

音楽における系列推定問題 ～自動採譜を例に～

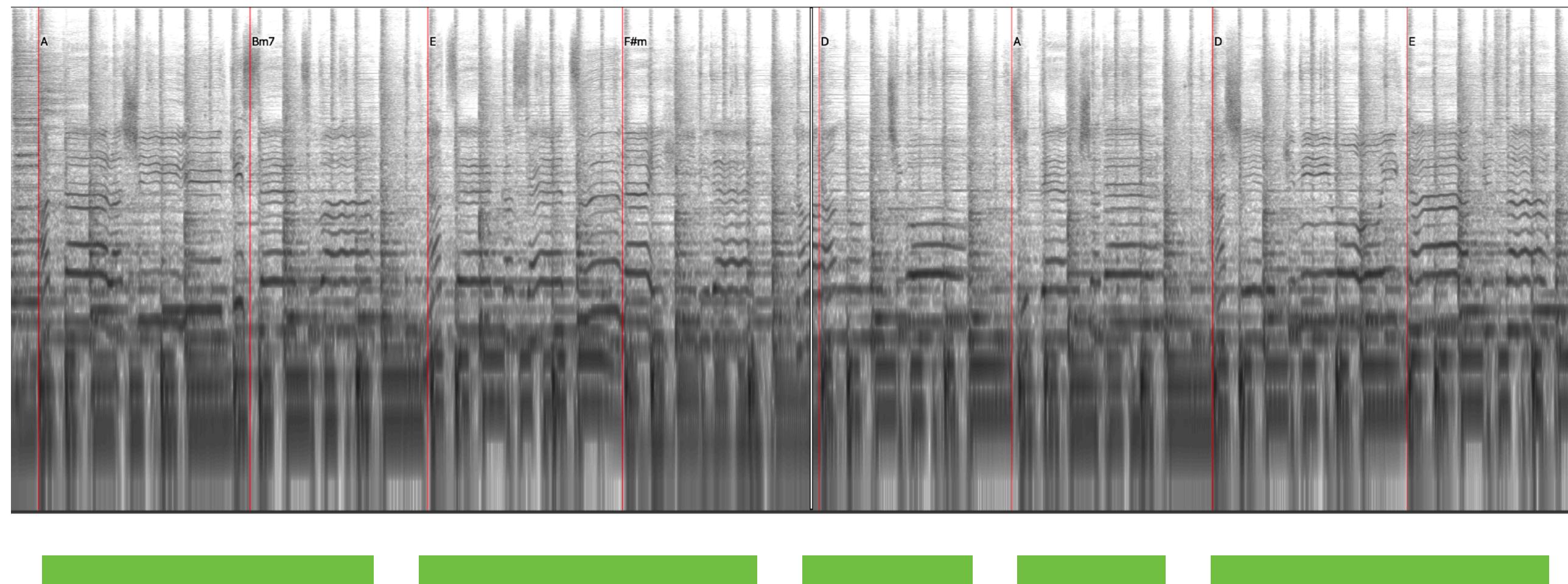
Deep-people #7

- **RNN**
 - 時系列を対象とするデータ（自然言語処理・音）のタスクでの利用
 - RNNのしくみ
 - ゲートつきRNN（GRU・LSTM）
 - そのほか時系列データに用いるモデルの紹介（TCN・Transformer）

- Juhanの資料
- [https://mac.kaist.ac.kr/~juhan/gct634/Slides/
\[week10-1\]%20automatic%20music%20transcription.pdf](https://mac.kaist.ac.kr/~juhan/gct634/Slides/[week10-1]%20automatic%20music%20transcription.pdf)

- **系列推定問題を解く方法の解説 & 音楽・音の識別タスクの実例**

- 特に音楽においては、曲全体で欲しい情報と各時刻ごとに欲しい情報がある
- 今までの分類は曲全体の情報の識別にフォーカス
- では、各時刻の情報を識別する方法は？ 自動採譜を例に身につけよう



Ex. 曲全体の情報 :

歌手 ->スピッツ, 楽曲 -> ロビンソン

ジャンル -> ポピュラー

Ex. 各時刻の情報 :

コード情報 (赤) -> A 35.0 - 37.3,

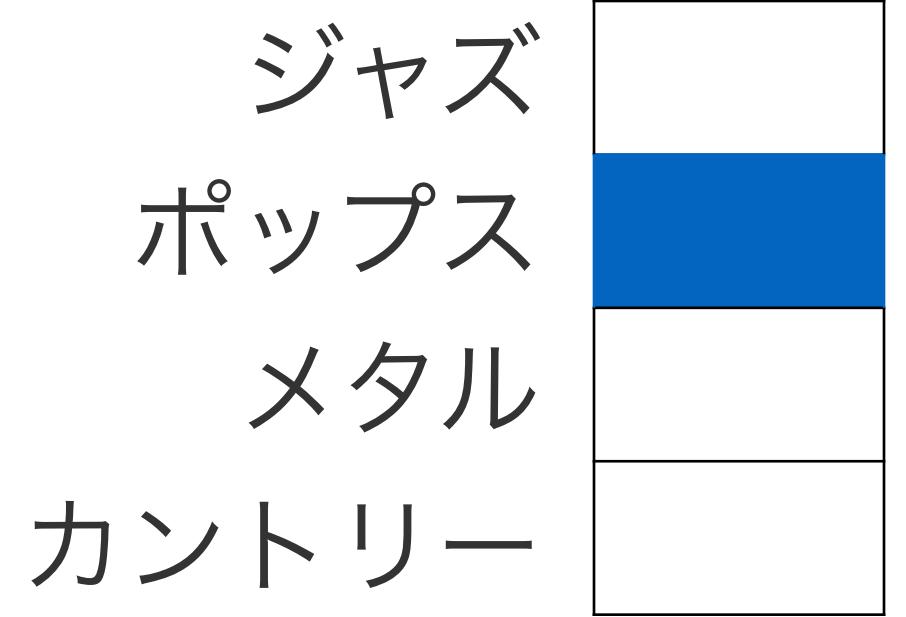
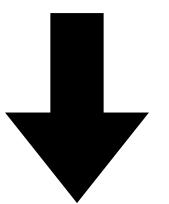
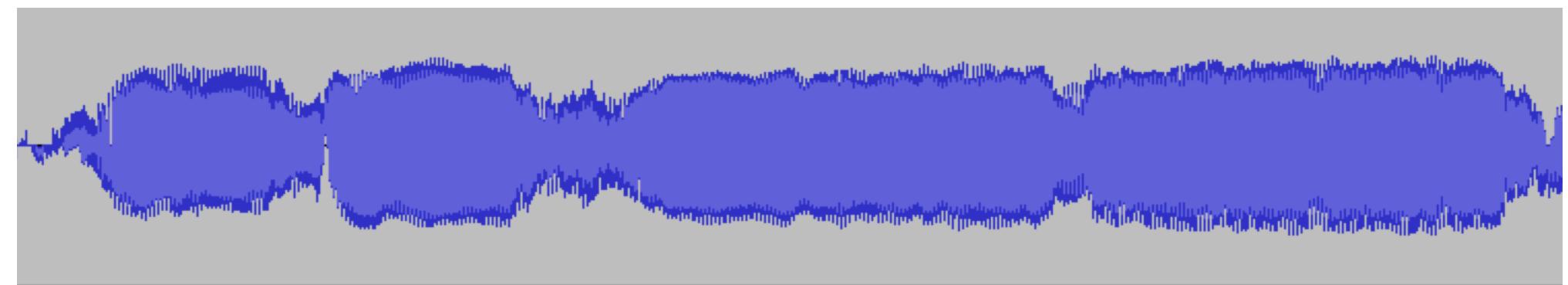
Bm7: 37.3 - 39.2 ...

歌声区間 (緑) -> 35.0 - 38.8, 39.2 - 40.9 ...

Ex. 音楽における系列推定

5

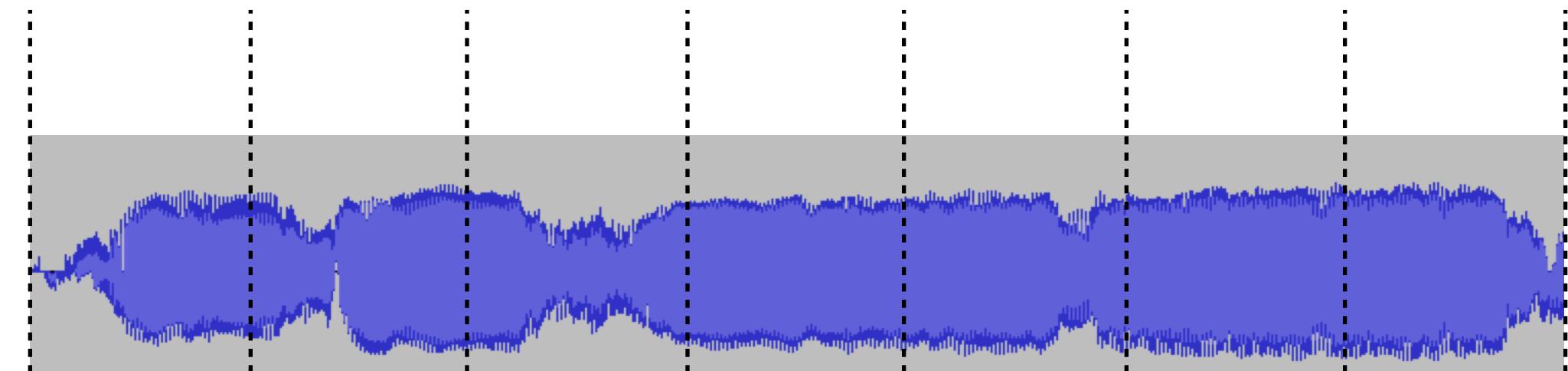
ファイル単位：音楽ジャンル推定



正解
ラベル

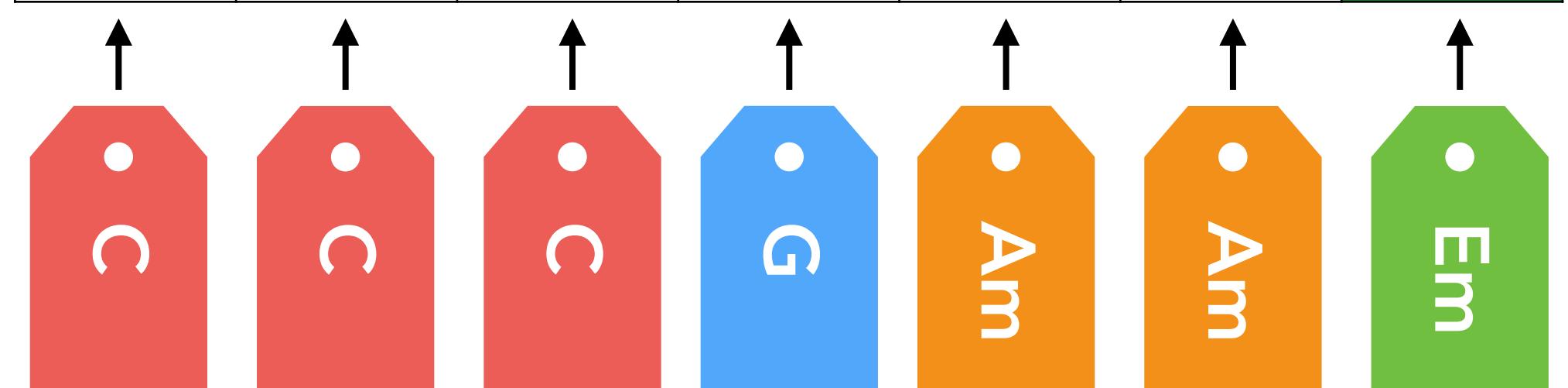


時系列：コード進行推定



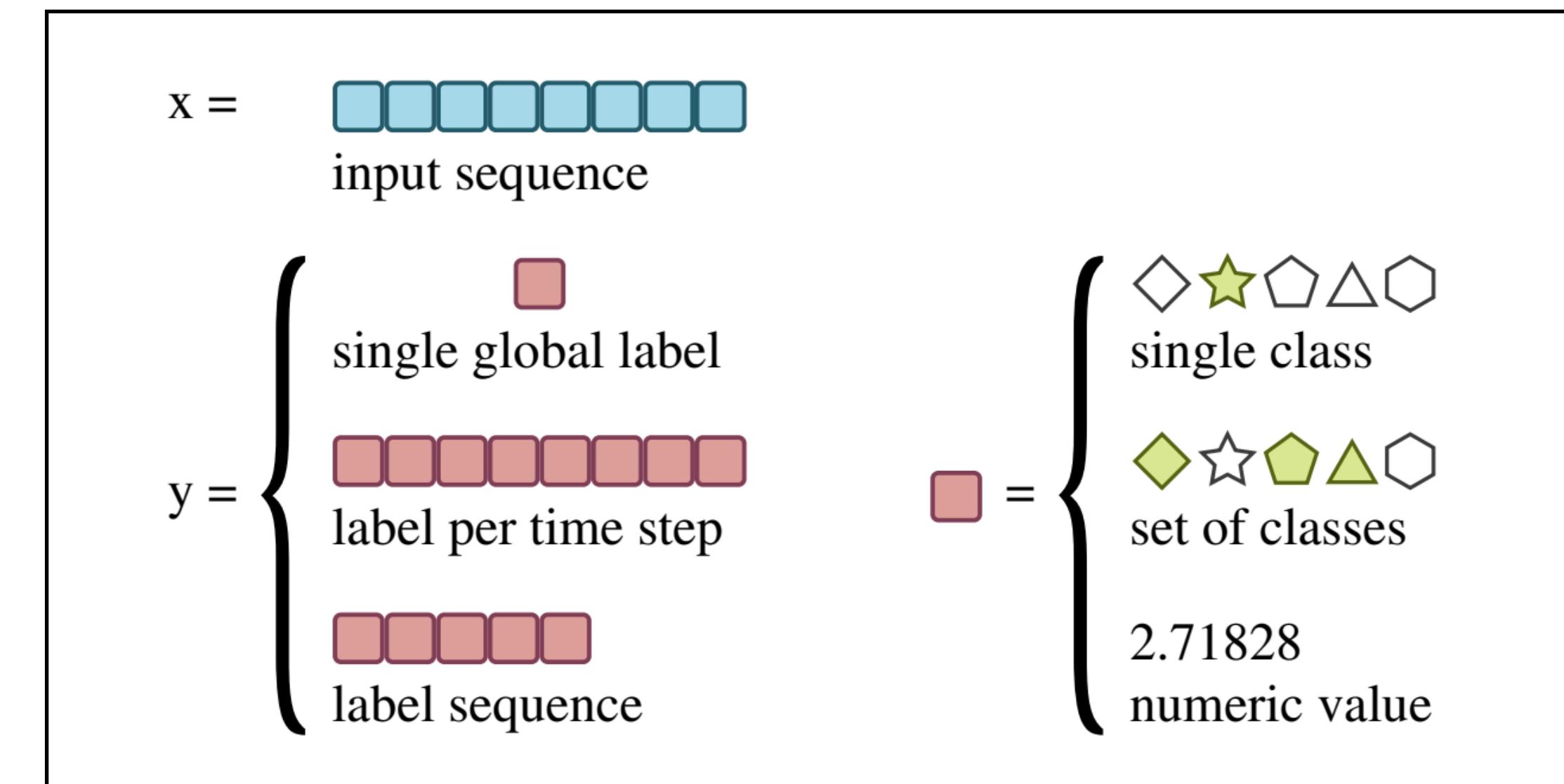
Am
G
C
Em

正解
ラベル



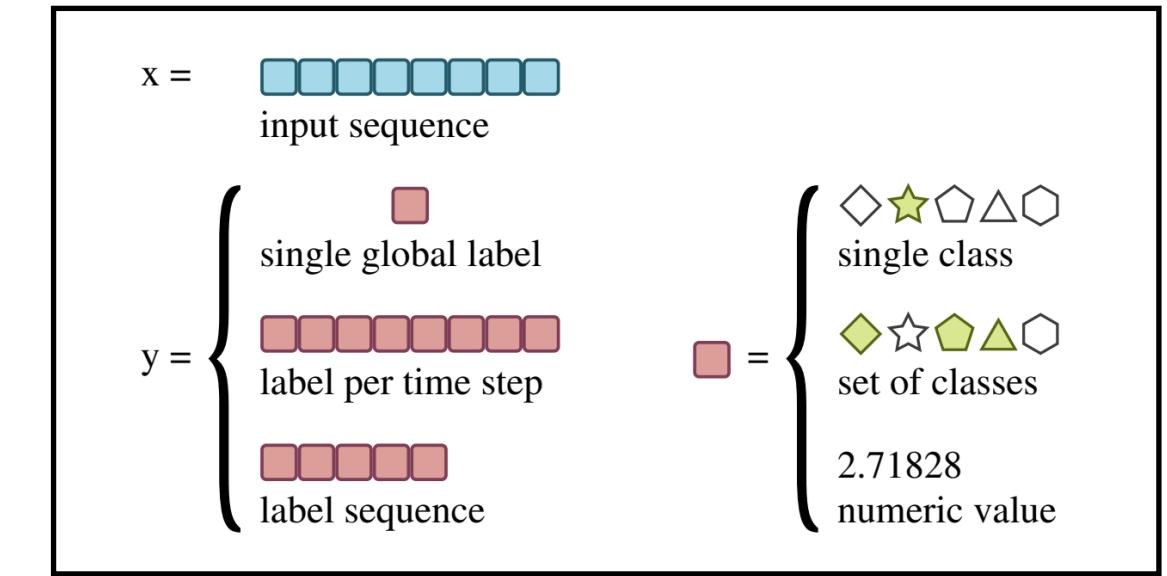
入力系列X（波形 or スペクトログラム）に対する出力Yの取り方の違いで分ける

- ラベルの種類
 - single global label: 系列全体に対するラベル
 - label per time step: 系列の各時刻に1対1対応するラベル
 - label sequence: 系列の各時刻に1対1対応しないラベル
- 何を推定するか？
 - single class: 分類クラスの中のどれか1つだけに分類する
 - set of classes: 複数のクラスの分類を許容, 各クラスの出現/非出現の分類を1つのモデルで独立に解く
 - numeric value: 連続値



音楽関連タスクの分類

7

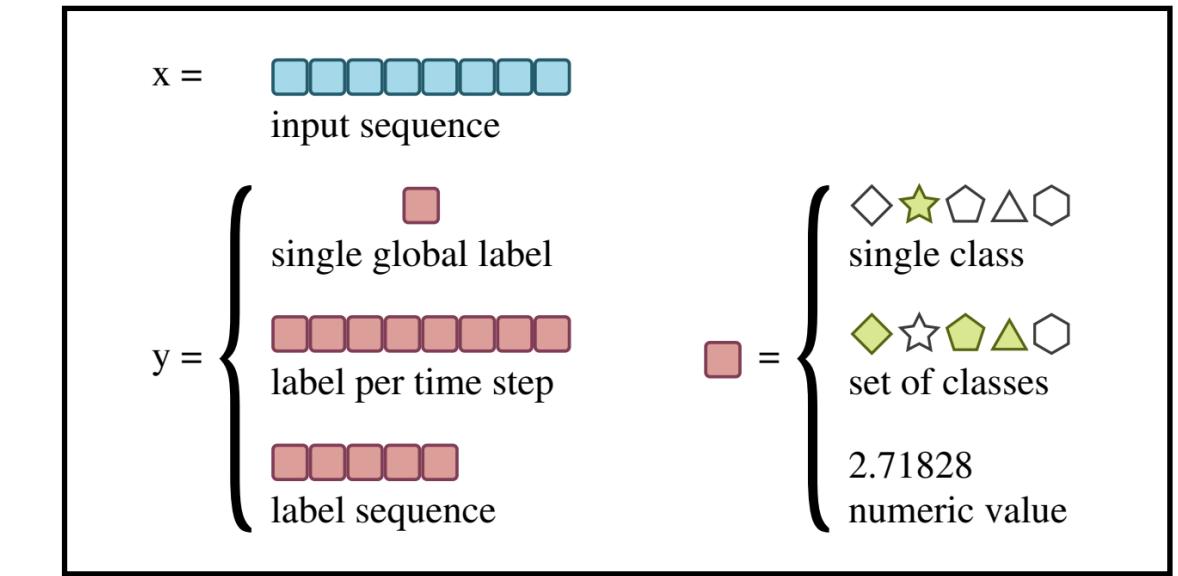


ファイル単位で推定 (Single global label)		時系列で推定 (Label per time step/ Label sequence)
单クラス分類 (single class)	音楽ジャンル分類 楽器・歌手識別	ビート推定 コード推定 自動採譜 (単音)
マルチクラス分類 (set of classes)	音楽タグ付け	自動採譜 (複音) 楽器アクティビティ識別
回帰 (numeric value)	音楽同士の類似度 テンポ推定	ピッチ推定 (音そのもの (音源分離等))

主に用いる最終層のアクティベーションとロス関数

8

- 問題設定に合わせて最終層をカスタマイズしよう

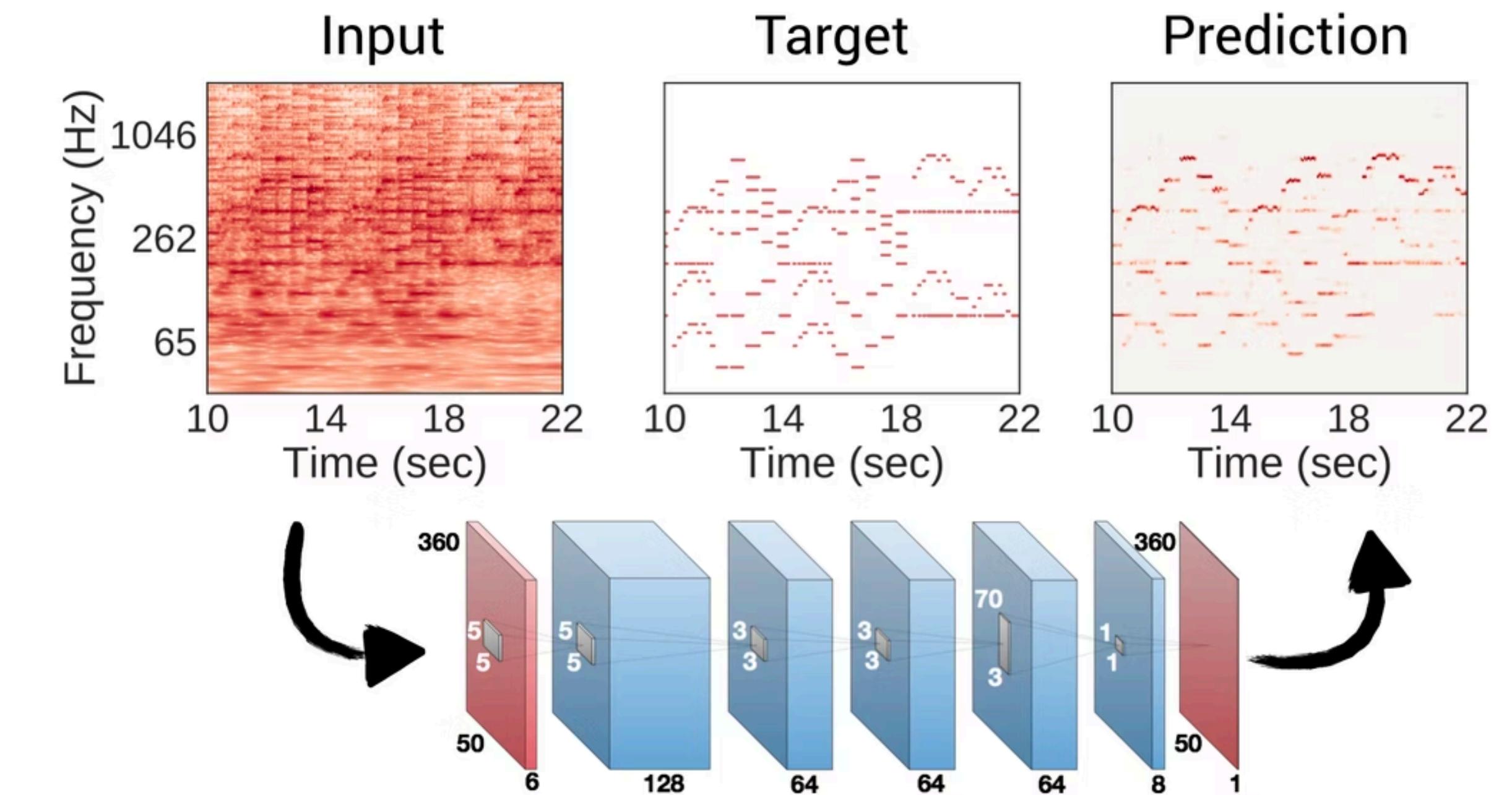


	ファイル単位で推定 (Single global label)	時系列で推定 (Label per time step/ Label sequence)
单クラス分類 (single class)	Softmax Cross-entropy	Softmax Cross-entropy
マルチクラス分類 (set of classes)	Sigmoid Binary-cross entropy	Sigmoid Binary-cross entropy
回帰 (numeric value)	(欲しい回帰値の値域による) L1, L2Loss	(欲しい回帰値の値域による) L1, L2Loss

NNのセッティング例①：CNN

9

- 時間方向にプーリングをせずに、同じ次元を出力するようにする
 - ex. Deep salience
 - 各時刻、各周波数ビンにおける、音高の存在確率の特徴表現



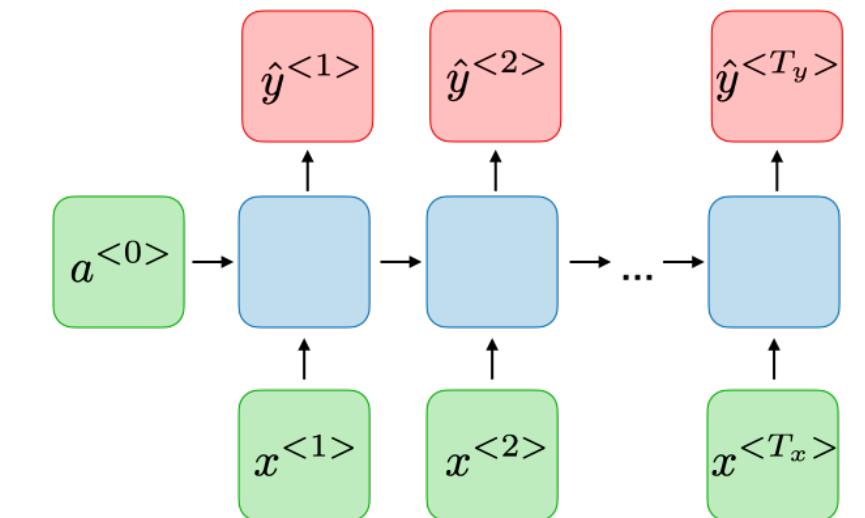
NNのセッティング例②：RNN/CRNN

- **Many-to-Many RNNを用いる**
- 出力の形式から対応あり/なしが決まるので、それに合うモデルを使う

音のフレーム単位で正解ラベルをつけることが可能である場合

(区間、ピアノロール等)
→対応あり

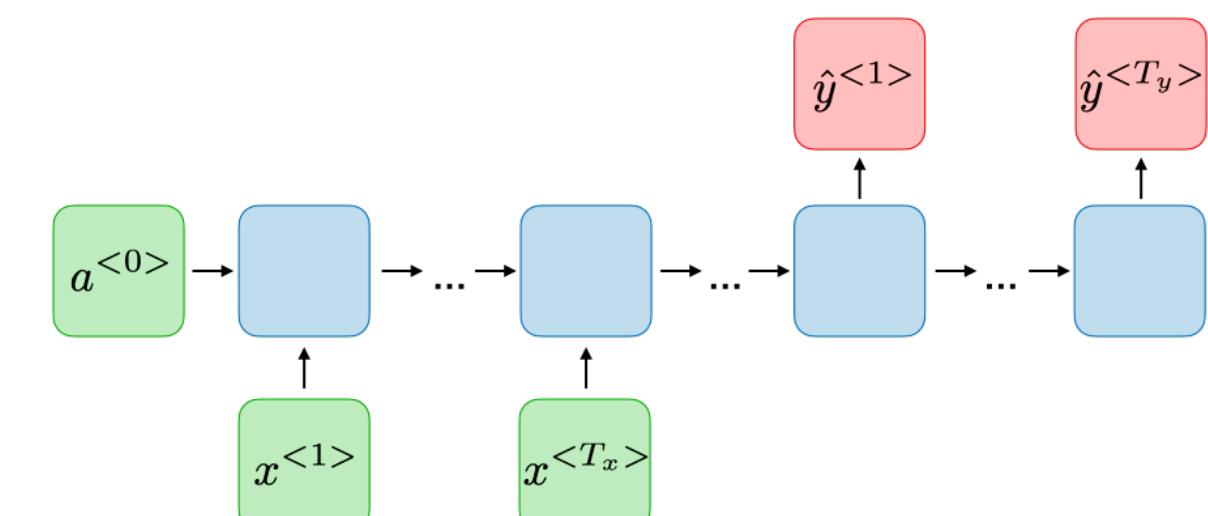
(RNNの各ステップの出力を用いる)



音のフレーム単位で正解ラベルをつけることが可能でない場合

(歌詞、音符等)
→対応なし

(Seq2Seqを使う)

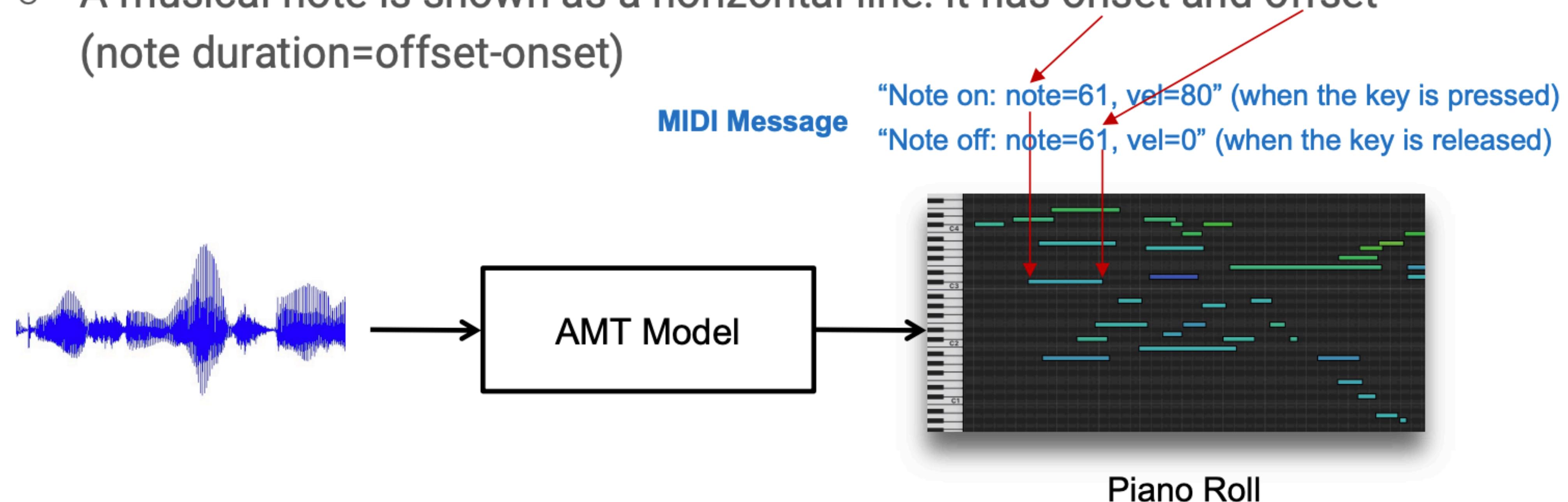


自動採譜の例

自動採譜 (Automatic Music Transcription; AMT) とは

12

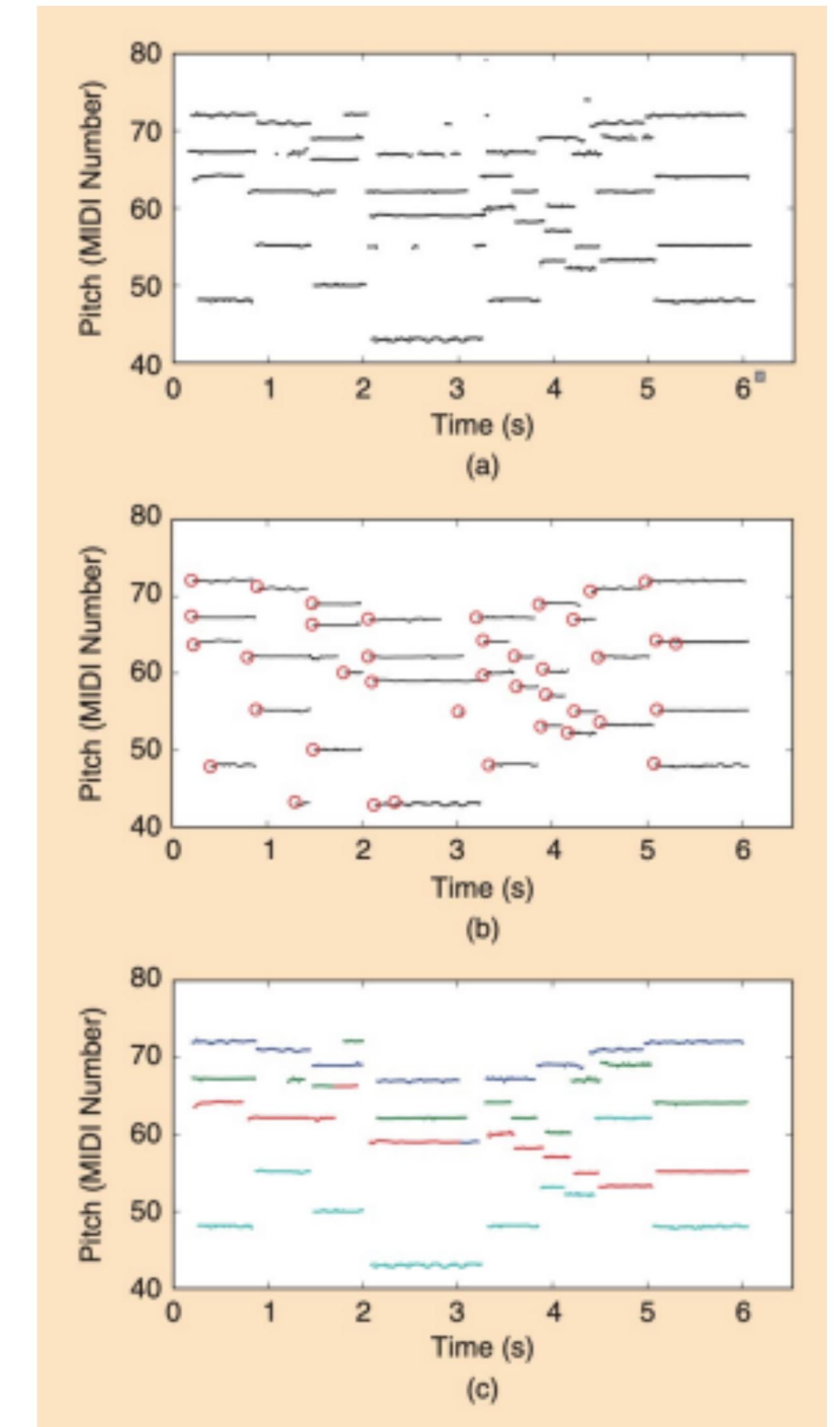
- 音響信号から音符への変換タスク
 - 多くの場合、音符そのものではなくピアノロールの形に変換
 - A musical note is shown as a horizontal line: it has onset and offset (note duration=offset-onset)



- 音楽情報処理における難しい課題の一つ

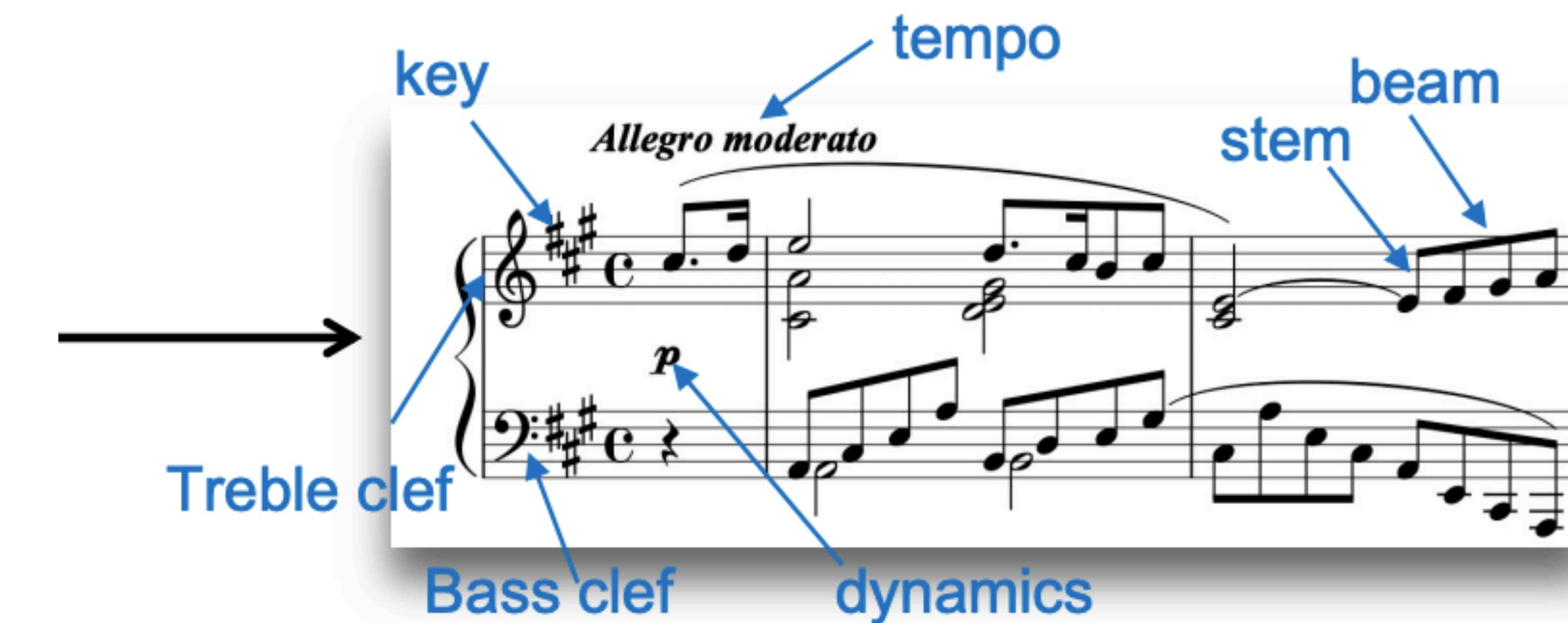
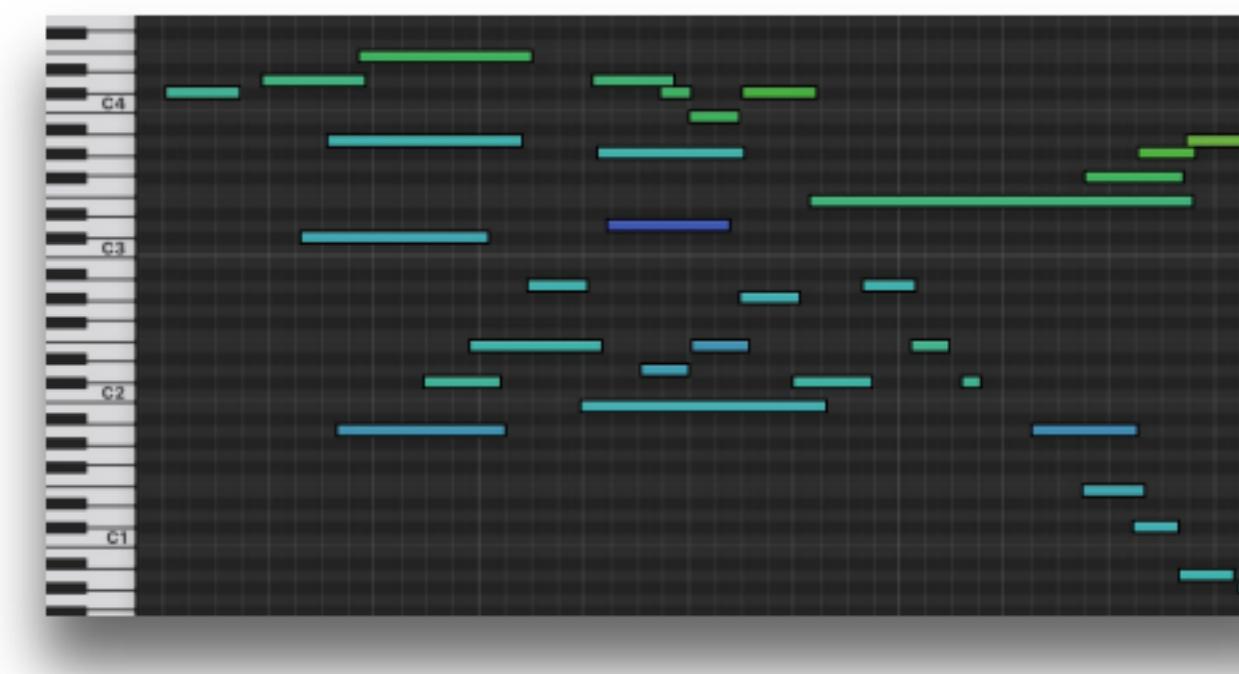
自動採譜が含むサブタスク

- ピッチ推定（フレーム毎）
 - ピッチ：離散値（半音 or cent 値）または連続値
 - 単音も複合音も考えられる（演奏音に依る）
- オンセット/オフセット検出（音符ごと）
 - （オフセット検出は考慮しない場合も）
- 楽器識別
 - 複数の楽器の演奏音が含まれる場合
 - 楽器の種類と楽器数は事前に与えることも
 - 音源分離と捉えることもできる



演奏音 -> 楽譜 (Sheet music) への変換

- 究極の目標。かなり難しく未到達
 - 拍節解析：テンポ，ビート/ダウンビートの推定
 - 調性：ハ長調，二短調というような，調の推定
 - 音符：音部，連符等，記譜に関わる部分の推定
 - 表現：ダイナミクス (f,p等)，アーティキュレーション (スタッカート等)，フレージング等



- **ピッチ推定**

- 単音の周波数推定：YIN, ケプストラム法, CREPE, etc...
- 複合音の周波数推定：反復推定, NMF, Deep salience, etc...
- ピアノロール推定：深層学習（Onset and Frames）
- 混合音からのメロディ検出：Melodia, 深層学習（JDC, Segnet等）

- **ビート推定**

- DP, HMM, 深層学習（Many-to-Many RNN, TCN等）

- **コード進行推定**

- HMM, CRF, 深層学習（CRNN 等）

事例① ピアノ採譜：Onset and Frames

16

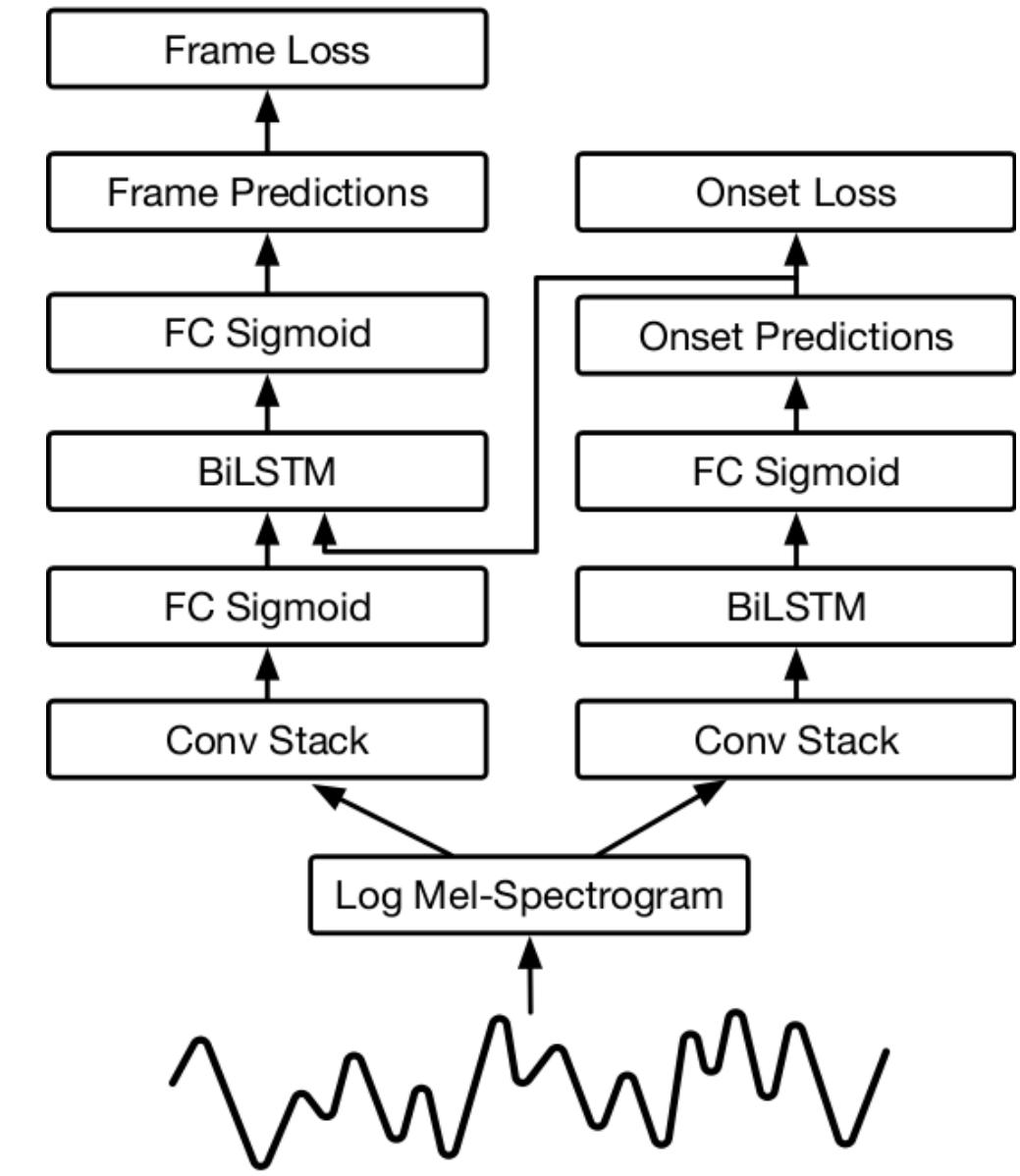
非常に高い精度での採譜を実現

- Googleが2018年に発表した「Onset and Frames」

- モーツアルト ソナタ K.331 第3楽章 (トルコ行進曲)

入力

出力

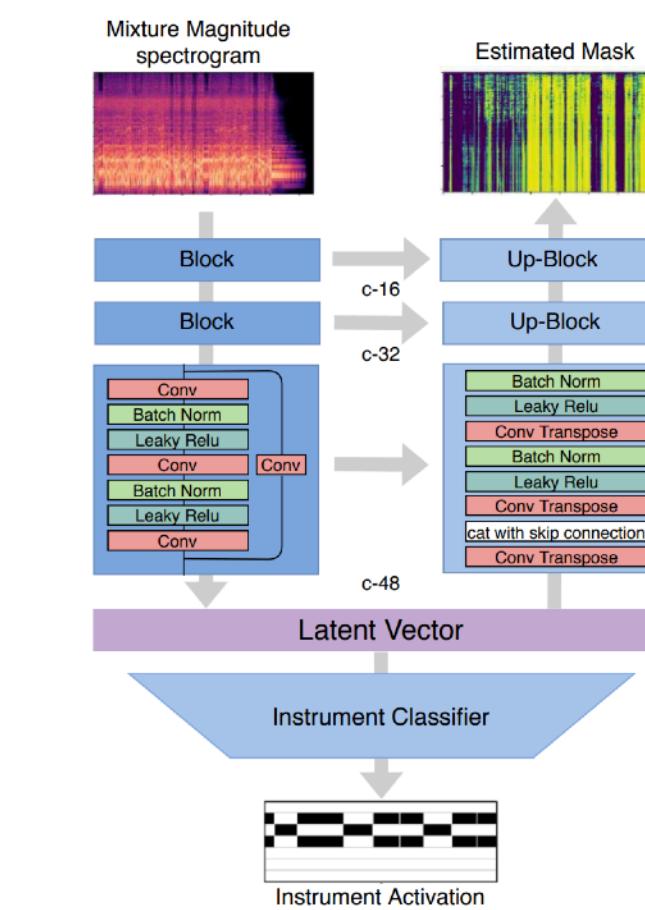


- 入出力：ピアノ演奏音源 -> 採譜結果（楽譜）
- ラベル：音の立ち上がり（オンセット）時刻 + その音高 （マルチタスク学習）

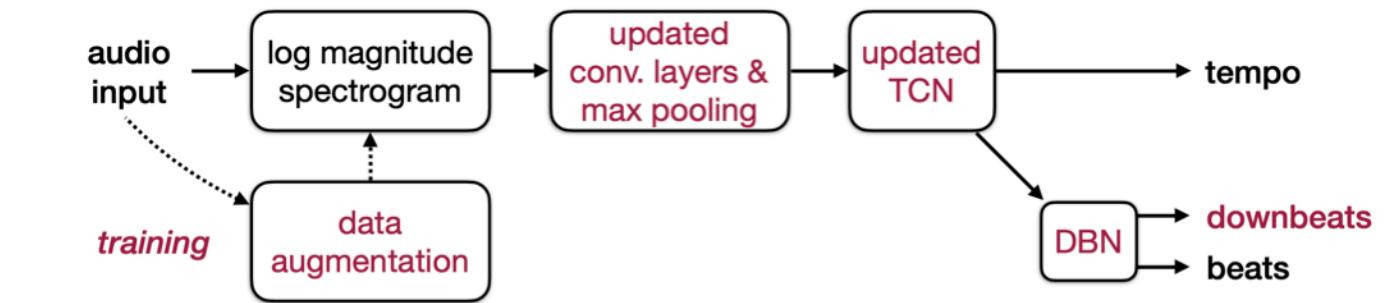
マルチタスク学習 (Multi-task learning)

17

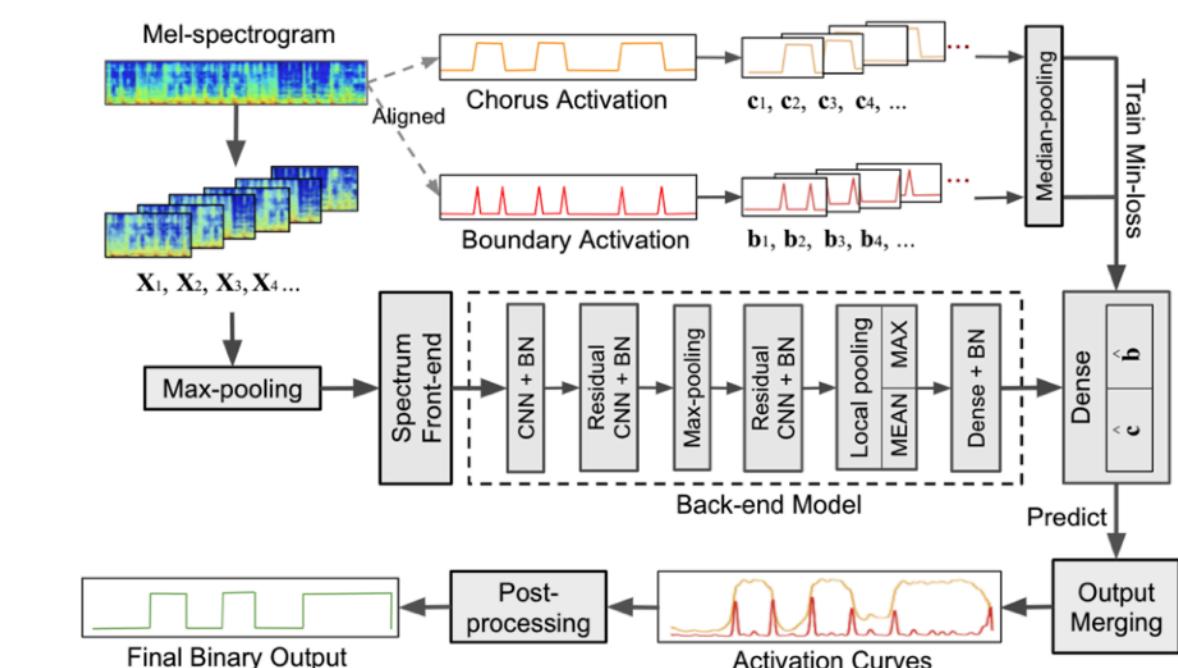
- 一つのモデルで複数のタスクを同時に解く枠組み
 - タスク間に共通点が多いとき、お互い共通に使える特徴を捉えることが可能
- 自動採譜では部分問題を追加タスクに加えることが多い
- モデルにとって学習するのが難しい場合も...



楽器識別と音源分離
MULTITASK LEARNING FOR INSTRUMENT ACTIVATION AWARE MUSIC SOURCE SEPARATION
Y. Hung et al. ISMIR 2020



テンポ、拍節とビート
Deconstruct, Analyse, Reconstruct: How to Improve Tempo, Beat, and Downbeat Estimation
S. Bock et al. ISMIR 2020

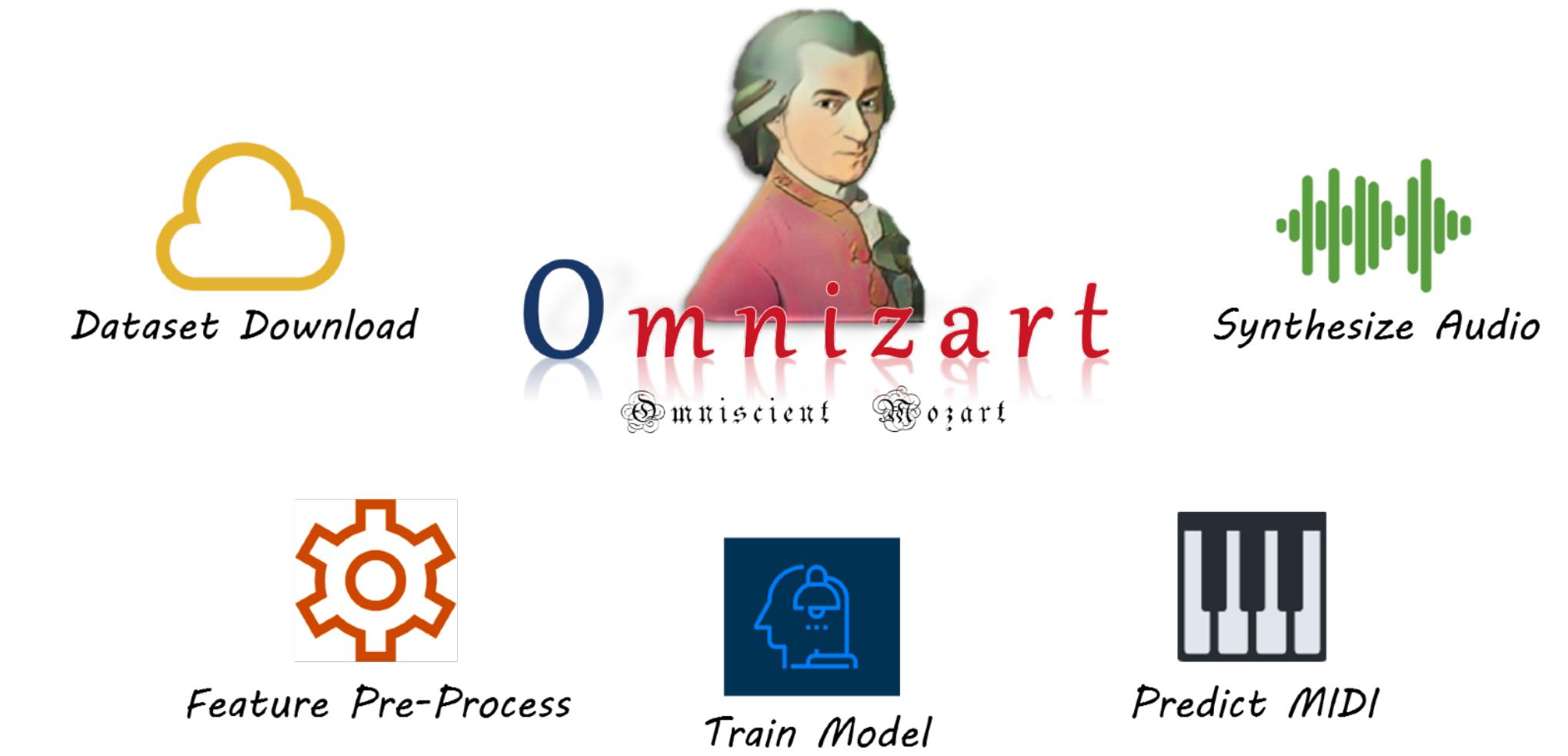


サビ検出とセクション変化検出
SUPERVISED CHORUS DETECTION FOR POPULAR MUSIC USING CONVOLUTIONAL NEURAL NETWORK AND MULTI-TASK LEARNING
J. Wang et al. ICASSP 2021

Fig. 1. The system diagram.

さまざまな楽器の採譜が可能

- 台湾・Academia Sinicaの「Omnizart」
 - 歌声(音符&ピッチ), ピアノ, ドラム, コード, 拍節を採譜可能
- 入出力：演奏音源 -> 採譜結果（楽譜）
- ラベル：楽譜（ピアノロール）



<https://music-and-culture-technology-lab.github.io/omnizart-doc/index.html>

- 産総研, 京大, 早稲田, 明治, クリプトンの共同プロジェクト
 - 自動採譜の部分は京大が担当
 - <https://www.youtube.com/watch?v=mmGCFzQFbJg>

- MIREX (Music Information Retrieval Evaluation eXchange)
 - 自動採譜のサブタスク等、さまざまなタスクを競うコンペ
 - https://www.music-ir.org/mirex/wiki/MIREX_HOME
 - いろんなタスクがあるので見るとよさそう
- チュートリアル（ビート推定）
 - （注：コードはkeras）
 - <https://tempobeatdownbeat.github.io/tutorial/intro.html>

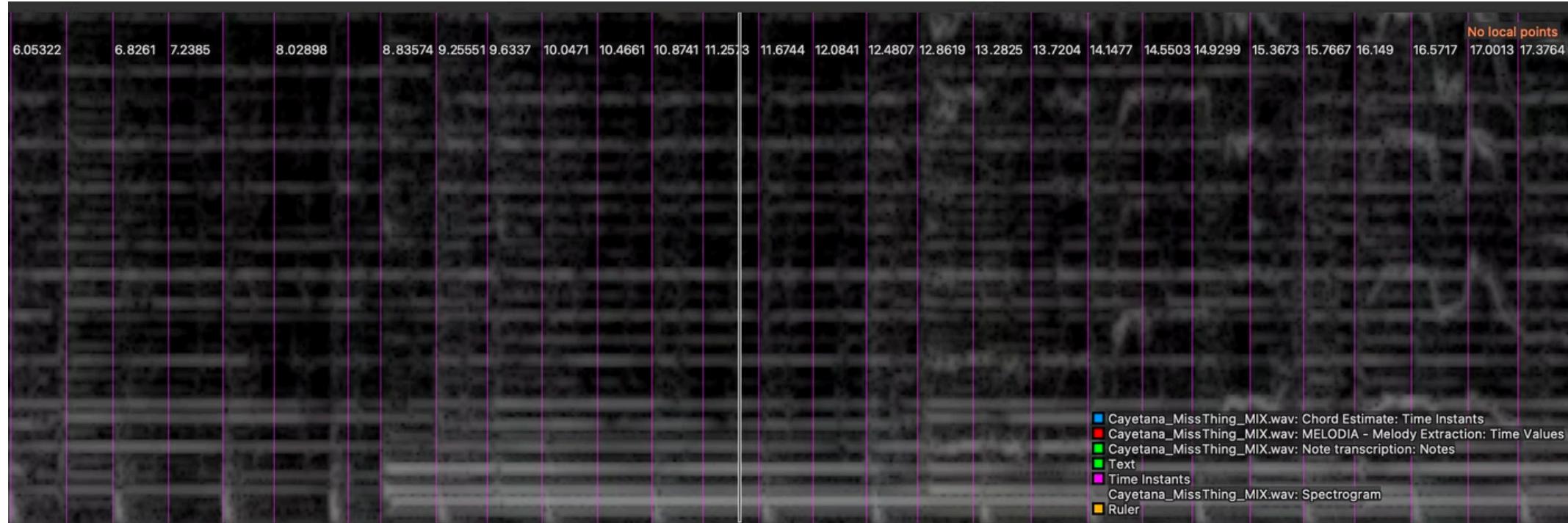
EOF

補足

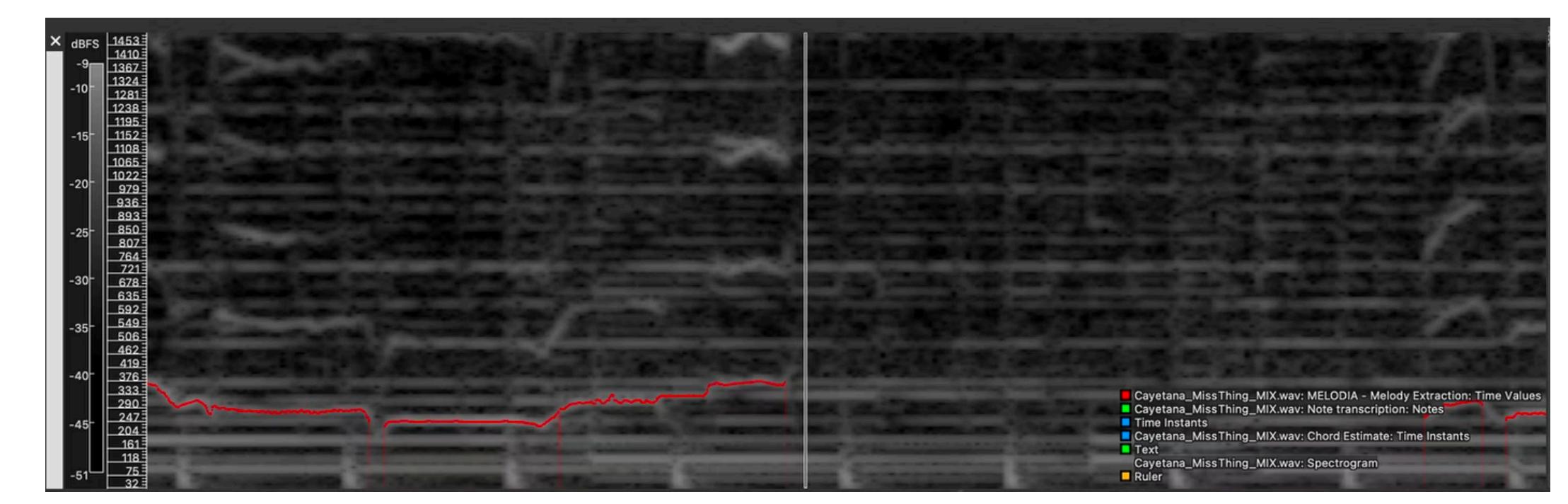
ラベルの種類

23

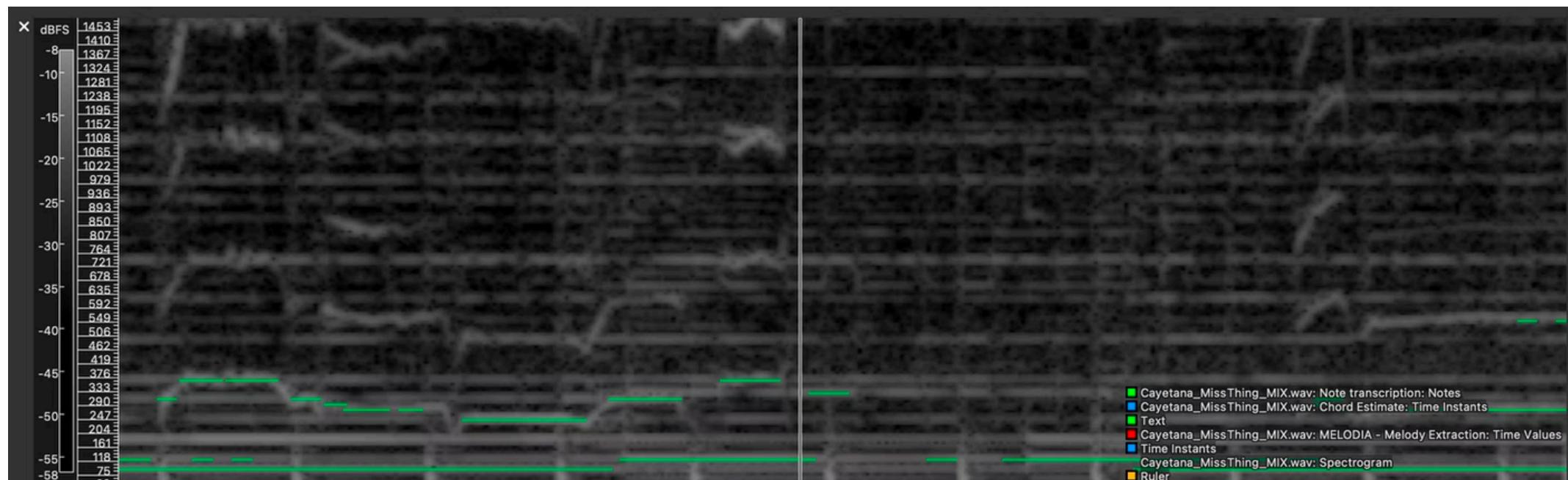
<https://qiita.com/yamathcy/items/db0626d01bb2c1f40107>



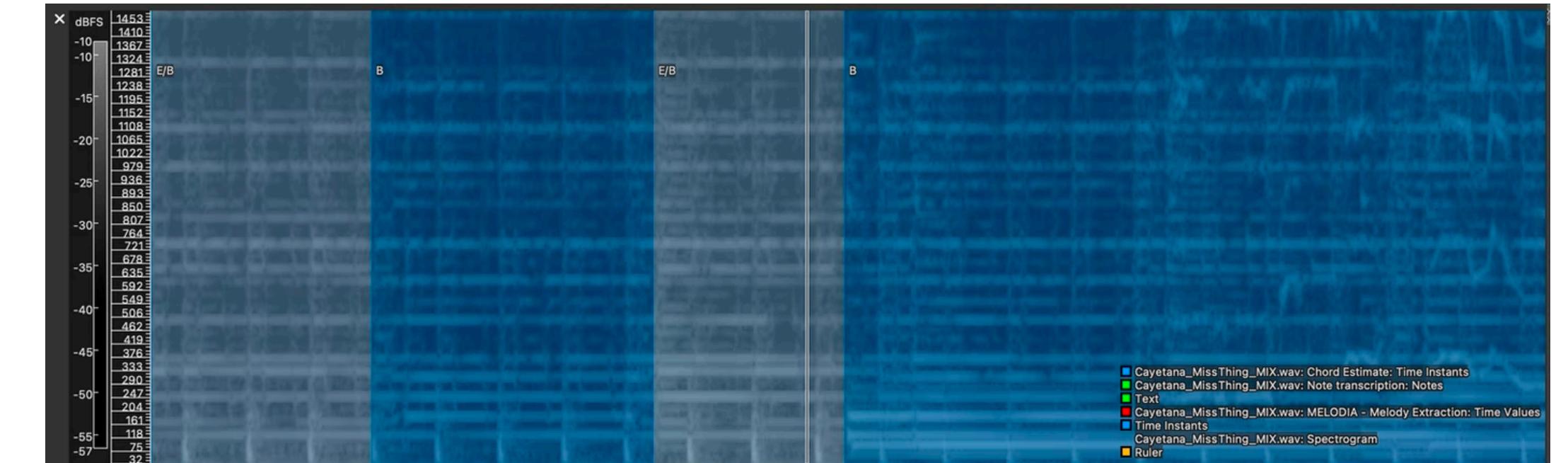
タイムインスタンス：ラベル+時刻
ビート， ドラムの音符等



タイムバリュー：連続値+時刻
ピッチ等



音符：音高+開始時刻+終了時刻

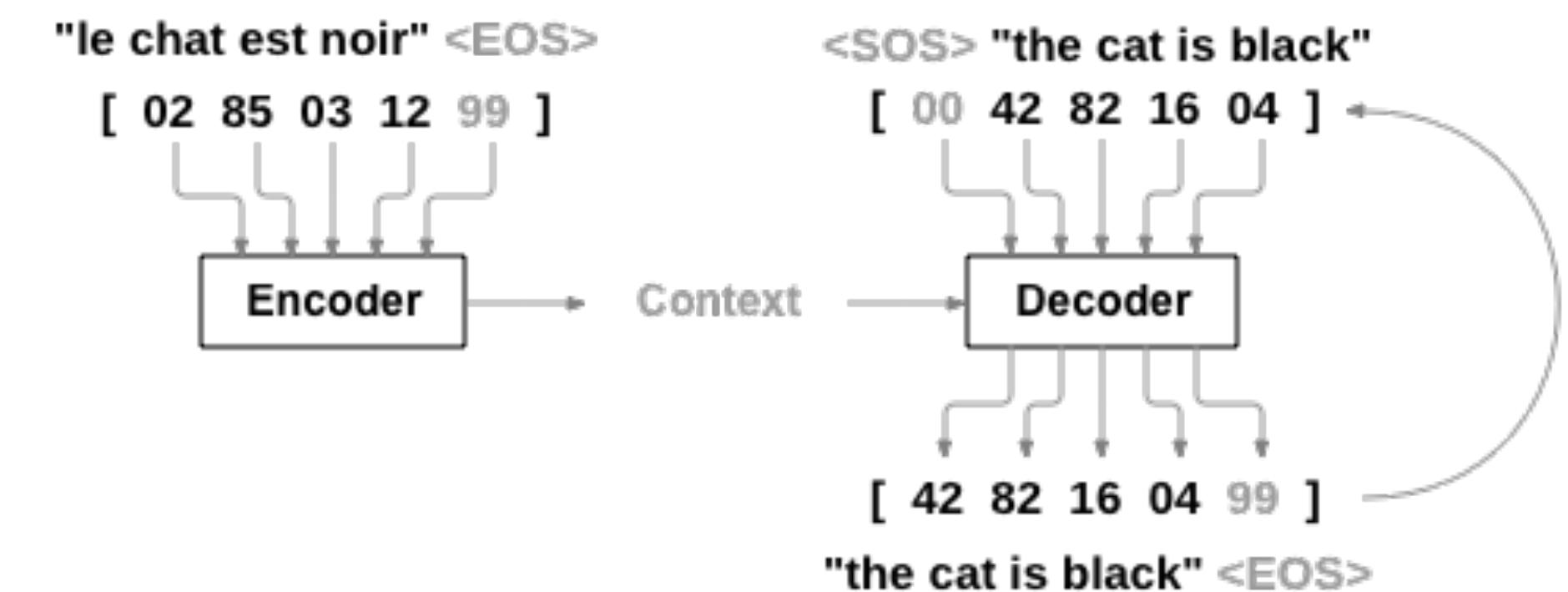
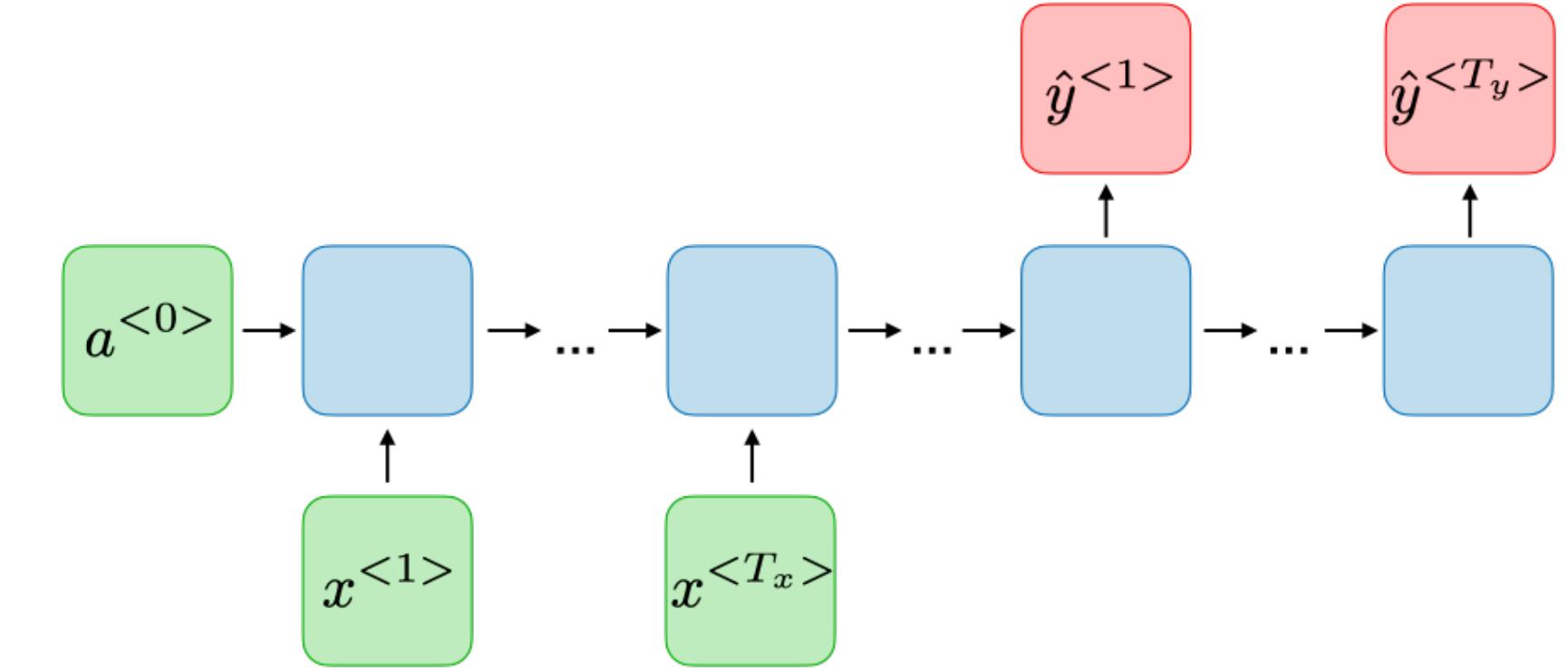


リージョン：ラベル+開始時刻+終了時刻
コード進行， セクション等

Seq2Seq (Many-to-Many RNN)

24

- 入出力が1対1対応しない系列の変換でのモデル
 - エンコーダー・デコーダーモデル（一般化）
- 実装上では2つの部分に分ける
 - エンコーダー -> 入力からコンテキストを得る
 - many-to-one
 - デコーダー -> コンテキストから出力を得る（自己回帰形式）
 - one-to-many

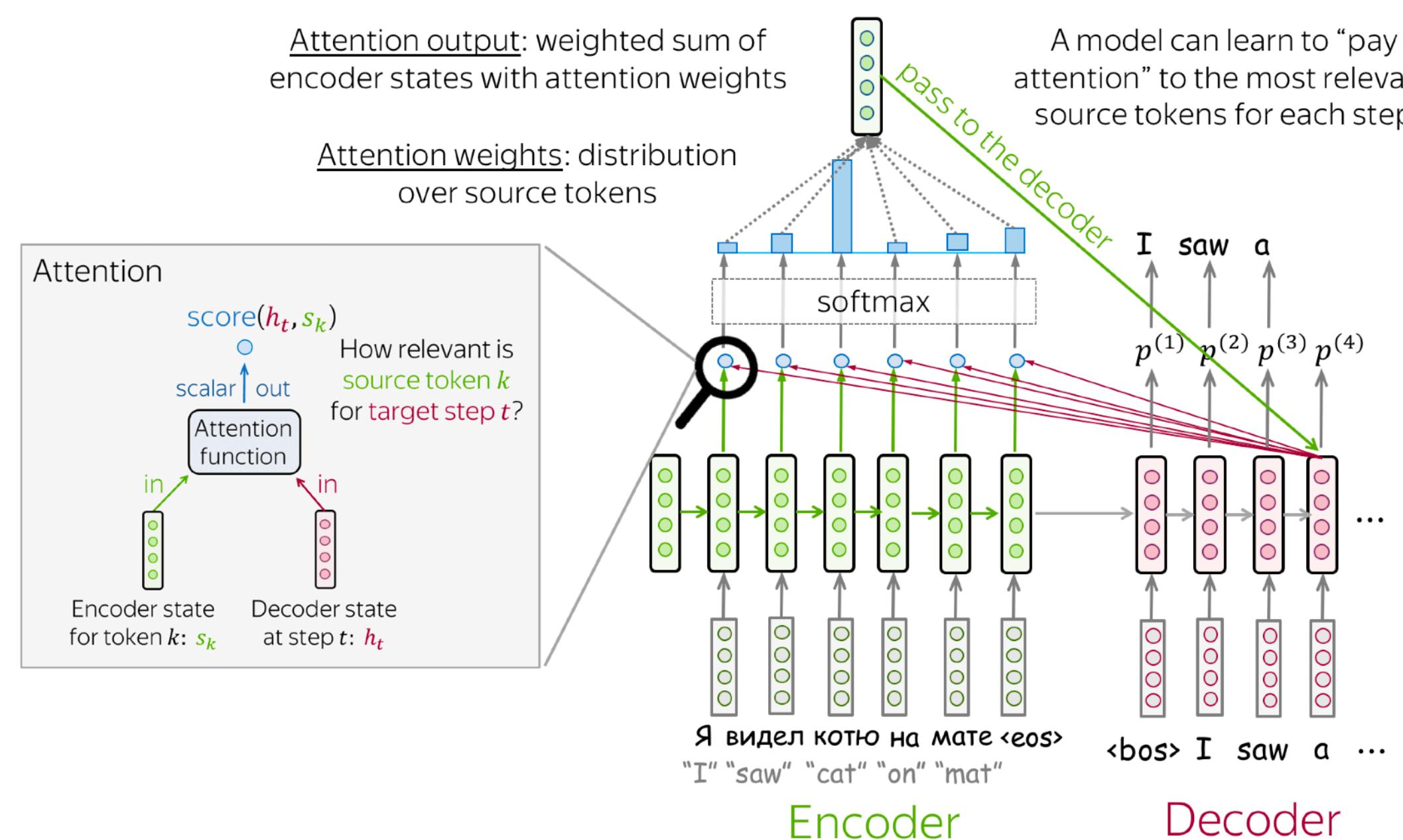


https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html

Attention付きSeq2Seq

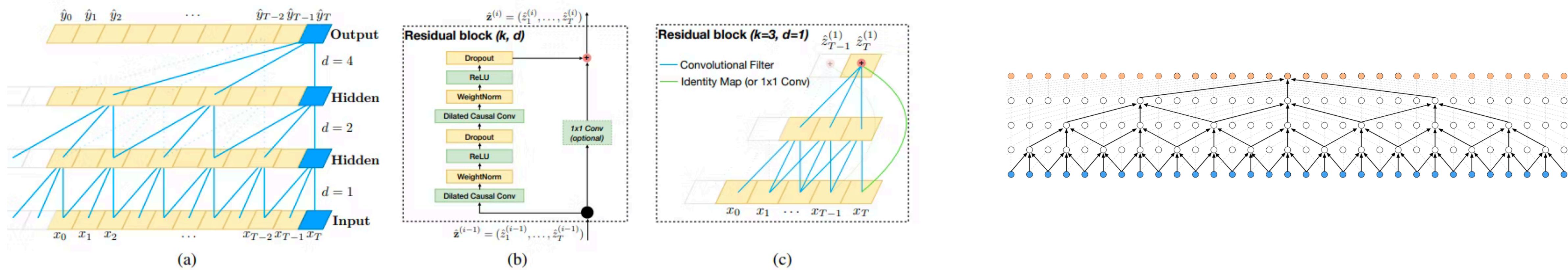
25

- Seq2Seqでは入力を全て固定長のベクトルにしていた
- →入力の長さが異なるが全て同じ情報量を保つ
- →Attention（注意機構）を用いて、入力のどこに着目するかを決定

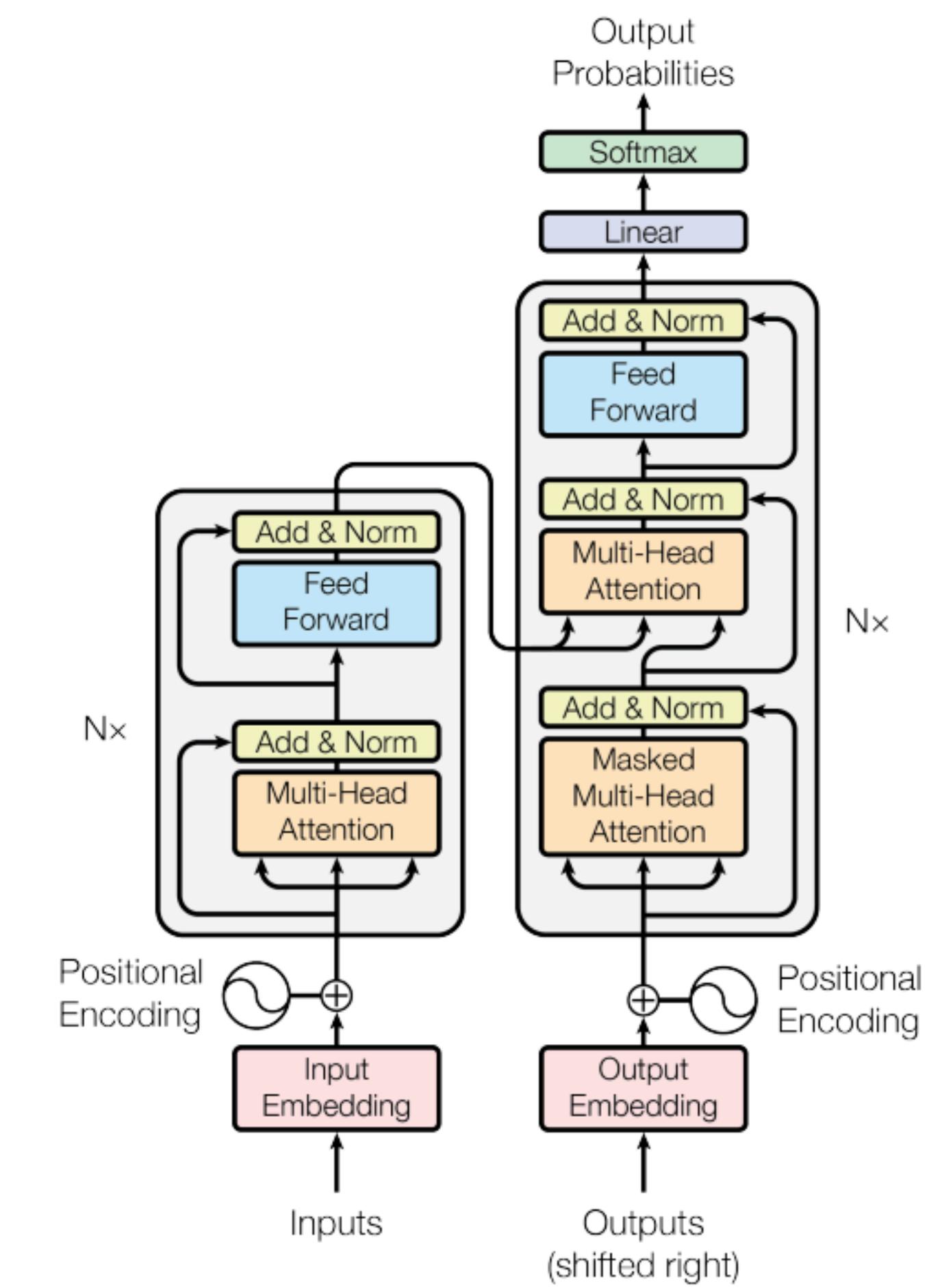


https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html

- **Temporal Convolutional Network (TCN)**
- 時刻方向に1次元畳み込みを行う + 受容野をdilatedして拡大
- 学習が速いし、精度もLSTMよりもよくなる場合もある（山本の感想）



- Transformer
 - 「なんか最近きてるやつ」
 - 注意機構（Attention）に基づくモデル
- CNNはカーネルの範囲内を、 RNNは系列の一つ一つを逐次に着目する
- Transformerは系列全てを一気に着目し、 Attentionによって注意するポイントを決める



詳しくは未来のTransformer回に

(略) :

下記資料のp.28-47をみてください

<https://speakerdeck.com/yushiku/end-to-end-object-detection-with-transformers> より

(Attentionの無い) CNNやRNNとの違い

Transformerは...

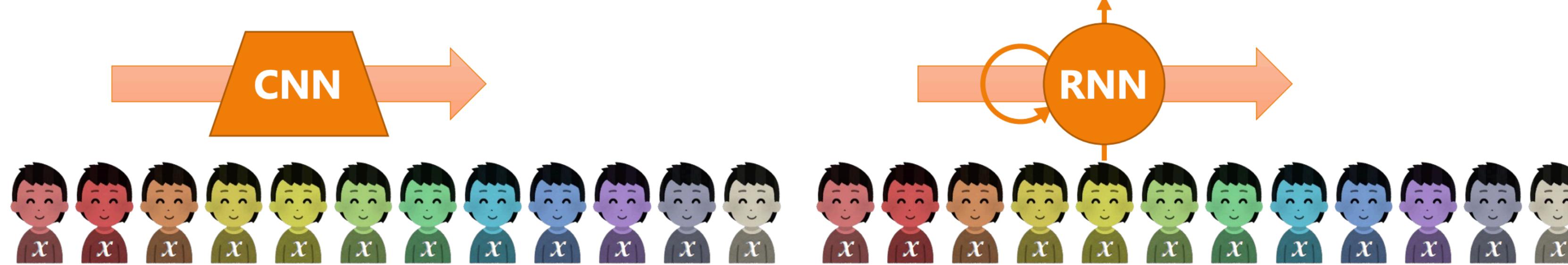
全 n 個のベクトルから全 n 個のベクトルへのアテンションを計算

$O(n^2)$ だが一番広域に情報がロス無く伝達可能

CNNは...

3つなど、全体からすれば少数のベクトルだけの畠込み計算を走査

$O(n)$ だが情報が伝わるのは近隣だけ



RNNは...

ベクトルを一つずつ走査しながら内部のセルに変数（記憶）を保存

$O(n)$ だが長い系列は不得手