AI Pianist:
Modeling Expressive Performance with Deep Learning

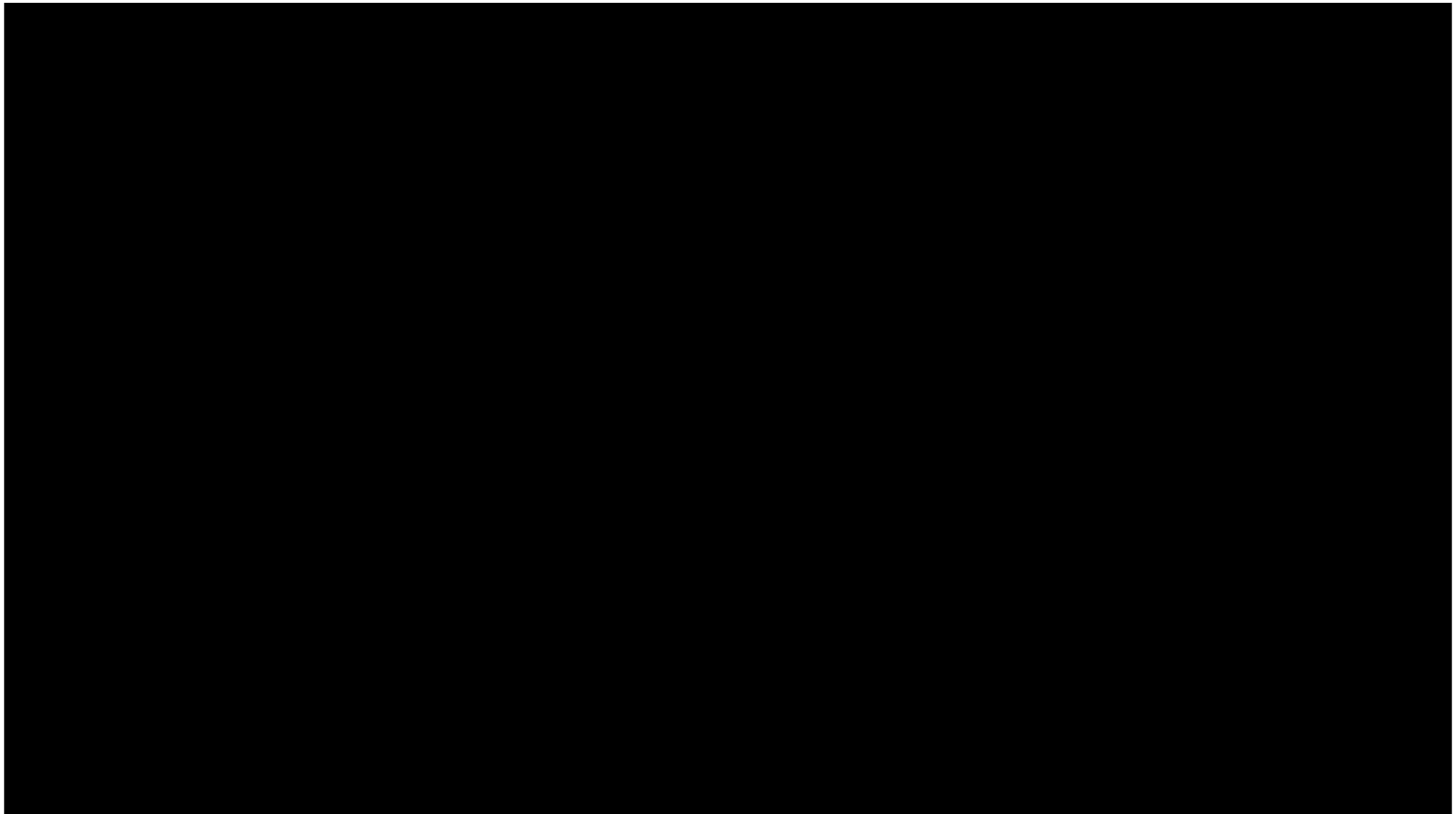
Dasaem Jeong,
Assistant Professor @ Dept. Art & Technology, Sogang University
2022. 4. 8

https://jdasam.github.io

Career

- Master @ KAIST (2013-2015):
  - Music visualization
- PhD @ KAIST(2015-2020)
  - Performance modeling, Performance transcription, Audio-to-score alignment
- T-Brain, SK Telecom (2020 - 2021.8):
  - Real-time AMT, MIDI beat-tracking, Singing Voice Synthesizer, Music representation learning, Query-by-humming
- Dept. of Art & Technology, Sogang University (2021.9 - Present)
  - Music and Art Learning (MALer) Lab

The Art of Performance

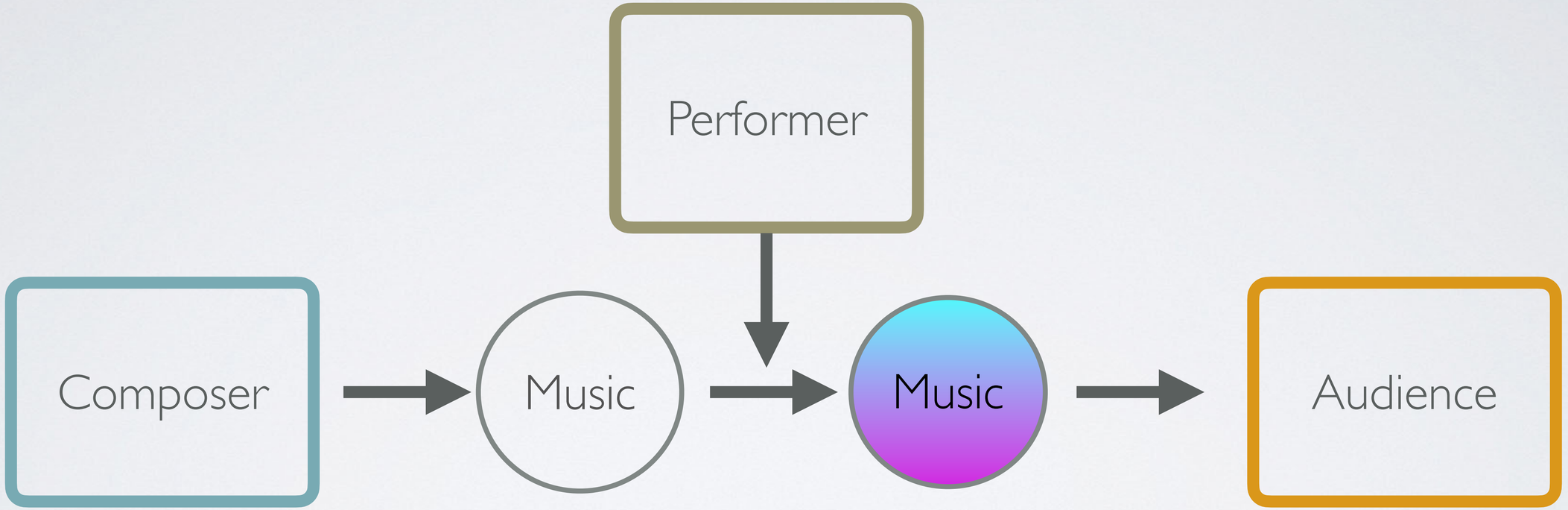


1. Introduction
2. Performance Modeling with RNN
3. Performance Modeling with GNN
4. Performance Style Analysis
5. Future Research

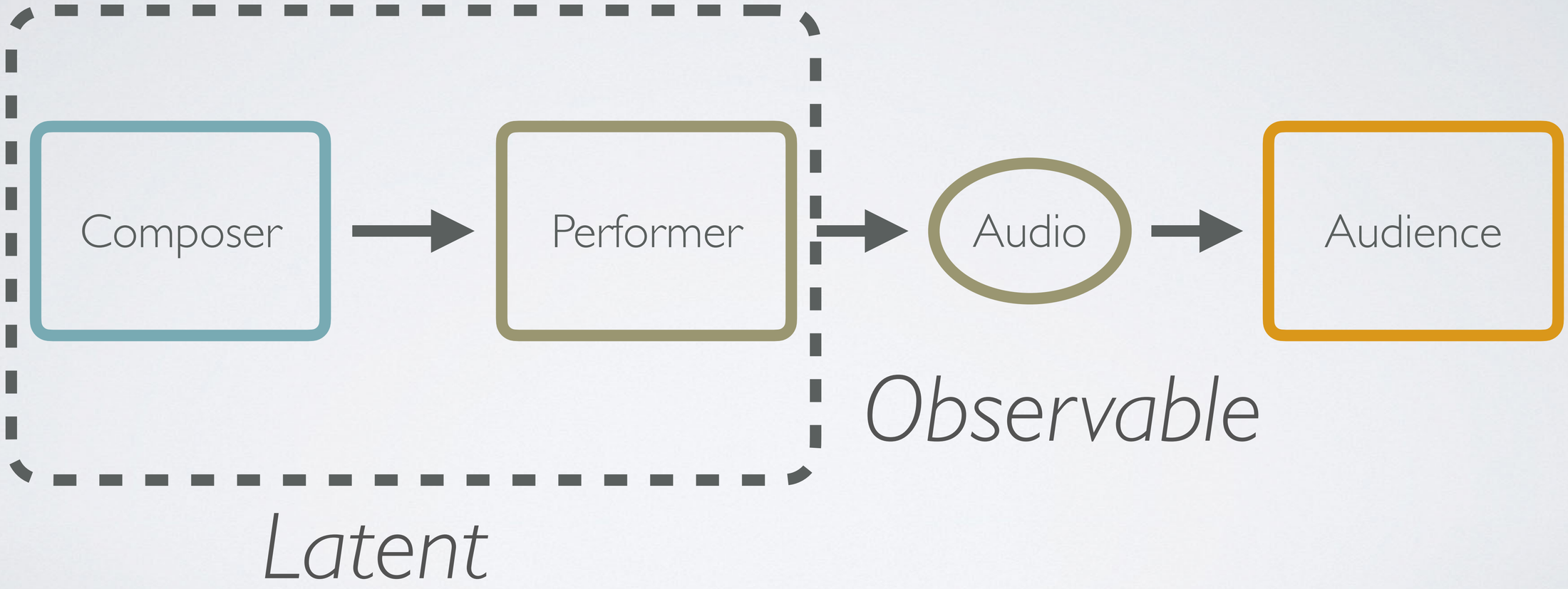
Primary Target of our Research

Pop	Rock	Jazz	**Classical**	Folk
Composing	**Performing**	Synthesizing	Listening	
Violin	Singing	**Piano**	Drum	Trumpet

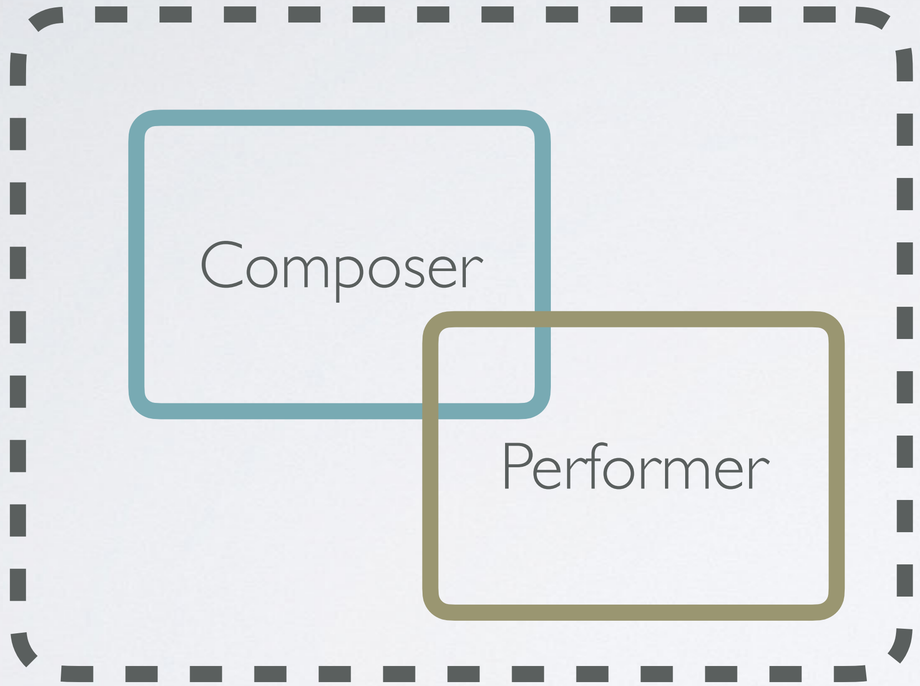
Role of Performer



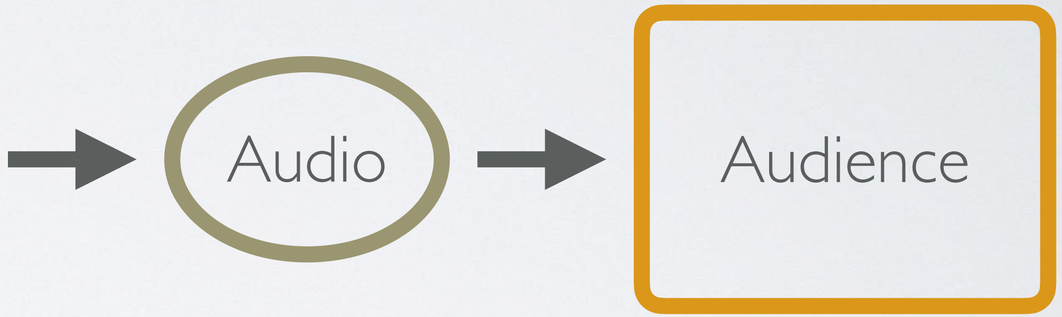
In Most of the Music Genres



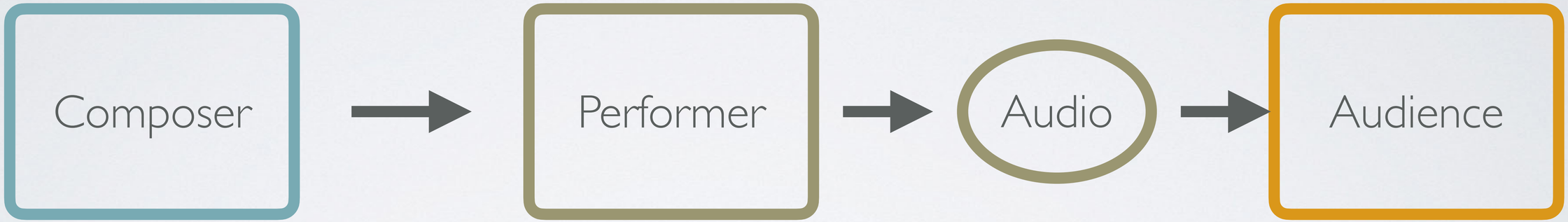
Why Classical Music?



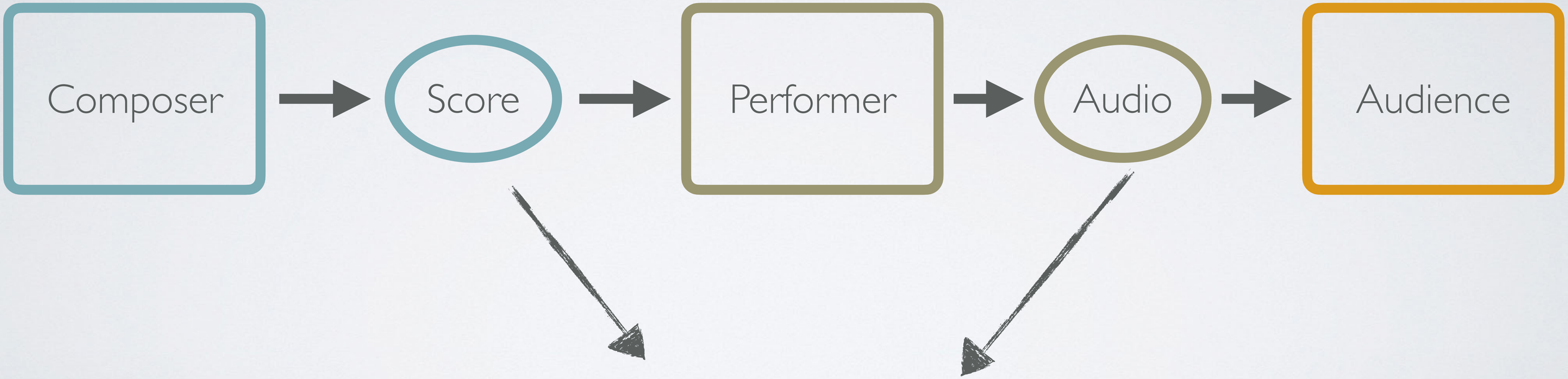
Entangled



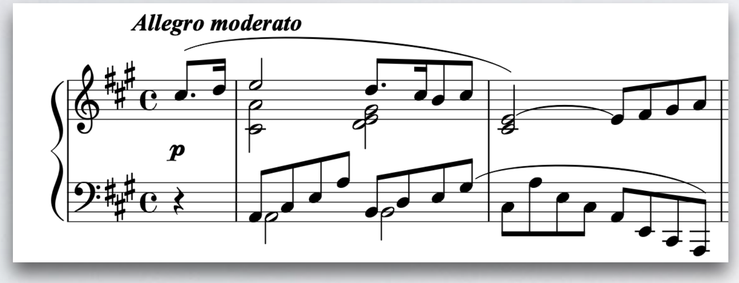
In Classical Music

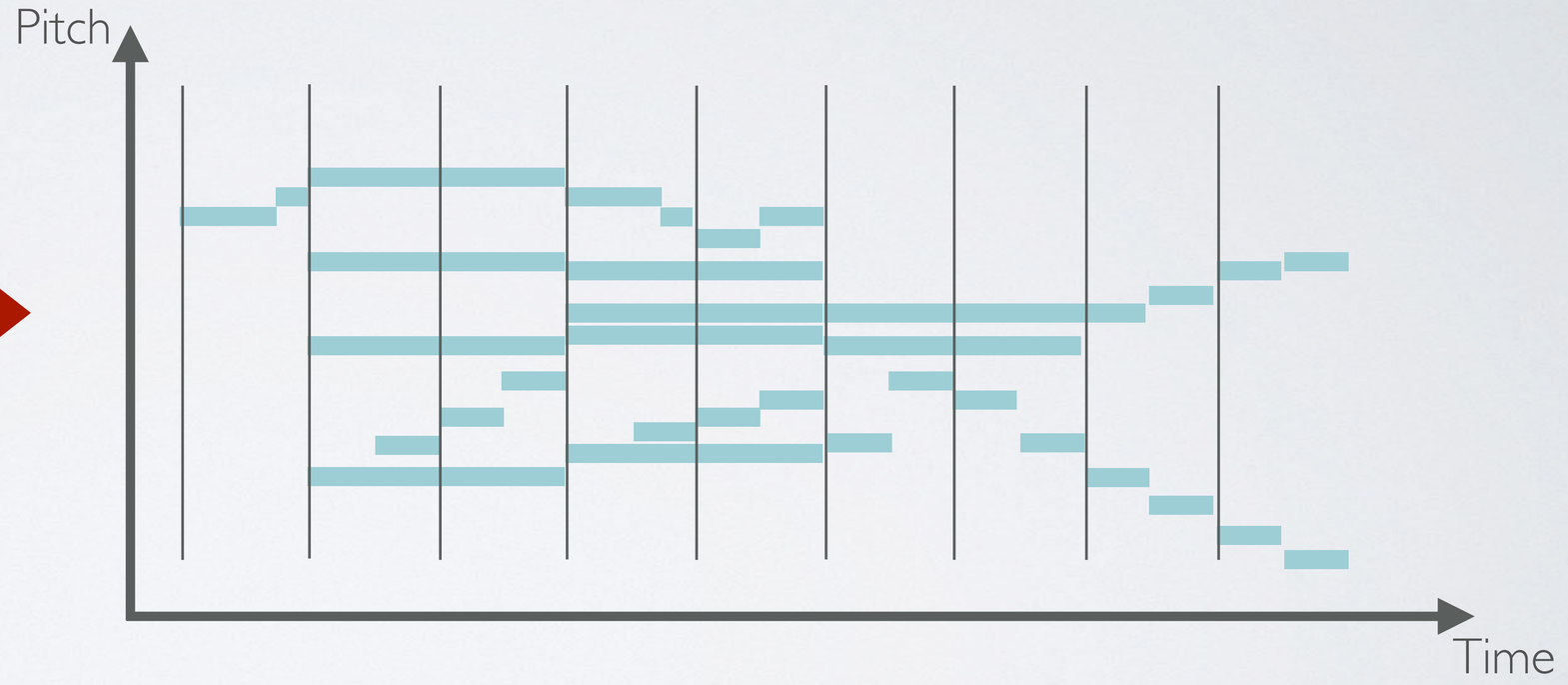


In Classical Music



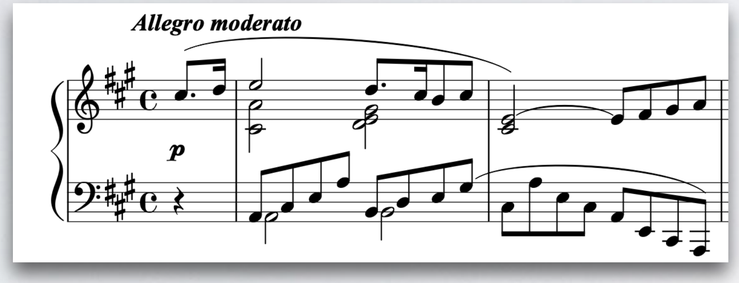
The Art of Performance





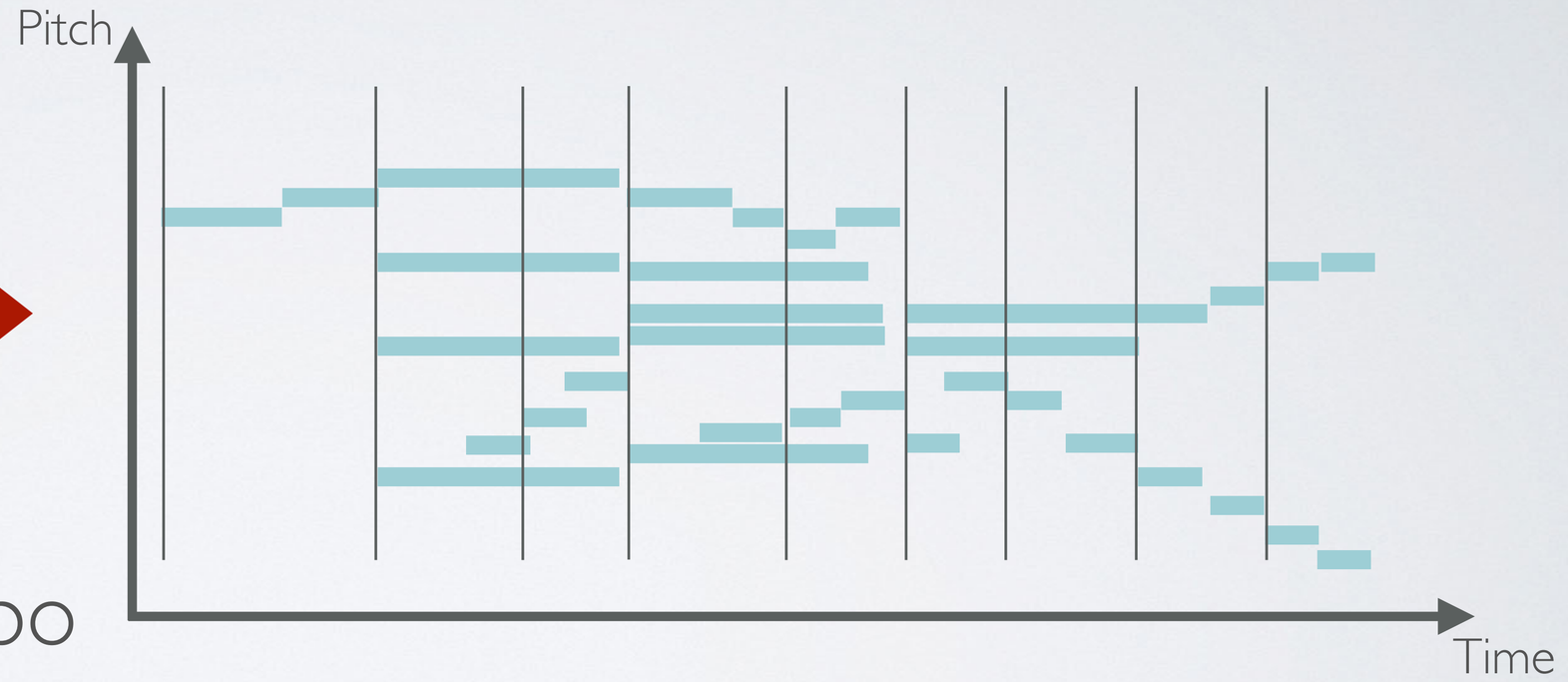
Mechanical Rendering

The Art of Performance

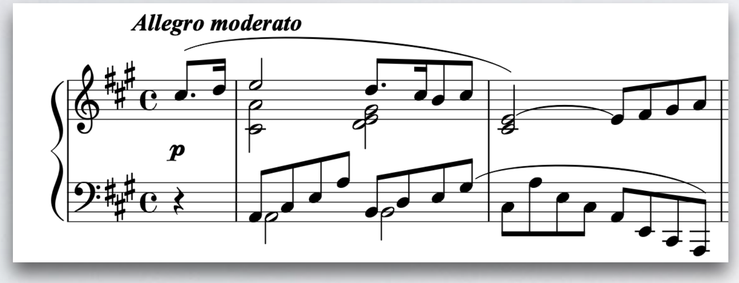


+

Tempo



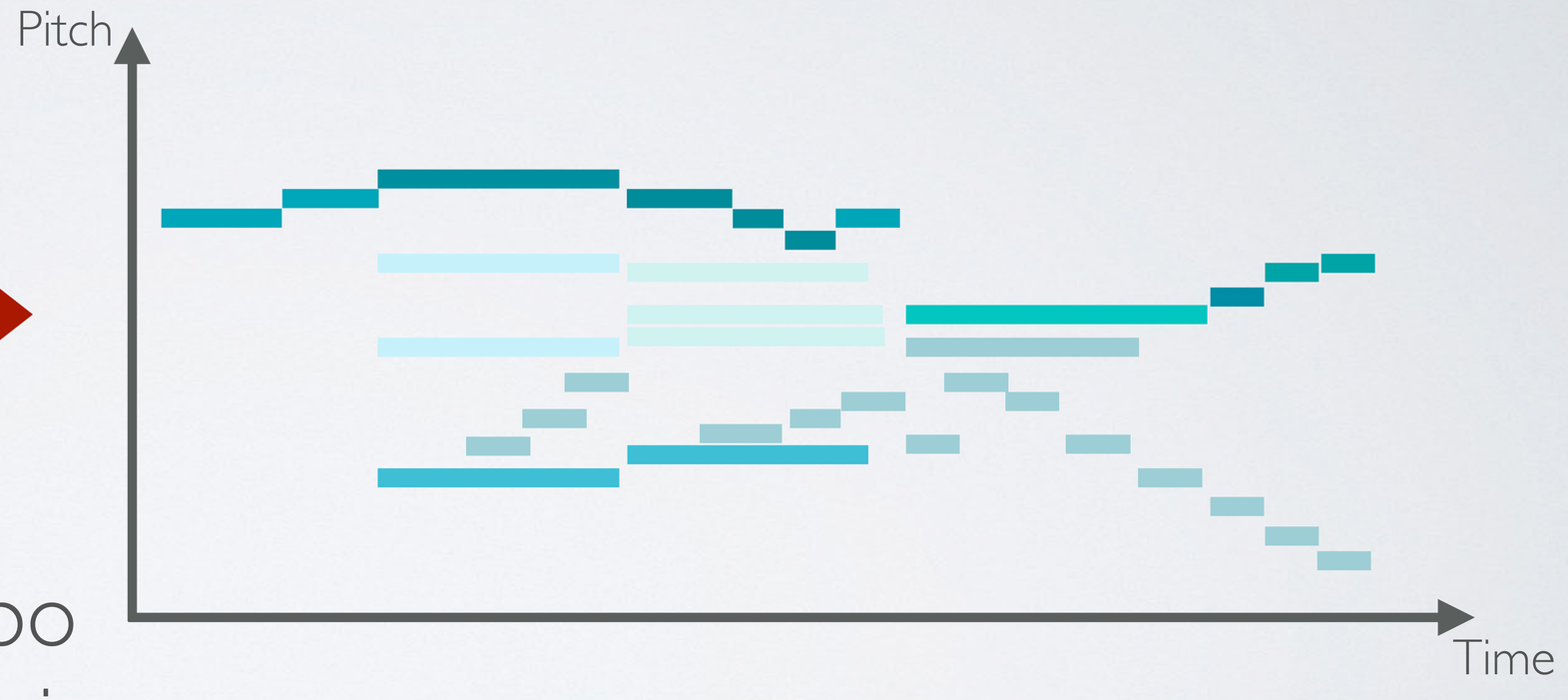
The Art of Performance



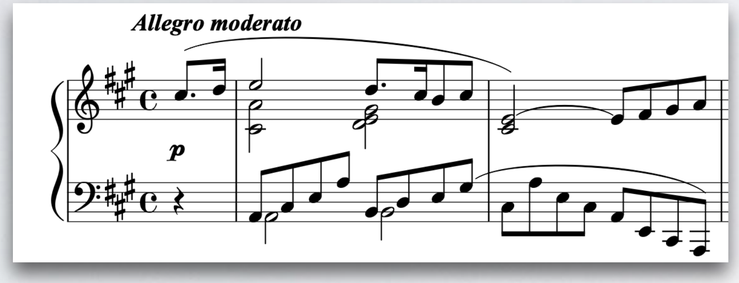
+

Tempo

Dynamics

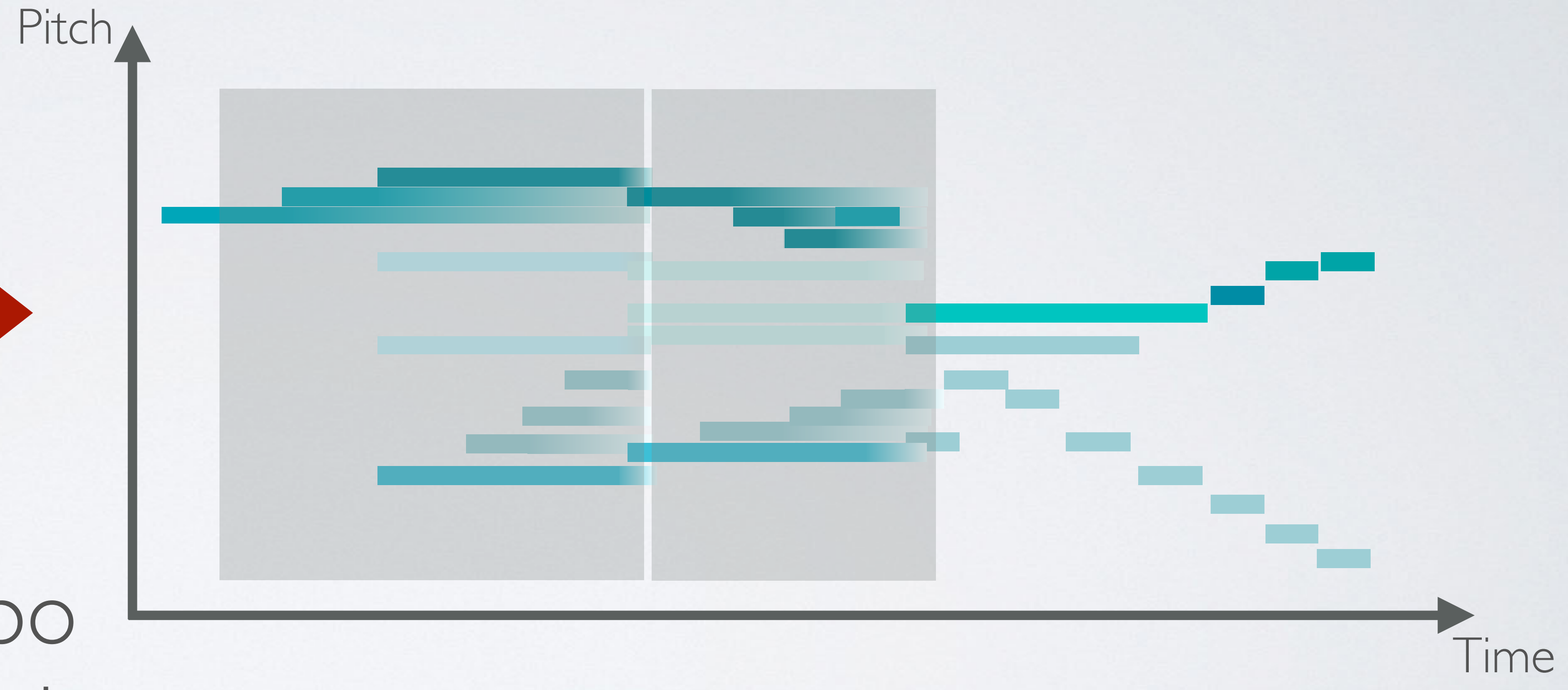


The Art of Performance



+

Tempo
Dynamics
Articulation and Pedals

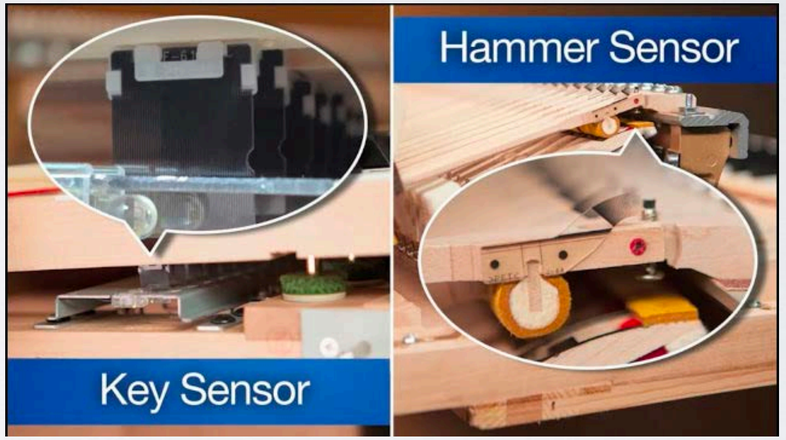


Human Performance

Why do we focus on Piano?

- Solo performance can have extreme expressivity
- Used in various genres
- Polyphonic (multiple notes at once)
- Used for both solo and ensemble performance

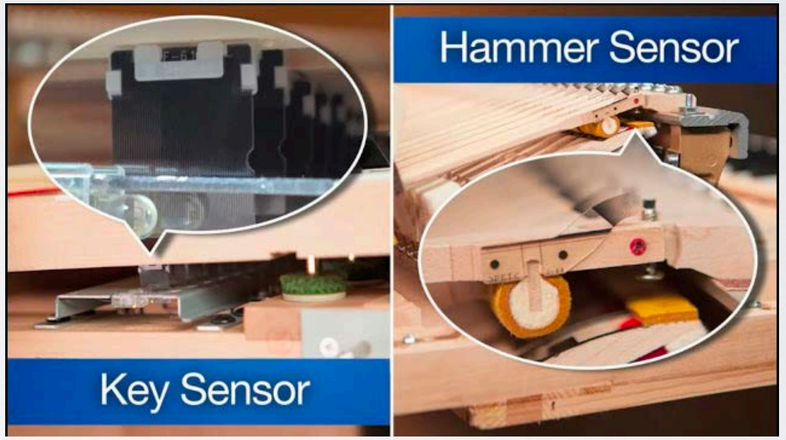
Computer-Controlled Piano



- The piano performance can be easily quantized by recording every mechanical movement of 88 keys and pedals

http://blog.yamahamusicsoft.com/2016/01/take-technical-tour-of-disklaviers.html

Computer-Controlled Piano



- Reconstruction of performance can be easily done with MIDI file and a computer-controlled piano

http://blog.yamahamusicsoft.com/2016/01/take-technical-tour-of-disklaviers.html

1. Introduction

2. Performance Modeling with RNN

3. Performance Modeling with GNN

4. Performance Style Analysis

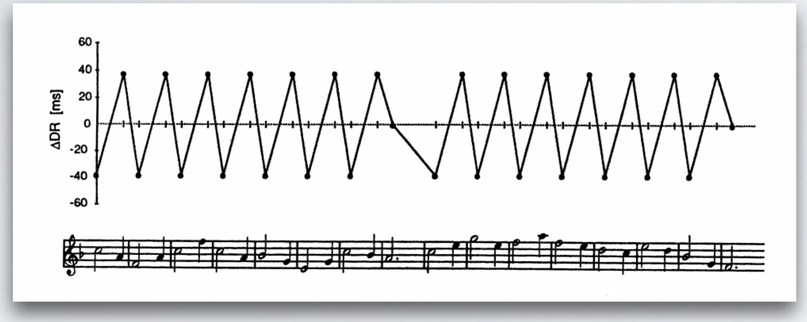
5. Future Research

Jeong et al, VirtuosoNet: A Hierarchical RNN-based System for Modeling Expressive Piano Performance (ISMIR, 2019)
Jeong et al, Score and Performance Features for Rendering Expressive Music Performances (Music Encoding Conference, 2019)

Task Definition

- Generating human-like performance for a given score

Previous Research



Generative rules for music performance: A formal description of a rule system

Friberg (1991)

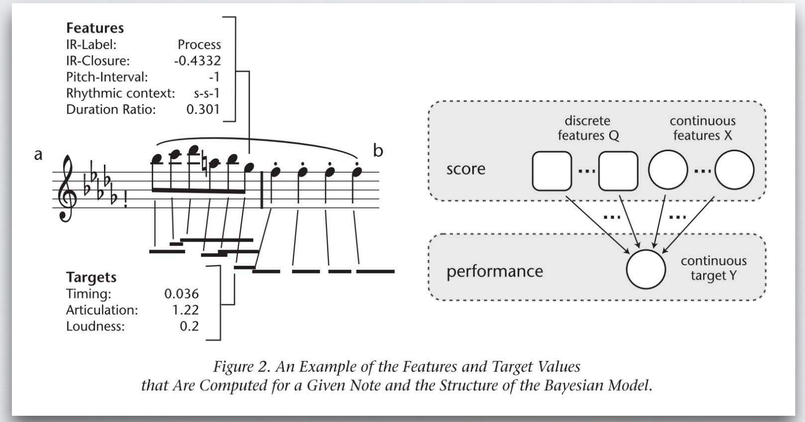
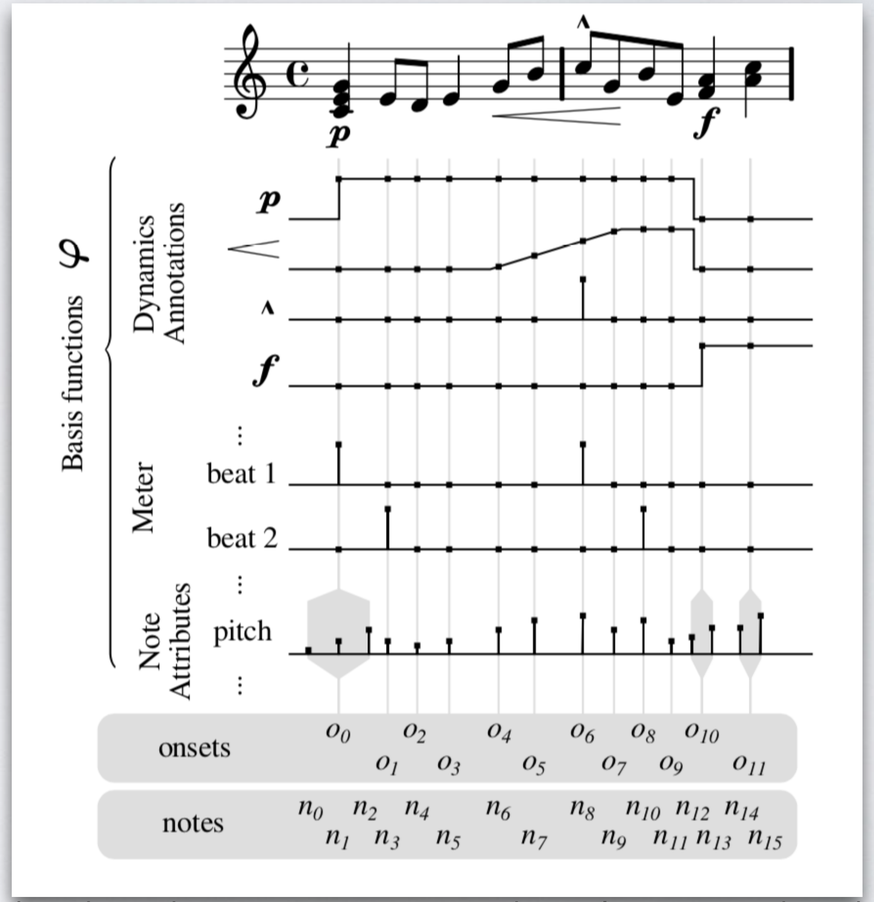


Figure 2. An Example of the Features and Target Values that Are Computed for a Given Note and the Structure of the Bayesian Model.

YQX plays Chopin

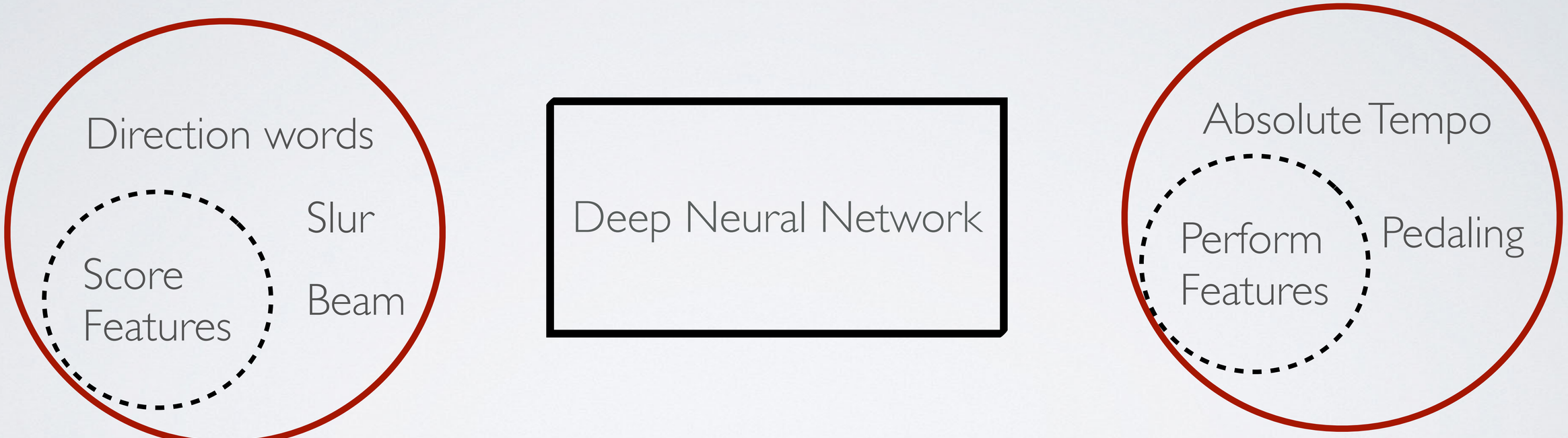
Widmer et al (2007)



The basis mixer: a computational romantic pianist

Cancino-Chacon et al (2016)

Our Aim


Direction words
Slur
Score Features
Beam
Deep Neural Network
Absolute Tempo
Perform Features
Pedaling

Performance Generation

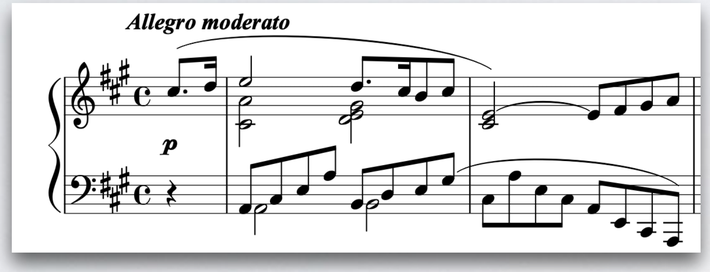
Score

MusicXML

Performance Generation System

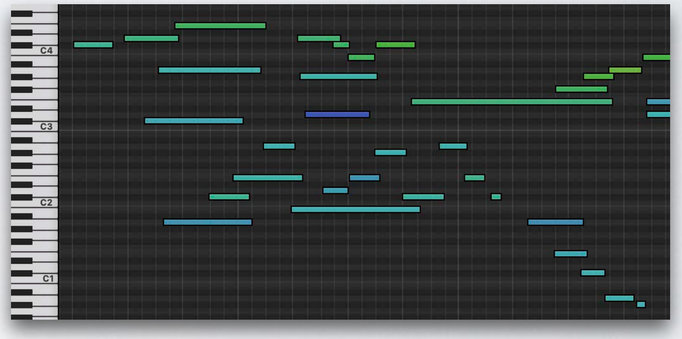
Performance

MIDI



Music Score

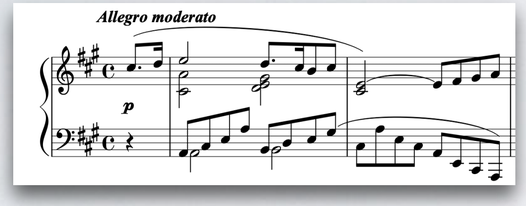
As a pair

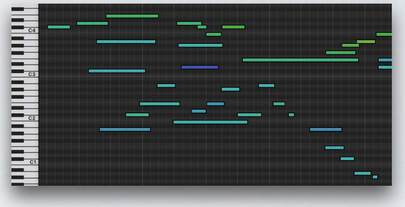


Human Performance

- We need pairs of music score and human performance for the training

Dataset



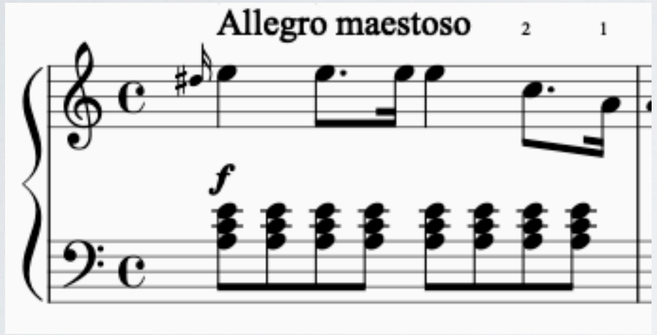


MusicXML

Performance MIDI

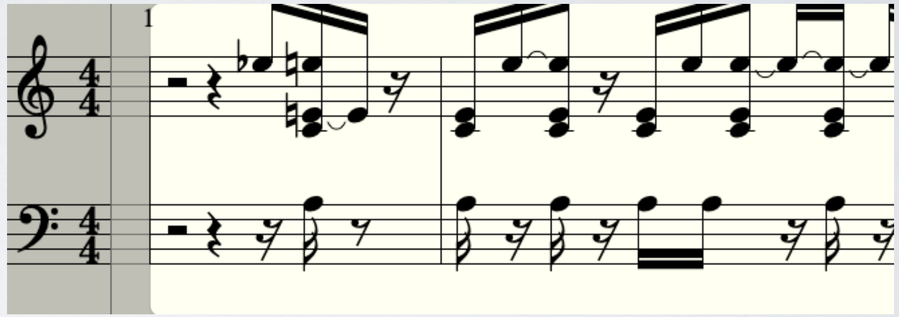
- Collected MusicXML and corresponding performance MIDI from public dataset

What is MusicXML?




```
<note default-x="85.76" default-y="-10.00">
  <grace/>
  <pitch>
    <step>D</step>
    <alter>1</alter>
    <octave>5</octave>
    </pitch>
  <voice>1</voice>
  <type>16th</type>
  <accidental>sharp</accidental>
  <stem>up</stem>
  <staff>1</staff>
  </note>
```

MusicXML



MIDI without Metric Info

- XML format for human-readable music score
- Include various information such as metric information, slur, rest, tempo and dynamic directions

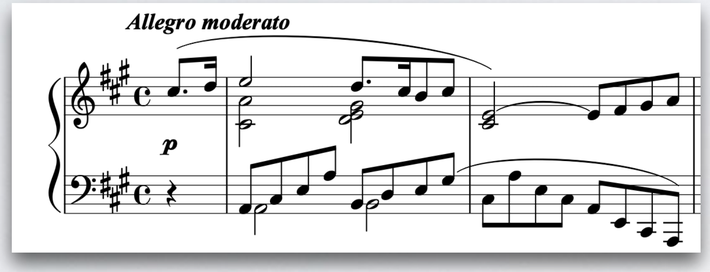
Score

MusicXML

Performance Generation System

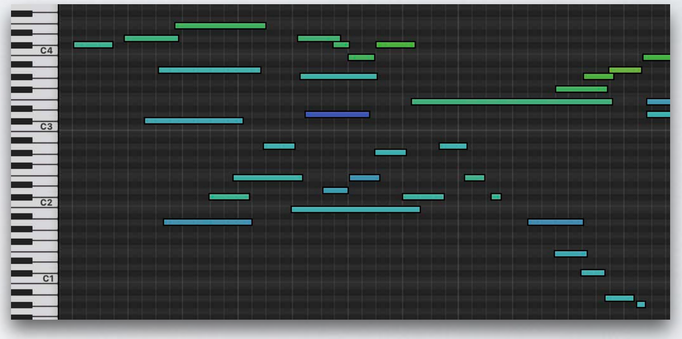
Performance

MIDI



Music Score

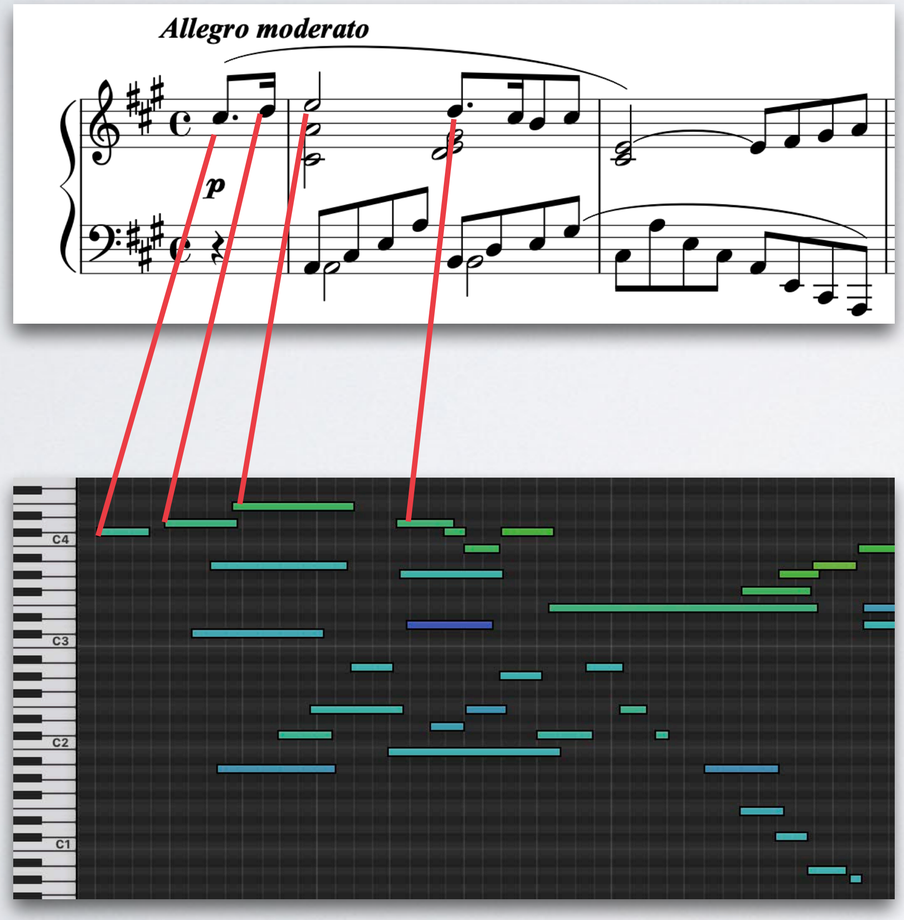
As a pair



Human Performance

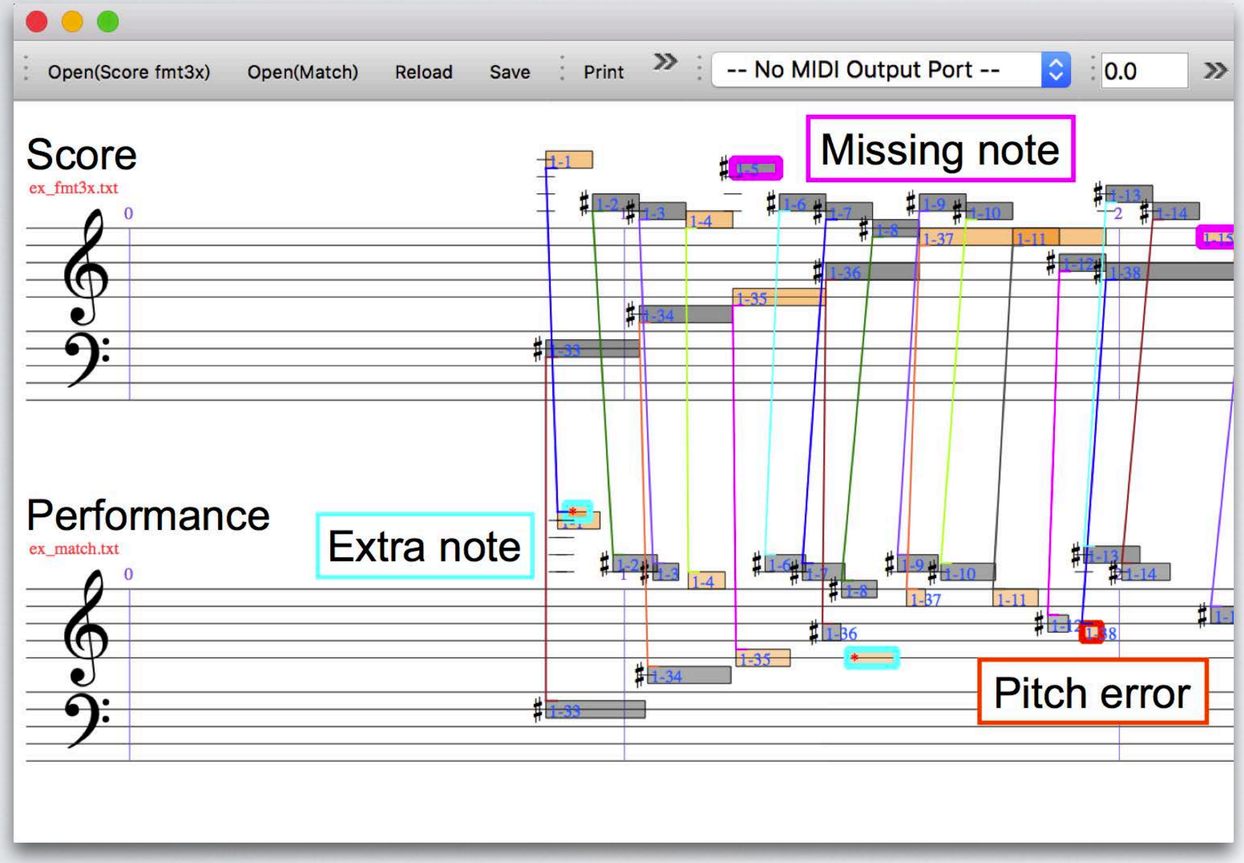
- We need pairs of music score and human performance for the training

Score-Performance Alignment



- Each note in the score has to be aligned with the performance note
- Human performance includes extra or missing notes

Score-Performance Alignment



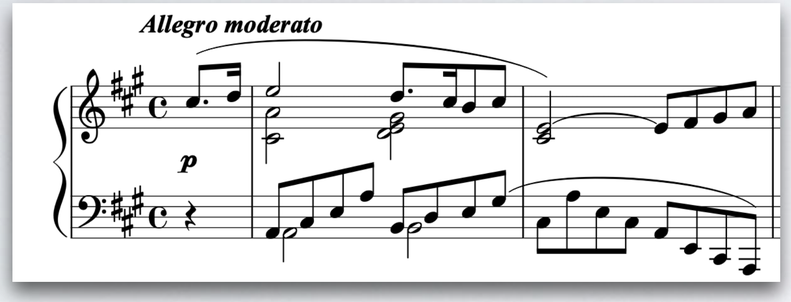
- Employ automatic alignment algorithm by Nakamura
- Additional refinement with rule-based approach

Nakmura et al., Performance Error Detection and Post-Processing for Fast and Accurate Symbolic Music Alignment, 2017

Dataset

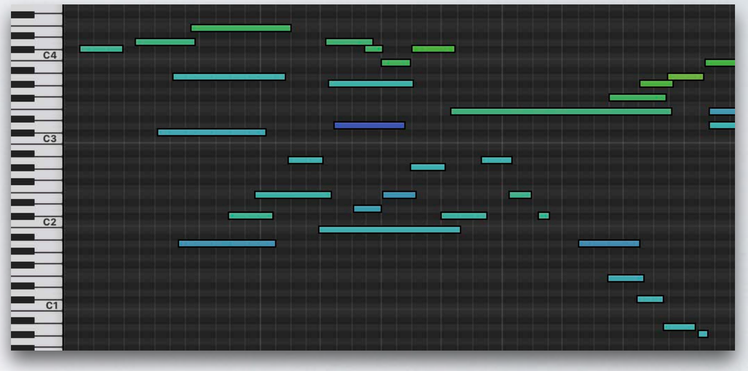
- 16 composers
- 226 pieces
- 1,052 performances
- 666,918 score notes
- 3,547,683 performance notes
  - 131,095 notes were failed to aligned with score notes
  - 114,914 notes were excluded during the refinement
  - 10 times larger than the previous dataset (Magaloff Dataset)

Simplifying the Task



MusicXML

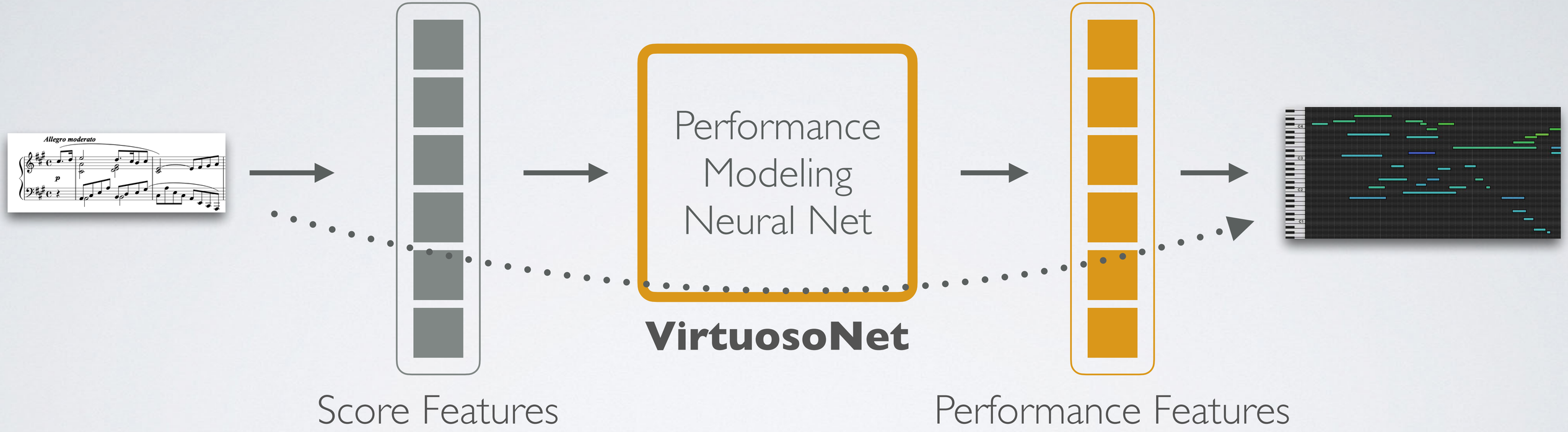
Performance Generation System



MIDI

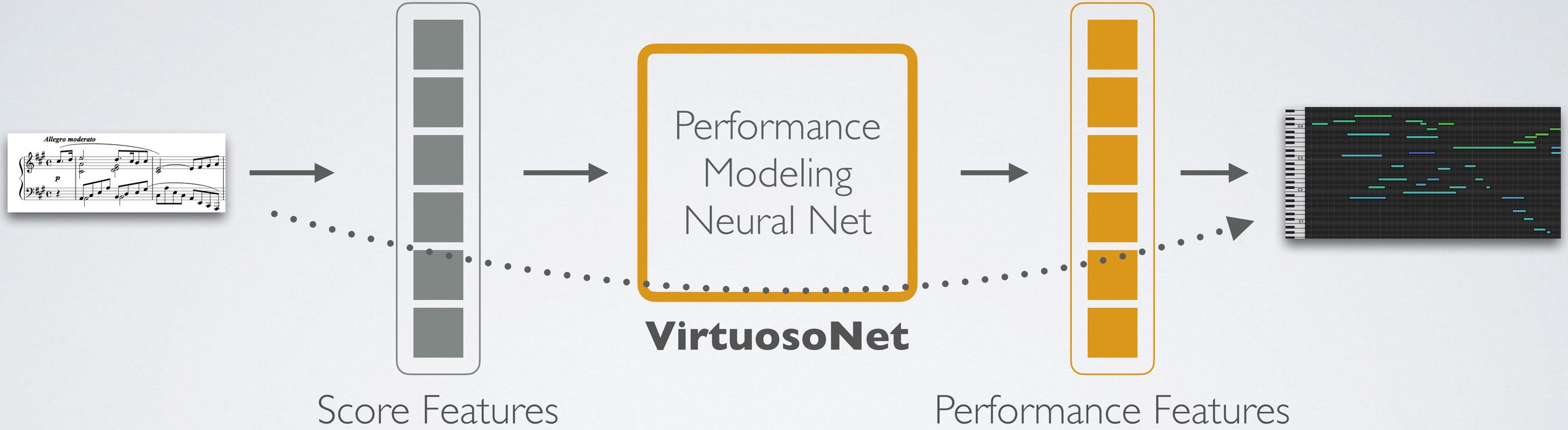
- Generating entire MIDI from scratch demands a system to understand how to "read" music, or converting notes in MusicXML into MIDI

Simplifying the Task



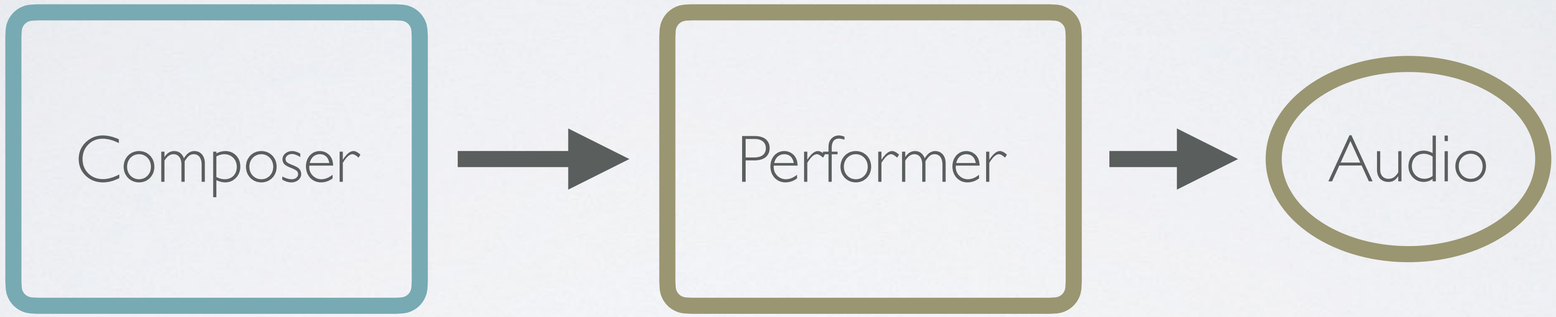
- Modeling system predicts performance features of each note from a given sequence of score features of each note

Simplifying the Task



- Neural network can focus only on modeling higher-level expression

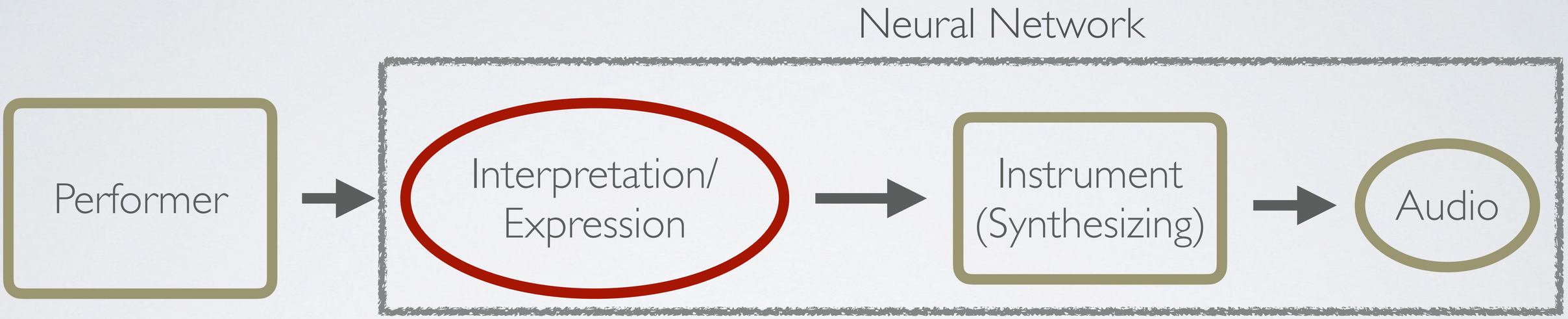
Dataset and Feature



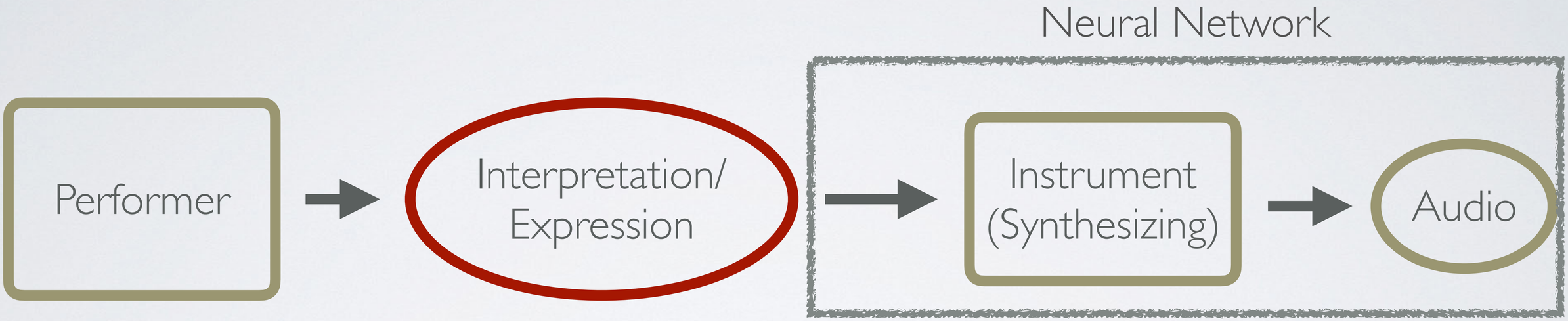
Why do we not directly model Audio of the performance?

Dataset and Feature

Dataset and Feature

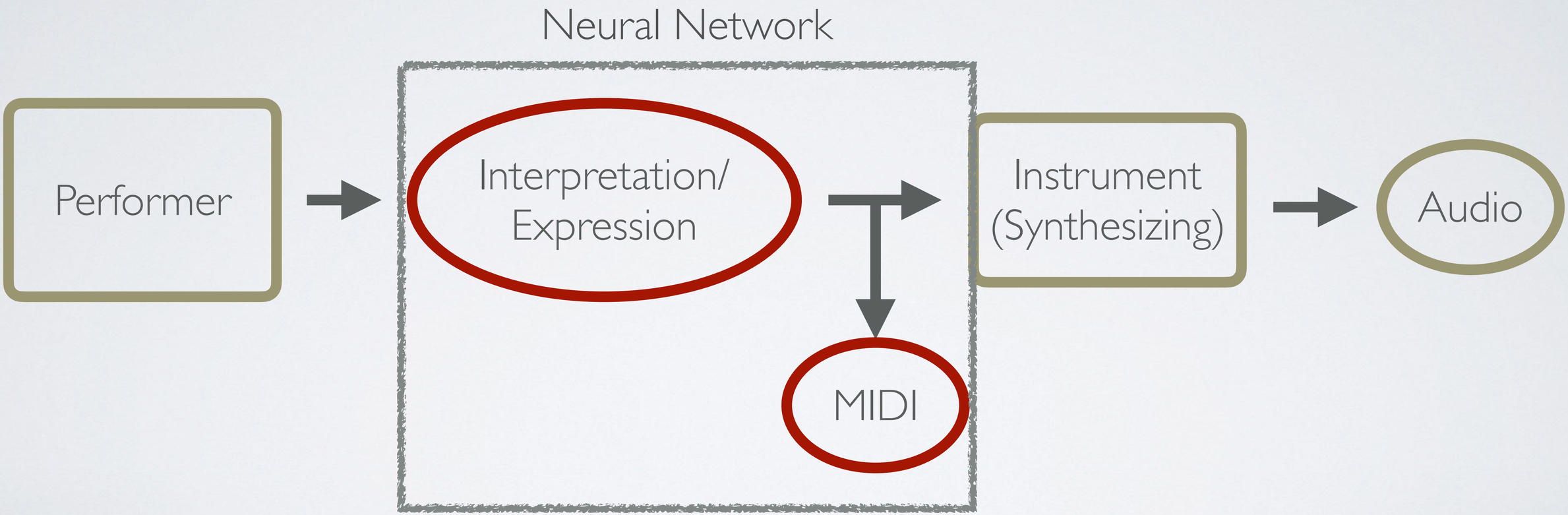


Dataset and Feature



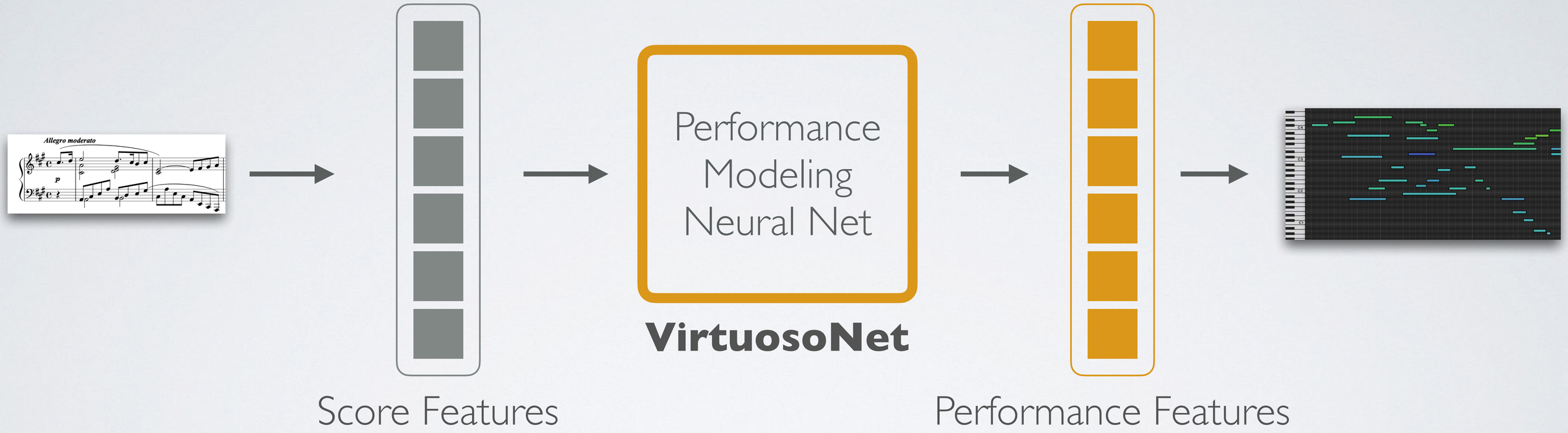
Neural network will use most of its capacity to learn acoustics

Dataset and Feature



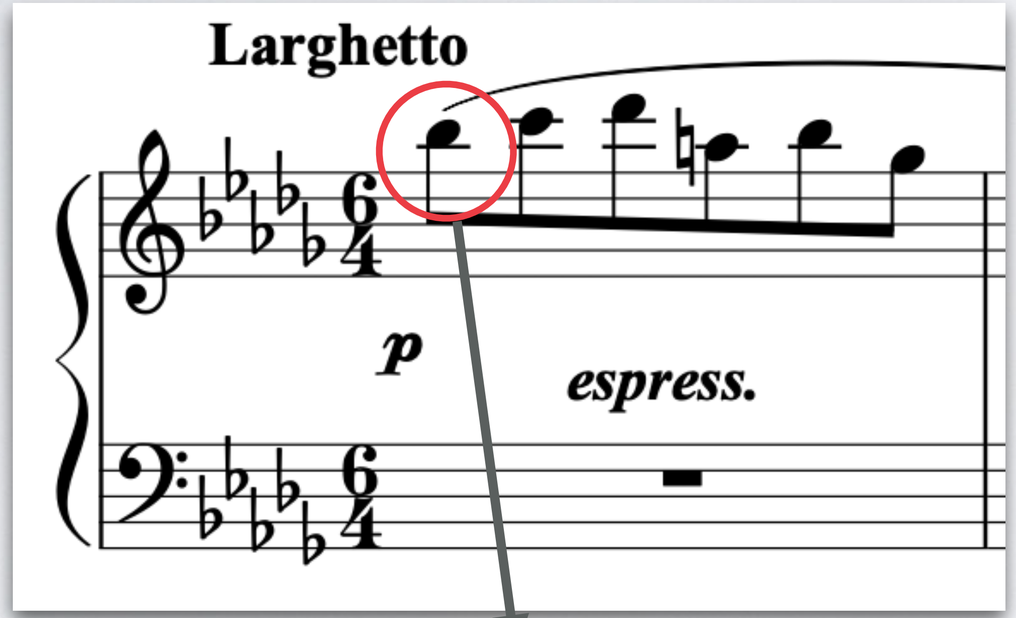
By using MIDI, we can focus more on performer's interpretation

Defining Features



- How do we define score and performance features

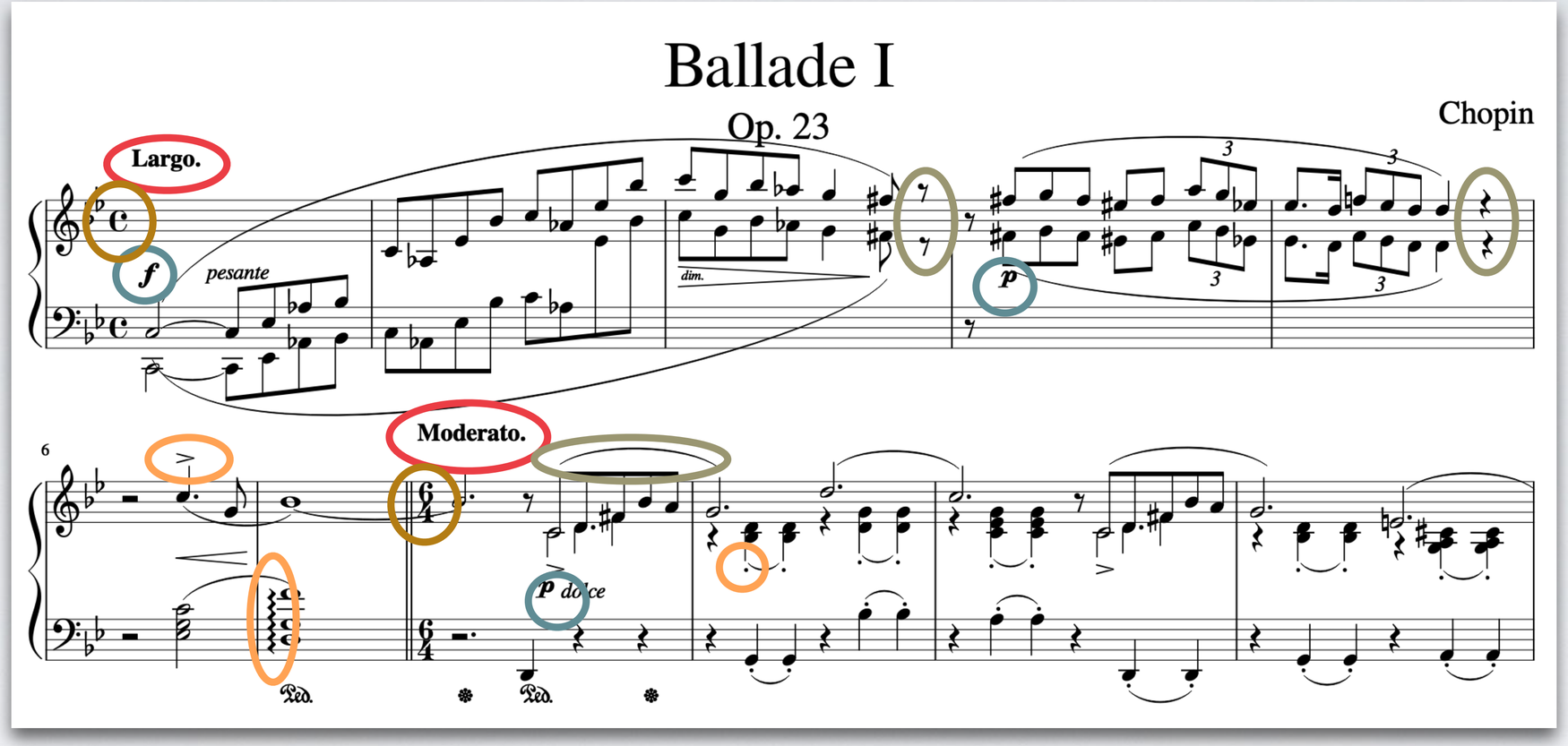
Score Features



Pitch: Bb5,
Duration: Eighth note
Metric Position: 0.5
Articulation: None

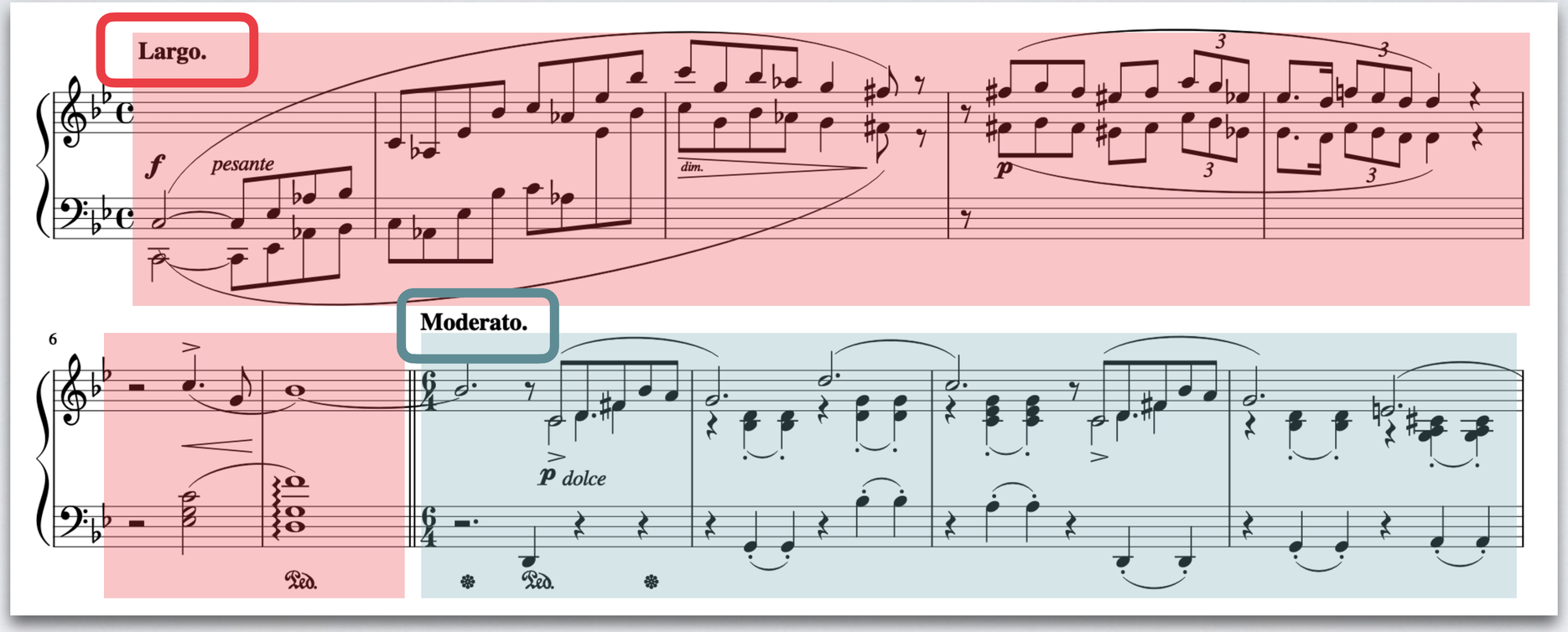
- Note-level
  - Pitch
  - Duration
  - Metric position
  - Articulation

Complexity of Music Score



- There are a lot of types of symbols in music score
- Each of them has a different effect range

Complexity of Music Score



- There are a lot of types of symbols in music score
- Each of them has a different effect range

One of the Solutions

Sequence of different types of symbol

Pitch

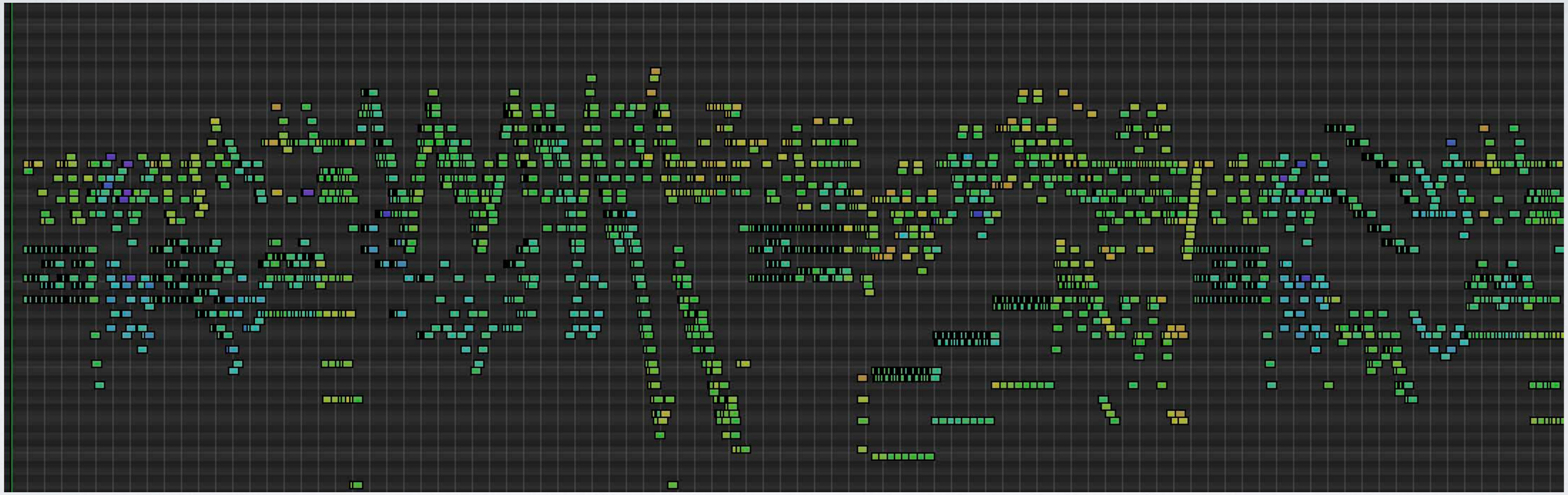
Duration

Articulation

Ornament

Problem of Data size

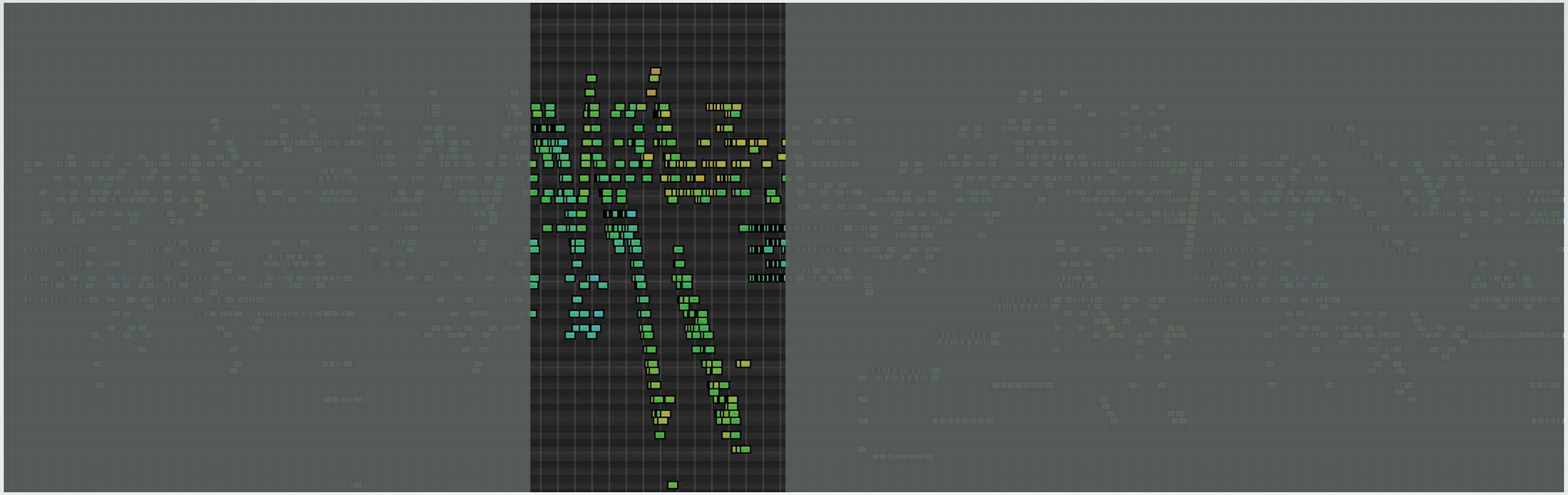
Allegro



We cannot train a neural network with too long sequence

Problem of Data size

~~Allegro~~



Slice the data

Our Idea: Note-level Encoding

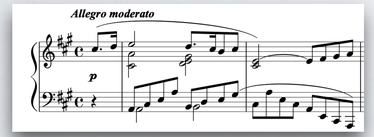
Our Idea: Note-level Encoding

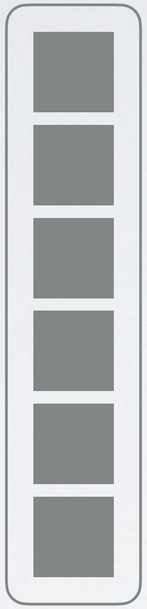
Encode every information in note-level

Our Idea: Note-level Encoding

- Advantages:
  - Can be sliced in any position
  - One-to-one mapping between score features and performance features

Defining Features

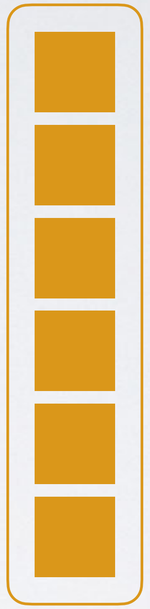




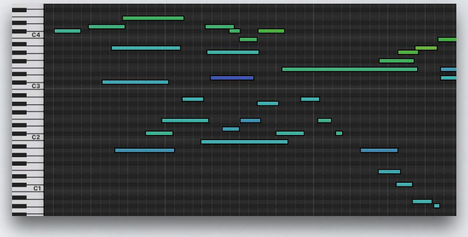
Score Features

Performance Modeling Neural Net

VirtuosoNet

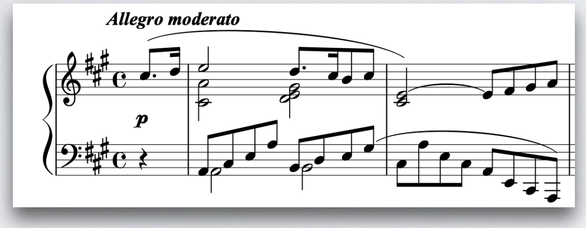


Performance Features

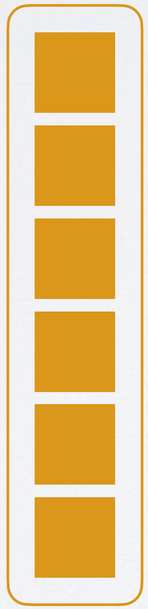


Defining Features

MusicXML

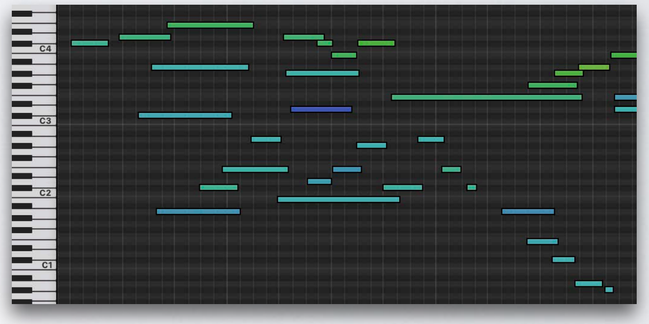


What to play



How to play

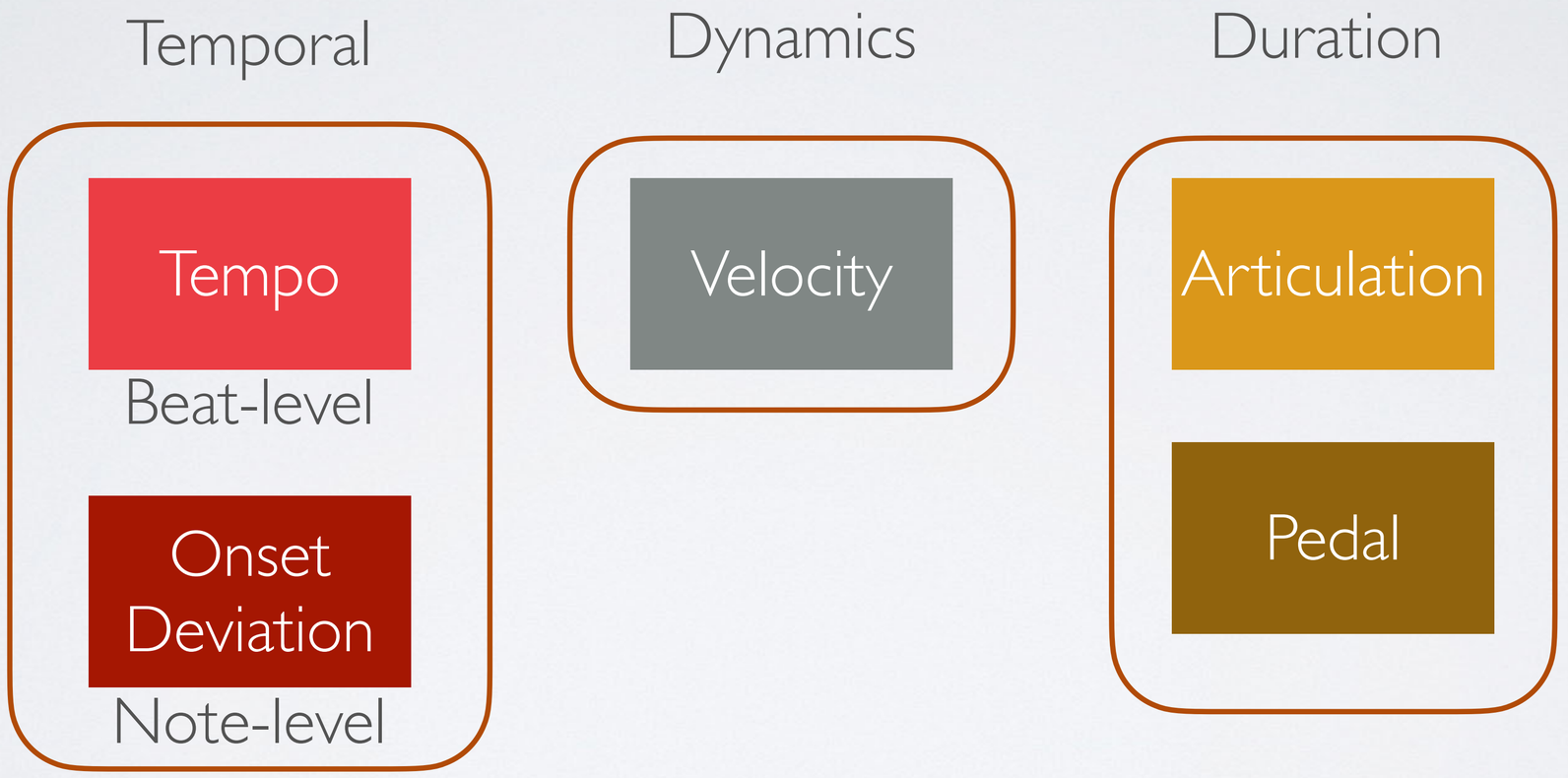
Rule-based Decoding



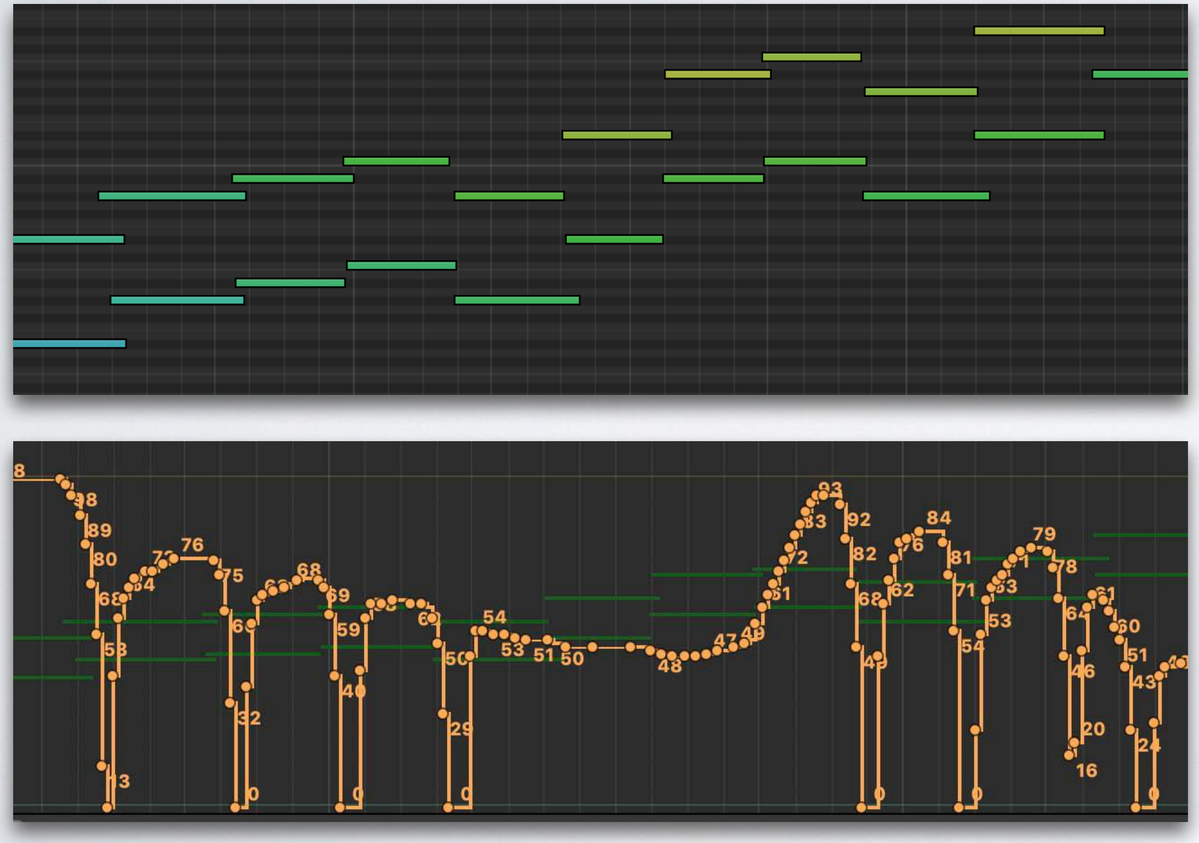
Performance MIDI

Performance Features

Performance Features for Piano



Pedaling



Pedaling changes continuously, regardless of note events

https://youtu.be/xwYBBWFDZRA

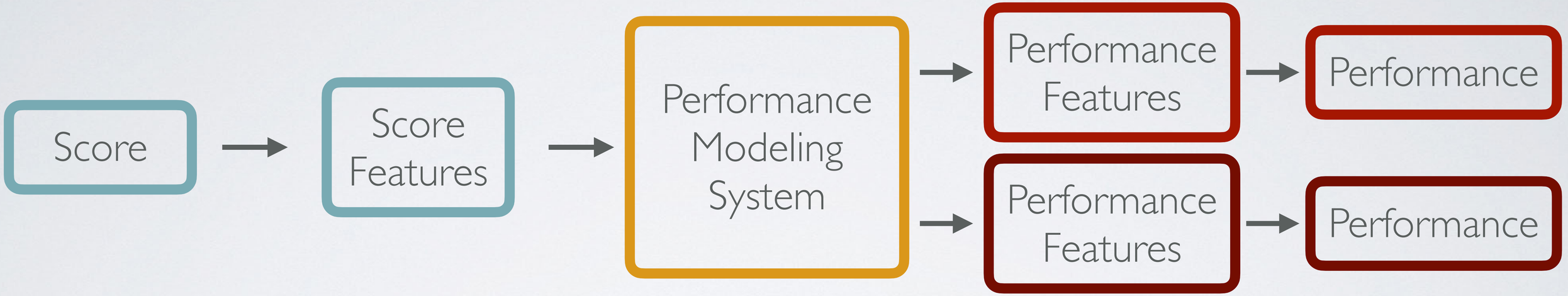
Note-level Pedal Encoding

1. Pedal value at onset
2. Pedal value at offset
3. Lowest pedal value within note
4. Elapsed time between note onset and 3)
5. Lowest pedal value after note offset and a new note onset
6. Elapsed time between note offset and 5)
7. Soft pedal value at start

System Structure

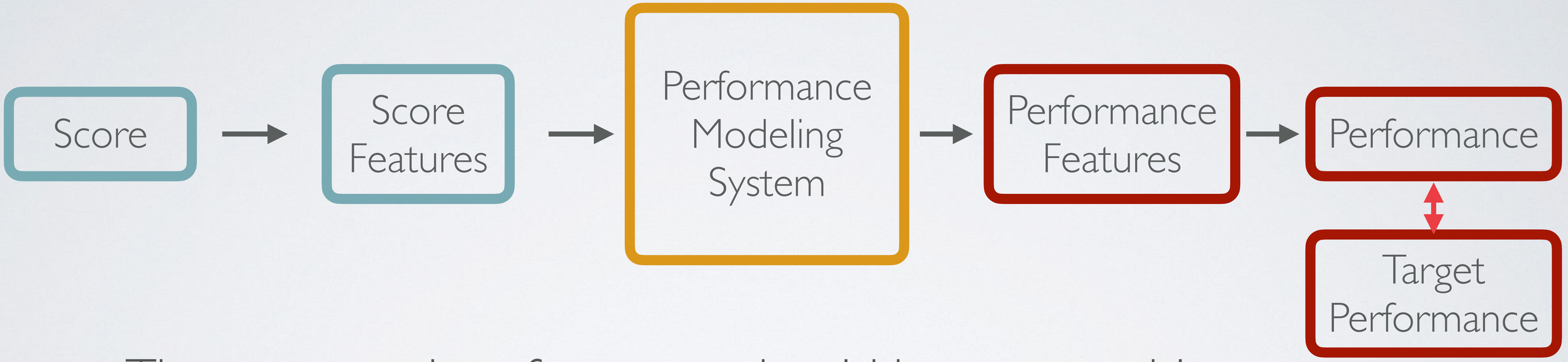
- What kind of problems do we have to consider?

System Structure



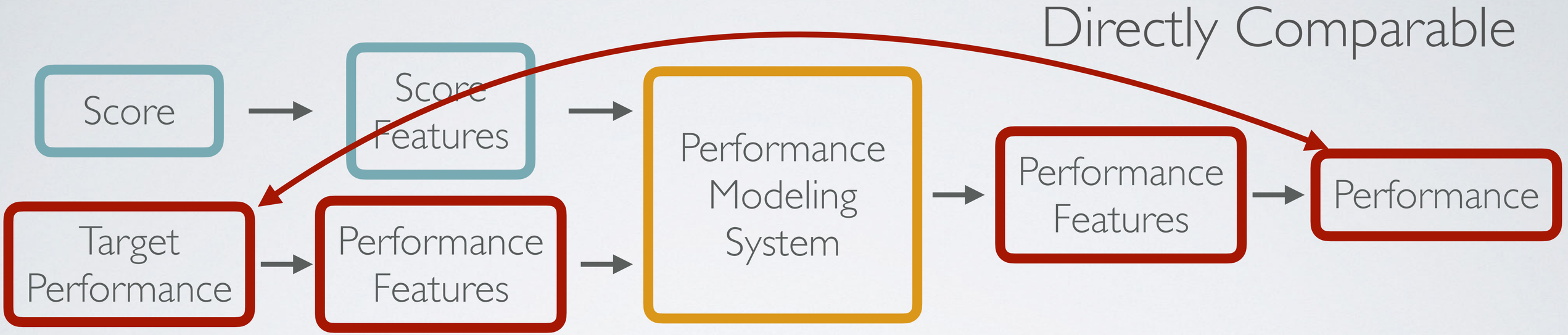
- There can be various styles of performance for given score
  - The system has to be able to handle the variety

System Structure



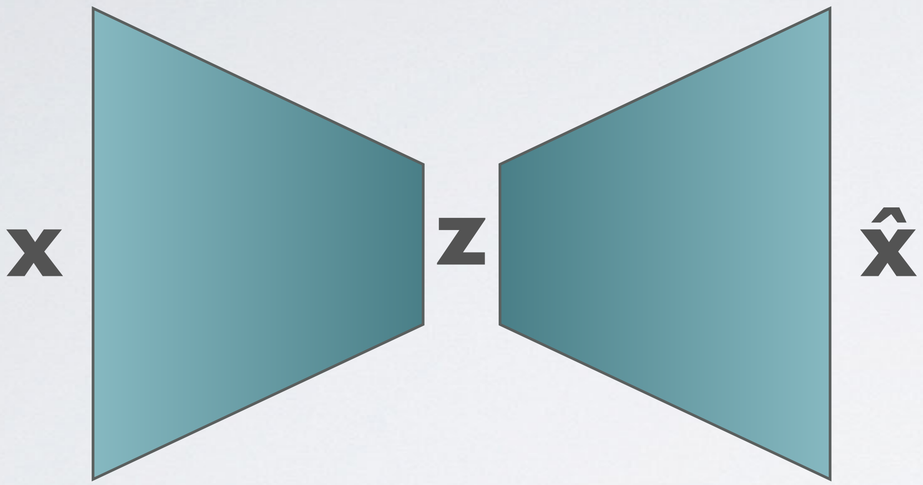
- The generated performance should be comparable to a target performance.

System Structure

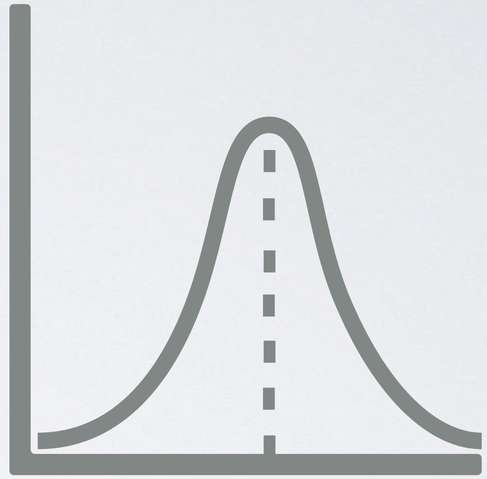


- The generated performance should be comparable to a specific target performance for quantitative evaluation
- The system has to know about the target performance

System Structure



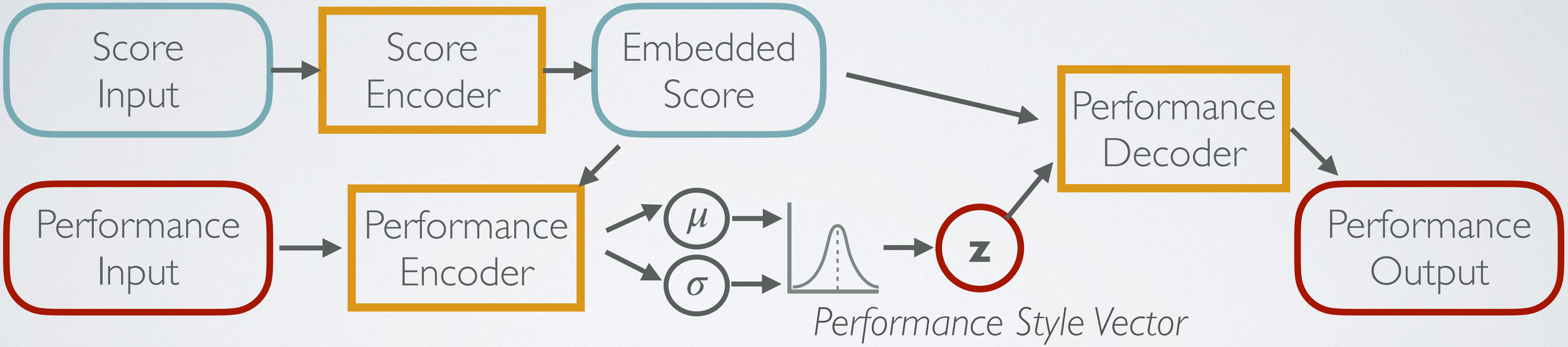
Autoencoder



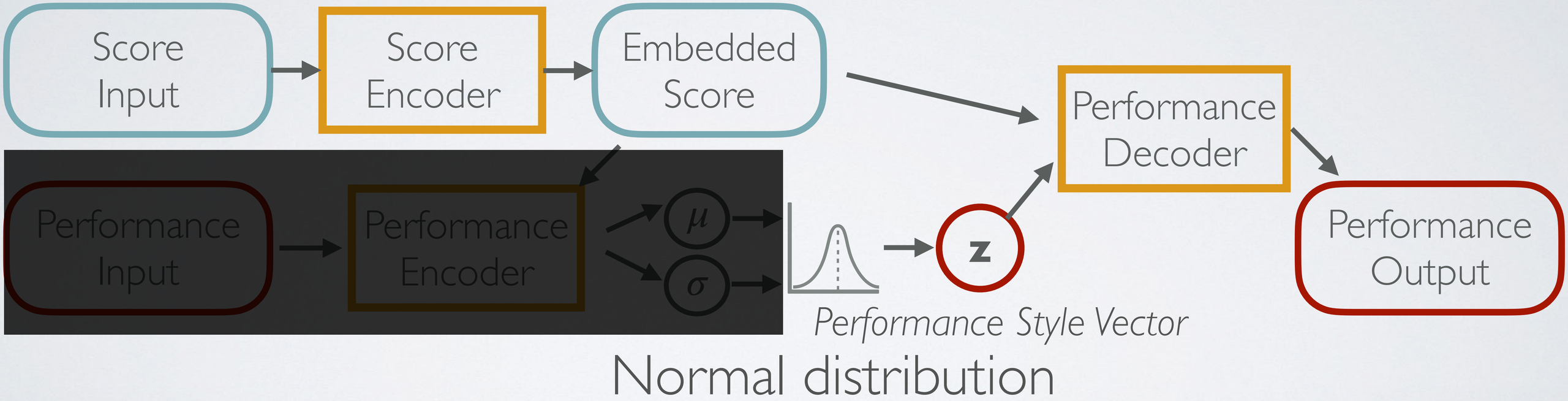
Variational Inference

- Conditional Variational Autoencoder (CVAE)
- Score as a condition, performance as an input and output

System Structure

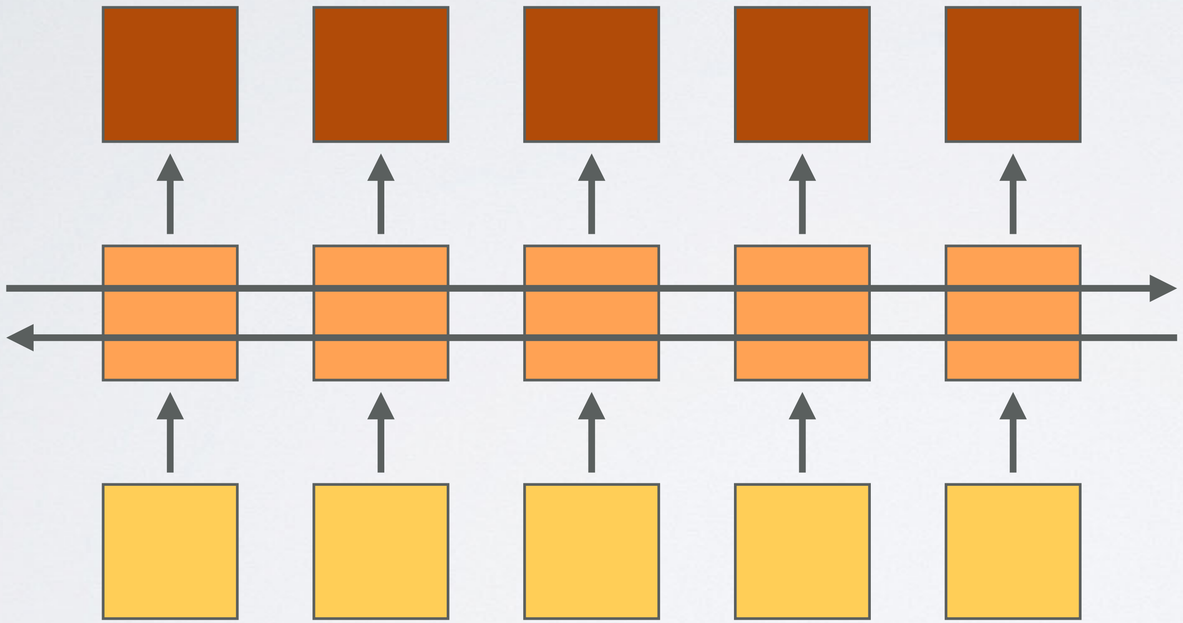


System Structure



- During the inference, we can sample **z** from normal distribution

Score Encoder

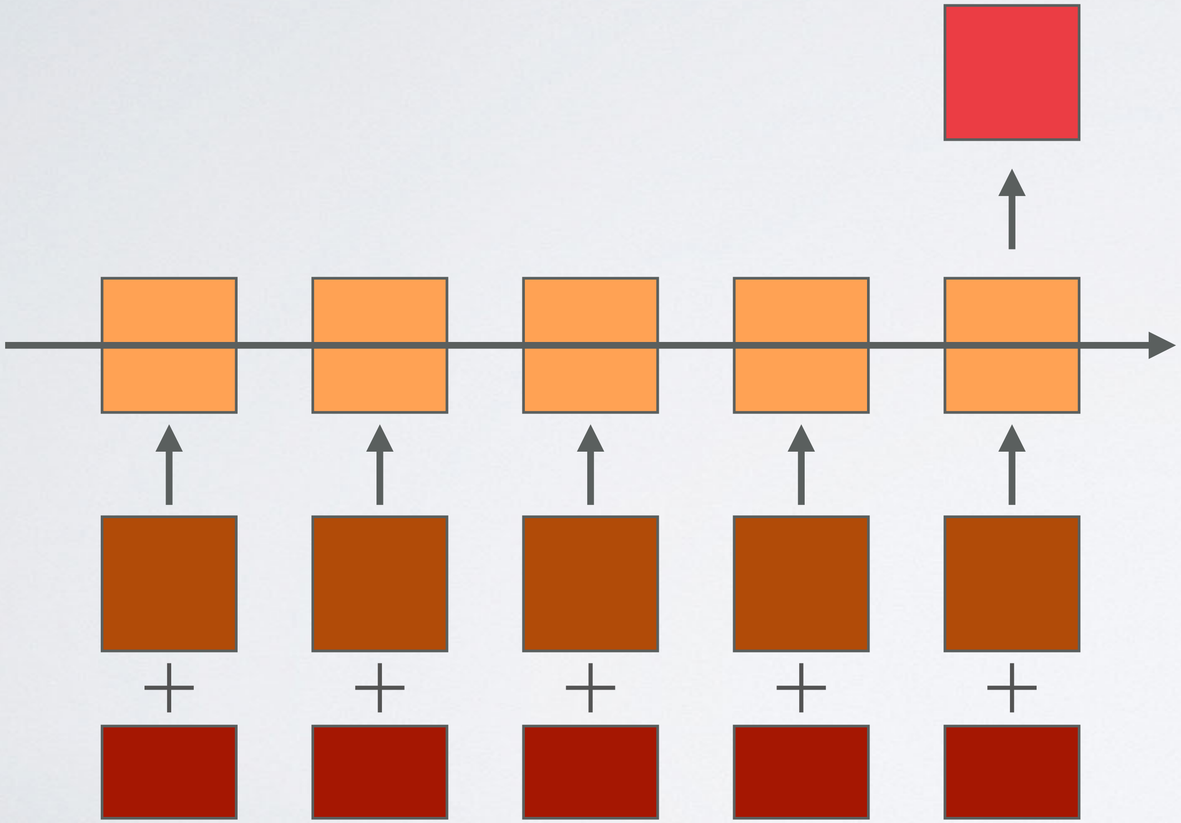


Note-level Representations

Bi-directional LSTM

Score Features

Performance Encoder

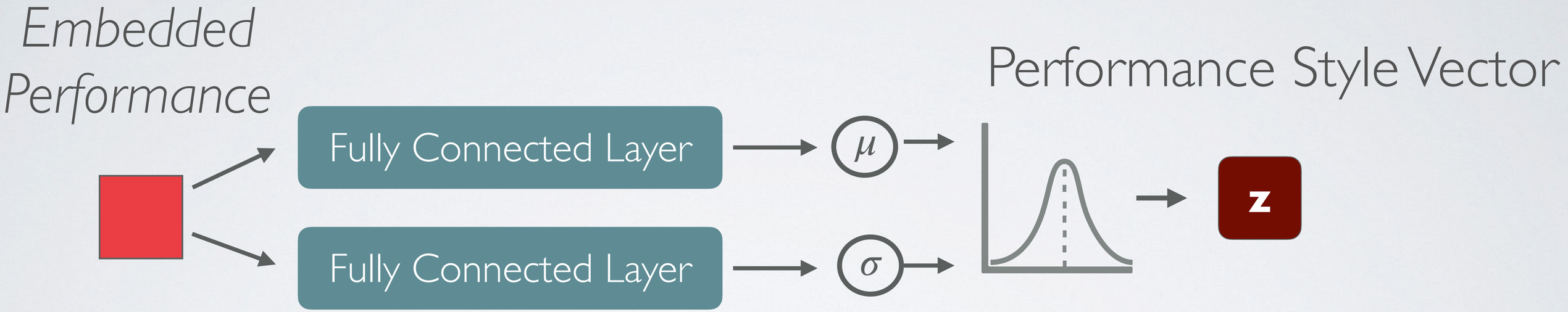


Embedded Performance

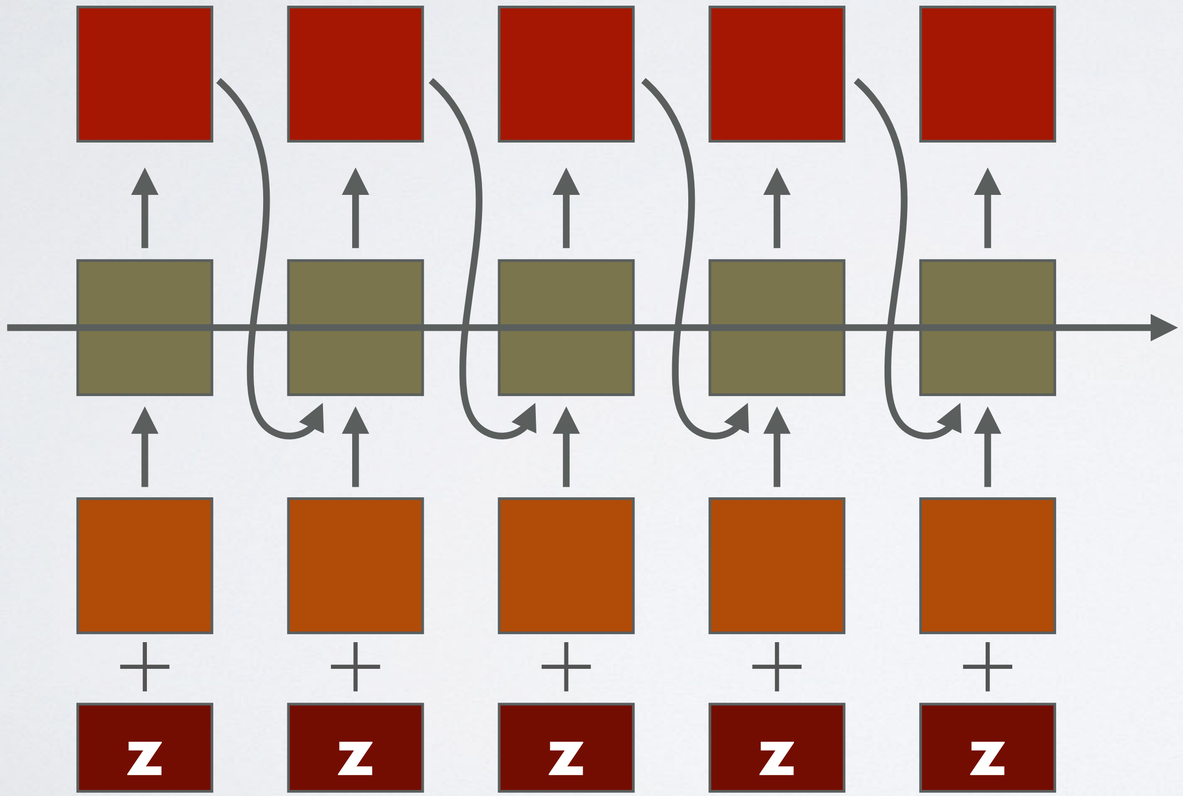
Uni-directional LSTM

Embedded Score
and Performance Features

Performance Encoder



Performance Decoder



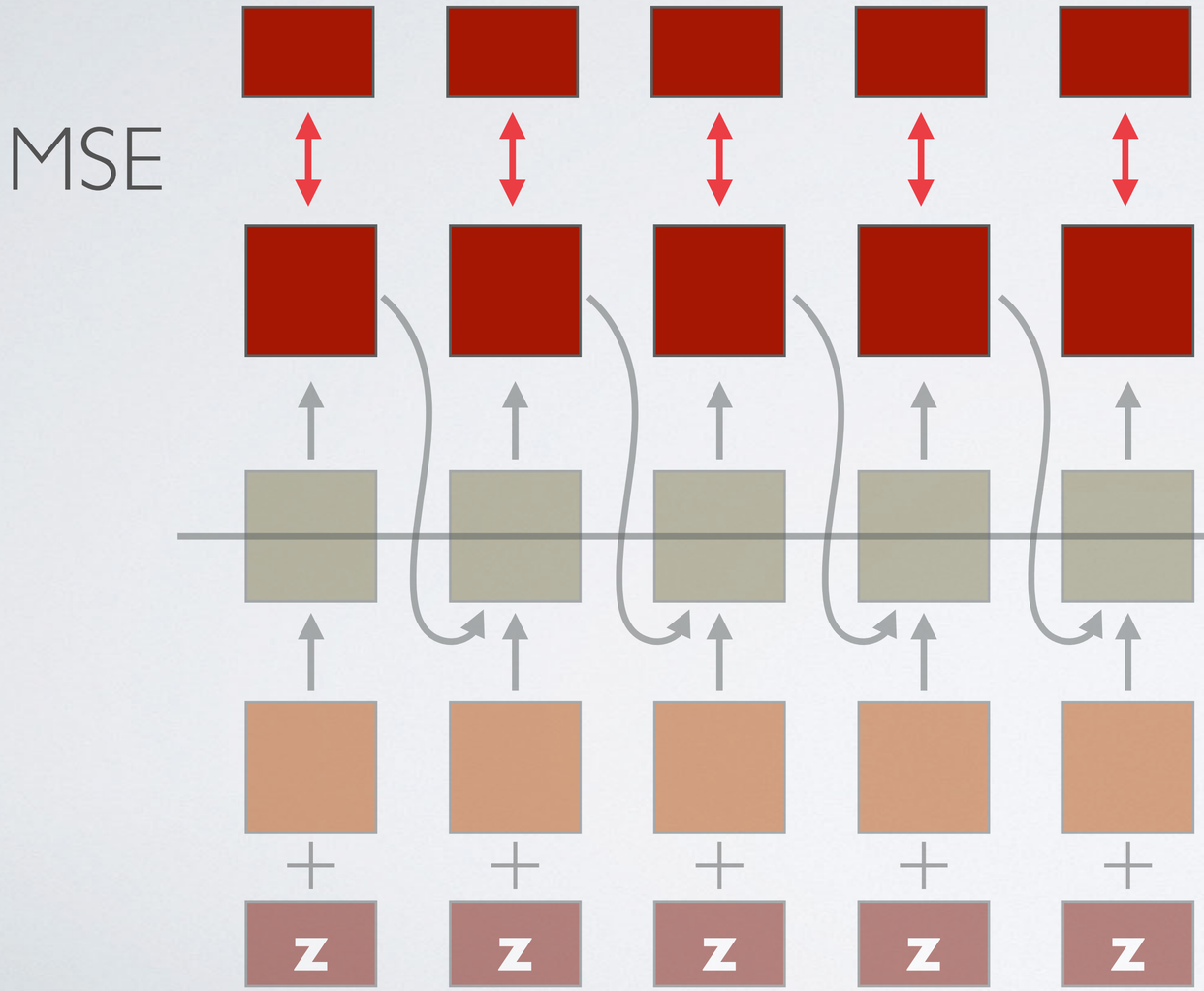
Performance Features

Uni-directional LSTM

Note-level Representations

Performance Style Vector

Loss Calculation



Target performance features

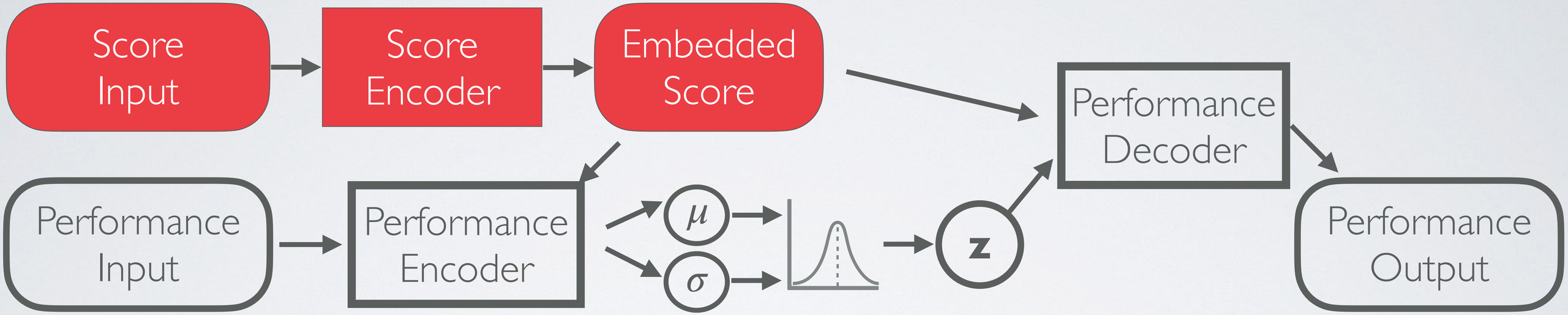
Performance Features

Uni-directional LSTM

Note-level Representations

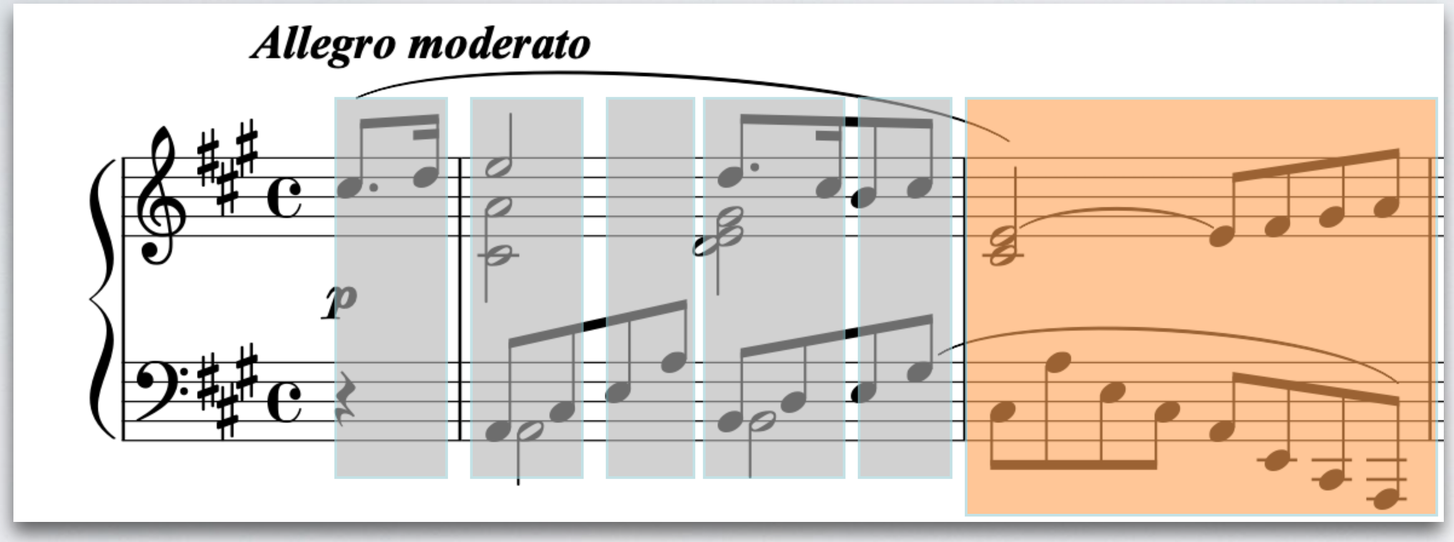
Performance Style Vector

Our Focus



- Understanding a score is the primary object
- Performance is largely decided by the input score
  - Commonality is stronger than diversity

Hierarchical Structure

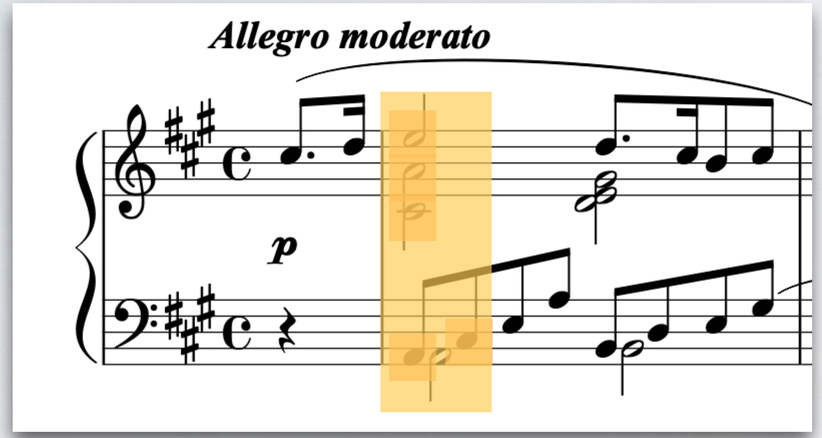


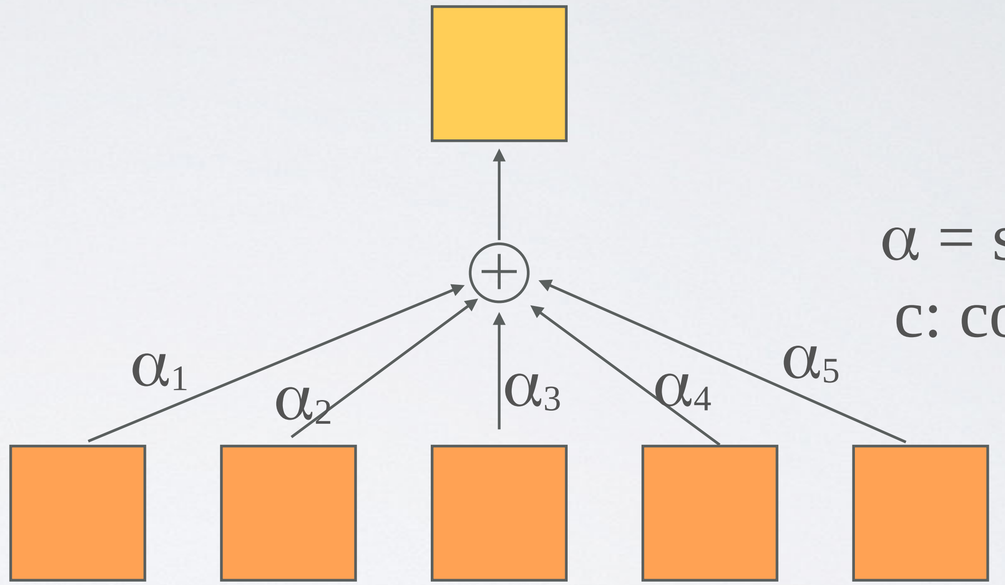
Beat-level

Measure-level

- Music has a clear hierarchical structure
- Beat and measure are units that are decided by the composer

Hierarchical Attention Network



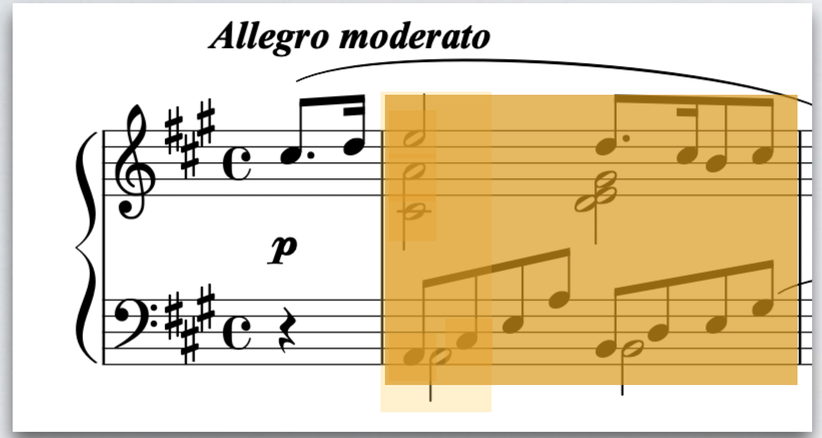


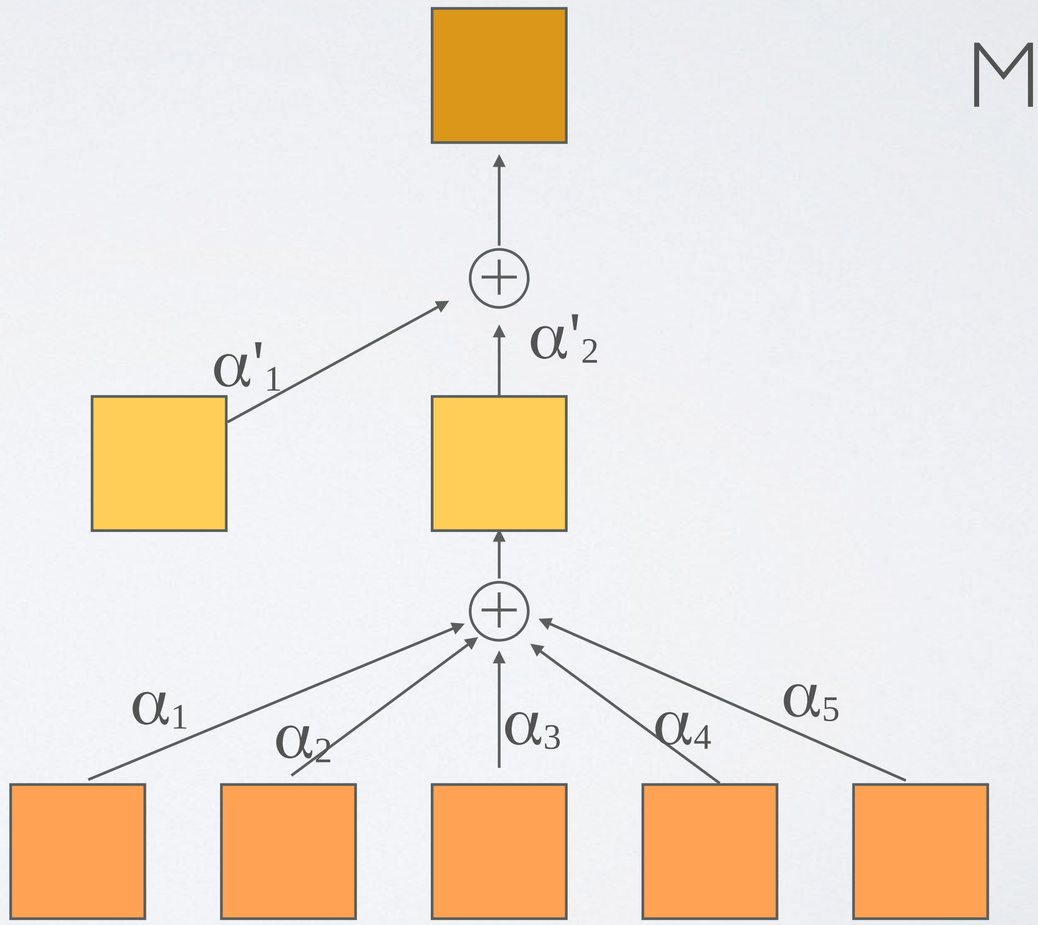
- First proposed for document classification (Yang et al, 2016)

  Hierarchical attention networks for document classification

- Summarize lower-level features by weighted sum

Hierarchical Attention Network



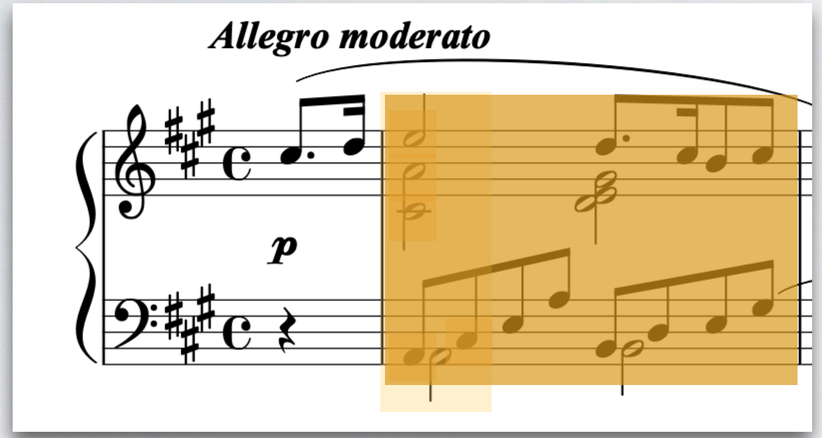


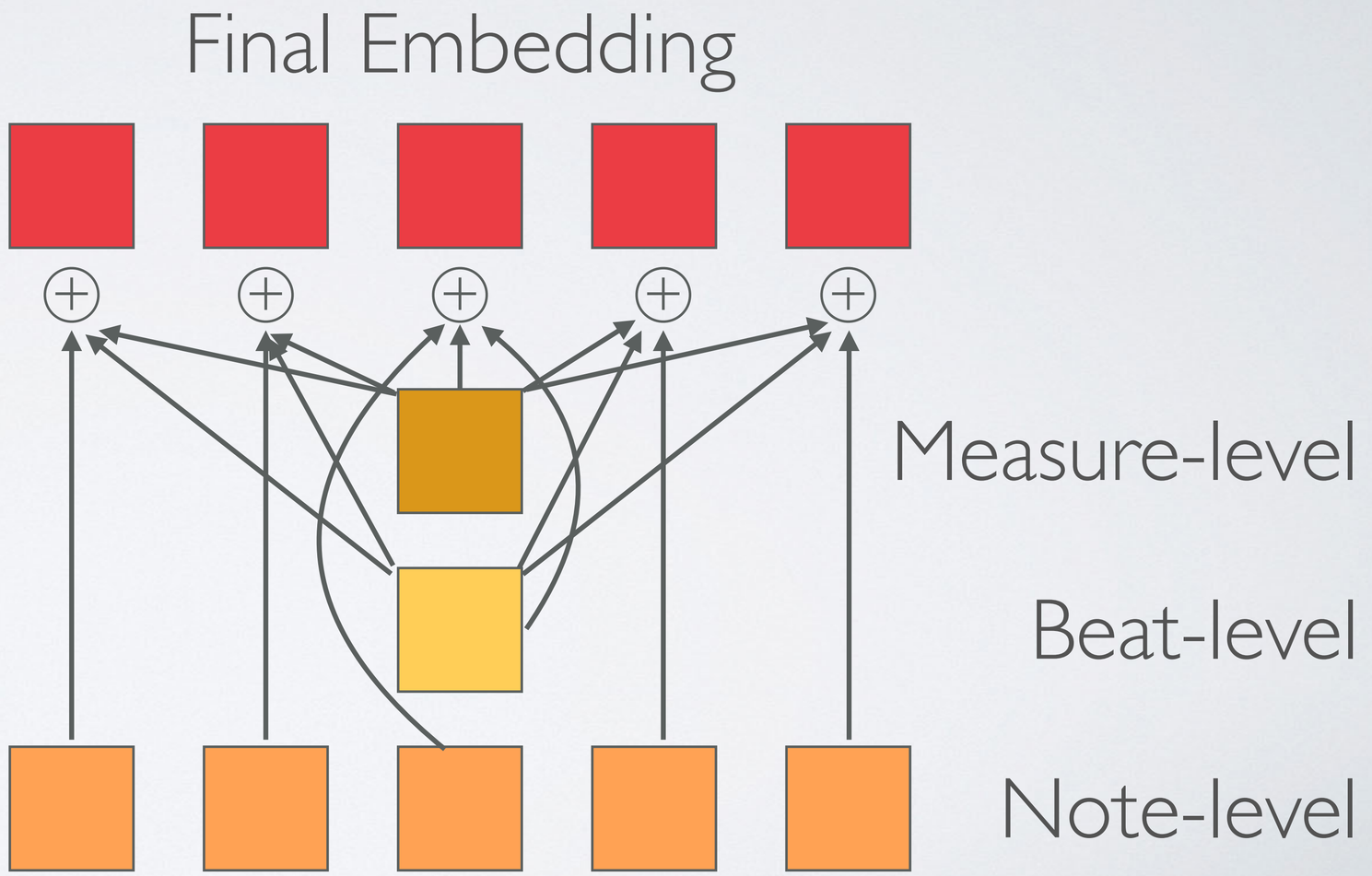
Measure-level

Beat-level

Note-level

Hierarchical Attention Network





Human Listening Test

Subjects

Five students
who are majoring in
piano performance

Pieces

1. Beethoven Piano Sonata No. 5, 1st mov
2. Chopin Etude op. 10-2
3. Schubert Piano Sonata D 664, 1st mov

Performances

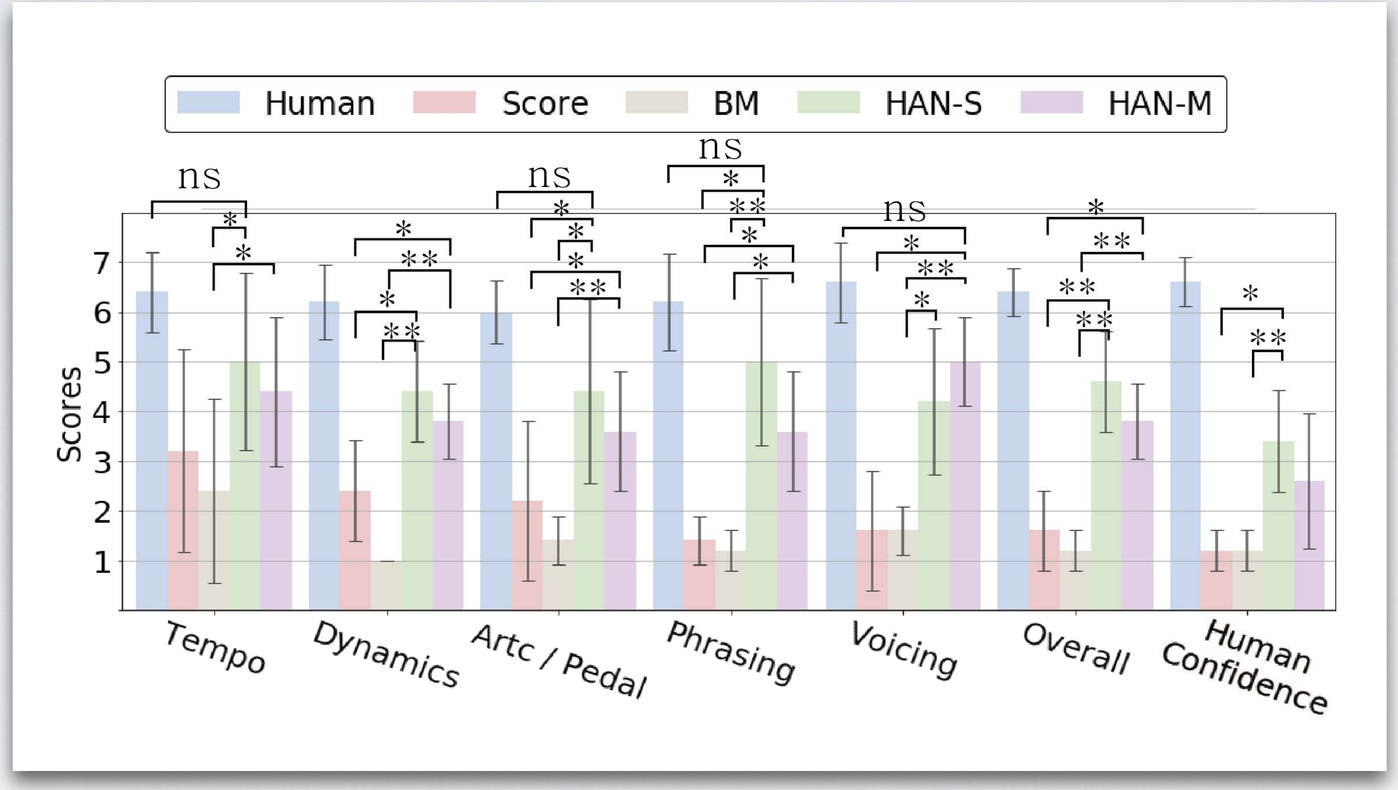
1. Human Performance
2. Direct Rendering
3. Basis Mixer
4. VirtuosoNet Single
5. VirtuosoNet Multi

Criteria

1. Tempo/Agogik
2. Dynamics
3. Articulation/Pedal
4. Phrasing
5. Voicing
6. Human Confidence
7. Overall

- Performances were played by Disklavier

Experiment Results



- Our model performed better than other models

Demo Video

- 1. Deadpan Rendering
  - Consistent tempo, Rule-based dynamics.
    Rendered by a notation program (MuseScore)
- 2. BasisMixer (Cancino-Chacón, 2016):
  - Does not use pedal. Phrasing boundary is little bit unnatural
- 3. VirtuosoNet (Proposed):
  - Use pedal, natural phrasing

1. Introduction

2. Performance Modeling with RNN

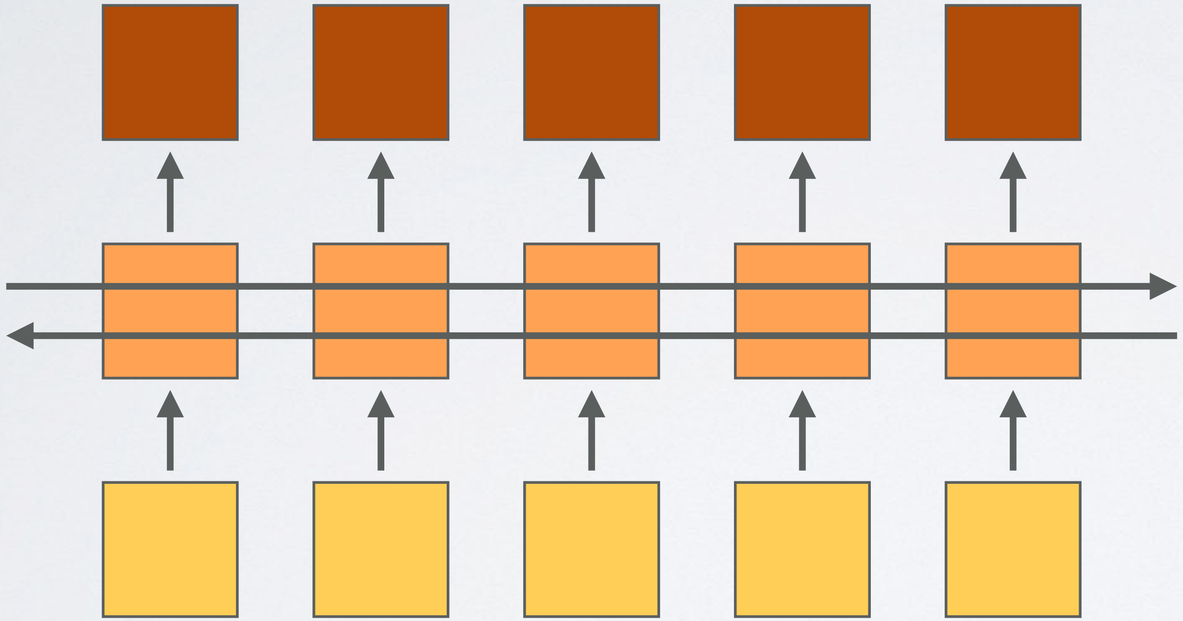
3. Performance Modeling with GNN

4. Performance Style Analysis

5. Future Research

Jeong et al, Graph Neural Network for Music Score Data and Modeling Expressive Piano Performance (ICML, 2019)

Previous Model

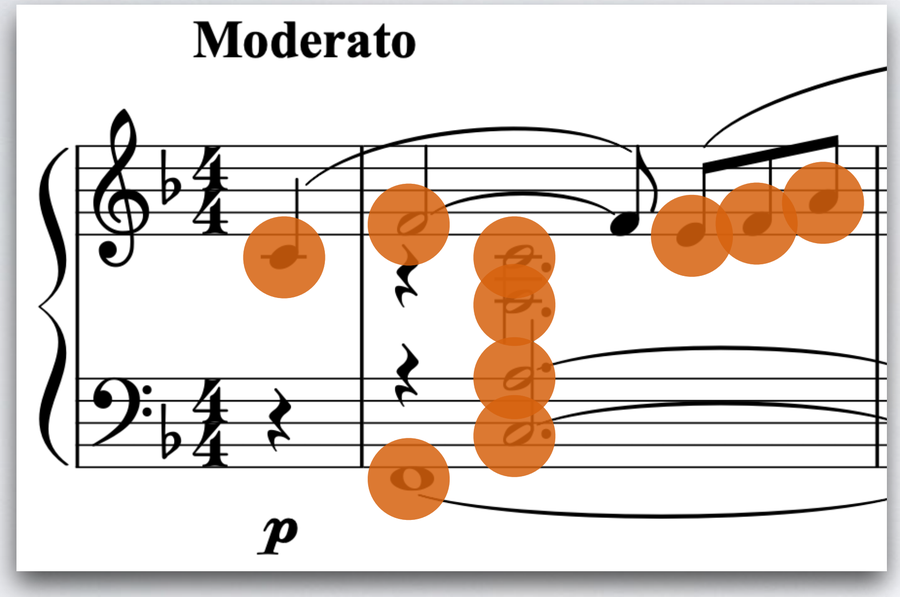


Note-level Representations

Bi-directional LSTM

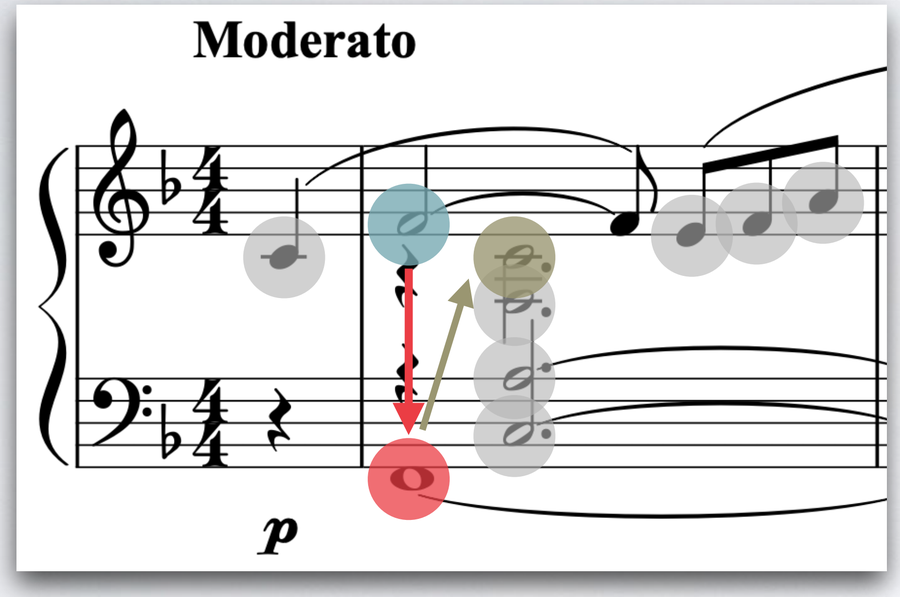
Score Features

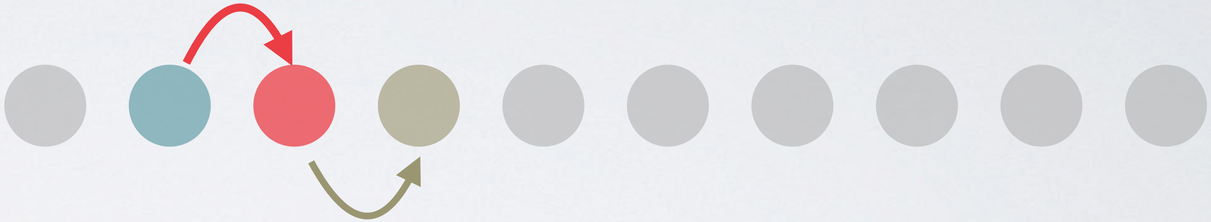
Previous Model



Flatten music score into an 1-D sequence
in order of time and pitch

Word-like Sentence



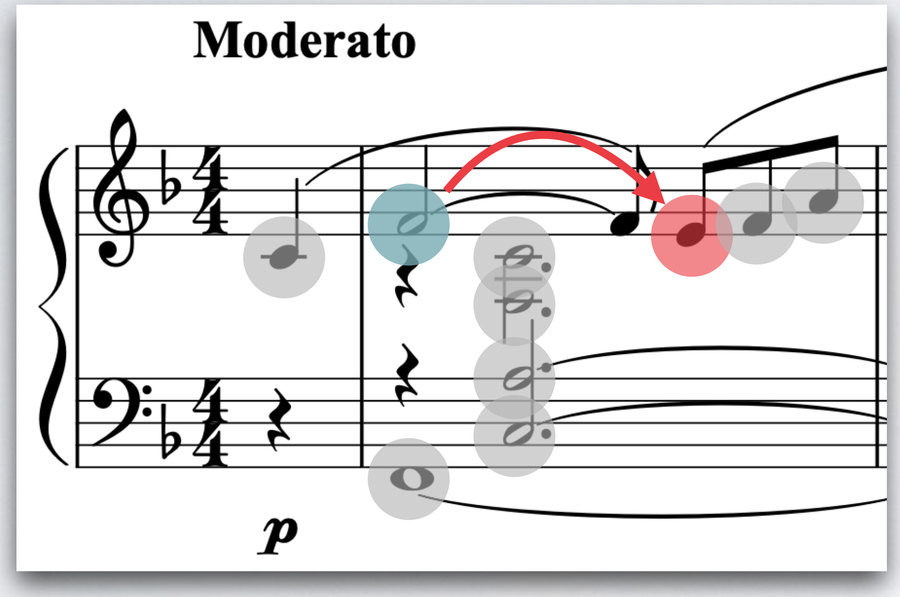


Appears Simultaneously

Appears while previous note is sustaining

The musical relations between adjacent notes in sequence are different.

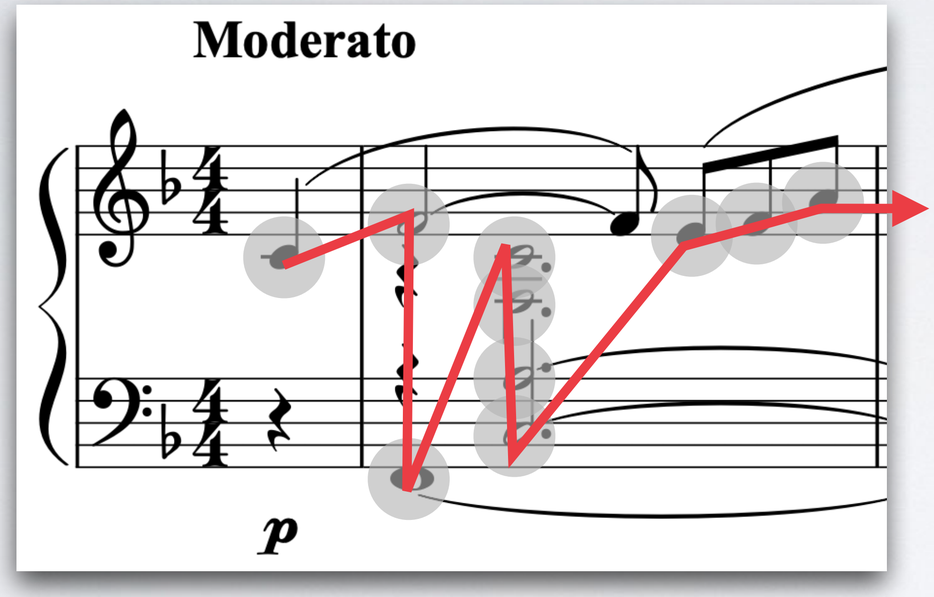
Word-like Sentence



Musically Neighboring Notes

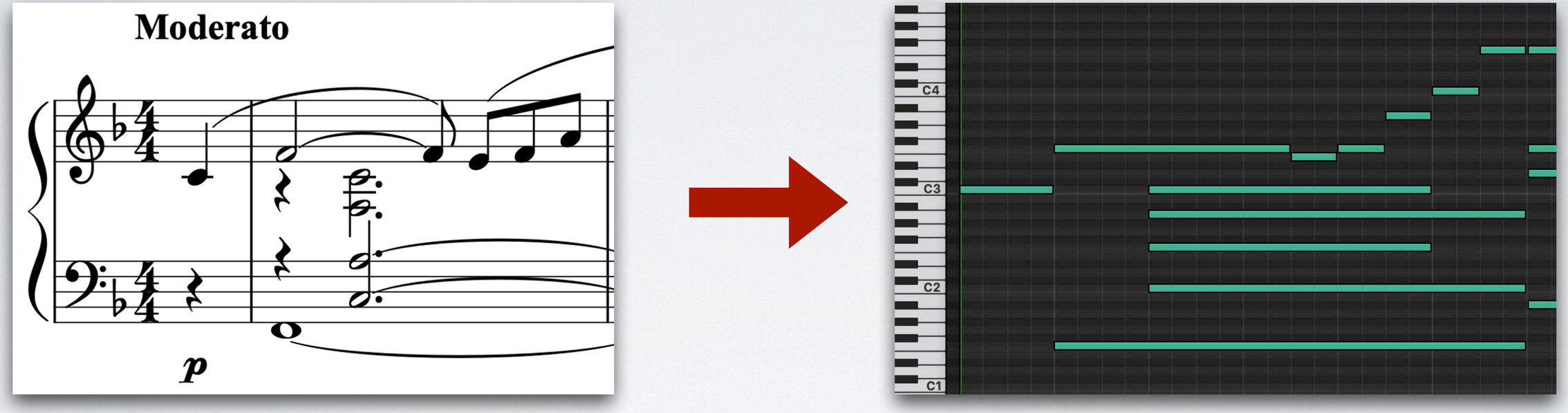
Musically neighboring notes can be located far apart

Word-like Sentence



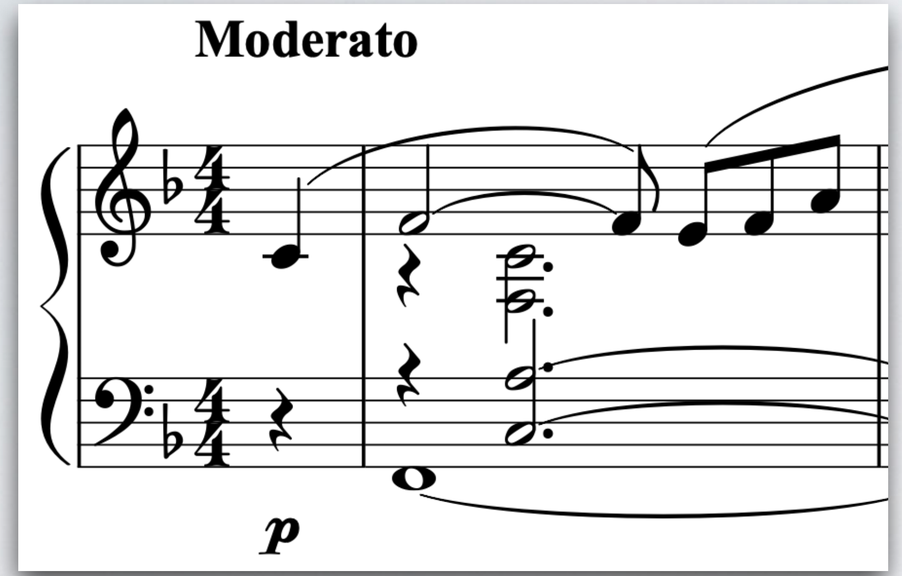
Does a human also read a score in this way?

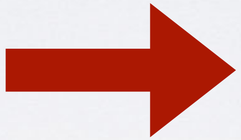
Piano Roll

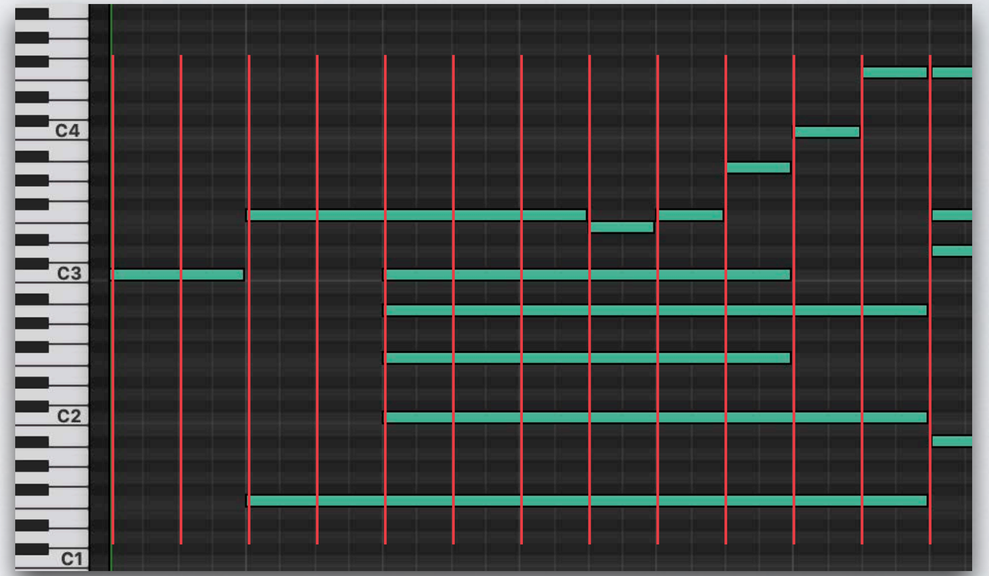


- Convert music score as a 2D matrix of note activation in time and pitch axis (piano-roll)
- Sampling-based representation rather than event-based

Piano Roll

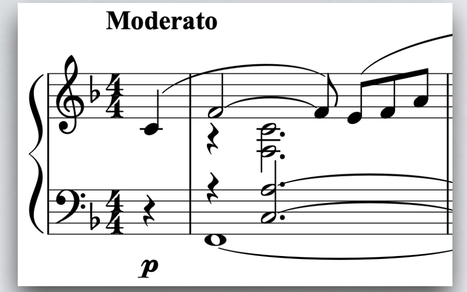


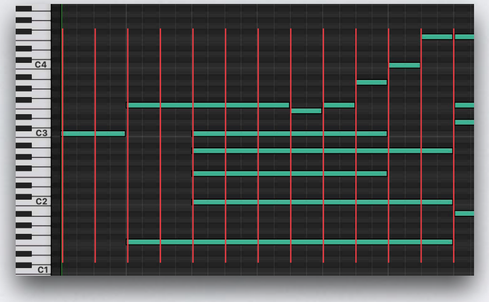




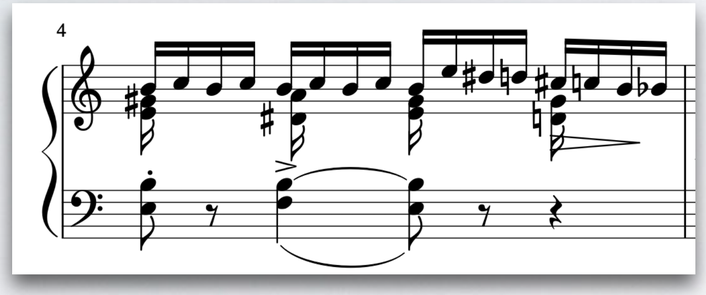
- How can we decide the sampling frequency?

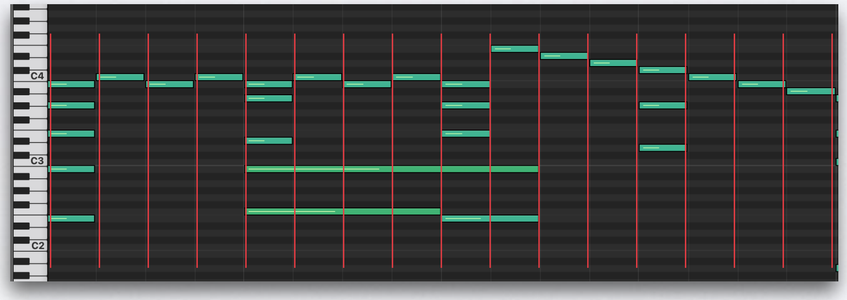
Sampling Frequency for Piano Roll



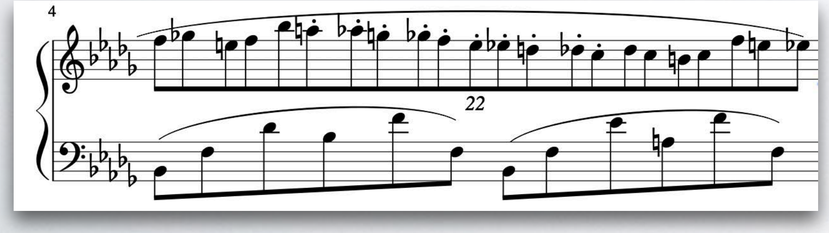


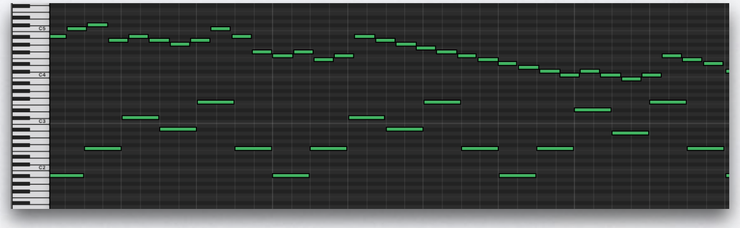
8 frames per measure





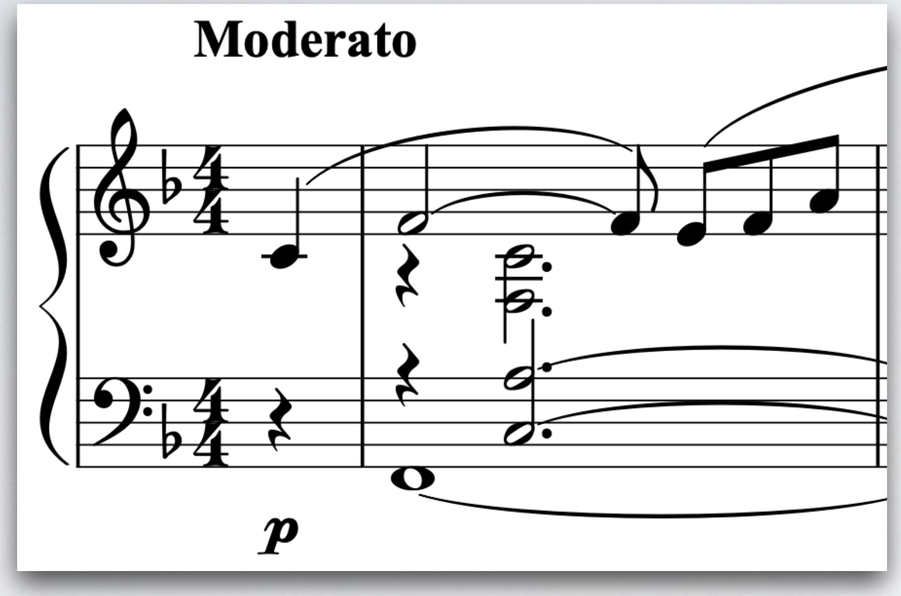
16 frames per measure

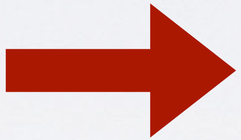


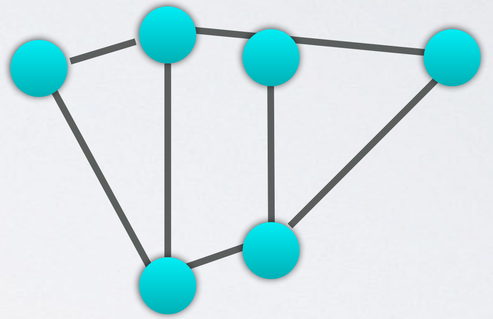


66 frames per measure

Music Score as a Graph

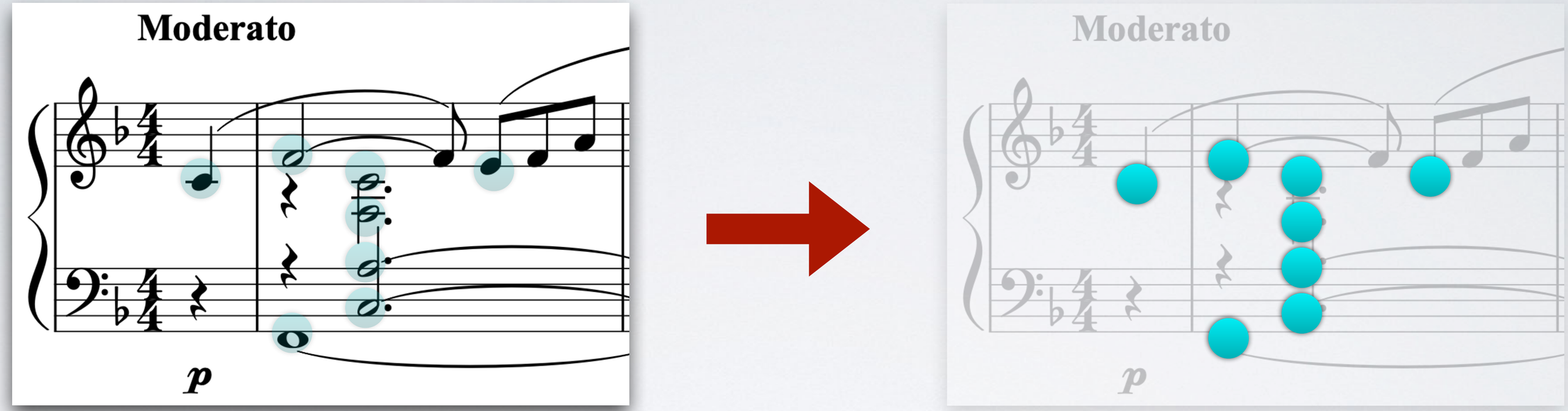






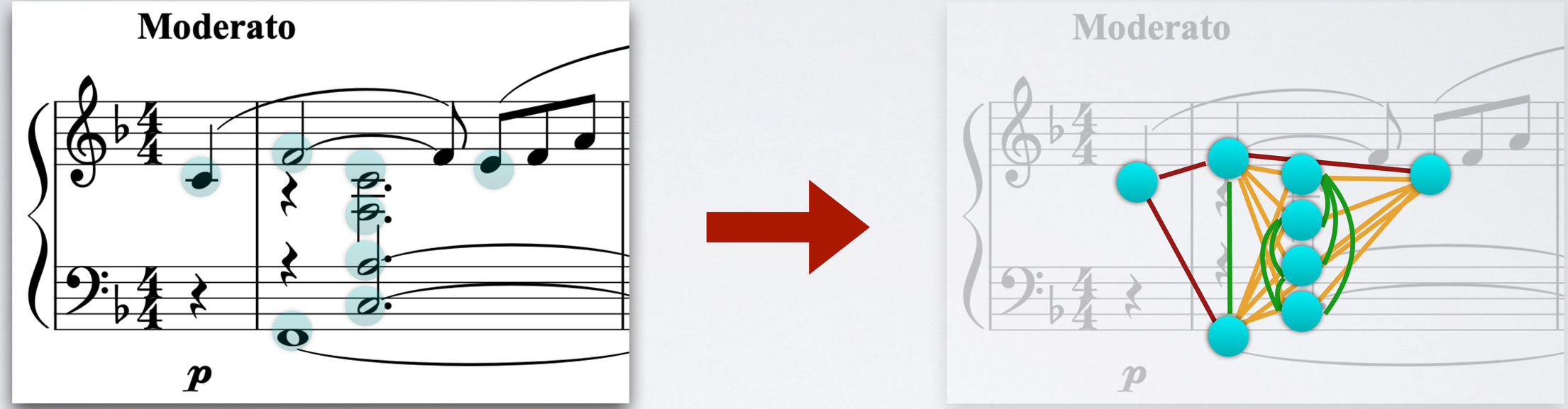
- How about handling a score as a graph?

Music Score as a Graph



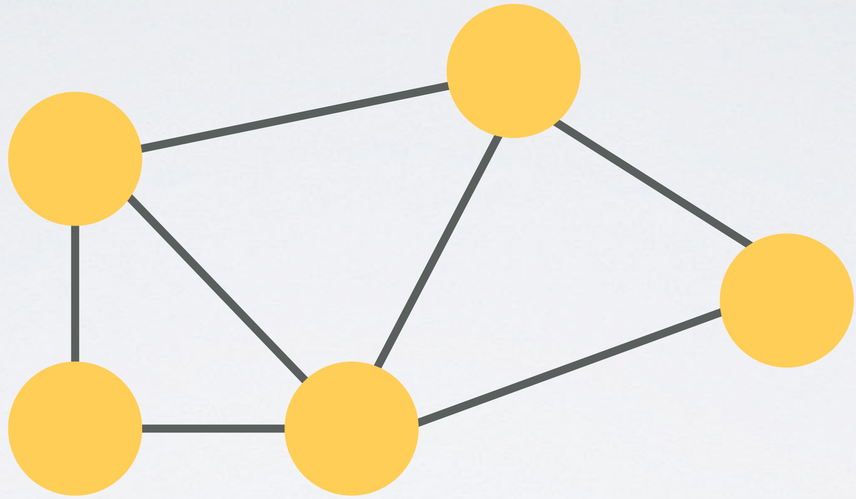
- Represent a music score as a graph of notes
- Each note becomes a node in a graph

Music Score as a Graph



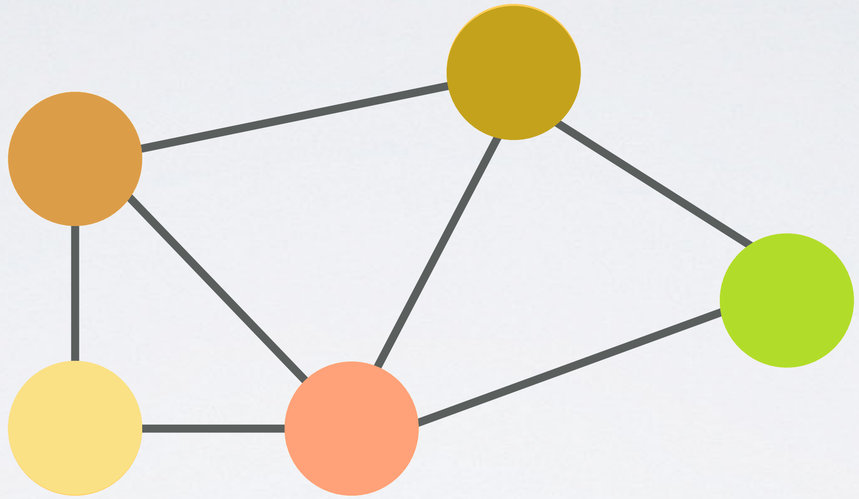
- Each node is connected with its musically neighboring notes
- Edge type is decided by notes' positional relation

Gated Graph Neural Network



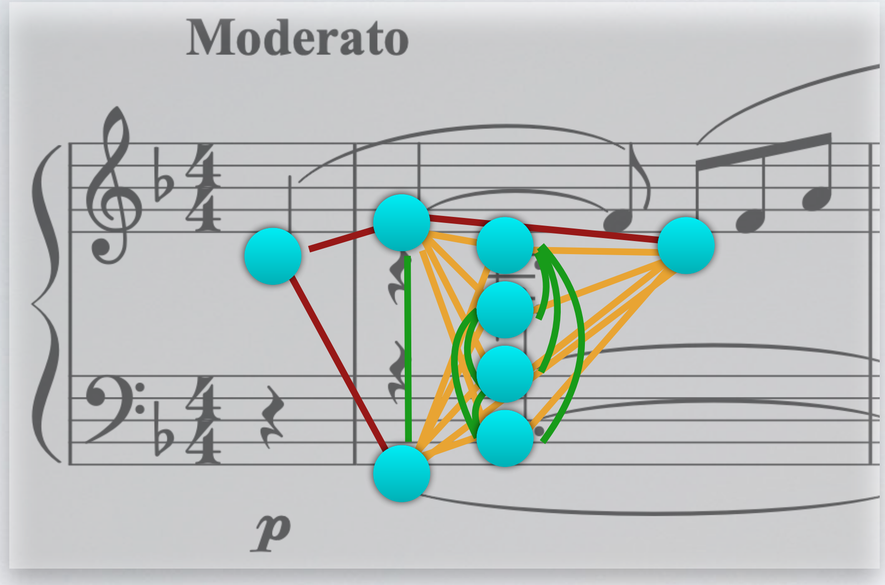
- Each node sends its information to neighboring nodes simultaneously with gated connection
- Each edge type has its own gate parameters

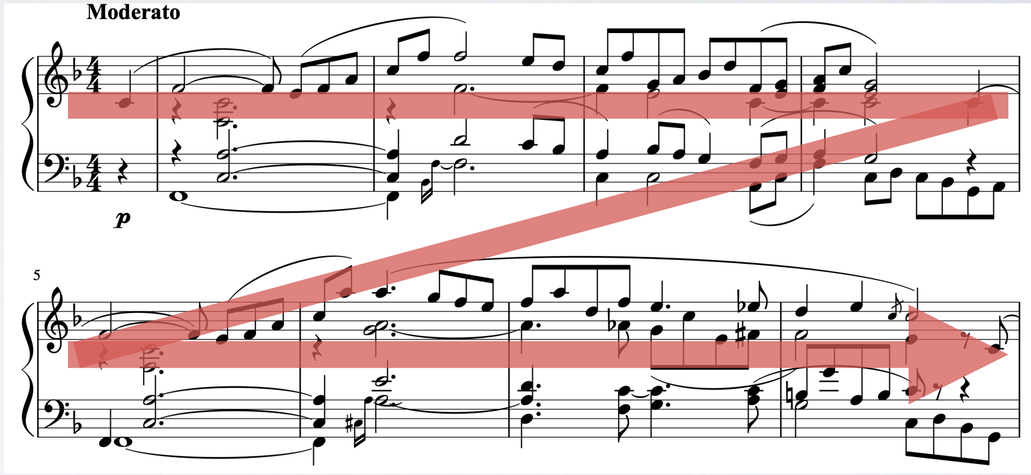
Gated Graph Neural Network



- Each node sends its information to neighboring nodes simultaneously with gated connection
- Each edge type has its own gate parameters

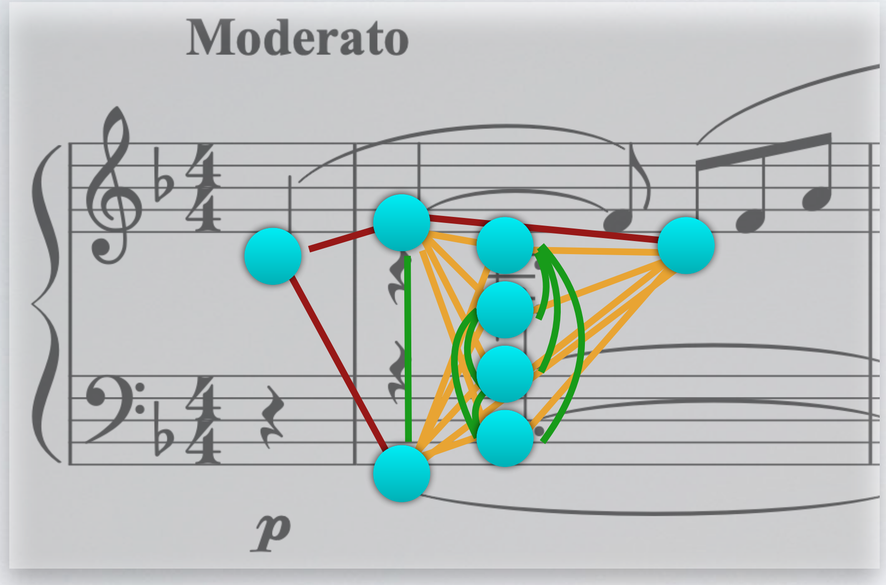
Music in Extended Context

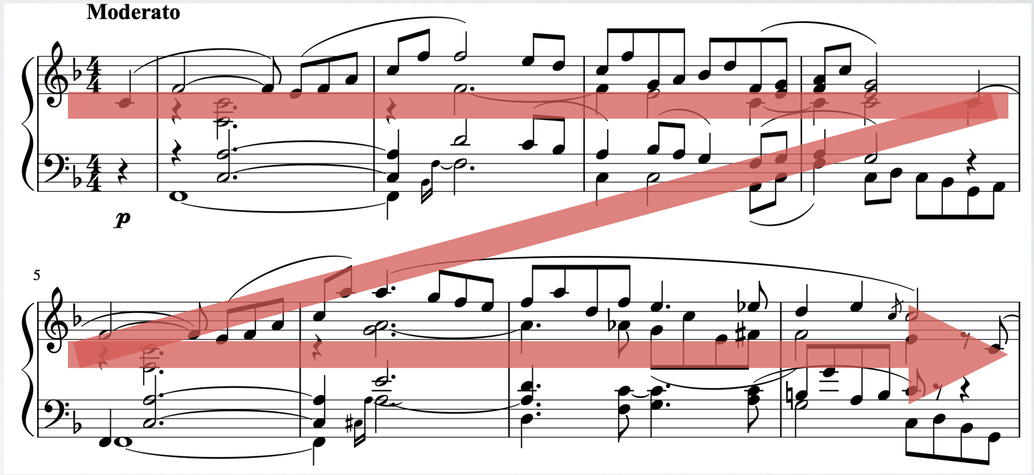




- GNN is suitable for handling the local context of each note.
- But music has sequence-like characteristics in extended context

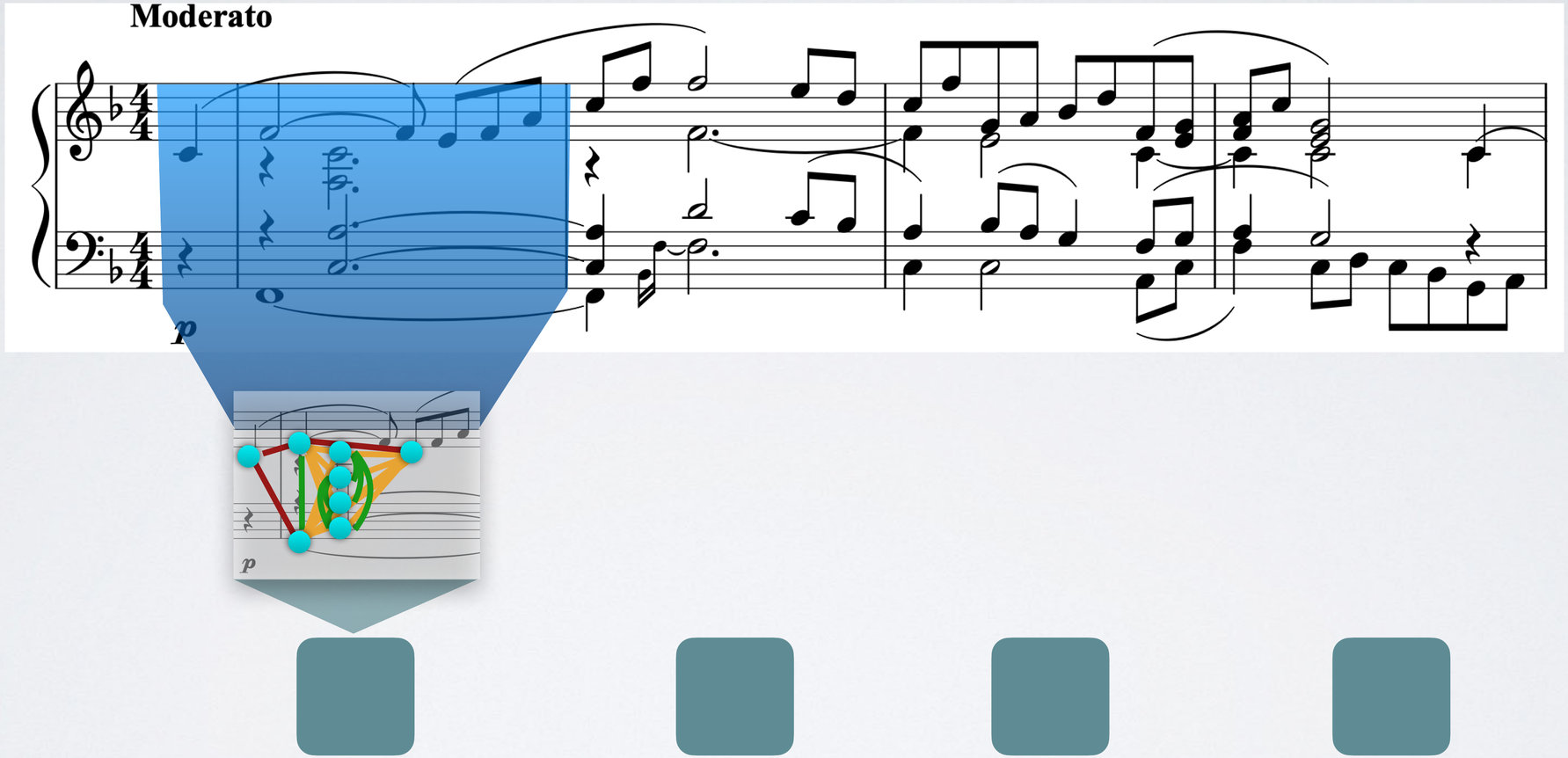
Music in Extended Context





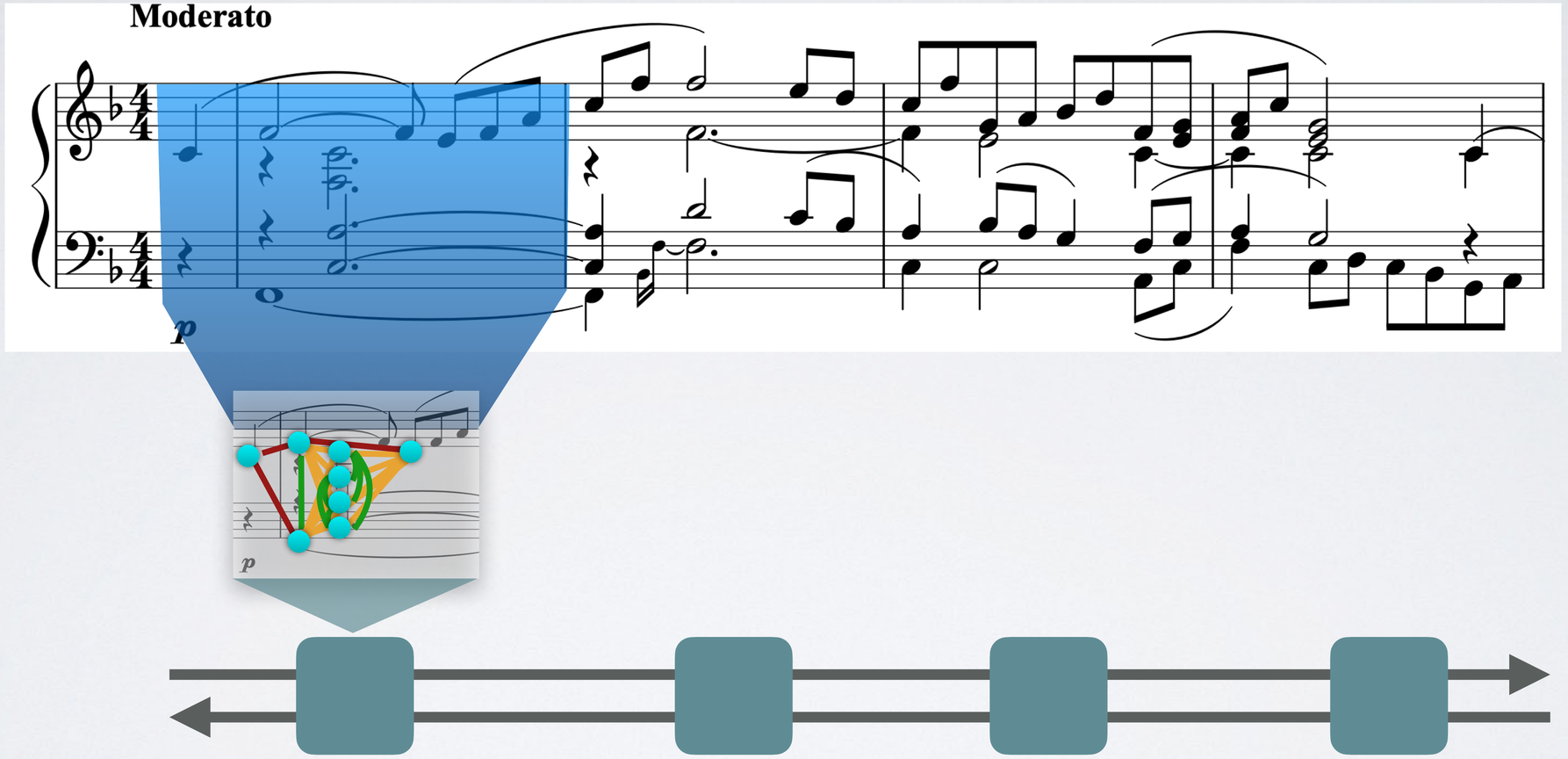
- We propose Iterative Sequential Graph Network (ISGN)

ISGN: Combining GNN with RNN



- Summarize note-level representations in a measure with Hierarchical Attention Network (HAN)

ISGN: Combining GNN with RNN



- Update measure-level representations with bi-directional RNN

ISGN: Iterative Update

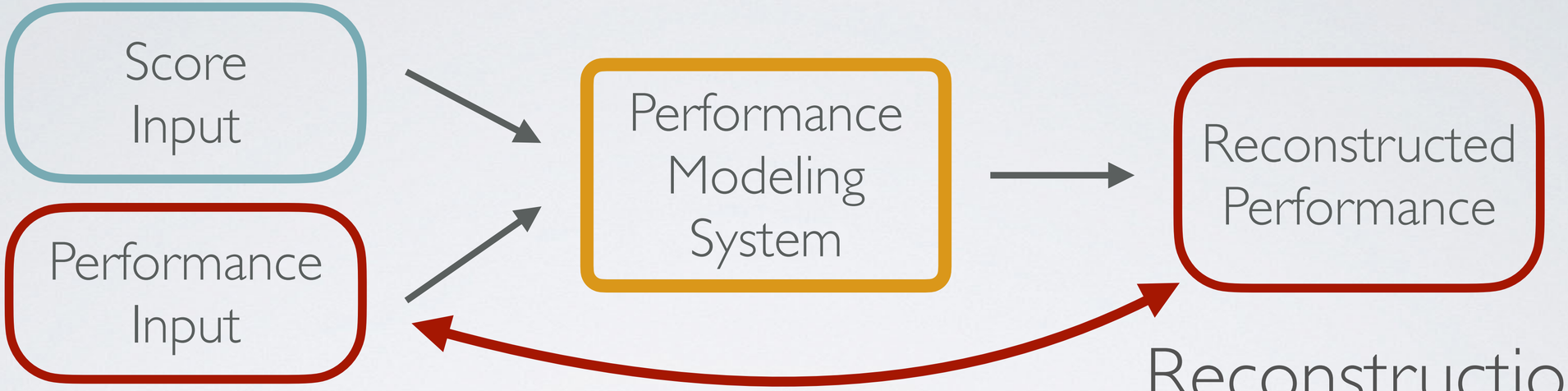
- Feed measure-level representations back into note-level representations
- Update note-level and measure-level representation again

Experiments

Baseline	HAN	G-HAN	ISGN
Note-level LSTM only	Note-level LSTM Beat-level LSTM Measure-level LSTM	Note-level GGNN Beat-level LSTM Measure-level LSTM	Note-level and Measure-level ISGN

- Trained 4 models with same module structure but different NN architecture.
- Trained with the same dataset and similar network size

Experiment Results



Model	Tempo	Vel	Dev	Pedal	KLD
BL	0.2721	0.6011	**0.7678**	0.8056	**2.2581**
HAN	0.2380	0.6290	0.7938	0.7681	13.666
G-HAN	0.2785	0.6212	0.7705	0.8092	7.1113
Proposed	**0.2379**	**0.5877**	0.7978	**0.7544**	3.7247

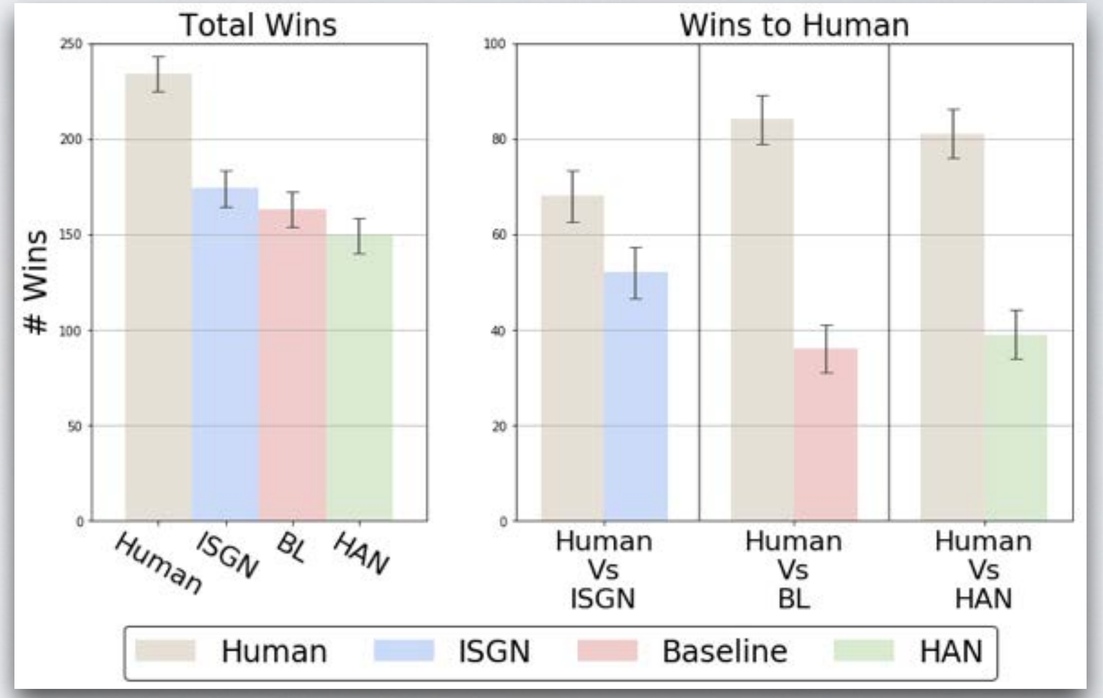
Reconstruction loss on test set

Loss values are in MSE
All features are normalized

Tempo: tempo in bpm
Vel: dynamics in MIDI velocity
Dev: onset deviation
Pedal: seven pedal parameters

Human Listening Test

- Pairwise Comparison in double-blind
- 40 subjects
- 6 pieces, about 30 seconds each
- 720 comparisons
  - Each model was included in 360 comparisons

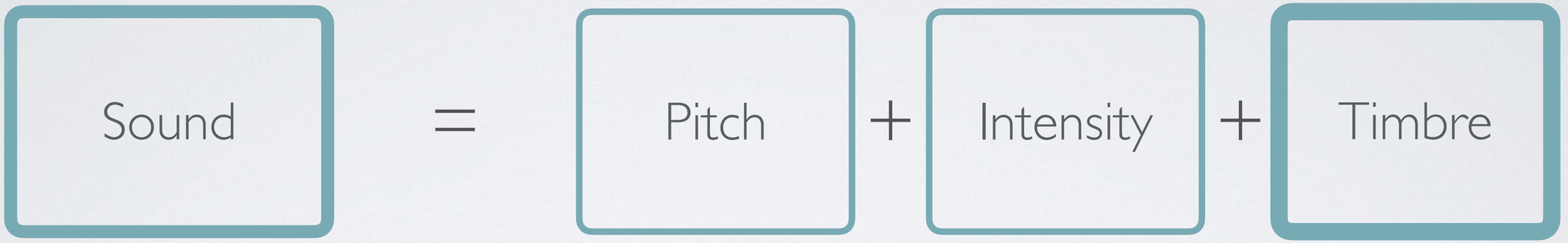


1. Introduction
2. Performance Modeling with RNN
3. Performance Modeling with GNN

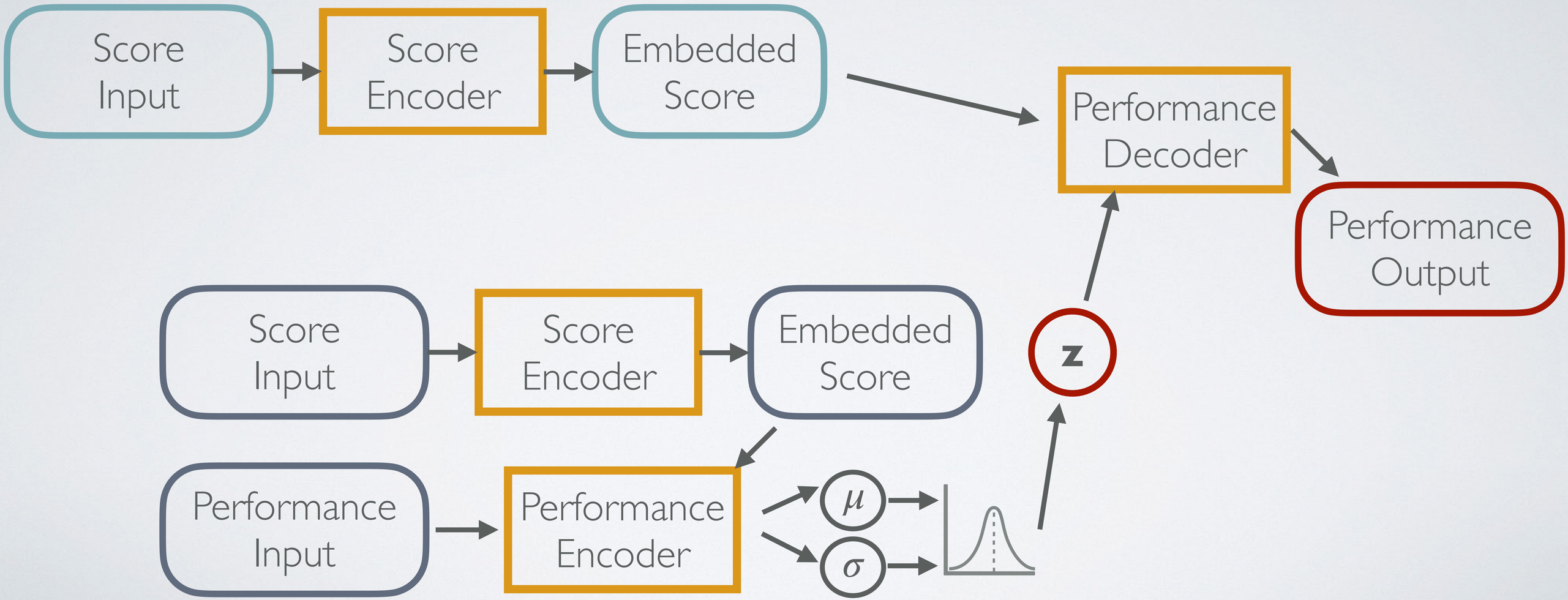
4. Performance Style Analysis

5. Future Research

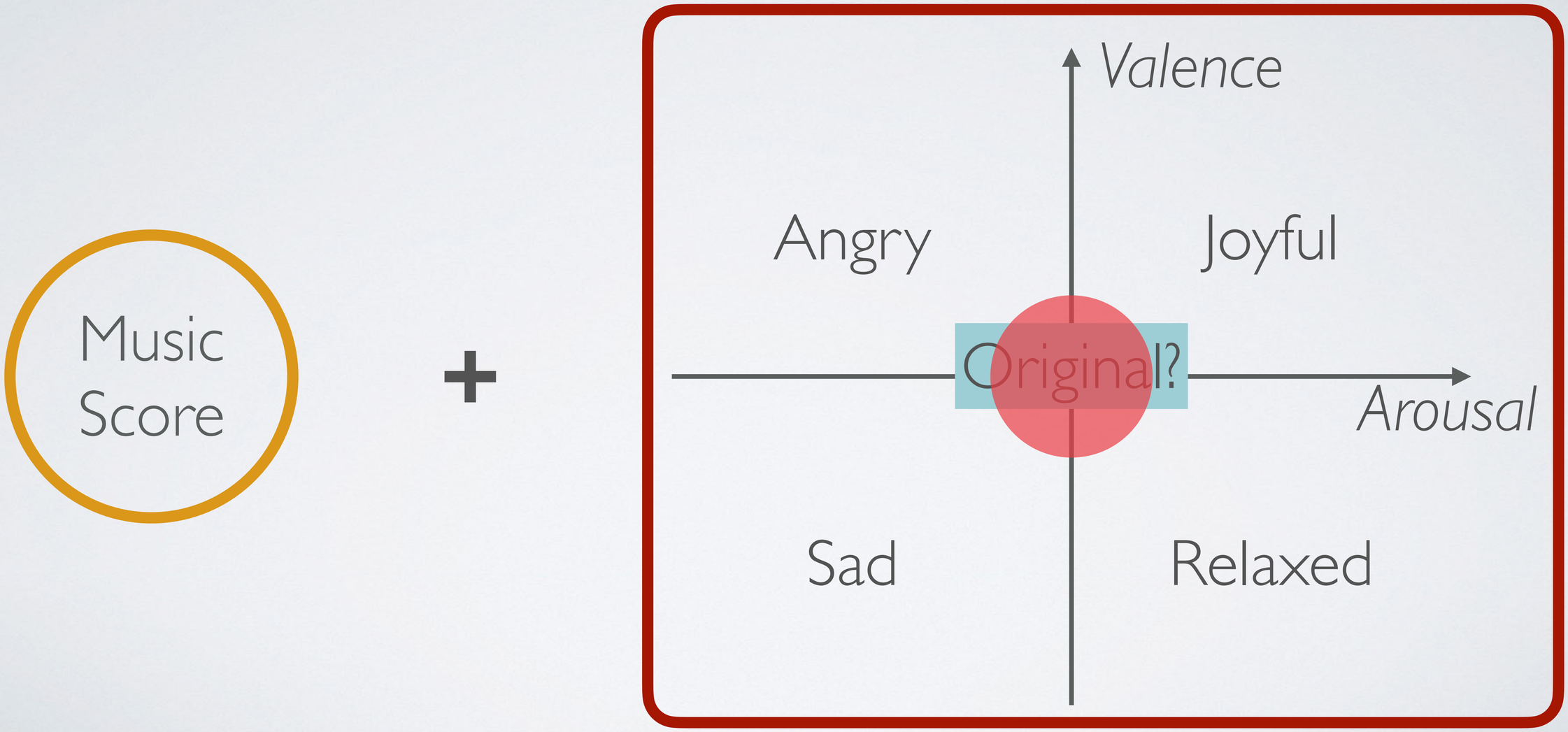
Performance Style Analysis



Performance Style Transfer

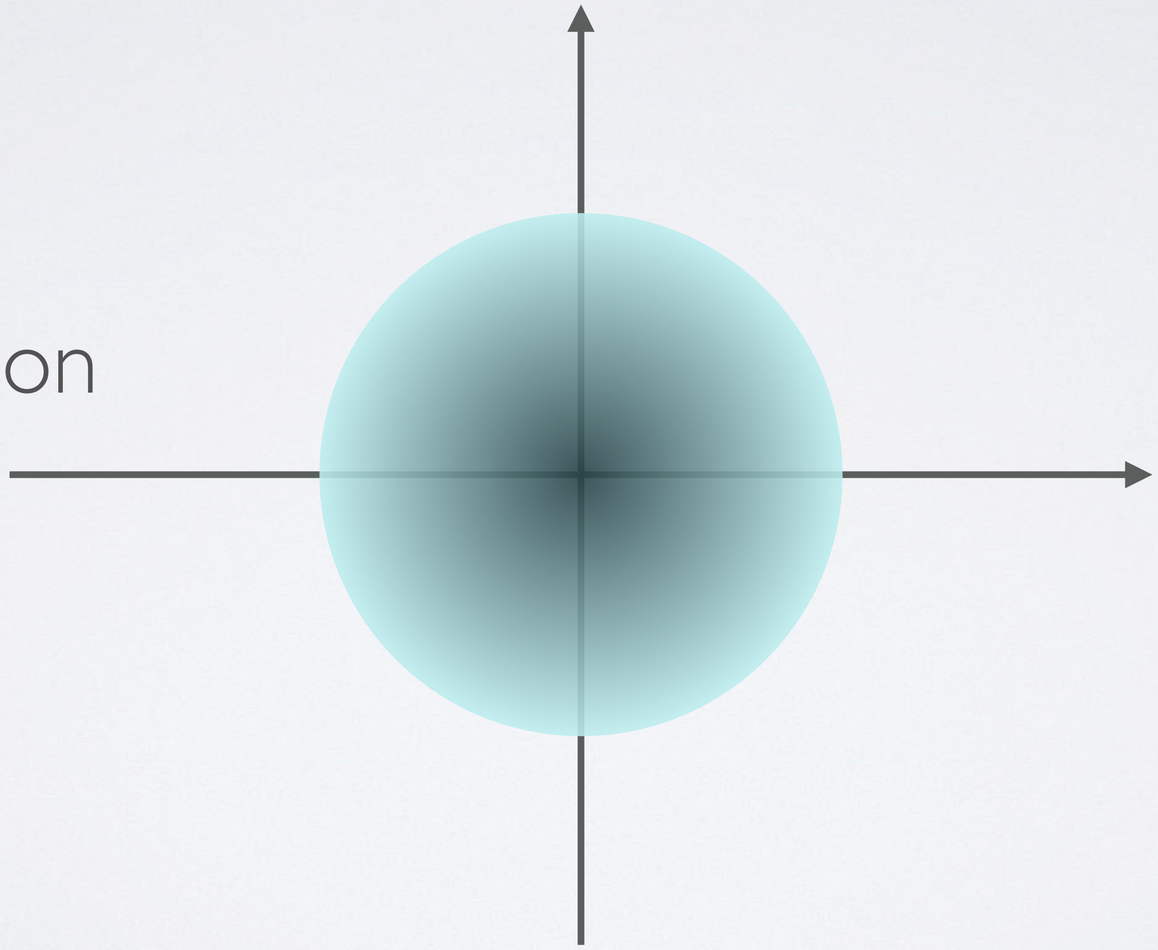


Emotion-cued Performance



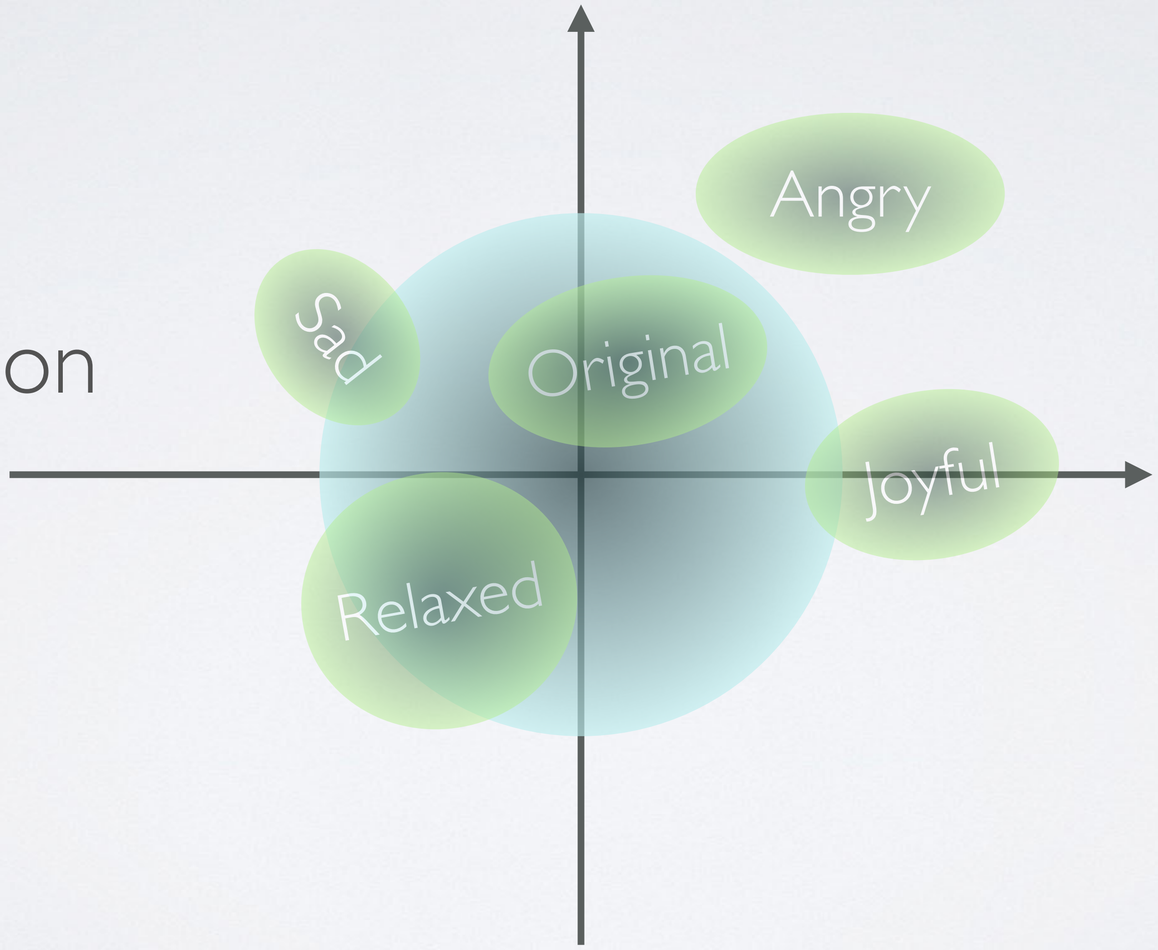
Performance Style Transfer

Normal Distribution



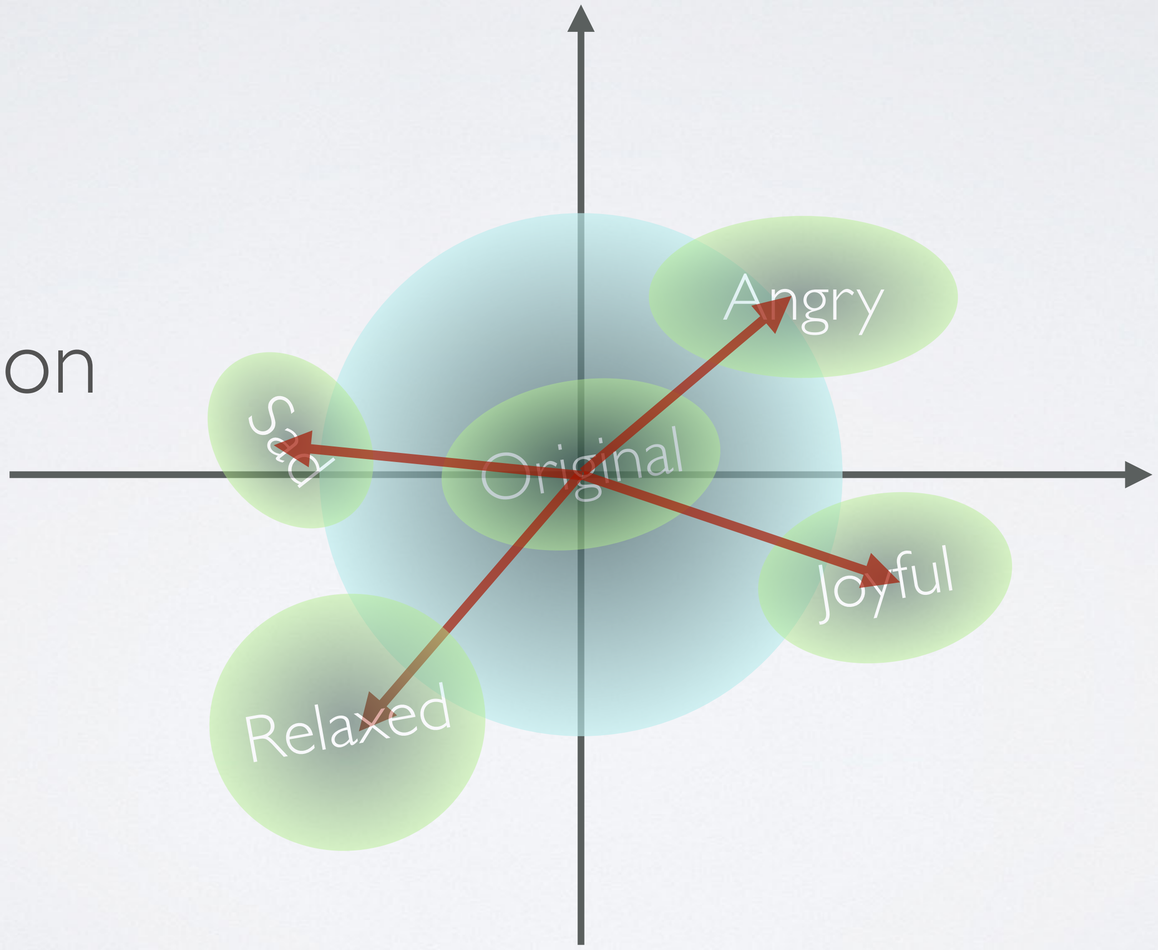
Performance Style Transfer

Normal Distribution

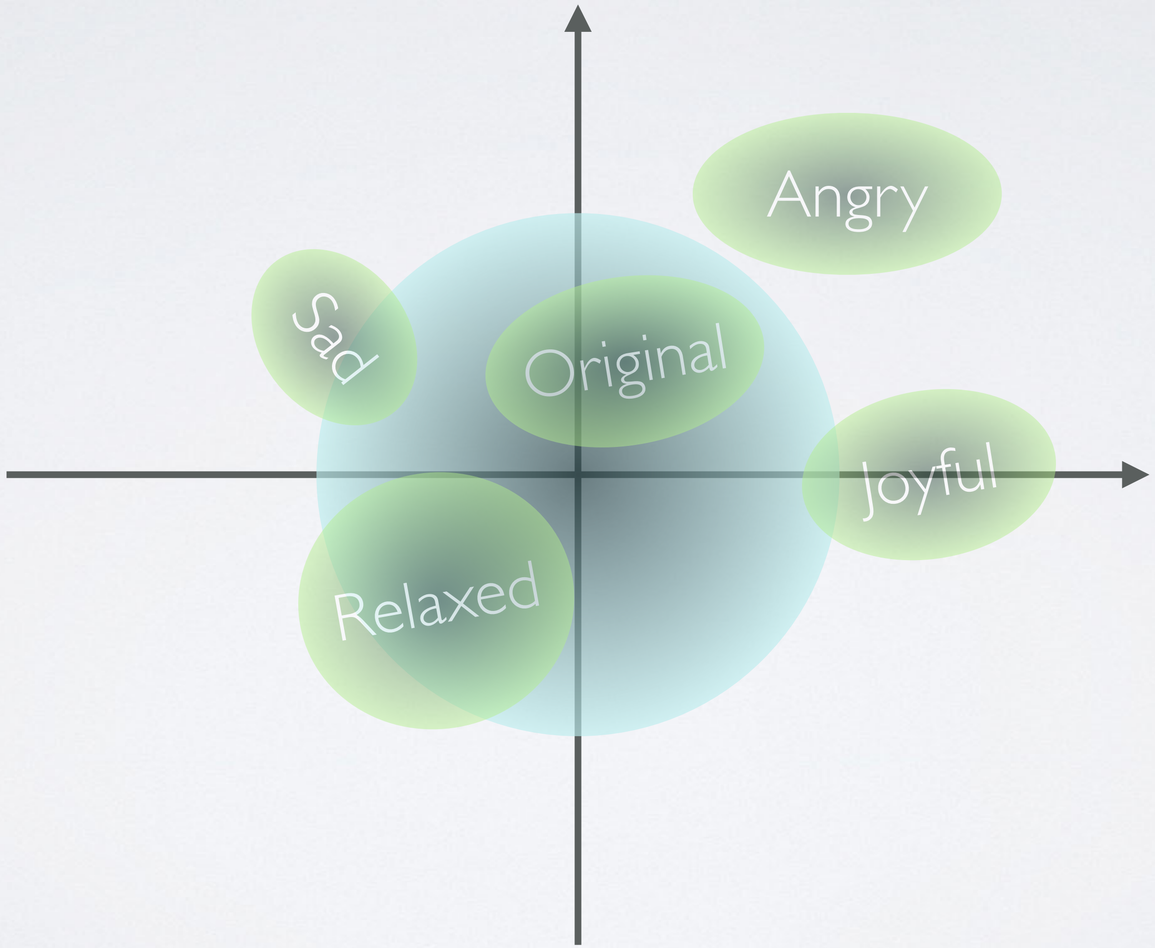


Performance Style Transfer

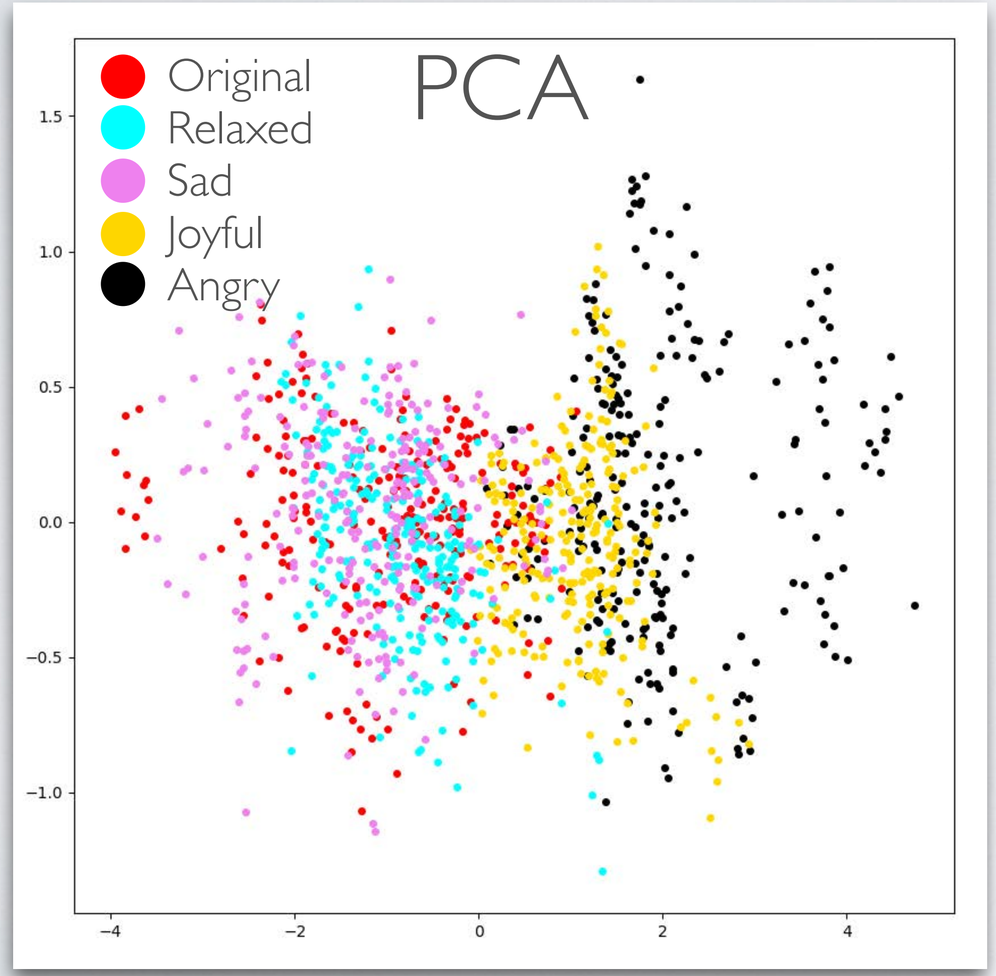
Normal Distribution

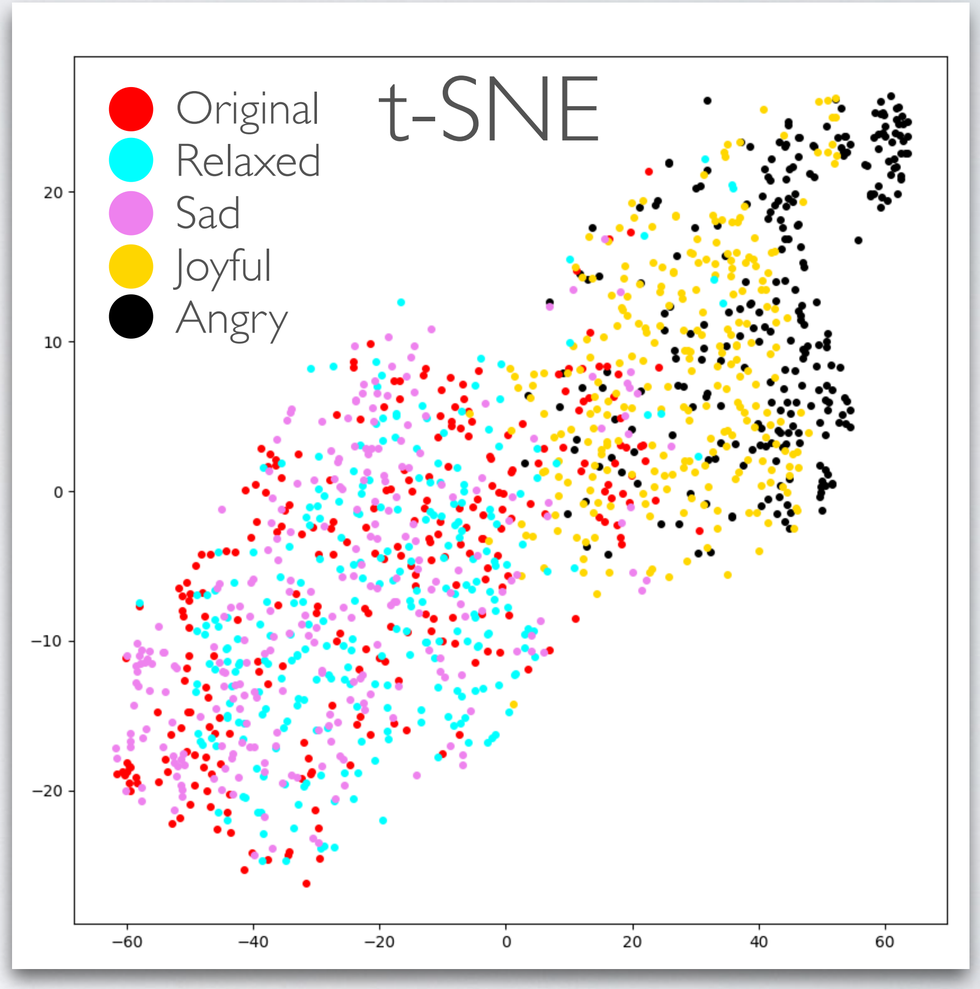


Performance Style Analysis

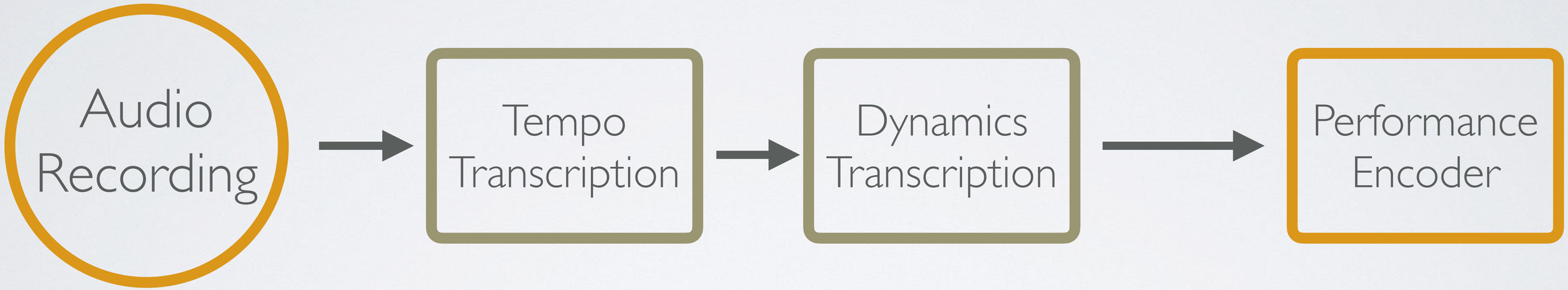


Performance Style Analysis



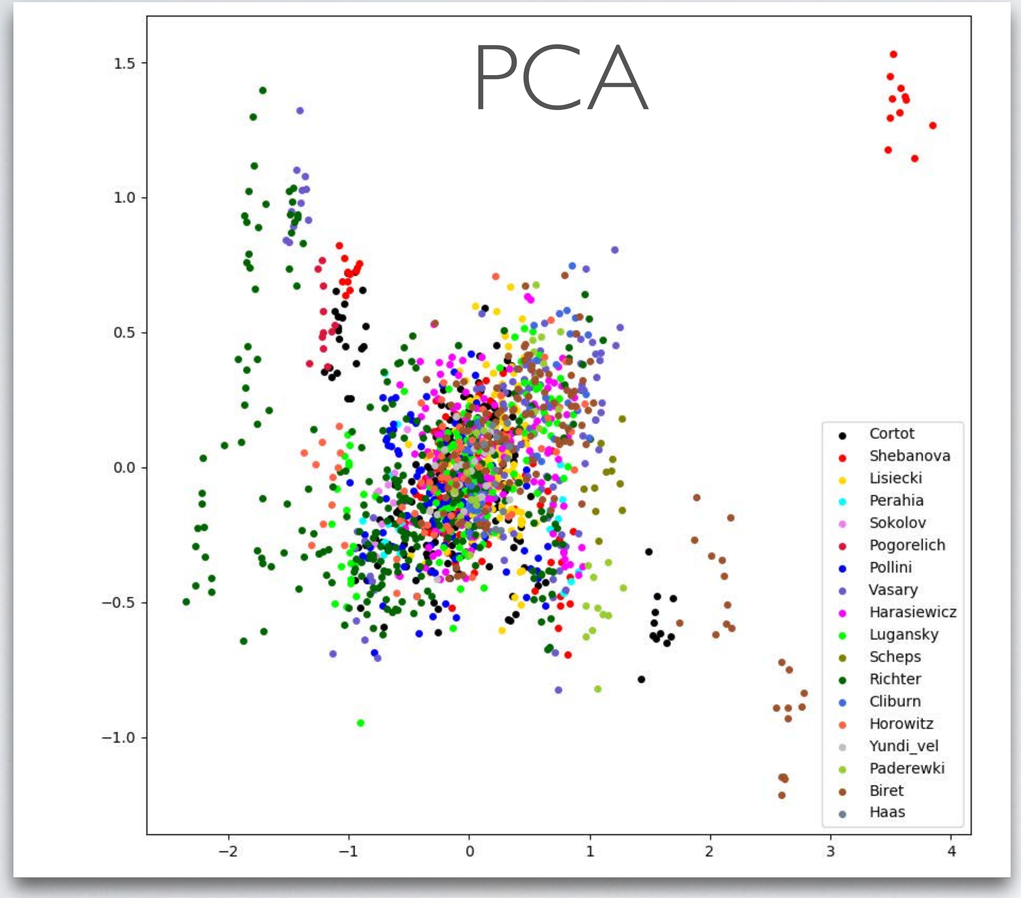


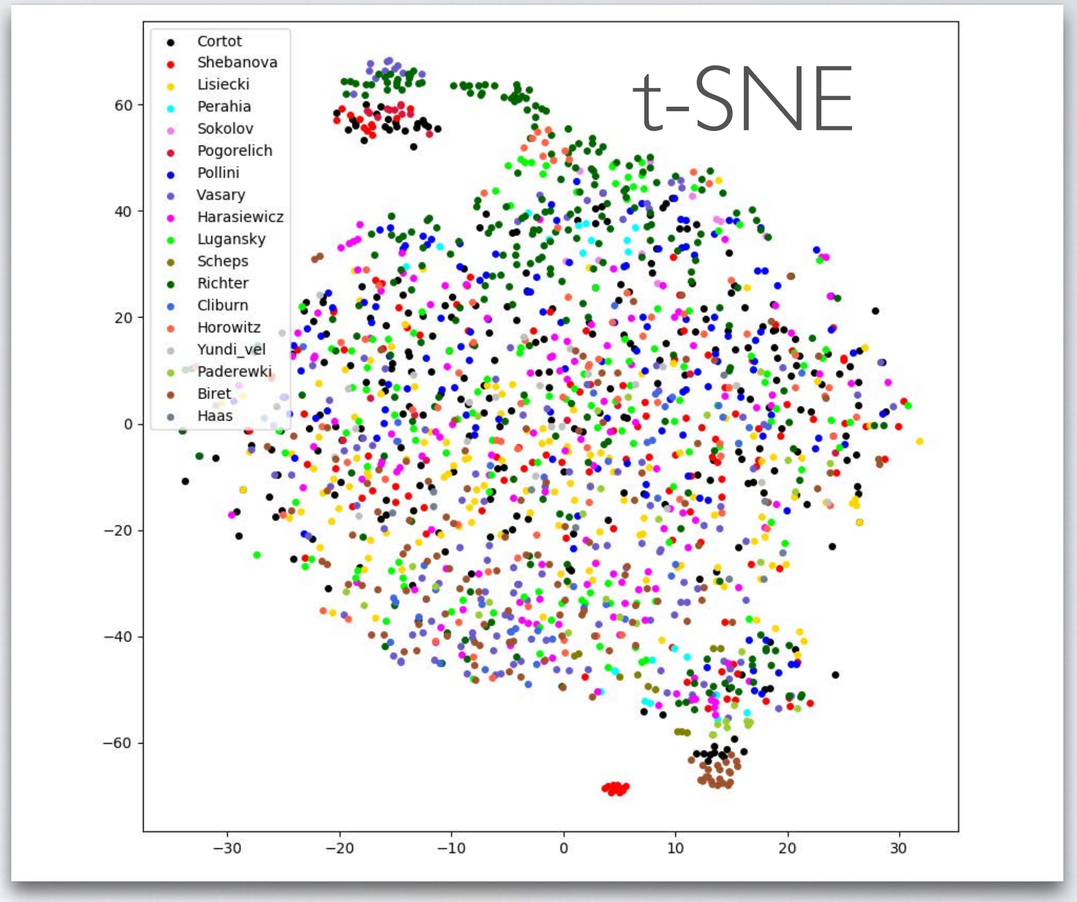
Performance Style Analysis



We transcribed audio recordings of 18 performers of 12 Etudes op. 10 by Chopin

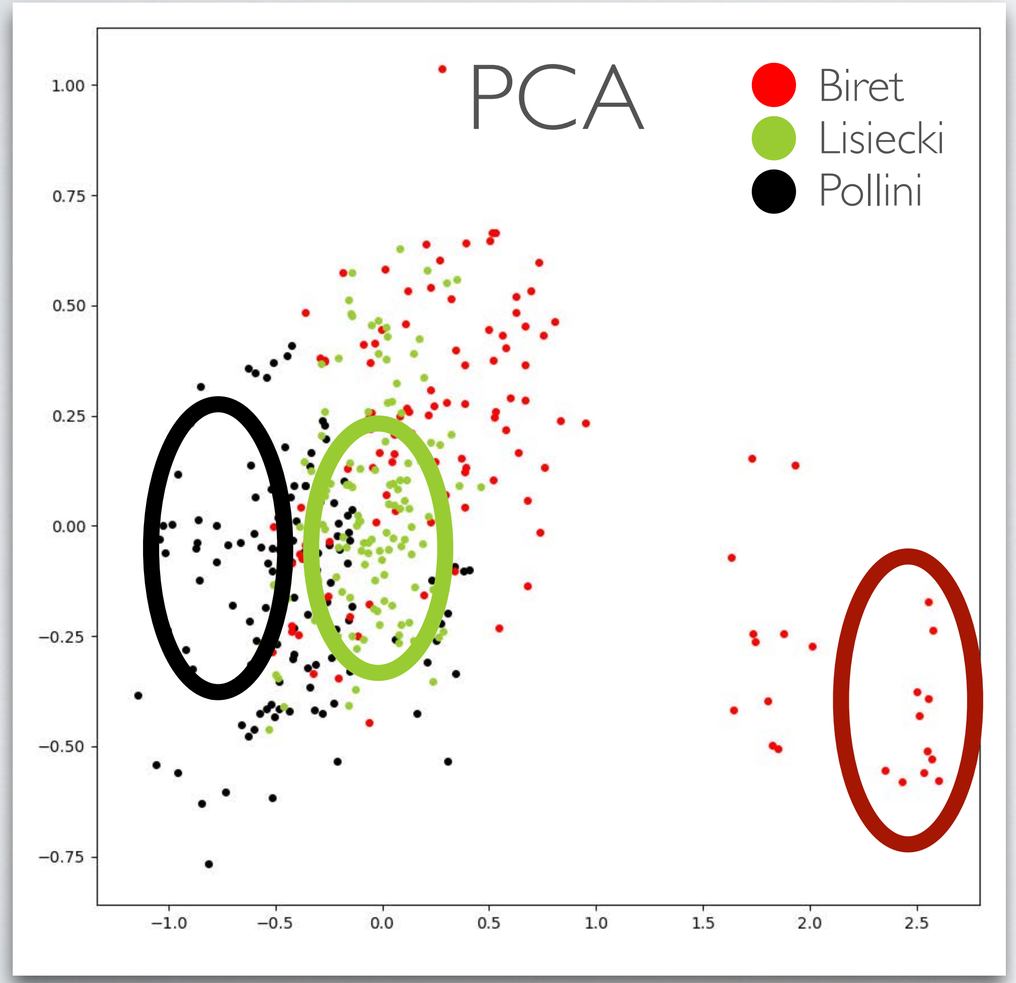
Performance Style Analysis

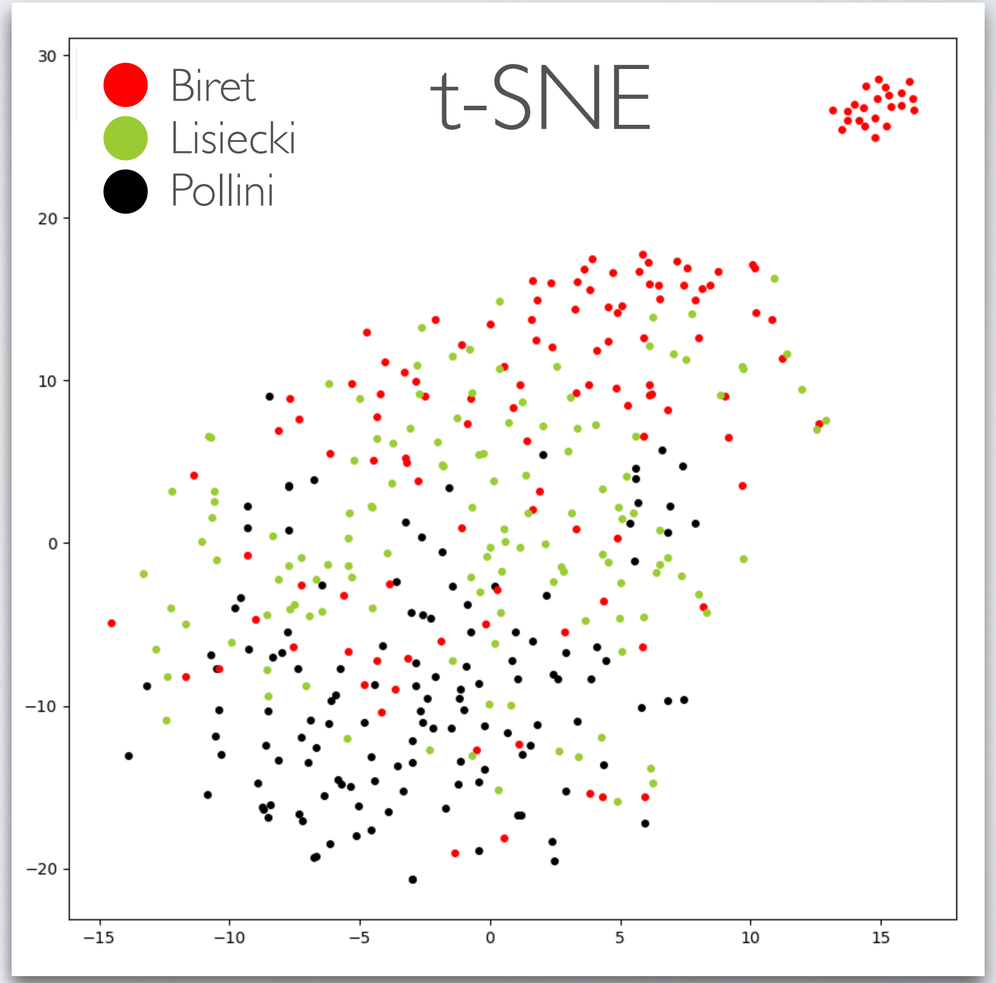




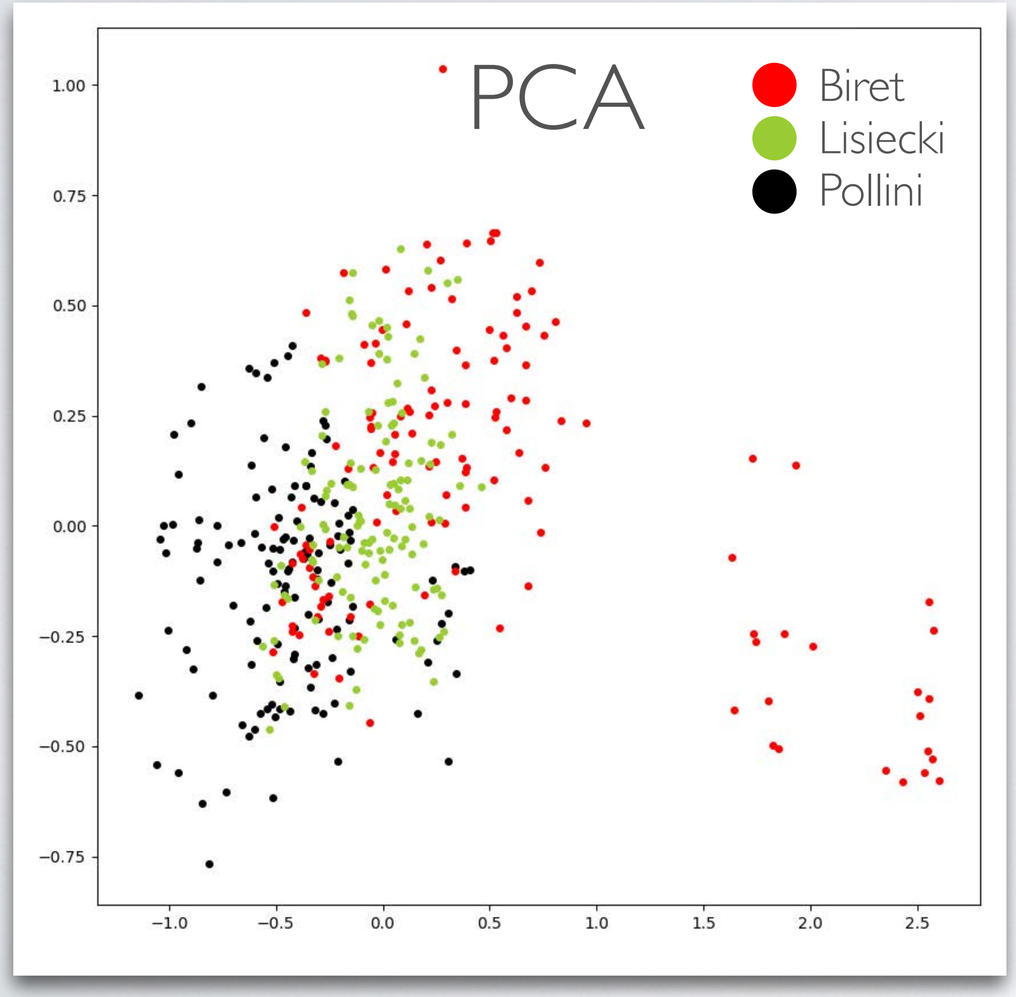
Each point represents a style vector of single piece

Performance Style Analysis

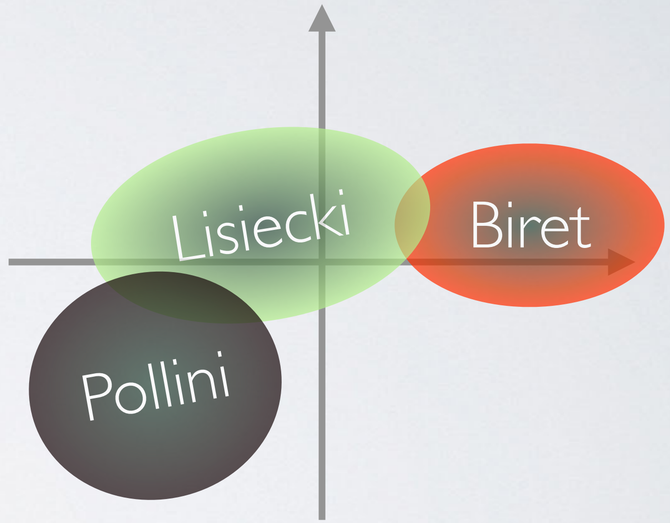




Performance Style Analysis

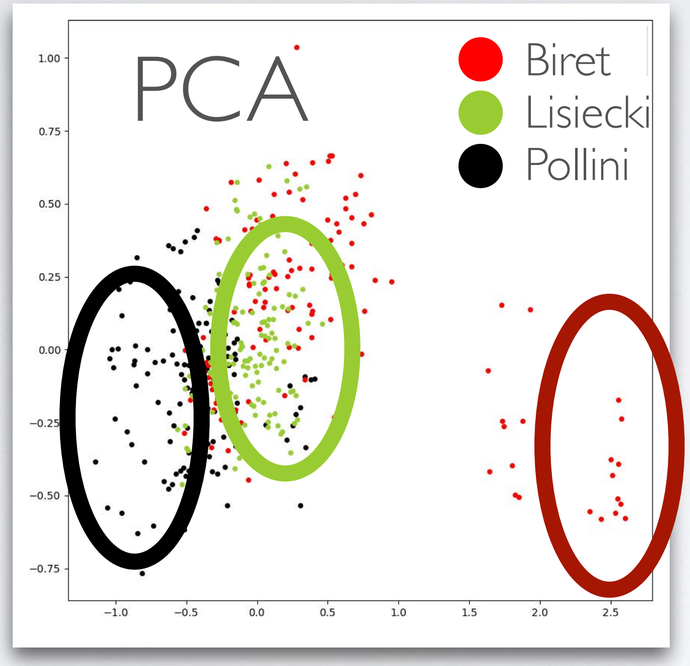


Reduction to 2D
for Visualization

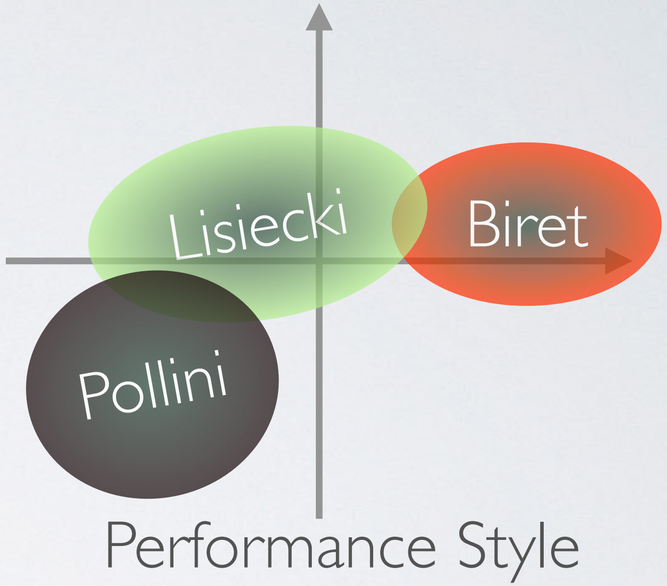


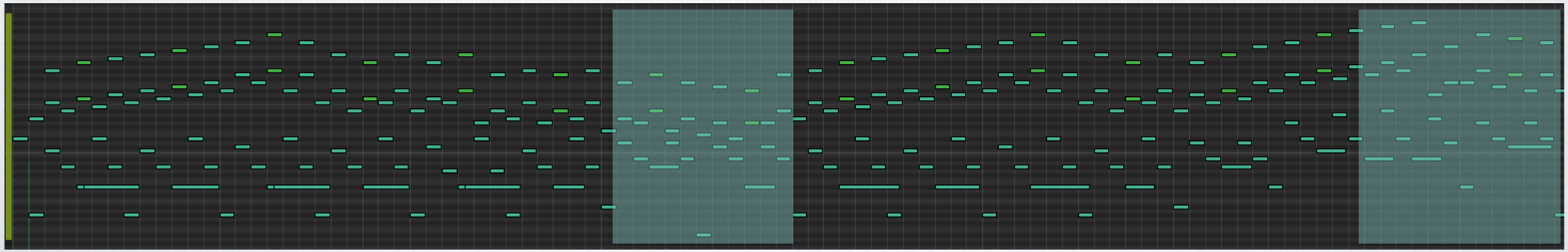
Performance Style

Performance Style Analysis



Reduction to 2D
for Visualization





1. Introduction
2. Performance Modeling with RNN
3. Performance Modeling with GNN
4. Performance Style Analysis

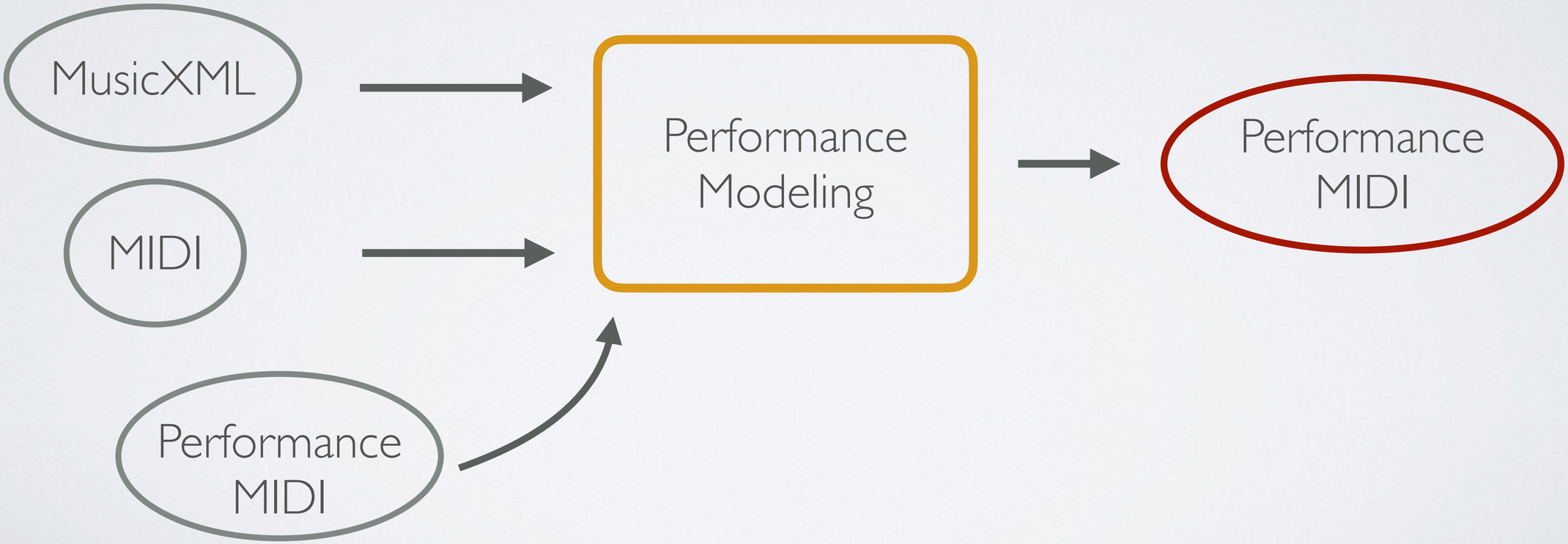
5. Future Research

Performing with Various Instruments

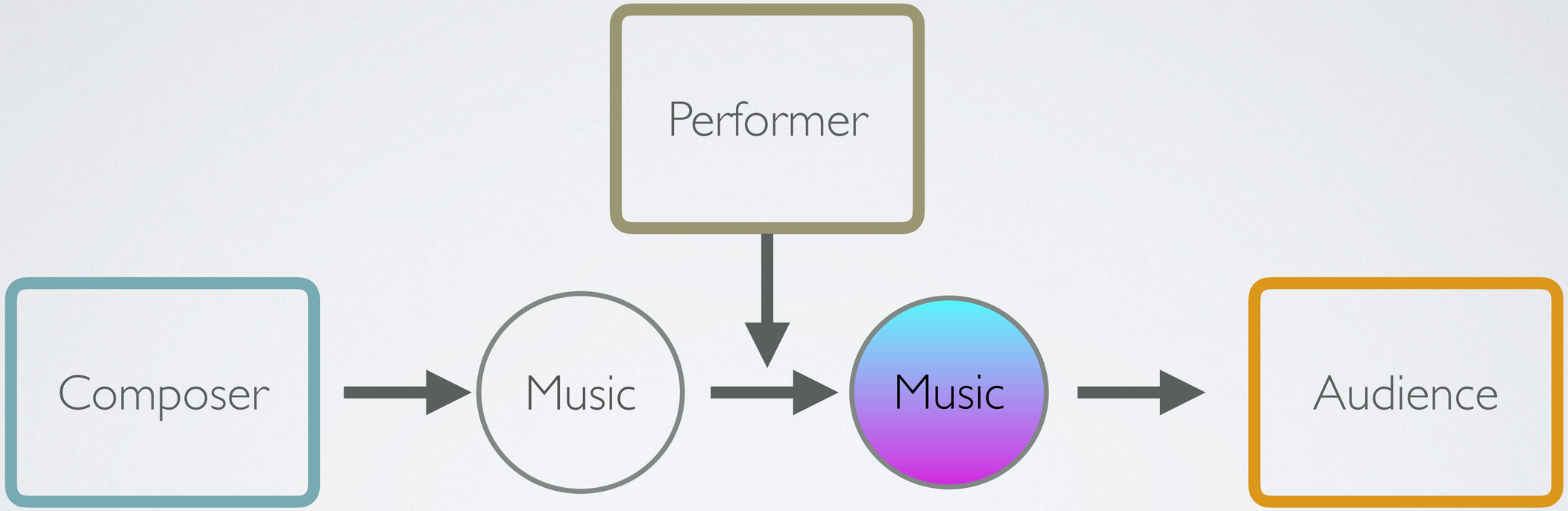
Performance Modeling

Performance MIDI

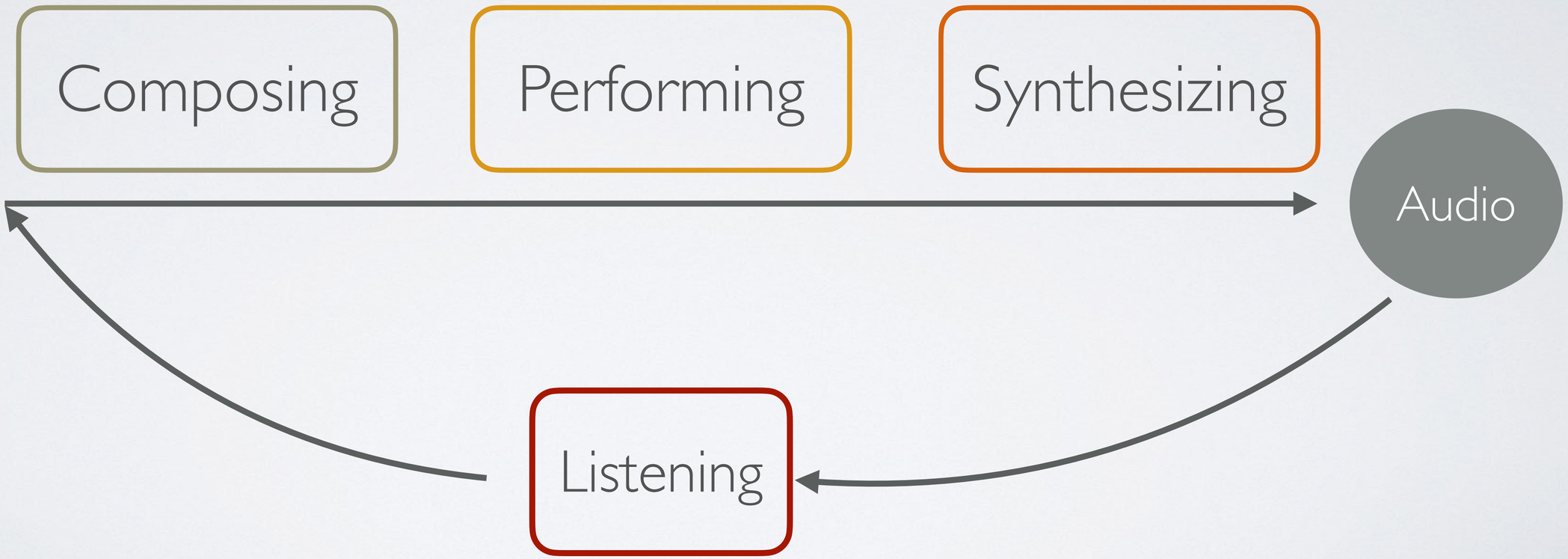
Performing from MIDI



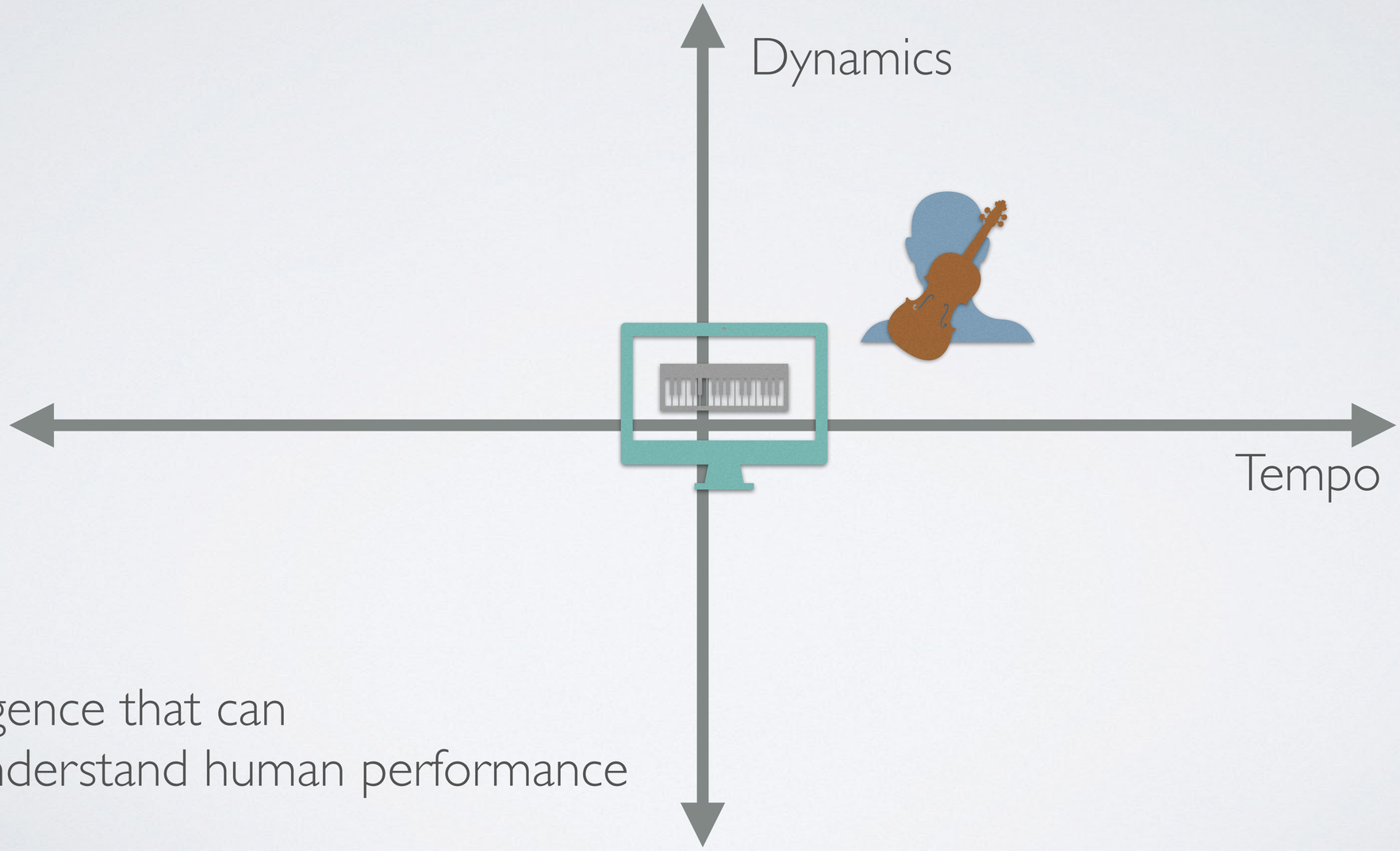
Every Music needs Performer



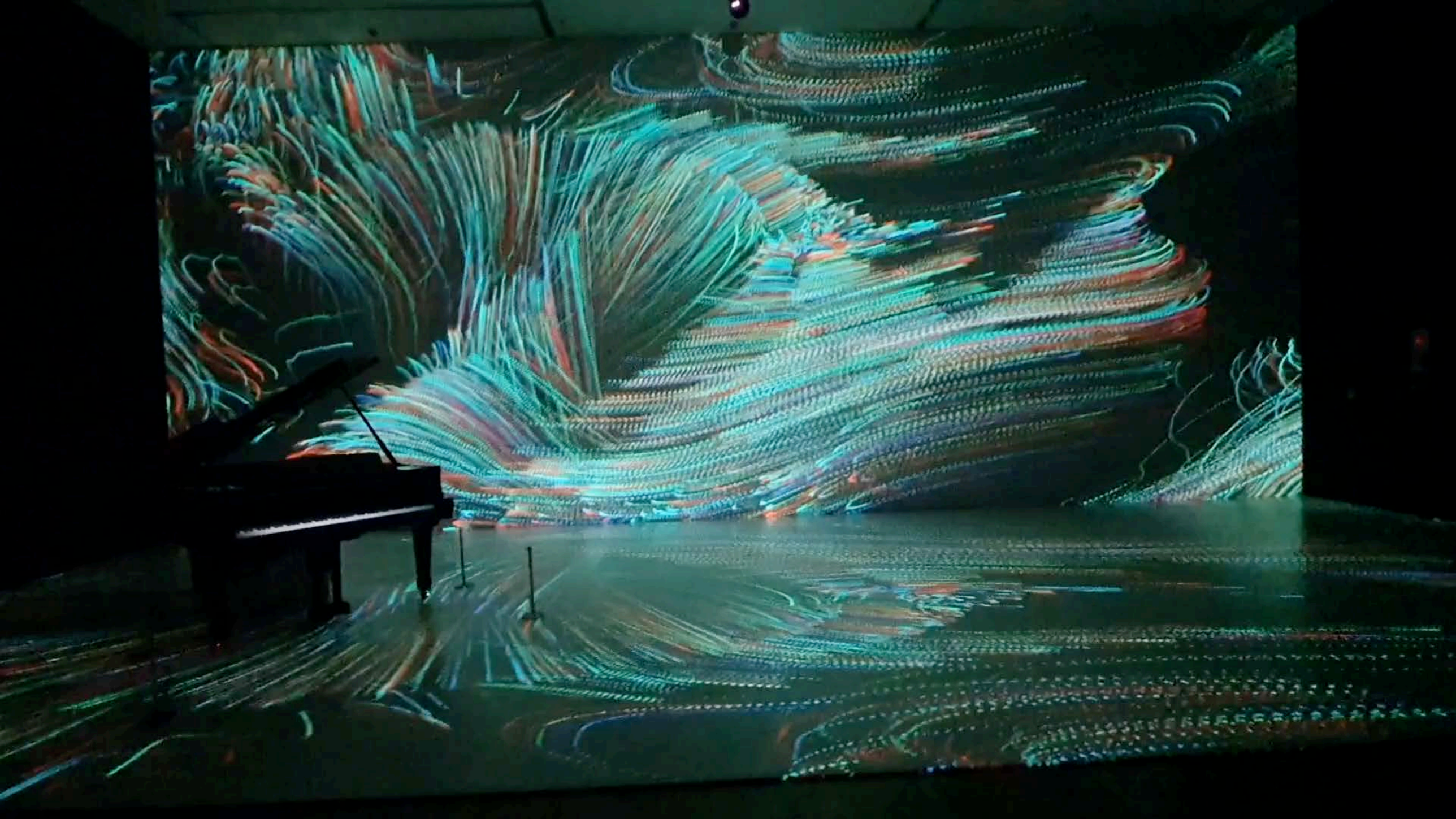
Complete Cycle



Human-Machine Ensemble



Music Intelligence that can
listen and understand human performance

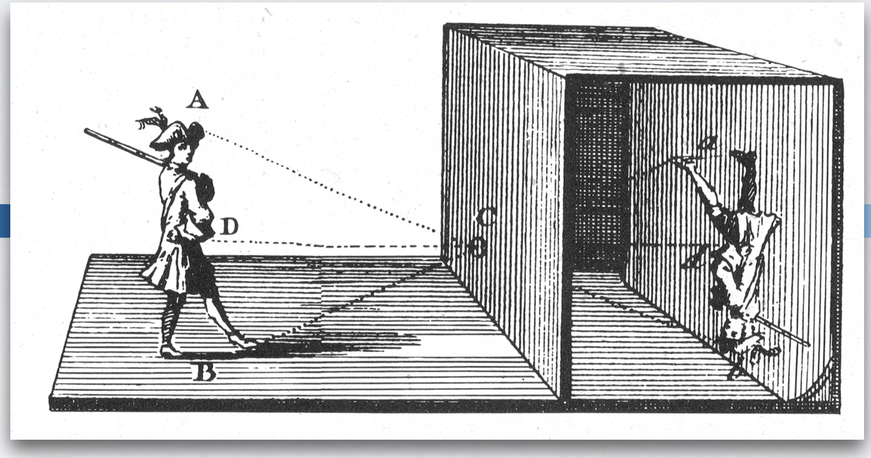



김포문화재단
Gimpo Cultural Foundation
Gimpo Philharmonic Orchestra
YAMAHA


2022 카이스트 봄 문화행사
SUMI JO
I'm a KAIST
조수미
KAIST
SPRING
CULTURE
FESTIVAL
17th

Will AI Threaten Pianists?

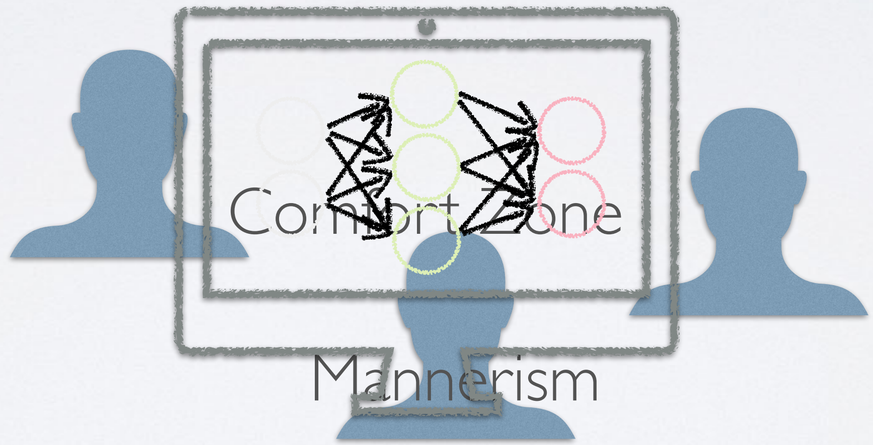
Paintings



Performance

Impact of AI on Music Performance

Bold



Creative

Fresh

생전에 인공지능에 지지 않으려면 아무래도 생각을 좀 바꿔야 할 것 같지 않나?

If you don't want to be defeated by AI,
why don't you change your goal?

오직 나 밖에는 하지 못할 독창적인 연주,
그것도 연주할 때마다 달라 데이터가 기록을 하다하다
포기할 만큼 매순간이 살아있는 연주,
경탄이 아닌 감동을 전하는 연주를 하는 인간으로 말이다.

One has to be a human who can demonstrate a performance,
which is so distinctive that can be only performed by himself/herself,
which is too dynamic that computer cannot catch,
which is moving rather than jaw-dropping

- 피아니스트 손열음

Pianist Yeol Eum Son

중앙선데이 473호 29면

Image from
http://www.hani.co.kr/arti/culture/music/863486.html

AI Pianist:
Modeling Expressive Performance with Deep Learning

Dasaem Jeong,
Assistant Professor @ Dept. Art & Technology, Sogang University
2022. 4. 8

https://jdasam.github.io