

Winter Term 2026/27

Selected Topics in Deep Learning for Audio, Speech, and Music Processing

Lecture

Automatic Music Transcription

Stefan Balke, Andreas Brendel, **Ben Maman**, Meinard Müller

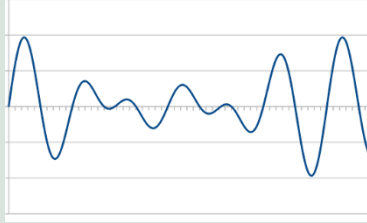
Agenda

1. Introduction to music information retrieval (MIR)
2. Automatic music transcription: data-driven methods
3. Score–audio alignment
4. Multi-instrument transcription

Introduction to Music Information Retrieval (MIR)

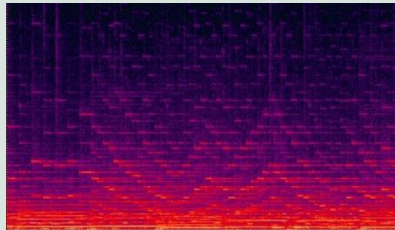
Music Representations

Waveform



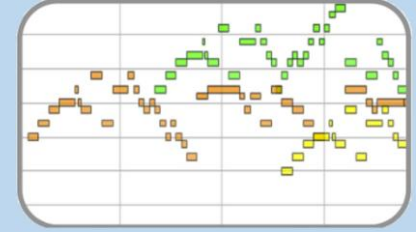
Spectrograms

STFT / Mel /
CQT



Continuous Signal

**MIDI /
Piano Roll**



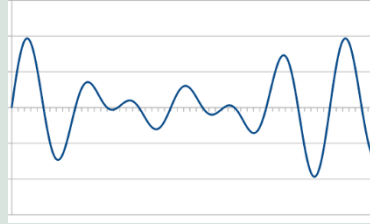
Sheet Music



Symbolic

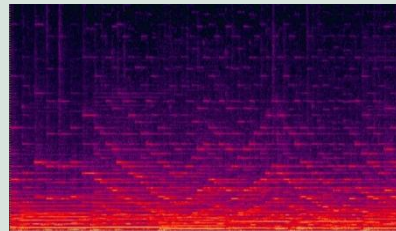
Music Representations

Waveform



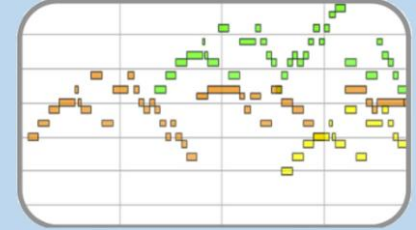
Spectrograms

STFT / Mel /
CQT



Continuous Signal

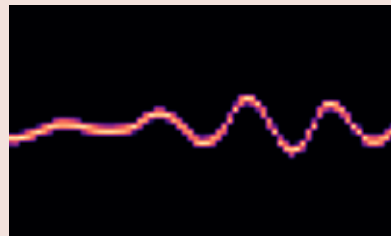
**MIDI /
Piano Roll**



Sheet Music



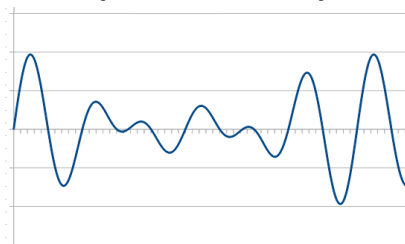
f_0 Contour



Symbolic

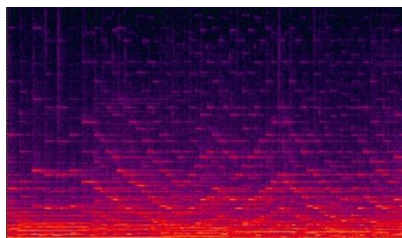
Levels of Abstraction: Continuous → Semantic

Amplitude samples



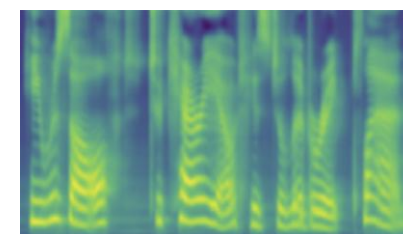
Raw Waveform

Linear frequencies



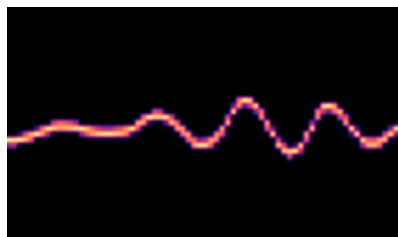
STFT

Perceptual (log) frequencies



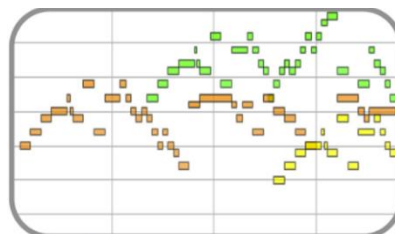
Mel / CQT

Pitch



f_0 contour

Notes, instruments,
physical time



MIDI / Piano Roll

Musical time

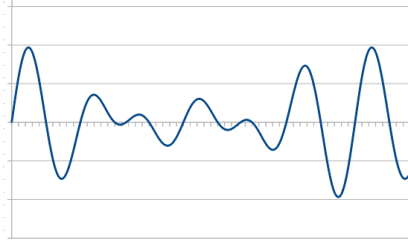


Sheet Music

Levels of Abstraction: Continuous → Semantic

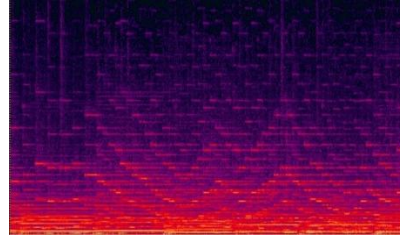
Not strict hierarchy!

Amplitude samples



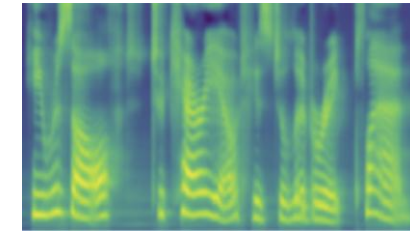
Raw Waveform

Linear frequencies



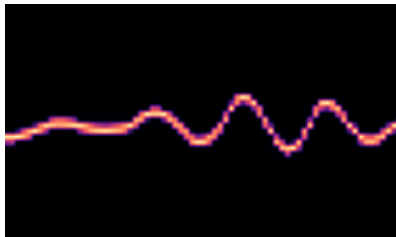
STFT

Perceptual (log) frequencies



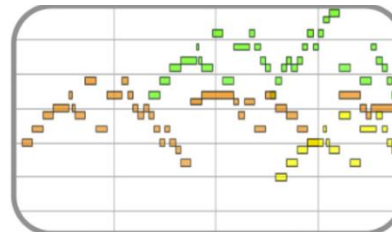
Mel / CQT

Pitch



f_0 contour

Notes, instruments,
physical time



MIDI / Piano Roll

Musical time



Sheet Music

Short-Time Fourier Transform (STFT)

$$X(n, w) = \sum_{m=-\infty}^{\infty} x(m) w(n-m) e^{-jw}$$

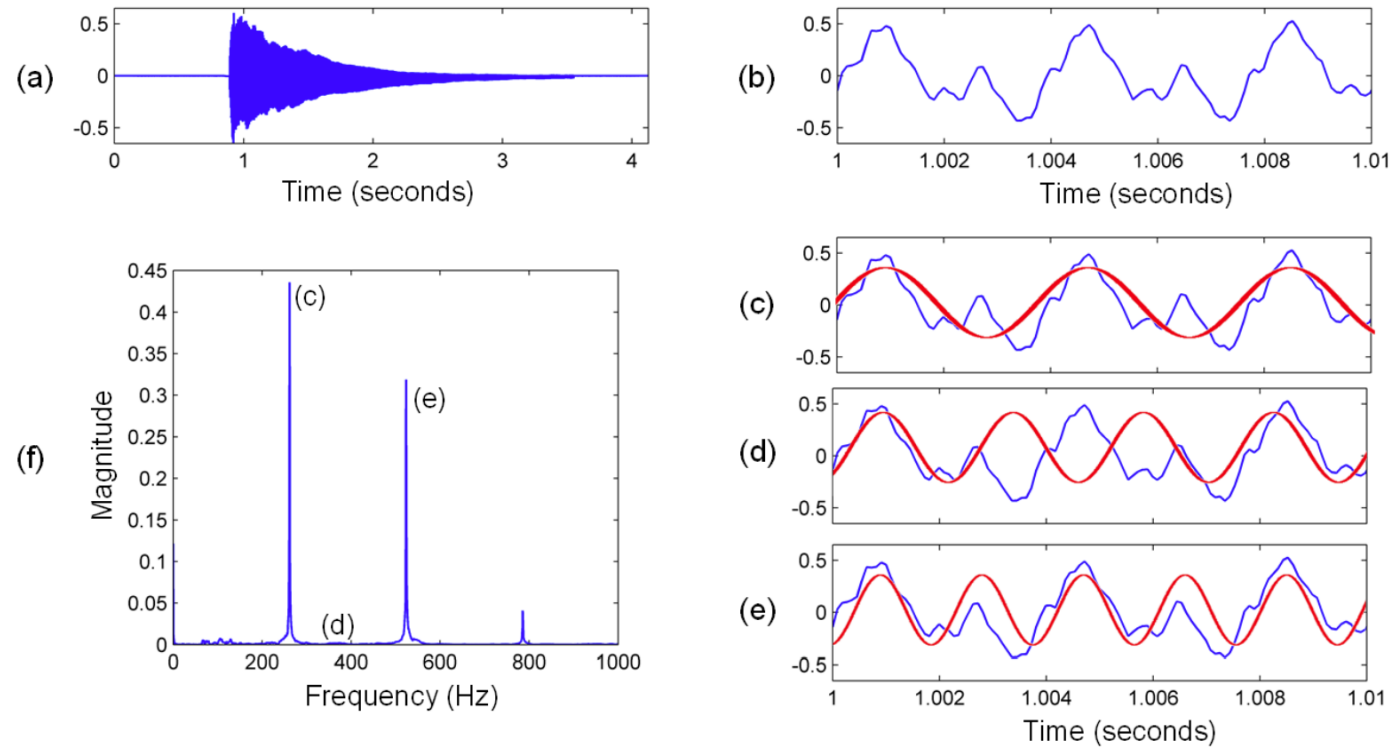
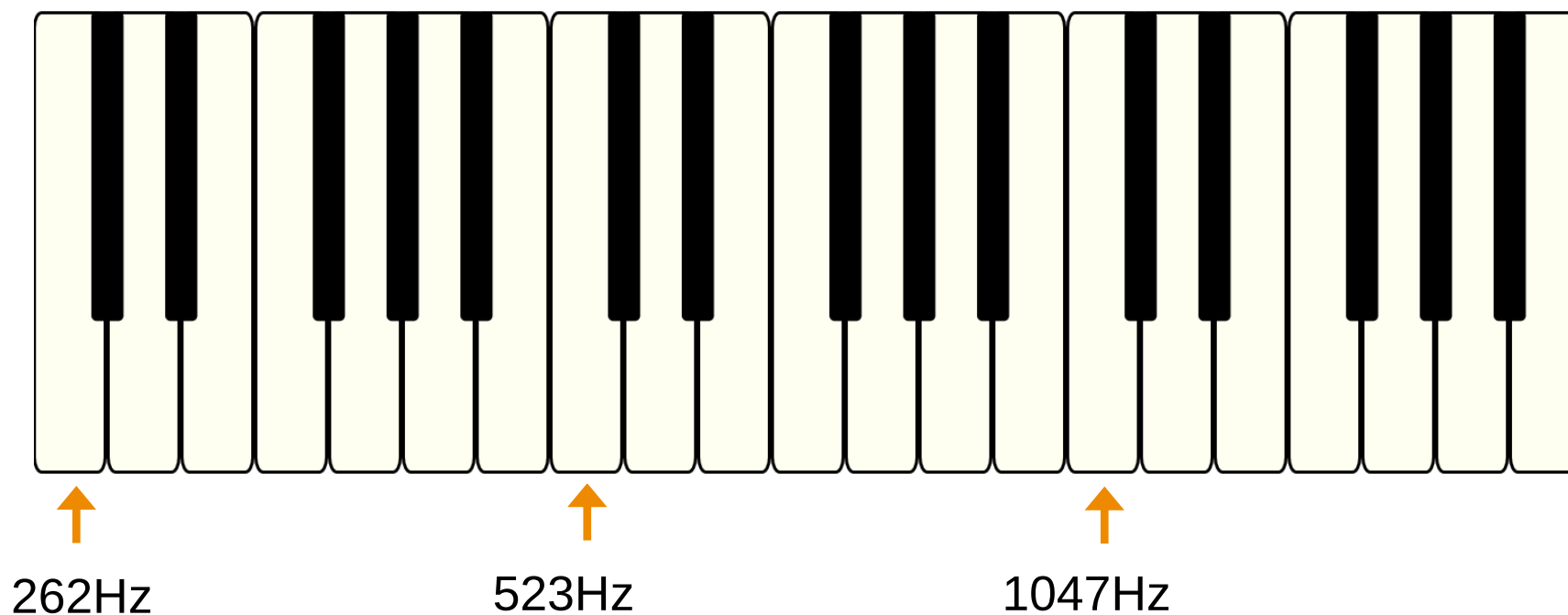


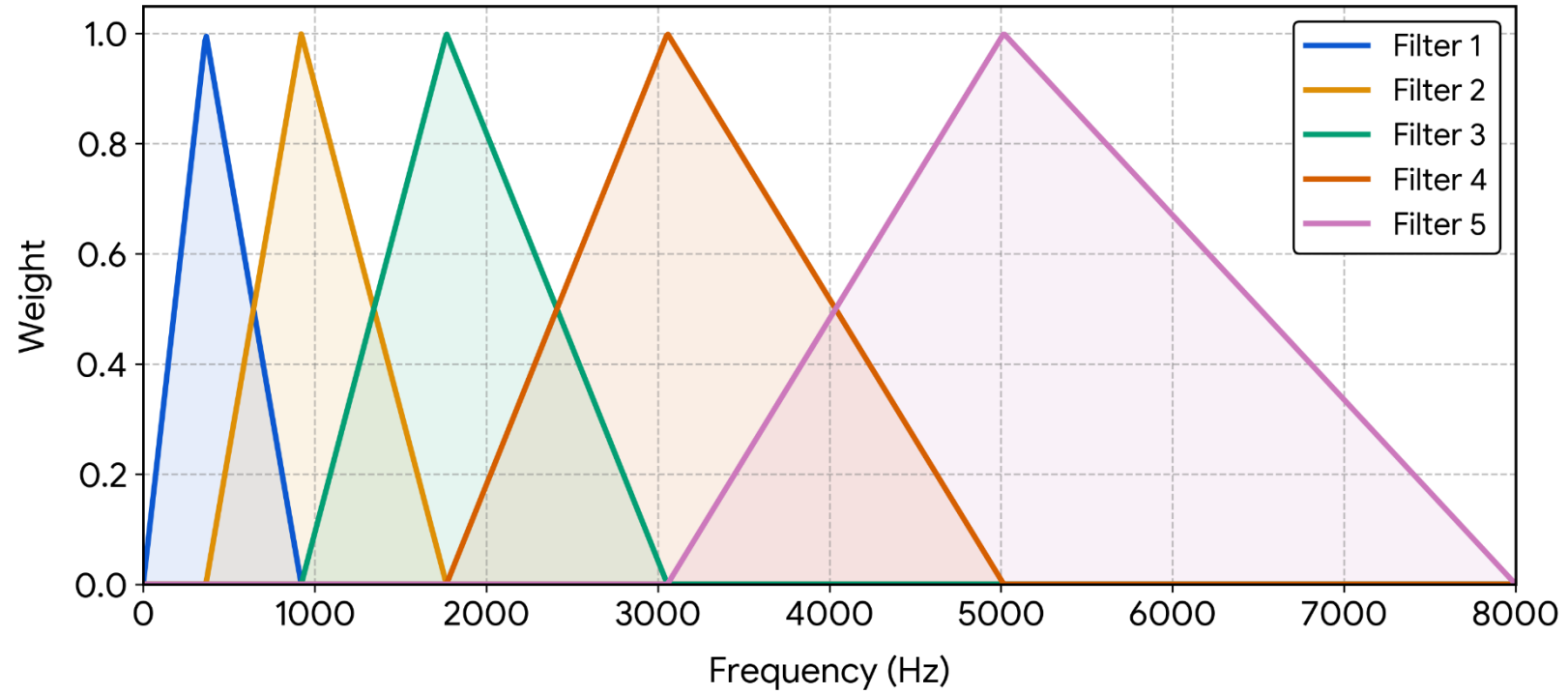
Figure 2.1 from [Meinard Müller, Fundamentals of Music Processing, Springer 2015]

STFT: Perceptual?

- Human perception—logarithmic



Mel Filterbank



$$m = 2595 \log_{10} \left(1 + \frac{f}{700} \right) = 1127 \ln \left(1 + \frac{f}{700} \right)$$

MIDI Scale

- Perceptual pitch scale, logarithmic
- Relation between frequency in Hertz and MIDI scale:

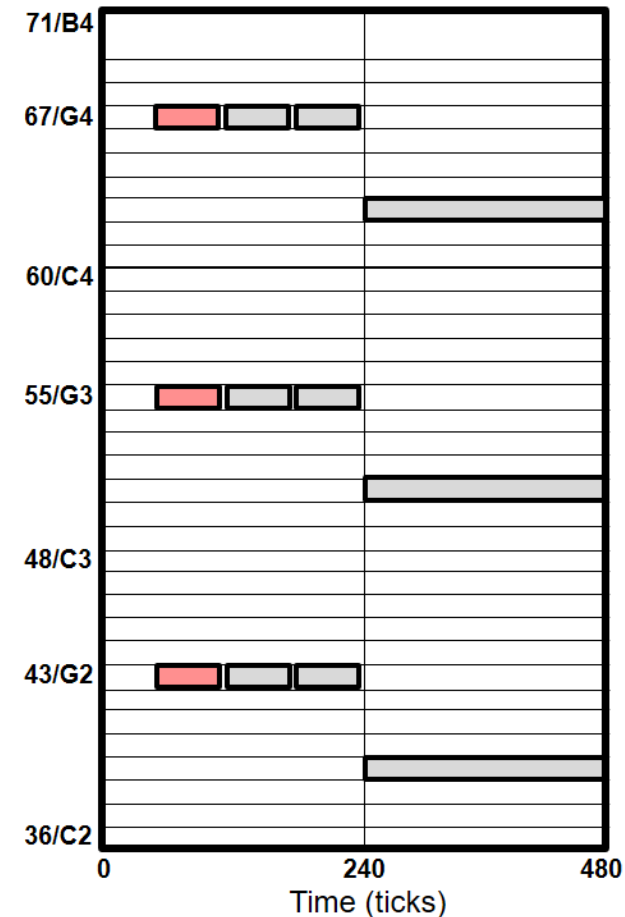
$$f = 440 \cdot 2^{(m-69)/12} \qquad m = 12 \log_2(f/440) + 69$$

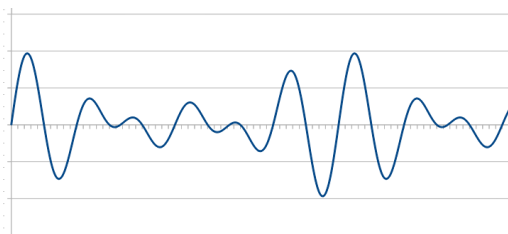
MIDI (Musical Instrument Digital Interface)

Figure 1.13 from
[Müller, FMP, Springer 2015]

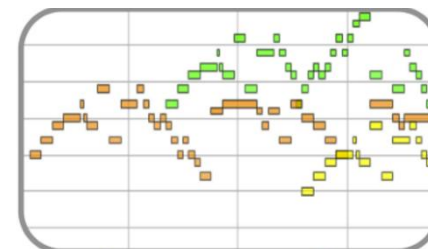


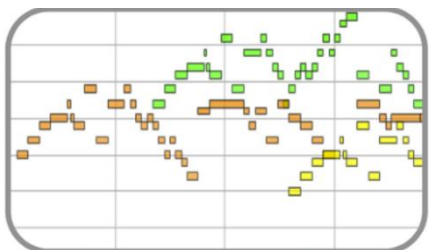
Time (Ticks)	Message	Channel	Note Number	Velocity
60	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	63	100
0	NOTE ON	2	51	100
0	NOTE ON	2	39	100
240	NOTE OFF	1	63	0
0	NOTE OFF	2	51	0
0	NOTE OFF	2	39	0



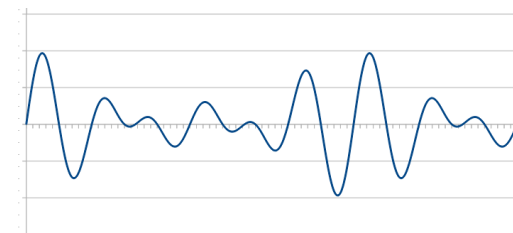


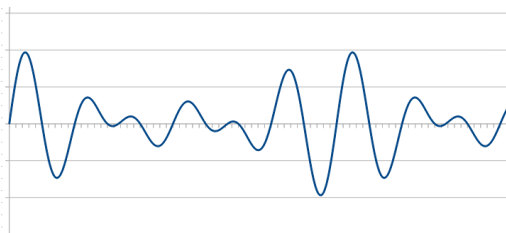
Transcription



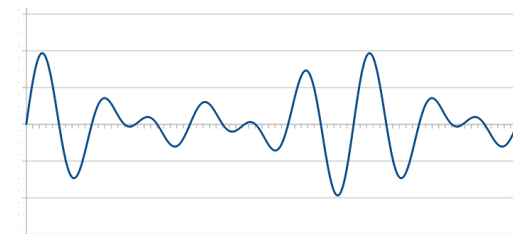


Synthesis





Conversion



Change pitch, singer, timbre...

Fundamental & Partial Frequencies

- Musical notes - **fundamental** frequencies
- Real-world instruments—many **partial frequencies**
Harmonics & non-harmonics
- **Harmonics**—integer multiples
- Pitch—**fundamental**
- Timbre—balance between **partials'** strength

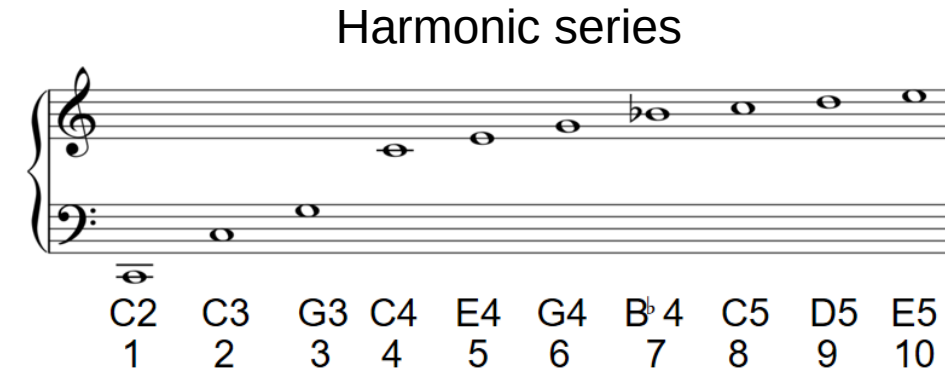
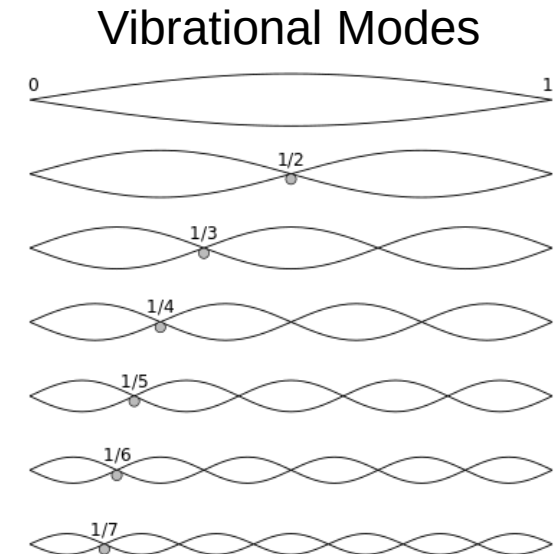
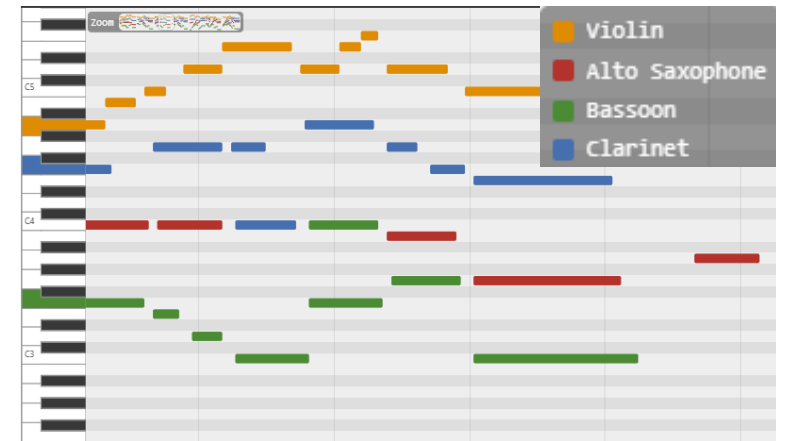
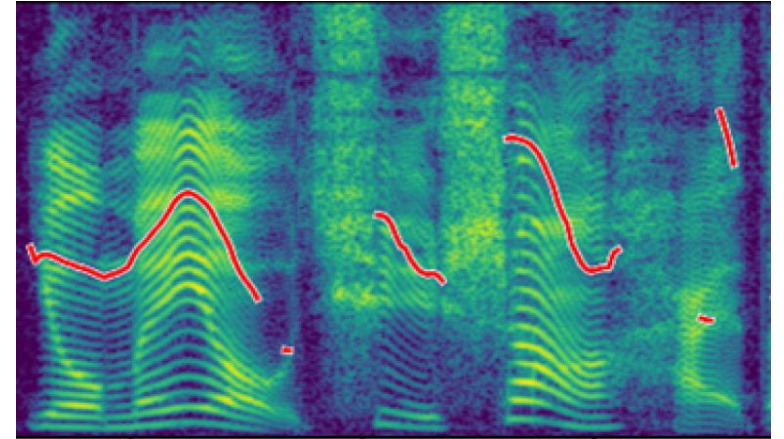


Figure 1.20 from [Meinard Müller, FMP, Springer 2015]



Music Transcription

- Separate fundamentals from other partials
- “Slicing” - note beginning (onset), end (offset)
- Monophonic (single-voice) → elementary algorithms (no ML)
- Polyphonic & multi-instrument → hard
Mixture of many frequencies
mostly partials, some fundamentals

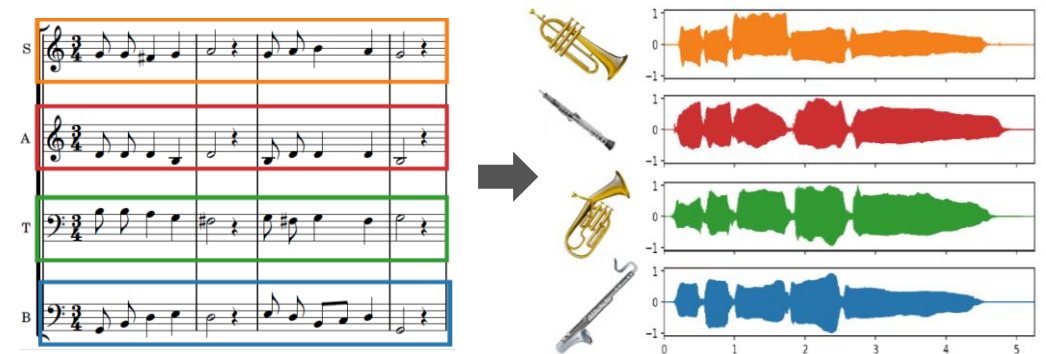


Music Synthesis

- Fundamental frequency known
- Concatenative (sampling-based) synthesis



- Challenge – realism
- What makes realistic sound?
 - Timbre (partial frequencies)
 - Expressivity—loudness & pitch fluctuation...



The Analysis–Synthesis Feedback Loop

- How to train analysis model (e.g., transcriber)?
 - ➔ Synthesizer: accurately annotated data
- How to learn synthesis from authentic data?
 - ➔ Parse with analysis model, e.g., transcriber
 - Score–audio training pairs



Automatic Music Transcription

Demo: <https://benadar293.github.io>

Automatic Music Transcription: What Is It Good For?

- Music education
- Musicology
- Control signals powering music generative models

Transcription Paradigms

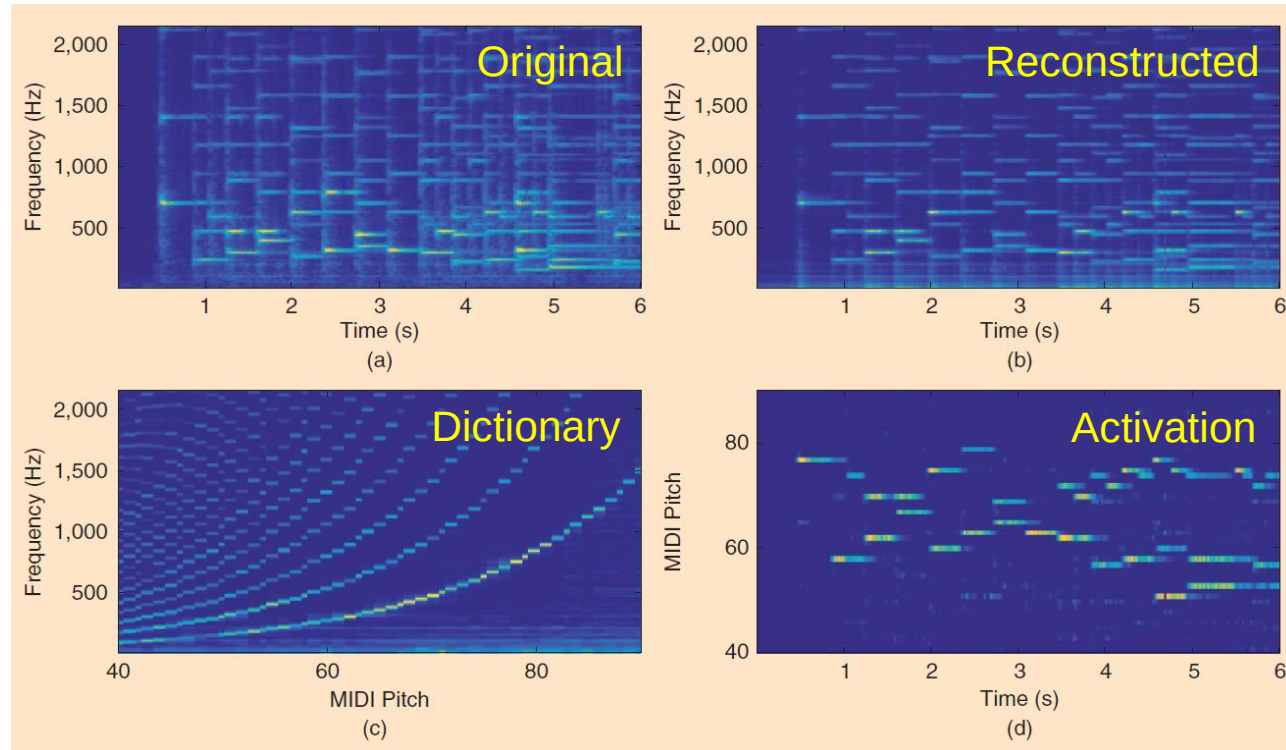
- Single-instrument: e.g., piano transcription
- Instrument-agnostic: detect pitch only within multi-instrument setting
- Instrument-aware: detect pitch & instrument
- Frame-level
- Note-level

Automatic Music Transcription: Data-Driven Methods

NMF (Non-Negative Matrix Factorization)

Smaragdis and Brown: Non-Negative Matrix Factorization for Polyphonic Music Transcription, WASPAA 2003
Benetos et al., Automatic Music Transcription: An Overview, IEEE Signal Processing Magazine, 2019

- **$S = DA$** Spectrogram = Dictionary x Activations

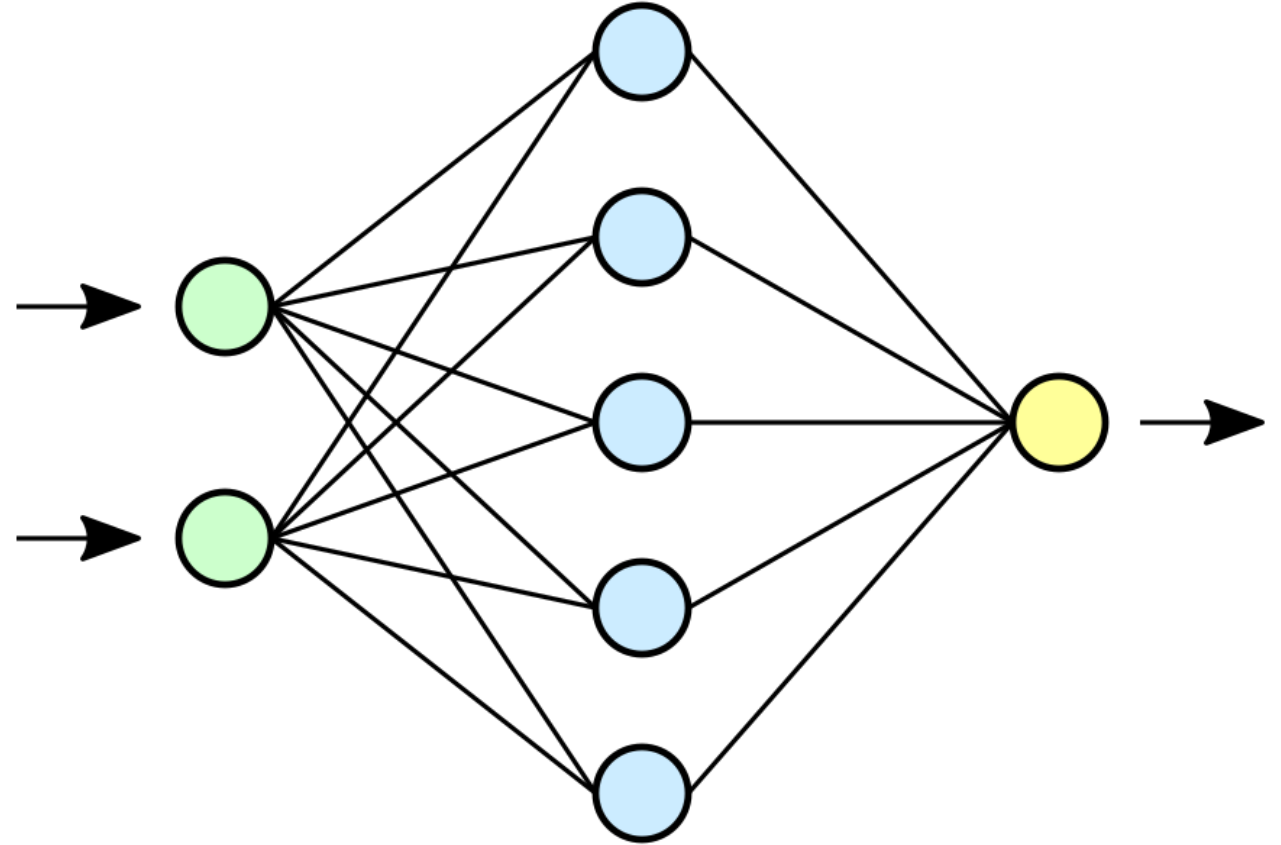


Transcription with Neural Networks

Neural Network

Composition of:

- Linear maps
Coefficients = network weights
- Non-linear activation functions

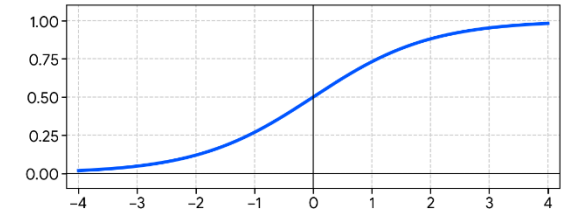


Activation Functions

- Make network non-linear → Expressive

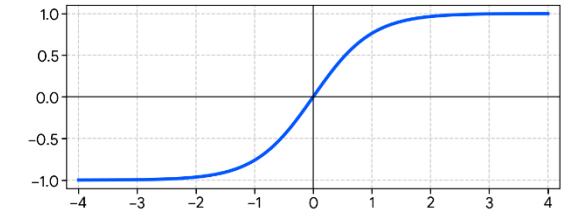
Sigmoid

$$\frac{1}{1 + e^{-x}}$$



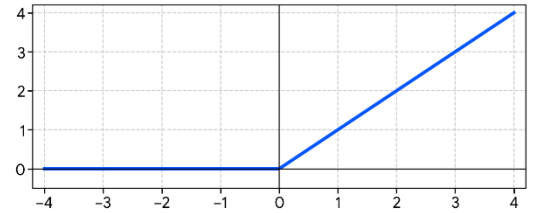
Tanh

$$\frac{e^x - e^{-x}}{e^x + e^{-x}}$$

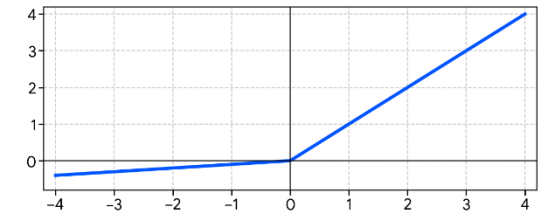


ReLU

$$\max(0, x)$$

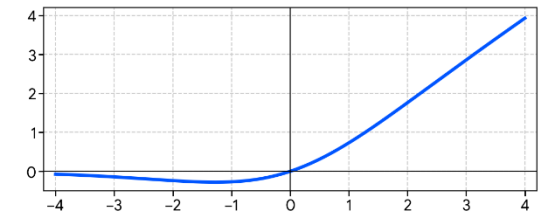


Leaky ReLU
 $\max(0.1x, x)$



Swish

$$x \cdot \frac{1}{1 + e^{-x}}$$



Loss Functions

- Measure difference between **current** and **target** prediction
- Example: Binary cross-entropy (binary classification)

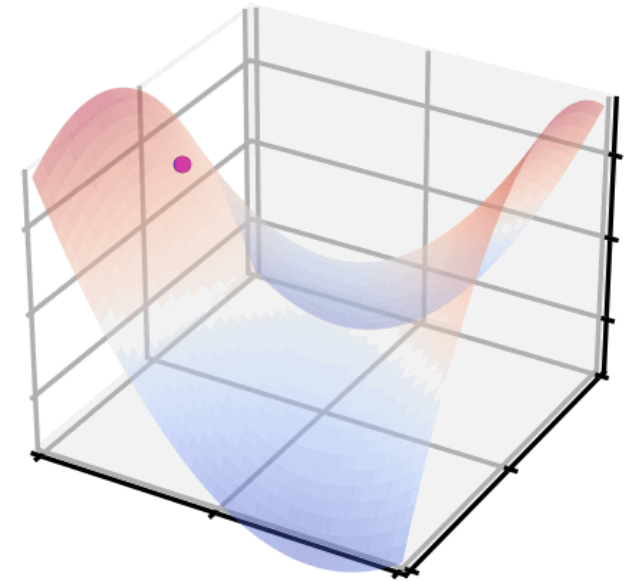
$$\text{BCE} = -\frac{1}{N} \sum_{i=1}^N \left[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \right]$$

- Objective: Find network weights that minimize loss

Stochastic Gradient Descent

- Optimization method
- Compute direction of best improvement in loss
- Update network weights

$$w_{t+1} = w_t - \alpha \frac{\partial L}{\partial w_t}$$



- All operations must be differentiable!

Transcription as Classification

- P pitches, T time frames
- E.g.: P = 88, T = 600
20 seconds, approx. 30 FPS
- Model decides which pitch active at each frame → classification!

Monophonic vs. Polyphonic Signals

- Monophonic signal (single pitch at a time):
 - Simple multi-class classification
 - Loss: Categorical Cross Entropy
- Polyphonic signal (multiple pitches at a time):
 - Multi-class, multi-label classification
 - Set of multiple binary classifications
 - Loss: Sum of Binary Cross Entropy

Piano Datasets: MAESTRO, MAPS

Hawthorne et al., Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset, ICLR 2019

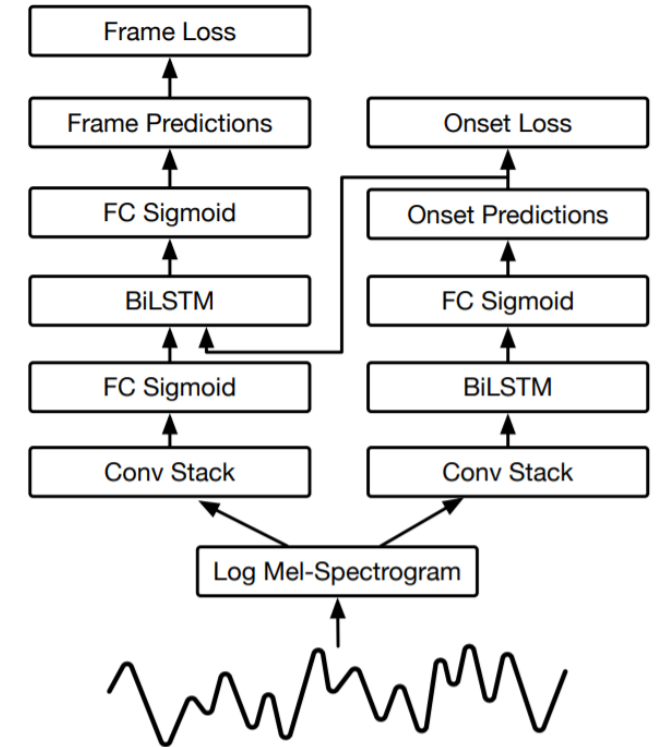
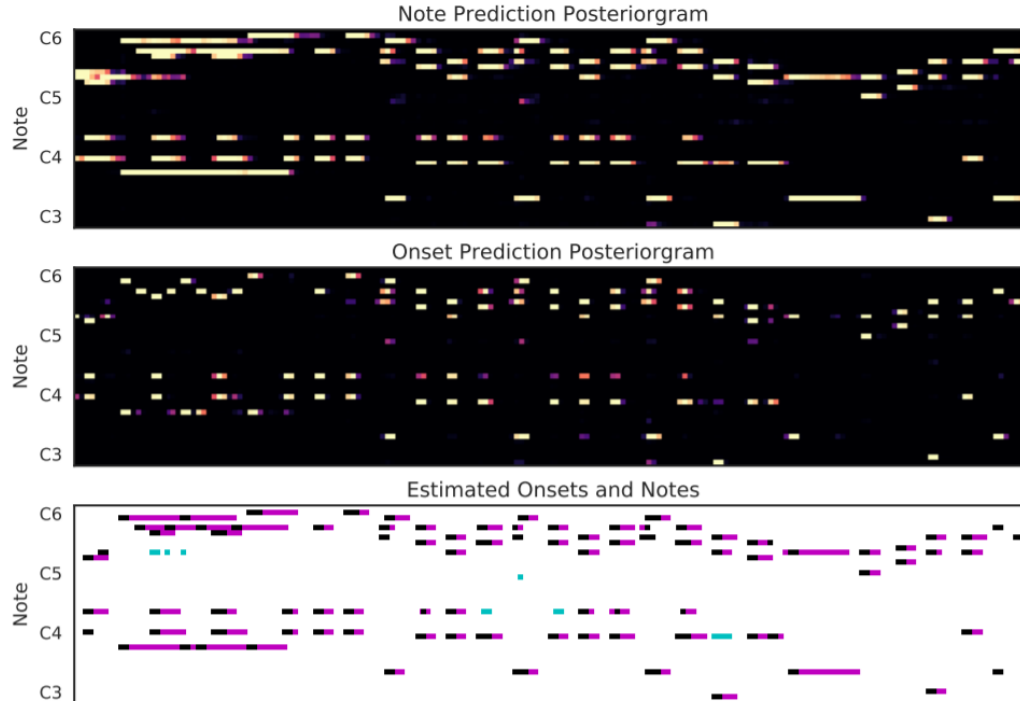
- Fully supervised datasets
- Audio & MIDI from Disklavier
- Nearly perfect alignment
- MAPS—train synth, test real.
- MAESTRO—130 hours real



Onsets and Frames

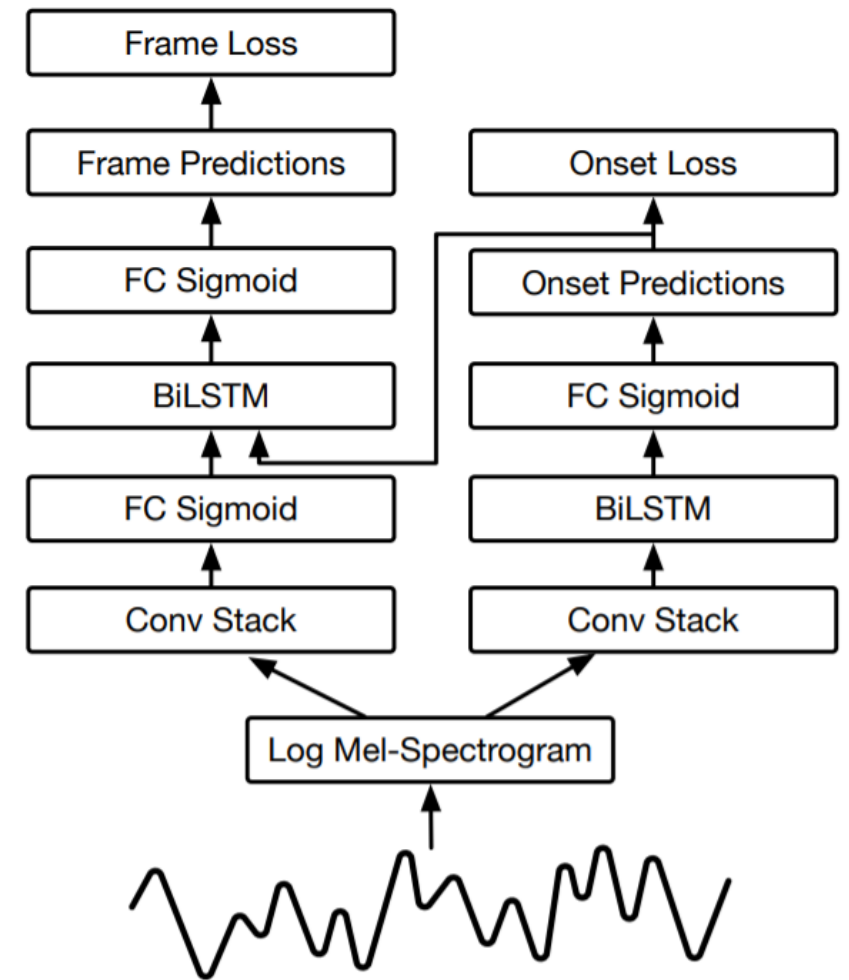
Hawthorne et al., Onsets and Frames: Dual-Objective Piano Transcription, ISMIR 2017

- Focus: Note onsets
- Complementary: Note frame-level activity detection



Onsets and Frames

- Convolutional layers & long short-term memory (LSTM) layers
- Convolution: Learned pattern-matching filters
- Pitch & time equivariance
- LSTM: Broader temporal context



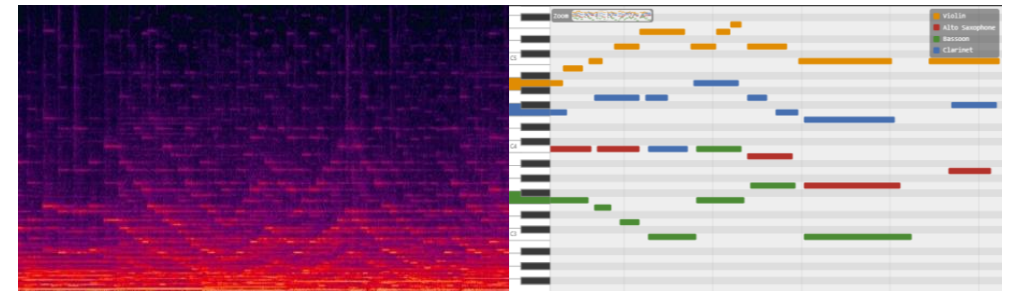
Other Instruments / Genres? Multi-Instrument?

No specialized hardware!



Data Problem

- We can't annotate everything
Expensive, sometimes impossible
Sequences – even worse
- Need unsupervised or weakly-supervised learning



Unaligned Supervision

- Full supervision from manual annotation infeasible
Often practically impossible (*precise* timing in milliseconds, e.g., music)
- **Unaligned / weak supervision easy**



Unaligned Supervision

A YouTube video thumbnail showing a conductor leading an orchestra in a concert hall.

Beethoven - Symphony No. 5 (Proms 2012)

11M views • 9 years ago

Mandetriens

Prom 12: Beethoven Cycle – Symphonies Nos. 5 & 6 Beethoven - Symphony No. 5 in C minor, Op. 67 1 - Allegro con brio 2 ...

A YouTube video thumbnail showing Herbert von Karajan conducting the orchestra.

Ludwig van Beethoven 5th Symphony By Herbert Von Karajan

304K views • 2 years ago

Champ

Quinta sinfonía de Beethoven dirigida por Herbert von Karajan.

A YouTube video thumbnail showing Iván Fischer conducting the Concertgebouworkest.

Beethoven - Symphony No. 5 - Iván Fischer | Concertgebouworkest

323K views • 2 years ago

Concertgebouworkest

Beethoven - Symphony No. 5 - Iván Fischer | Concertgebouworkest Watch more videos free of charge on our...



Beethoven Symphony 5.mid

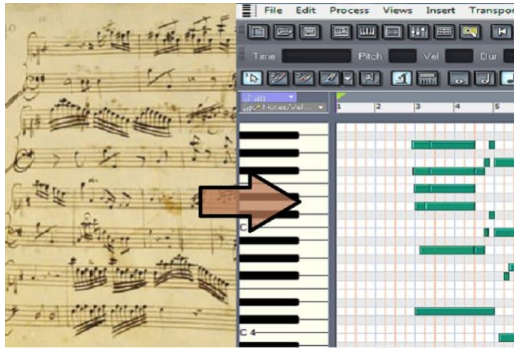
Unaligned Supervision

- Create accurate supervision for easily attainable data



Real-world
recording

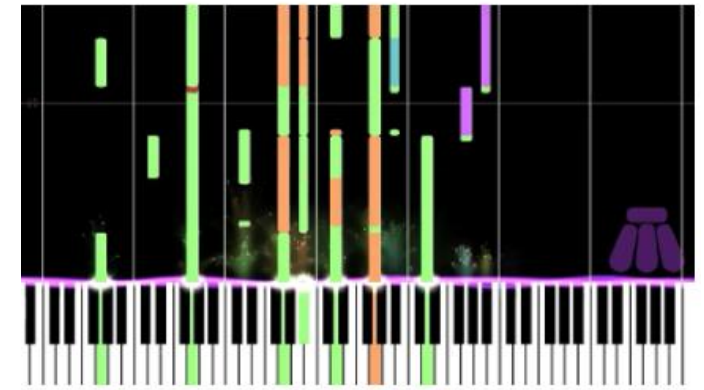
+



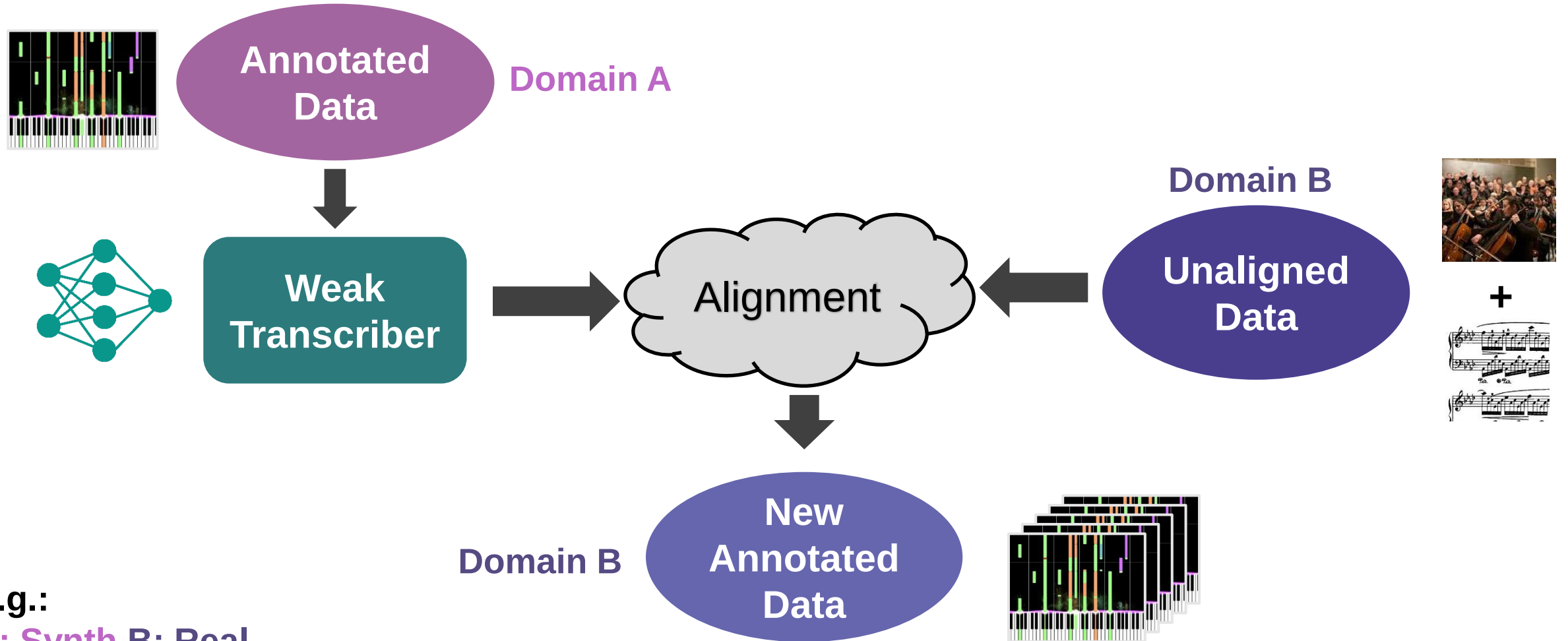
Unaligned
digitized
score

Train

Transcriber



Unaligned Supervision



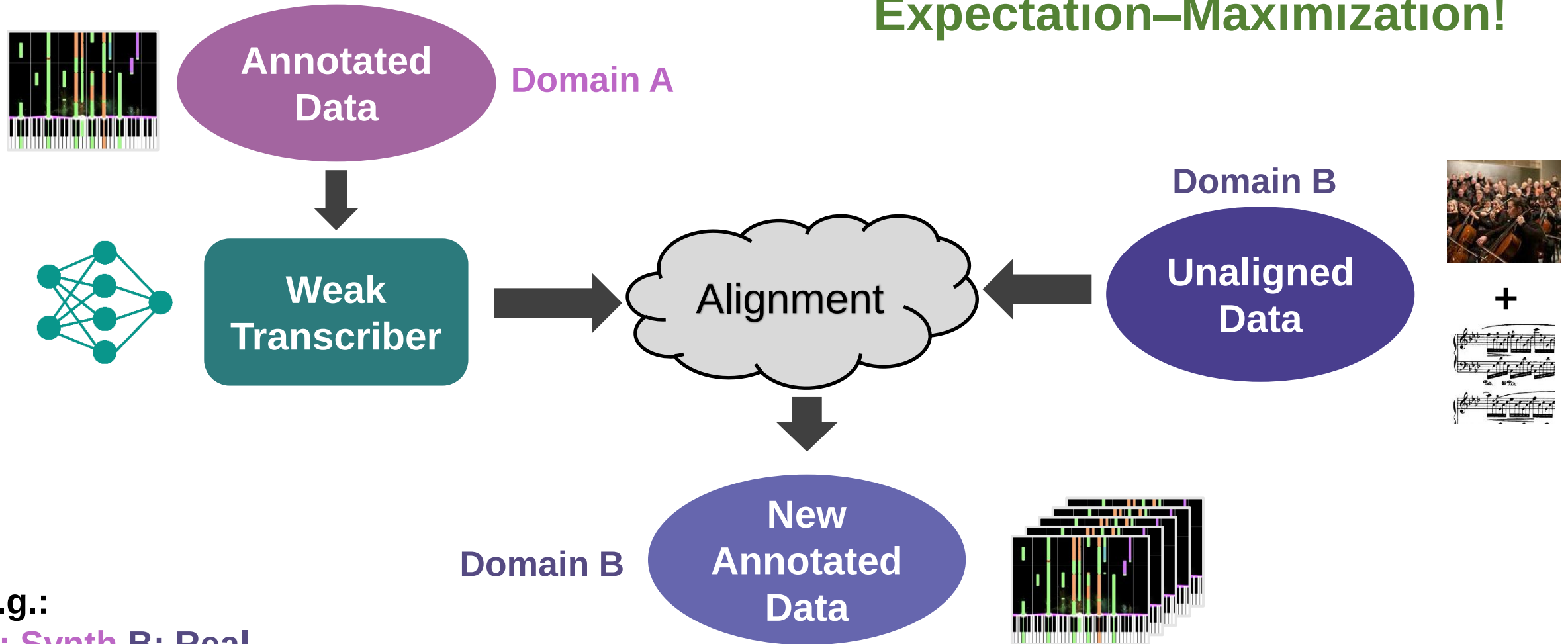
E.g.:

A: Synth B: Real

A: Piano B: Guitar

Unaligned Supervision

Expectation–Maximization!



E.g.:

A: Synth B: Real

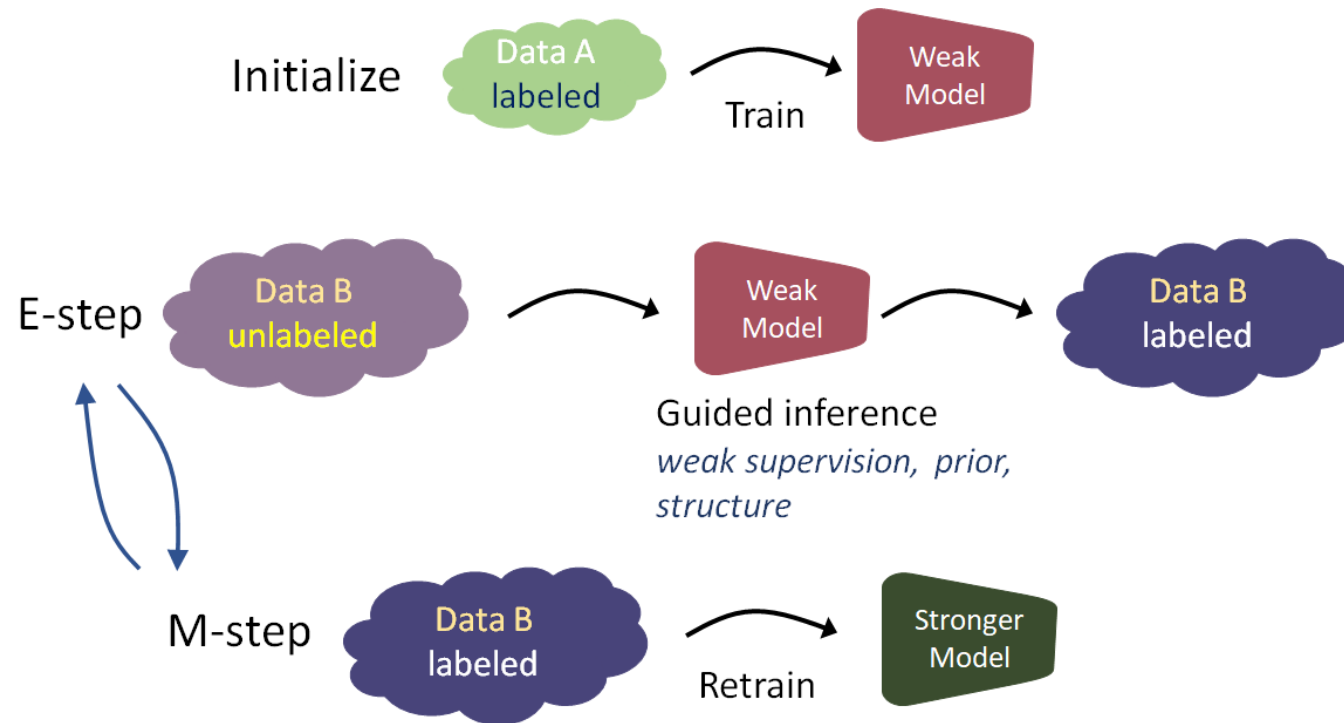
A: Piano B: Guitar

Unaligned Supervision

- Key insight: Alignment easier than transcription
- Detect **When** note occurs easier than detecting **whether or not** it occurs
- Weak model can still produce good alignment!

Expectation–Maximization

Expectation–Maximization (EM)



- Unsupervised / weakly-supervised learning
- label and train “simultaneously”
- E.g., K-Means, Baum-Welch, Inside-Outside, Deep Clustering

EM-Based Machine Translation

- IBM Alignment Model

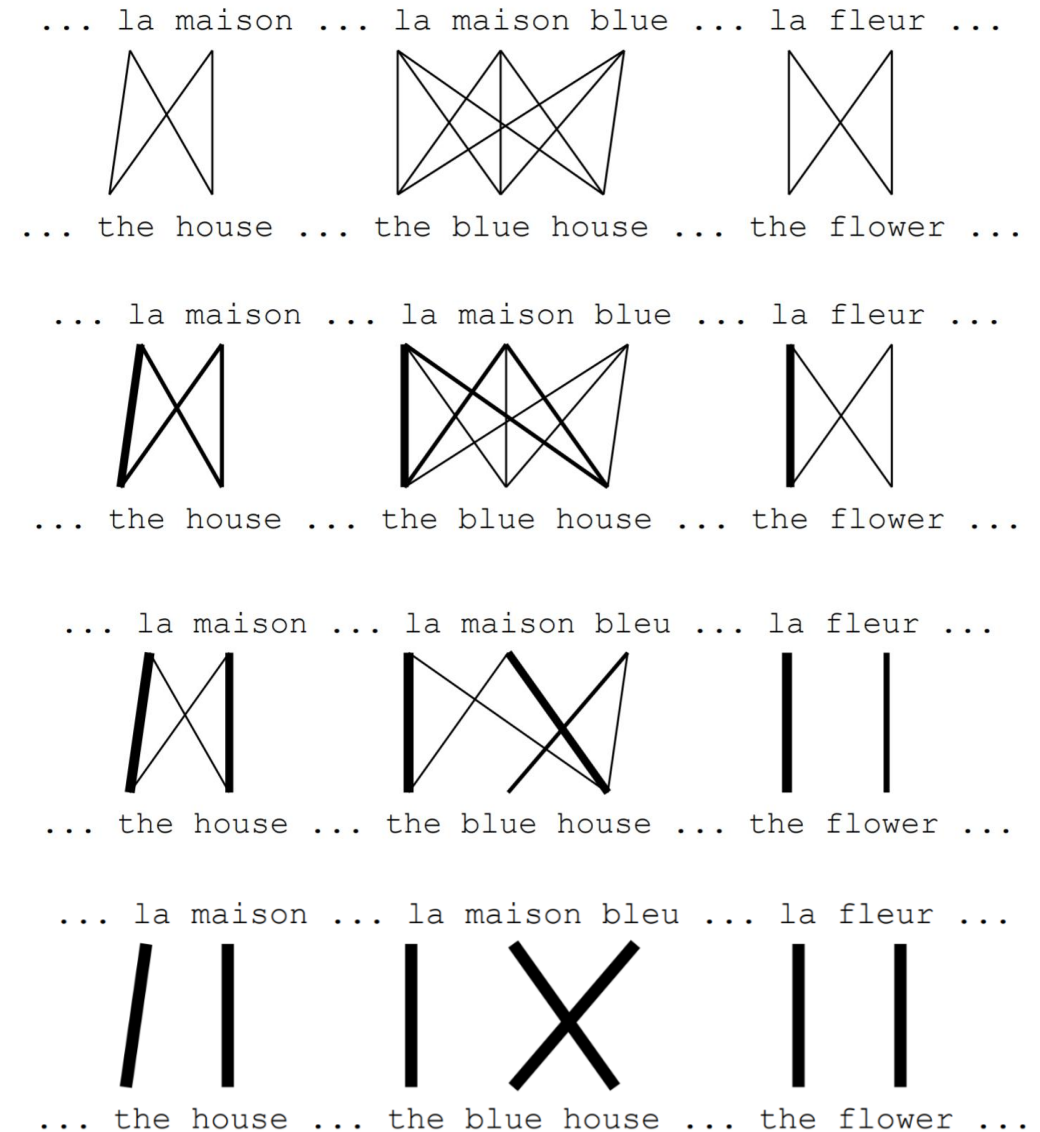
- Learn word-level translation

Haus => house, building, home...

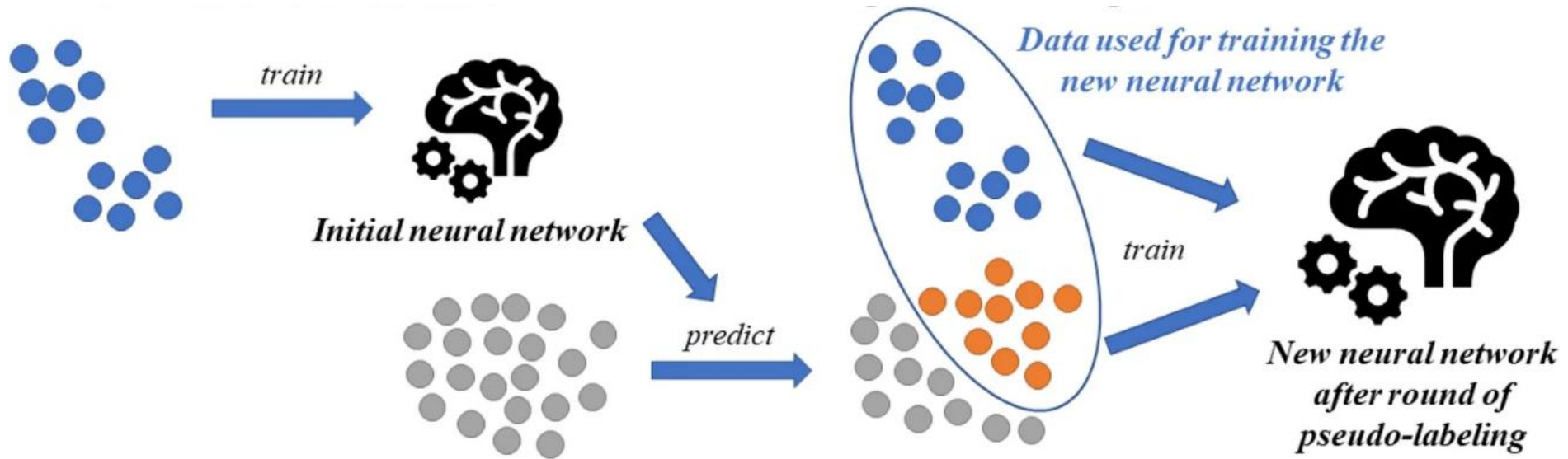
- Only paired sentences

- Weakly-supervised

Brown et al.: The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics 1993



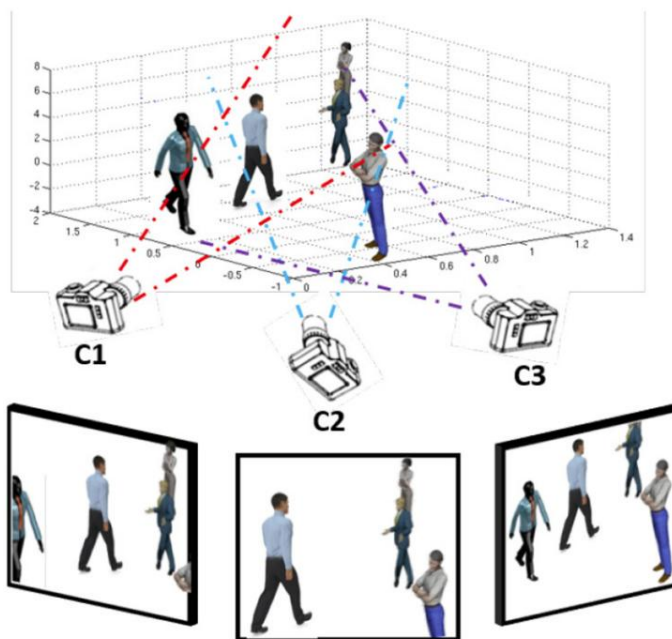
Self-Refinement



- Can network self-annotate data? Naïve approach—pseudo-labels
- Noisy—perpetuate mistakes
- Can still improve

Caron et al., Deep Clustering for Unsupervised Learning of Visual Features, ICCV 2018
Xie et al., Self-training with Noisy Student improves ImageNet Classification, CVPR 2020

Self-Refinement



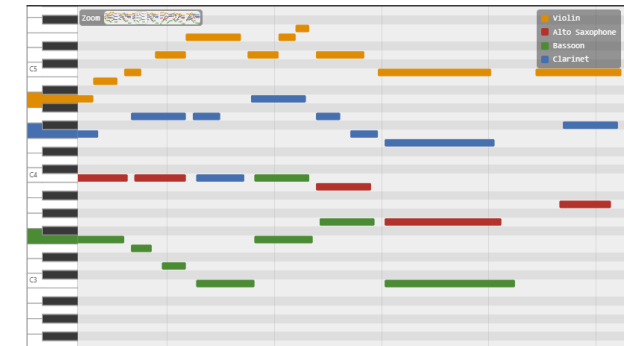
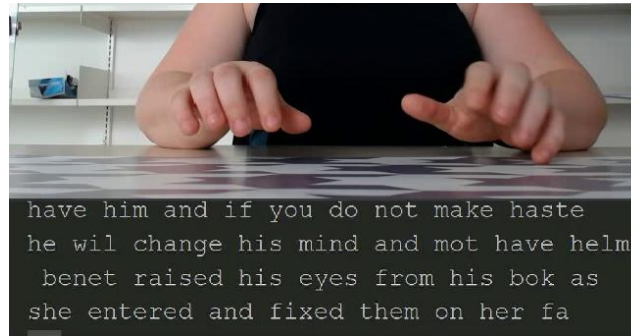
- Structure / prior?

Multi-view, musical score...

Higher accuracy, less noise, new knowledge

Continuous signal → Natural Language

- Speech recognition
 - Lip reading
 - Flat-surface text entry
 - Automatic music transcription
-
- Unaligned labels often available
 - Good prior for EM

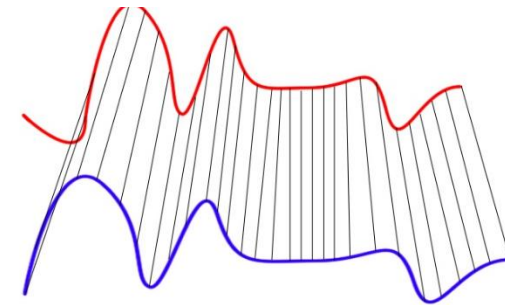
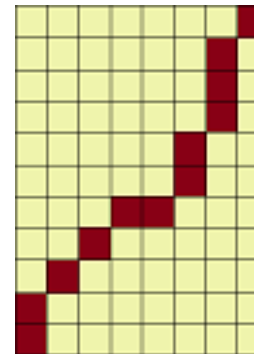
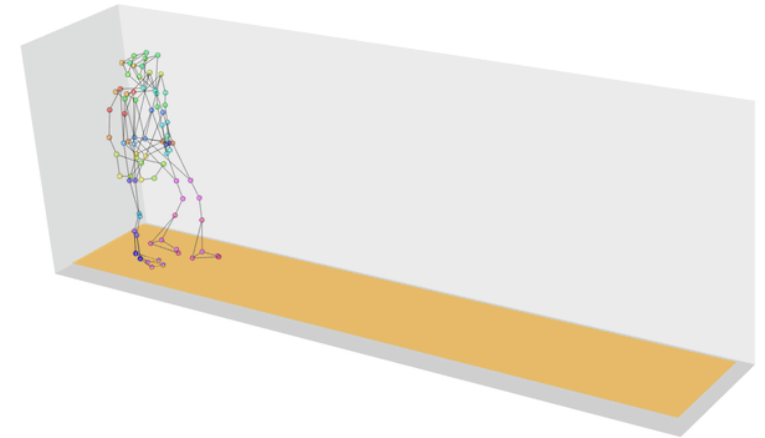


Score–Audio Alignment

Dynamic Time Warping (DTW)

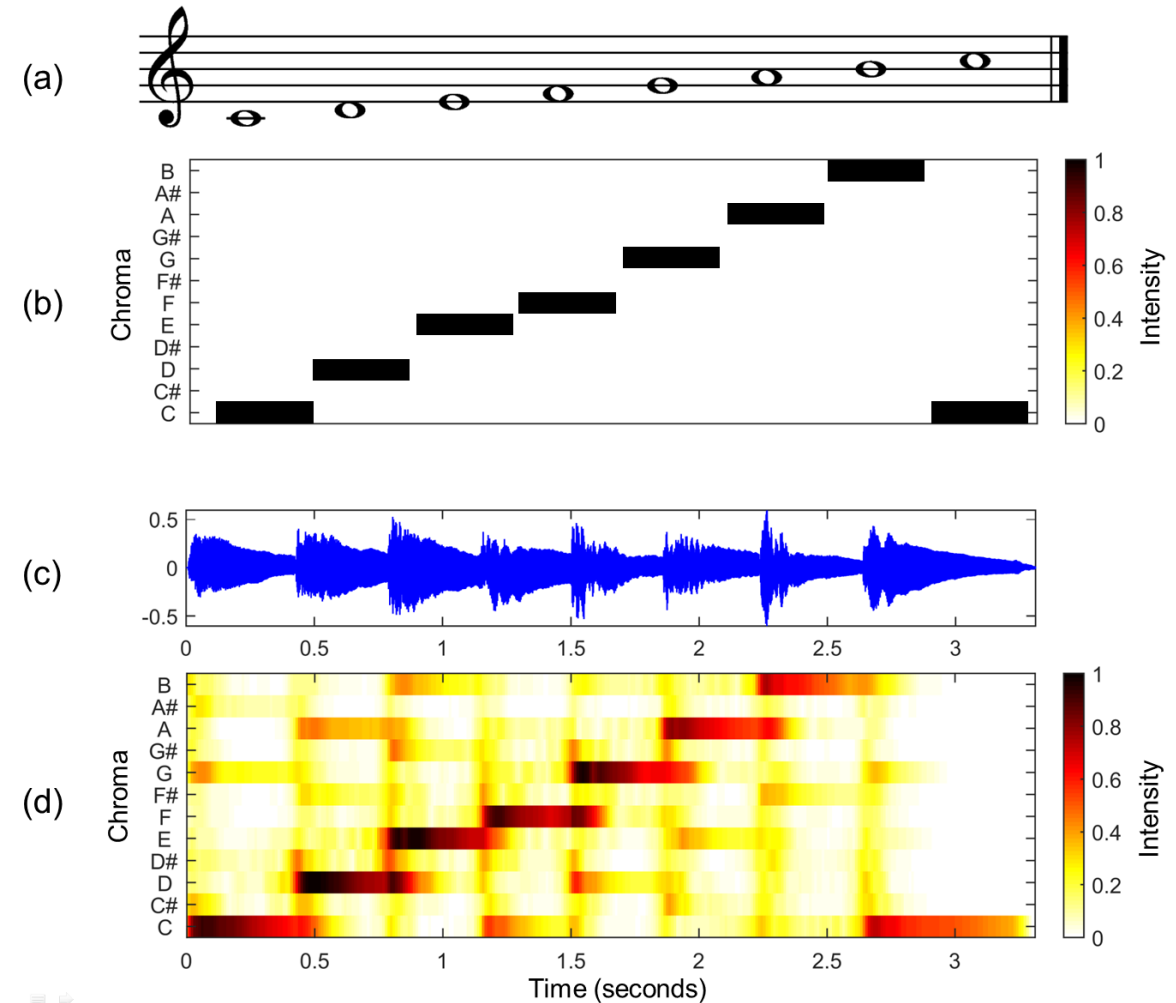
- Optimal monotonic warping between two sequences
- Our case: Warp score to audio
- Dynamic programming solution
- Need local cost function

$$\begin{aligned} & \underset{t \in \mathbb{Z}^n}{\text{minimize}} && \sum_{i=1}^n C(\mathbf{X}_{t_i}, \mathbf{Y}_i) \\ & \text{subject to} && t_0 = 0, \\ & && t_n = m, \\ & && t_i \leq t_j \quad \text{if } i < j. \end{aligned}$$



DTW: Local Cost Function

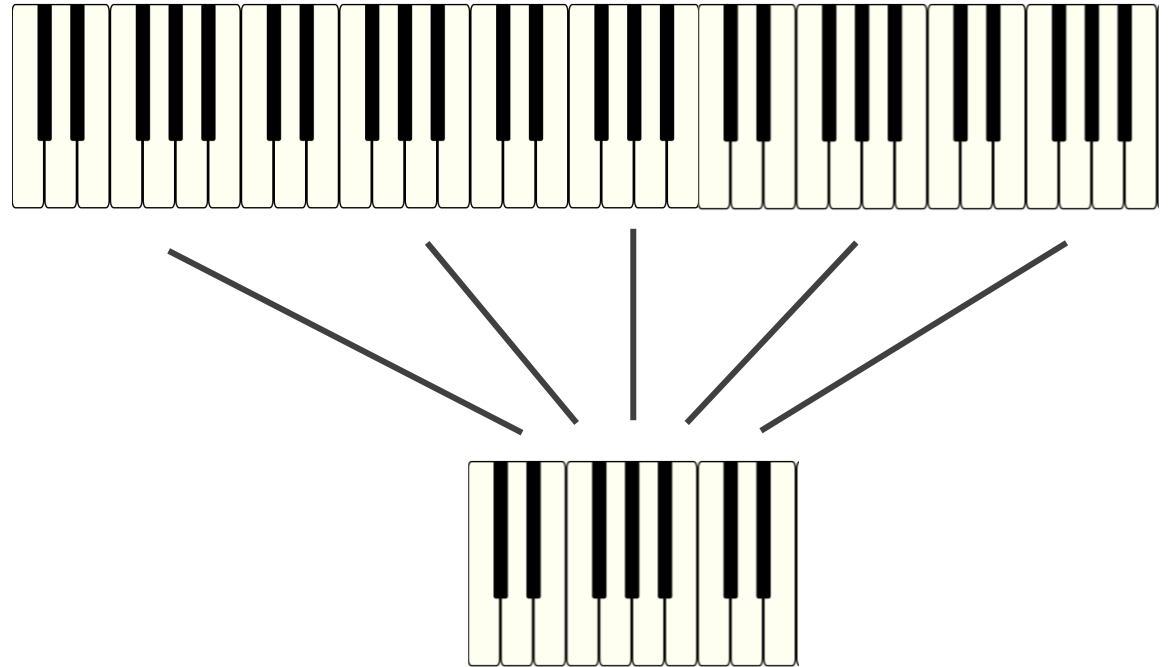
- Compare sequence elements, e.g., audio & score
- Audio frame & musical score frame—how similar?
 - Low-frequency STFT features
 - Transcription features
 - Chroma Features



Müller et al., Audio Matching via Chroma-Based Statistical Features, Ismir 2005
Thickstun et al., Learning Features of Music from Scratch, ICLR 2016
Maman and Bermano, Unaligned Supervision for Automatic Music Transcription in The Wild, ICML 2022

Chroma Features

- Maximize activation across octaves
- DTW quadratic time
- 88 pitches → slow
- 12 chroma classes → fast



DTW in Polyphonic Settings: Problem

- DTW only sequence-level, not individual notes
- Chord from score not exactly simultaneous in audio? E.g., arpeggio
- Pitch activity detection: Most of note correct ✓
- Onset detection: Many wrong onsets ✗
- Offset detection: Same ✗

Sequence- vs. Note-Alignment

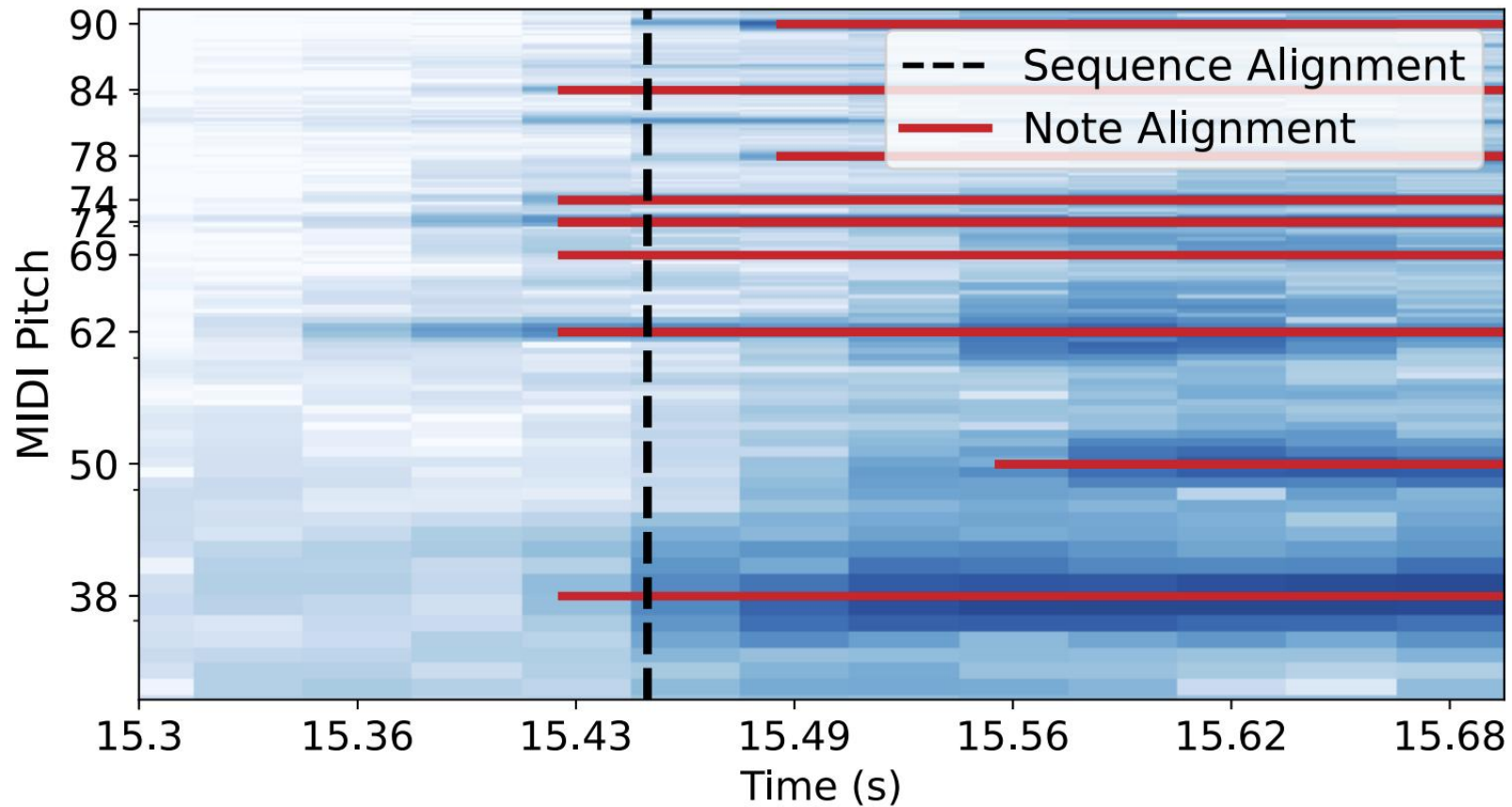
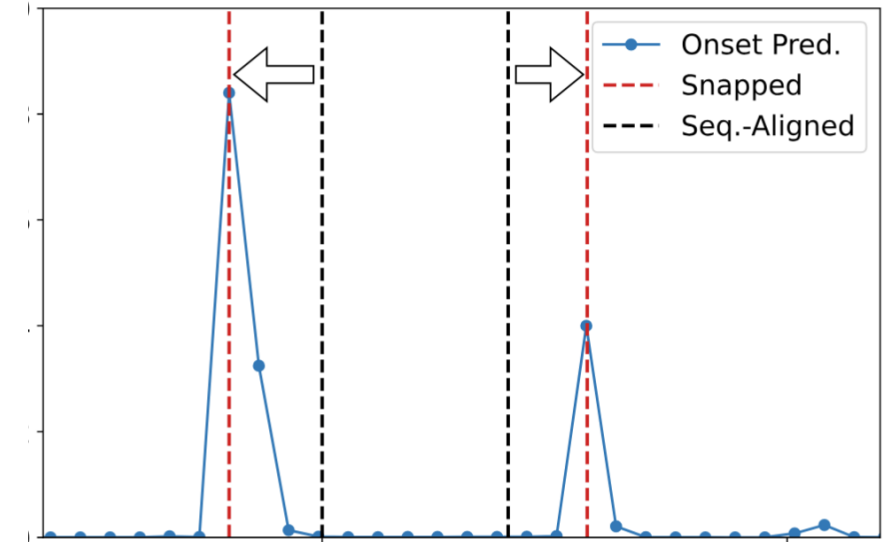


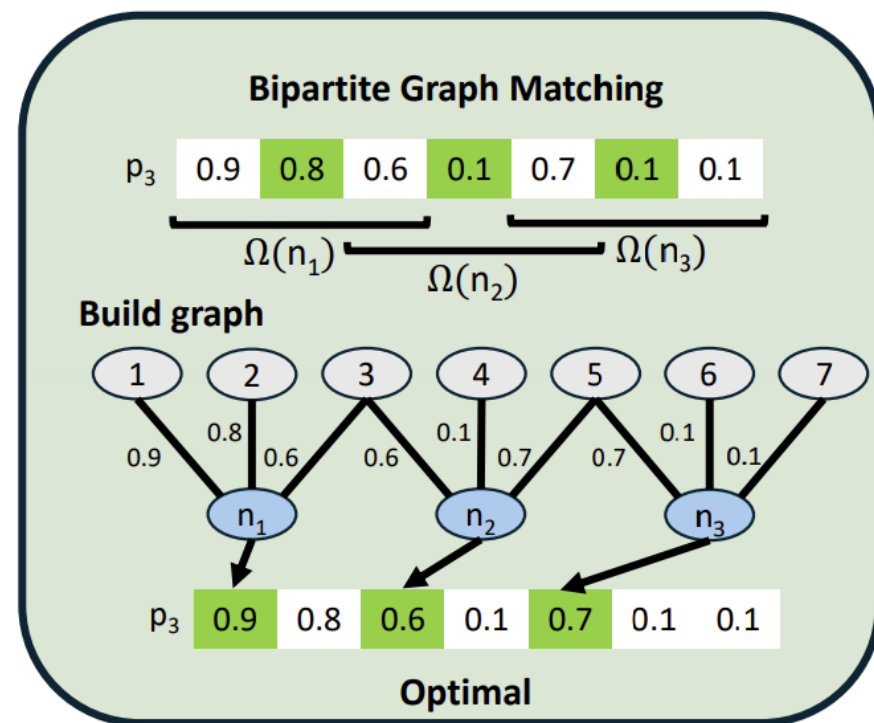
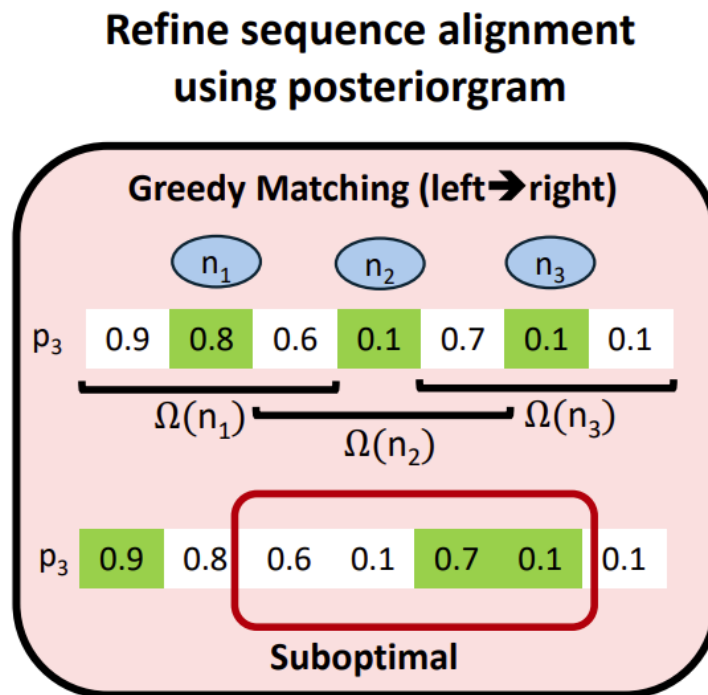
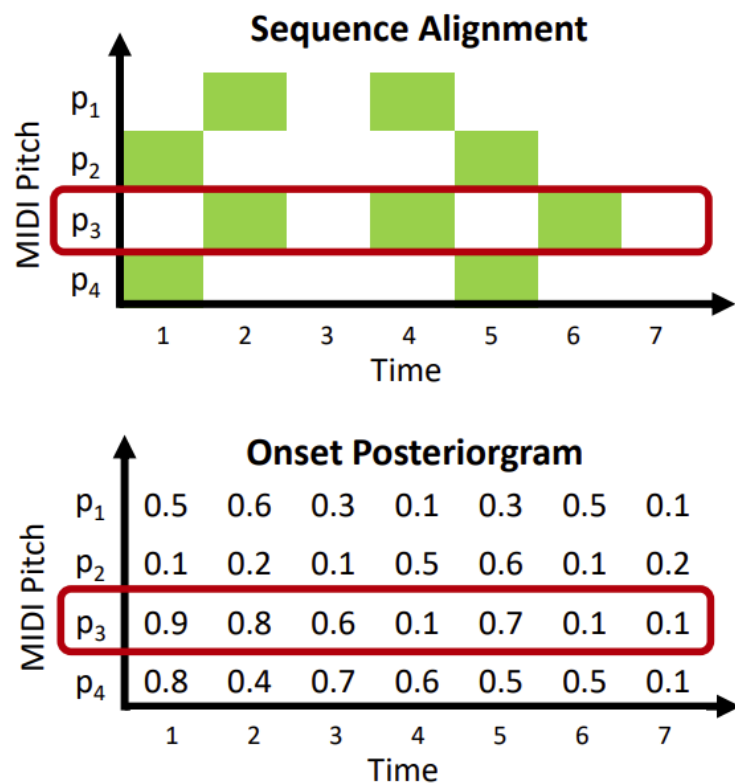
Figure 5 from Berendes et al.: Beethoven Symphony Excerpt Dataset (BSED): An Evaluation Dataset for Orchestral Music Transcription

Onset “Snapping”

- Locally adjust to timesteps with high onset activation
- Onset features from pre-trained model
- Snapping strategies
Histogram-based, greedy nearest-peak, graph-based
- DTW—improve localization



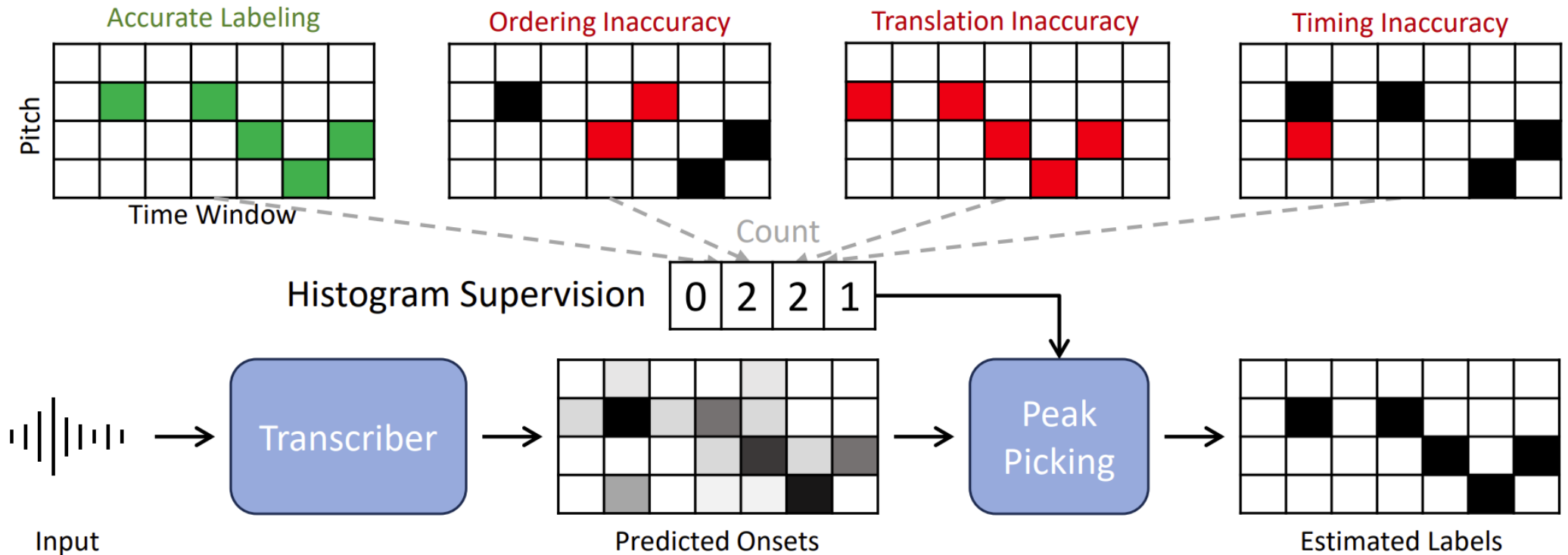
Onset “Snapping”: Greedy vs. Graph-Based



Saha et al.: Snapping Matters: Context-Aware Onset Refinement for Automatic Music Transcription, ICMC 2026

Demo: <https://abhirupsaha8.github.io>

Onset “Snapping”: Histogram-Based Top-K Selection



Yaffe et al.: Count The Notes: Histogram-Based Supervision for Automatic Music Transcription, ISMIR 2025

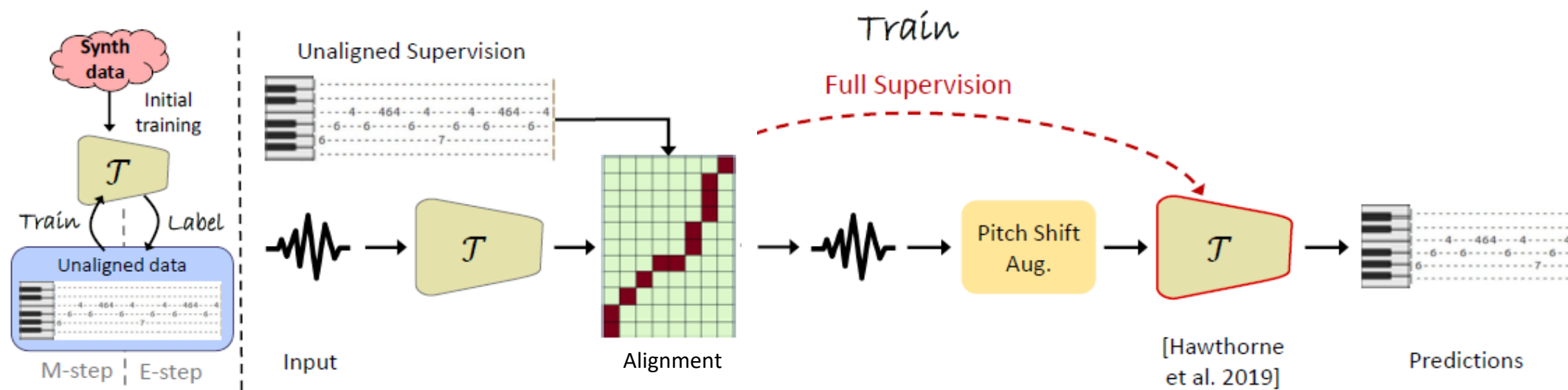
Demo: <https://yoni-yaffe.github.io/count-the-notes>

Note Duration, Note Offsets

- Remains hard problem!
- Score: only approximate duration
- Offsets hard to detect, even hard to define
- Partial solutions:
 - Real data, approximate duration from score
 - Synthetic supervised data
 - Offset snapping?

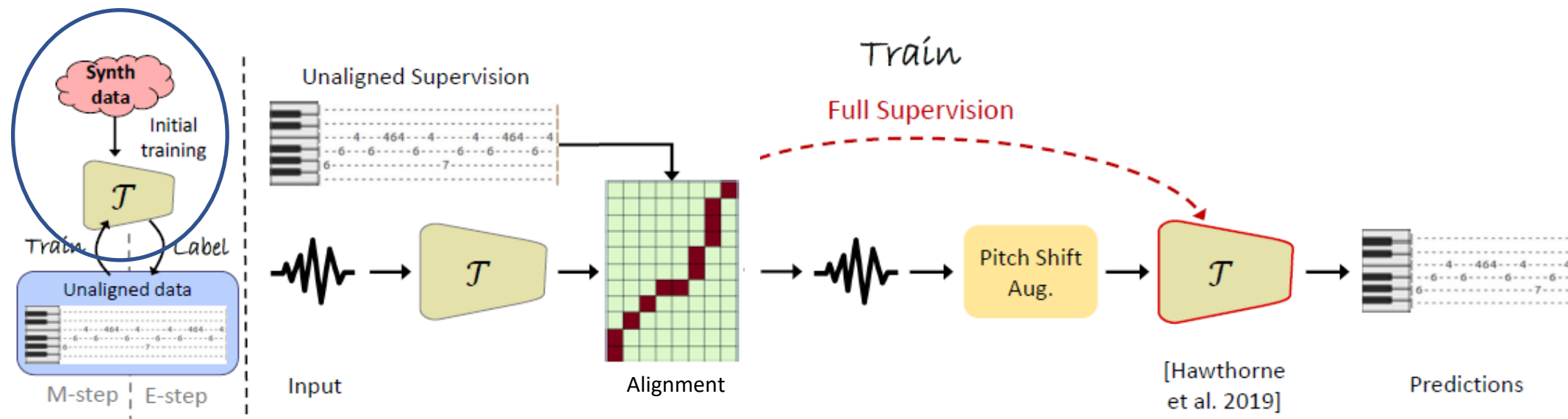
NoteEM

Pipeline

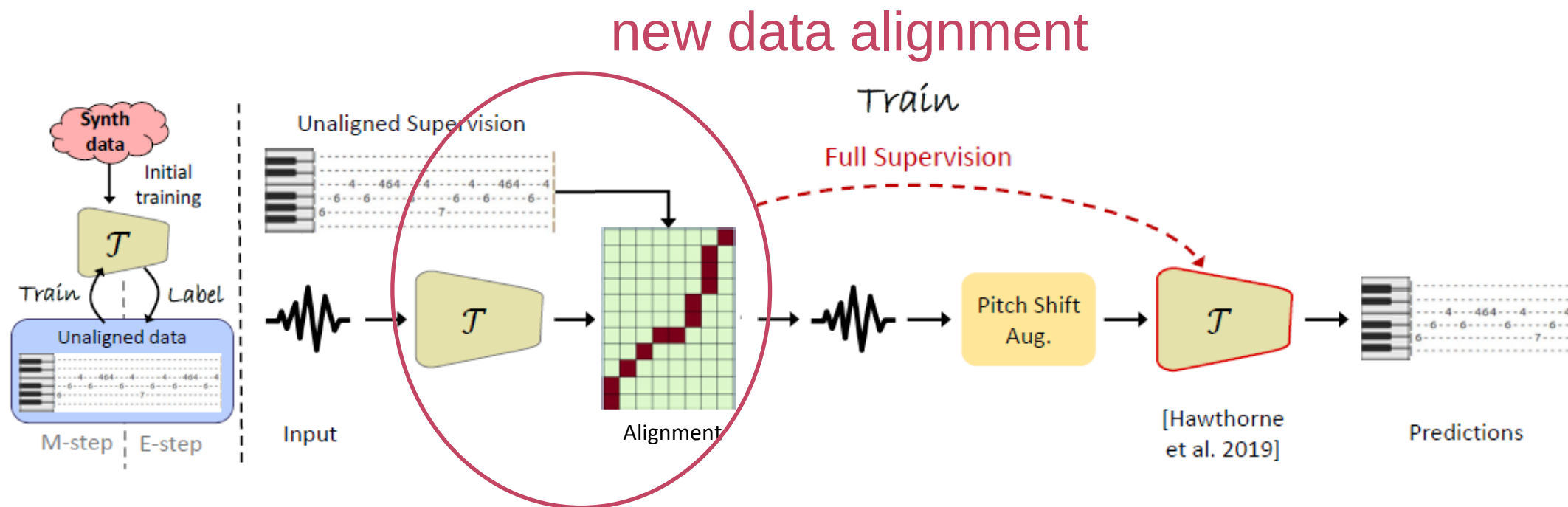


Pipeline

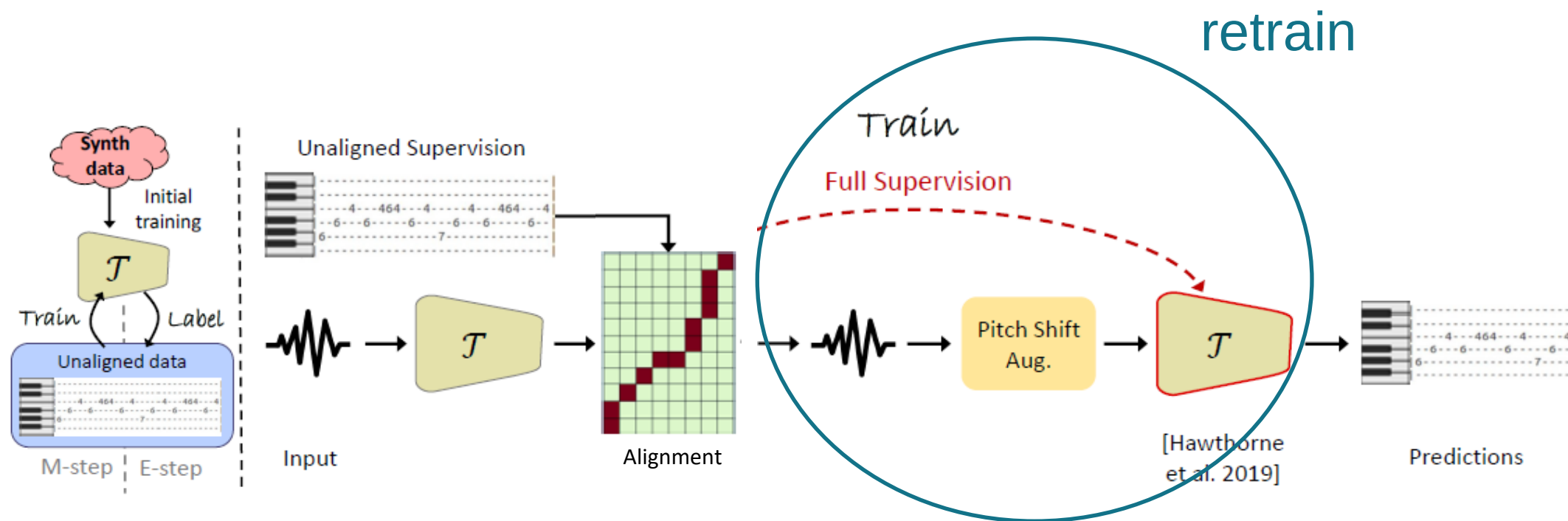
initial train



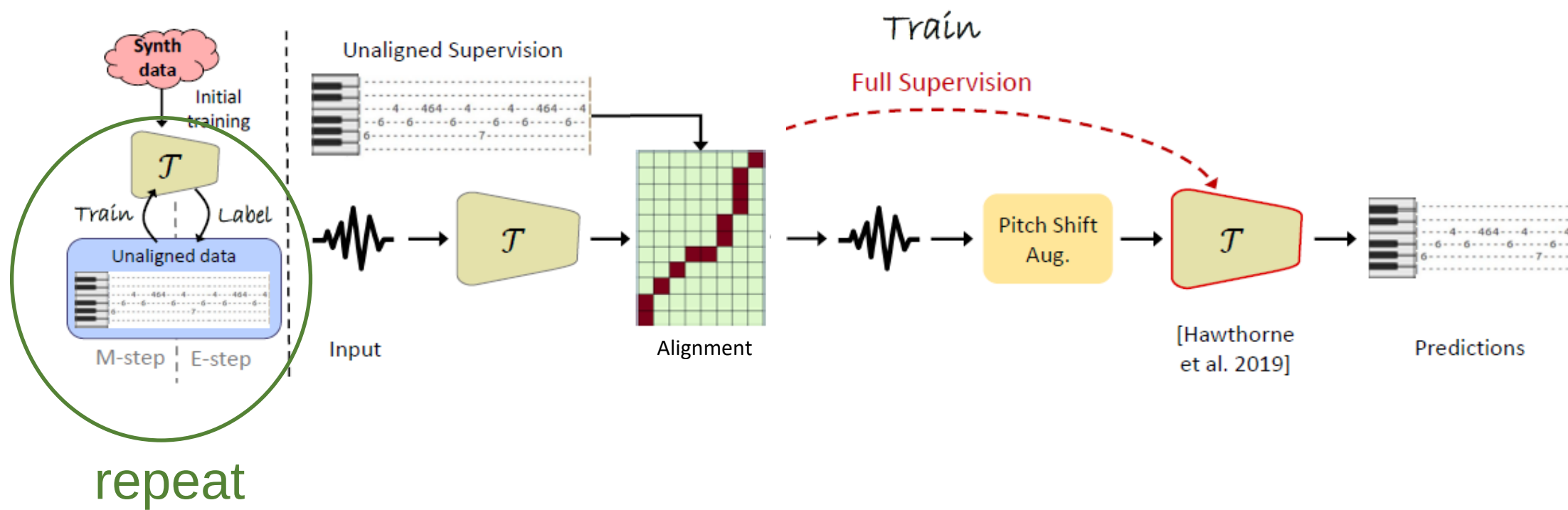
Pipeline



Pipeline



Pipeline

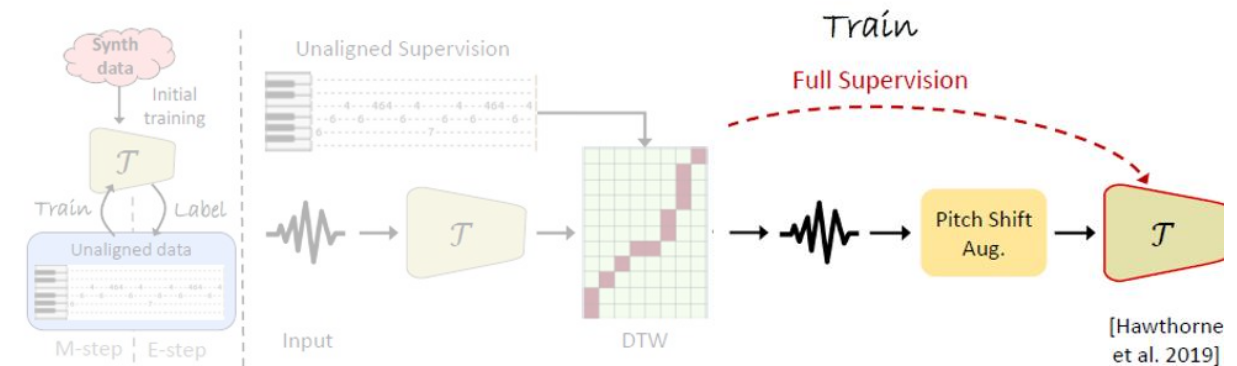
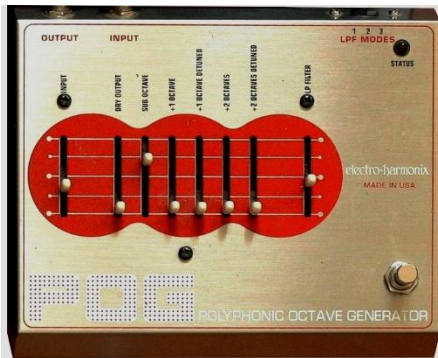


Pitch-Shift Augmentation

- 11 pitch shifted copies (semi-tones), label translated

$$s = i + \alpha_i, \quad -5 \leq i \leq 5, \quad \alpha_i \sim \mathcal{U}(-0.1, 0.1)$$

- ✓ Data augmentation
- ✓ Consistency $\rightarrow$ regularization



Metrics

- Precision: $\text{true positive} / \text{predicted positive}$
- Recall: $\text{true positive} / \text{reference positive}$
- F1-score: harmonic mean of precision & recall
- Frame-level metrics: activation over time, no individual note events
- Note-level metrics: Individual note events, need to match predicted & reference

Metrics

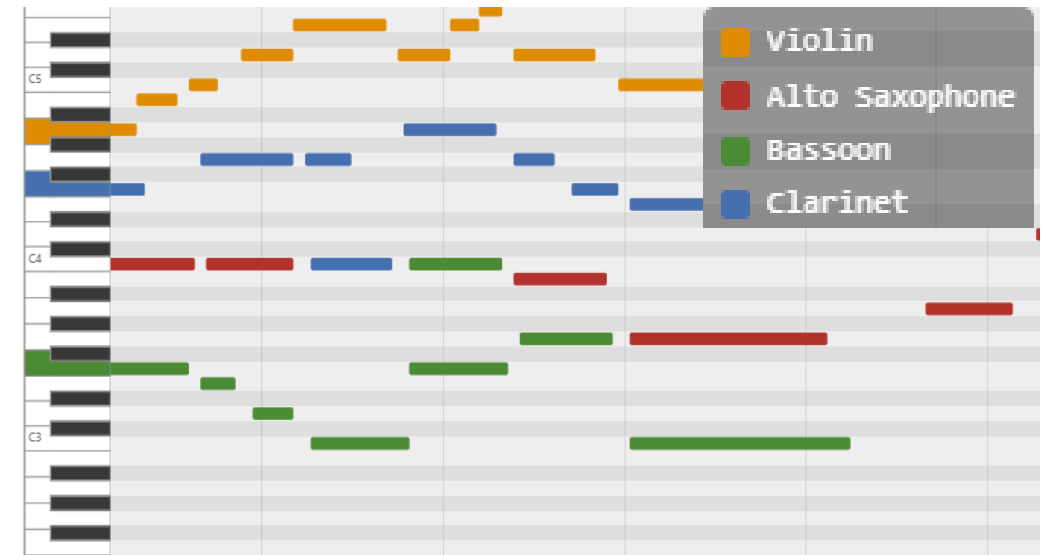
- Frame-level: Note activation / Multi-Pitch Estimation (MPE)
- Note-level: Onset within 50ms
- Note-with-offset: Offset within $\max(50\text{ms}, 20\% \text{ note length})$
- Note-with-instrument: Onset within 50ms + correct instrument
- Note-with-offset-and-velocity: Require correct velocity (note “strength”)

Multi-Instrument Transcription

Demo: <https://audiolabs-erlangen.de/resources/MIR/2026-IAT>

Instrument-Aware Transcription

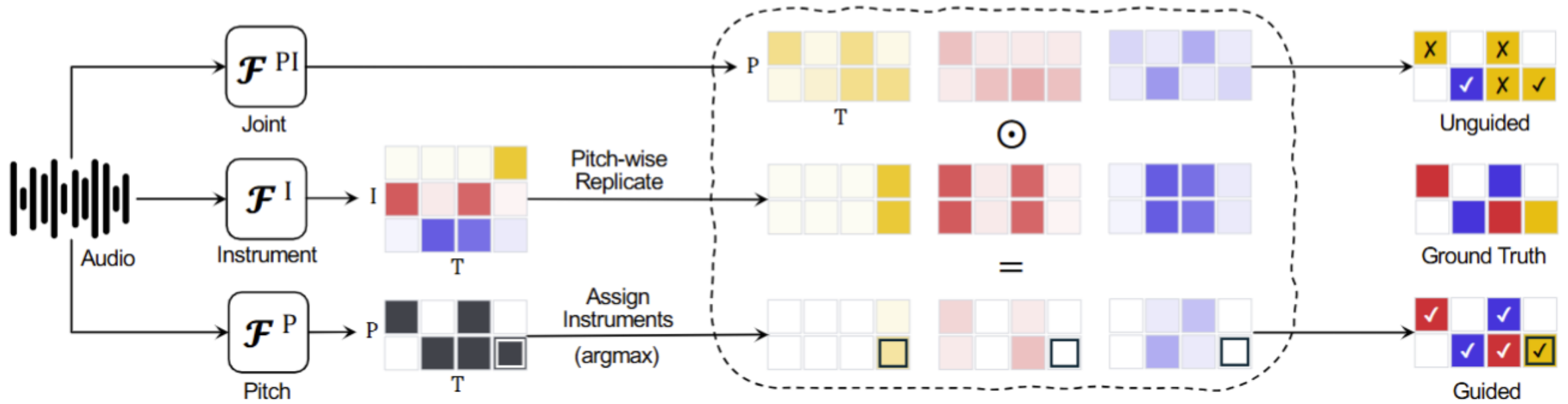
- Task: Detect pitch & instrument
- Joint modelling task → complex
- P pitches, I instruments → P x I classes!
- Disadvantage: No interclass similarity
Same pitch different instrument, same instrument different pitches



Instrument-Aware Transcription: Subtasks

- Instrument-agnostic transcription: Model $\mathcal{F}^P$ predicts $\hat{M}^P \in [0, 1]^{T \times P}$
- Instrument detection: Model $\mathcal{F}^I$ predicts $\hat{M}^I \in [0, 1]^{T \times I}$
- Instrument-aware transcription: Model $\mathcal{F}^{PI}$ predicts $\hat{M}^{PI} \in [0, 1]^{T \times P \times I}$

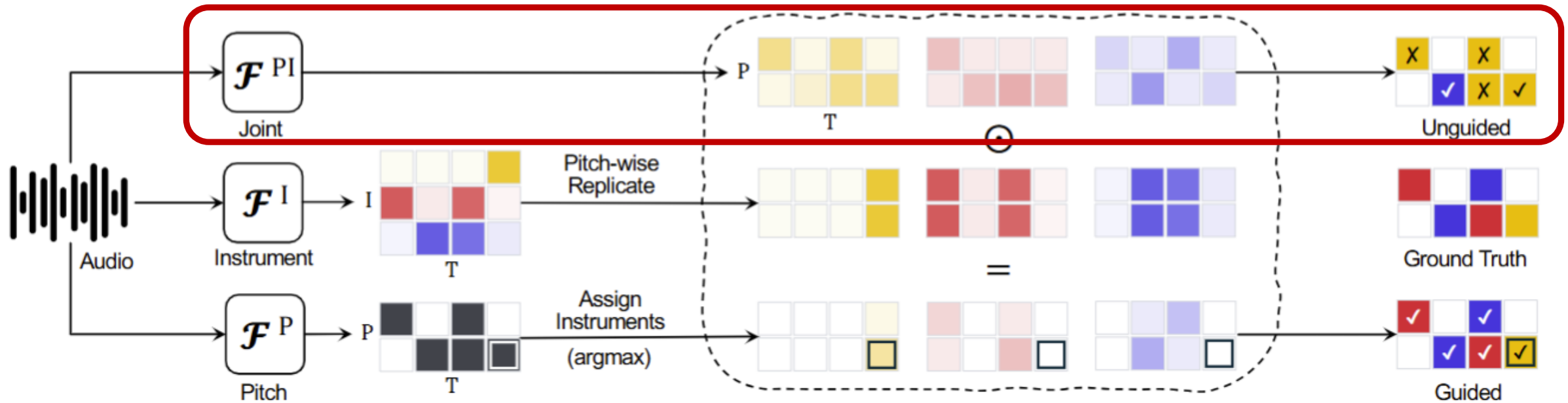
Pitch & Instrument Guidance



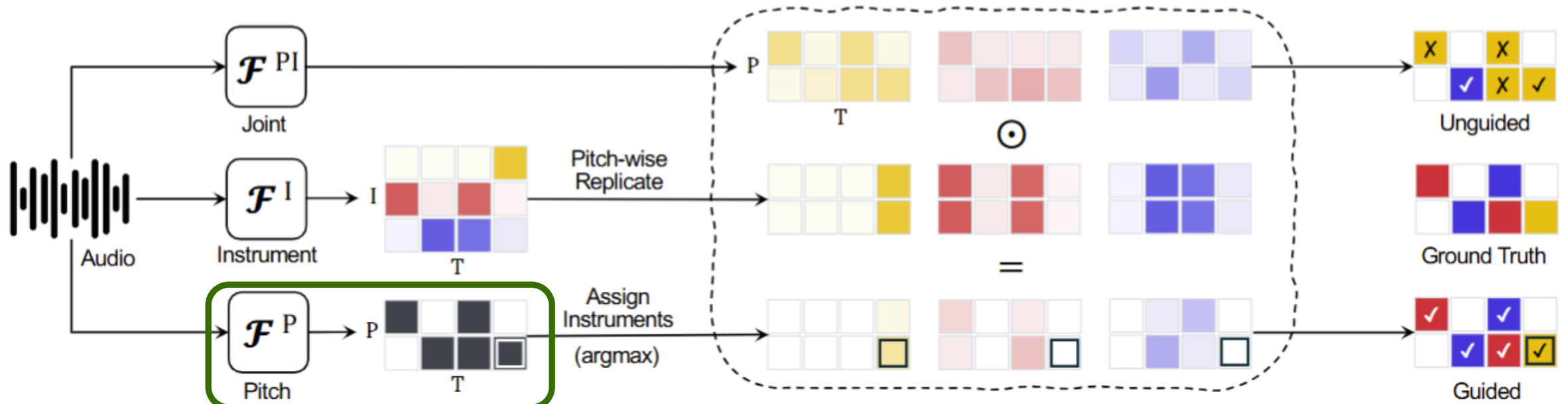
Pitch & Instrument Guidance

Joint pitch–instrument prediction (unguided)

$$\mathcal{N} = \left\{ (t, p, i) : \hat{M}_{t,p,i}^{\text{PI}} \geq \tau \right\}$$



Pitch & Instrument Guidance



Pitch-only prediction (pitch guidance)

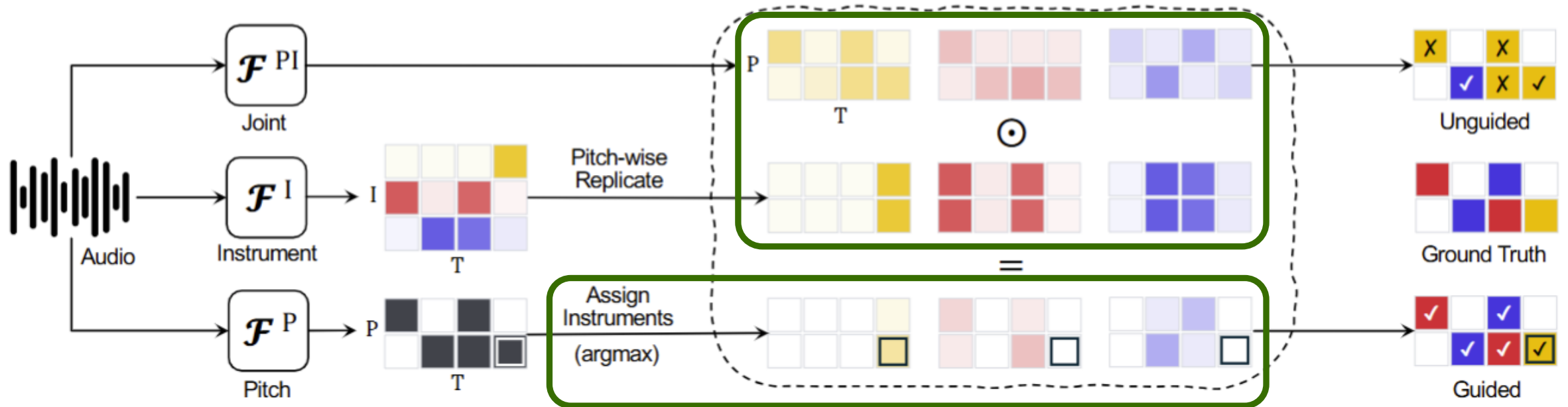
$$\mathcal{N} = \left\{ (t, p) : \hat{M}_{t,p}^P \geq \tau \right\}$$

Pitch & Instrument Guidance



Assign instrument (unguided) $i^* = \arg \max_{i \in \mathcal{I}} \hat{M}_{t,p,i}^{PI}$

Pitch & Instrument Guidance



Assign instrument (instrument guidance)
$$i^* = \arg \max_{i \in \mathcal{I}} \hat{M}_{t,p,i}^{PI} \hat{M}_{t,i}^I$$

Pitch & Instrument Guidance

- Joint pitch–instrument prediction: $\mathcal{N} = \left\{ (t, p, i) : \hat{M}_{t,p,i}^{\text{PI}} \geq \tau \right\}$
 τ : Threshold, e.g., 0.5

Pitch-guided approach:

- Step 1: pitch-only prediction $\mathcal{N} = \left\{ (t, p) : \hat{M}_{t,p}^{\text{P}} \geq \tau \right\}$
- Step 2: assign instrument $i^* = \arg \max_{i \in \mathcal{I}} \hat{M}_{t,p,i}^{\text{PI}}$

Pitch & Instrument Guidance

- Instrument guidance: mask instrument assignment using $\mathcal{F}^I$

$$i^* = \arg \max_{i \in \mathcal{I}} \hat{M}_{t,p,i}^{\text{PI}} \hat{M}_{t,i}^{\text{I}}$$

- Incorporate instrument model prior

Pitch & Instrument Guidance

- Alternative approach: mask using both $\mathcal{F}^I$ and $\mathcal{F}^P$

$$\mathcal{N} = \left\{ (t, p, i) : \hat{M}_{t,p}^P \hat{M}_{t,i}^I \hat{M}_{t,p,i}^{PI} \geq \tau \right\}$$

- In practice, pitch guidance followed by instrument guidance works better
- Pitch—strong & robust cue, generalizes well across environments

Further Reading

- Ewert et al., “High Resolution Audio Synchronization using Chroma Onset Features”, ICASSP 2009
- Benetos et al., “Automatic Music Transcription: An Overview”, IEEE Signal Processing Magazine, 2018
- Hawthorne et al., “Onsets and Frames: Dual-objective Piano Transcription”, ISMIR 2018
- Hawthorne et al., “Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset”, ICLR 2019
- Maman and Bermano, “Unaligned Supervision for Automatic Music Transcription in The Wild”, ICML 2022

Thank You!

Figure Sources

Slide 4: https://commons.wikimedia.org/wiki/File:Spectrogram_of_Bach%27s_Chorales_for_Organ.jpg,
https://commons.wikimedia.org/w/index.php?title=File%3AMoonlight_sanata.pdf&page=5

Slide 6: <https://commons.wikimedia.org/wiki/File:TalkNet.png>

Slide 9: https://commons.wikimedia.org/wiki/File:Piano_88_keyboard_template.svg

Slide 16: https://commons.wikimedia.org/wiki/File:Harmonicpartials_on_strings.svg

Slide 17: Li et al.: End-to-End Mandarin Speech Reconstruction Based on Ultrasound Tongue Images Using Deep Learning, TNSRE 2024

Slide 18: https://commons.wikimedia.org/wiki/File:Bob_Moog3.jpg, Figure 2 from Berndt et al.: ChoraleWind: An Expressive Wind-Quartet Dataset for End-to-End Rendering from the Neues Thüringer Choralbuch, ICMC 2026

Slide 26: https://commons.wikimedia.org/wiki/File:Neural_network.svg

Slide 29: https://commons.wikimedia.org/wiki/File:Optimizer_Animations.gif

Slide 32: <https://commons.wikimedia.org/wiki/File:DisklavierView2.jpg>

Slide 35: <https://commons.wikimedia.org/wiki/File:Aleppomusic.jpg>, <https://commons.wikimedia.org/wiki/File:Cordesteatre.jpg>

Slide 36: https://commons.wikimedia.org/wiki/File:Image_segmentation.png

Slide 37: [https://commons.wikimedia.org/wiki/File:Example_of_subtitles_\(Charade,_1963\).jpg](https://commons.wikimedia.org/wiki/File:Example_of_subtitles_(Charade,_1963).jpg)

Slide 45: Kevin Knight, Statistical Machine Translation lecture slides, USC / Information Sciences Institute

Slide 46: Ferreira et al.: Using pseudo-labeling to improve performance of deep neural networks for animal identification, Scientific Reports 2023,

Slide 47: Charco et al.: Camera pose estimation in multi-view environments: From virtual scenarios to the real world, Image and Vision Computing, 2021

Slide 48: <https://commons.wikimedia.org/wiki/File:Speech-to-text.svg>, Erbey and Barışçı : Lip-Reading Classification of Turkish Digits Using Ensemble Learning Architecture Based on 3DCNN, Applied Sciences 2025

Slide 50: https://commons.wikimedia.org/wiki/File:Two_repetitions_of_a_walking_sequence_of_an_individual_recorded_using_a_motion-capture_system.gif,
https://commons.wikimedia.org/wiki/File:Euclidean_vs_DTW.jpg

Slide 51: <https://commons.wikimedia.org/wiki/File:ChromaFeatureCmajorScaleScoreAudioColor.png>

Slide 65: https://commons.wikimedia.org/wiki/File:EH_POG-2.jpg