

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

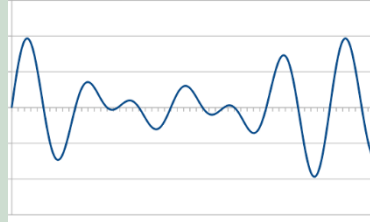
Agenda

1. Introduction to music information retrieval (MIR)
2. Data-driven methods for automatic music transcription
3. Score–audio alignment
4. Multi-instrument transcription

Introduction to Music Information Retrieval (MIR)

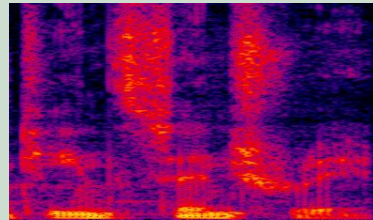
Music Representations

Waveform



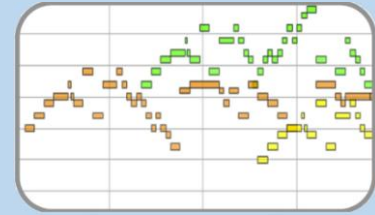
Spectrograms

STFT / Mel /
CQT



Continuous Signal

**MIDI /
Piano Roll**



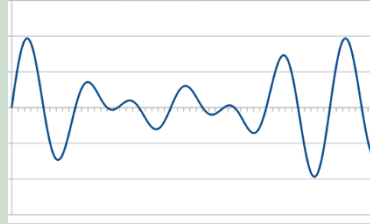
Sheet Music



Symbolic

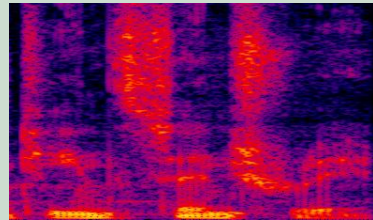
Music Representations

Waveform



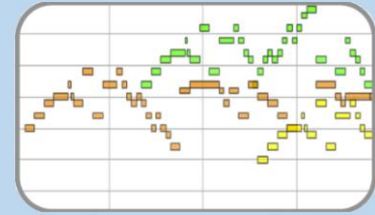
Spectrograms

STFT / Mel /
CQT



Continuous Signal

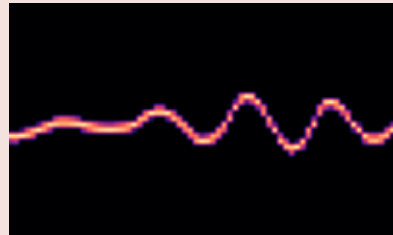
**MIDI /
Piano Roll**



Sheet Music

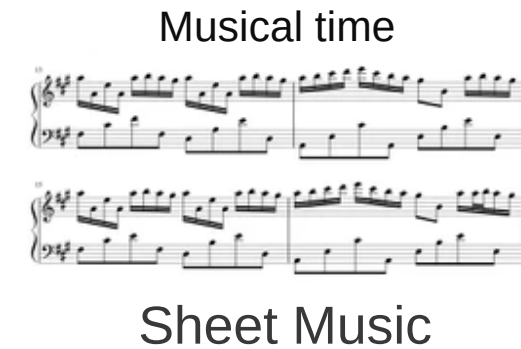
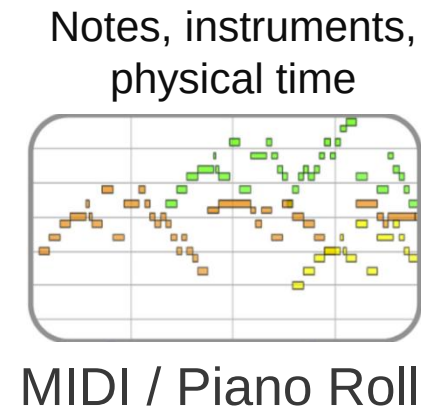
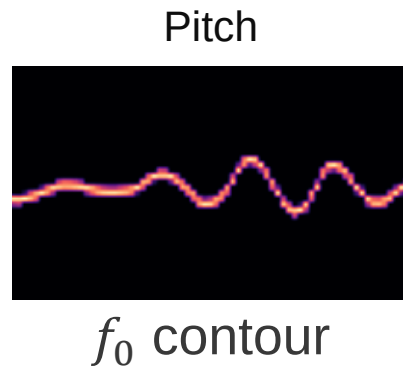
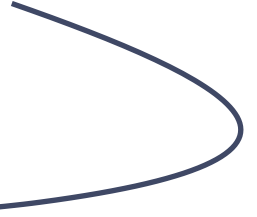
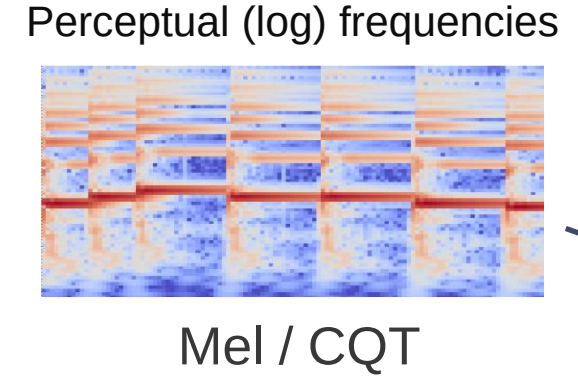
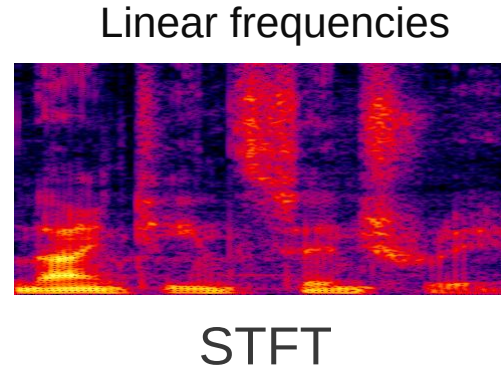
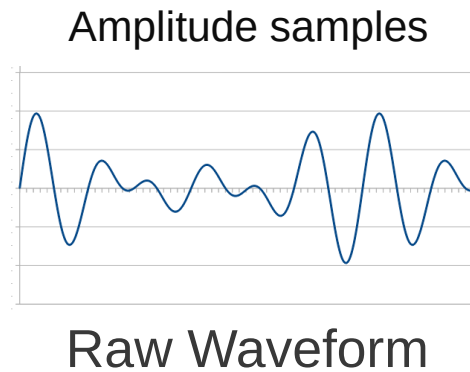


f_0 Contour



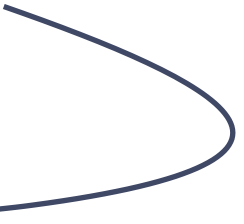
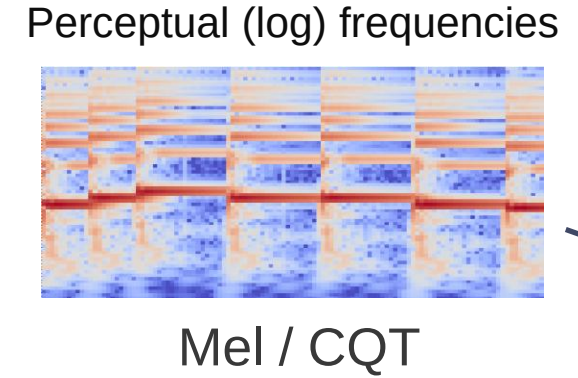
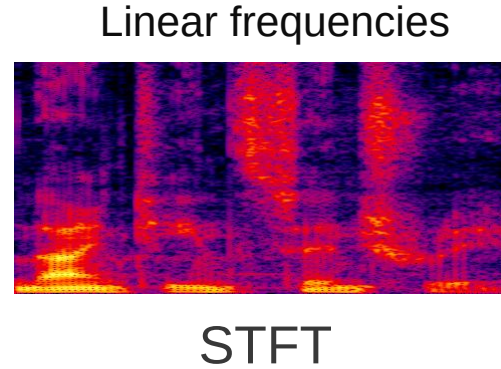
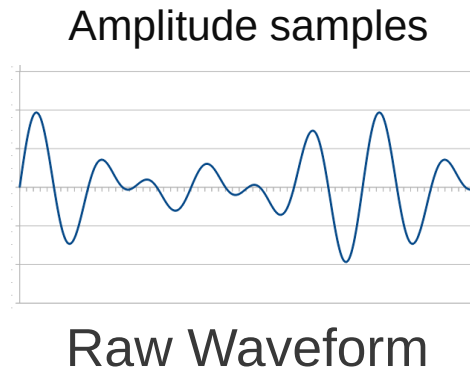
Symbolic

Levels of Abstraction: Continuous → Semantic

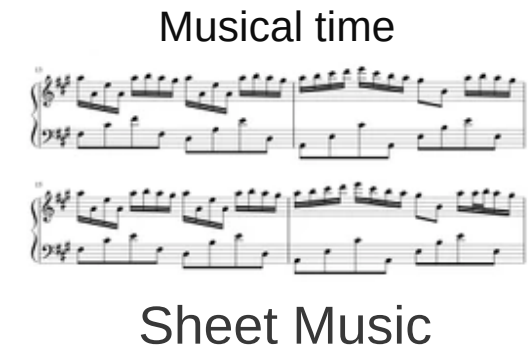
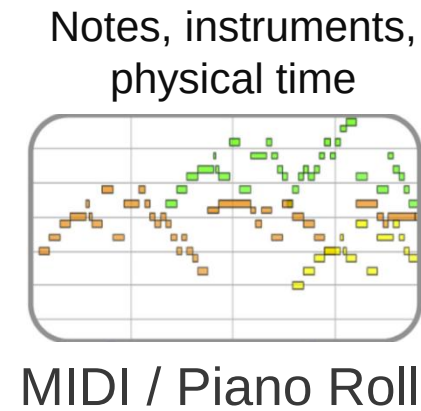
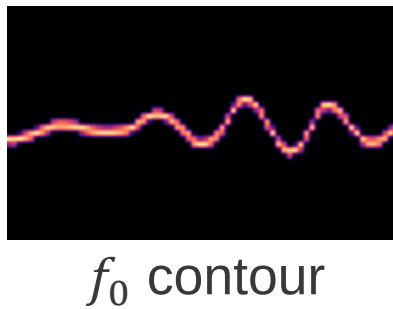


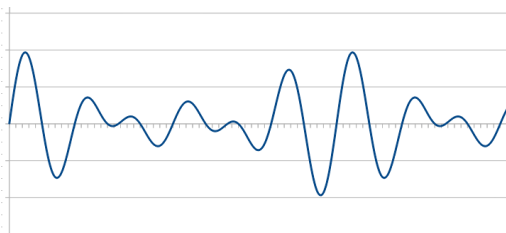
Levels of Abstraction: Continuous → Semantic

Not strict hierarchy!

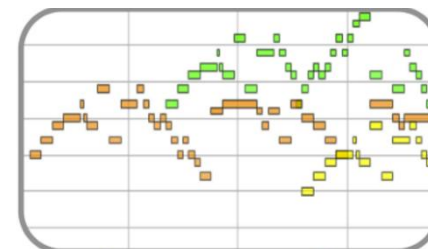


Pitch

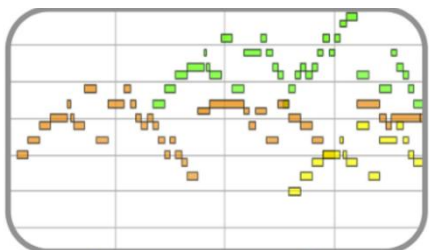




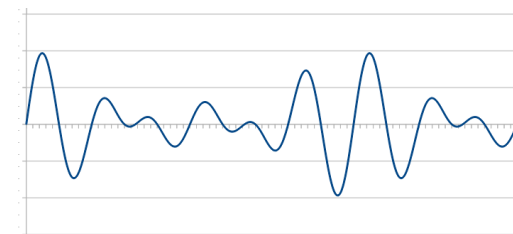
Transcription

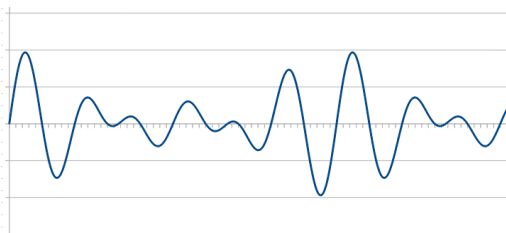


NoteEM

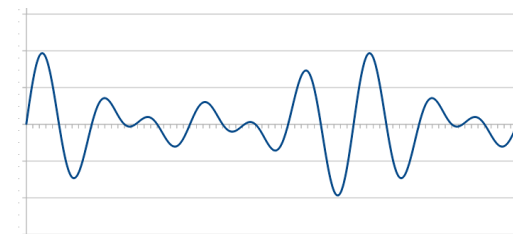


Synthesis





Conversion



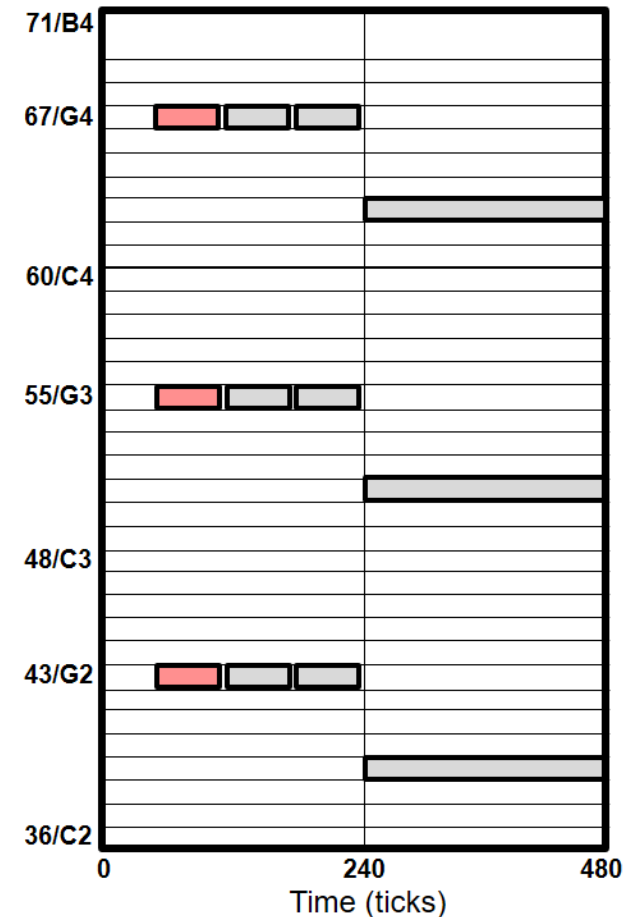
Change pitch, singer, timbre...

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



STFT: Perceptual?

- Human perception—logarithmic

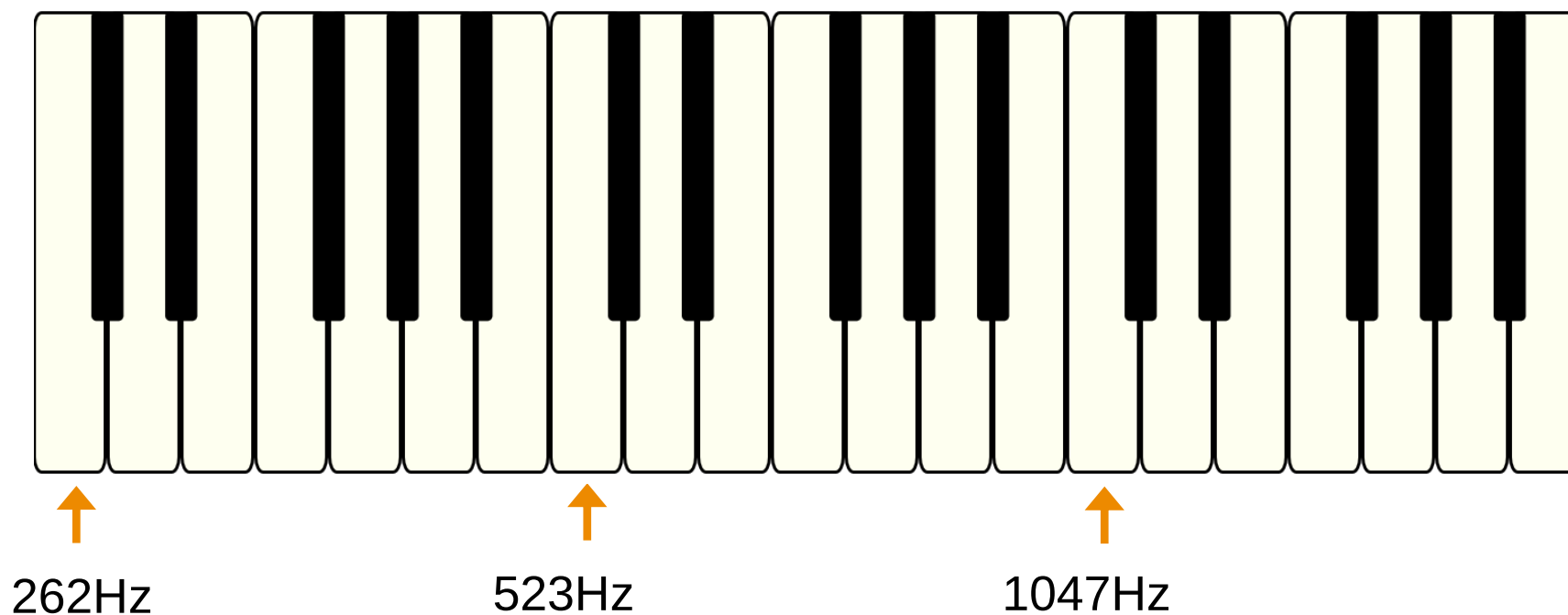
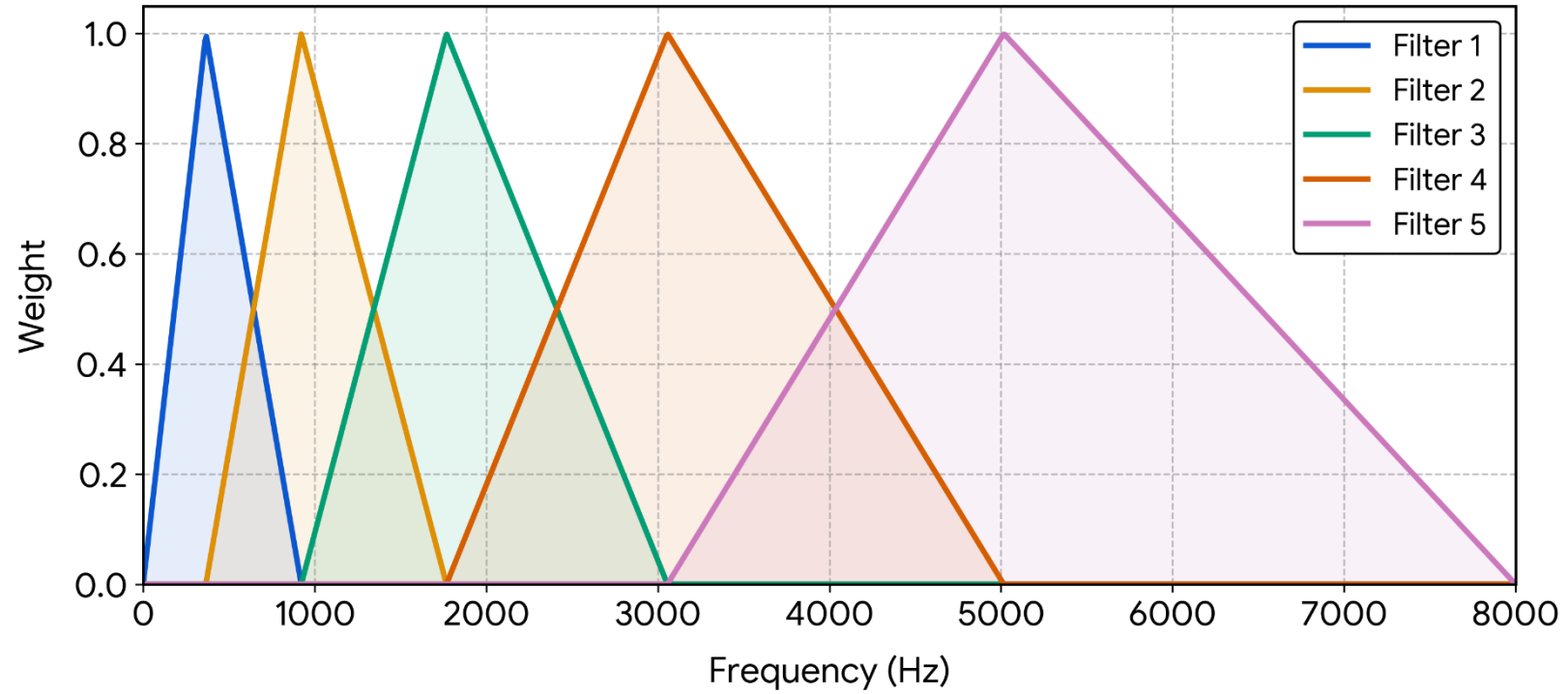


Figure from https://commons.wikimedia.org/wiki/File:Piano_88_keyboard_template.svg

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:

$$f = 2^{(m-69)/12}$$

$$m = 12 \log_2(f/440) + 69$$

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

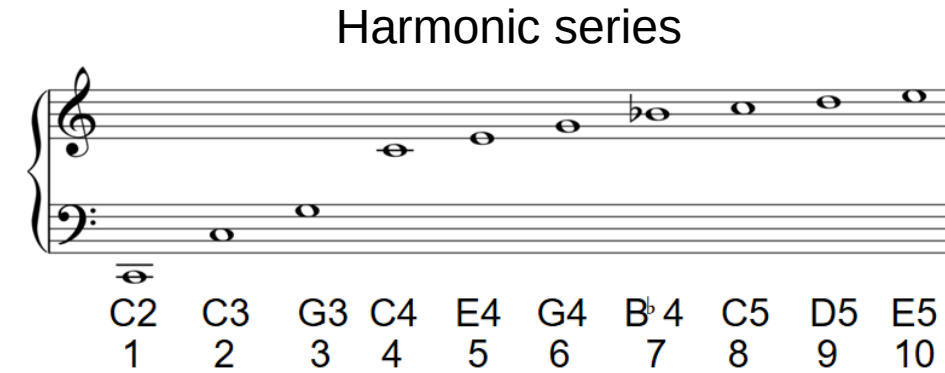


Figure 1.20 from [Meinard Müller, FMP, Springer 2015]

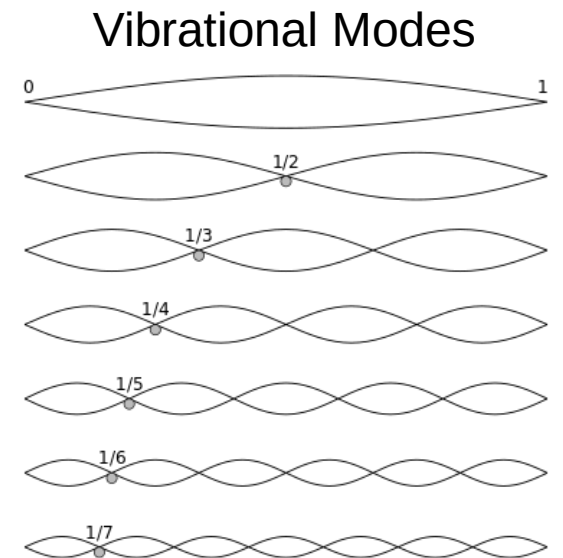
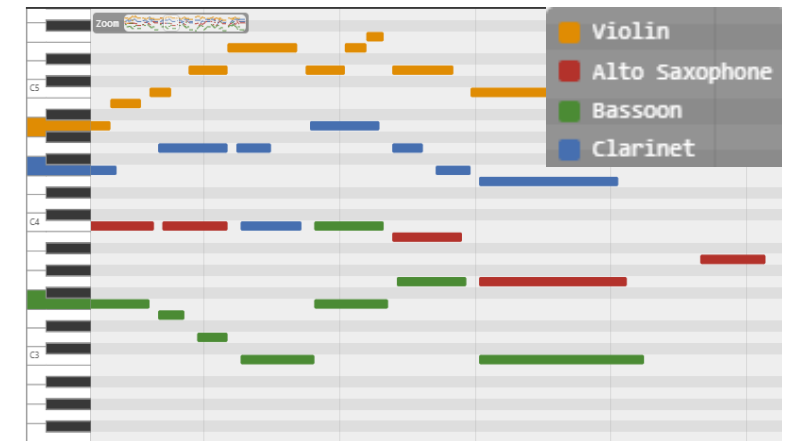
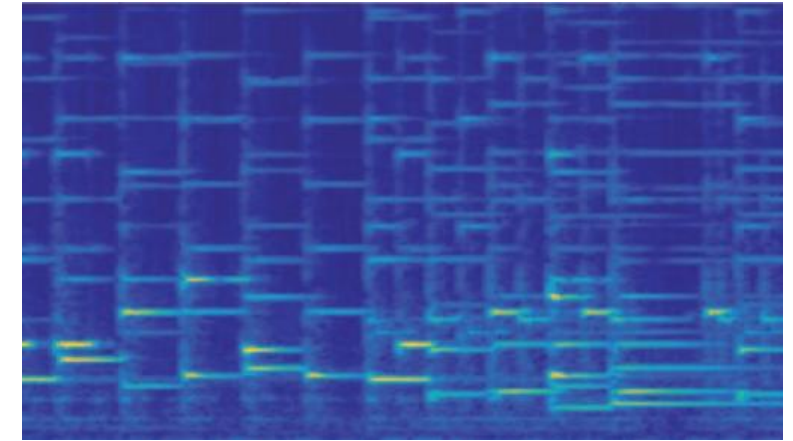


Figure from: commons.wikimedia.org/wiki/File:Harmonic_partials_on_strings.svg

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—intensity & pitch fluctuation...

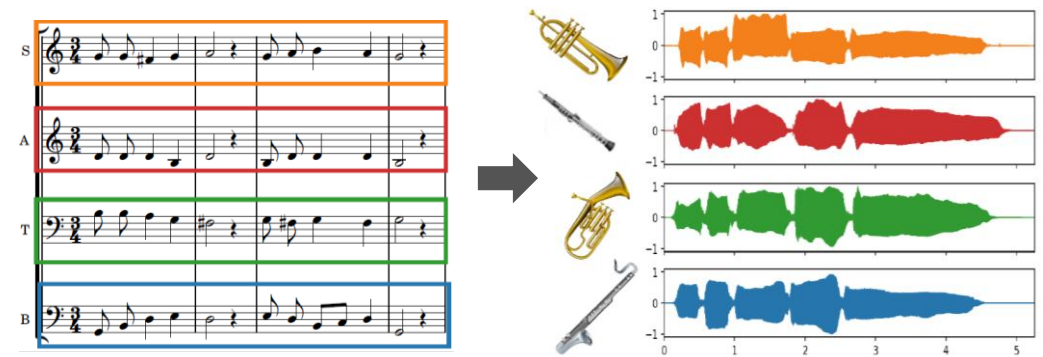


Figure 2 from Berndt et al.: ChoraleWind: An Expressive Wind-Quartet Dataset for End-to-End Rendering from the Neues Thüringer Choralbuch, ICMC 2026

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



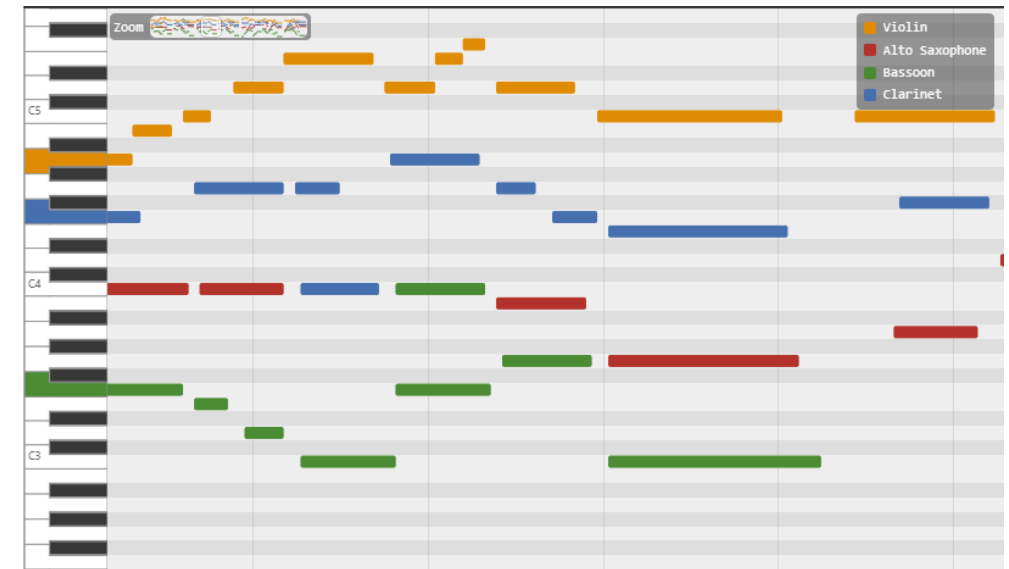
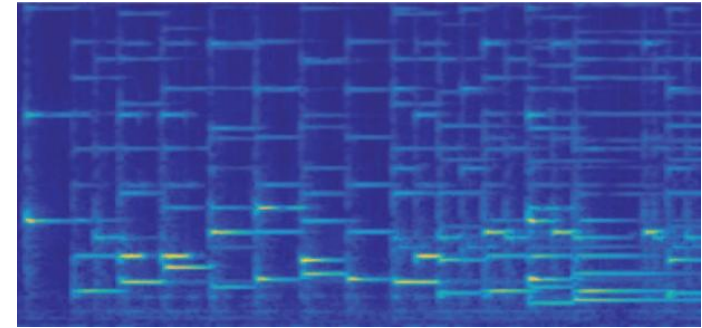
Automatic Music Transcription

Automatic Music Transcription: What Is It Good For?

- Music education
- Musicology
- Control signals powering synthesis models

Automatic Music Transcription

- Input: Audio (spectrogram)
- Output: Note events (MIDI / Piano Roll)
- Challenges:
 - Separate fundamental & partial
 - Timing—note begin / end
 - Polyphony, multi-instrument
 - Instrument detection
 - Dynamics (velocity)



Automatic Music Transcription

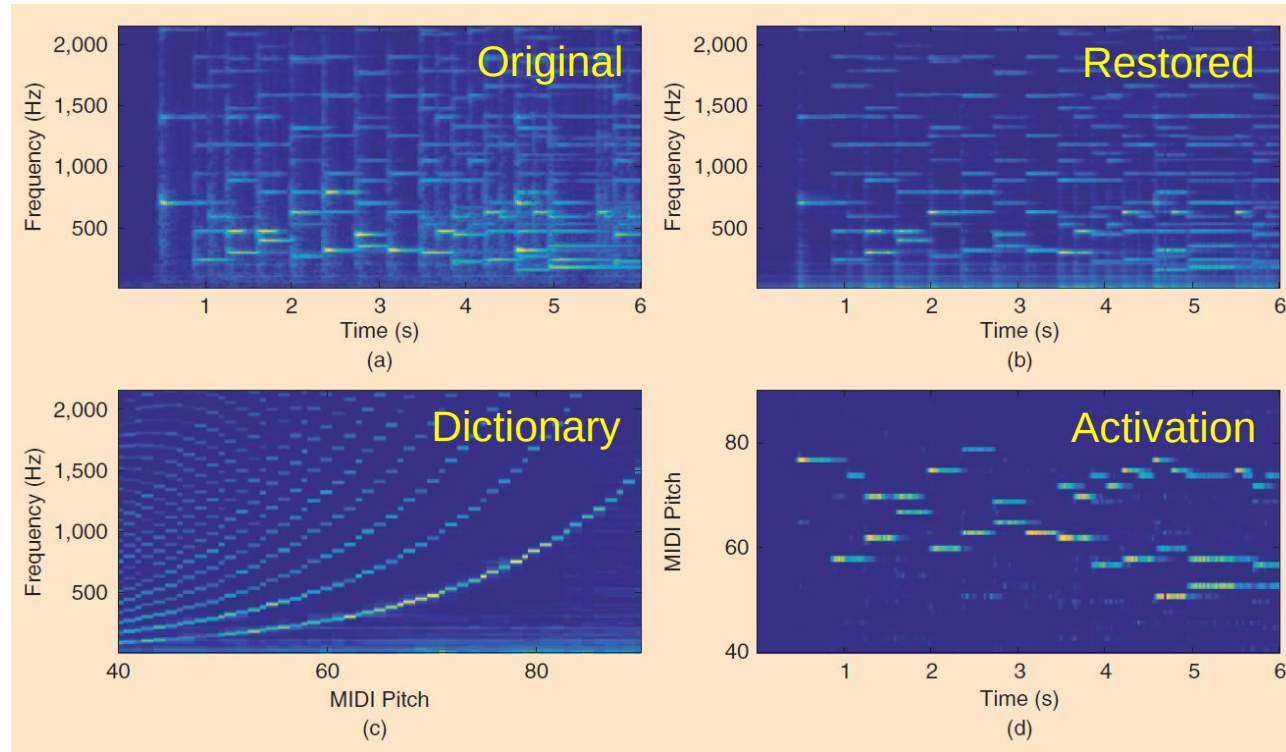
- Single-instrument: e.g., piano
- Instrument-agnostic: detect pitch only within multi-instrument setting
- Instrument-aware: detect pitch & instrument

Data-Driven Methods for Automatic Music Transcription

NMF (Non-Negative Matrix Factorization)

Benetos et al., Automatic Music Transcription: An Overview, IEEE Signal Processing Magazine, 2019

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

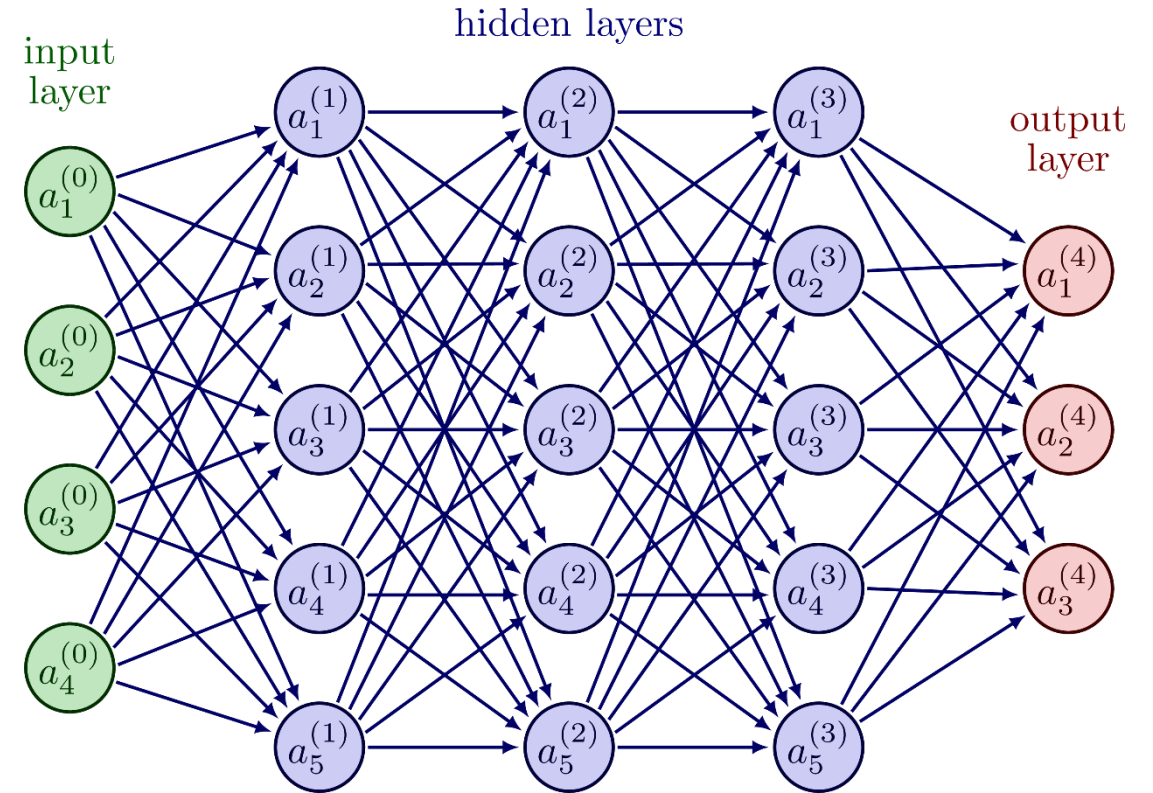


Transcription with Neural Networks

Neural Network

Composition of:

- Linear maps
- Non-linear activation functions
- Linear maps ➔ **weights**

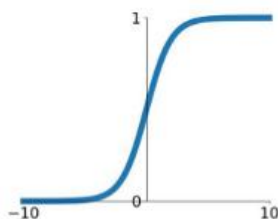


Activation Functions

- Make network non-linear → Expressive

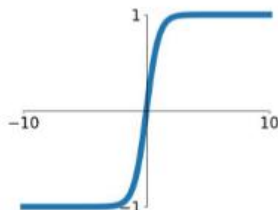
Sigmoid

$$\sigma(x) = \frac{1}{1+e^{-x}}$$



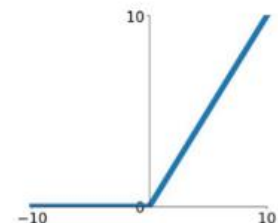
tanh

$$\tanh(x)$$



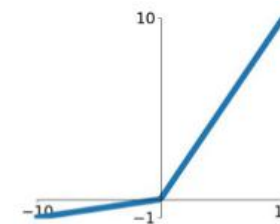
ReLU

$$\max(0, x)$$



Leaky ReLU

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

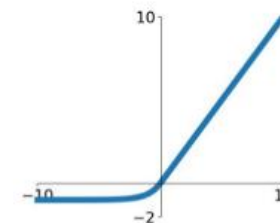


Maxout

$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Loss Functions

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

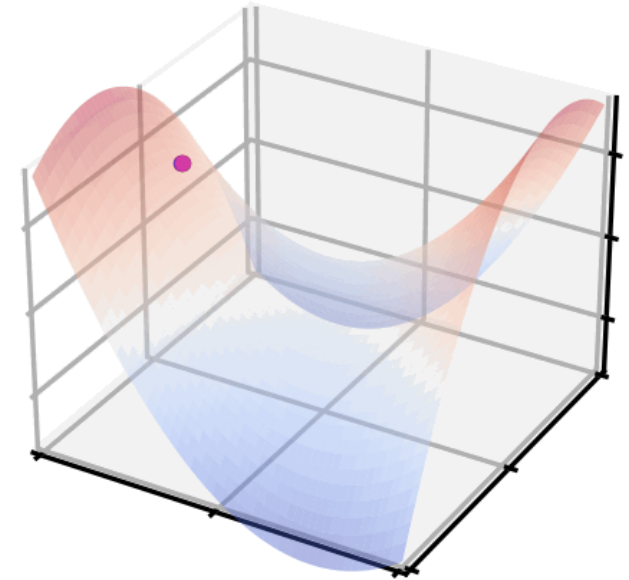
$$H_p(q) = -\frac{1}{N} \sum_{i=1}^N y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$$

- 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!

Classification Types

- 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 → multiple binary classifications
 - Loss: Sum of multiple Binary Cross Entropy

Piano Datasets: MAESTRO, MAPS

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

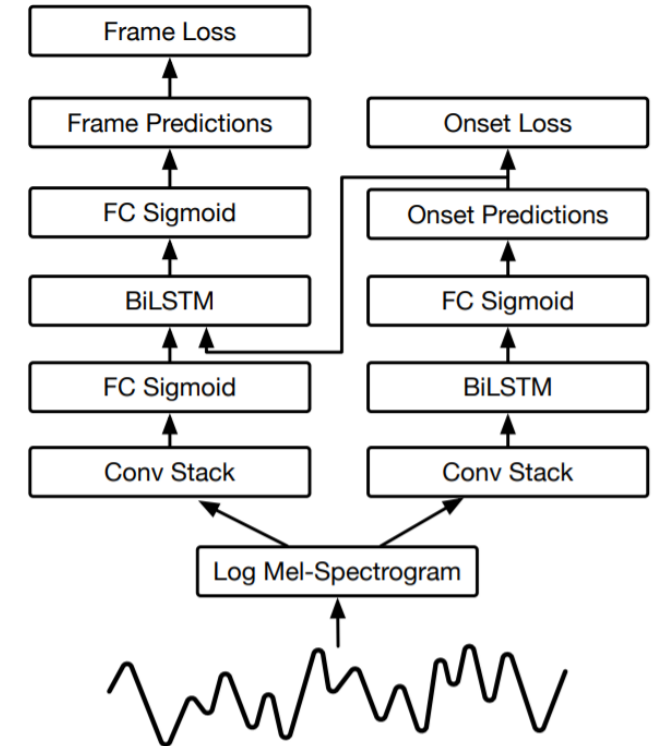
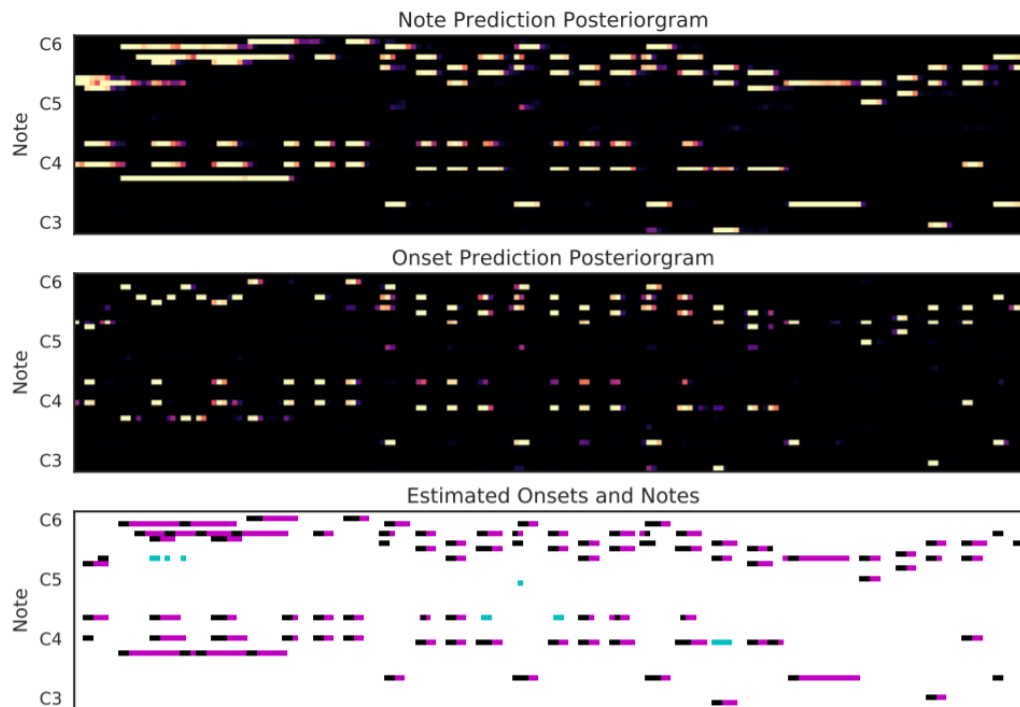
- Fully supervised datasets
- Audio & MIDI from Disklavier
- Nearly perfect alignment
- MAPS—train synth, test real.
- MAESTRO—130 hours real



Onsets and Frames

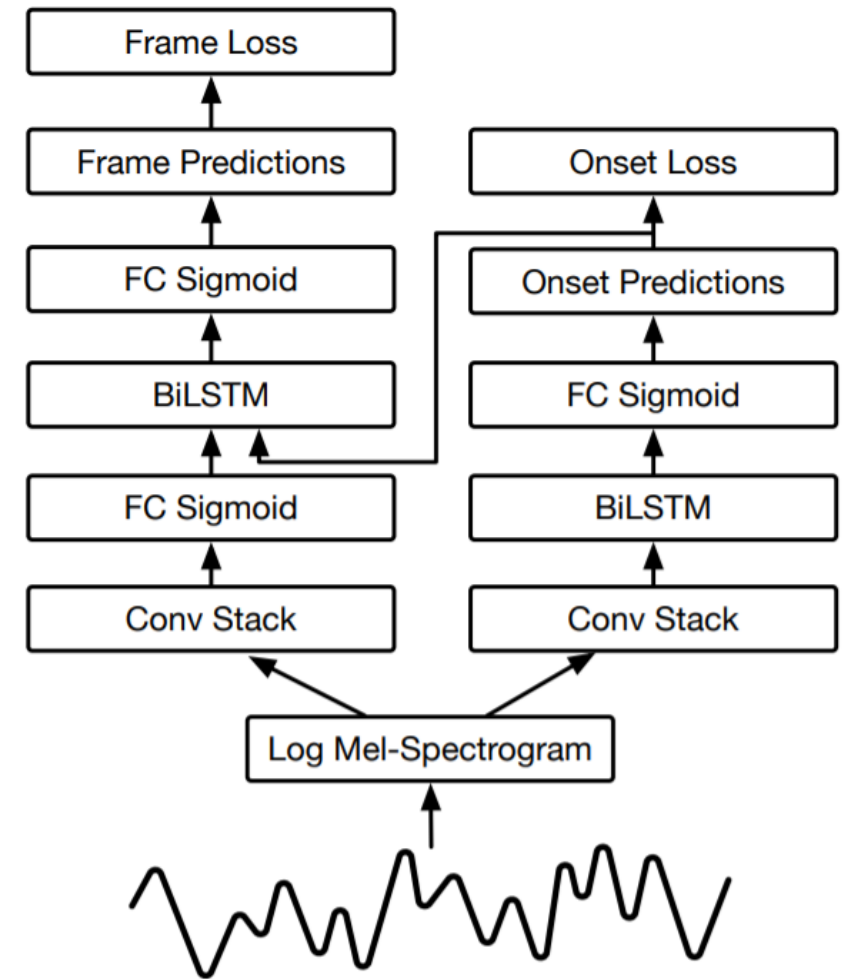
Hawthorne et al., Onsets and Frames: Dual-Objective Piano Transcription, ISMIR 2017

- 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



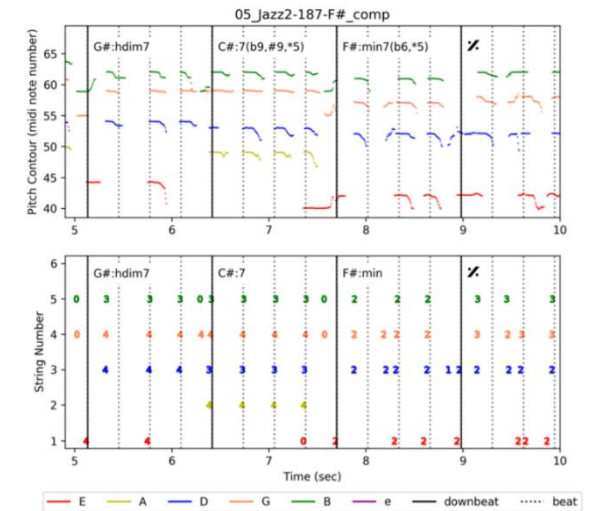
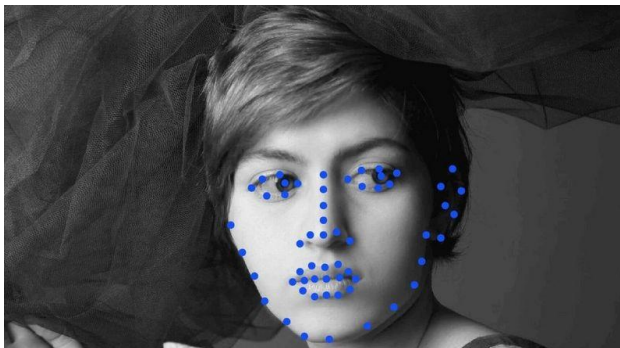
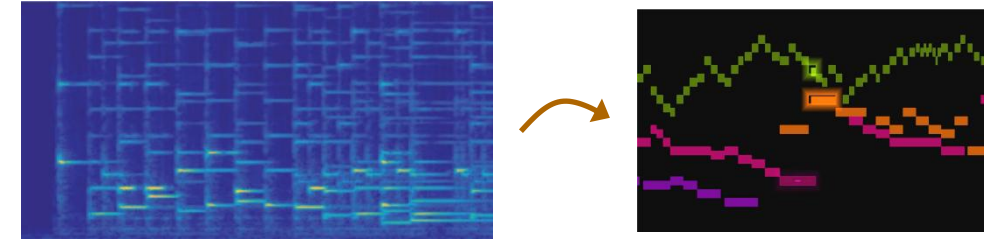
Other Instruments / Genres? Multi-Instrument?

- Only poorly supervised data



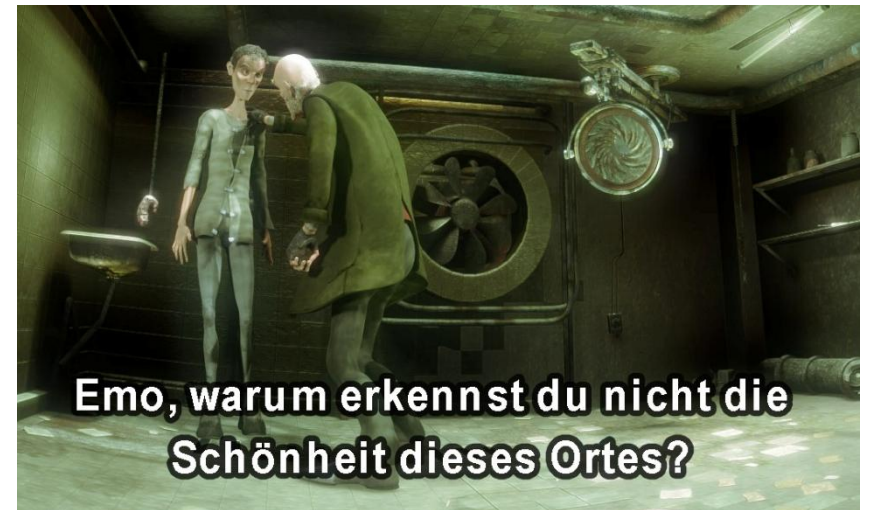
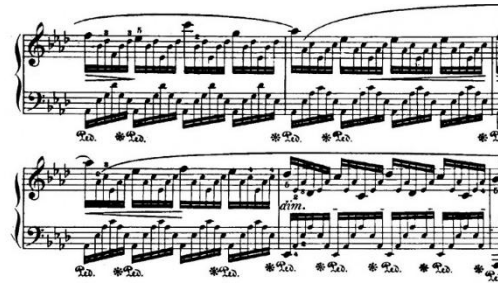
Data Problem

- 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**



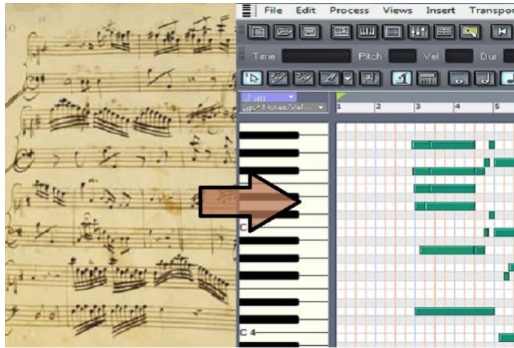
Unaligned Supervision

- Create accurate supervision for easily attainable data



Real-world
recording

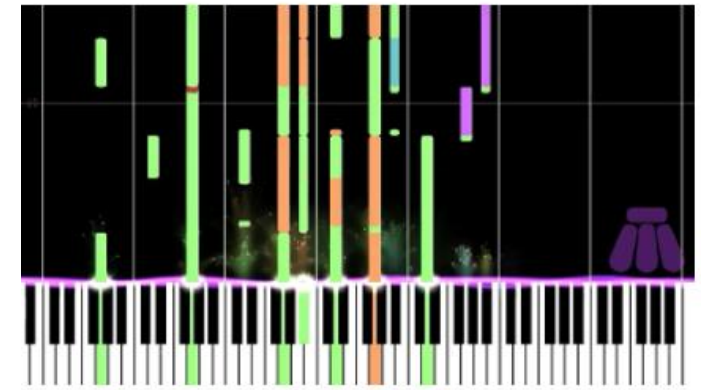
+



Unaligned
digitized
score



Transcriber



Unaligned Supervision



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 ...



Ludwig van Beethoven 5th Symphony By Herbert Von Karajan
304K views • 2 years ago
Champ

Quinta sinfonía de Beethoven dirigida por Herbert von Karajan.



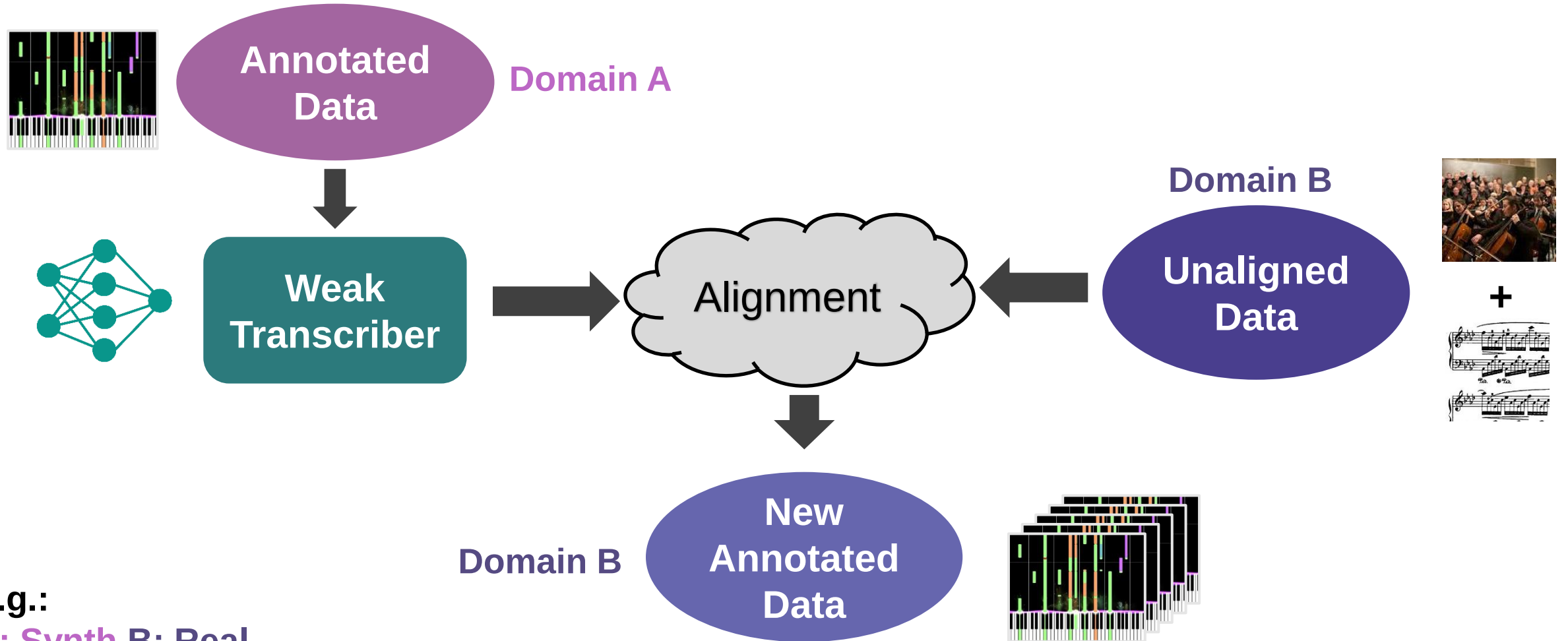
Beethoven - Symphony No. 5 - Iván Fischer | Concertgebouworkest
323K views • 2 years ago
Concertgebouworkest

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



Beethoven Symphony 5.mid

Unaligned Supervision



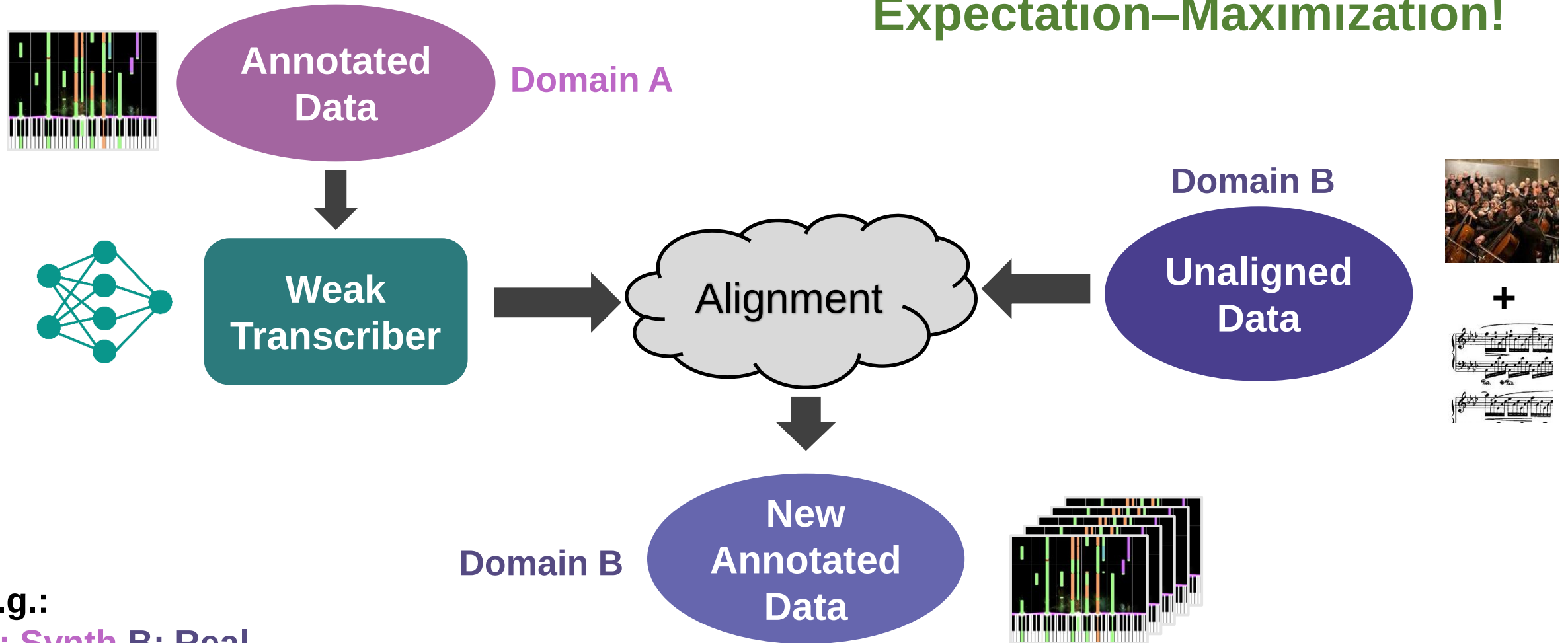
E.g.:

A: Synth B: Real

A: Piano B: Guitar

Unaligned Supervision

Expectation–Maximization!



E.g.:

A: Synth B: Real

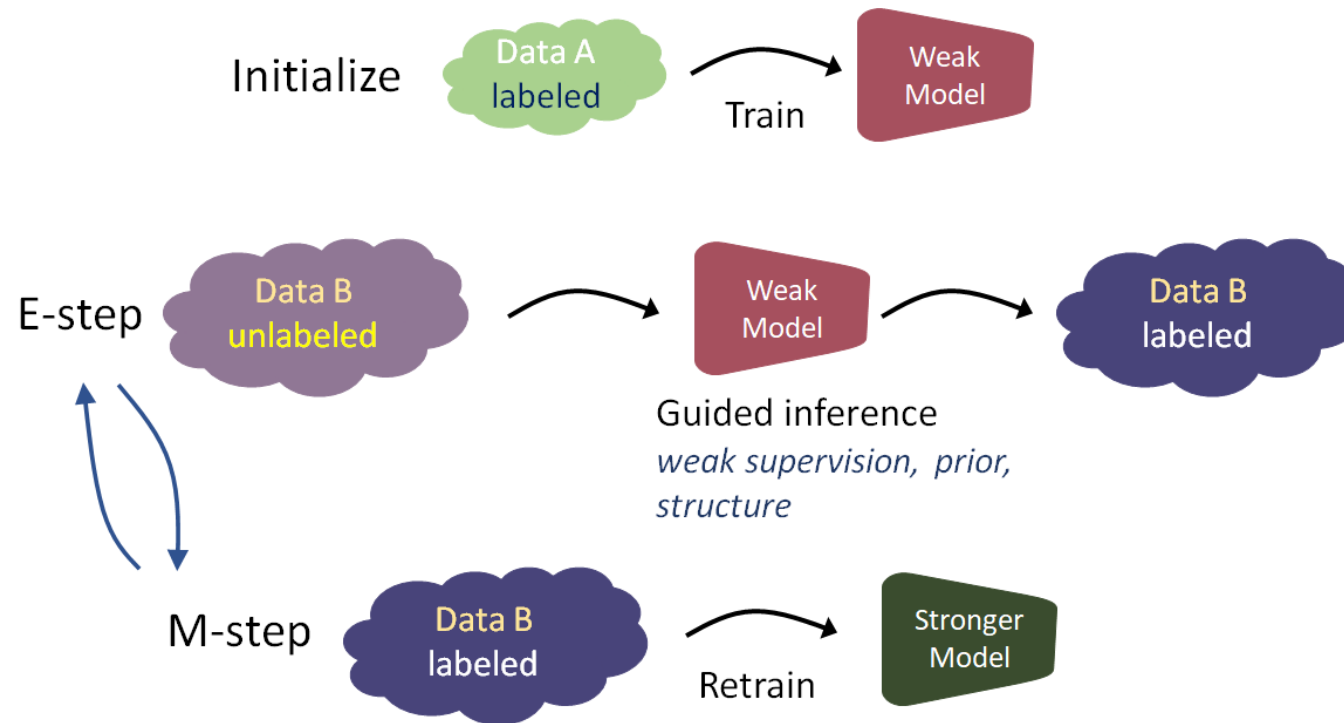
A: Piano B: Guitar

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

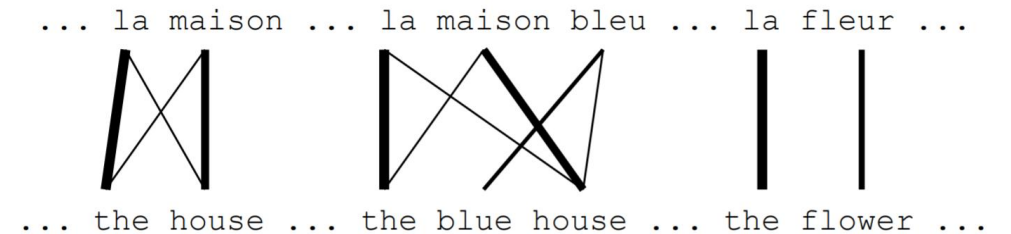
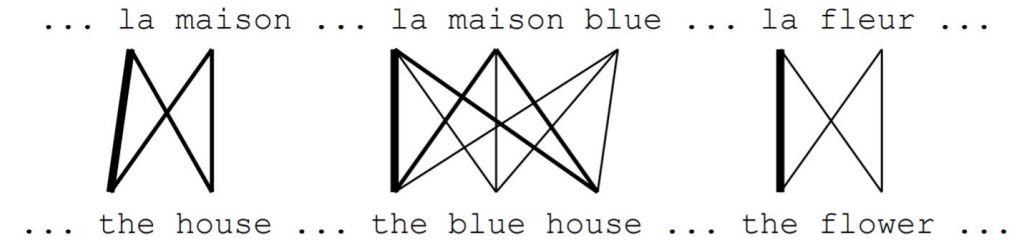
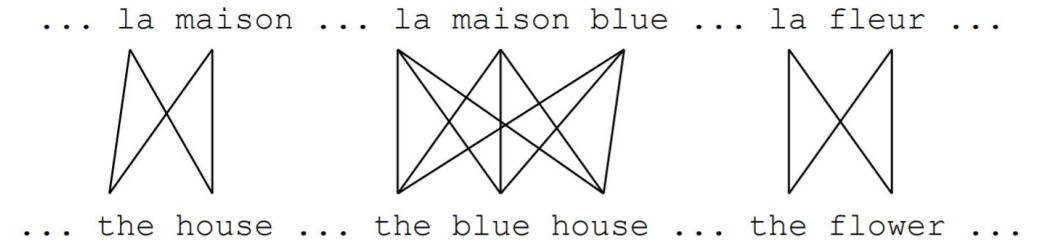
- IBM Alignment Model

- Learn word-level translation

Haus => house, building, home...

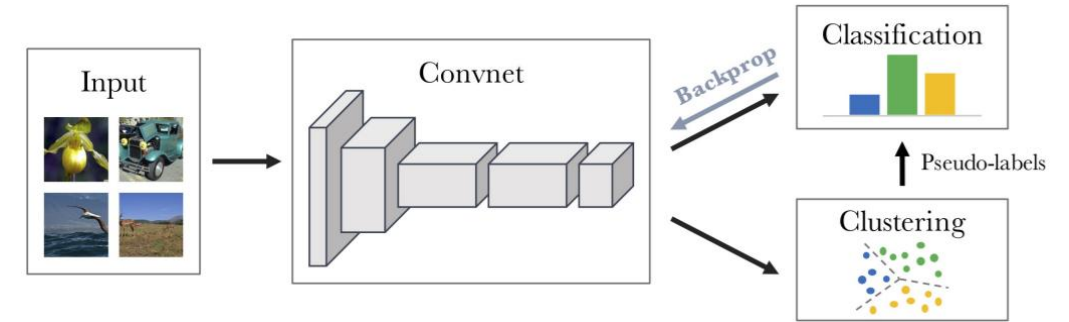
- Only paired sentences

- Weakly-supervised



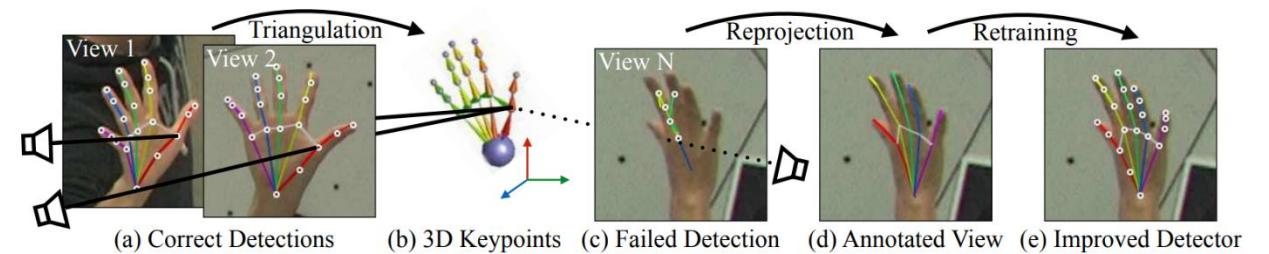
Self-Refinement

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



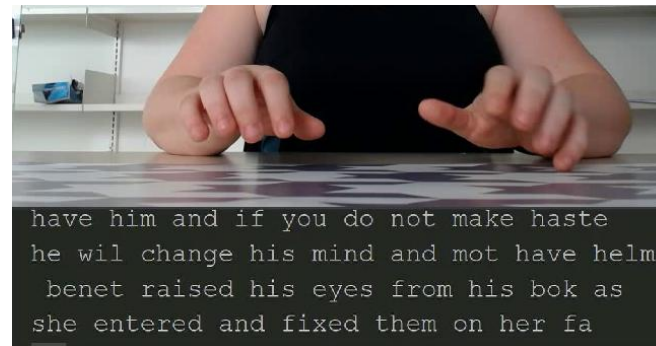
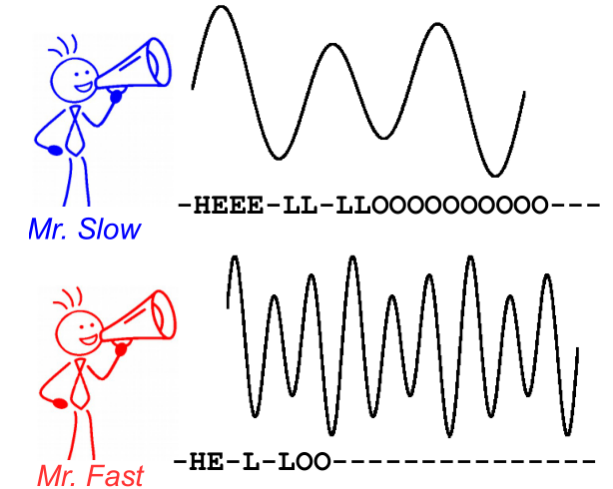
Caron et al., Deep Clustering for Unsupervised Learning of Visual Features, ICCV 2018
Xie et al., Self-training with Noisy Student improves ImageNet Classification, CVPR 2020

- Structure / prior?
Higher accuracy, less noise



Continuous signal → Natural Language

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

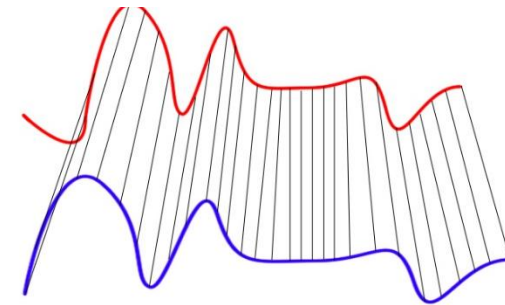
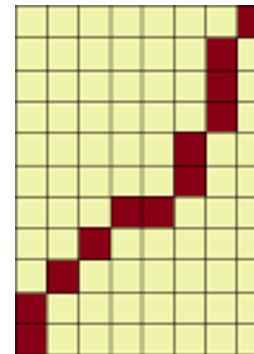
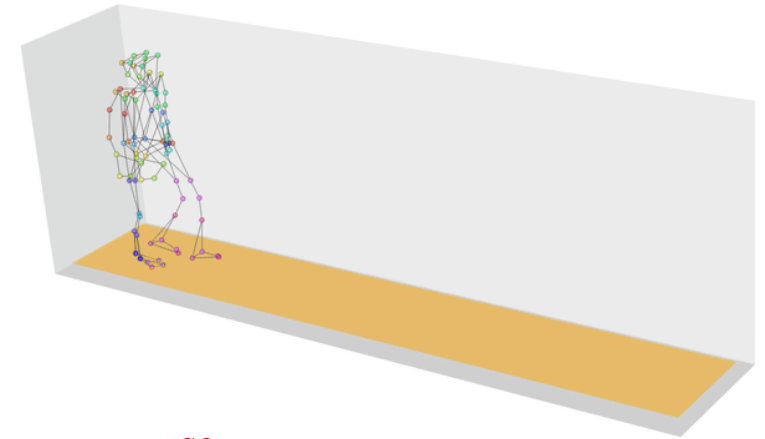


Score–Audio Alignment

Dynamic Time Warping (DTW)

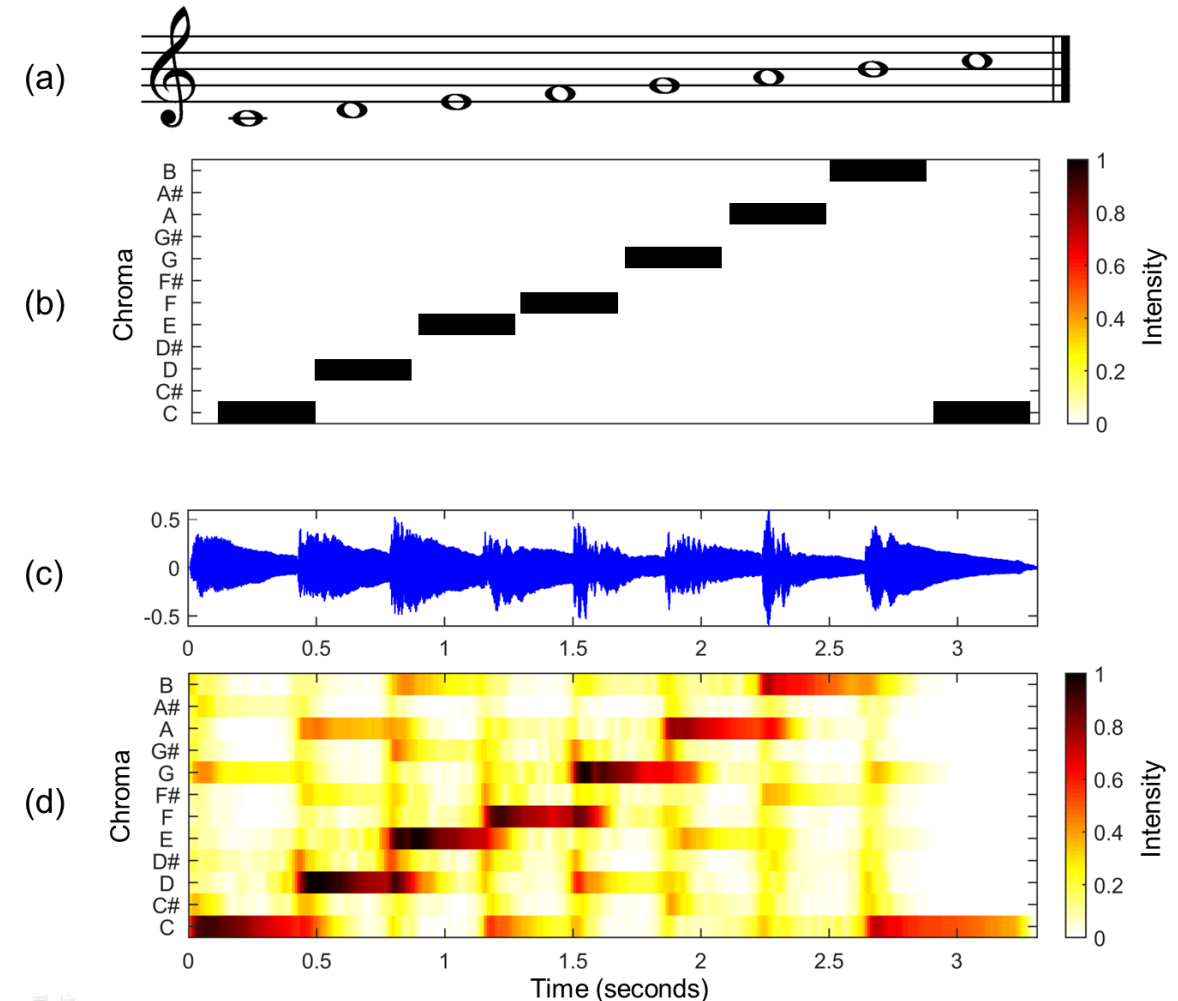
- Optimal monotonic warping between two sequences
- Our case: Warp score to audio
- Dynamic programming solution
- Need local cost function

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



DTW: Local Cost Function

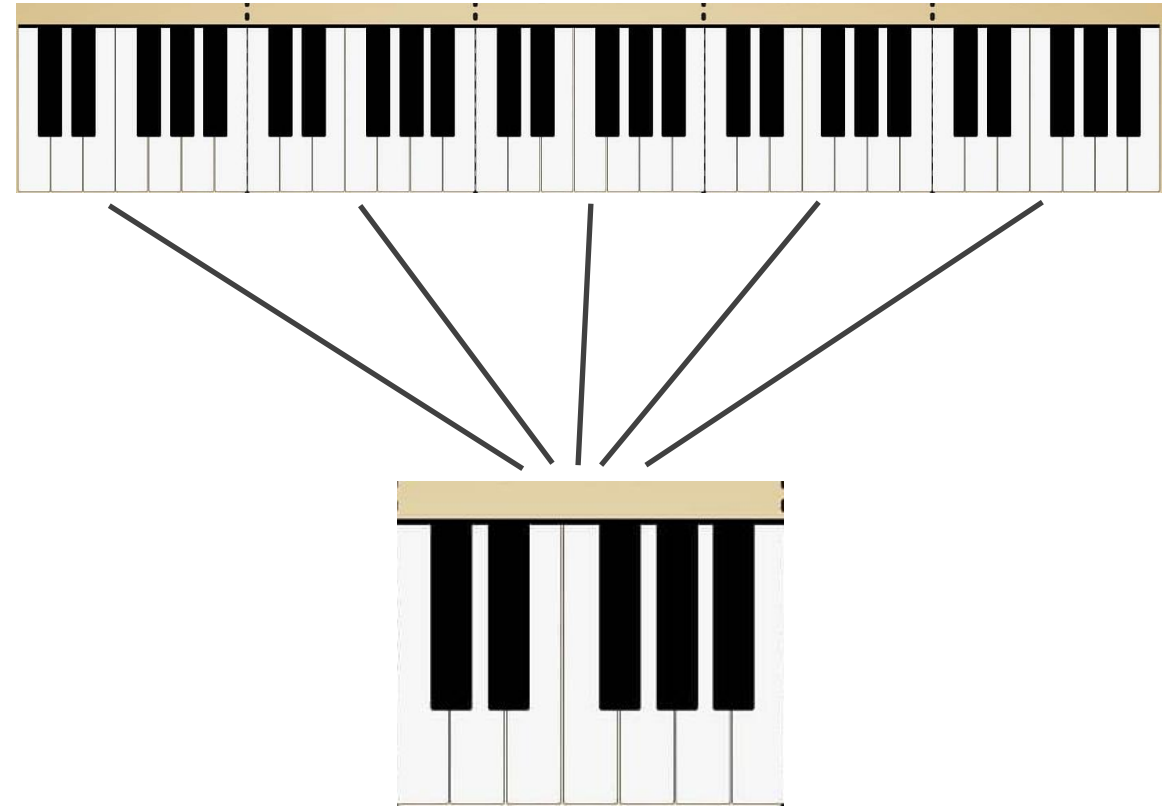
- Compare sequence elements, e.g., audio & score
- Audio frame & musical score frame—how similar?
 - Low-frequency STFT features
 - Chroma Features
 - Transcription features



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

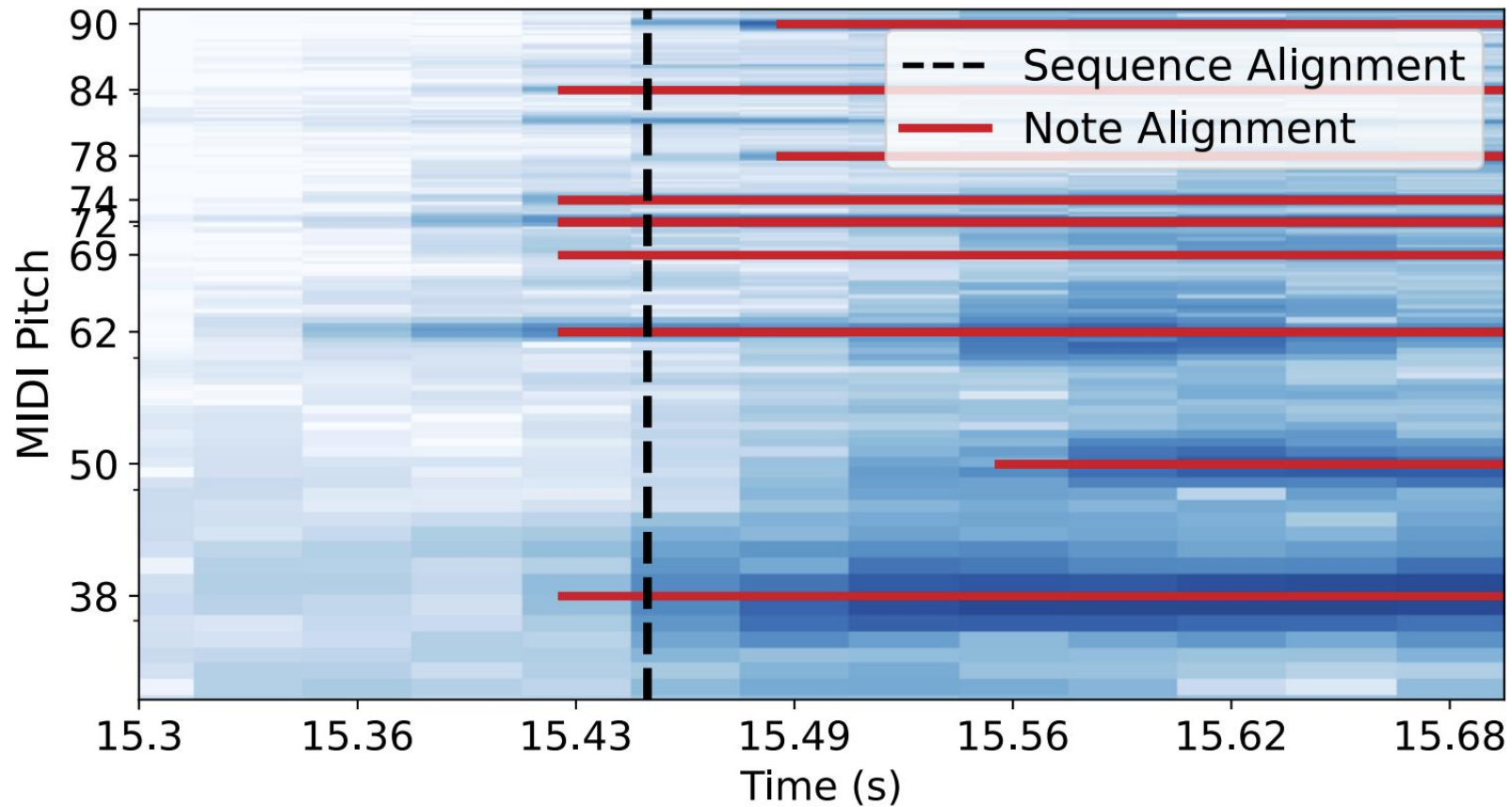
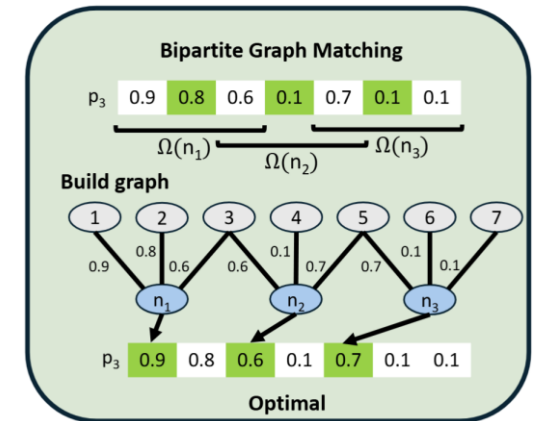
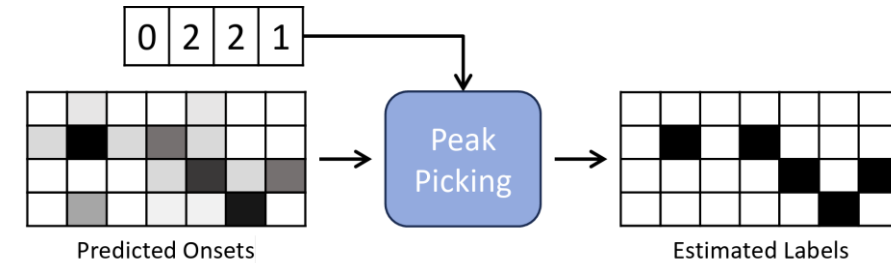
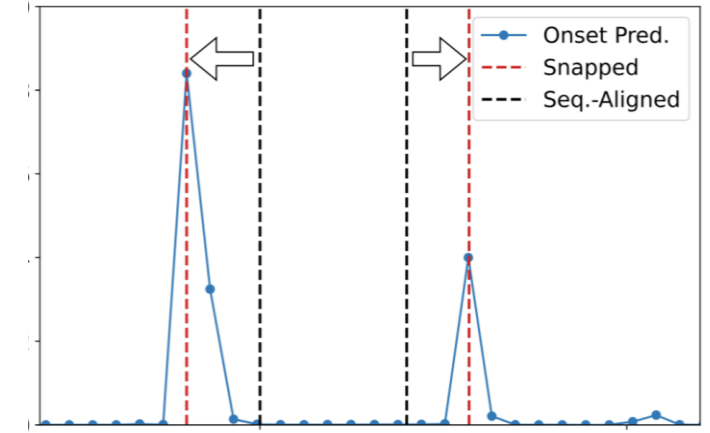


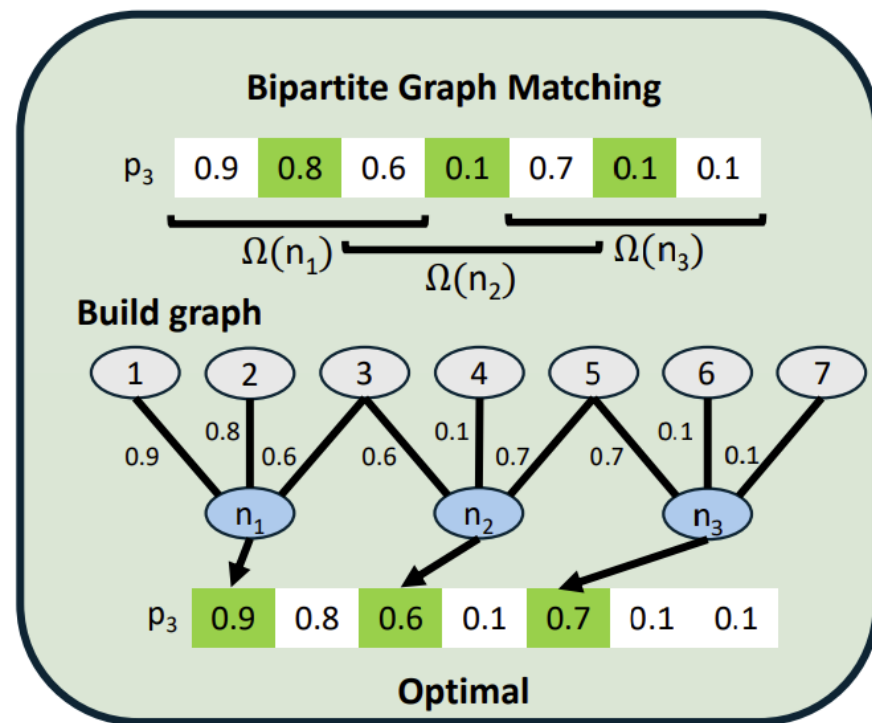
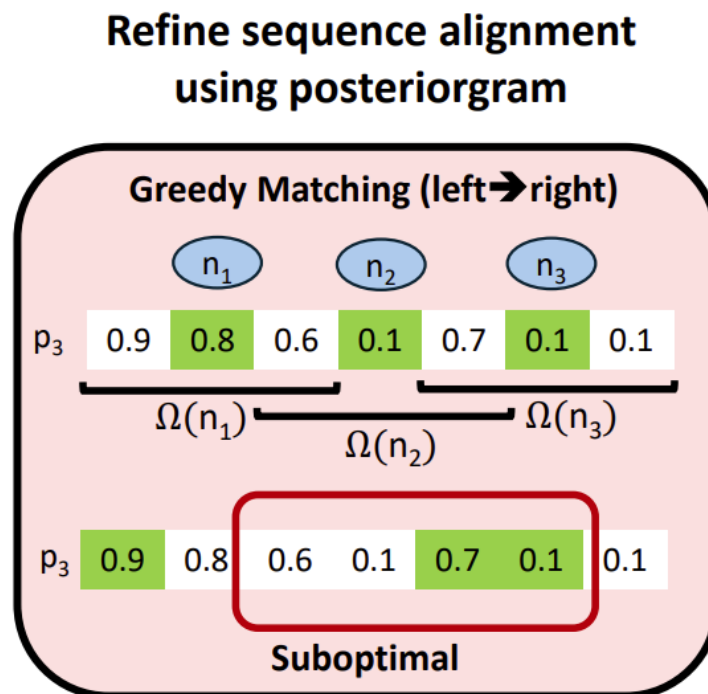
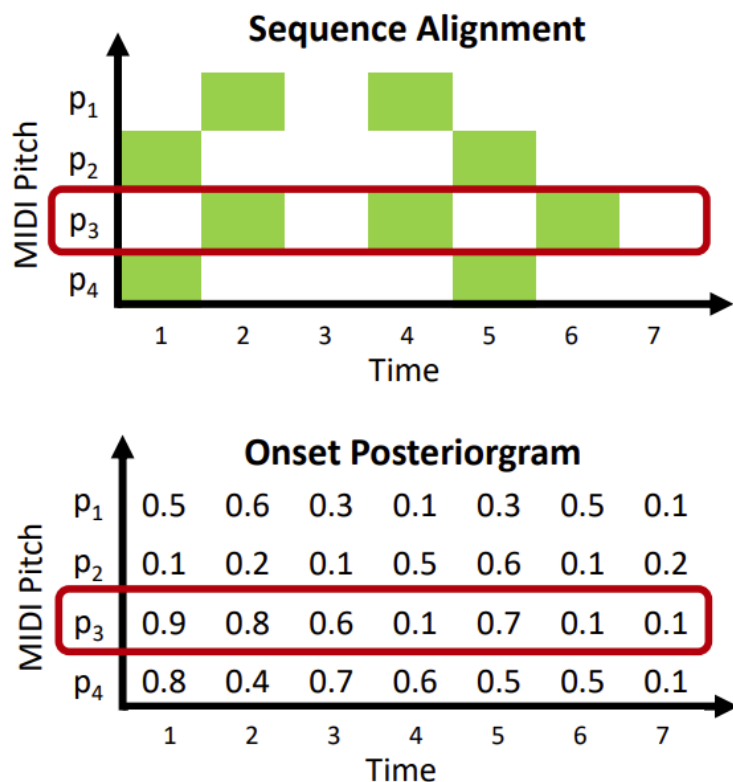
Figure 5 from Berendes et al.: Beethoven Symphony Excerpt Dataset (BSED): An Evaluation Dataset for Orchestral Music Transcription

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

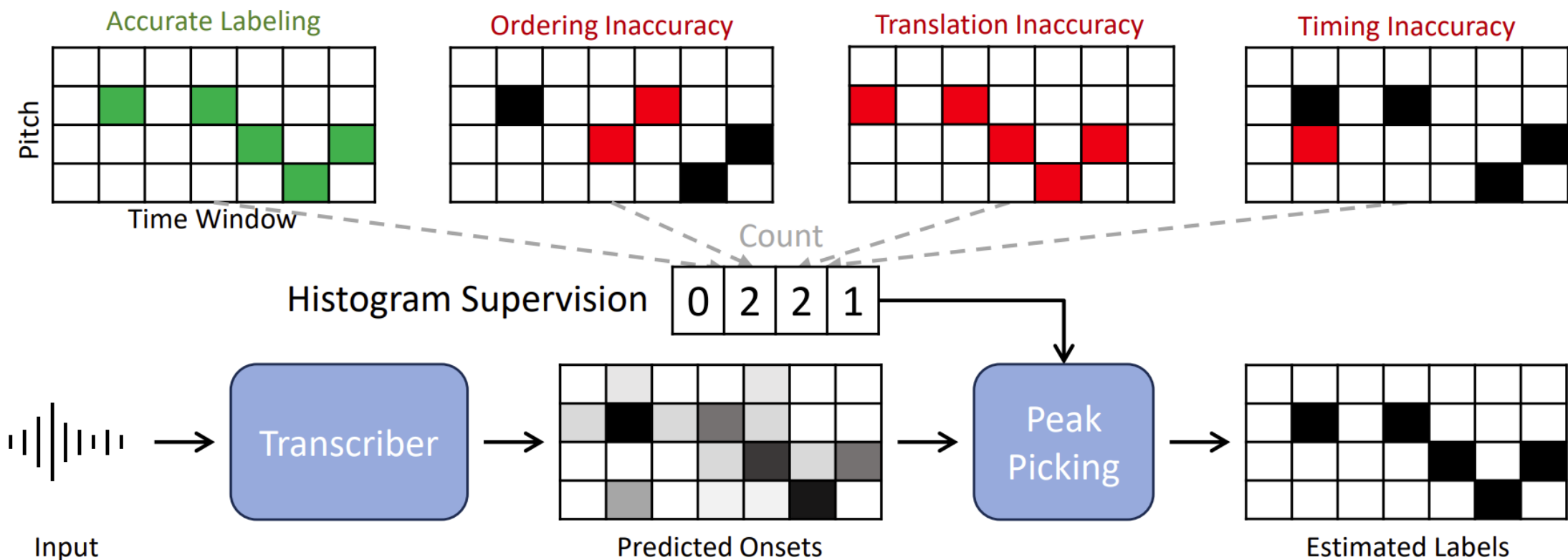


Onset “Snapping”: Greedy vs. Graph-Based



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

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



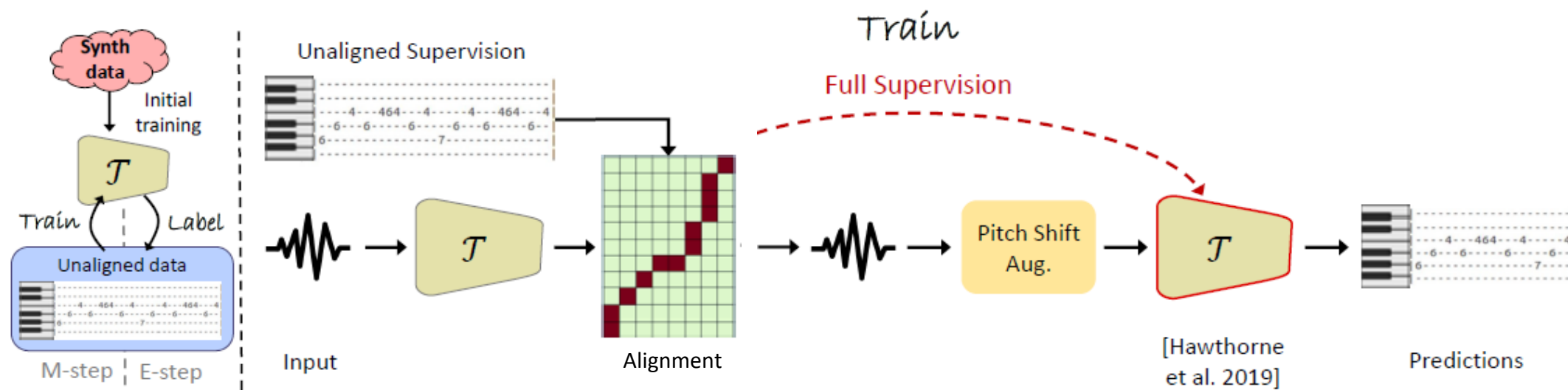
Yaffe et al.: Count The Notes: Histogram-Based Supervision for Automatic Music Transcription, ISMIR 2025

Note Duration, Note Offsets

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

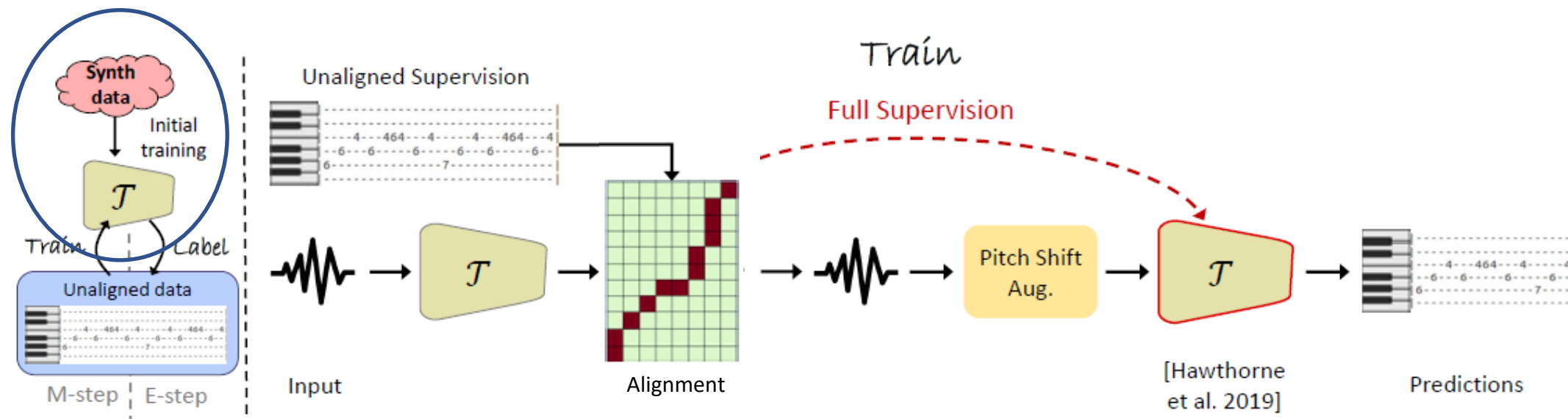
NoteEM

Pipeline

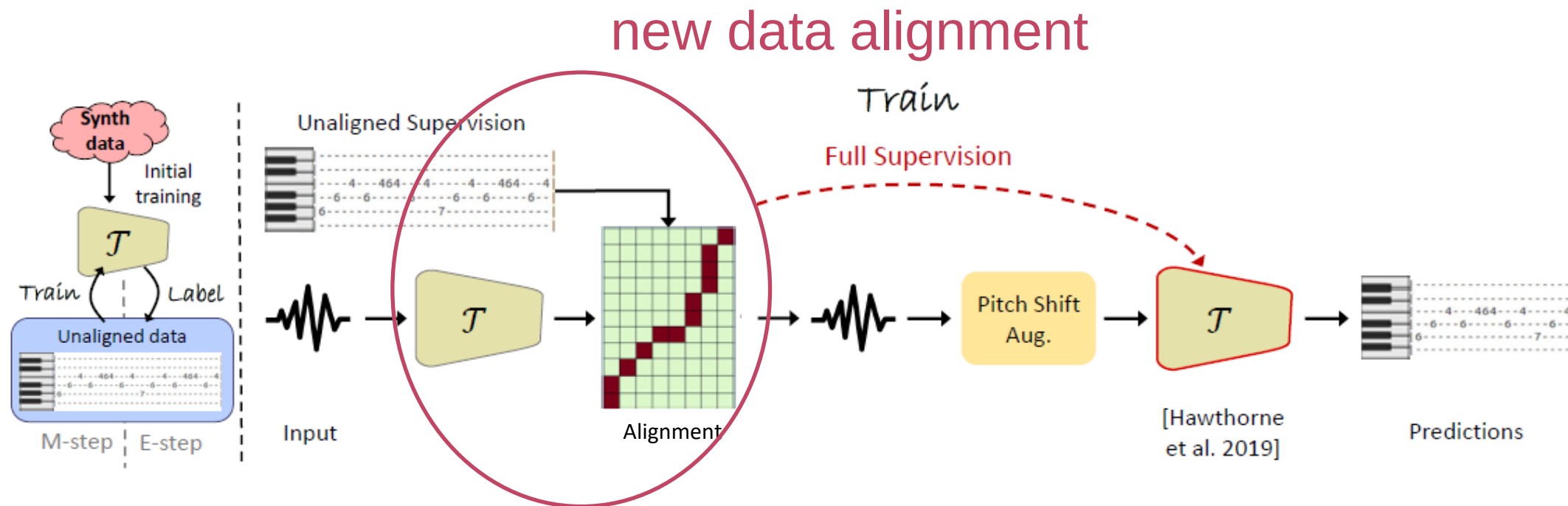


Pipeline

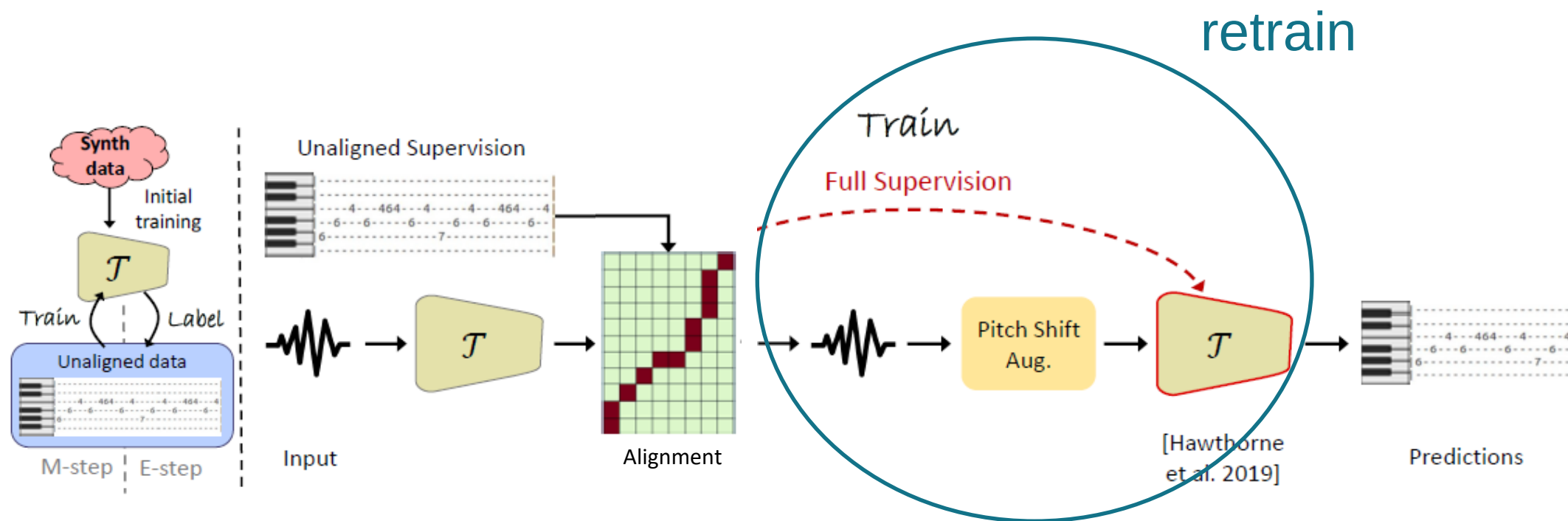
initial train



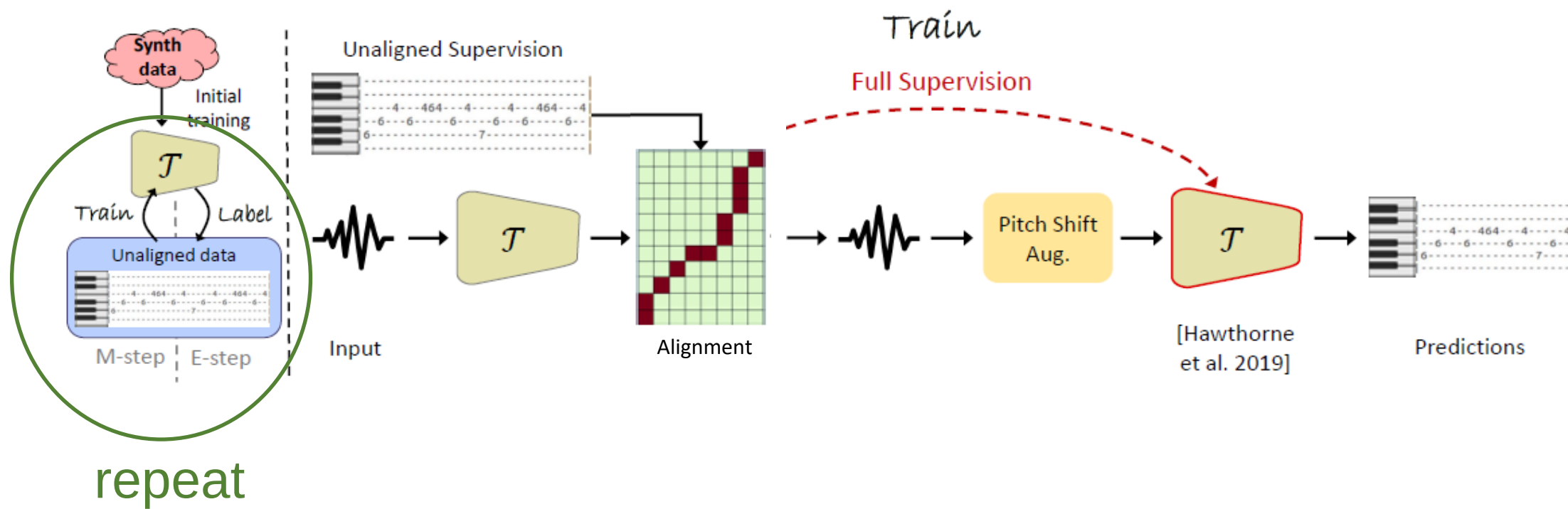
Pipeline



Pipeline



Pipeline

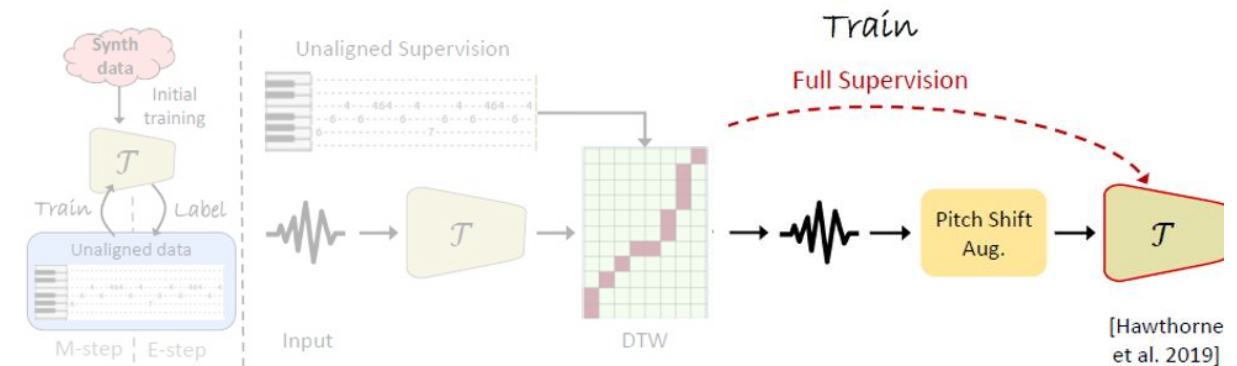
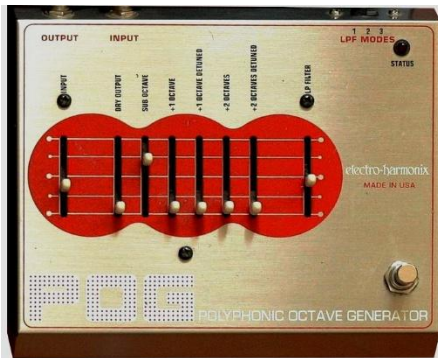


Pitch-Shift Augmentation

- 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



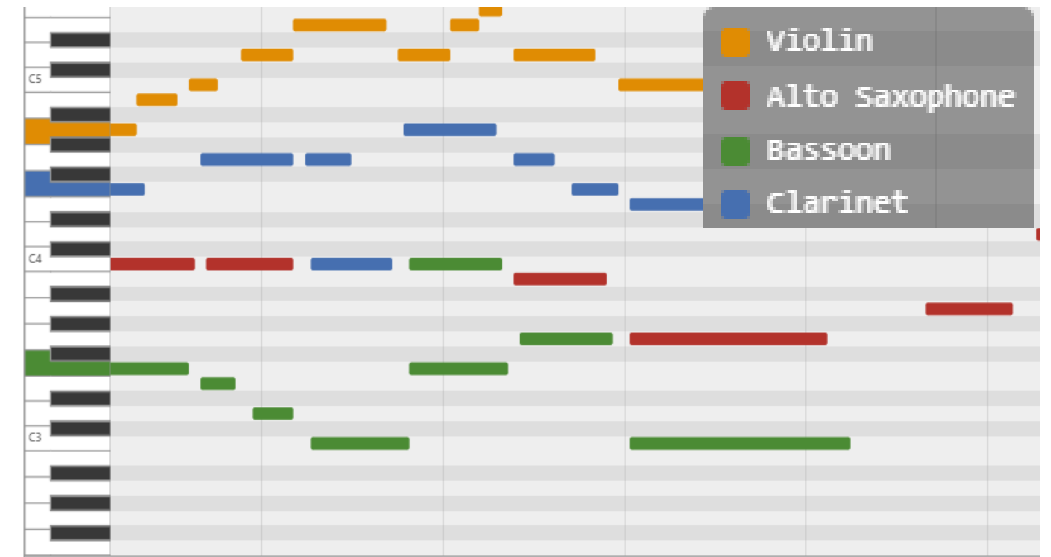
Metrics

- Note-level: Onset within 50ms
- Frame-level: Note activation / Multi-Pitch Estimation (MPE)
- 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

Instrument-Aware Transcription

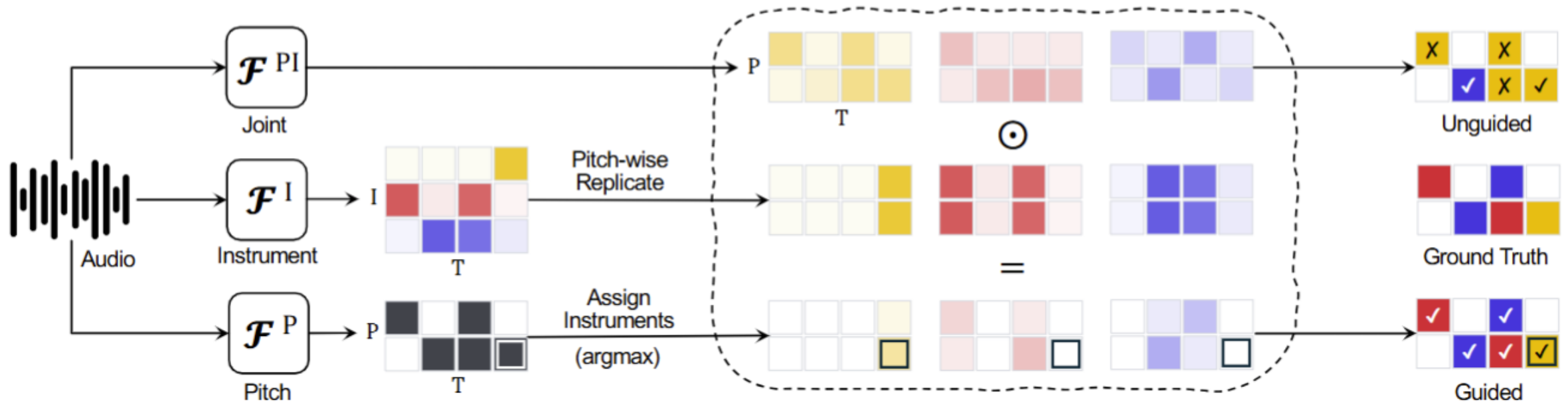
- Task: Detect pitch & instrument
- Joint modelling task → complex
- P pitches, I instruments → $P \times I$ classes!



Multi-Instrument Transcription

- 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 $\hat{M}^{PI} \in [0, 1]^{T \times P \times I}$

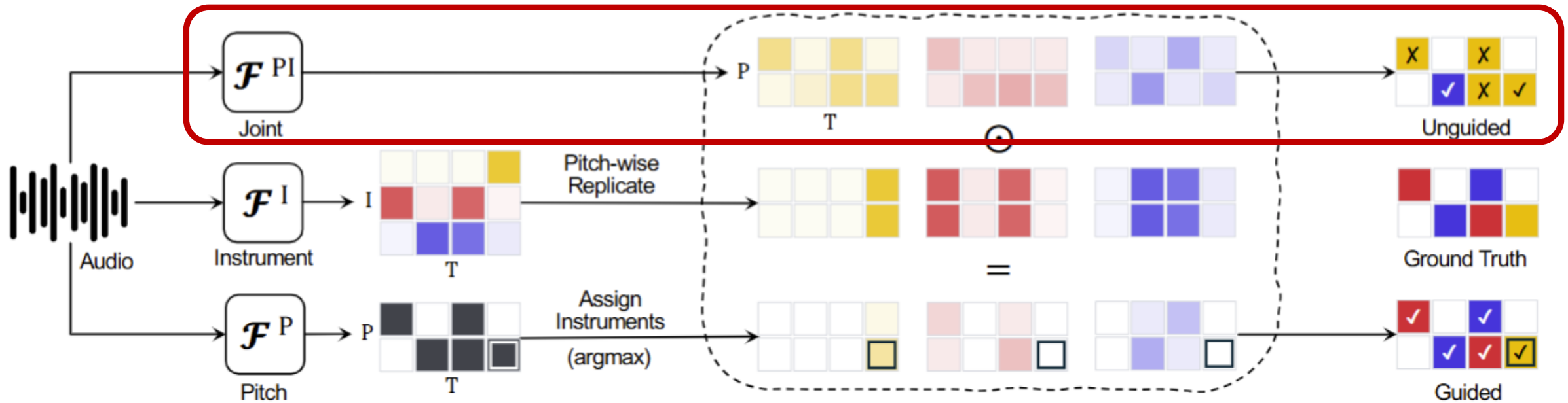
Pitch & Instrument Guidance



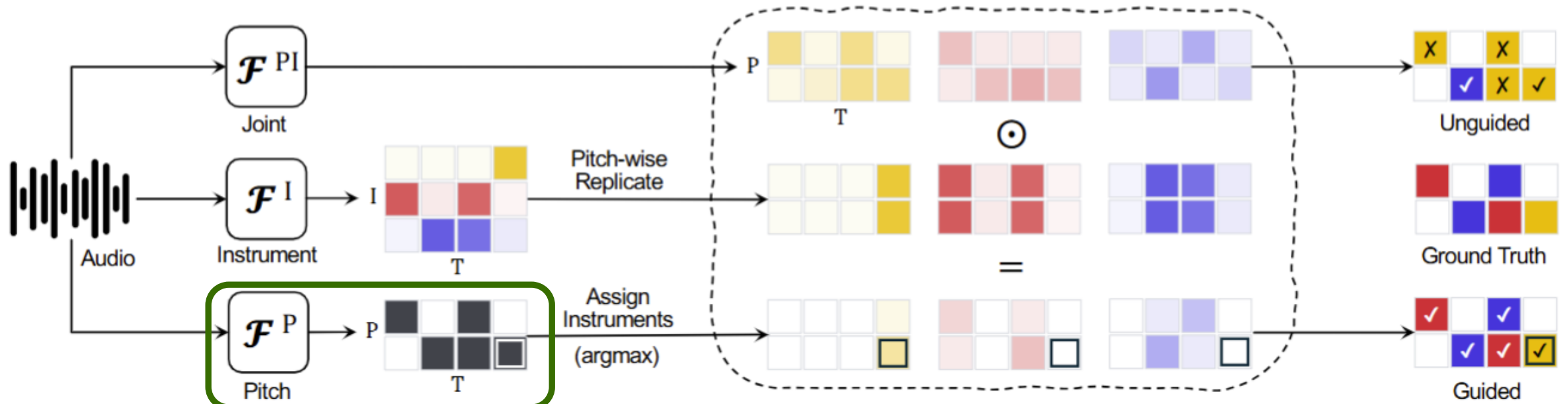
Pitch & Instrument Guidance

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)

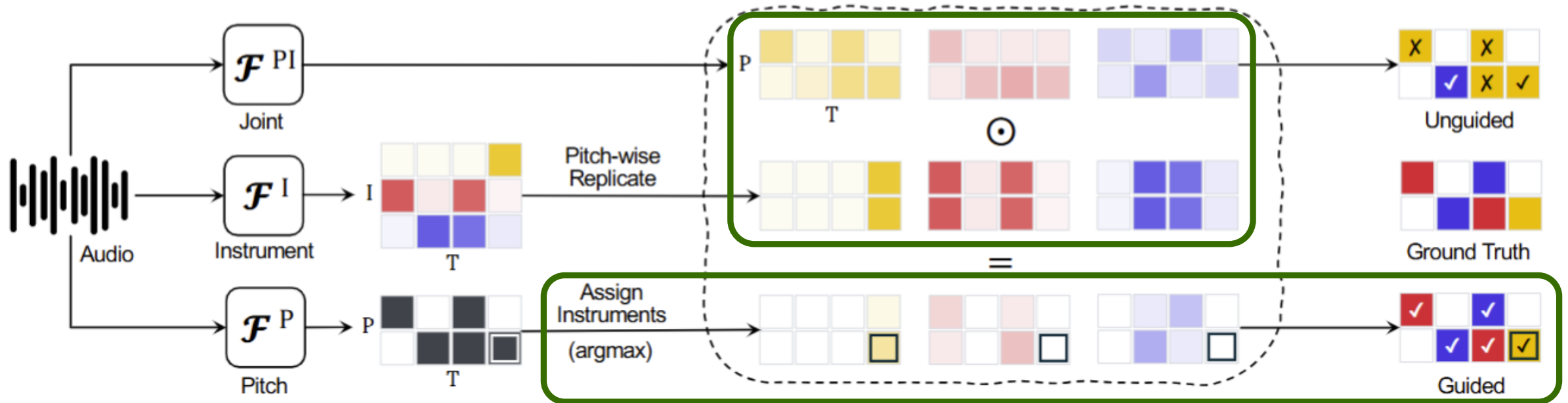
$$\mathcal{N} = \left\{ (t, p) : \hat{M}_{t,p}^P \geq \tau \right\}$$

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

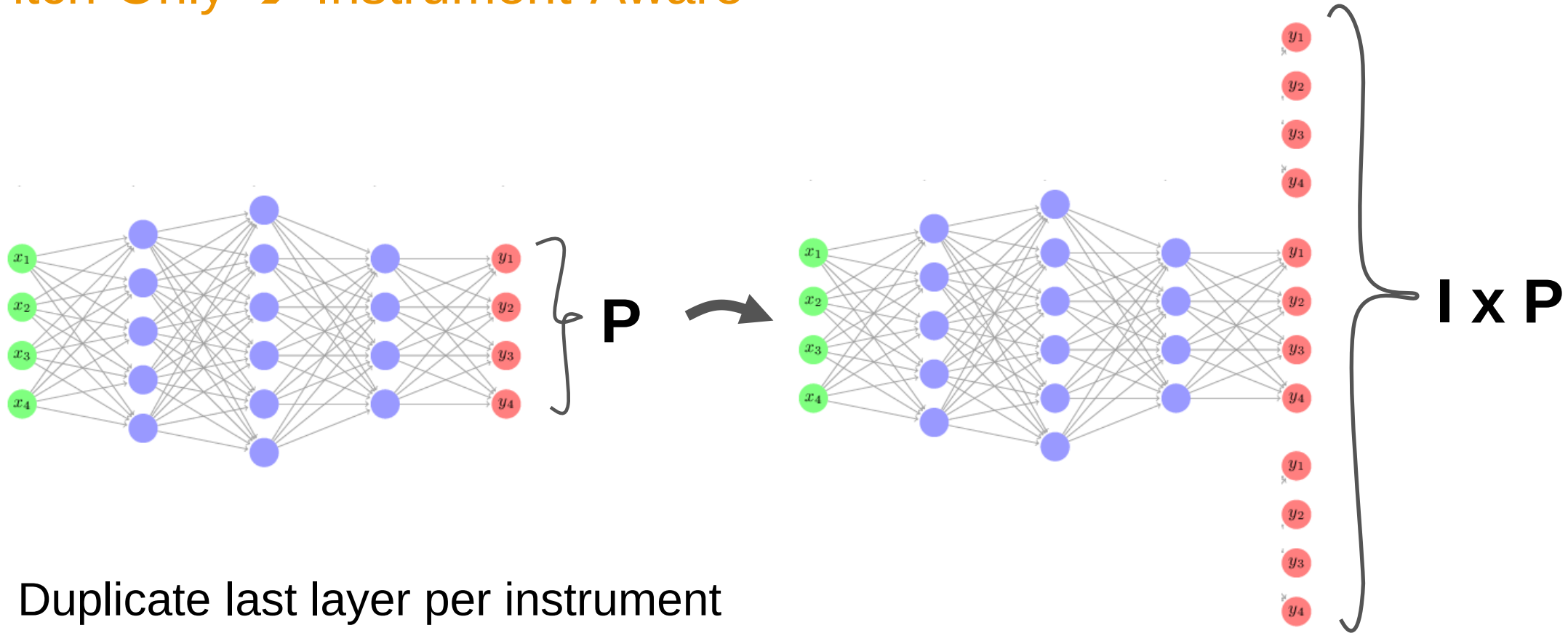
Pitch & Instrument Guidance

- 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

Pitch-Only → Instrument-Aware



- Duplicate last layer per instrument
- Initialize—all instruments active, labels—instrument-aware

Thank You!