Tutorial
Cross-Modal Music Retrieval and Applications

Part II: Fingerprinting 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

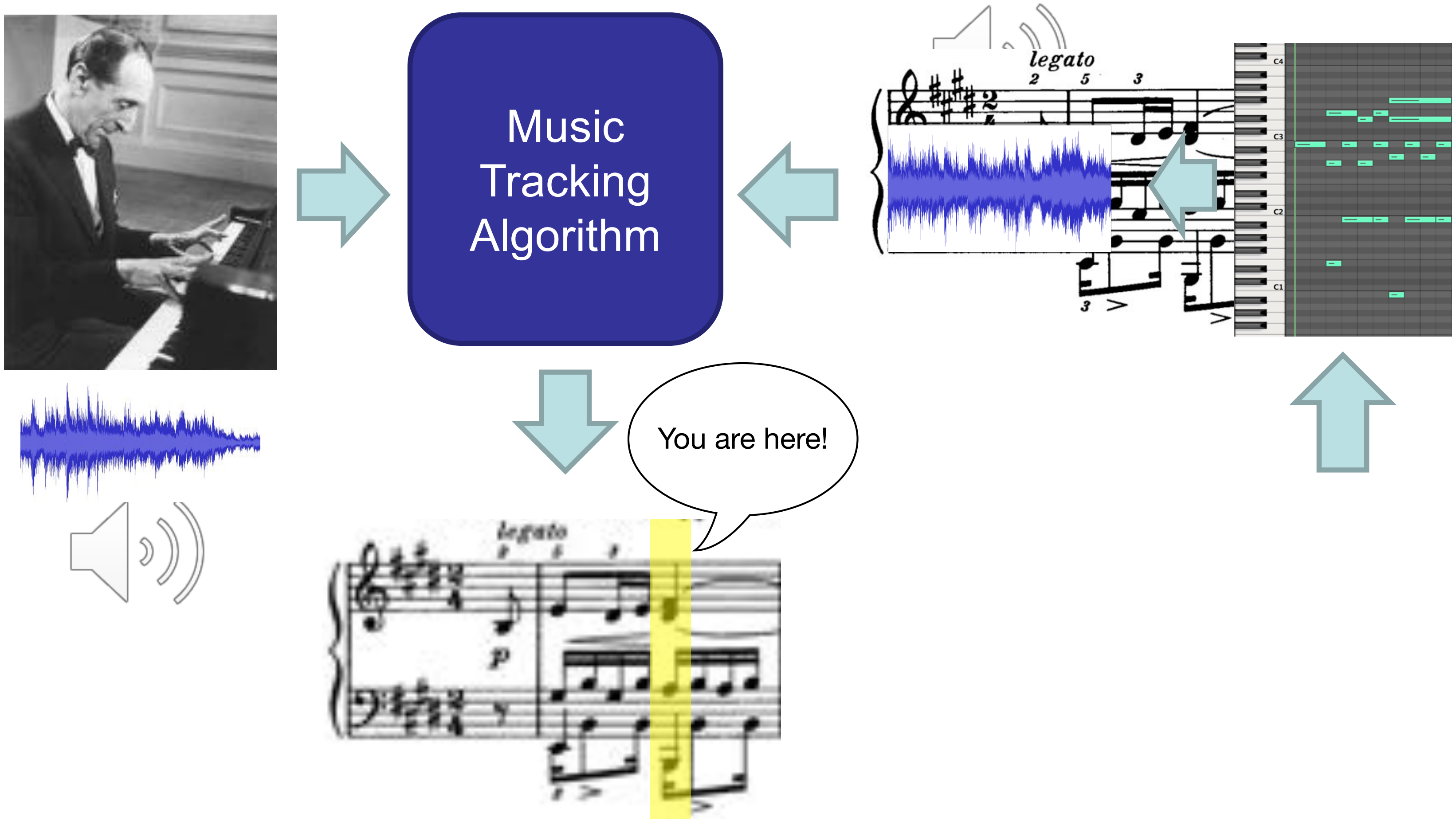
Overview (Part II)

- An Application Scenario: Flexible Music Tracking
- Automatic Music Transcription
  - Task Description
  - Recent Developments
- Fingerprinting
  - The “Shazam” Algorithm
  - Generalized Fingerprinting
- Flexible Music Tracking Re-visited

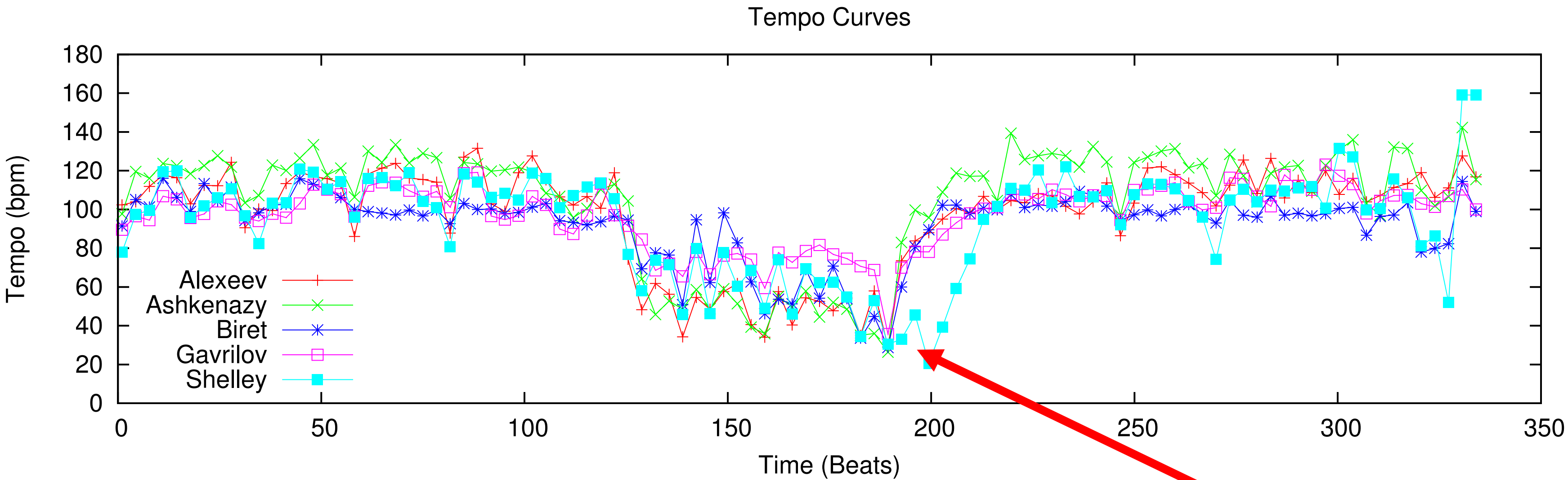
Application Scenario

MUSIC TRACKING

What is Music Tracking (Score Following)?



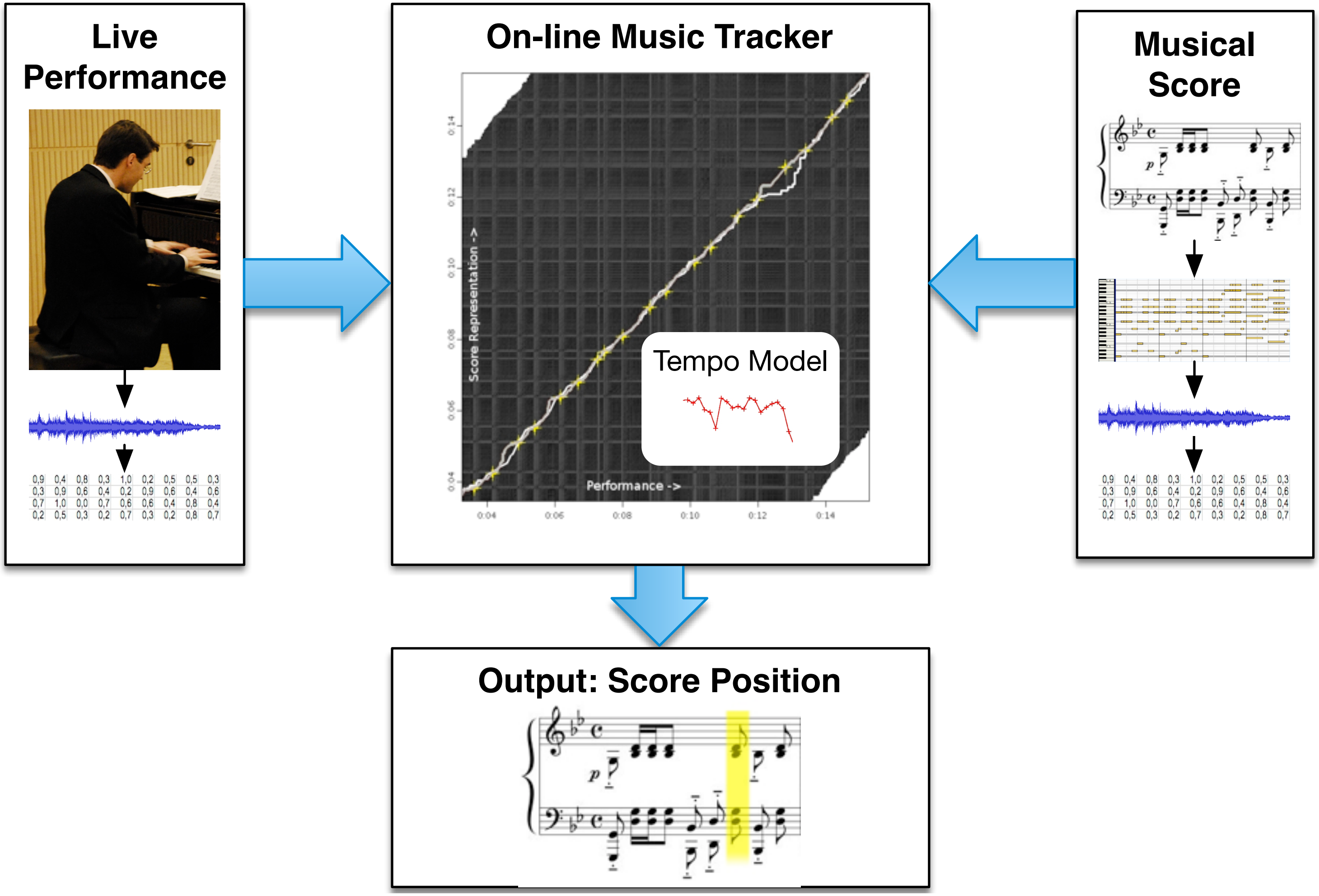
Why is Music Tracking Difficult?



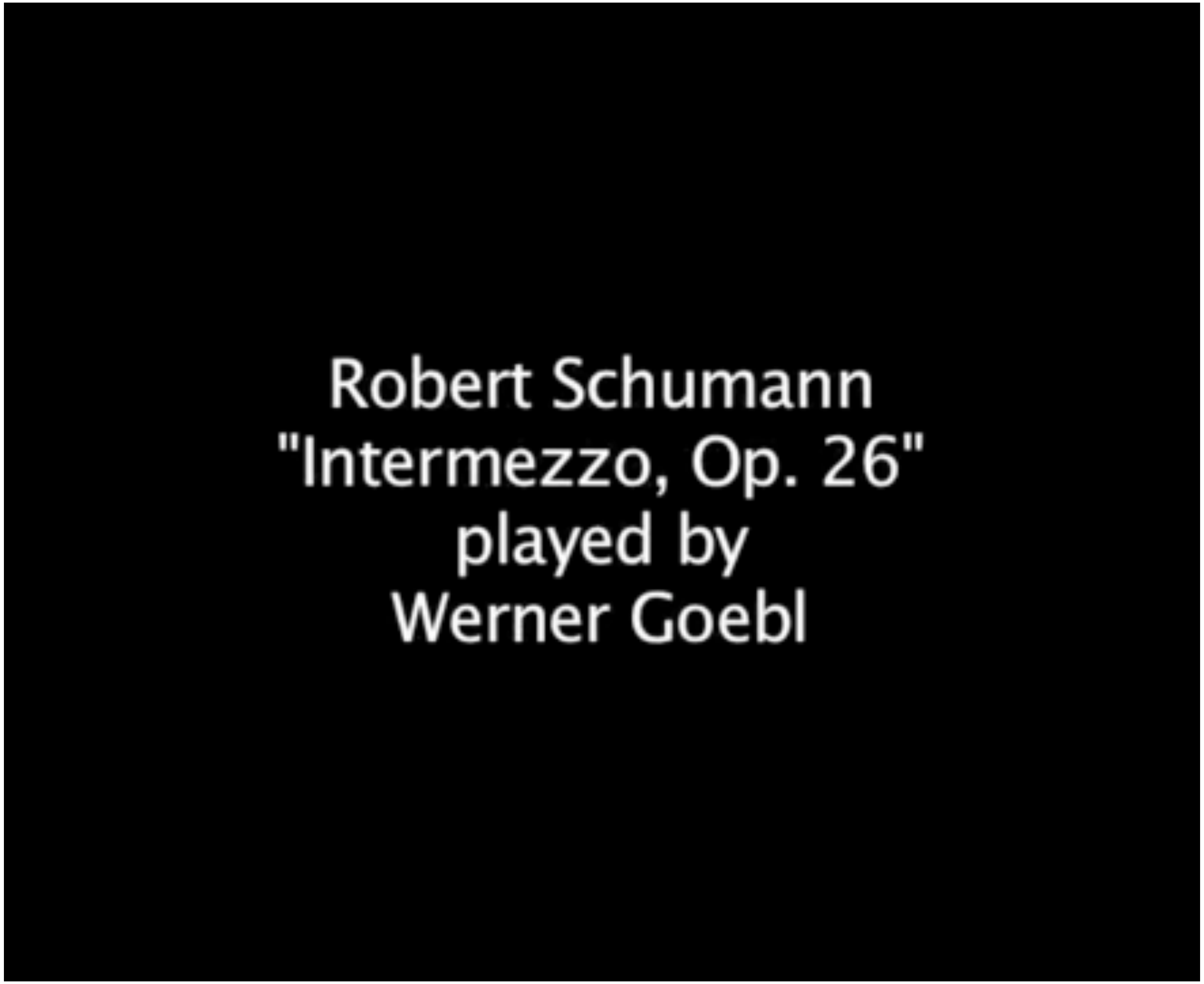
- Tempo curves extracted from 5 different performances of Rachmaninoff’s Prelude Op. 23 No. 5

[Arzt, Widmer: SMC 2010]

Music Tracking System

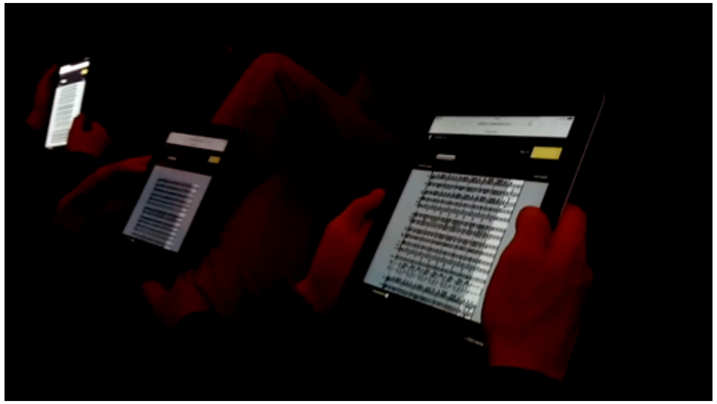


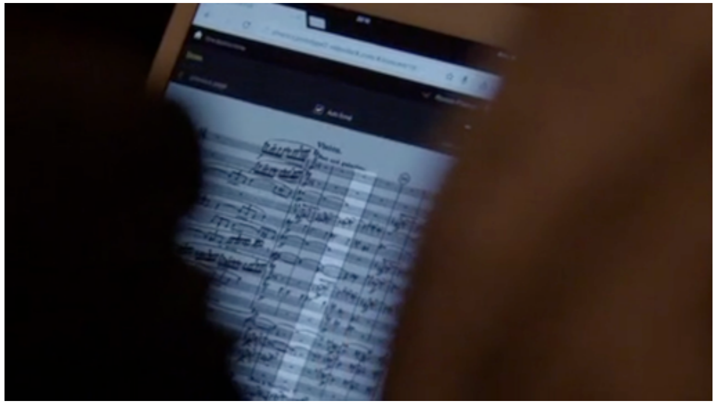
Demo: An Automatic Page Turner



[Arzt, Widmer, Dixon: ECAI 2008]

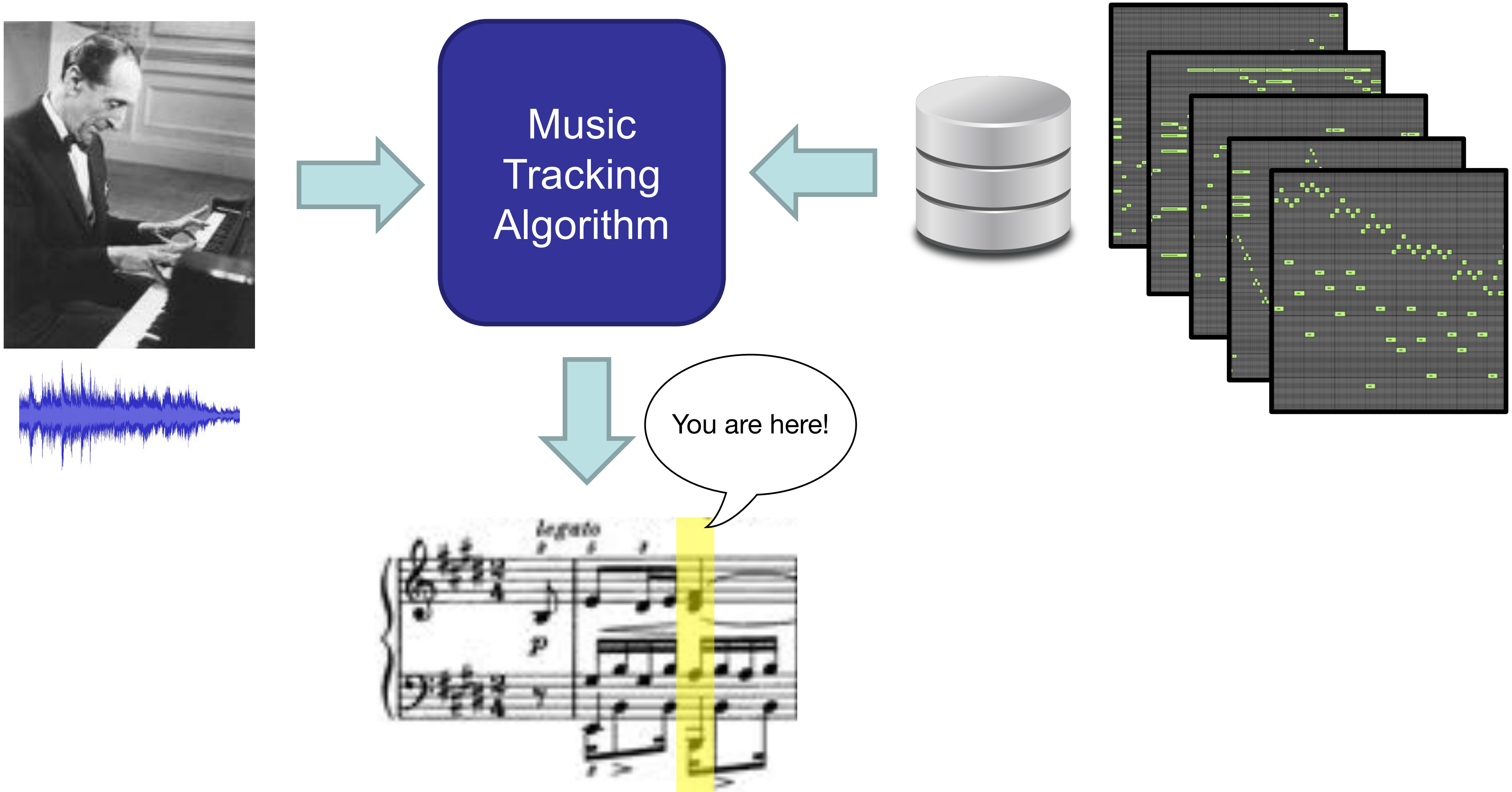
Demo: Music Tracking in the Concertgebouw





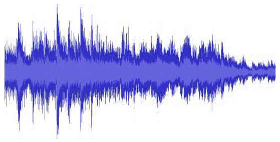
[Arzt, Frostel, Gadermaier, Gasser, Grachten, Widmer: IJCAI 2015]

Flexible Music Tracking?

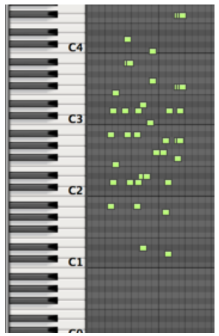


Fast Music Retrieval Based on Short Excerpts

- Matching in the Audio Domain:
  - long queries needed (15-20 seconds)
  - computationally costly

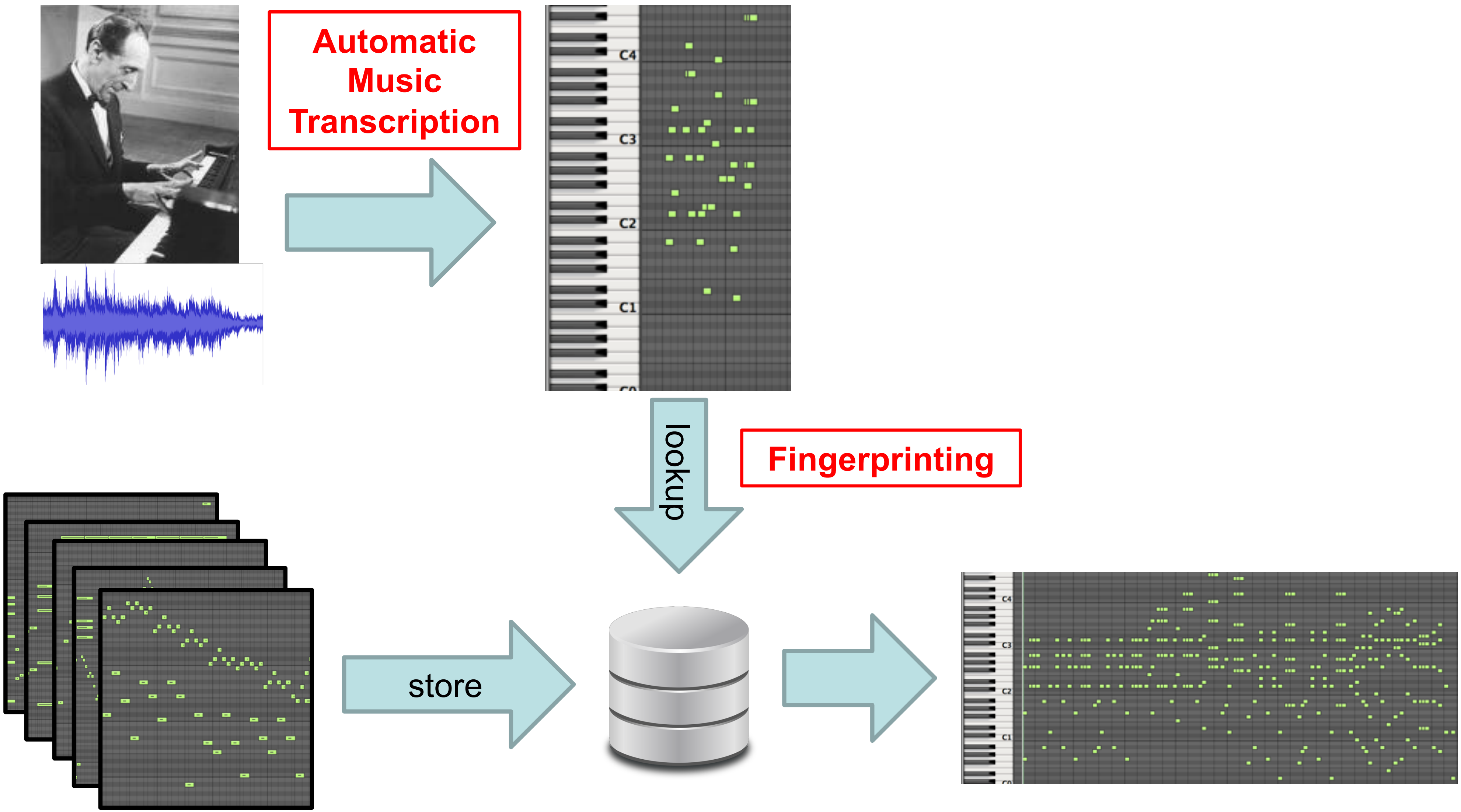


- Matching in the Symbolic Domain:
  - more compact, reduced to the essential information
  - fast algorithms



- How to transfer data to the symbolic domain?
- How to perform fast lookup?

Retrieval via Automatic Music Transcription and Fingerprinting



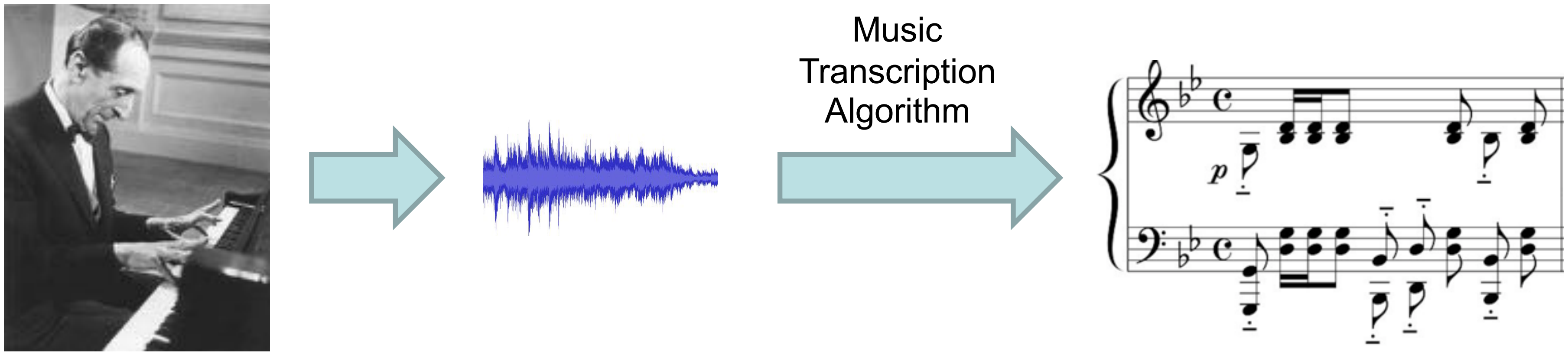
Output: Rachmaninoff Prelude Op.23 No. 5

AUTOMATIC MUSIC TRANSCRIPTION

Automatic Music Transcription

Task

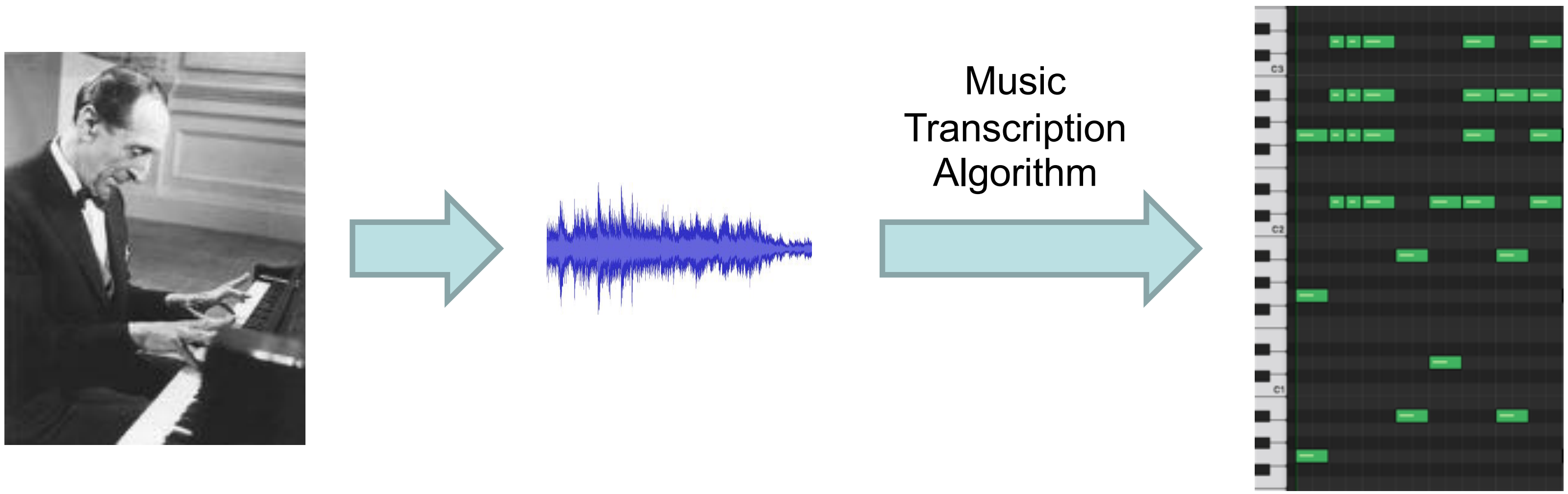
- **Given:** Audio Recording of a Piece of Music
- **Goal:** Create Sheet Music (or some symbolic representation) of the recording



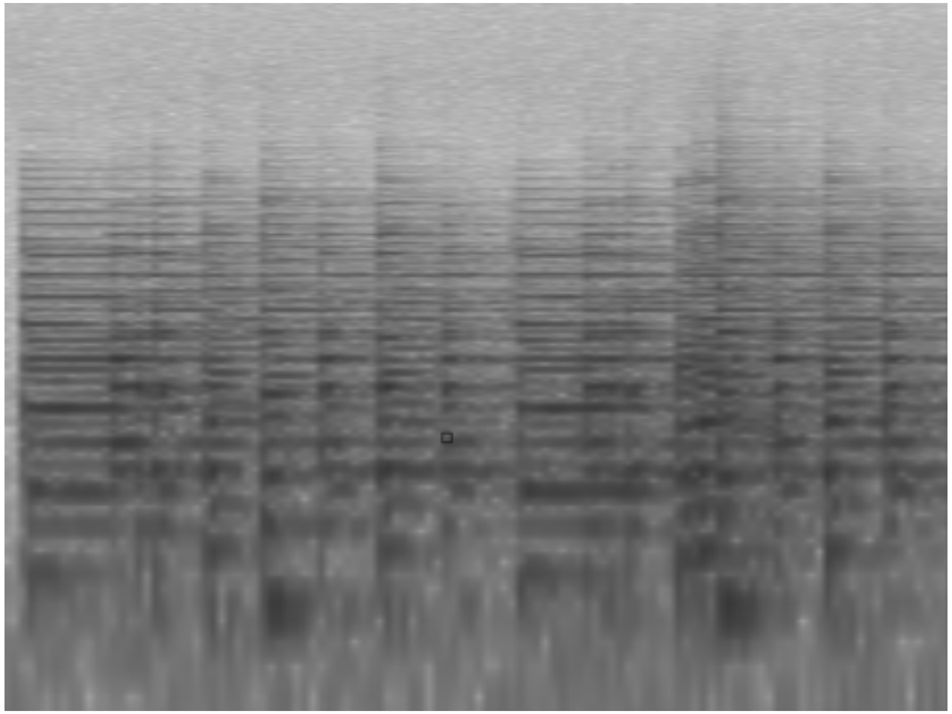
Automatic Music Transcription

Task

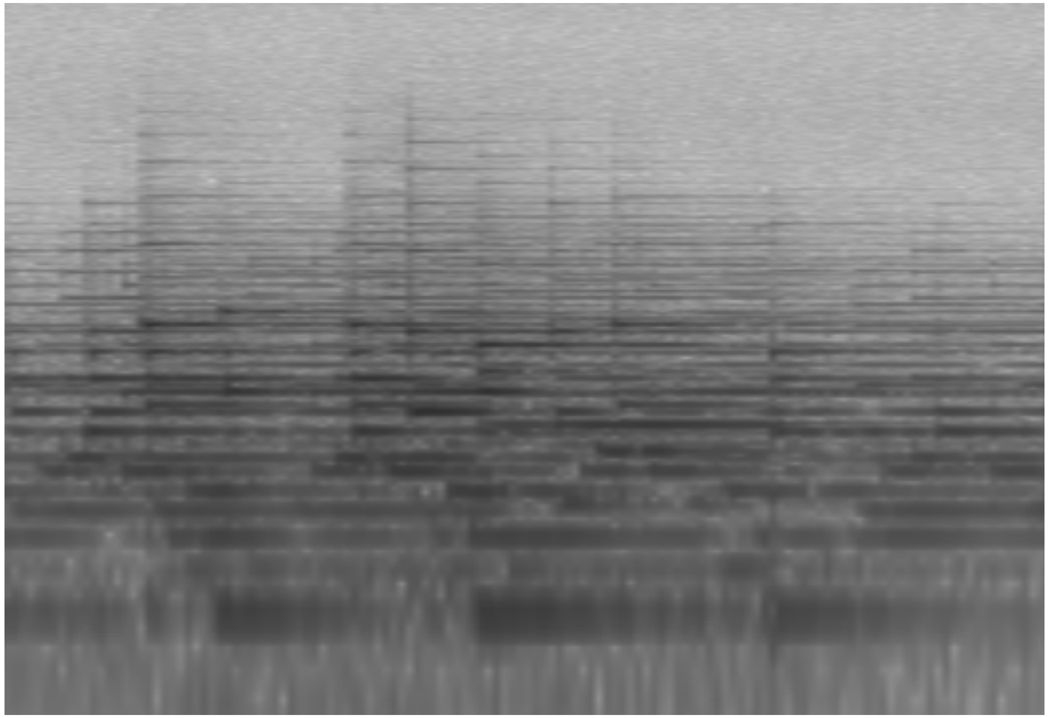
- **Given:** Audio Recording of a Piece of Music
- **Goal:** Create Sheet Music (or some symbolic representation) of the recording



Automatic Music Transcription



Automatic Music Transcription



Automatic Music Transcription

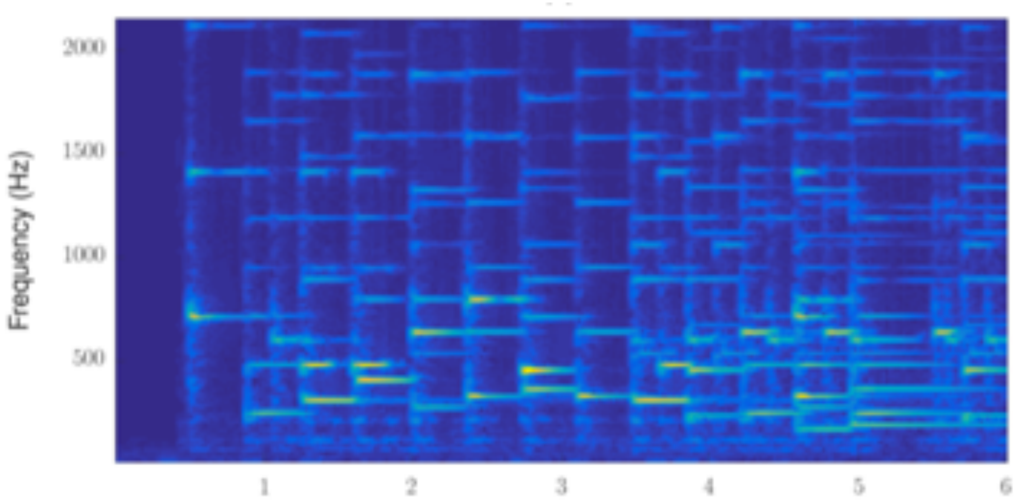
Automatic Music Transcription: Key Challenges

- **Polyphonic music** is a **mixture of multiple simultaneous sources** with different pitch, loudness and timbre. **Inferring musical attributes** (e.g., pitch) from the mixture signal is an **under-determined problem**.
- The **harmonics** of overlapping sound events often **overlap in frequency**, making the separation of the voices even more difficult.
- **Timing** of musical voices is **governed by the regular metrical structure** of the music. This **violates** the **assumption of statistical independence** between sources.
- **Annotation** of ground-truth transcriptions for polyphonic music is very **time consuming** and requires **high expertise**.

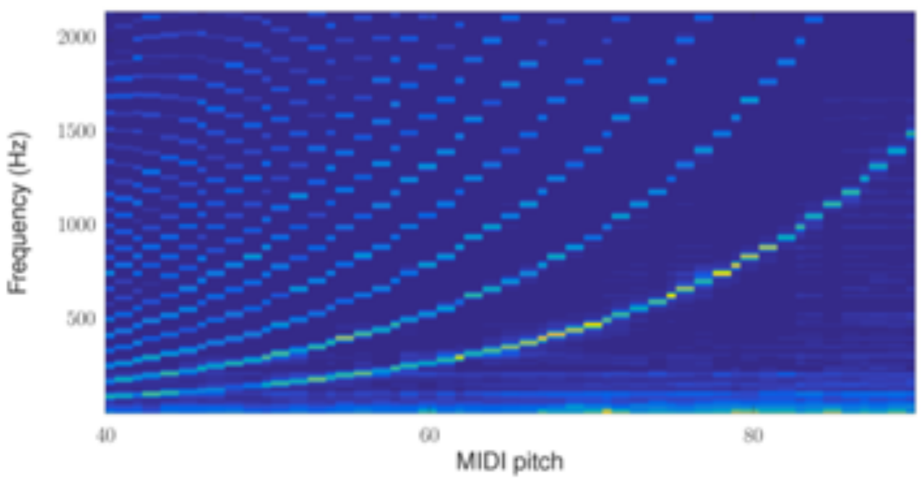
after [Benetos, Dixon, Duan, Ewert: IEEE SPM, 2019]

Automatic Music Transcription: Non-negative Matrix Factorization

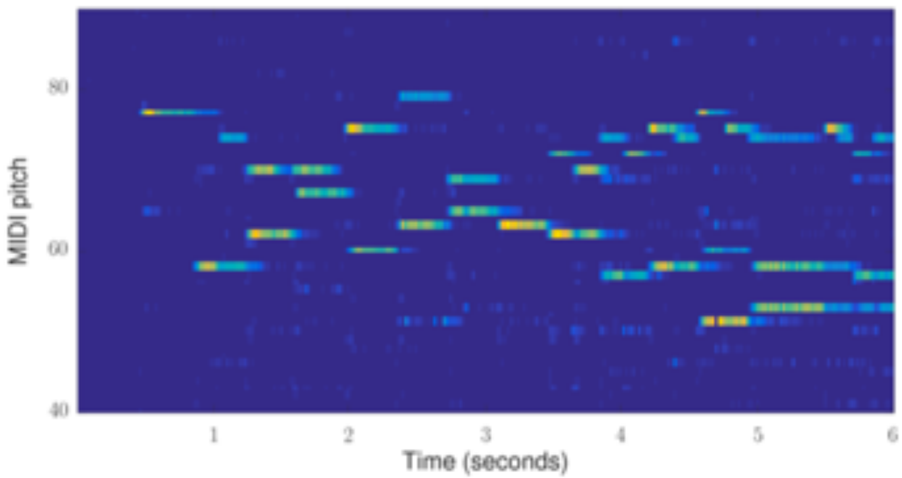
adapted from
[Benetos, Dixon, Duan, Ewert: IEEE SPM, 2019]



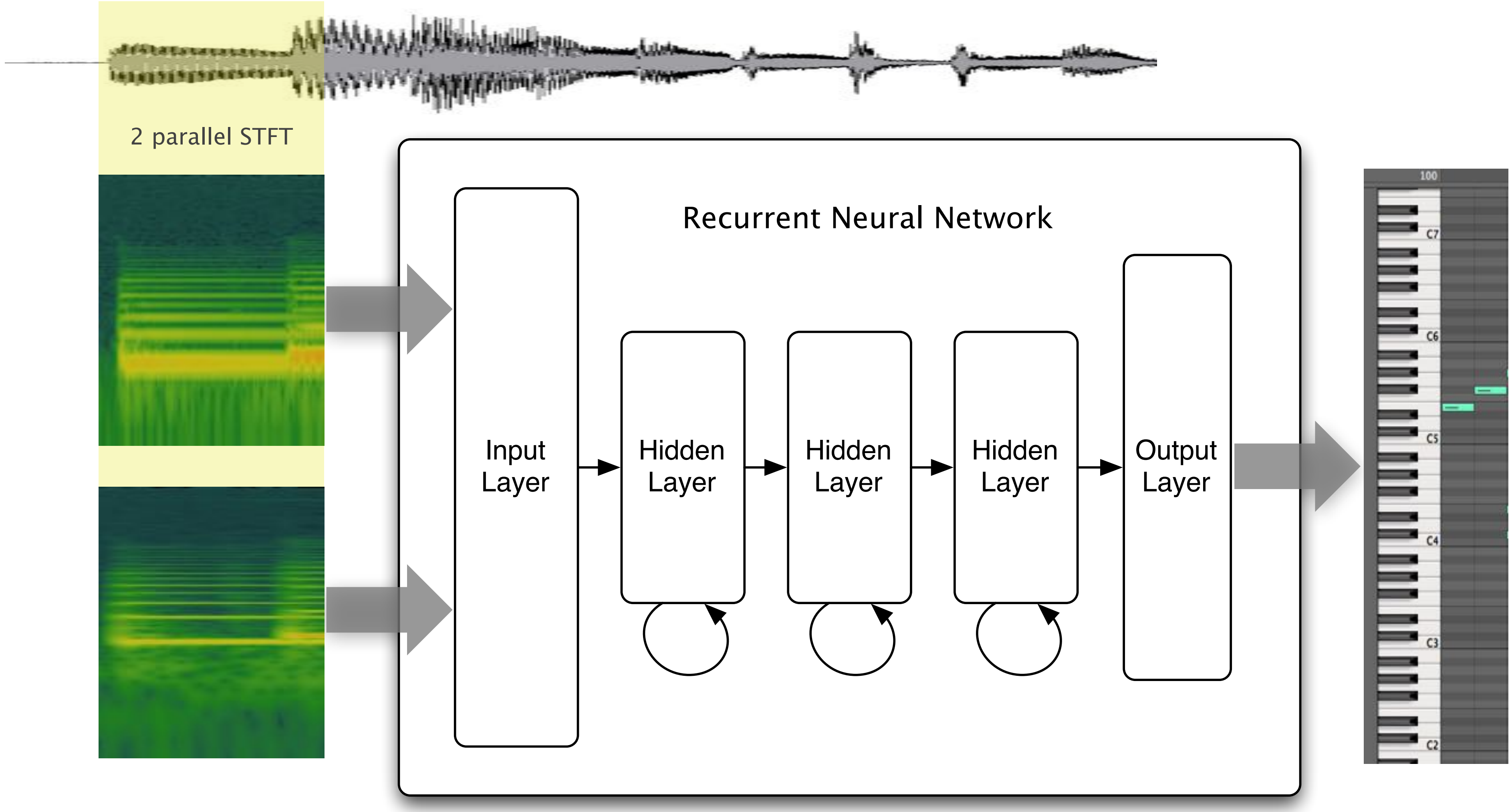
=



•



Automatic Music Transcription: Neural Networks

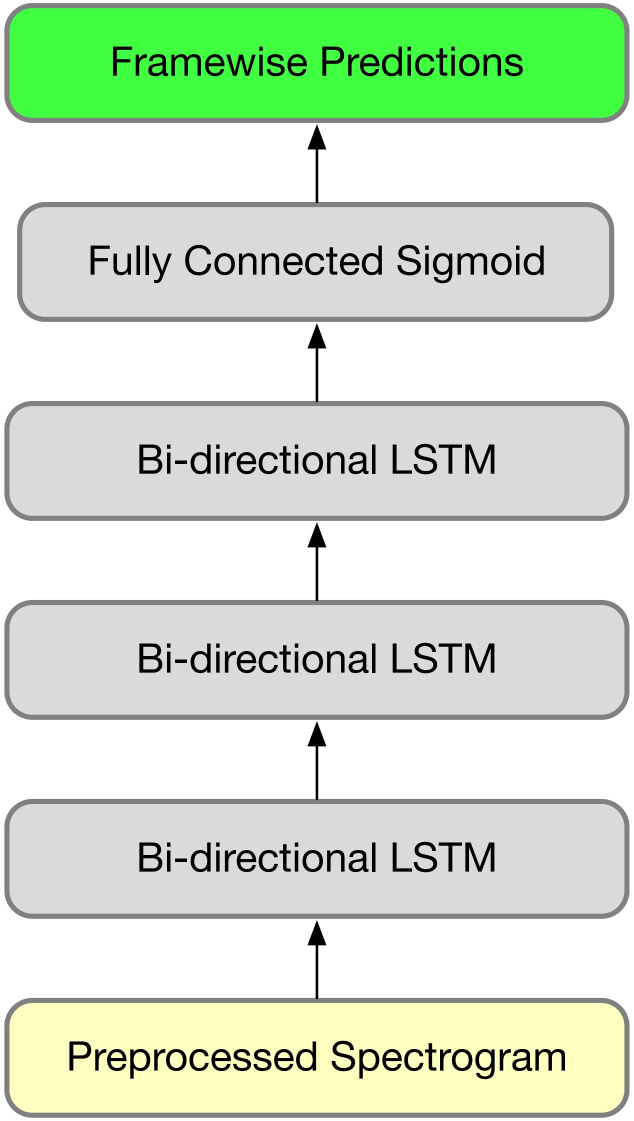


[Böck, Schedl: ICASSP 2012]

Automatic Music Transcription: Neural Networks

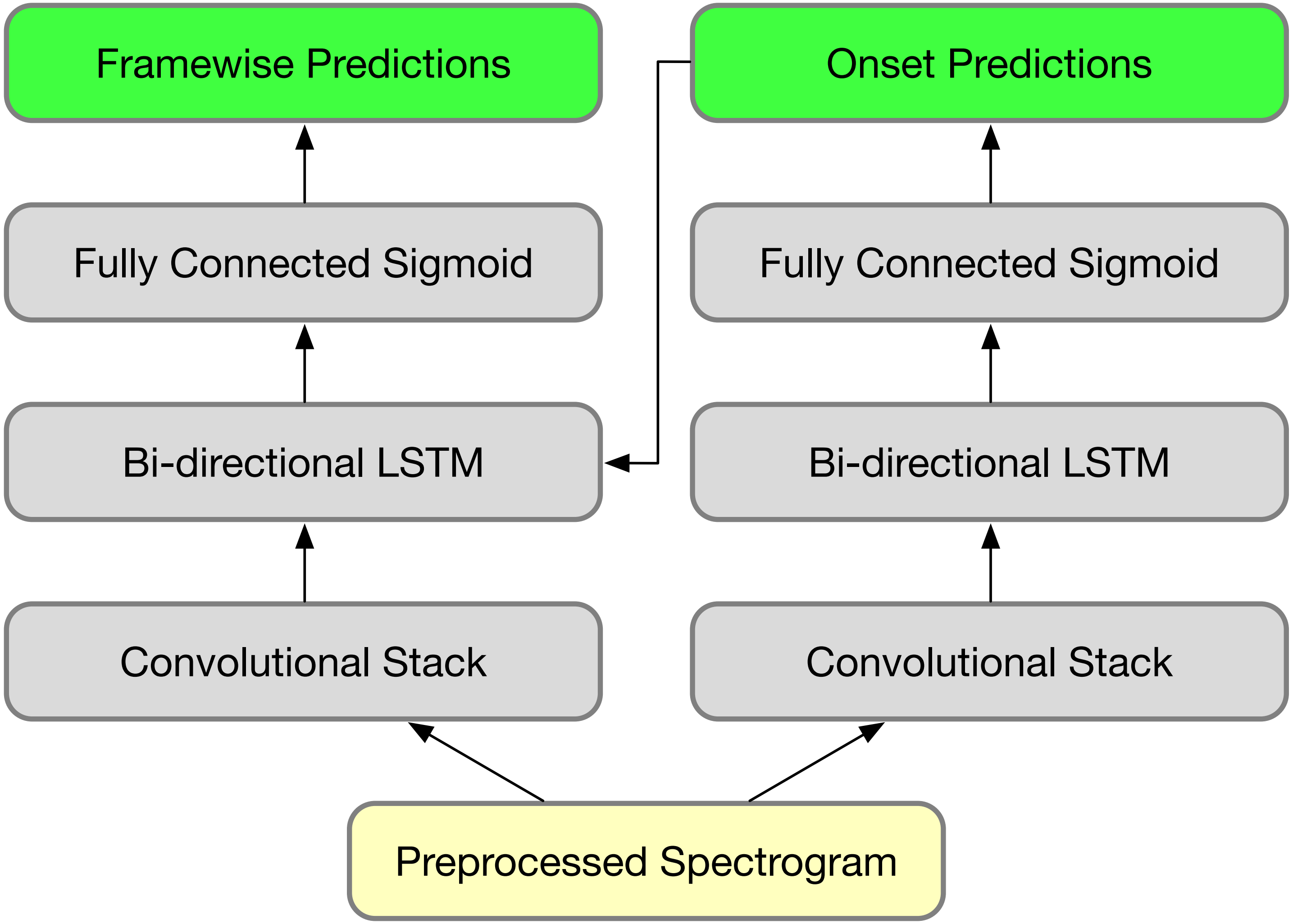
- Needed: Annotated Training Data
- (Large enough) Datasets for General Music Transcription are virtually non-existent
- Exception: Piano Music Transcription
  - MAPS Dataset [Emiya, Bertin, David, Badeau: TR 2012]
  - MAESTRO Dataset [Hawthorne, Stasyuk, Roberts, Simon, Huang, Dieleman, Elsen, Engel, Eck: CoRR 2018]

Automatic Music Transcription: Neural Network Architectures



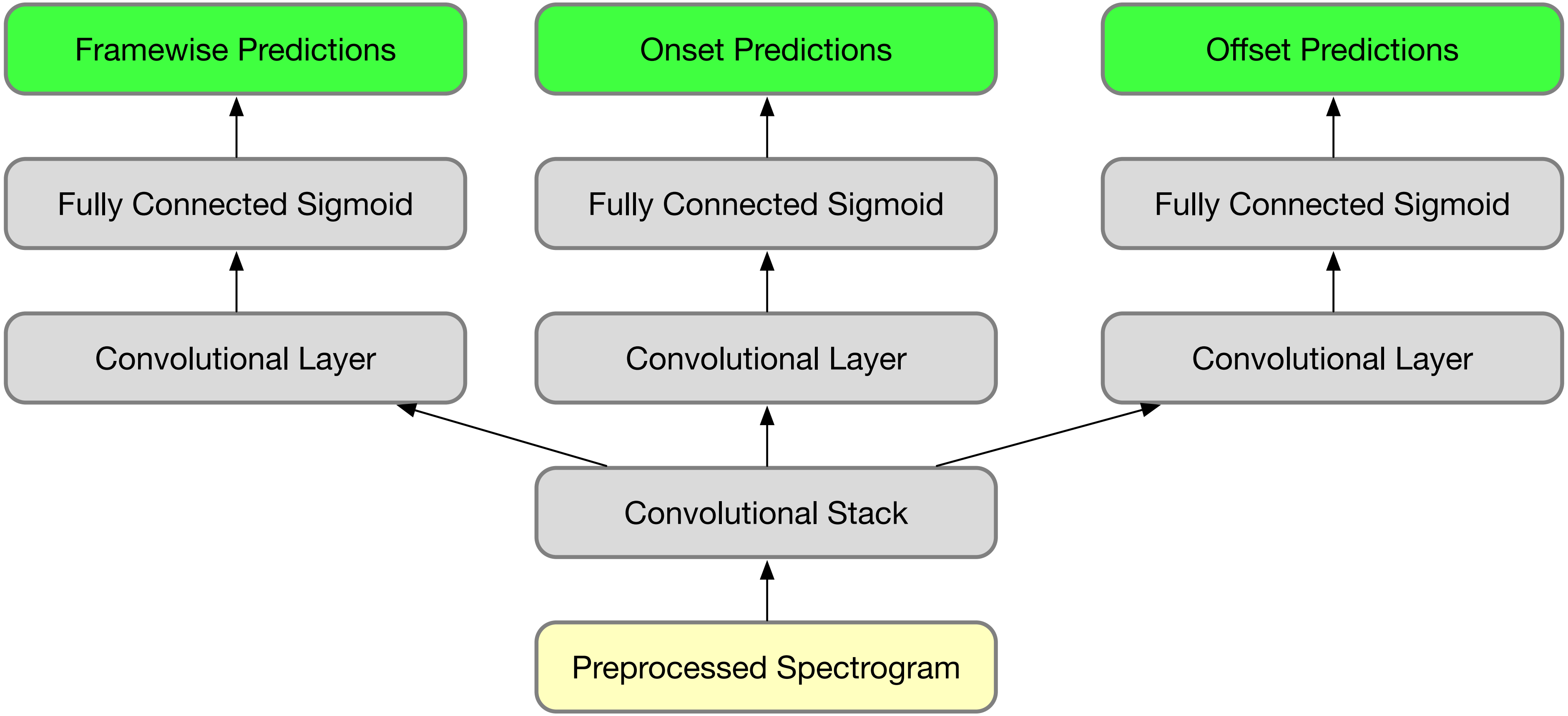
[Böck, Schedl: ICASSP 2012]

Automatic Music Transcription: Neural Network Architectures



[Hawthorne, Elsen, Song, Roberts, Simon, Raffel, Engel, Oore, Eck: ISMIR 2018]

Automatic Music Transcription: Neural Network Architectures



[Kelz, Böck: ICASSP 2019]

Automatic Music Transcription: Examples

Original Audio

Re-synthesized Transcription

Examples produced using the algorithm presented in [Hawthorne, Elsen, Song, Roberts, Simon, Raffel, Engel, Oore, Eck: ISMIR 2018] (https://magenta.tensorflow.org/onsets-frames)

Automatic Music Transcription: Examples

Yefim Bronfman playing the Cadenza from Rachmaninov's Piano Concerto No. 3
[https://www.youtube.com/watch?v=yh4_63ugeho]

FINGERPRINTING

Audio Fingerprinting

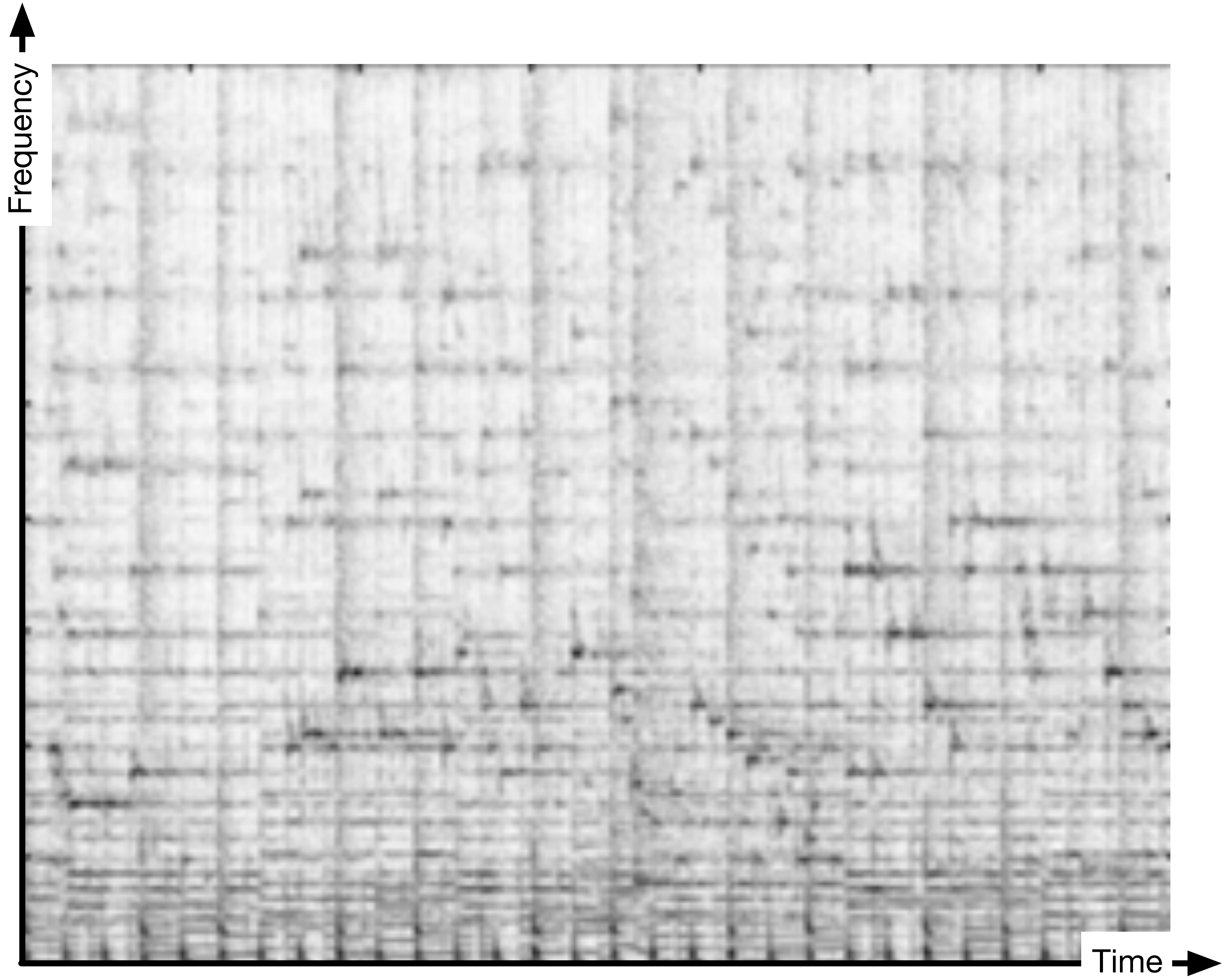
Task

- **Given:** Short Excerpt of Audio Recording of a Piece of Music
- **Goal:** Find Corresponding Instance in Database of Pieces (Audio Recordings) of Music

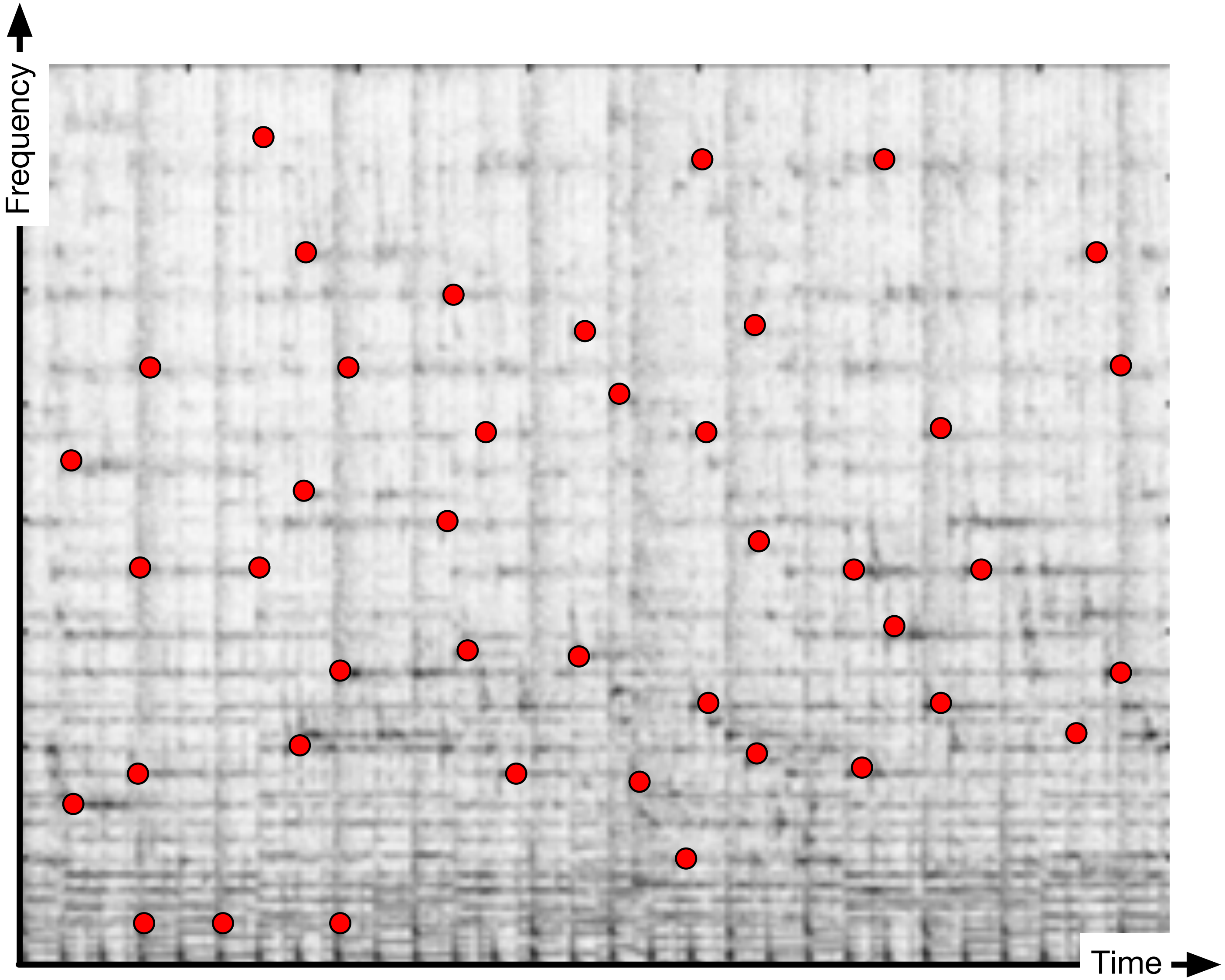
Idea

- Describe Sequences via so-called “Fingerprints”
  - local, translation-invariant, robust, compact and discriminative features
- Common Approach: Use a “Constellation Map” as basis for the Fingerprinting Algorithm → “Landmark-based Fingerprinting”

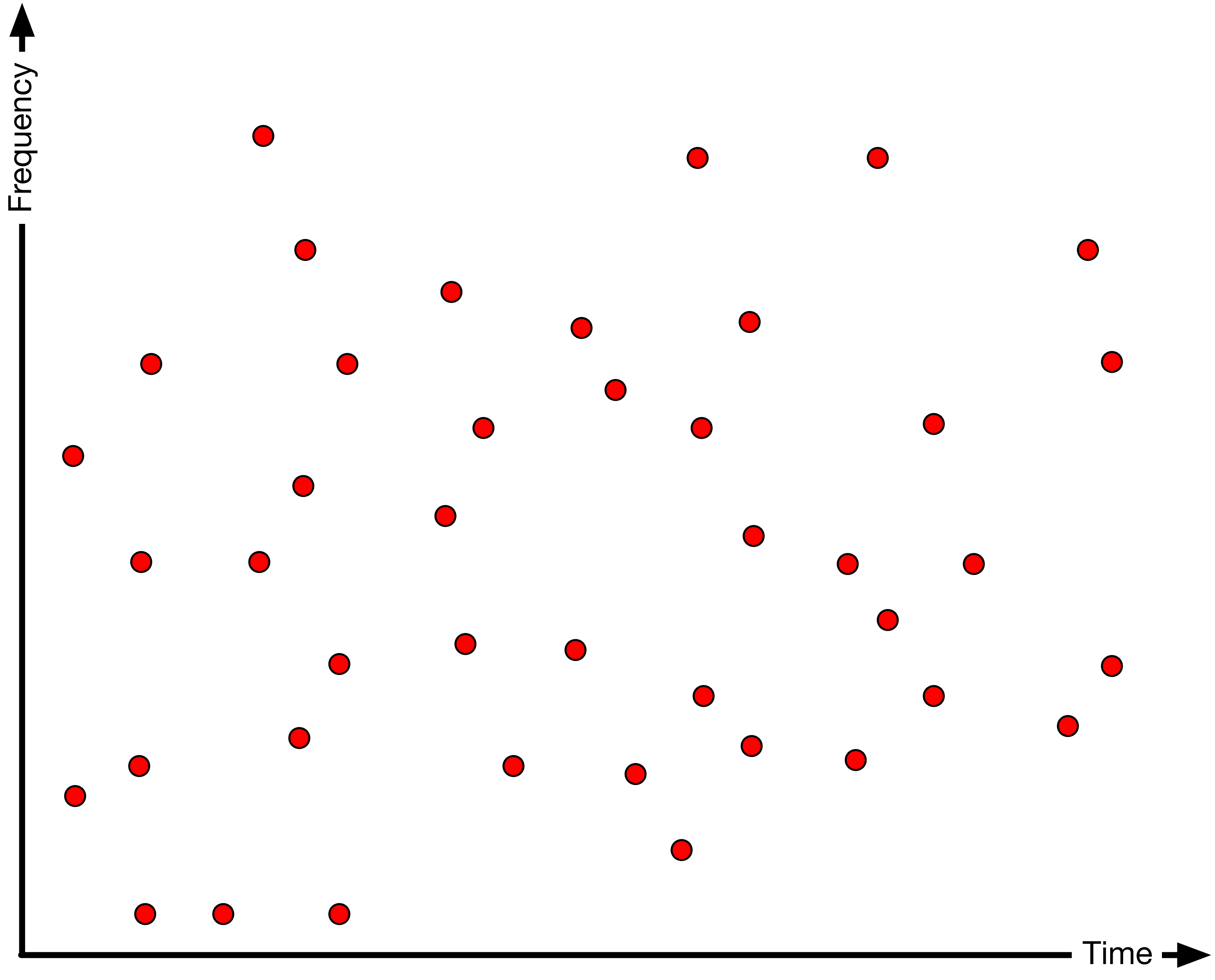
Constellation Map from Peaks in the Audio



Constellation Map from Peaks in the Audio



Constellation Map from Peaks in the Audio

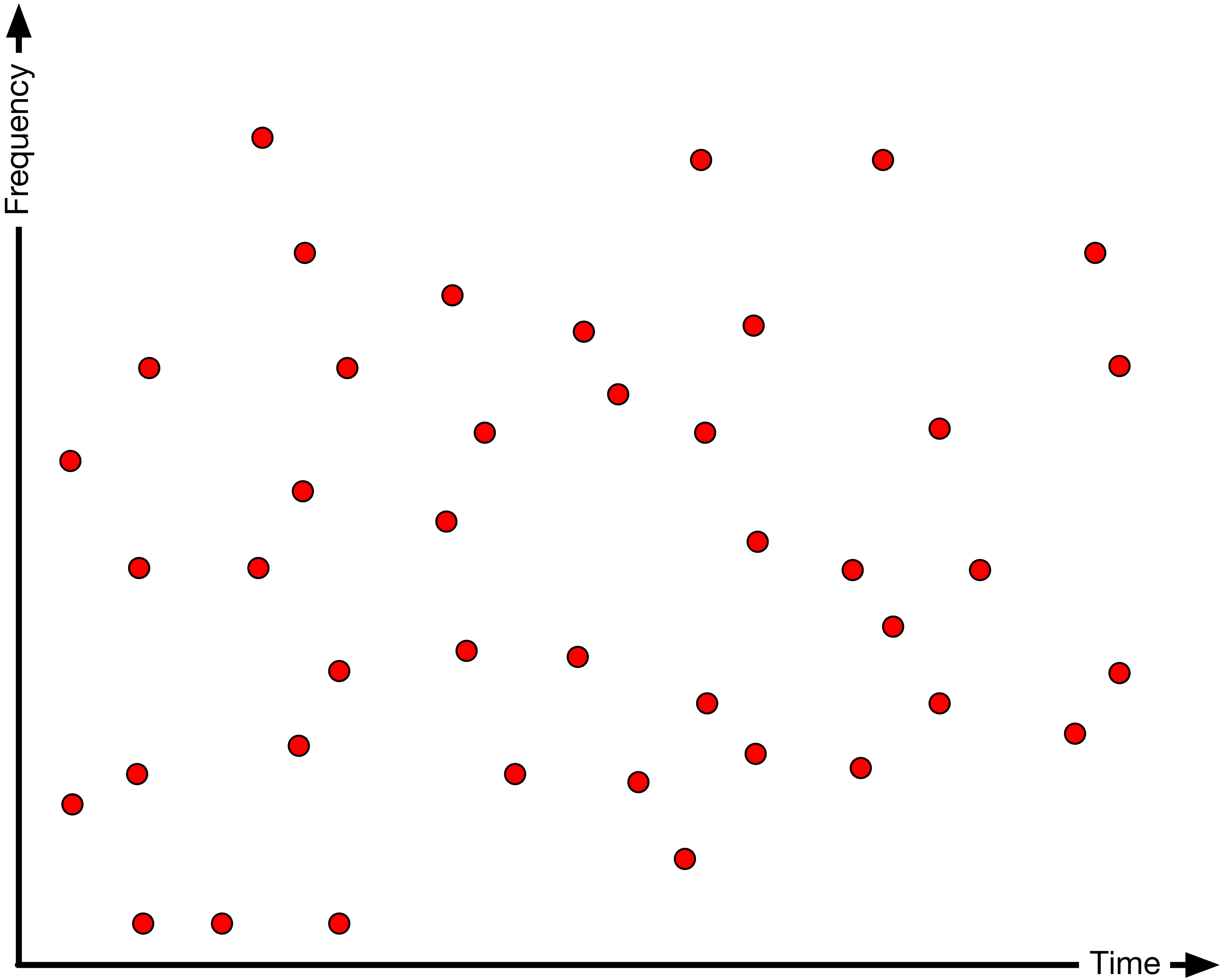


The “Shazam” Algorithm: Basic Idea

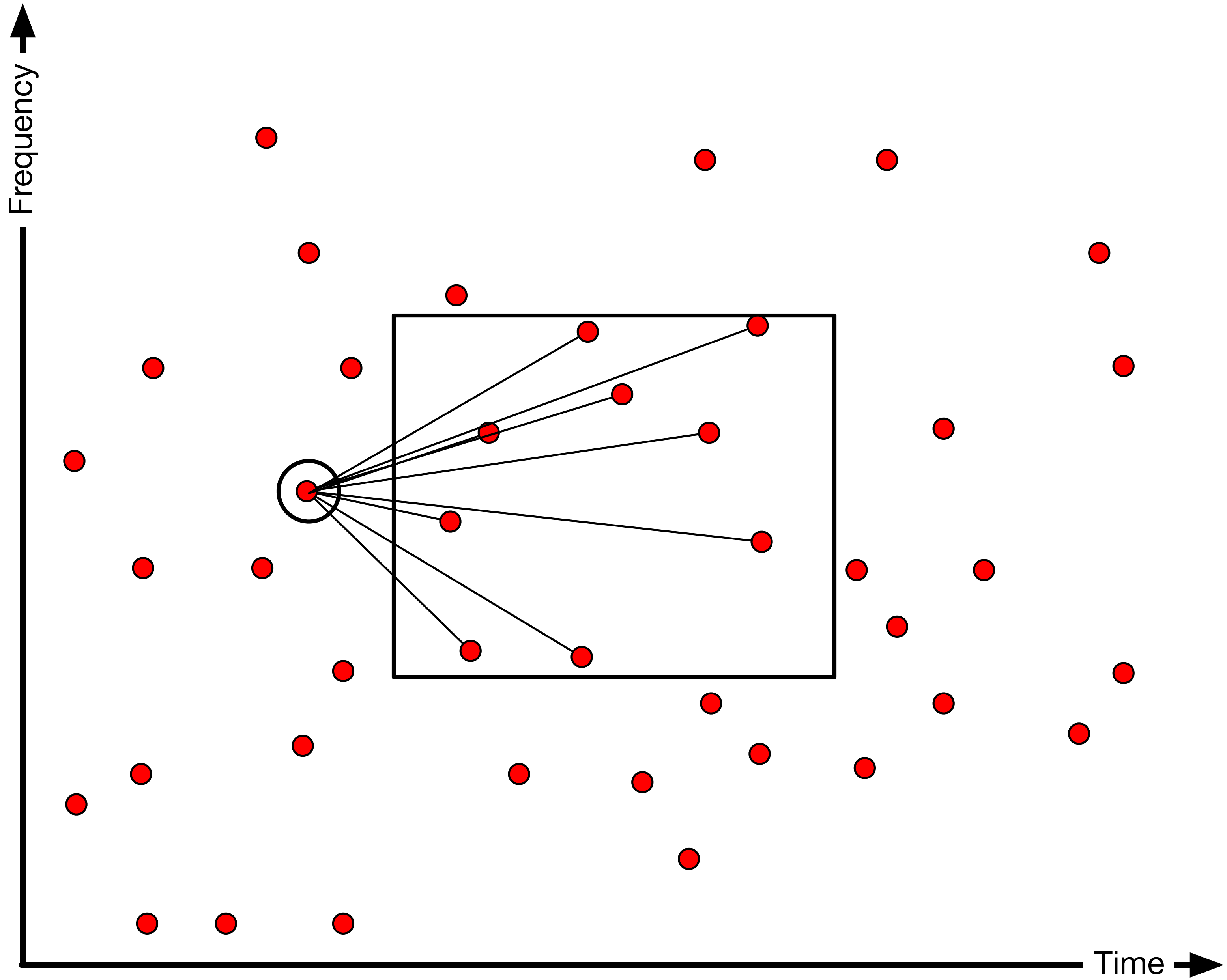
- For all Items in the Database
  - compute constellation map (Shazam: spectral peaks)
  - create local pairs from points in the constellation map
  - describe the pairs in a compact fashion (via hashes)
  - store them in a fast database (hash table)
- For the Query
  - compute constellation map (Shazam: spectral peaks)
  - create local pairs from points in the constellation map
  - describe the pairs in the same compact fashion (hashes)
  - query the database for matching pairs
  - find consecutive sequences of matching pairs
  - return item which contains the best matching sequence of pairs

[Wang: ISMIR 2003]

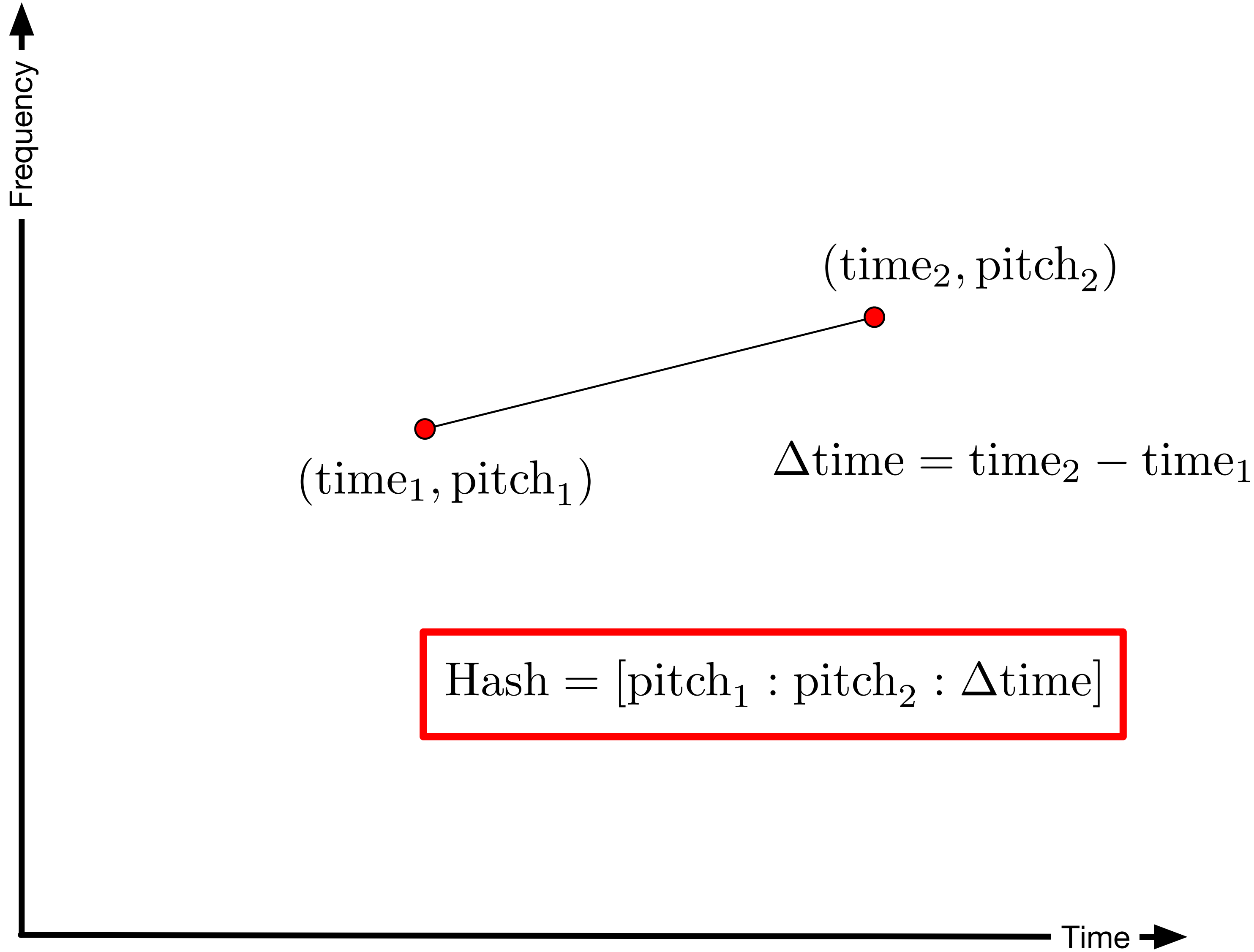
The “Shazam” Algorithm



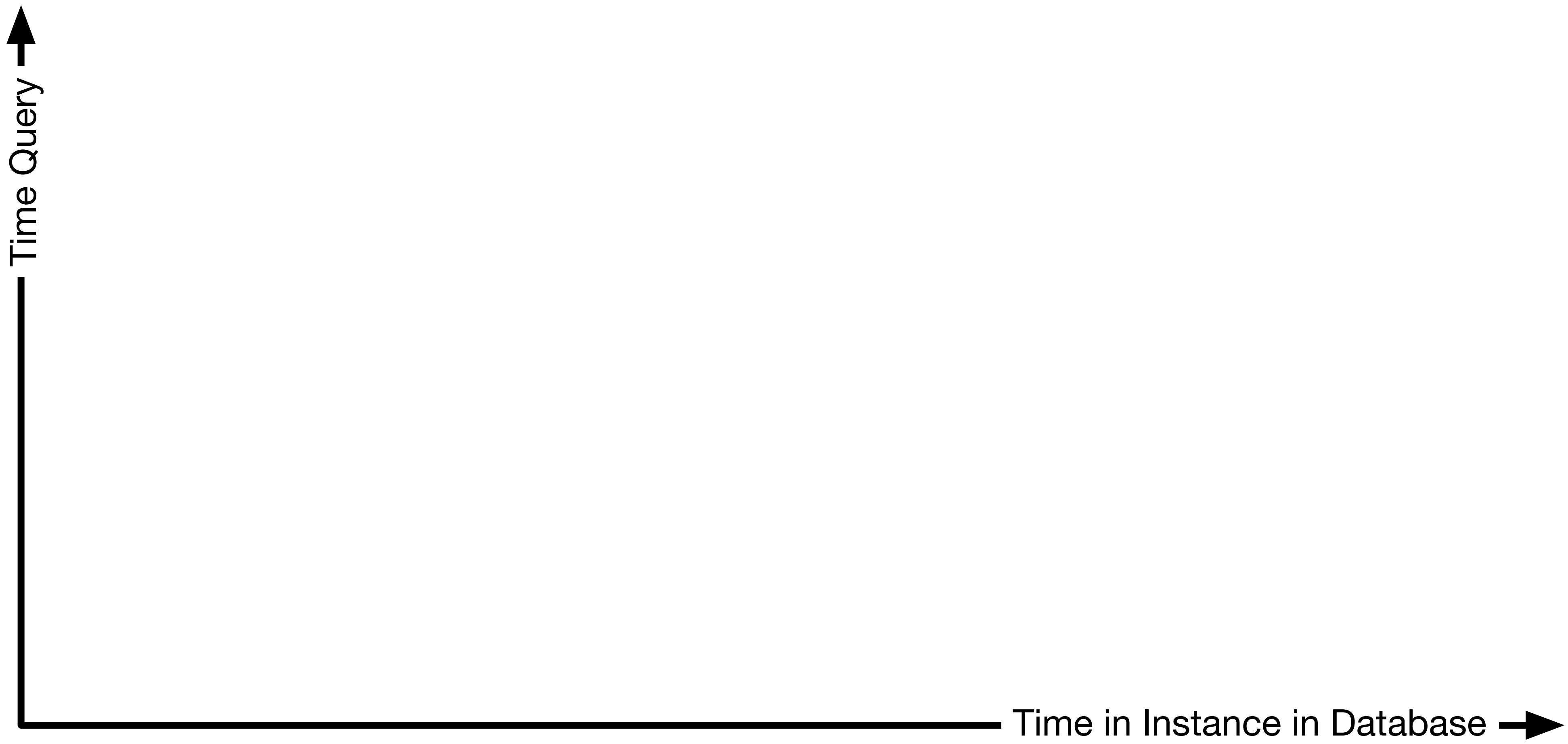
The “Shazam” Algorithm



The “Shazam” Algorithm

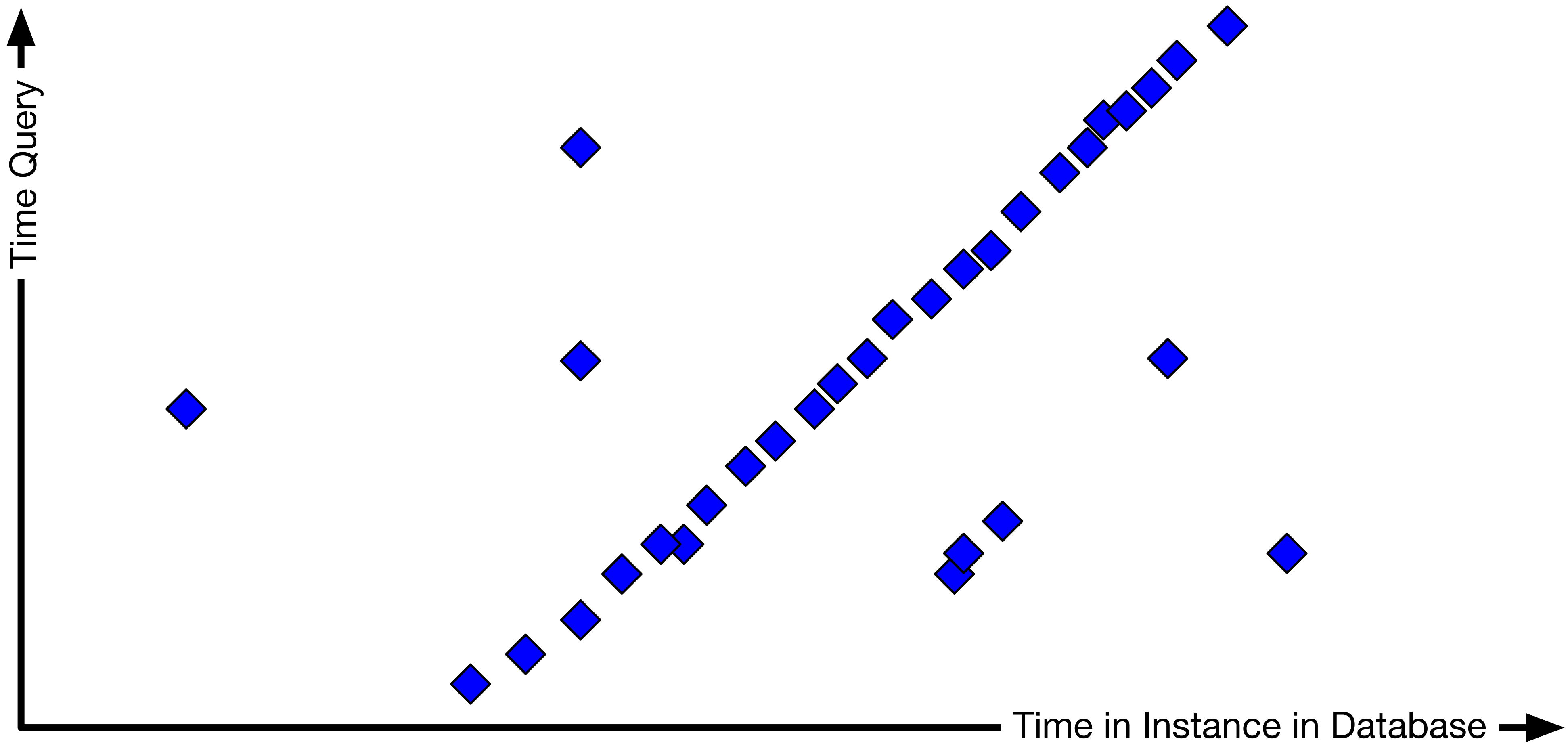


The “Shazam” Algorithm: Lookup



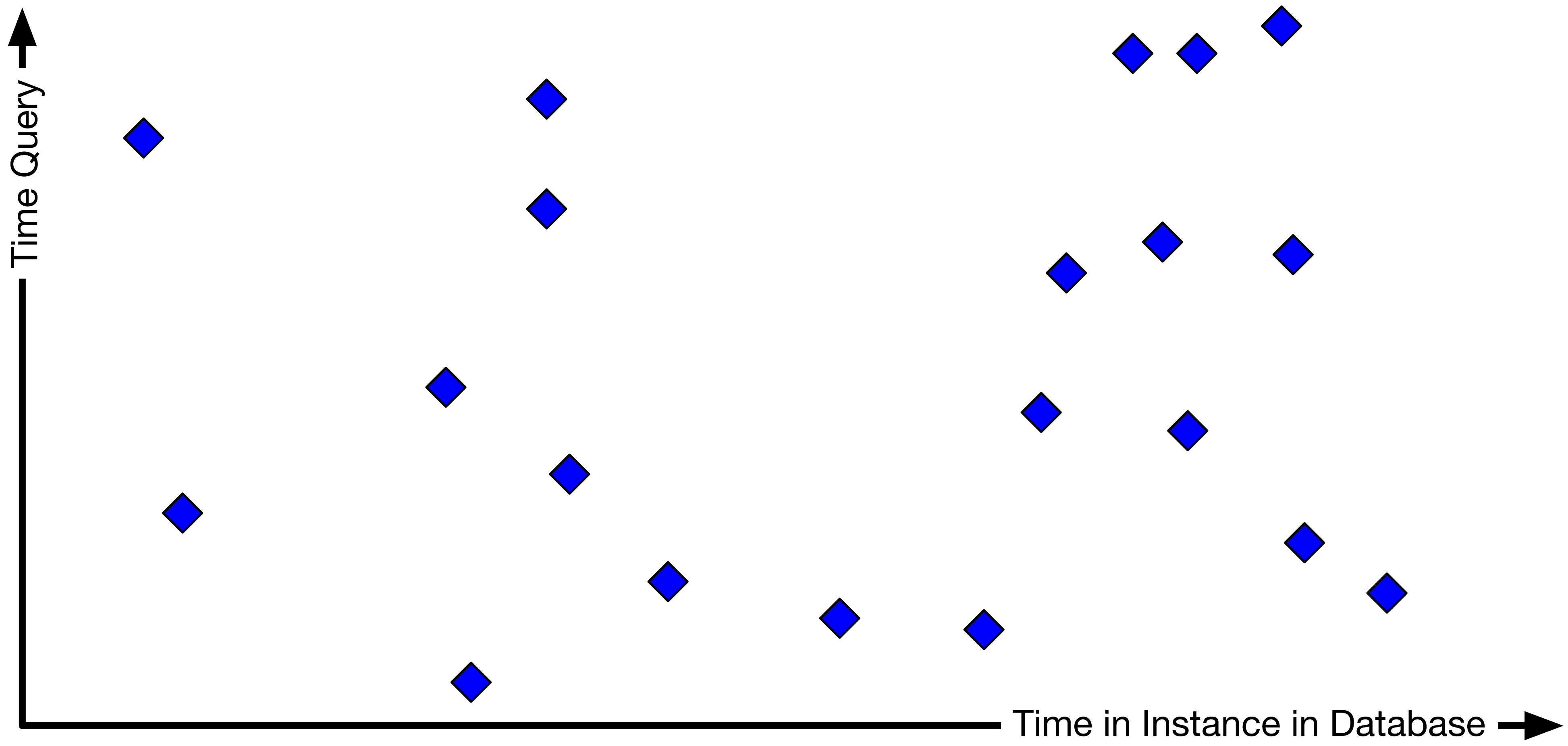
Scatterplot of Matching Hash Locations of Query with Instance in Database

The “Shazam” Algorithm: Lookup



Scatterplot of Matching Hash Locations of Query with Instance in Database
Match (Diagonal)

The “Shazam” Algorithm: Lookup



Scatterplot of Matching Hash Locations of Query with Instance in Database

No Match (No Diagonal)

The "Shazam" Algorithm

- Industry-strength Algorithm for Music Identification from Audio, scales well to Millions of Audio Files
- Invariant to
  - noise
  - most distortions
- Not Invariant to
  - tempo variations
  - transpositions
  - different instrumentations
  - …

→ The Shazam Algorithm can only detect **exact duplicates** (regarding the musical content)

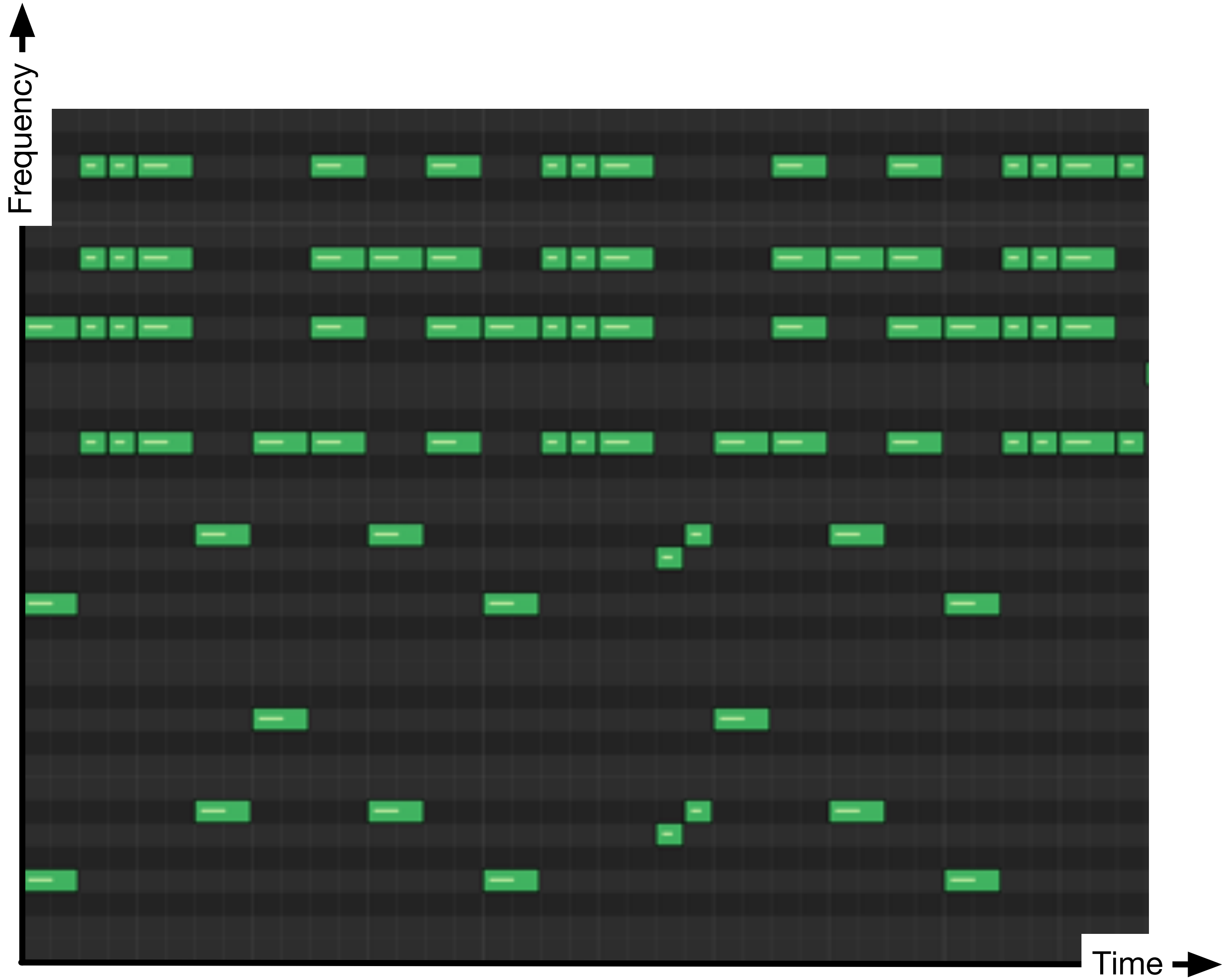
[Wang: ISMIR 2003]

Generalized Fingerprinting

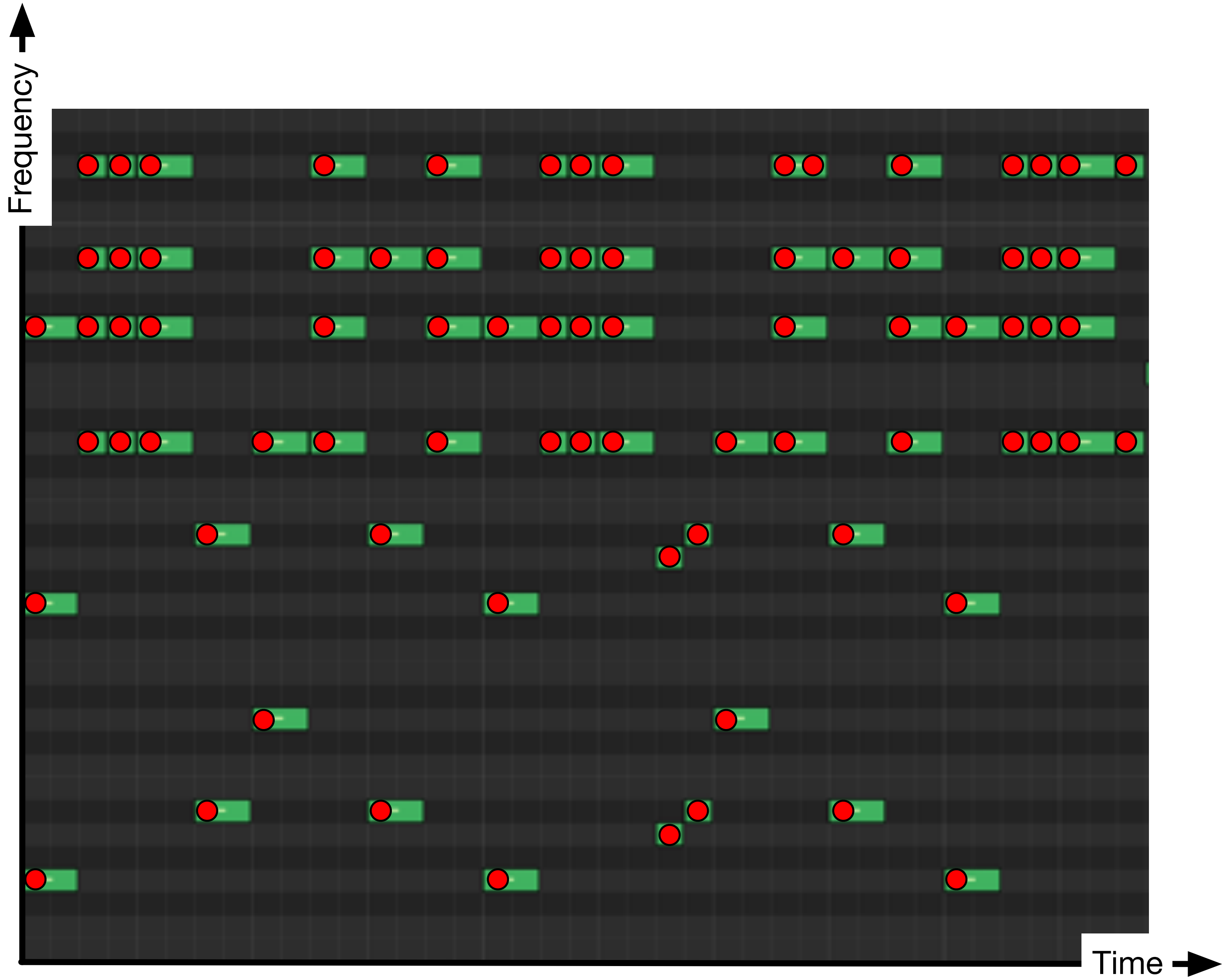
- Apply Fingerprinting to
  - audio representations and
  - symbolic representations

- Add Invariances
  - to transpositions
  - to tempo
  - to instrumentation (given a good-enough transcription algorithm exists)

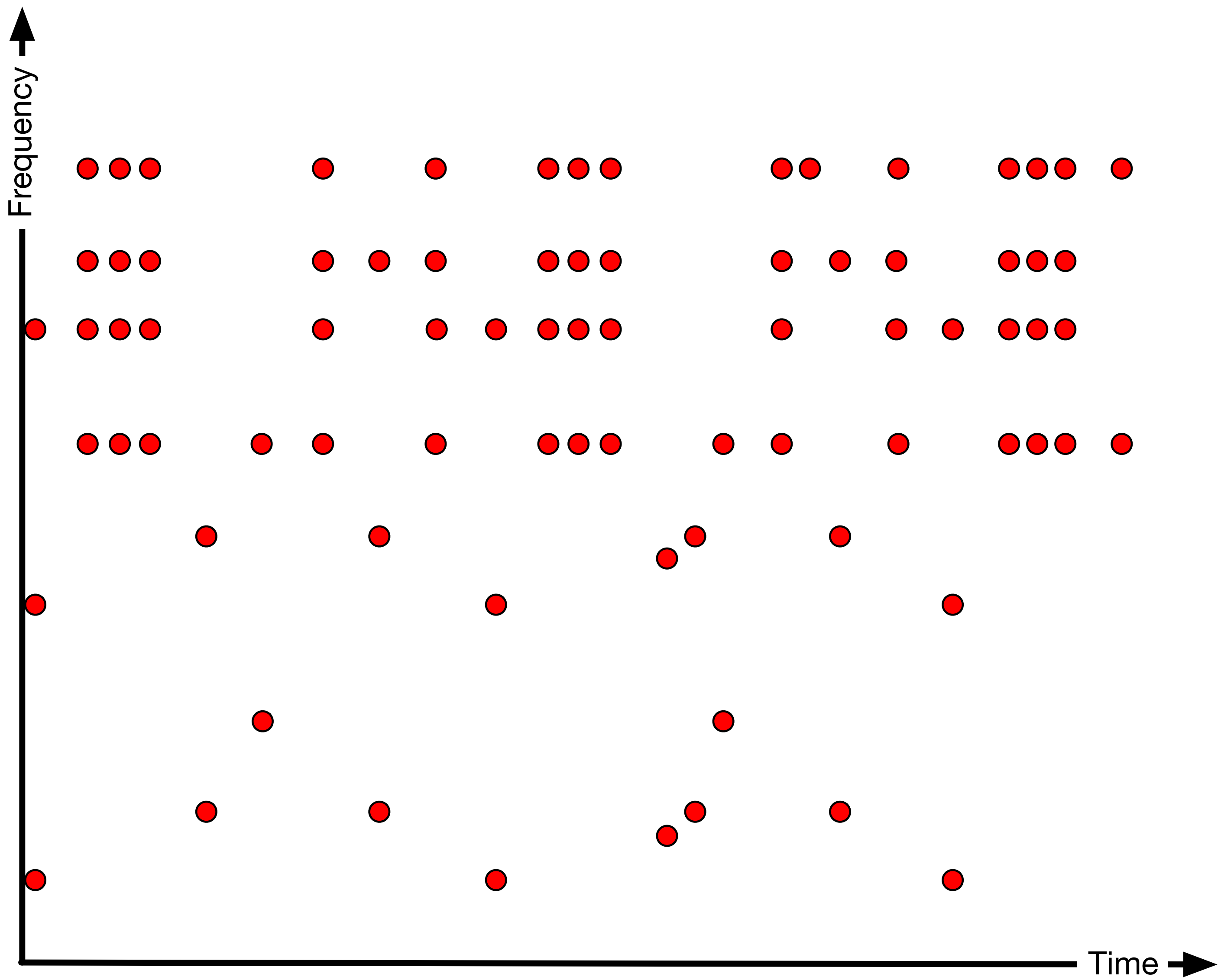
Constellation Map from Symbolic Representation



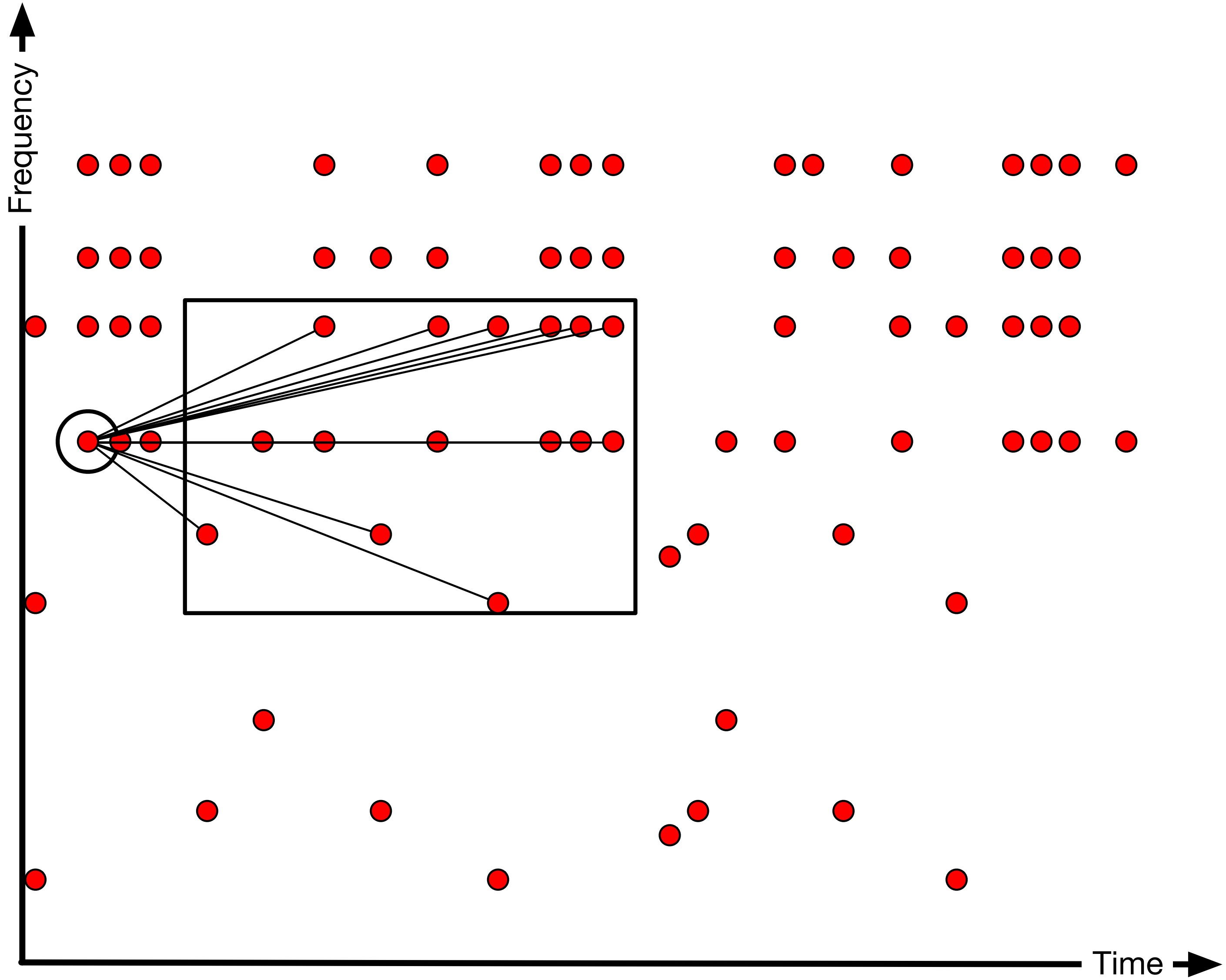
Constellation Map from Symbolic Representation



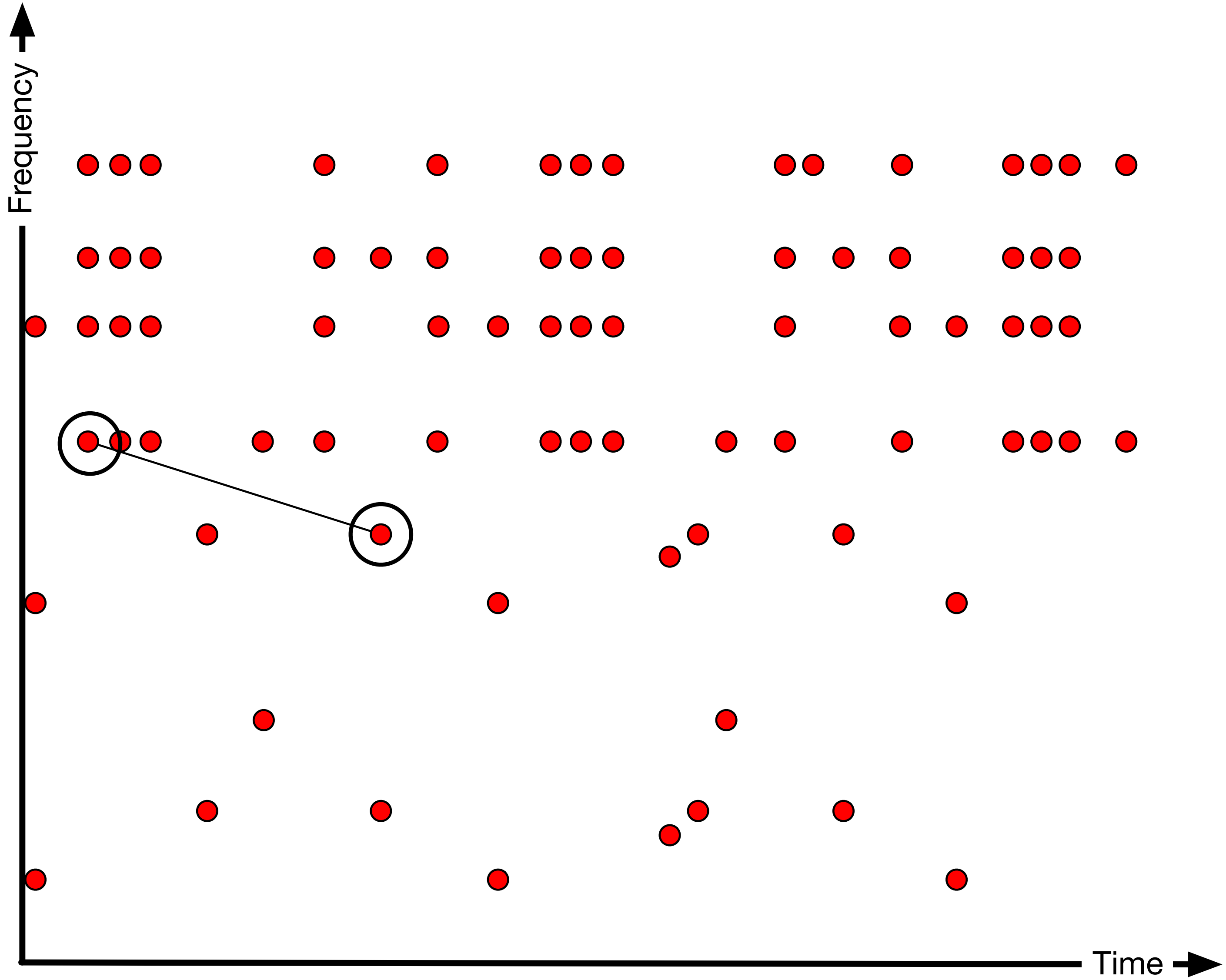
Constellation Map from Symbolic Representation



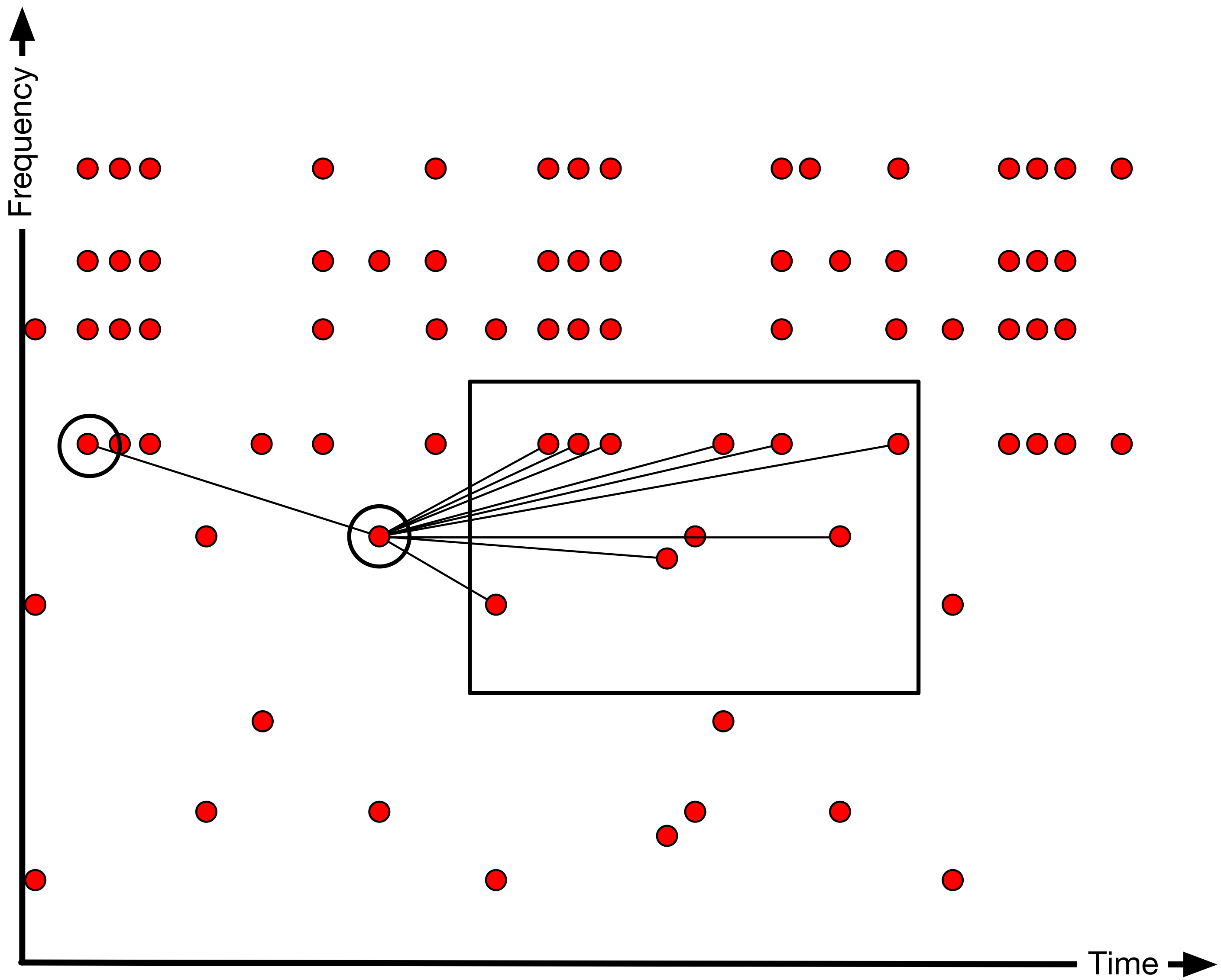
Generalized Fingerprinting



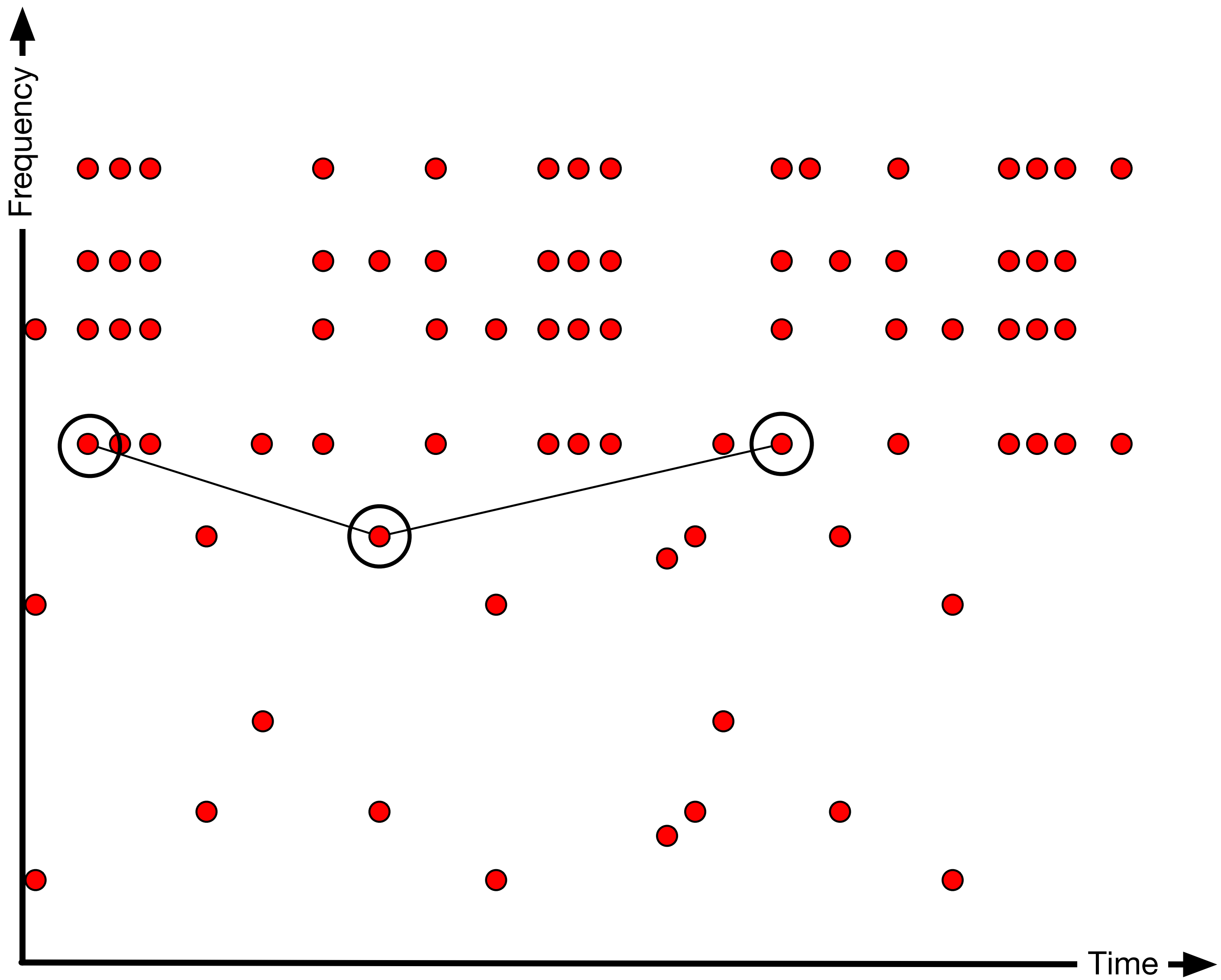
Generalized Fingerprinting



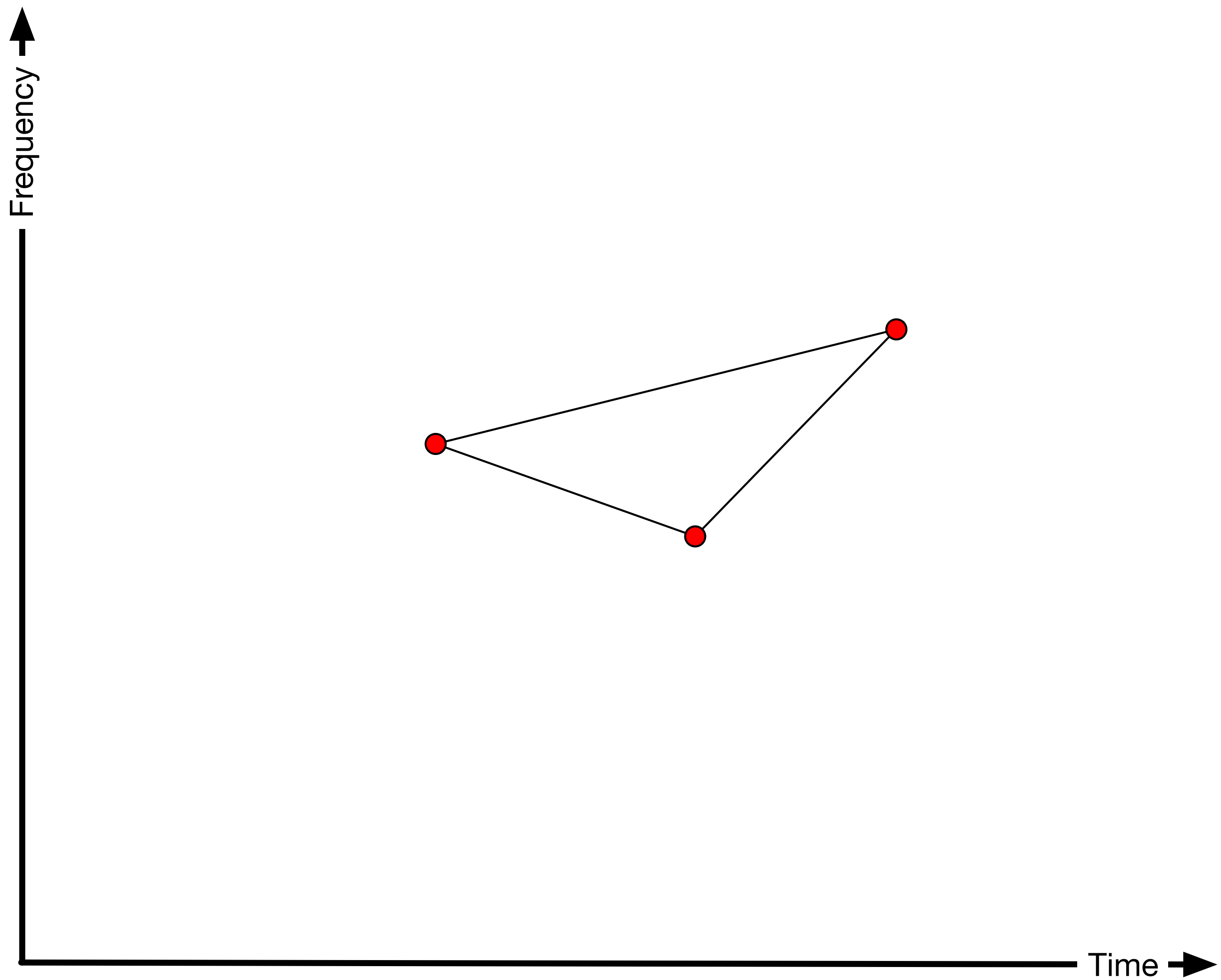
Generalized Fingerprinting



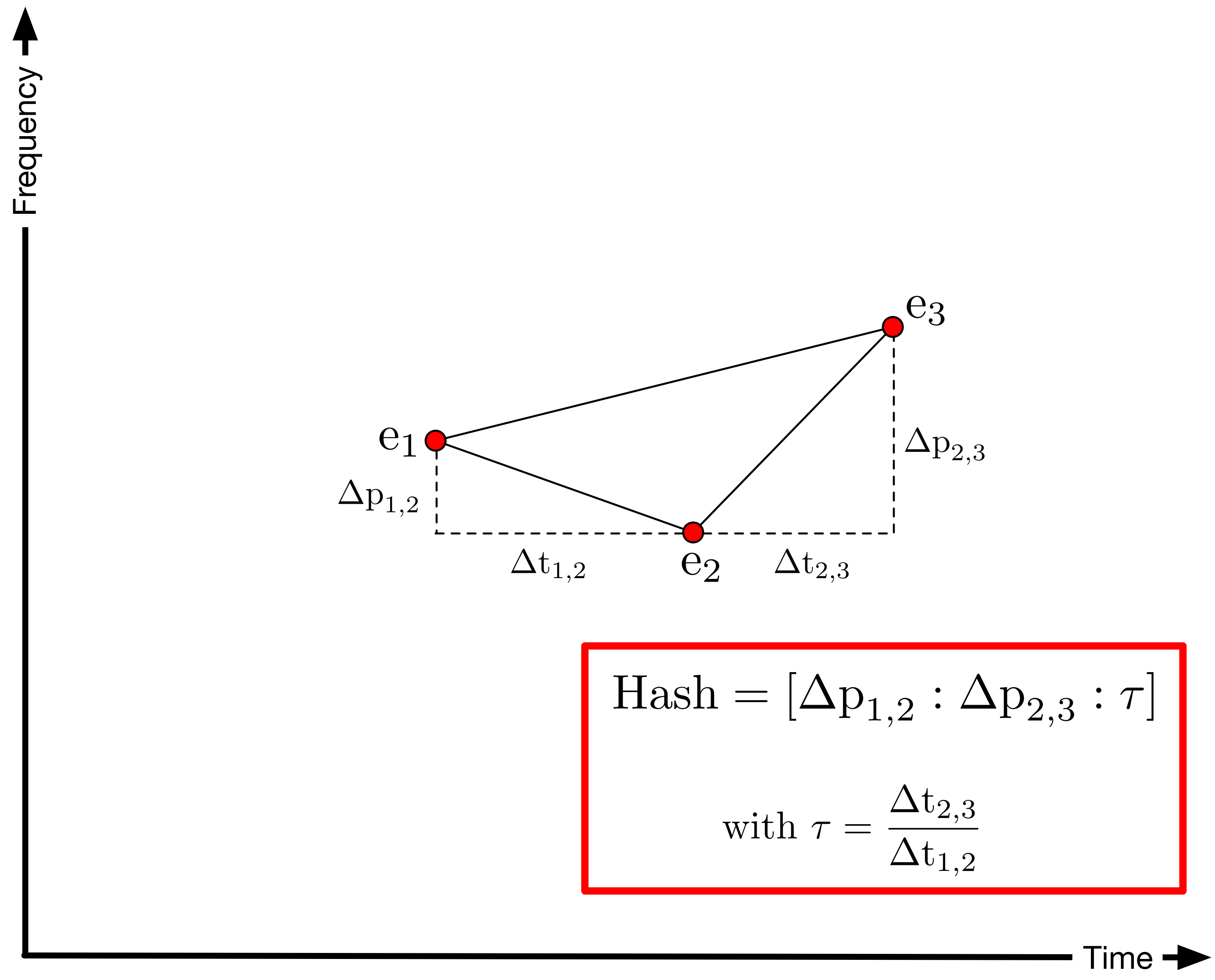
Generalized Fingerprinting



Generalized Fingerprinting



Generalized Fingerprinting

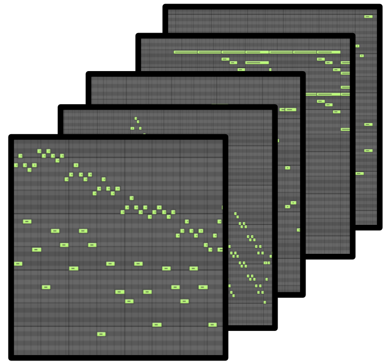


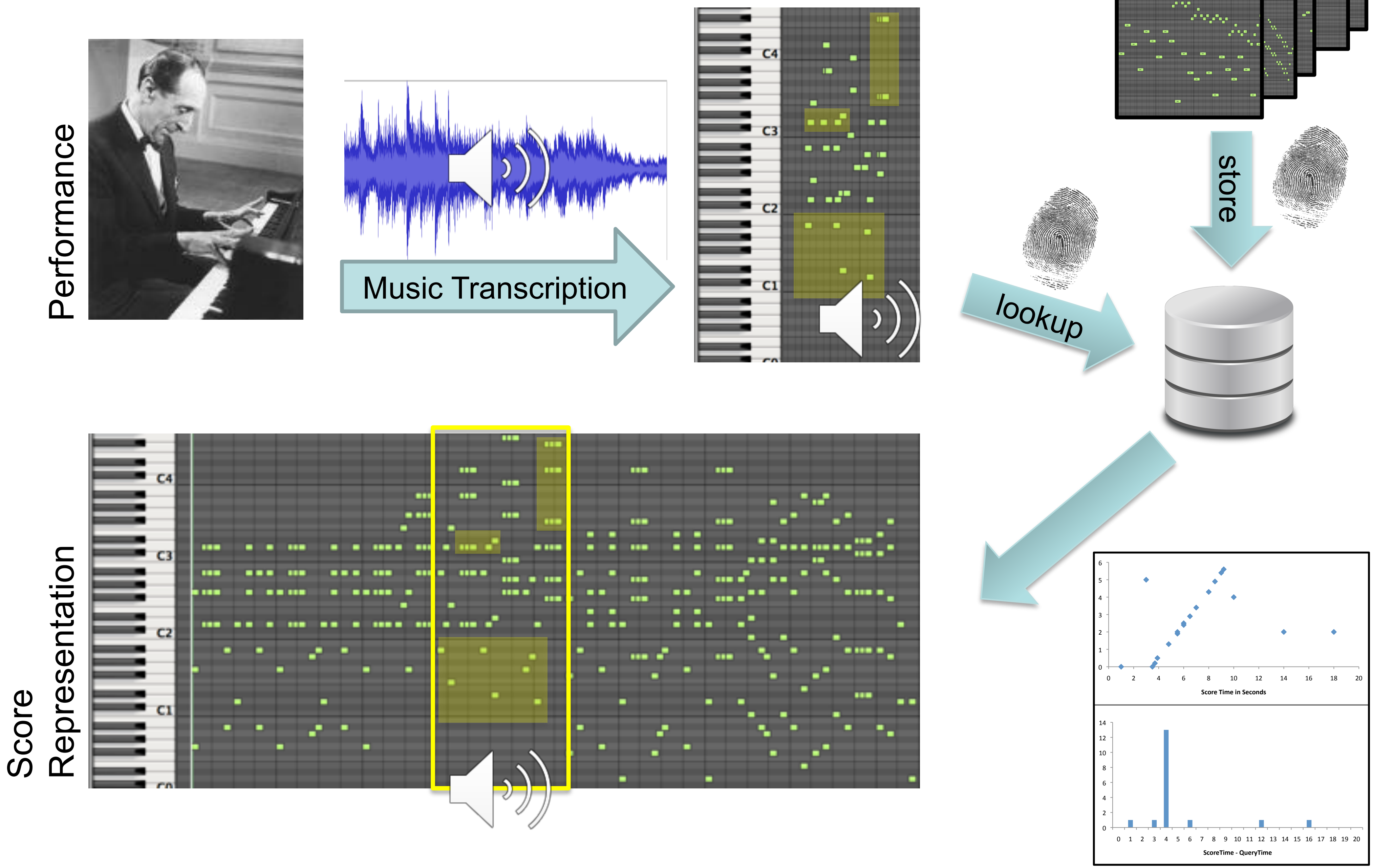
[Arzt, Widmer, Sonnleitner: ISMIR 2015]

Generalized Fingerprinting

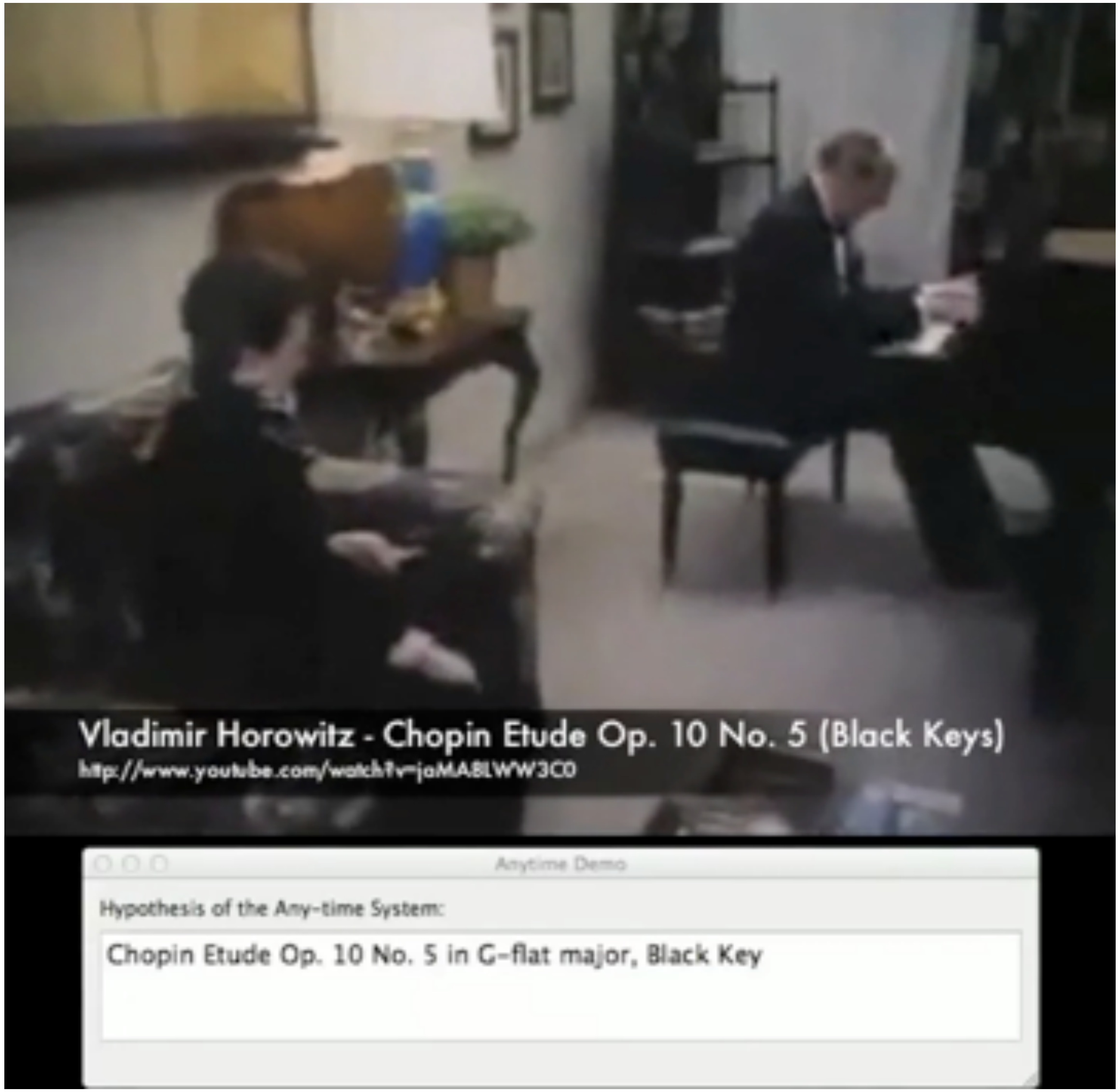
- Relative Representations lead to Tempo- and Transposition-Invariance
- Number of Events per Fingerprint-Token: Trade-off between Discriminative Power and Robustness
  - e.g. Quad-based Fingerprinting [Sonnleitner, Widmer: TASLP 2016]
- Can be used to identify different Performances of the same Piece (“Cover Versions”)
- … and to identify the (symbolic) Score a Performance is based on!

Fast Performance-to-Score Matching





Demo: Fast Performance-to-Score Matching



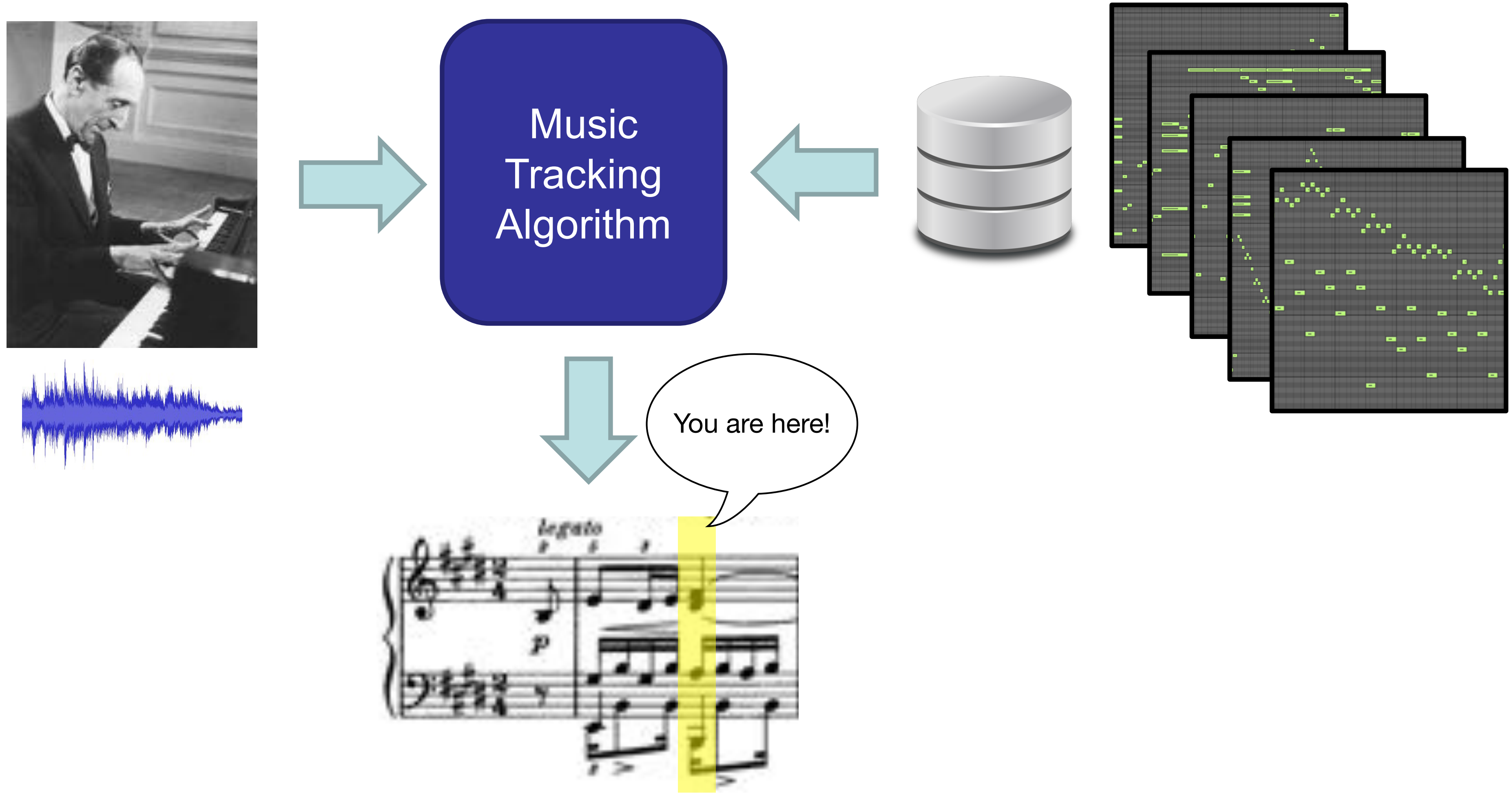
Evaluation (Tempo-invariant Fingerprinting)

- Database Size: more than 1,000,000 notes
  - Mozart, Chopin, Beethoven, …
- For queries with a length of 25 notes:
  - 91% correct piece as top match
  - 0.16 sec. mean execution time
- For queries with length 50 notes (using shingling and other extensions):
  - 98% correct piece as top match
  - 0.49 sec. mean execution time
- With additional transposition-invariance, length 50 notes:
  - 92% correct piece as top match
  - 3.21 sec. mean execution time

Application Scenario

FLEXIBLE MUSIC TRACKING RE-VISITED

Flexible Music Tracking



Flexible Music Tracking Re-visited