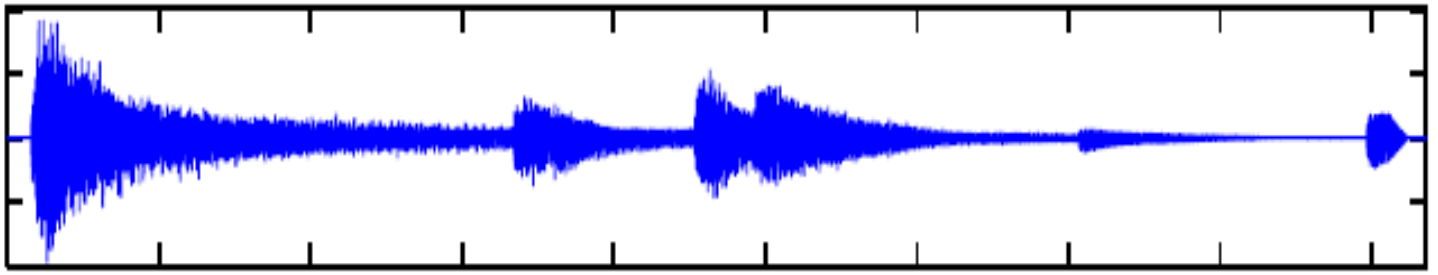


How to make the data comparable?

Image



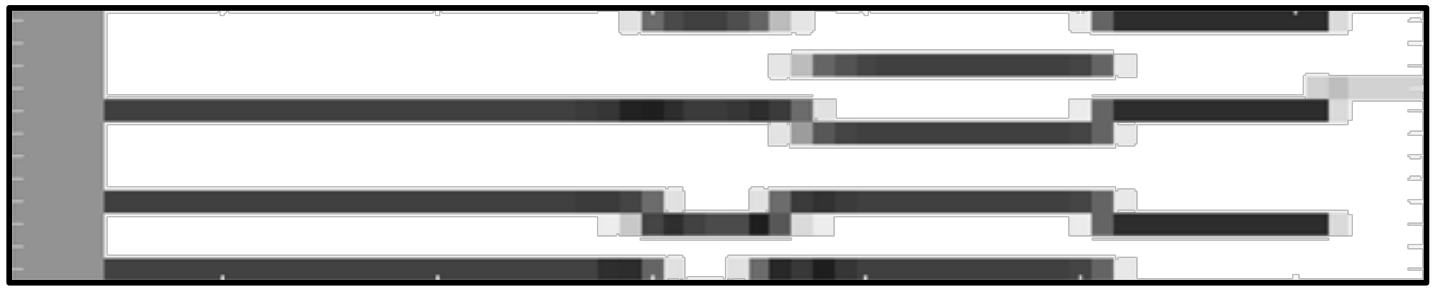
Audio



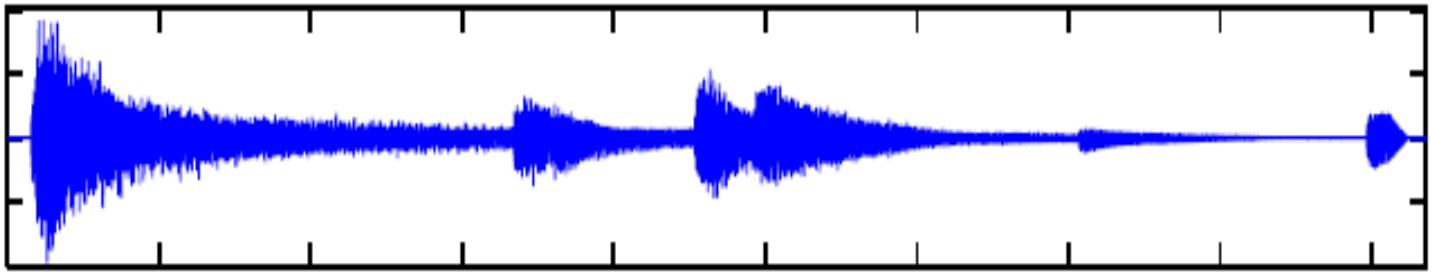
How to make the data comparable?

Image Processing: Optical Music Recognition

Image



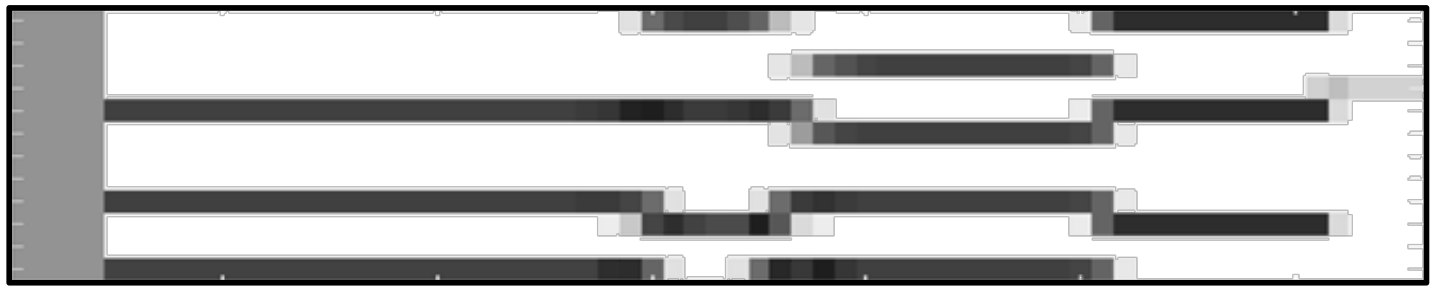
Audio



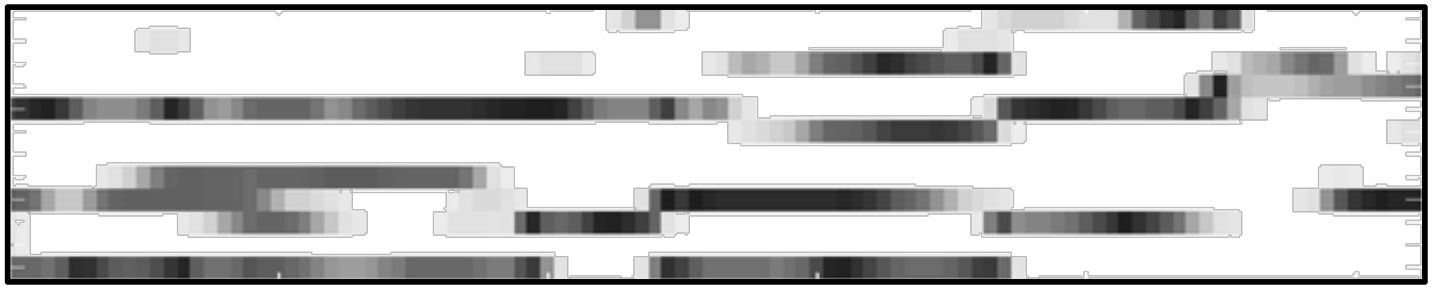
How to make the data comparable?

Image Processing: Optical Music Recognition

Image



Audio

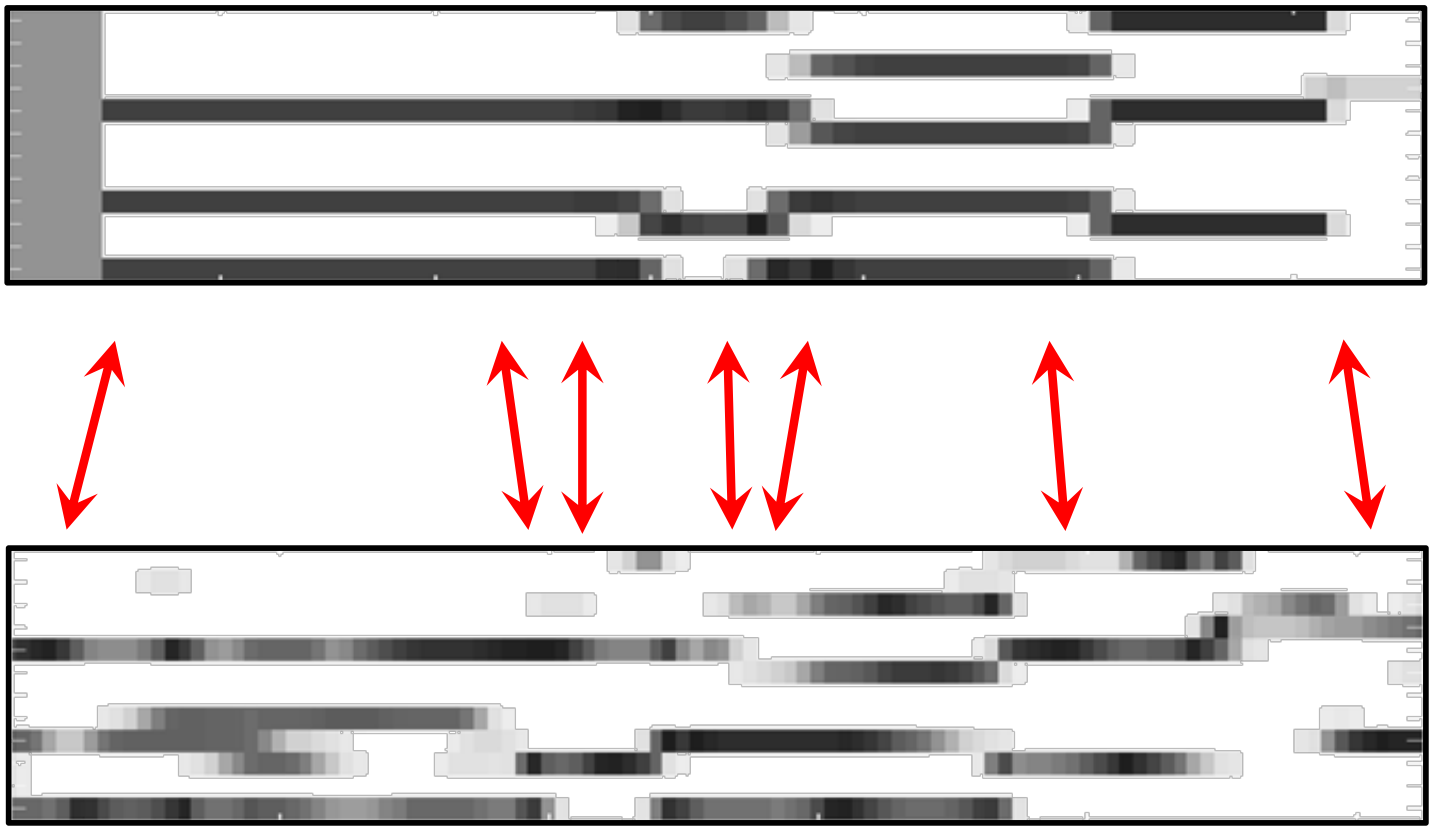


Audio Processing: Fourier Analysis

How to make the data comparable?

Image Processing: Optical Music Recognition

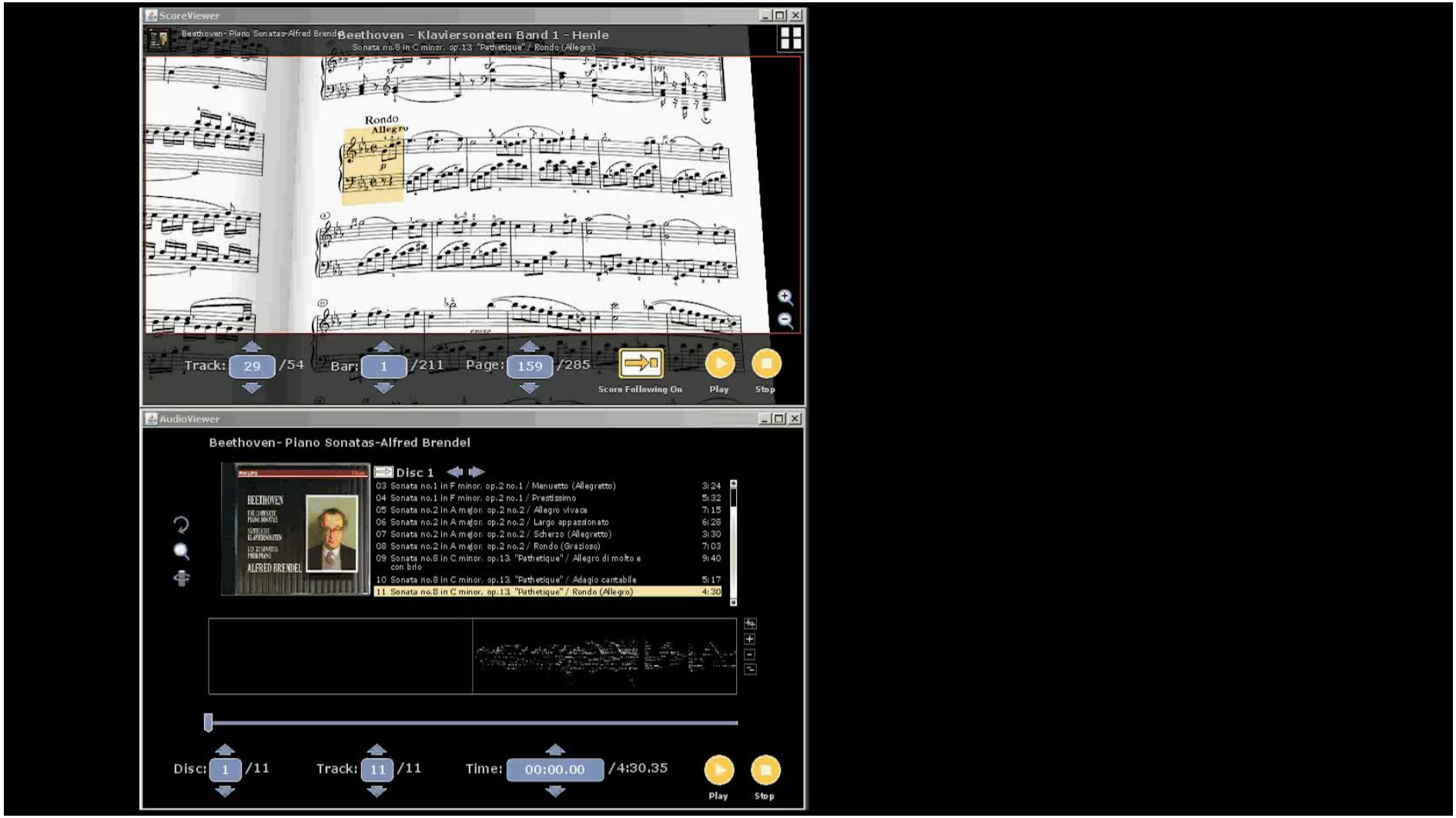
Image



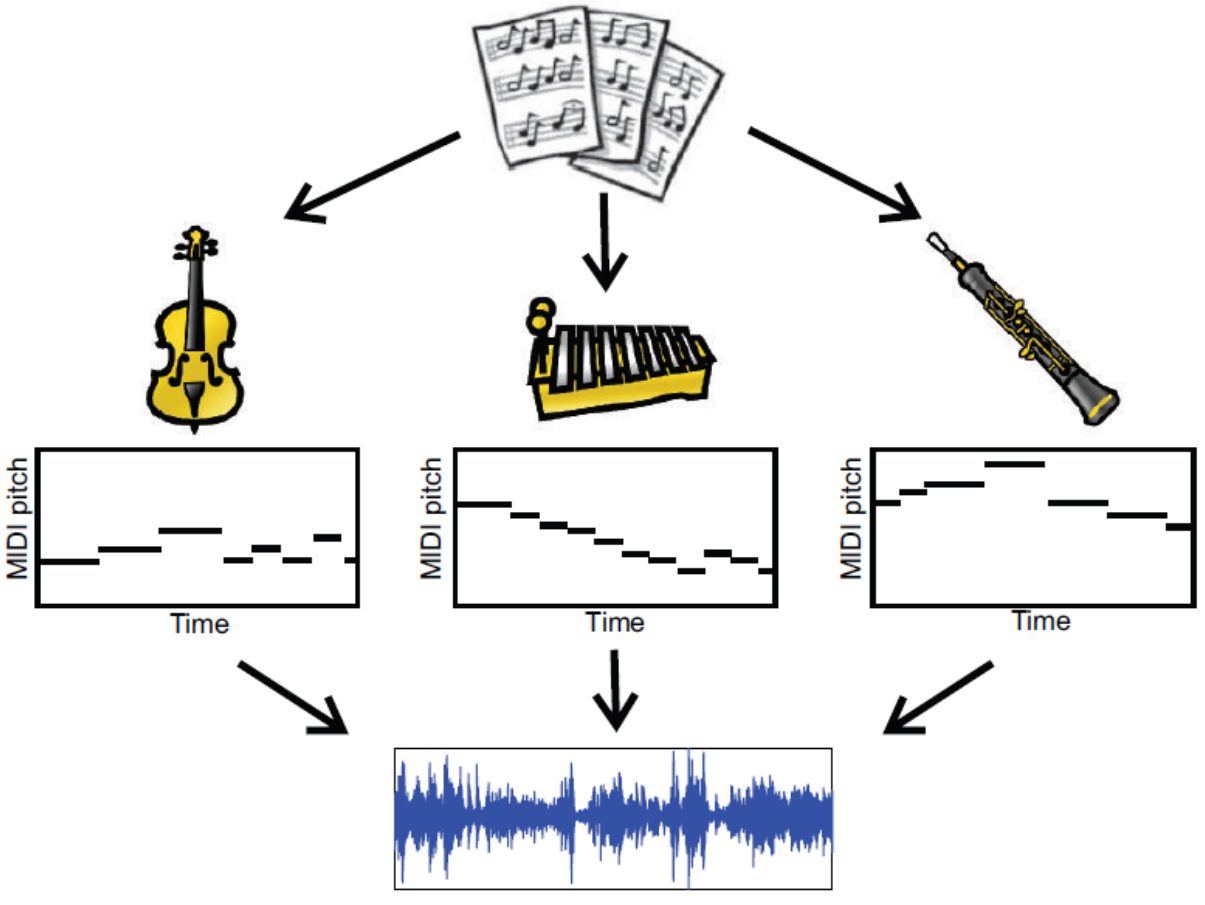
Audio

Audio Processing: Fourier Analysis

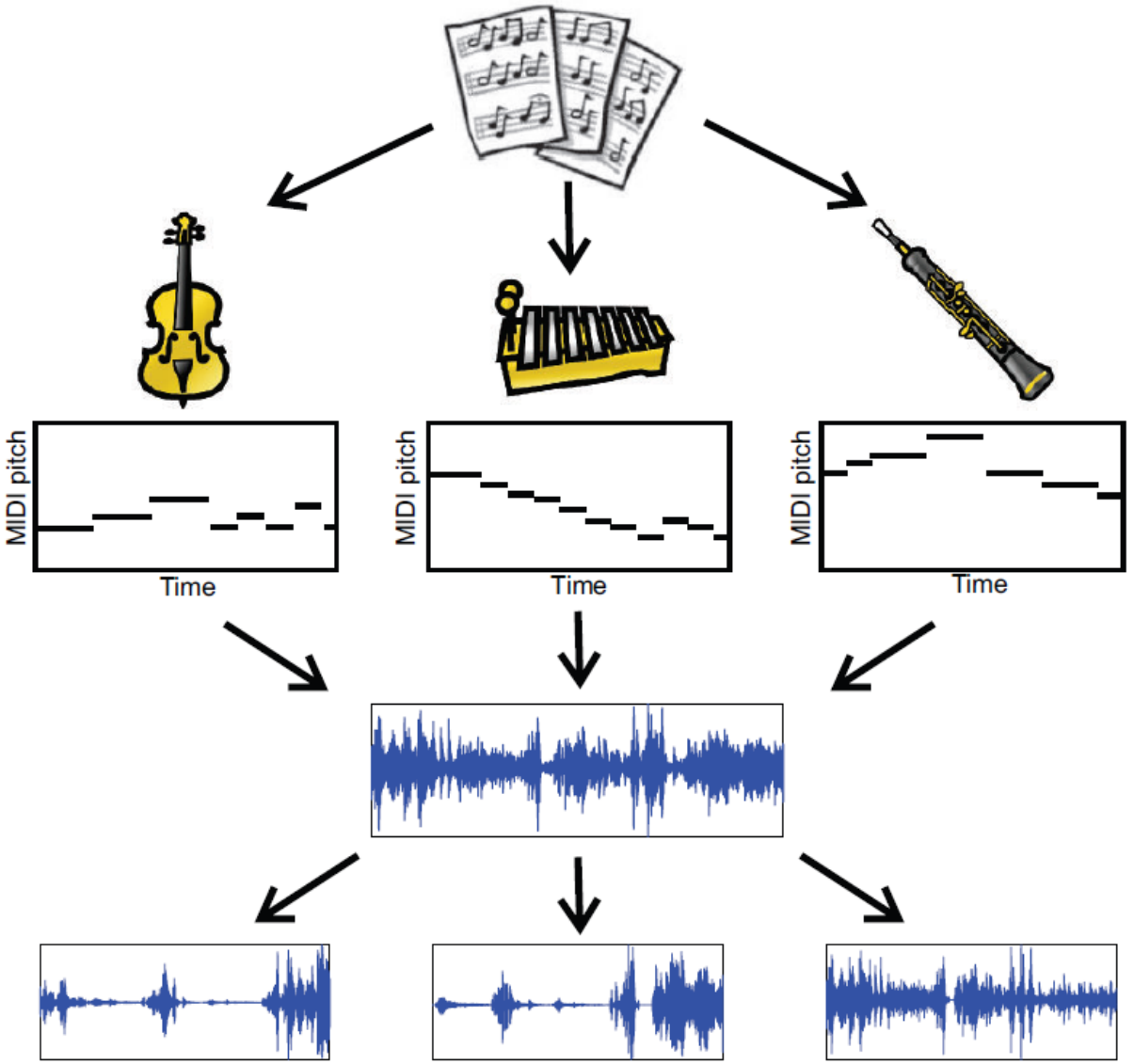
Application: Score Viewer



Application: Score-Informed Source Separation

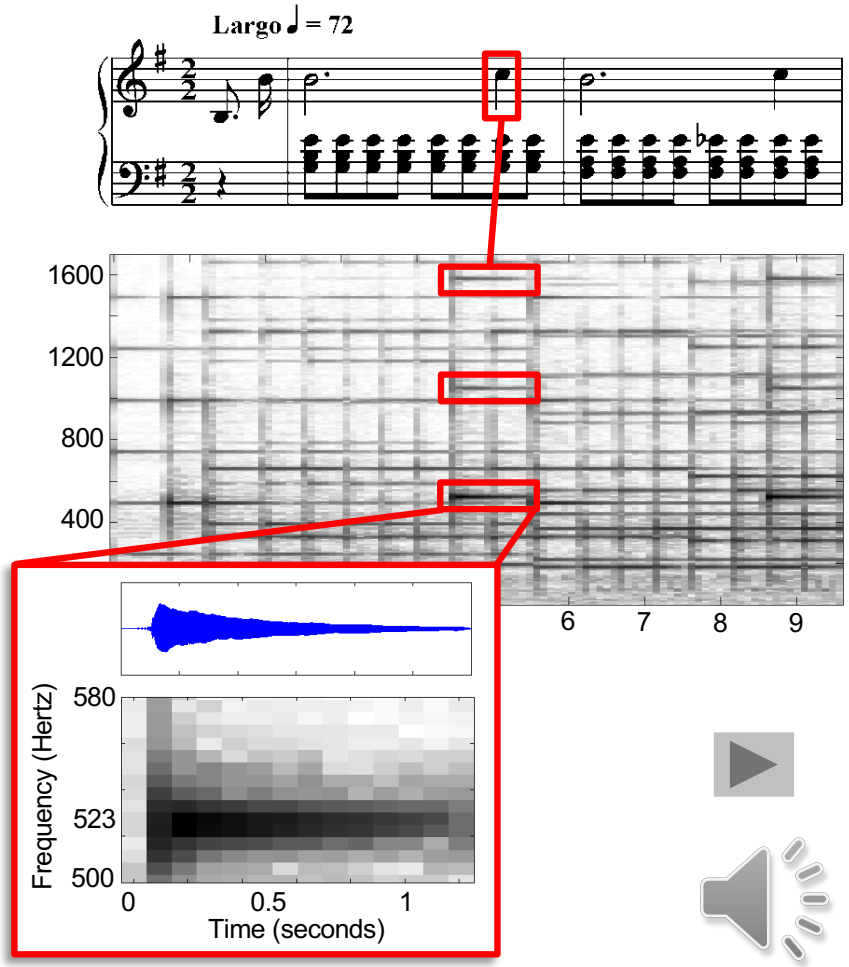


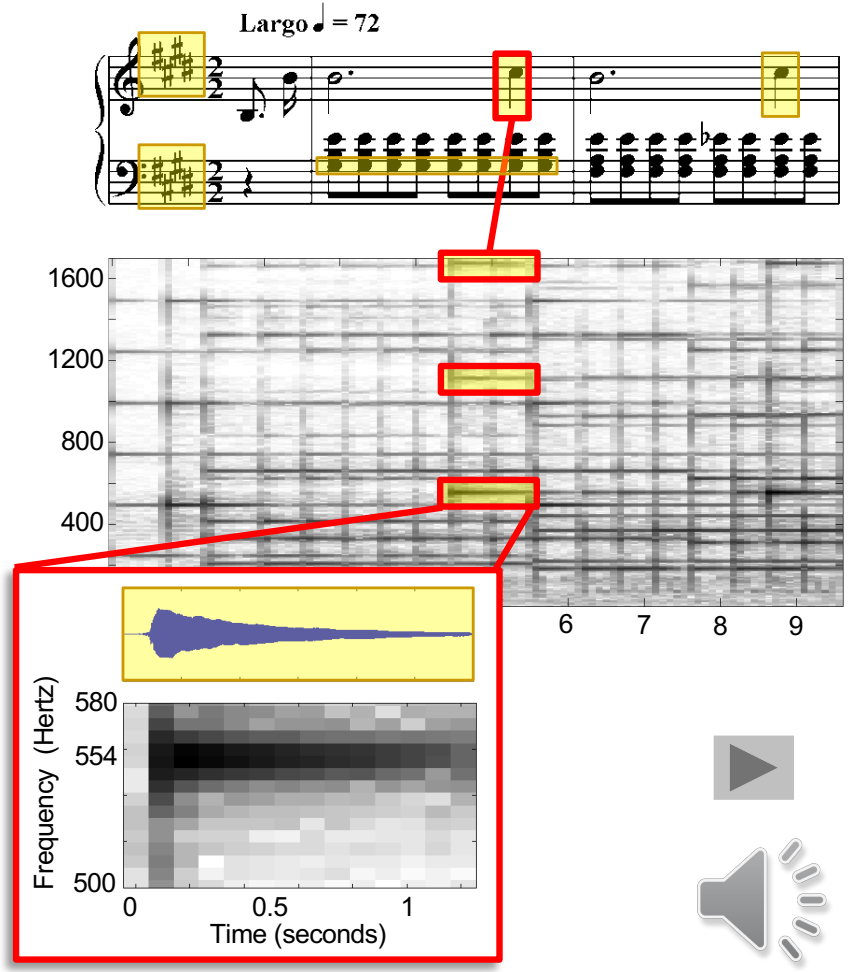
Application: Score-Informed Source Separation



Application: Score-Informed Source Separation

Audio editing

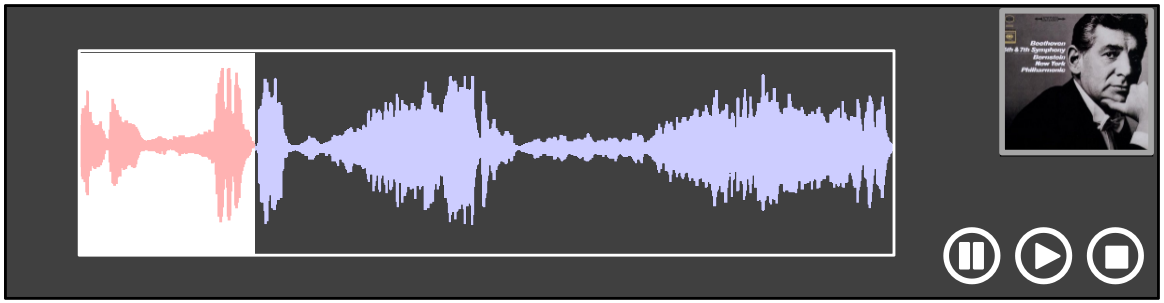




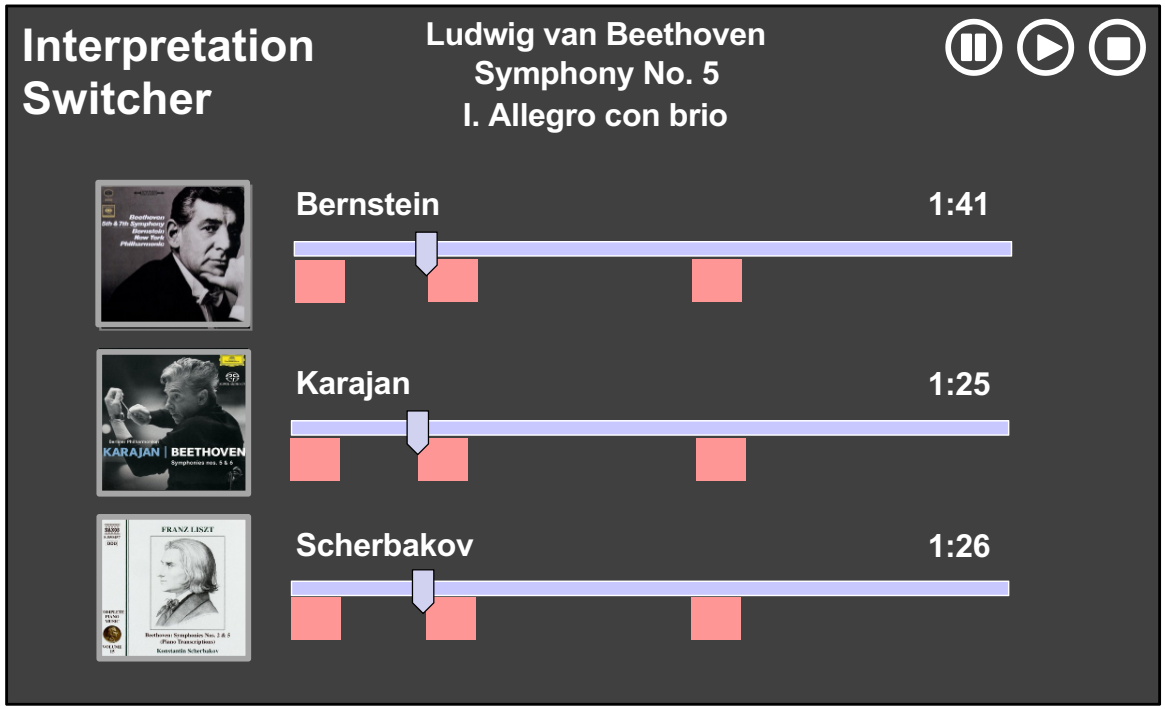
Audio Matching

Task

Query:



Database: Matches

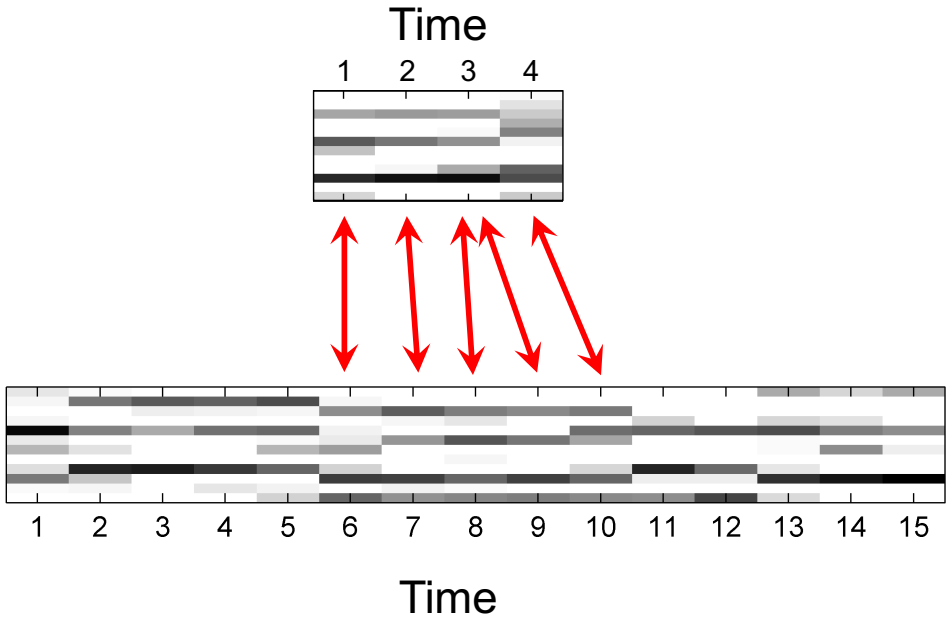


Audio Matching

Task

Query: Sequence *X*

Database: Sequence *Y*

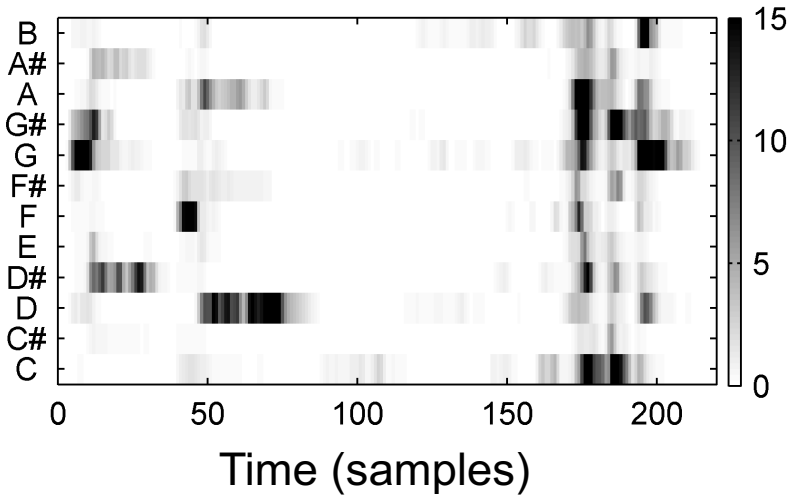


Subsequence matching

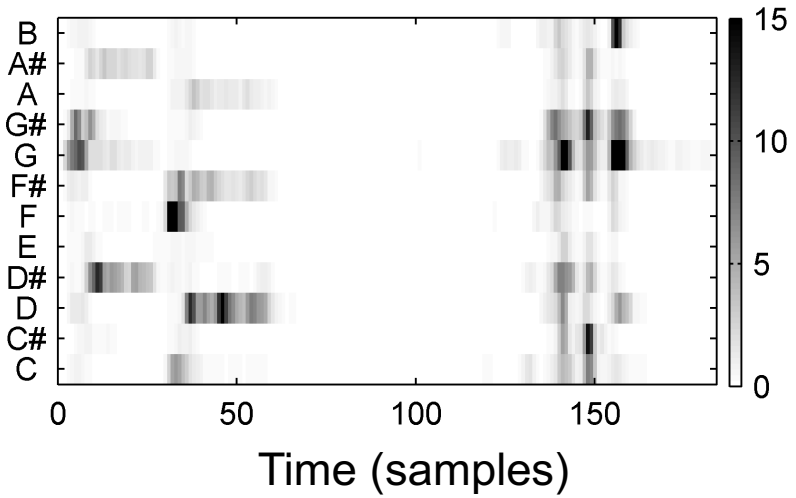
Audio Features

Example: Beethoven's Fifth

Bernstein



Karajan

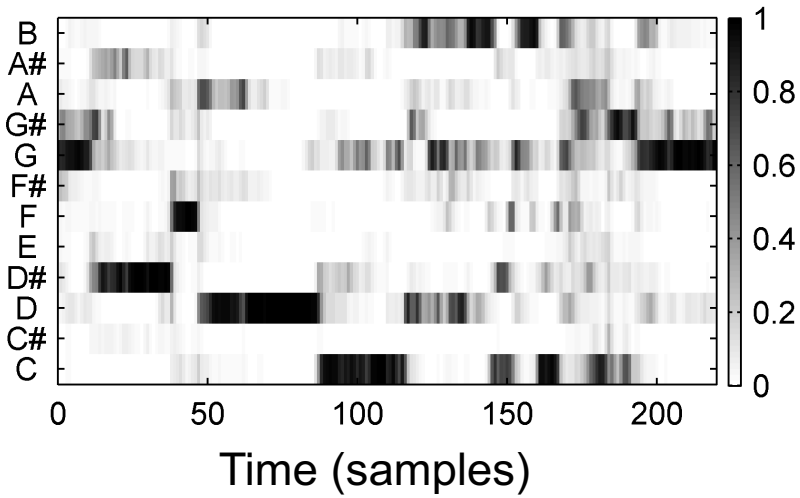


Chroma representation (10 Hz)

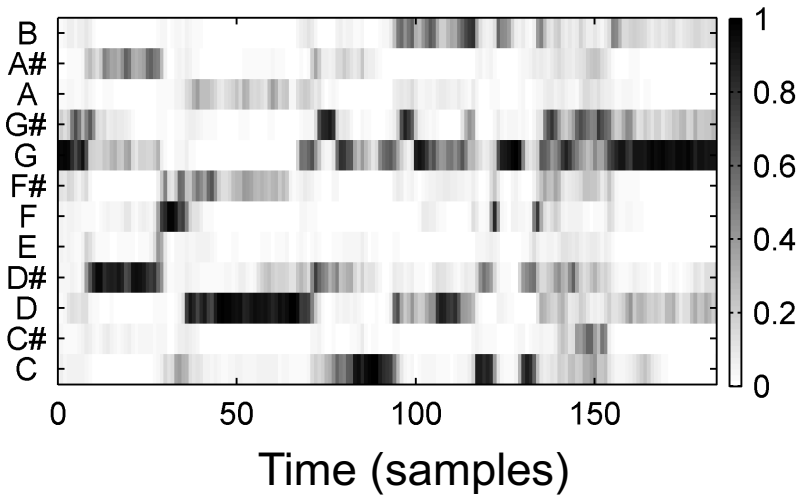
Audio Features

Example: Beethoven's Fifth

Bernstein



Karajan



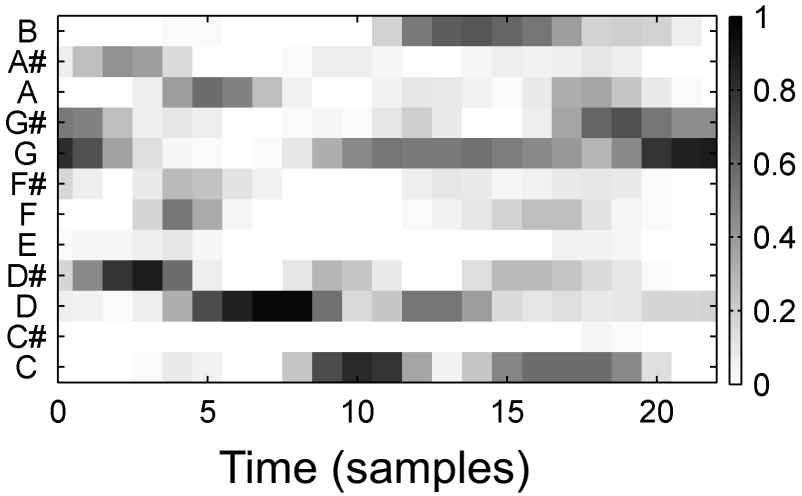
Chroma representation (10 Hz)

- Normalization

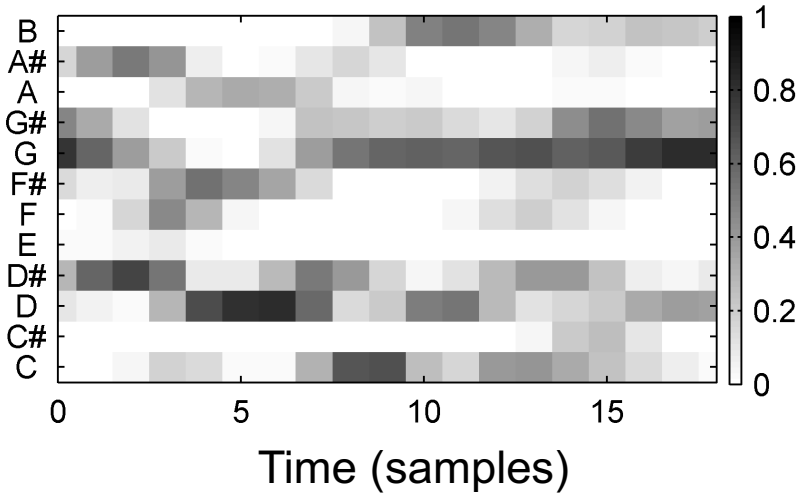
Audio Features

Example: Beethoven's Fifth

Bernstein



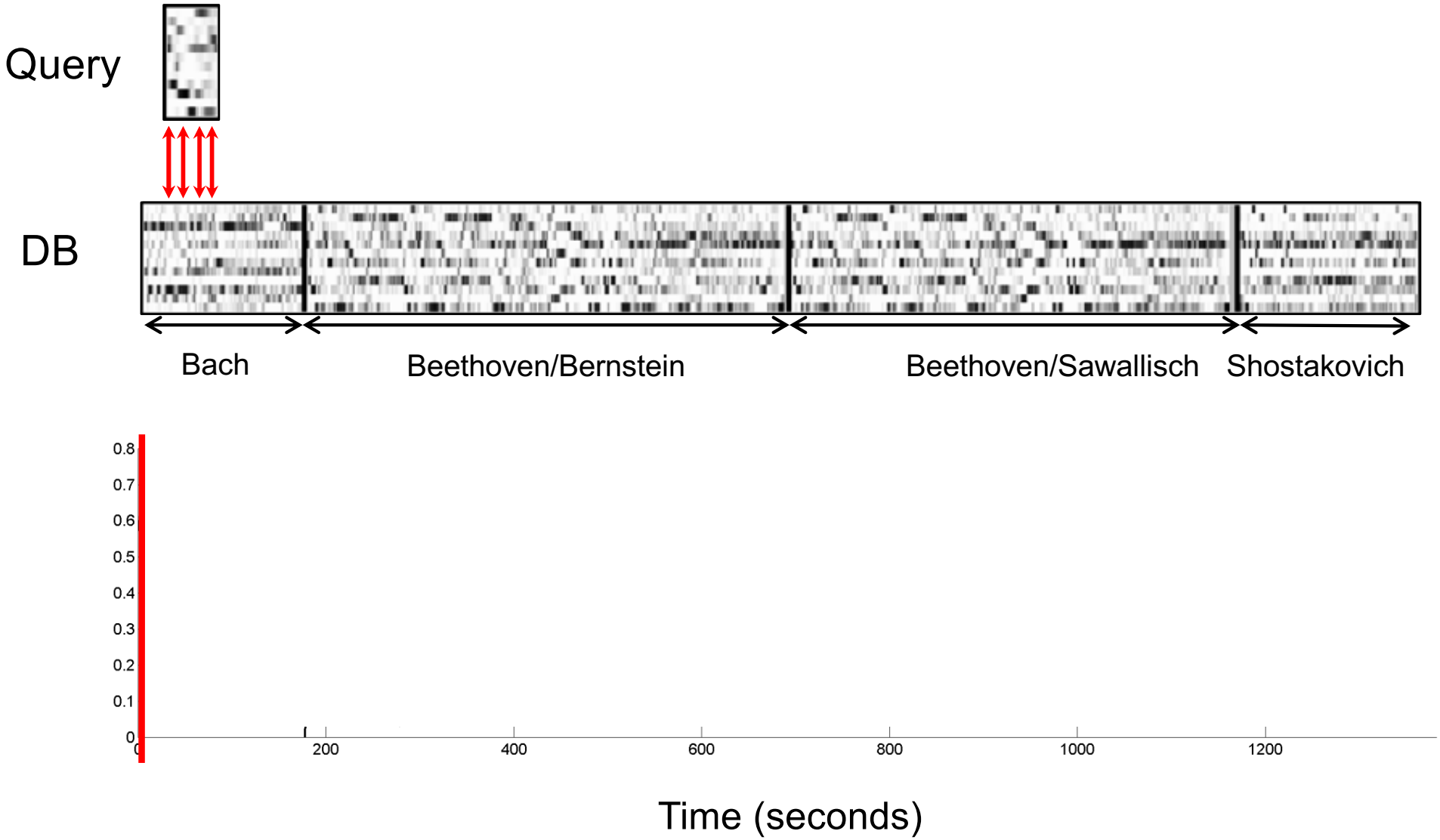
Karajan



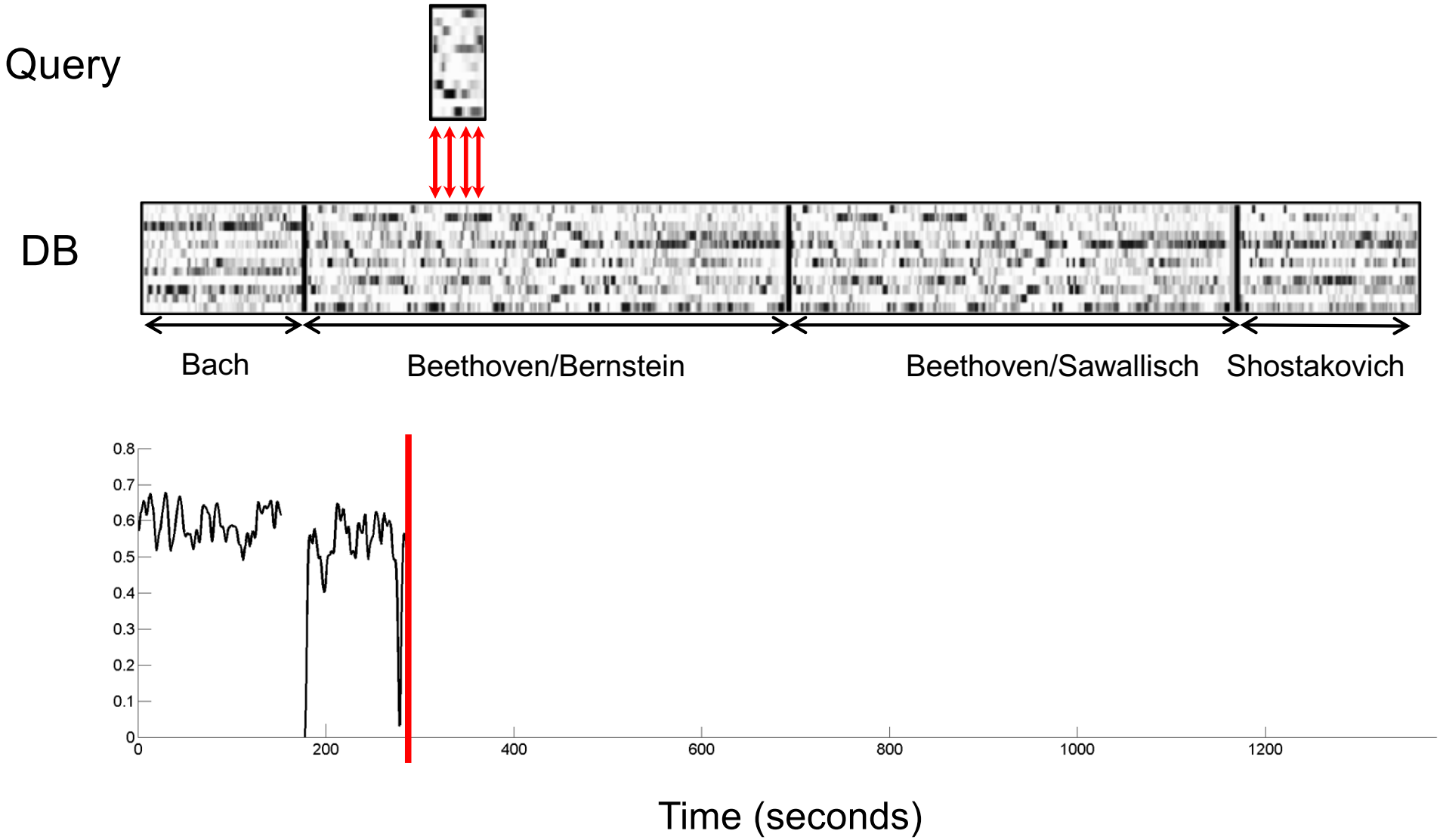
Chroma representation (1 Hz)

- Normalization
- Smoothing & downsampling

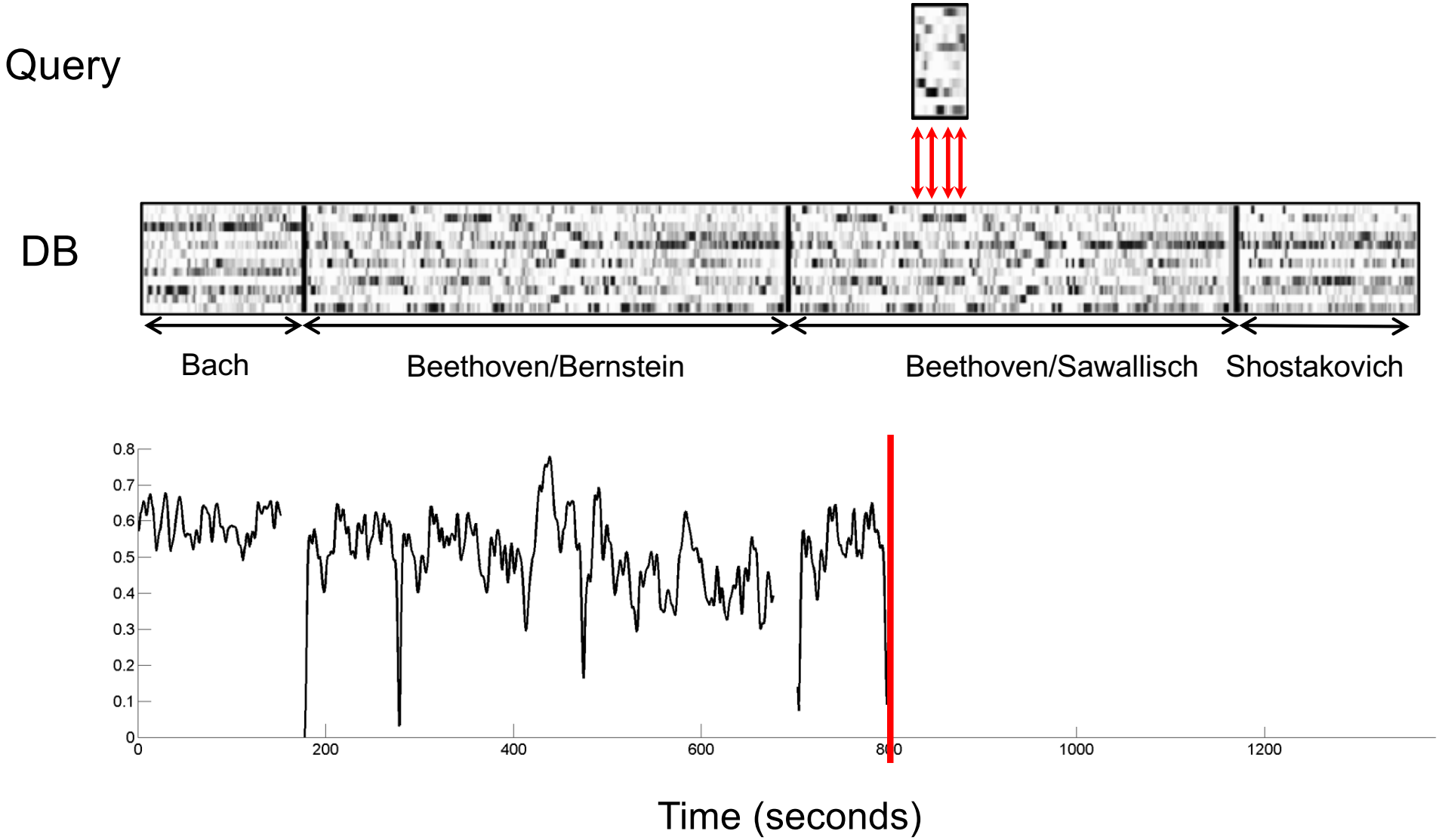
Matching Procedure



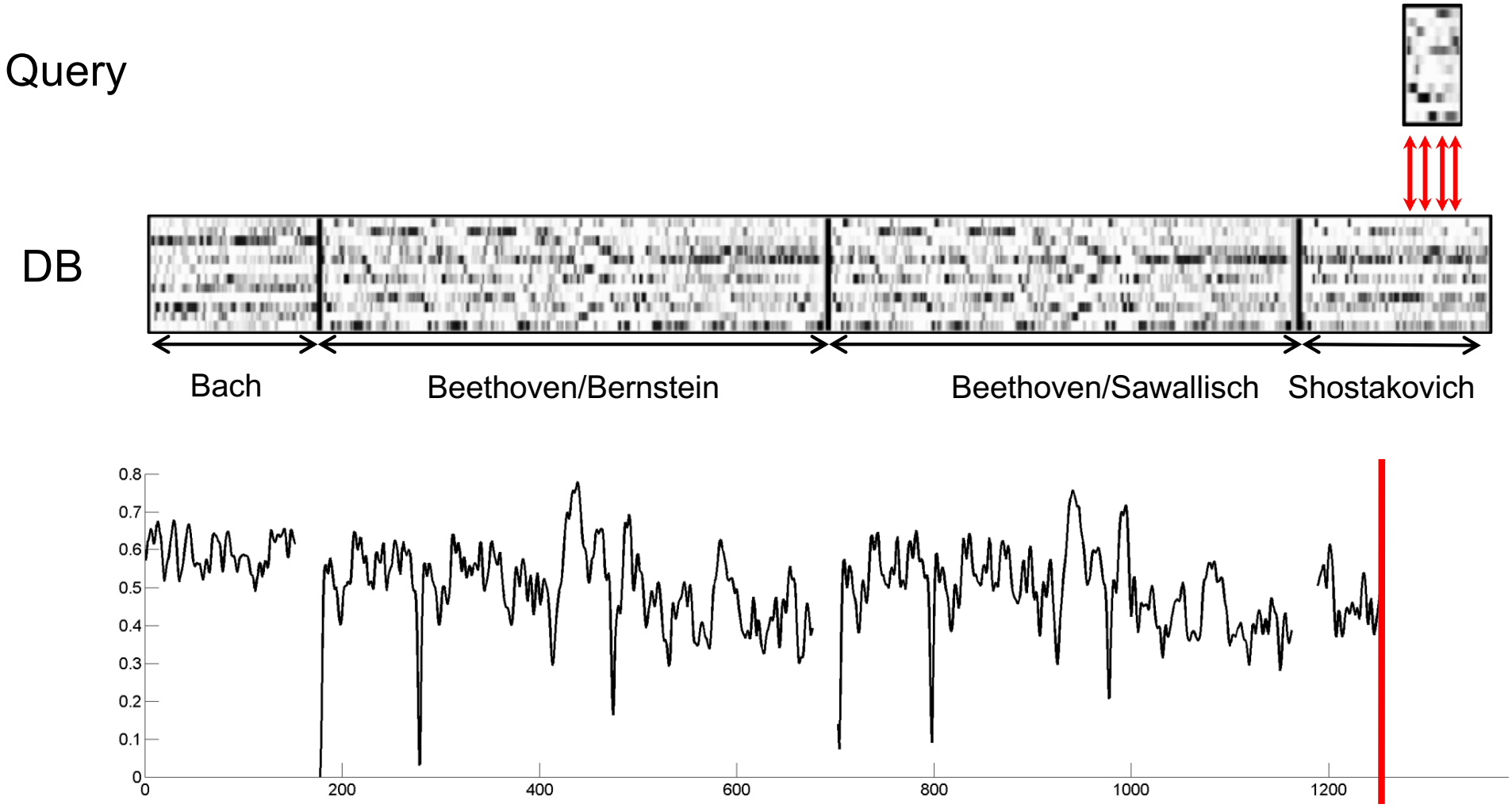
Matching Procedure



Matching Procedure



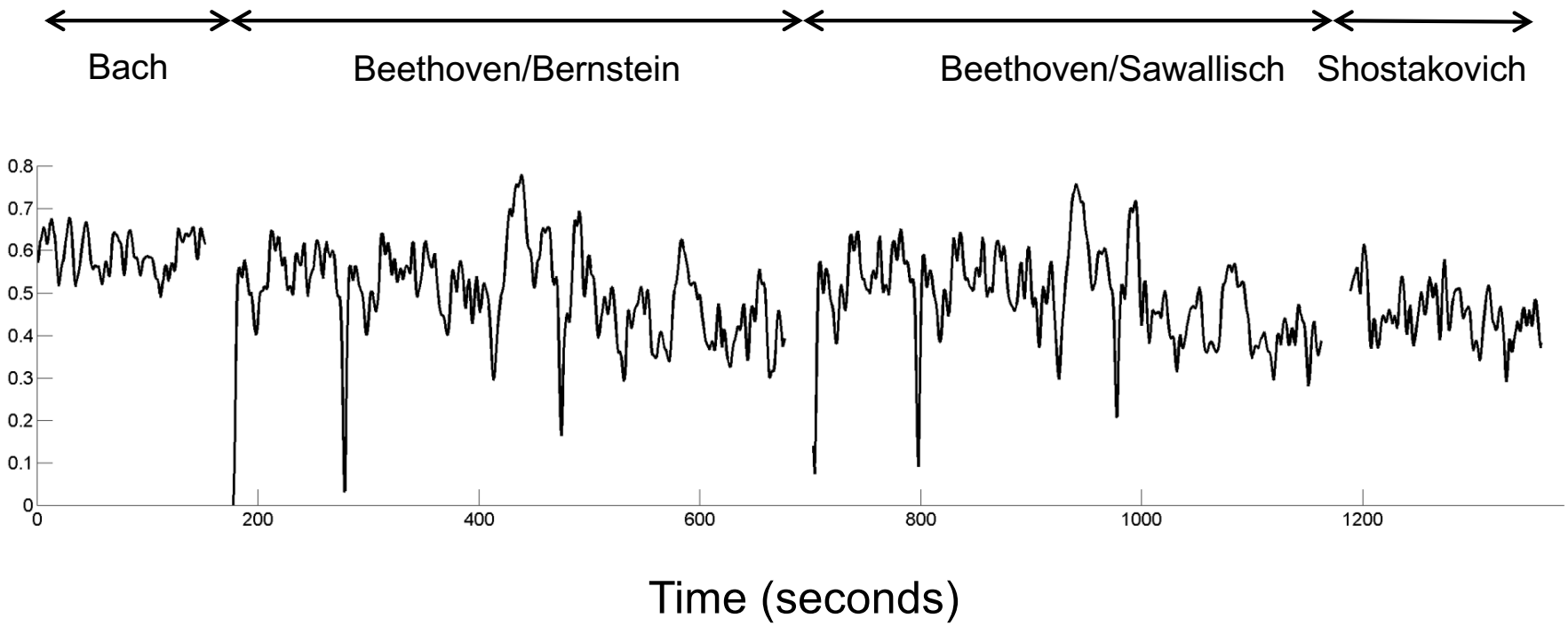
Matching Procedure



Matching Procedure

Matching curve

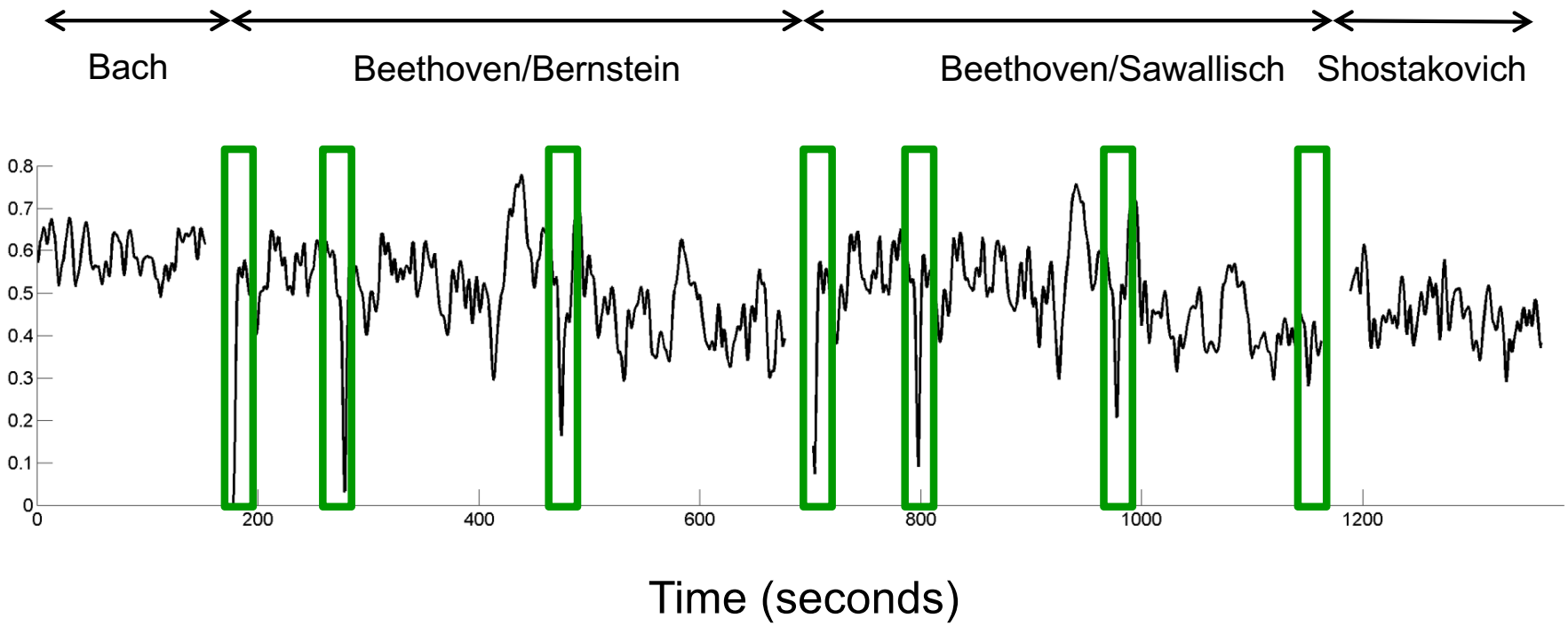
Query: Beethoven's Fifth / Bernstein (first 20 seconds)



Matching Procedure

Matching curve

Query: Beethoven's Fifth / Bernstein (first 20 seconds)

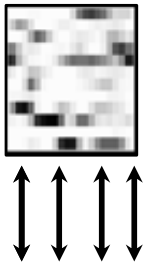


Matches 1 2 5 3 4 6 7

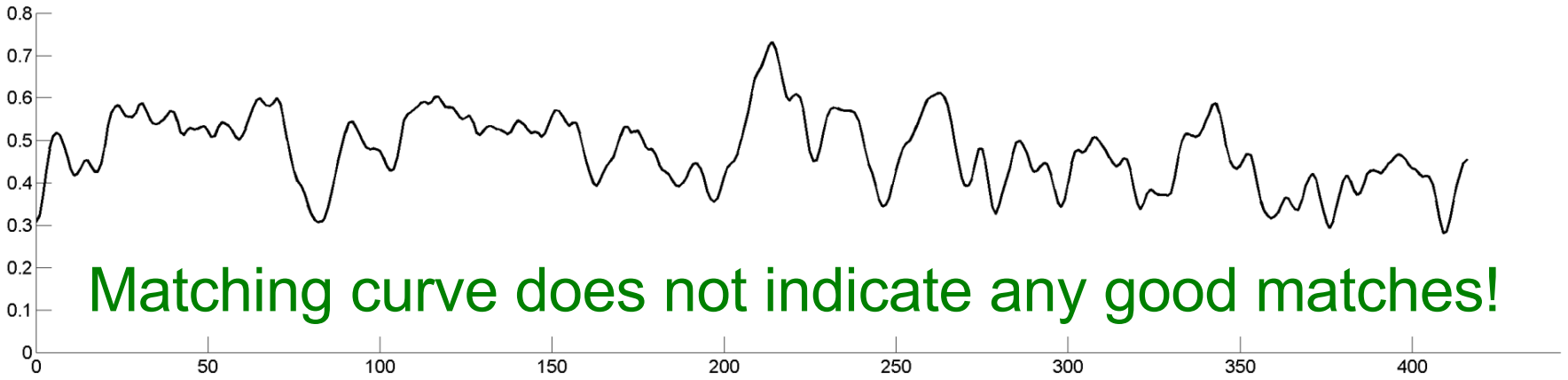
Matching Procedure

Problem: How to deal with tempo differences?

Karajan is much
faster than Bernstein!



Beethoven/Karajan



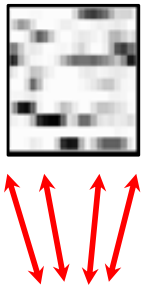
Matching curve does not indicate any good matches!

Time (seconds)

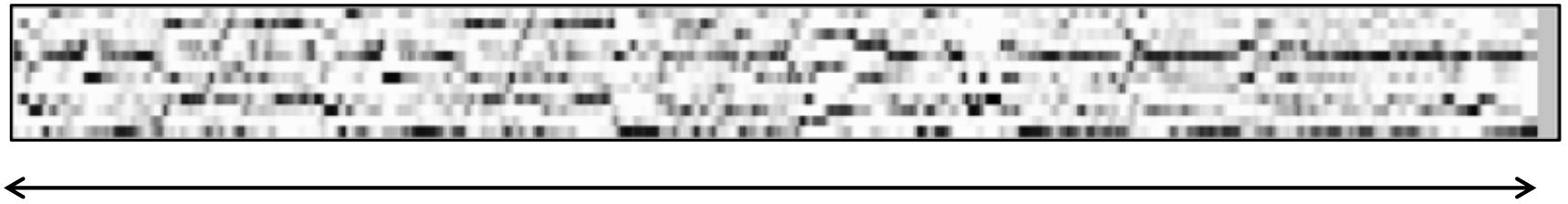
Matching Procedure

1. Strategy: Usage of local warping

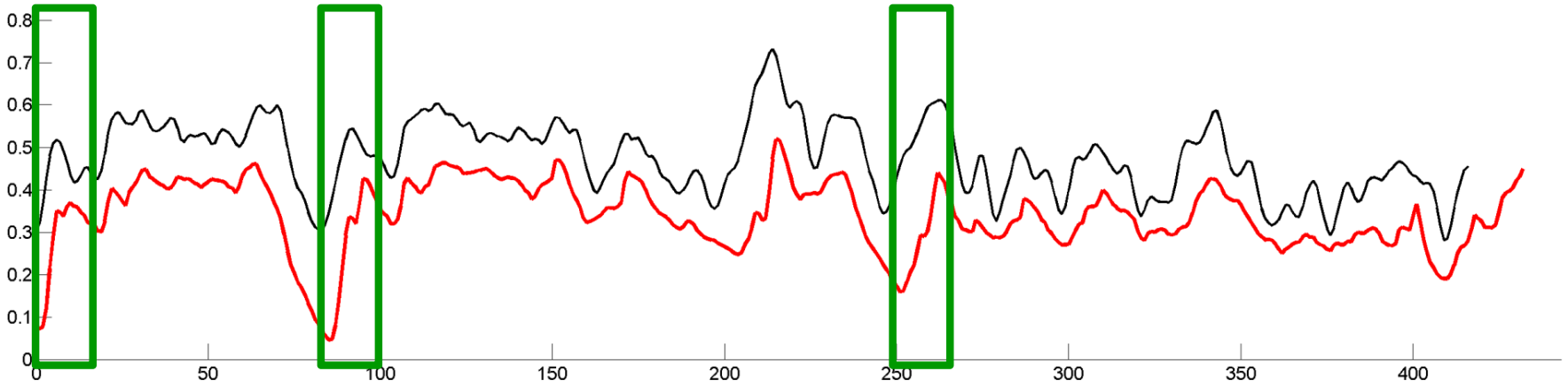
Karajan is much
faster than Bernstein!



Warping strategies are computationally expensive and hard for indexing.



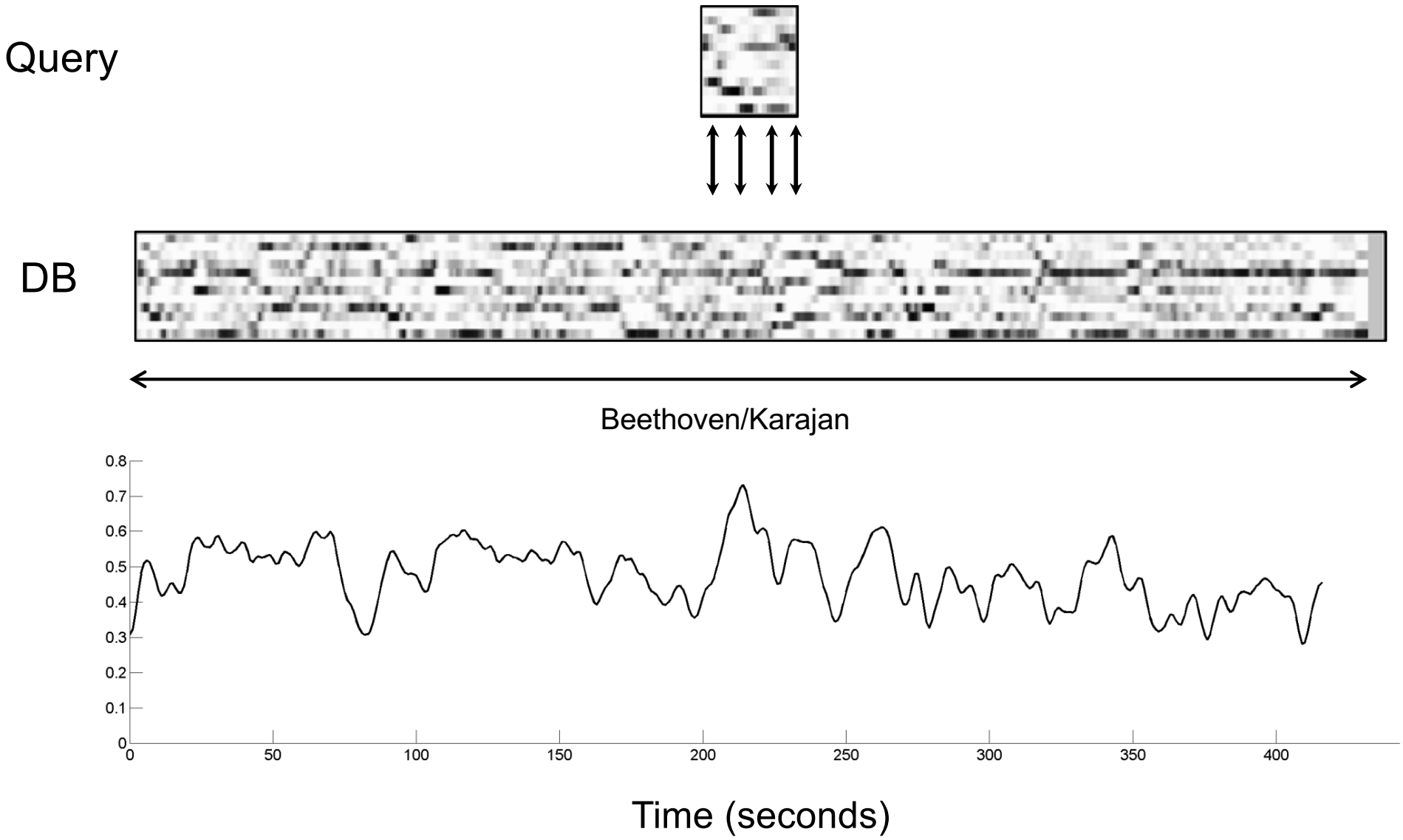
Beethoven/Karajan



Time (seconds)

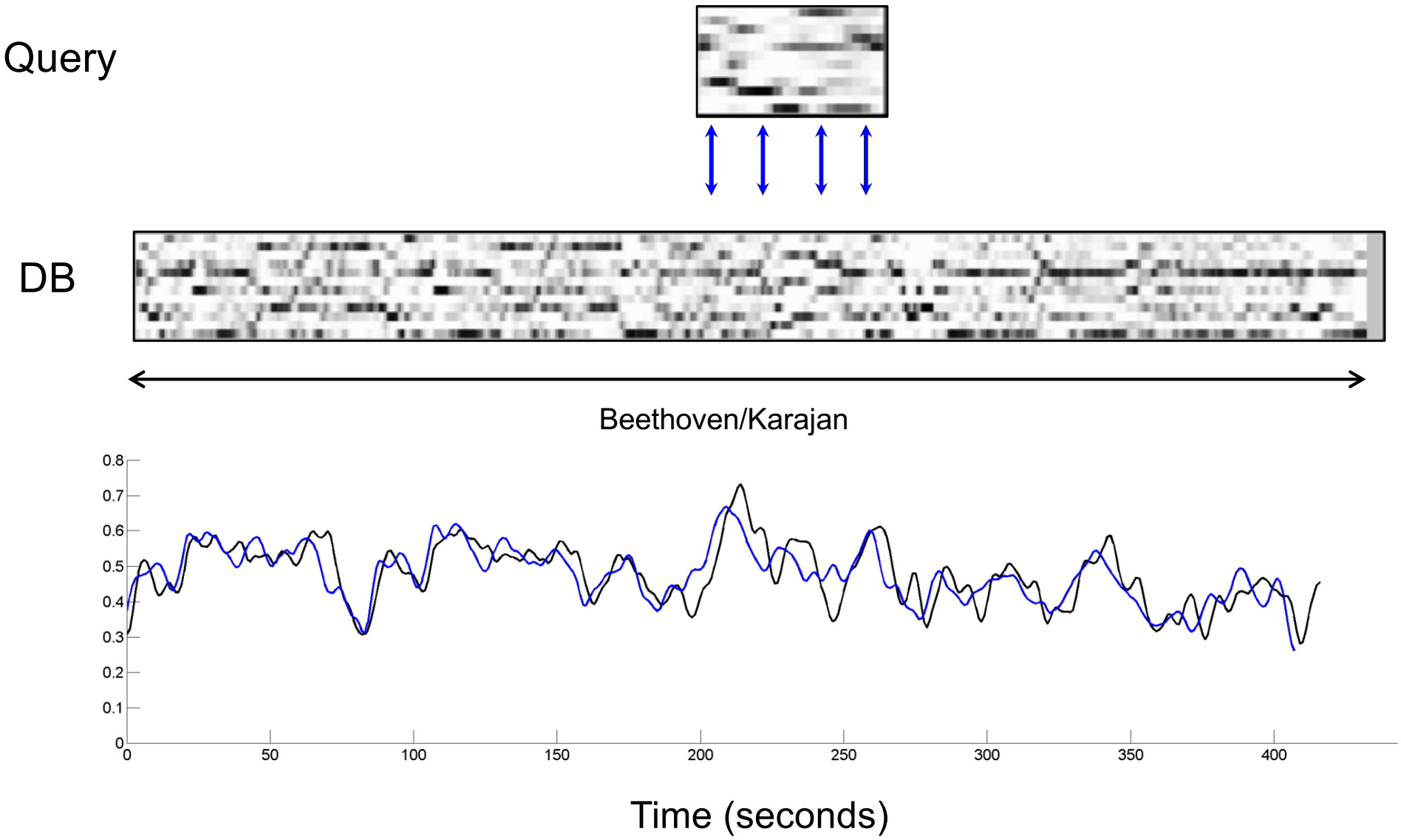
Matching Procedure

2. Strategy: Usage of multiple scaling



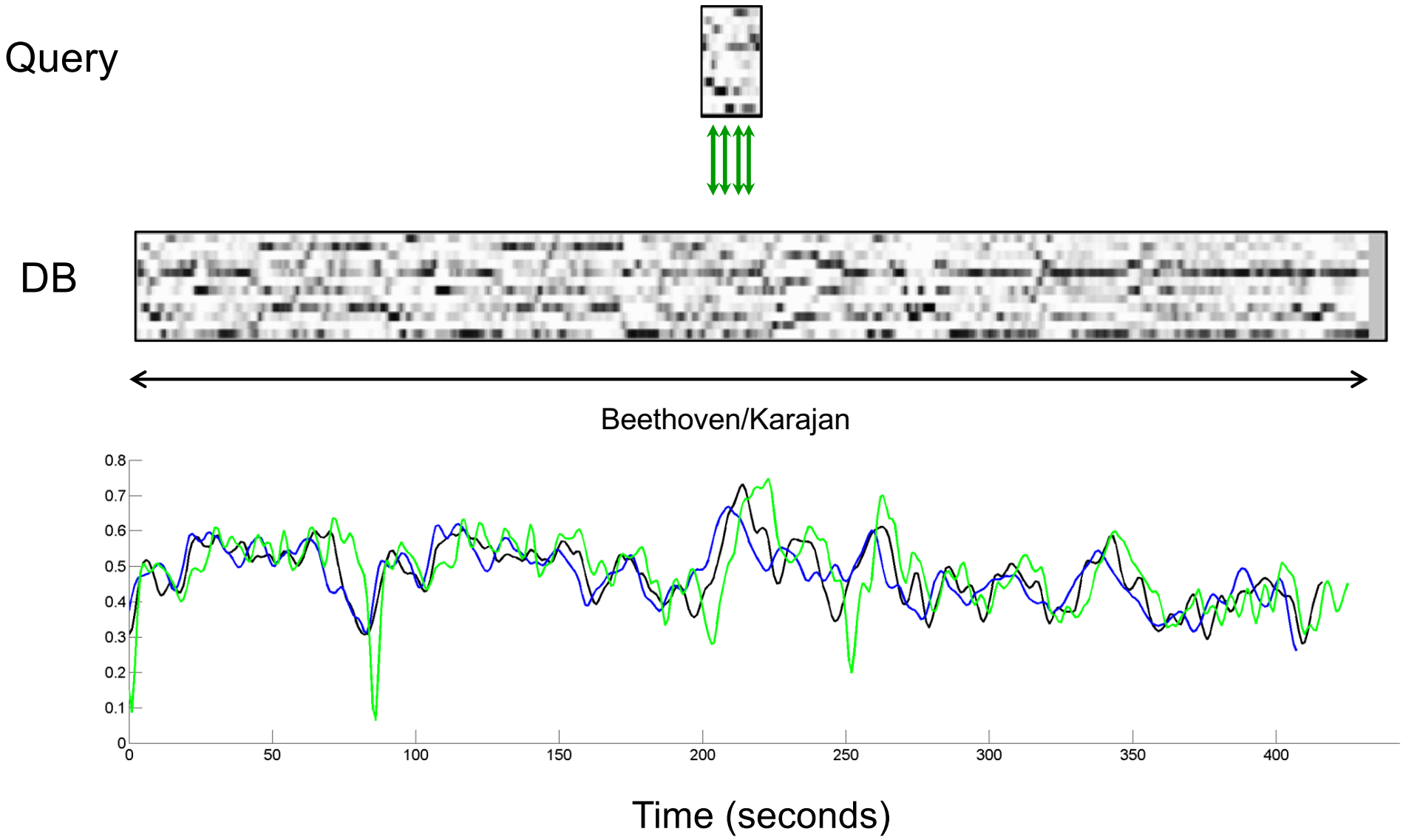
Matching Procedure

2. Strategy: Usage of multiple scaling



Matching Procedure

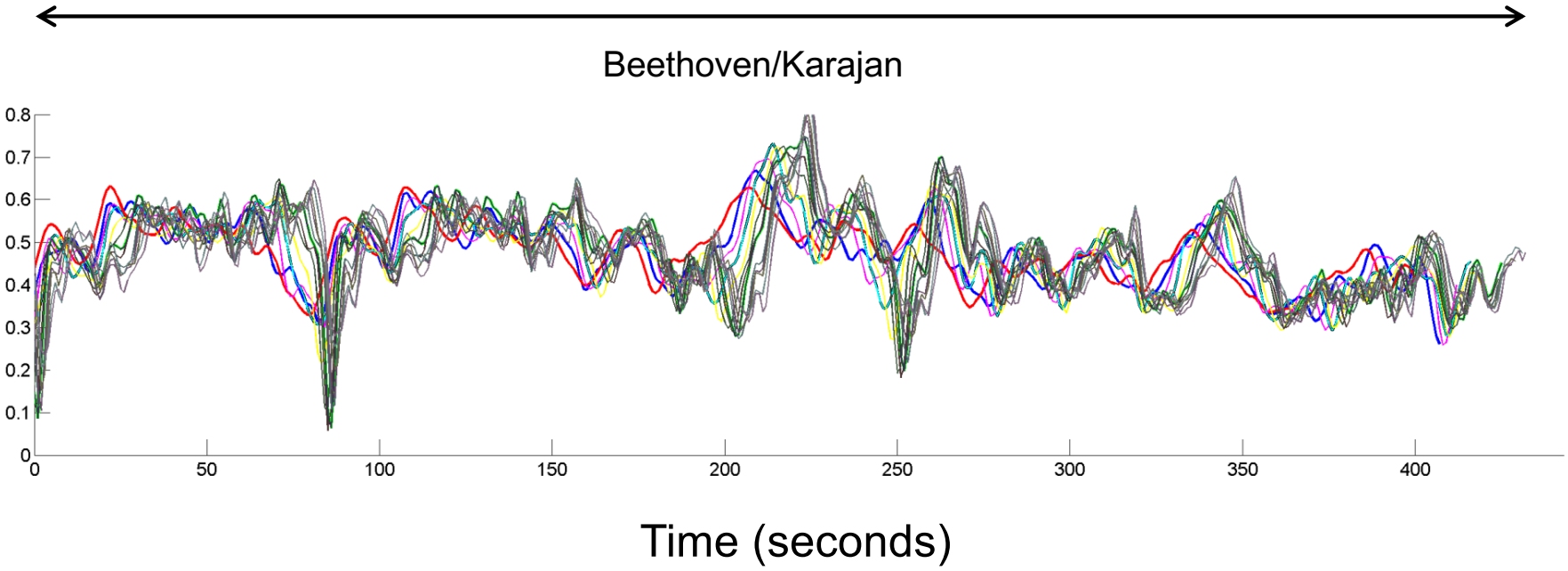
2. Strategy: Usage of multiple scaling



Matching Procedure

2. Strategy: Usage of multiple scaling

Query resampling simulates tempo changes

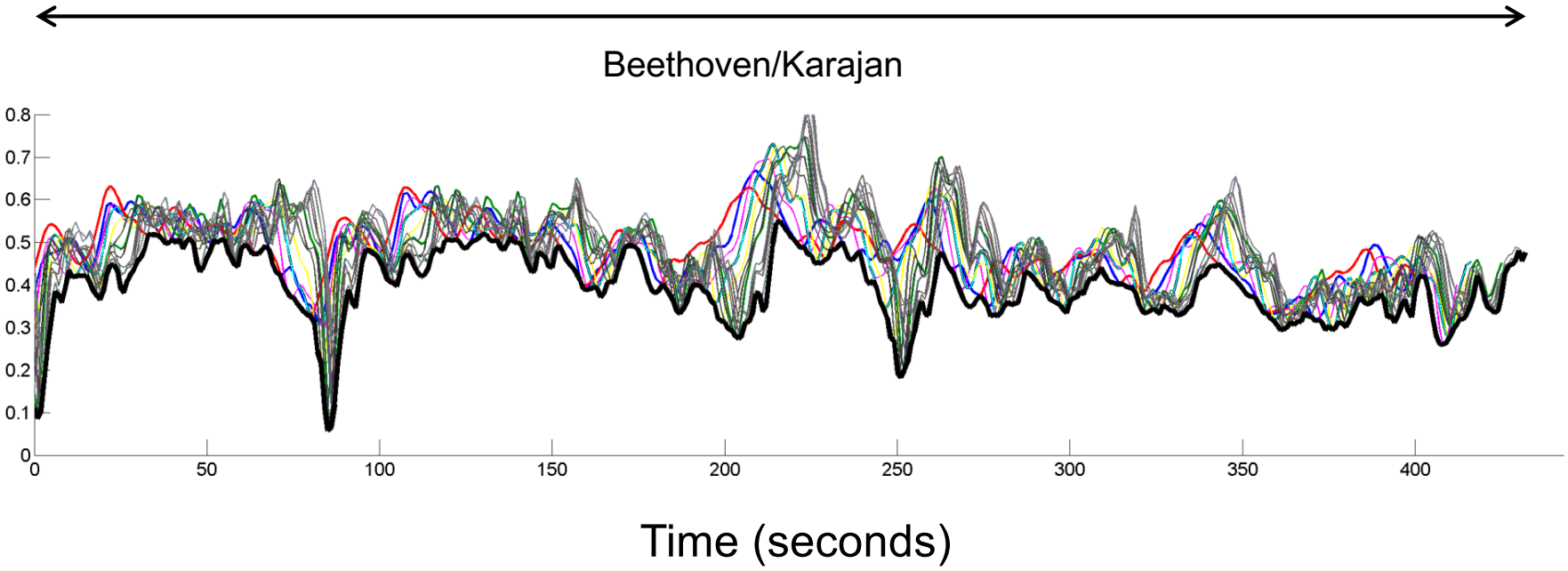


Matching Procedure

2. Strategy: Usage of multiple scaling

Query resampling simulates tempo changes

Minimize over all curves



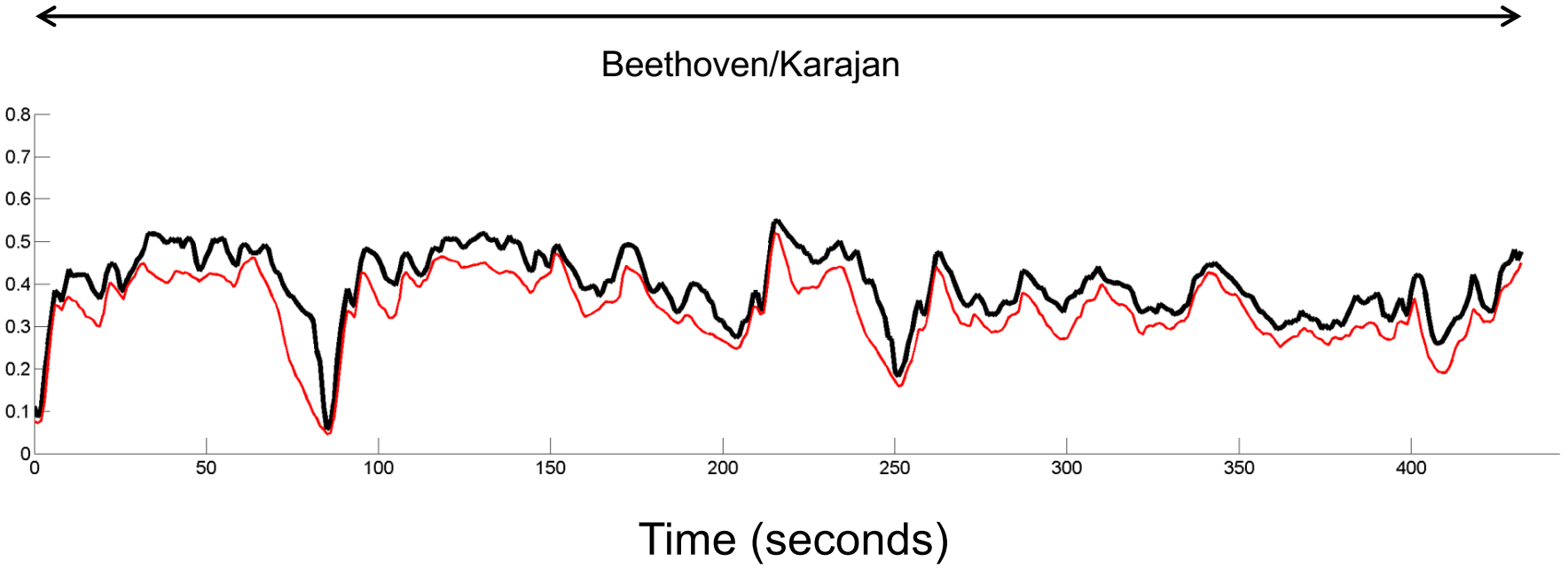
Matching Procedure

2. Strategy: Usage of multiple scaling

Query resampling simulates tempo changes

Minimize over all curves

Resulting curve is similar to warping curve



Audio Matching

Query: Beethoven's Fifth / Bernstein (first 20 seconds)

Rank	Piece	Position
1	Beethoven's Fifth/Bernstein	0 - 21
2	Beethoven's Fifth/Bernstein	101- 122
3	Beethoven's Fifth/Karajan	86 - 103
⋮	⋮	⋮
⋮	⋮	⋮
10	Beethoven's Fifth/Karajan	252 - 271
11	Beethoven's Fifth/Scherbakov	0 - 19
12	Beethoven's Fifth/Sawallisch	275 - 296
13	Beethoven's Fifth/Scherbakov	86 - 103
14	Schumann Op. 97,1/Levine	28 - 43

Audio Matching: Conclusions

Strategy: Handle variations at various levels

- Chroma → invariance to timbre
- Normalization → invariance to dynamics
- Smoothing → invariance to local time deviations
- Multiple queries → invariance to global tempo

Audio Matching: Conclusions

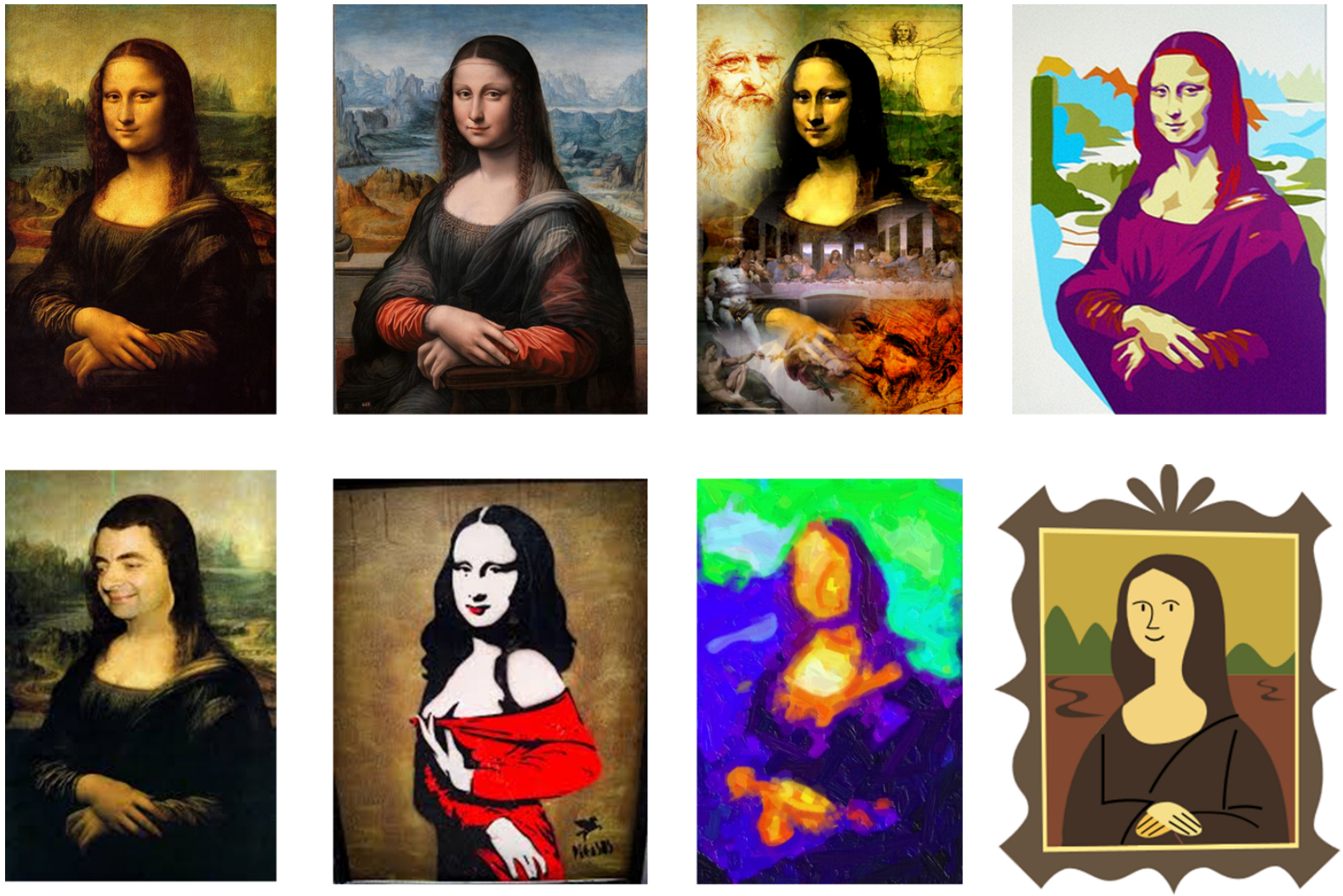
Strategy: Handle variations at various levels

- Chroma → invariance to timbre
- Normalization → invariance to dynamics
- Smoothing → invariance to local time deviations
- Multiple queries → invariance to global tempo

Notes:

- There is no “standard” chroma feature.
  → Variants can make a huge difference!
- Learn invariance from examples
  → “Deep Chroma” [Korzeniowski, Widmer; ISMIR 2016]
- Temporal warping makes problem hard
- Efficiency

Version (Cover Song) Identification



Version (Cover Song) Identification

Nearly anything can change! But something doesn't change.

Often this is chord progression and/or melody

Bob Dylan Knockin' on Heaven's Door	▶	key	▶	Avril Lavigne Knockin' on Heaven's Door
Metallica Enter Sandman	▶	timbre	▶	Apocalyptica Enter Sandman
Nirvana Poly [Incesticide Album]	▶	tempo	▶	Nirvana Poly [Unplugged]
Black Sabbath Paranoid	▶	lyrics	▶	Cindy & Bert Der Hund Der Baskerville
AC/DC High Voltage	▶	recording conditions	▶	AC/DC High Voltage [live]
		song structure		

Version (Cover Song) Identification

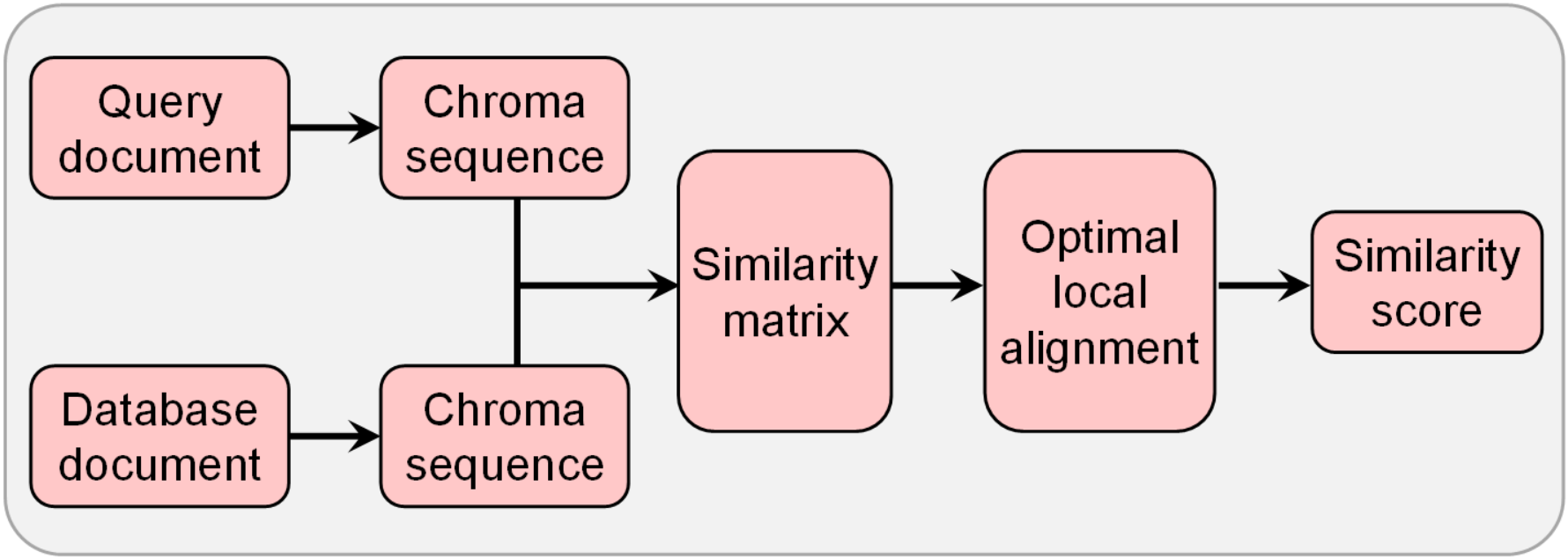
Task

Given a music recording of a song or piece of music as query, find all “similar” music recordings (versions) such as:

- Live versions
- Different interpretations
- Cover songs
- Versions adapted to particular country/region/language
- Contemporary versions of an old song
- Radically different interpretations of a musical piece
- …

Instance of document-based retrieval

Version (Cover Song) Identification



Alignment Strategies

Classical DTW

Music synchronization

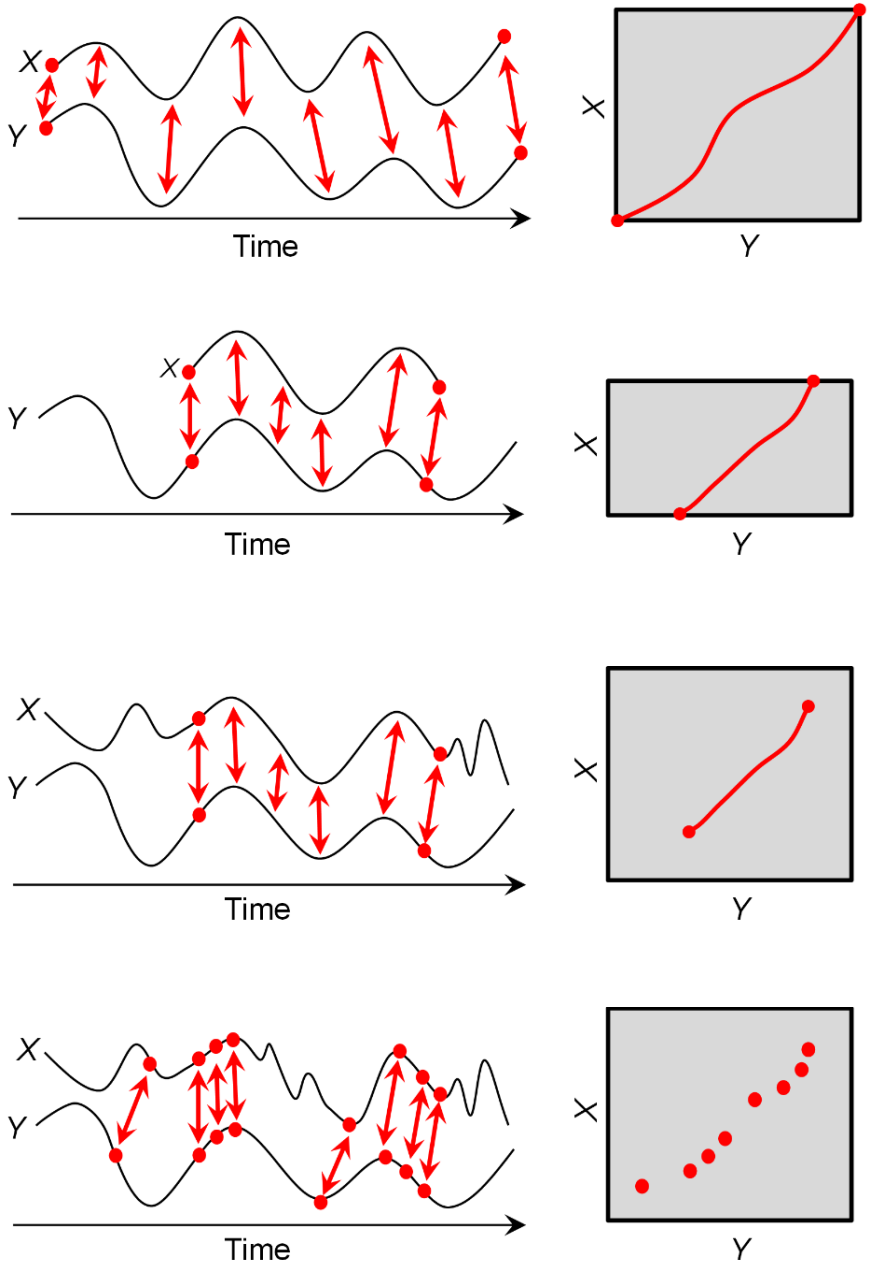
Subsequence DTW

Audio matching

Local alignment

Version (cover song) identification

Partial alignment



Shingle-Based Retrieval

Idea

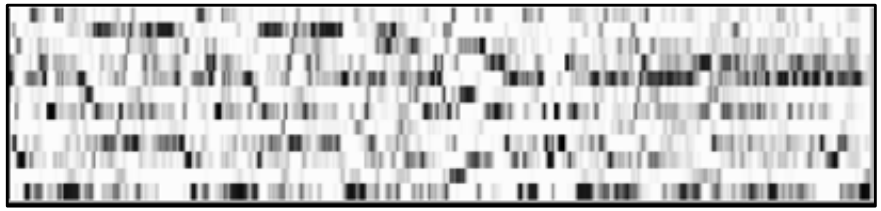
- Query and database are split up into small overlapping shingles that consist of short chroma feature subsequences.
- Shingles can be matched using efficient nearest neighbor retrieval.
- Trade-off:
  - Large shingles have high musical relevance
  - High shingle dimensionality makes indexing difficult

[Casey, Rhodes, Slaney; IEEE TASLP, 2008]
[Grosche, Müller; ICASSP 2012]

Shingle-Based Retrieval

Database

Chroma sequence



Query

Chroma sequence

(ca. 10 to 30 seconds)

Shingle-Based Retrieval

Database
Chroma sequence

Chroma shingles

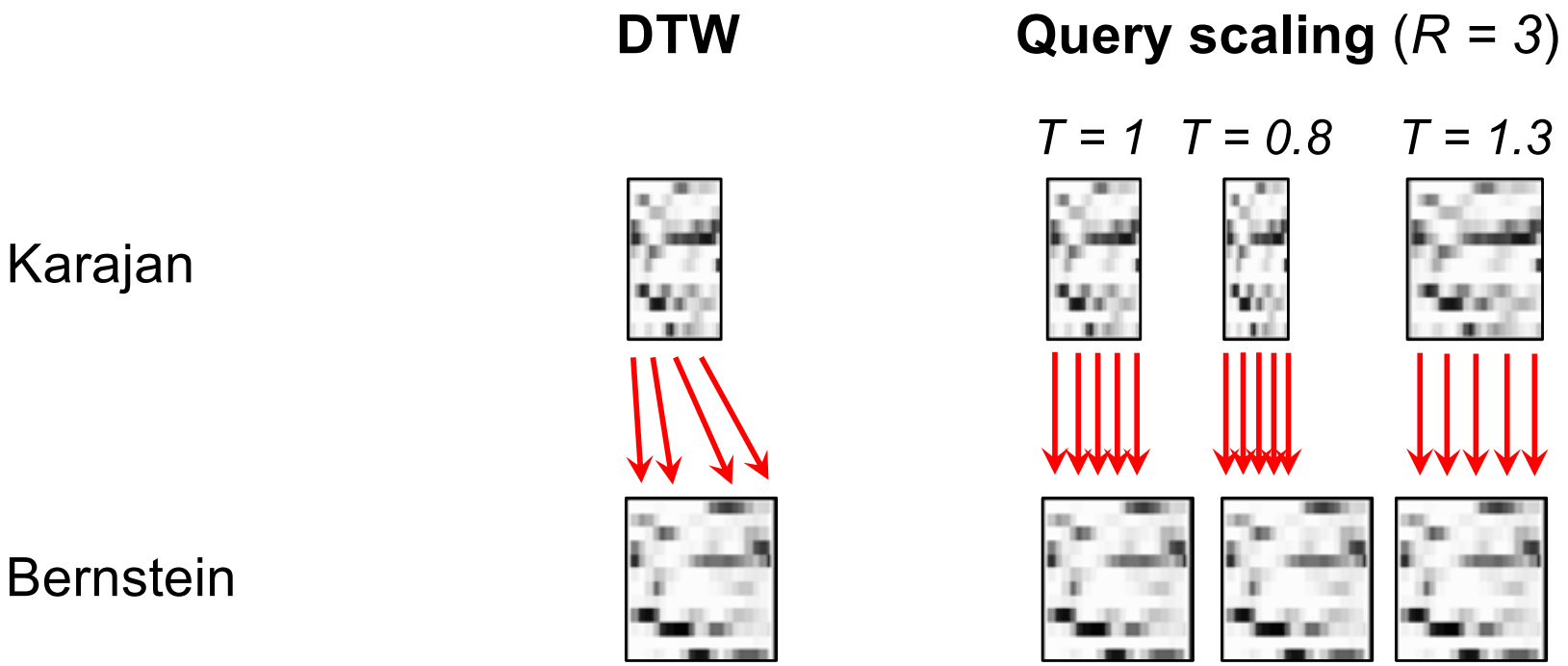
Retrieval
(index-based)

Query
Chroma sequence
(ca. 10 to 30 seconds)

Shingle-Based Retrieval

Tempo-invariant matching

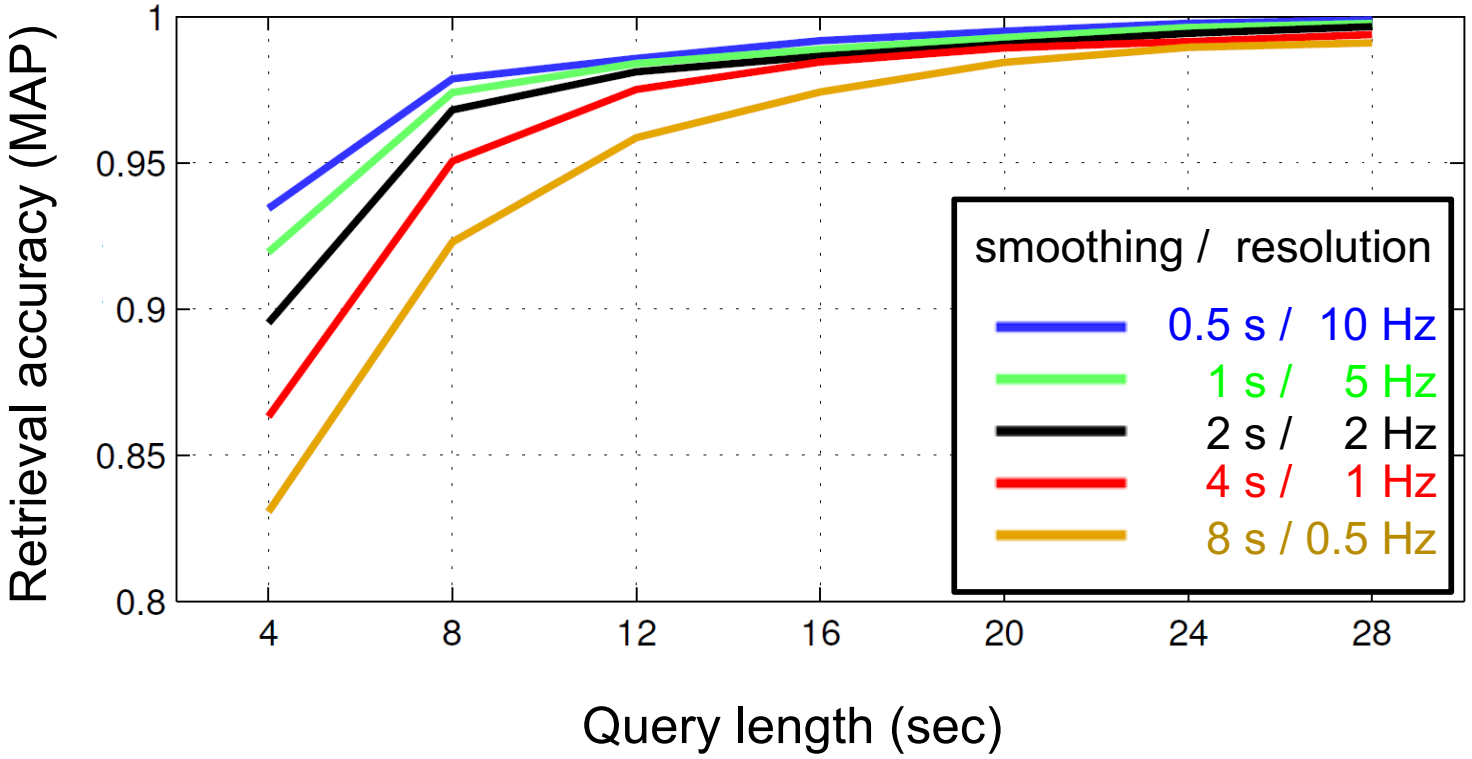
Avoiding expensive temporal warping, tempo differences are handled by creating *R* scaled variants of the query, each simulating a global change in tempo of up to ± 50 %.



Shingle-Based Retrieval

Query length and feature type

Dataset: 2484 recordings (162 hours)
Per query 10-50 relevant documents

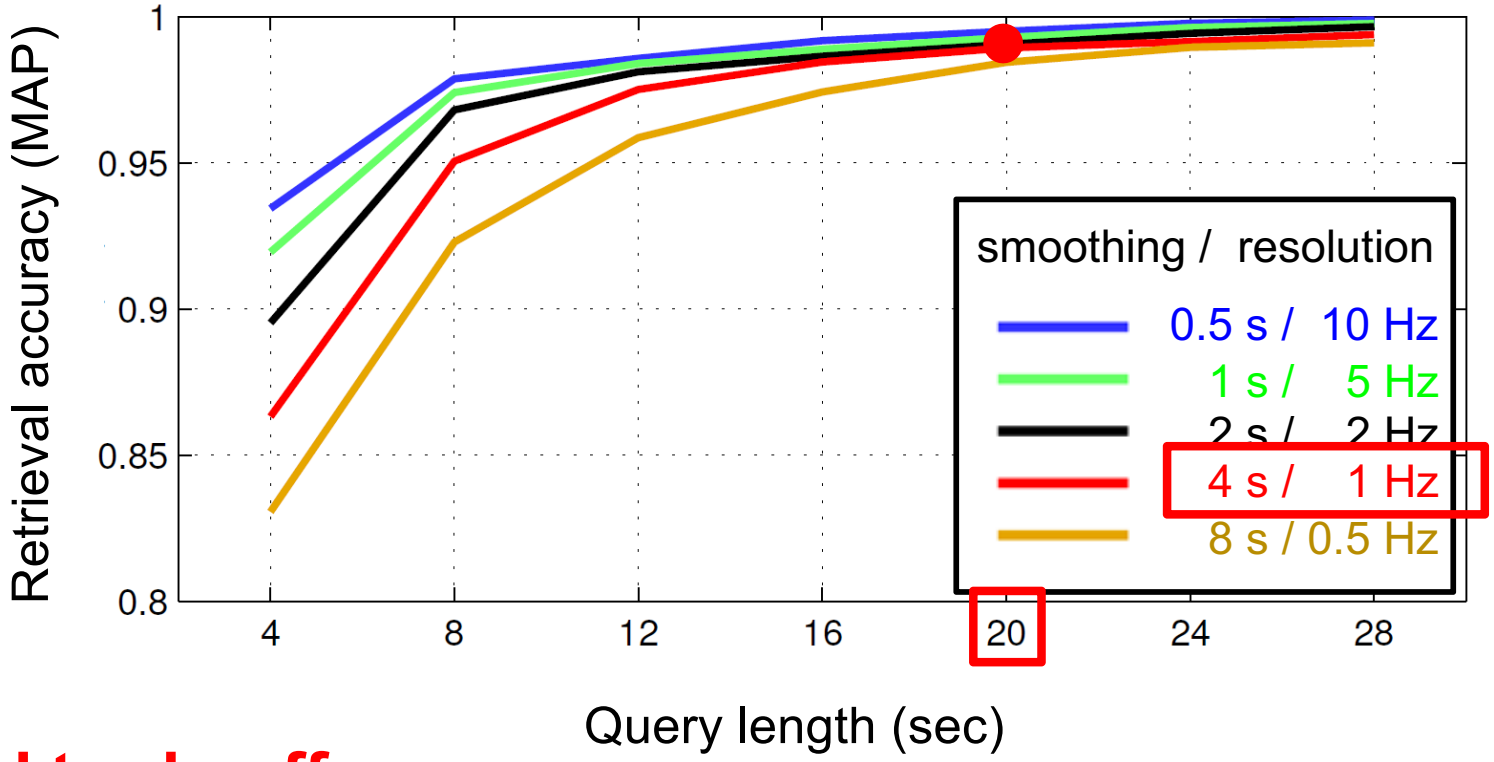


[Grosche, Müller; ICASSP 2012]

Shingle-Based Retrieval

Query length and feature type

Dataset: 2484 recordings (162 hours)
Per query 10-50 relevant documents



Good trade-off

- Query length = 20 sec
- Feature type: 4 s / 1 Hz

[Grosche, Müller; ICASSP 2012]

Shingle-Based Retrieval

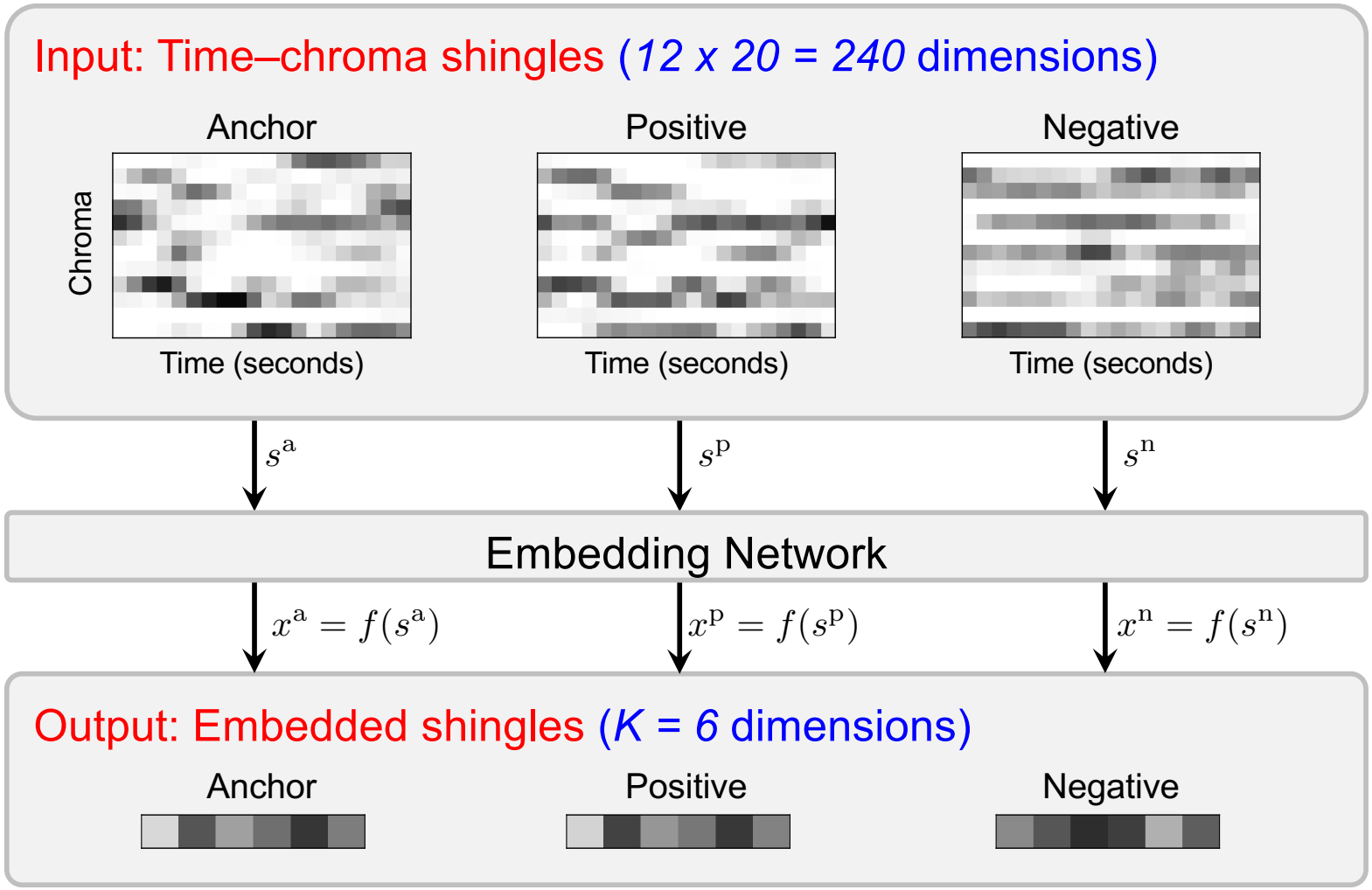
- Time-chroma shingle: 12 x 20 = 240 dimensions
- Indexing via Locality-Sensitive Hashing
  - Speedup factor of 25 with MAP > 0.9
  - Speedup factor of 100 with MAP > 0.8
- Further reduction of shingle dimensionality?
  - Linear embedding using PCA
  - Non-linear embedding using deep learning

[Grosche, Müller; ICASSP 2012]

Shingle-Based Retrieval

[Schroff et al.; CVPR 2015 (FaceNet)]
[Zalkow, Müller; submitted]

Strategy: Learn embedding using Siamese NN with triplet loss



Shingle-Based Retrieval

[Zalkow, Müller; submitted]

Retrieval accuracy (MAP)

Dimension K	240	30	20	15	10	7	5	3
PCA	0.97	0.95	0.93	0.90	0.87	0.79	0.76	0.58
NN	0.97	0.96	0.94	0.94	0.92	0.87	0.80	0.68

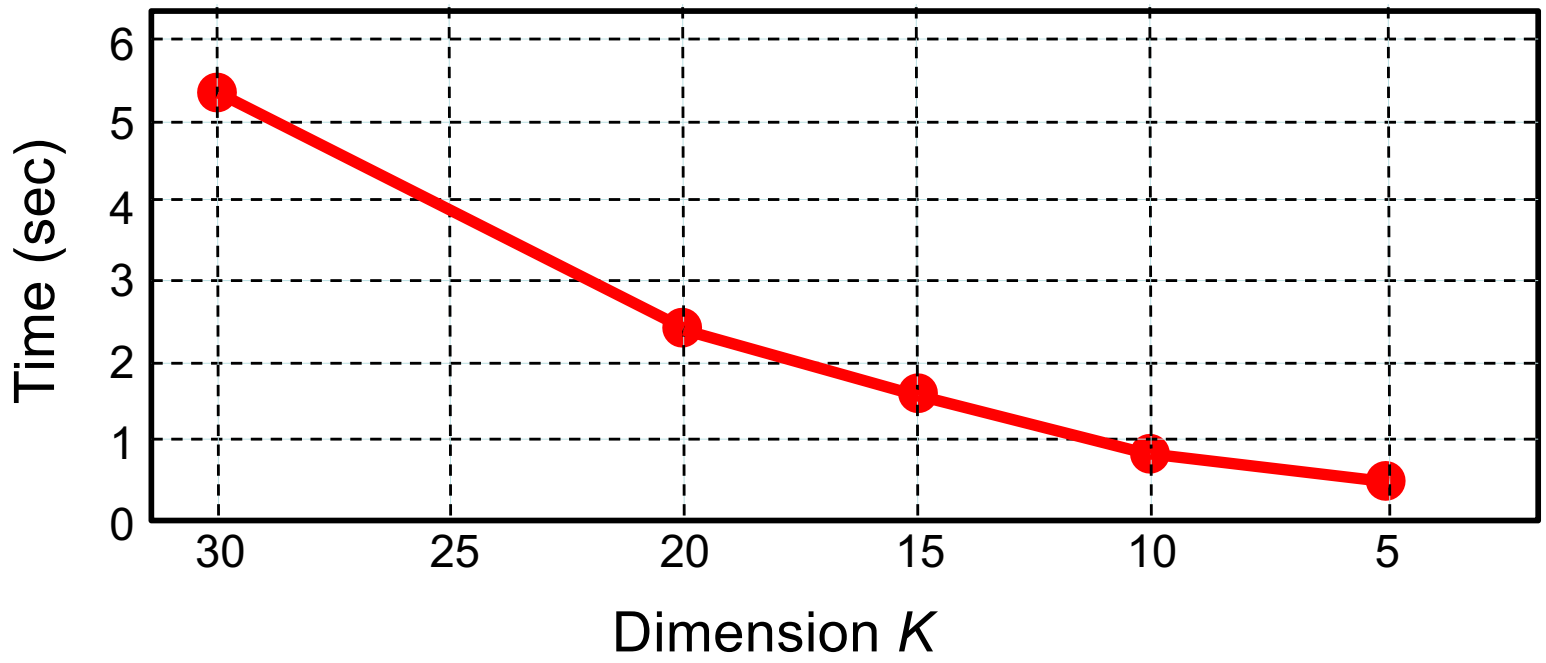
Shingle-Based Retrieval

[Zalkow, Müller; submitted]

Retrieval accuracy (MAP)

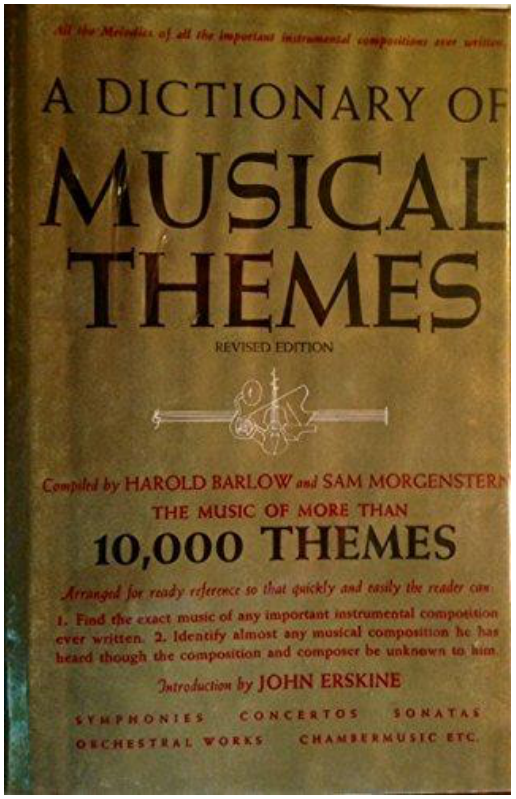
Dimension K	240	30	20	15	10	7	5	3
PCA	0.97	0.95	0.93	0.90	0.87	0.79	0.76	0.58
NN	0.97	0.96	0.94	0.94	0.92	0.87	0.80	0.68

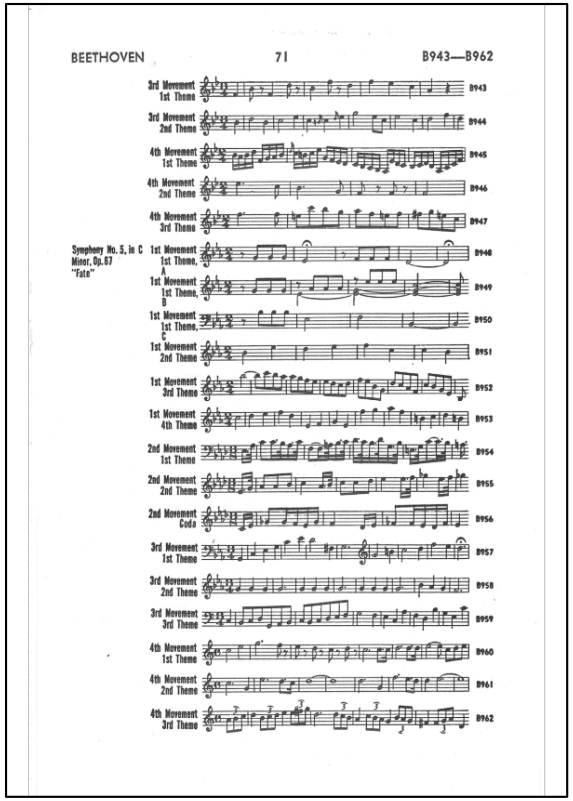
Running time (k-d tree implementation, 3300 queries)



Cross-Modal Music Retrieval

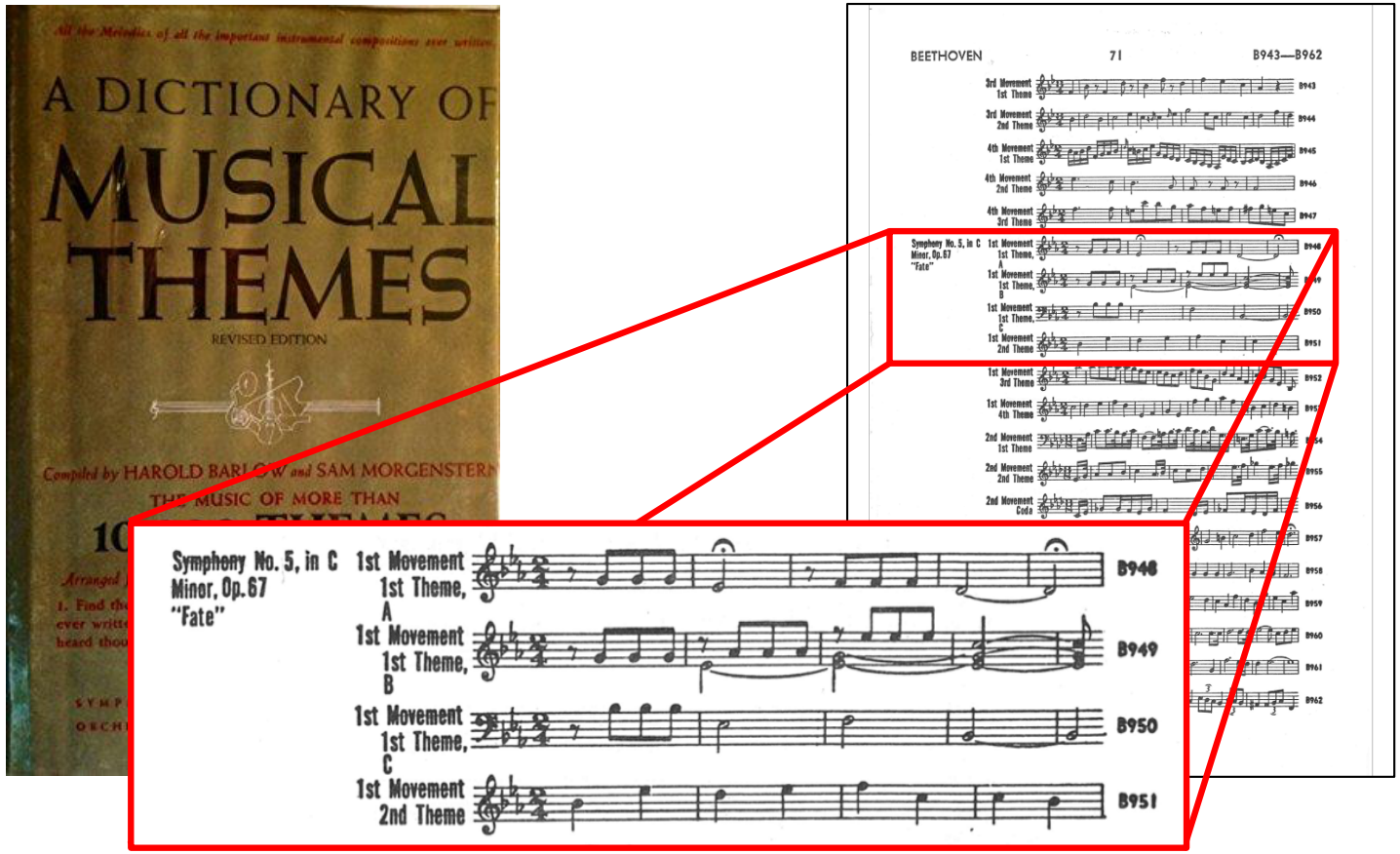
Barlow & Morgenstern (1949): A Dictionary of Musical Themes





Cross-Modal Music Retrieval

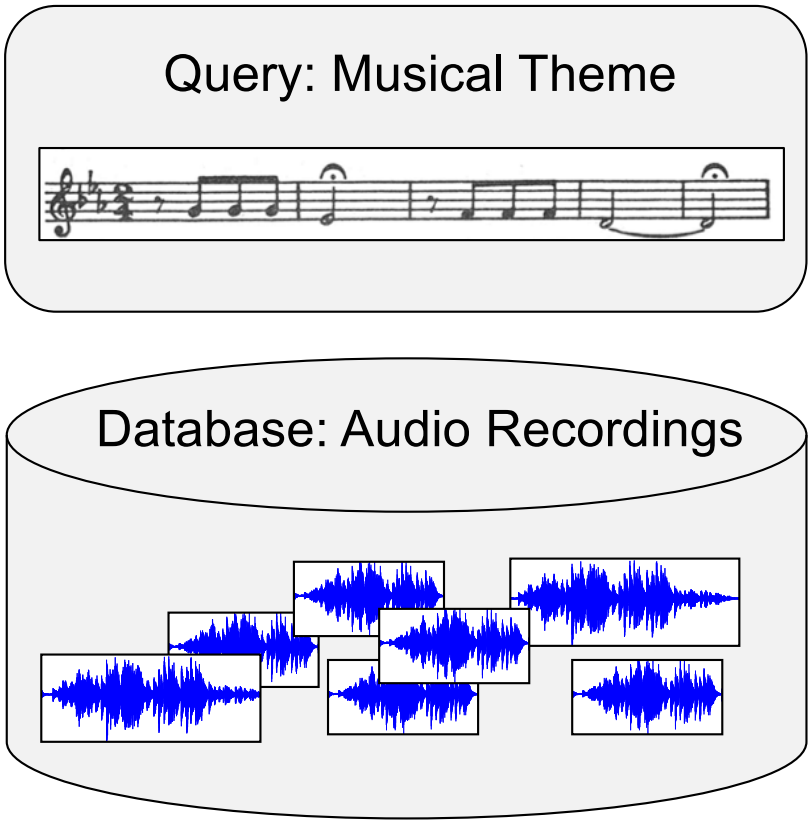
Barlow & Morgenstern (1949): A Dictionary of Musical Themes



10,000 Themes from Western classical music

Cross-Modal Music Retrieval

Barlow & Morgenstern (1949): A Dictionary of Musical Themes



Challenges

- **Cross-modality**
  Symbolic vs. audio data
- **Tuning**
  Deviations from standard tuning
- **Transposition**
  Played key vs. written key
- **Tempo**
  Local & global tempo deviations
- **Polyphony**
  Monophonic query vs. polyphonic audio

Cross-Modal Music Retrieval

Retrieval Experiment

#Queries: 2045 themes

#Database: 1114 recordings (120 hours)

Balke et al.; ICASSP 2016]

	Top-1	Top-20	Top-50
Tuning	18.3	29.2	46.1
Transposition & query length	39.5	66.9	76.1

Cross-Modal Music Retrieval

Retrieval Experiment

#Queries: 2045 themes

#Database: 1114 recordings (120 hours)

[Balke et al.; ICASSP 2016]

	Top-1	Top-20	Top-50
Tuning	18.3	29.2	46.1
Transposition & query length	39.5	66.9	76.1

[Zalkow, Balke, Müller; ICASSP 2019]

Feature Type	Top-1	Top-20	Top-50
Chroma (filter bank, IIS)	47.0	70.0	79.2
Chroma (melody extraction, MEL)	23.1	50.0	59.9
Chroma (salience, BG1)	75.4	88.5	91.3
Chroma (deep learning, CNN)	69.3	85.3	89.6

https://www.audiolabs-erlangen.de/resources/MIR/2019-ICASSP-BarlowMorgenstern

Cross-Modal Music Retrieval