Representation learning for symbolic music

PhD thesis

Mathieu Prang

Supervised by : **Philippe Esling**

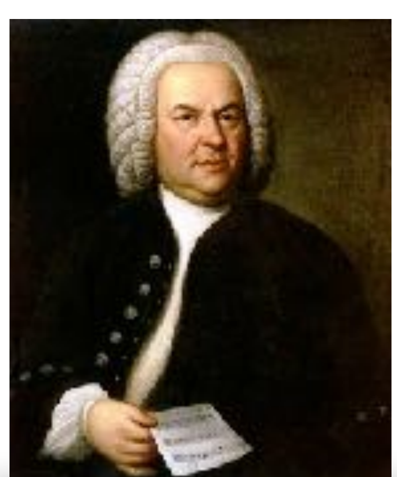
Directed by : **Carlos Agon**

IRCAM, Equipe Représentation Musicale

Introduction - musical representation

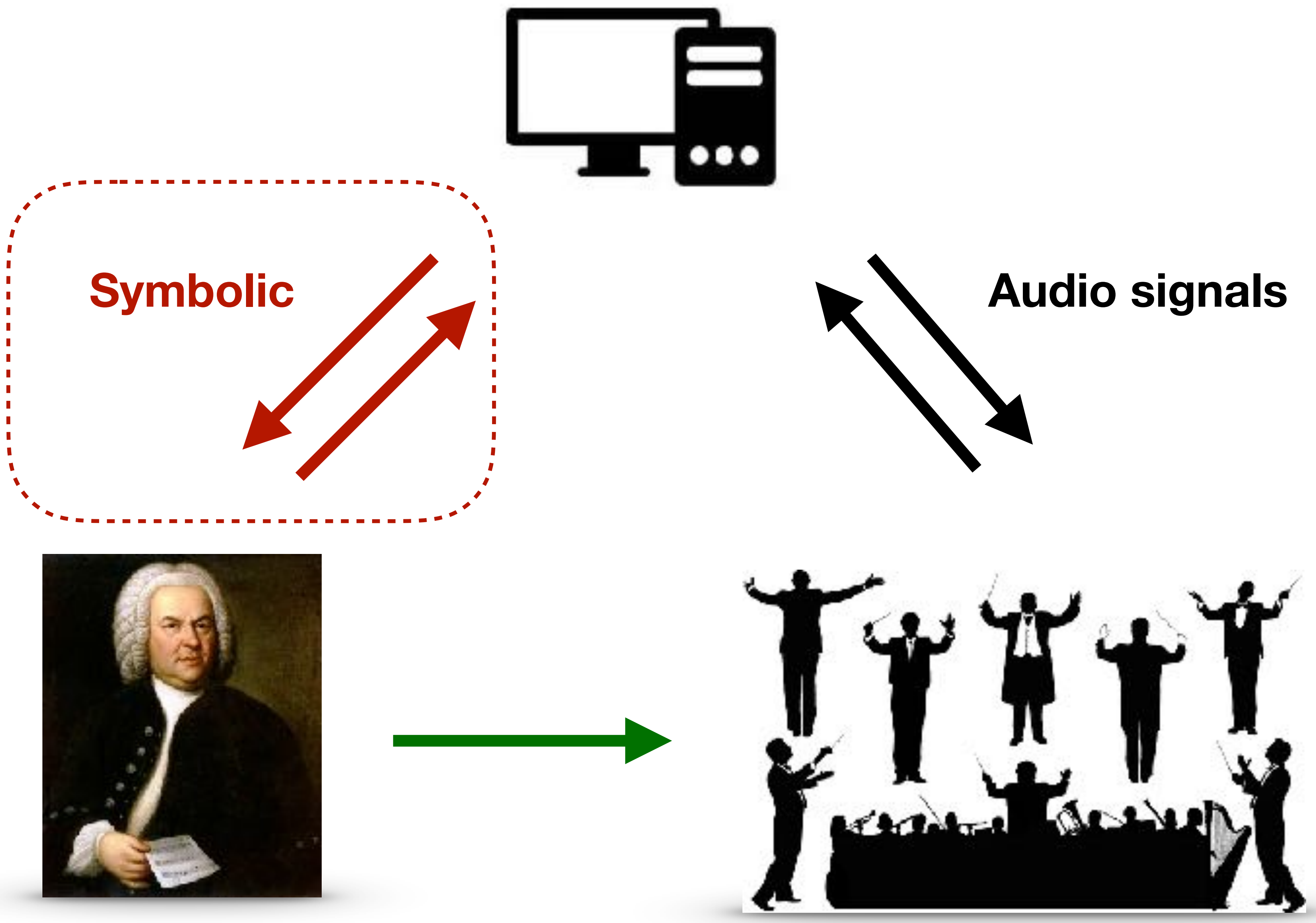
- Main symbolic representation : **the score**

- Provide information to **musicians** for reproducing the music as intended by **composers**



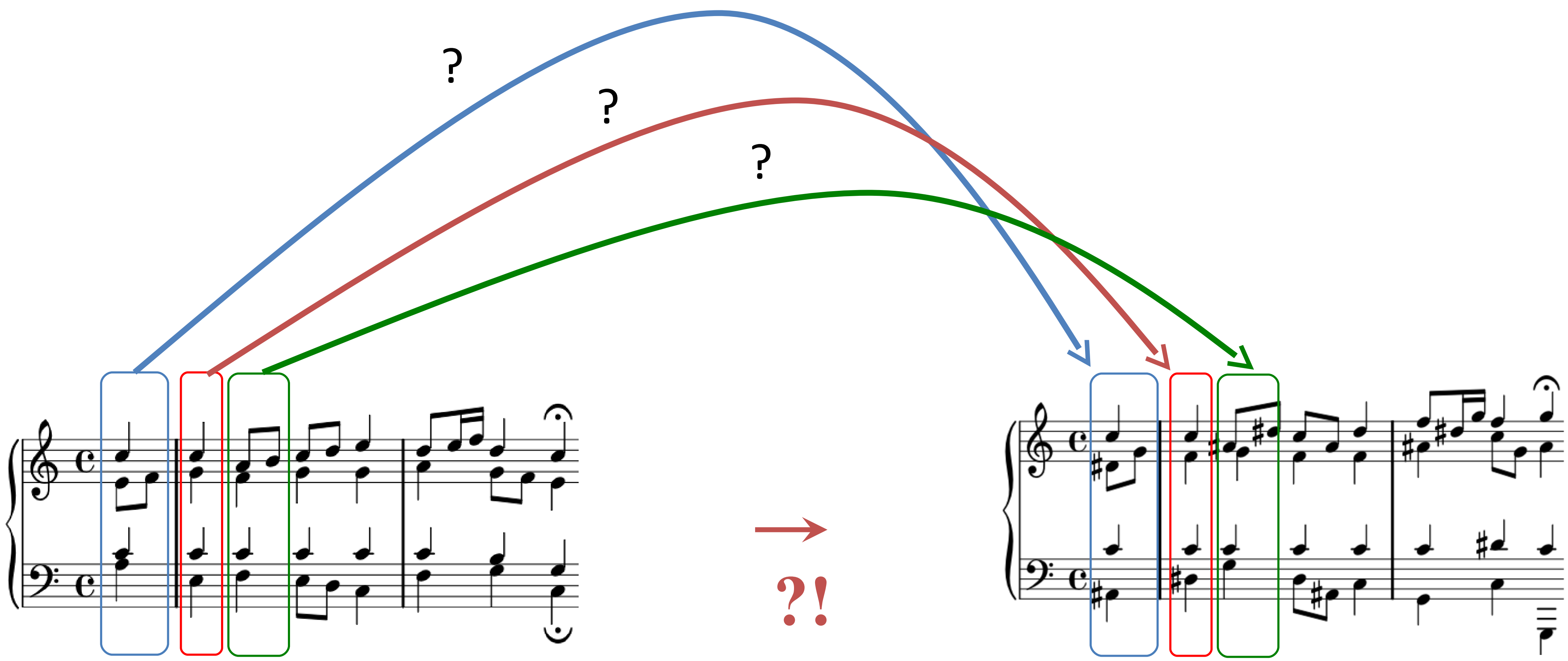
Introduction - musical representation

- Since the second half of the 20th century, the rise of **computer science** has opened new possibilities and new scientific challenges



Introduction - musical representation

- Example of a reharmonization



- The manipulation of scores is a **difficult task**
- Is done **sequentially**, required **multiple operations** and **musical knowledge**

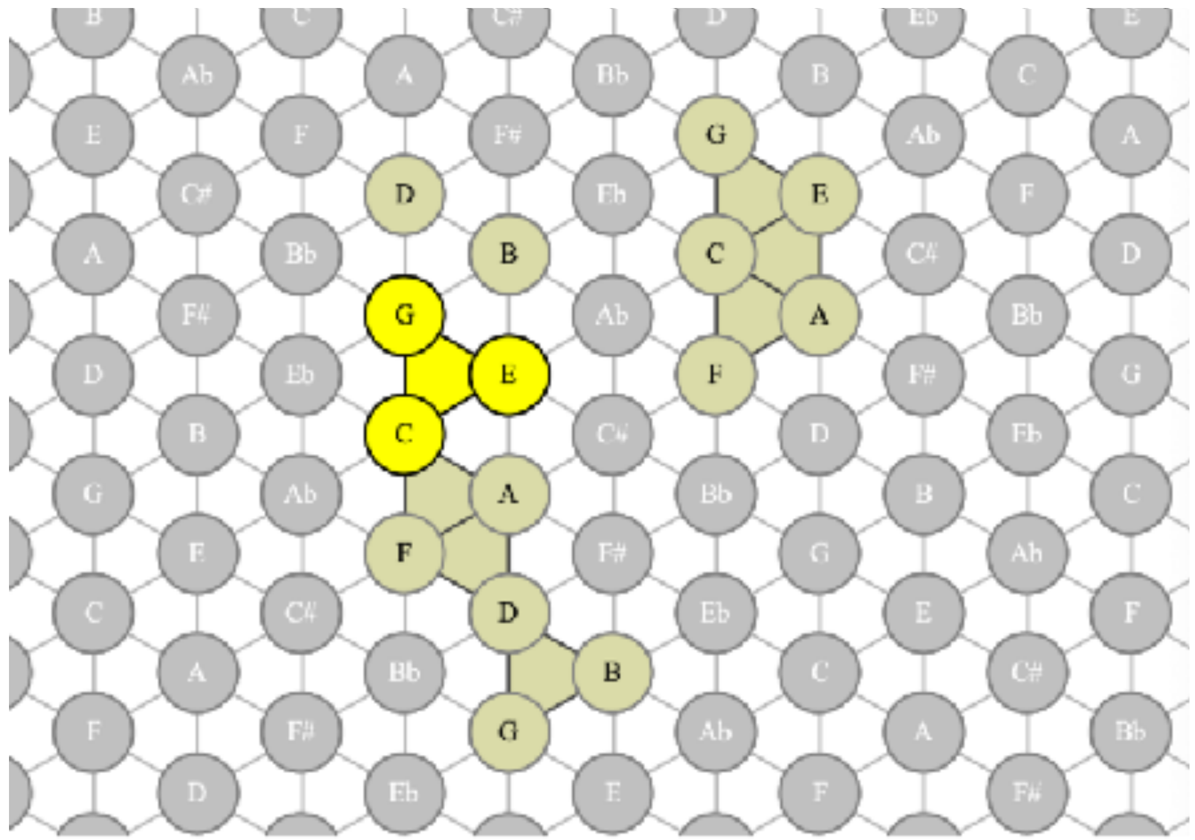
Introduction - musical spaces

- Use of musical spaces **structured** according to **music theory**

R chord - relative

L chord - counter relative

P opposite mode

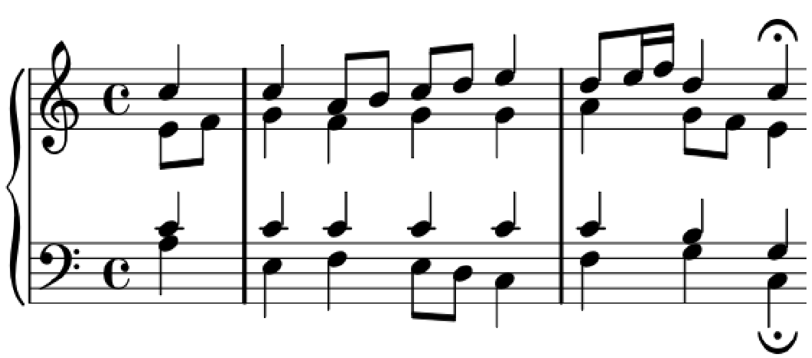


Structured ✓

*The **Tonnetz** : Pitch classes spatial representation*

K[3,4,5] ↑

↓ **K[2,3,7]**

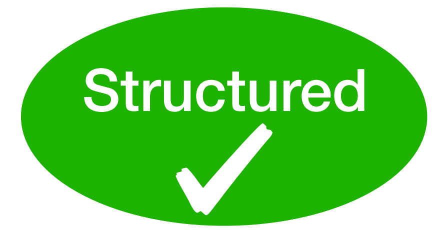


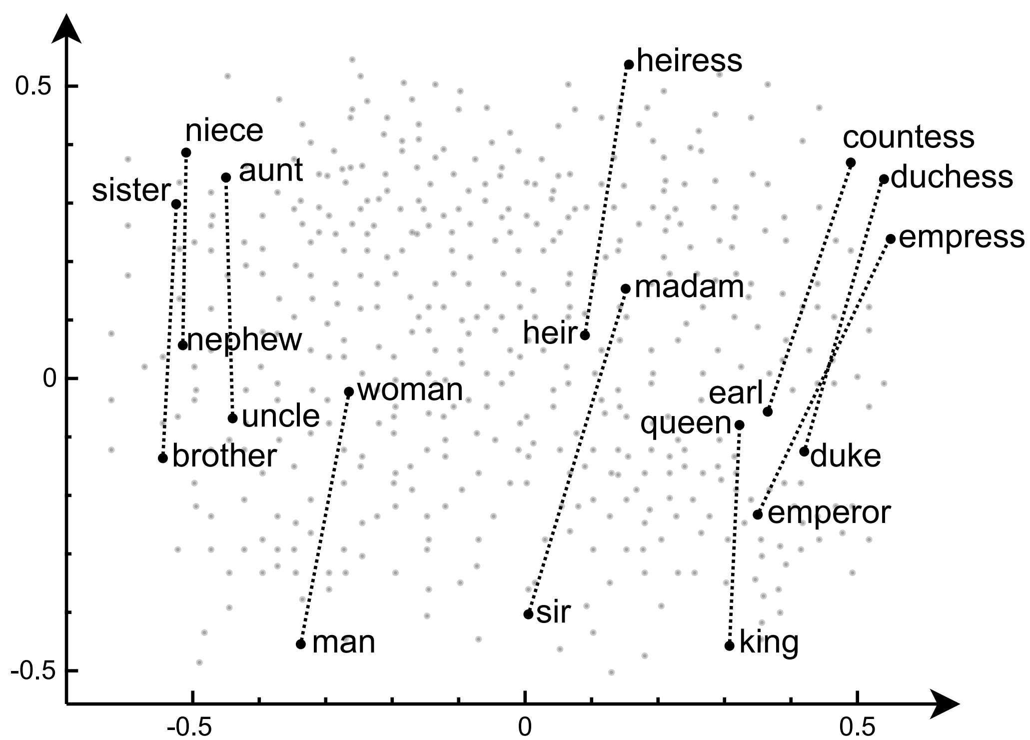
- The manipulation is **easy** and **fast**

Bigo, Louis, and Moreno Andreatta. "Topological structures in computer-aided music analysis." Computational music analysis. Springer, Cham, 2016. 57-80.

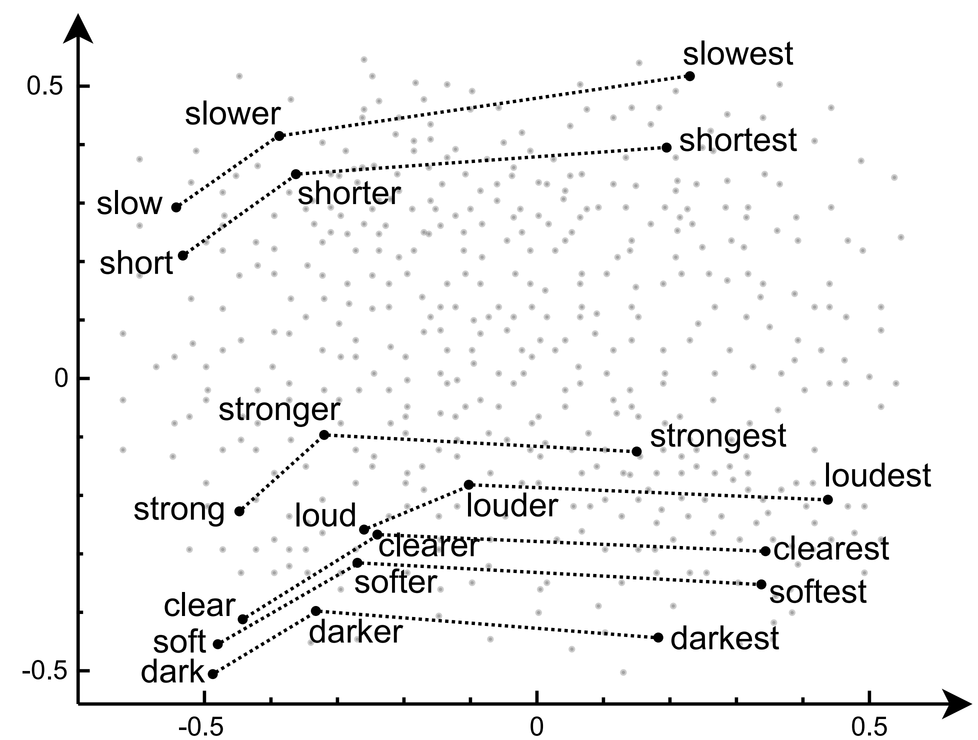
Introduction - embedding spaces

- **Machine learning** framework
- Design of **structured spaces**
- In the **Natural Language Processing** field : **word embeddings**
- Capture the **semantic relationships** between words
- Reflected in the **geometric structure** of the space.





Woman - man



comparative - superlative

- Carry **high level concepts** of the language
- Used as **input representation**

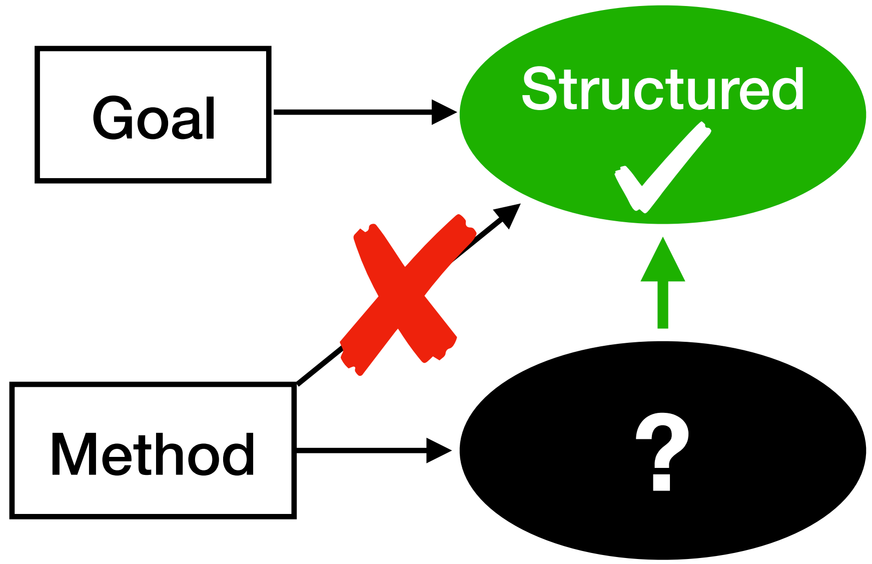
Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.

Introduction - Organization

- **Objectives** : develop ML algorithms for learning **self-structured embedding spaces** for **symbolic music**

- **No direct method**

- **Proxy task** required



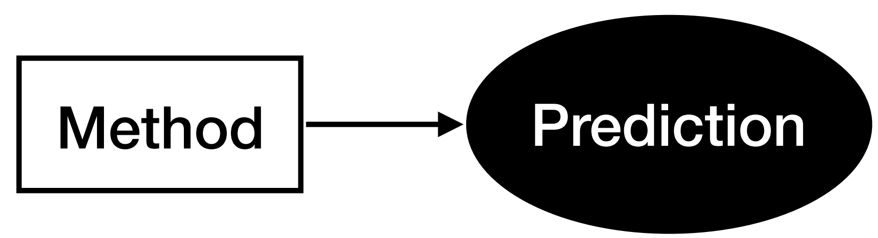
- **Dependent** on the quality of the space

- **Jointly** improve during training

- **Plan** :

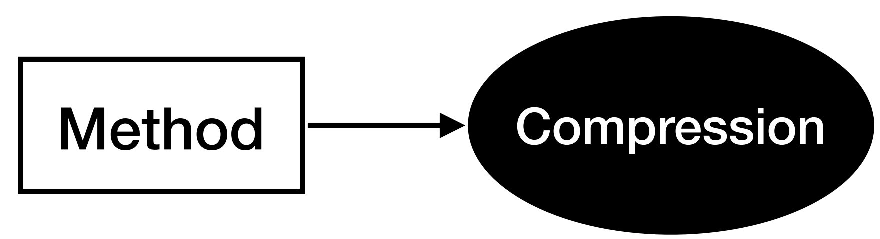
  I) First method - Adapt the NLP approach

  Overview - Contributions



  II) Second method - Variational Auto-Encoders

  Overview - Contributions



  III) Applications

  IV) Conclusion

I) First method - Adapt the NLP methods

II) Second method - Variational Auto-Encoders

III) Applications

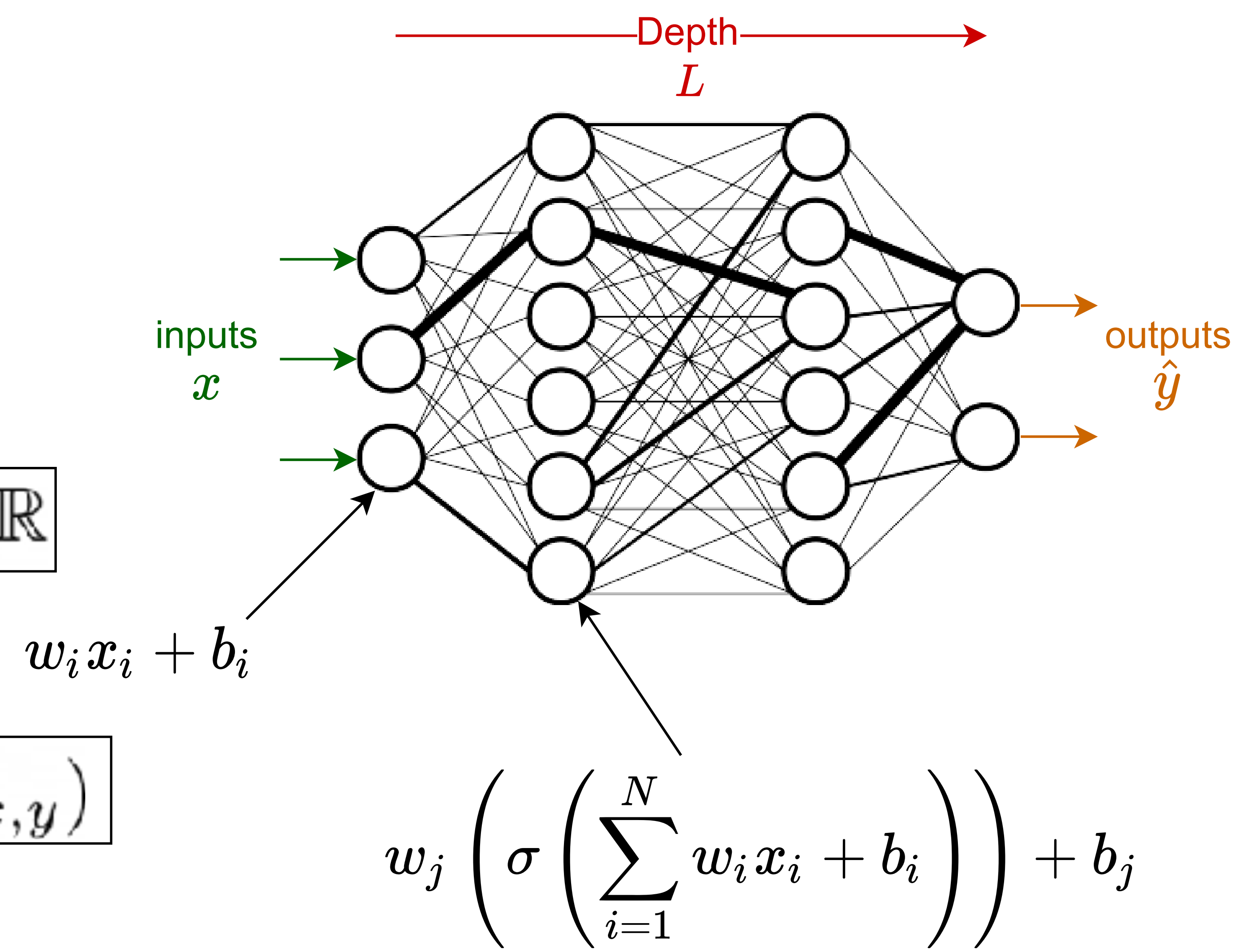
IV) Conclusion

Overview - Machine learning

- **Machine learning** framework
- **Neural networks**
- Artificial neuron : **affine transform**
- Stacked in successive **layers**
- Model with **parameters :** $\boxed{\theta \in \Theta}$
- **Loss function :** $\boxed{\mathcal{L} = \mathcal{L}_{x,y} : \Theta \to \mathbb{R}}$
- **Objective :** $\boxed{\hat{\theta} = argmin_{\theta \in \Theta}(\mathcal{L}_{x,y})}$
- **Output :** $\boxed{\hat{y} = \hat{\Gamma}_\theta(x)}$

- **Approximate** an **unknown function**

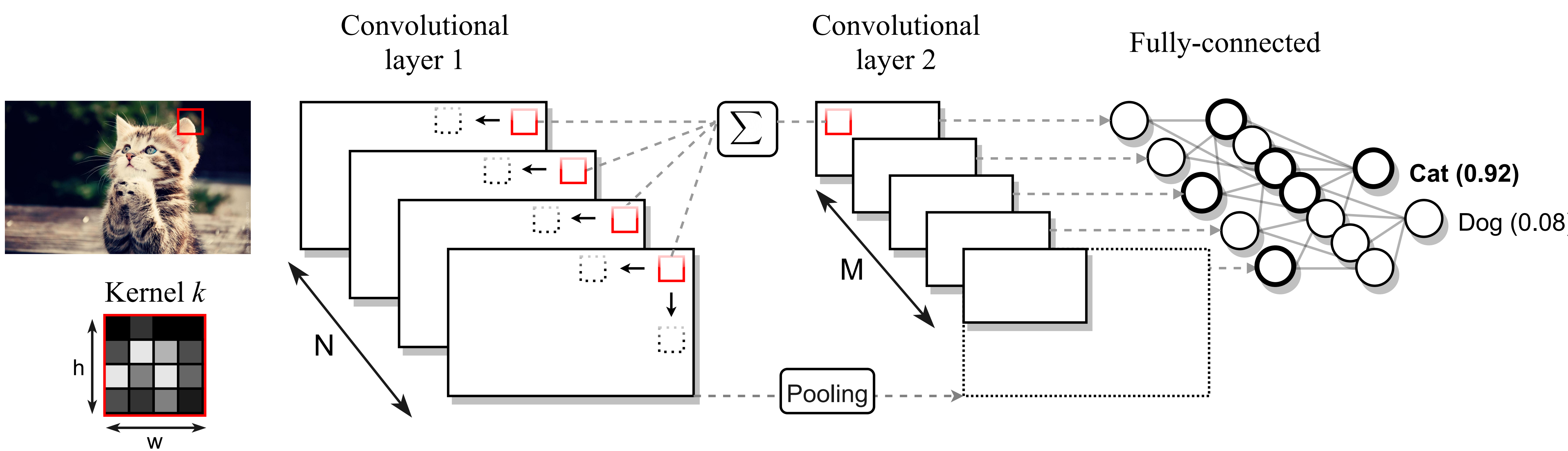
$$\boxed{y = \Gamma(x)}$$



Fully-connected network

Overview - Convolutional Neural Network

- Very efficient in **visual features recognition**



- Neurons : small **kernels**
- Producing **features maps**
- **Convolved** across the input image

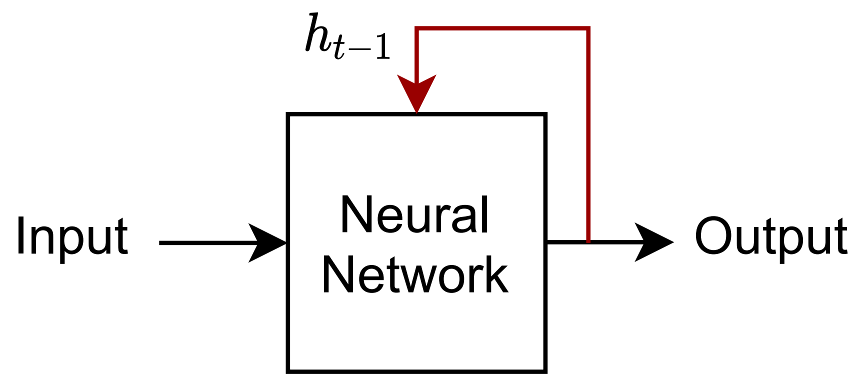
$$h_{ij}^{k} = (W^k * x)_{ij} + b_k$$

- **Pooling** operation to maintain a reasonable **dimensionality**
- Followed by a fully-connected network for classification task

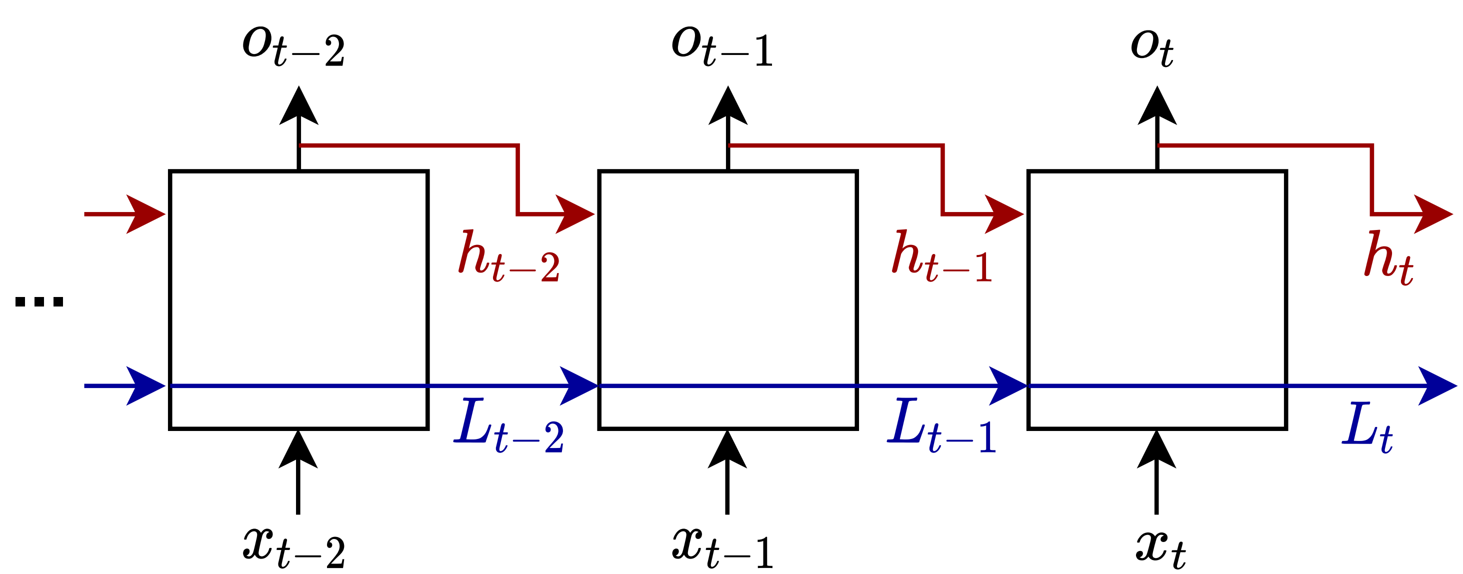
LeCun, Yann, and Yoshua Bengio. "Convolutional networks for images, speech, and time series." The handbook of brain theory and neural networks 3361.10 (1995): 1995.

Overview - Recurrent Neural Network

- **Feed-forward** networks do not retain past information
- Recurrent Neural Network handle **temporal structures** thanks to a **loop**



"Unfolded"



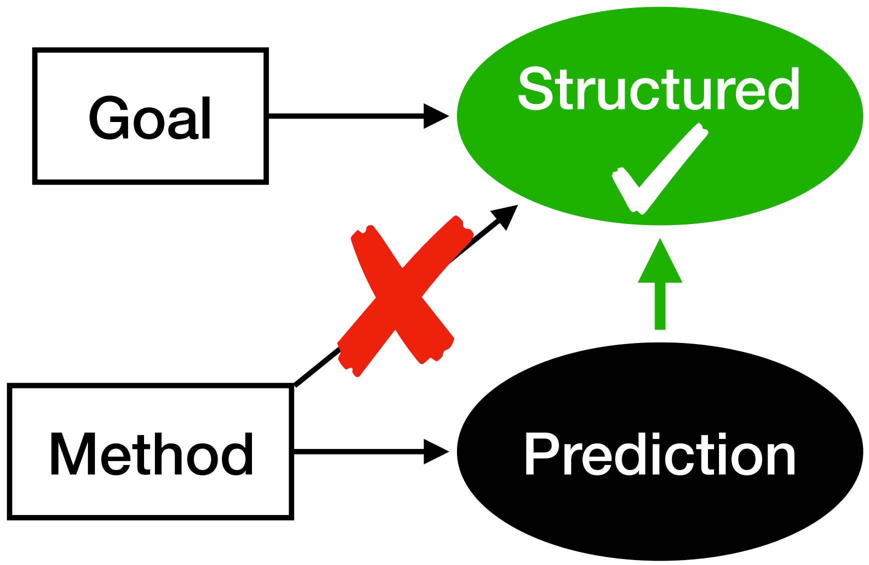
- Improvement for **longer-term** structures : **LSTM**

Elman, Jeffrey L. "Finding structure in time." Cognitive science 14.2 (1990): 179-211.

Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.

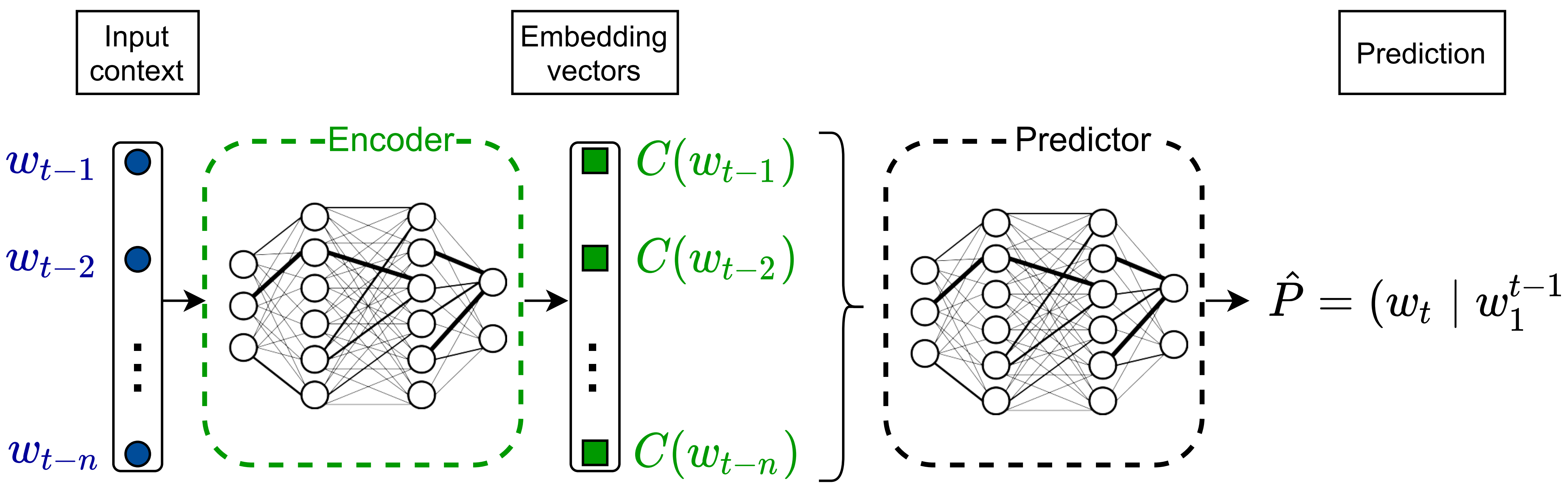
Overview - NLP method

- **No direct way** to learn **to structure** the output space as desired



*To correctly **predict** word occurrences, it is necessary to **capture** the global concepts of the language*

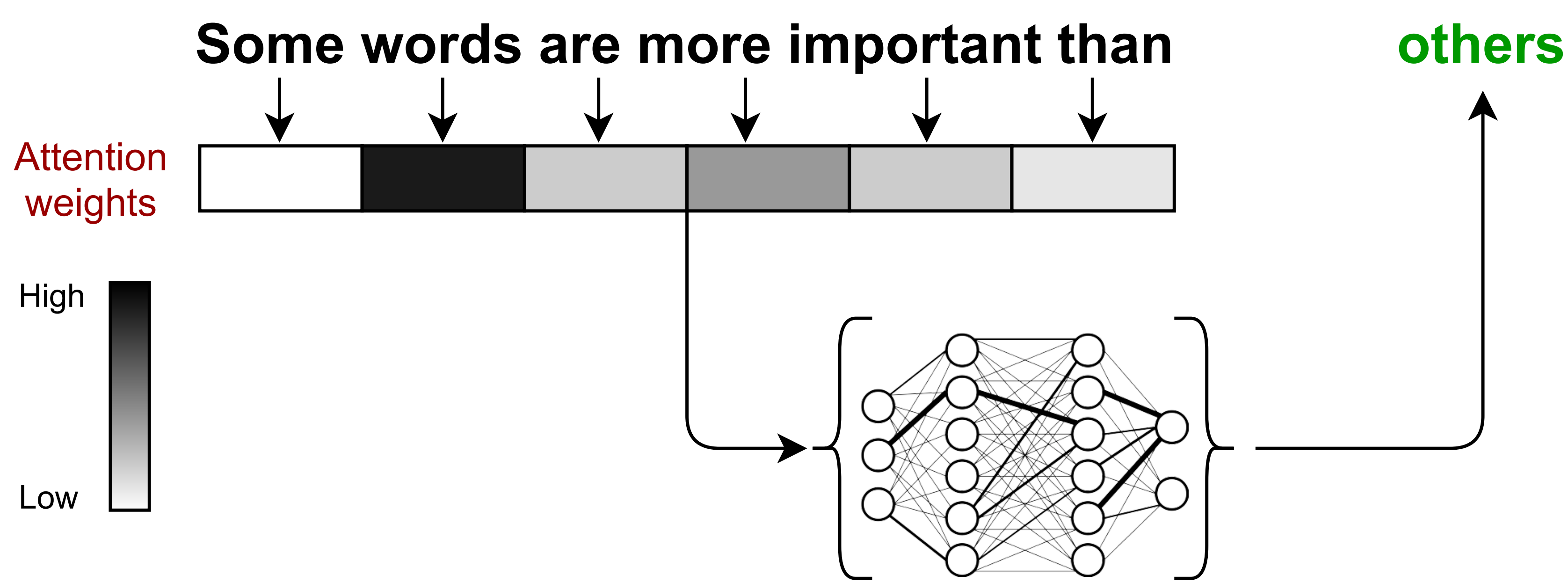
- Necessity to rely on a proxy task : **the prediction**



Bengio, Yoshua, et al. "A neural probabilistic language model." The journal of machine learning research 3 (2003): 1137-1155.

Overview - Attention mechanism

- **Different significances** between words in a sentence

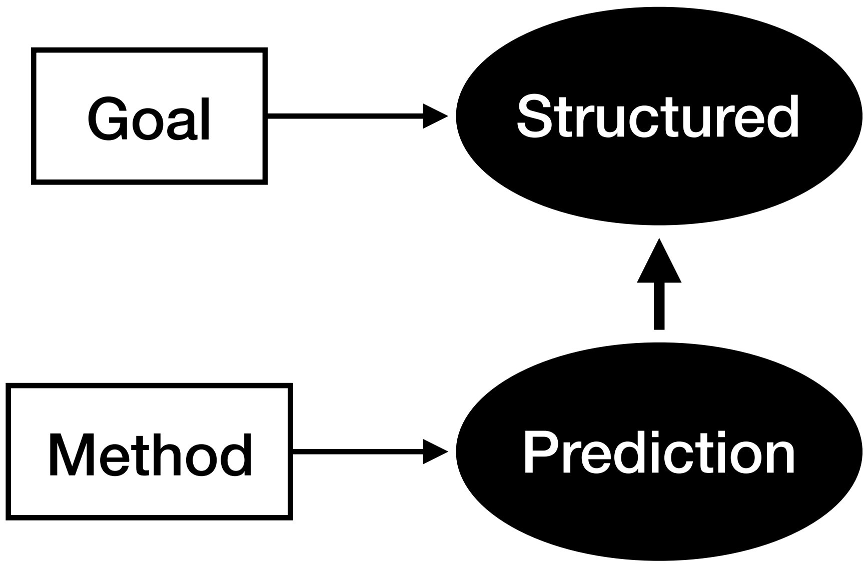


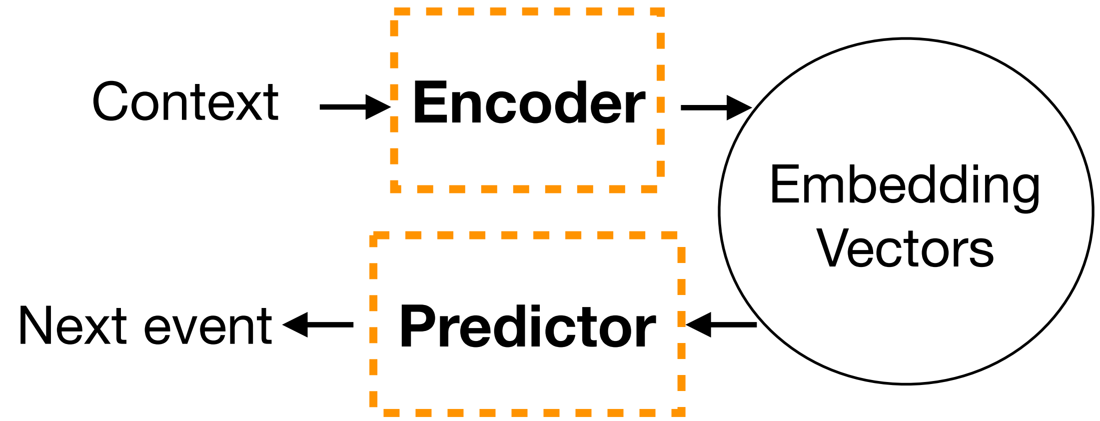
- **Weight** an input sequence according to the **relevance** of each step.
- **Jointly optimized** with the other network parameters
- **Strong improvement** in the overall performances of the system

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

Proposal - Approach description

- **Adapt** the word embedding approach **to musical data**





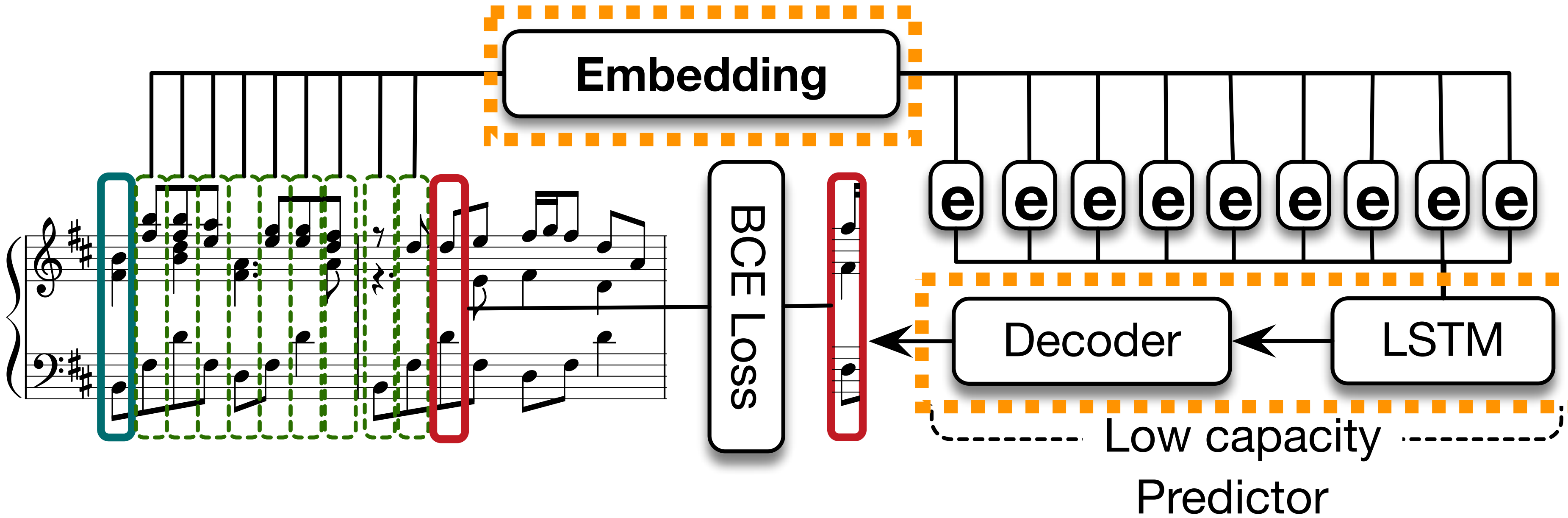
- Architecture of the **encoder** and the **predictor** designed to target **specificities of musical data**

Intervals, transpositions, octaves → ***Spatial features***

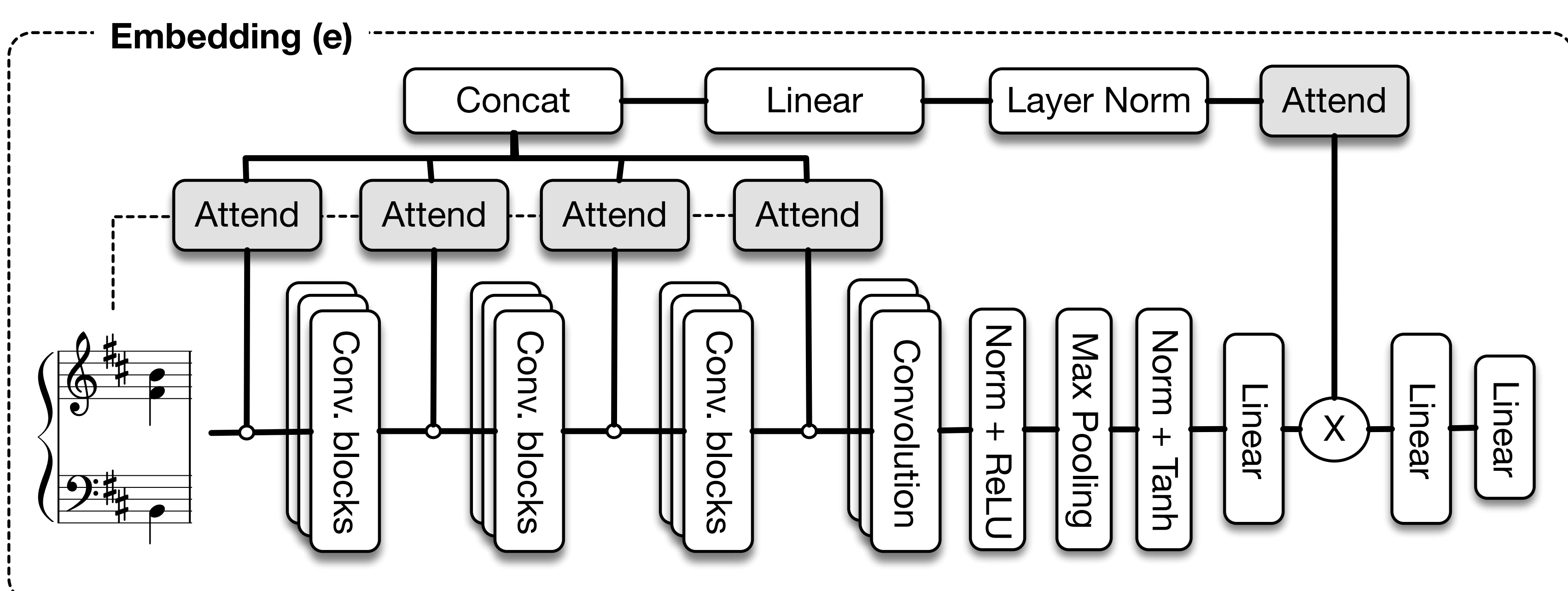
Rhythm → *Complexity of the* ***temporal relationships***

Proposal - Model architecture

- Trained to **predict** the current event in the **embedding space**

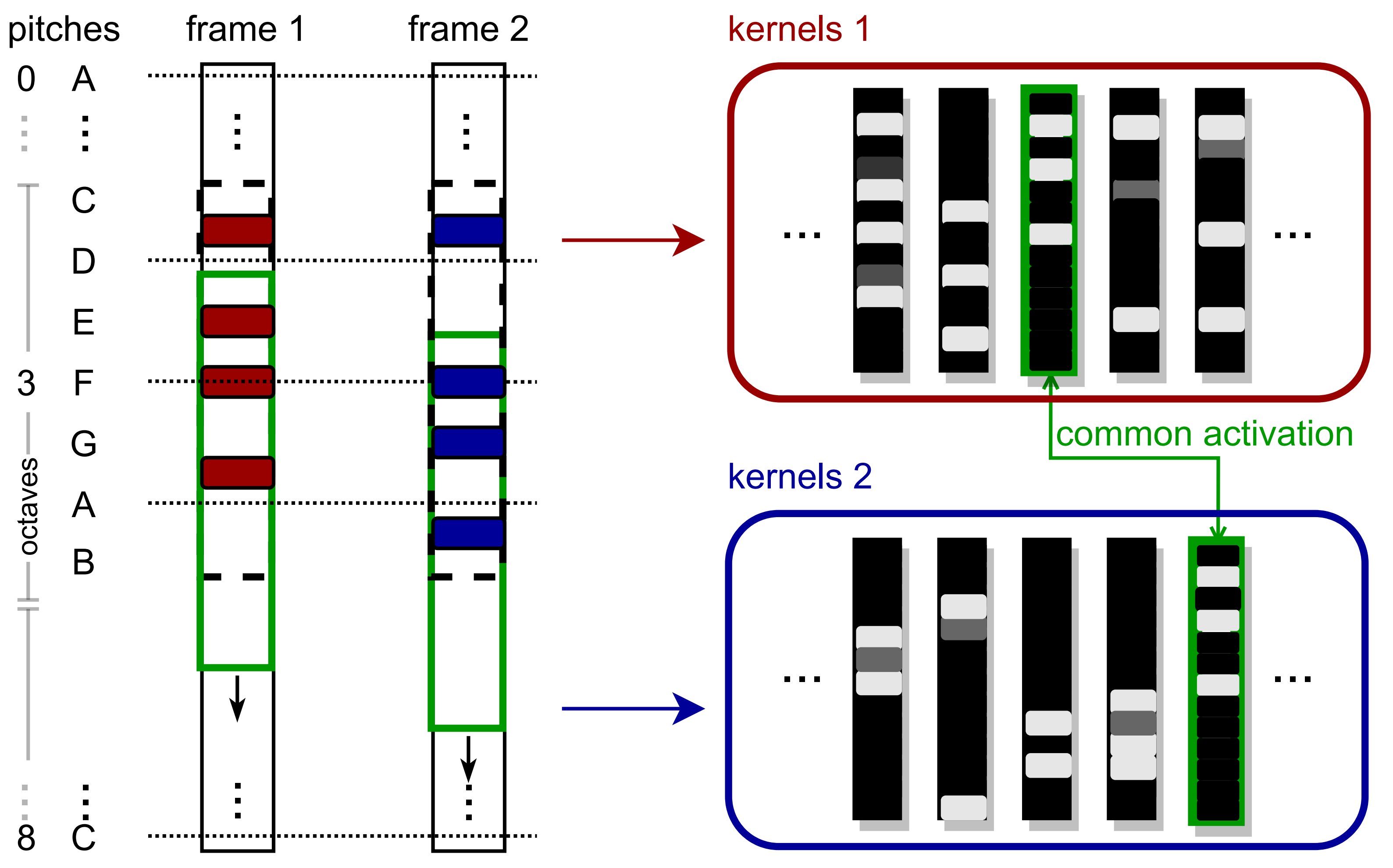


- **CNN-based encoder**
- With an **Attention mechanism**



Proposal - CNN for Piano-roll

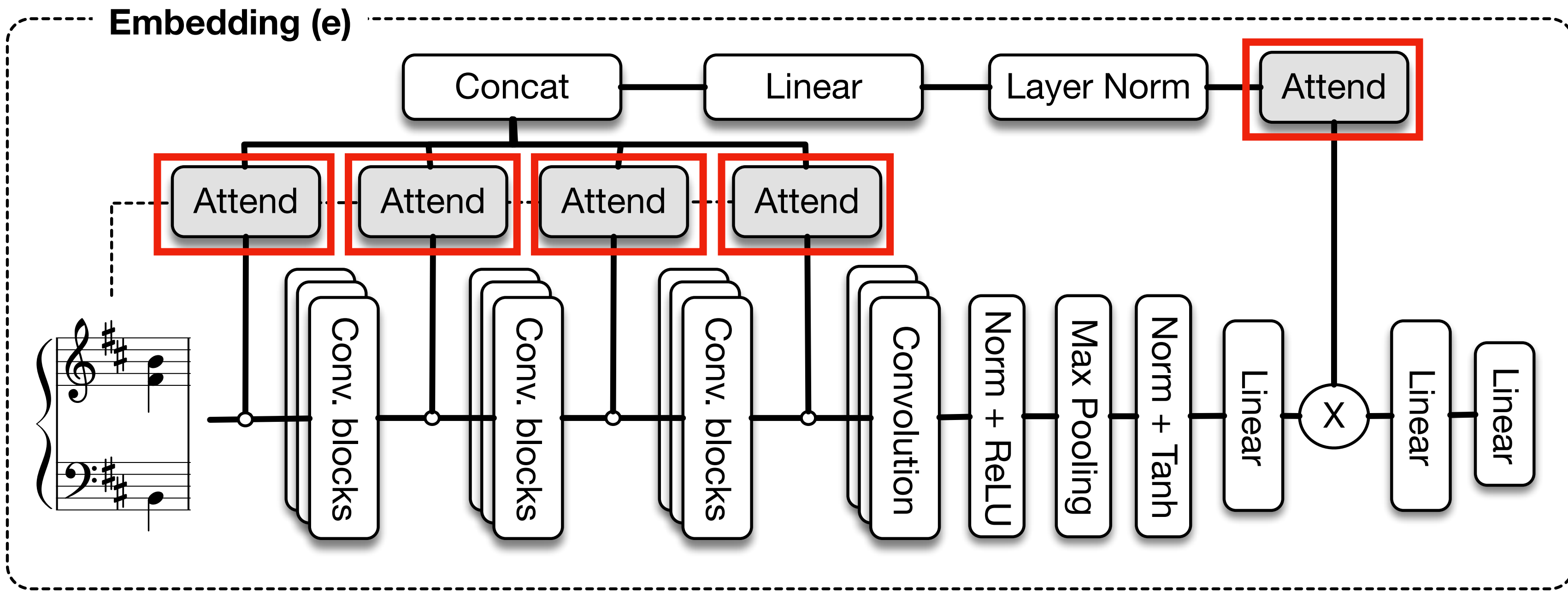
- Applied to musical data - Piano-roll frames



- Core properties of chord reflected in the kernels

Proposal - Hierarchical Attention Mechanism

- Differen



- Our proposal : **Hierarchical Attention Mechanism (HAM)**

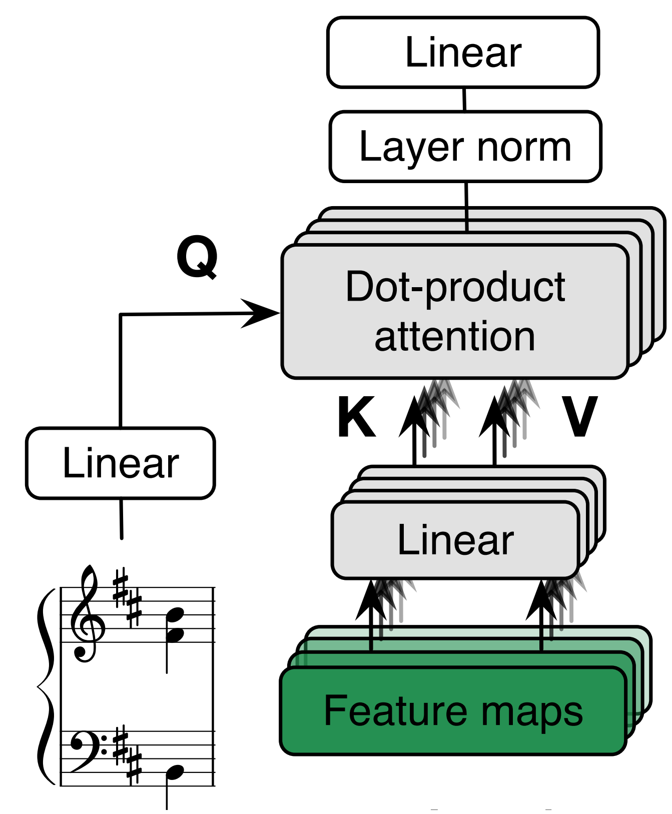
*On each **features maps** separately*

→ *Weight the kernels*

*At all **layers***

→ *Weight the different levels of abstraction*

***Modulate** the fully-connected network*



Proposal - Prediction results

- 4 datasets of **different complexity**

JSB Chorales *: Four-parts chorales by Bach*

Nottingham : *British and American folk tunes*

Piano-midi.de : *Classical music played on piano*

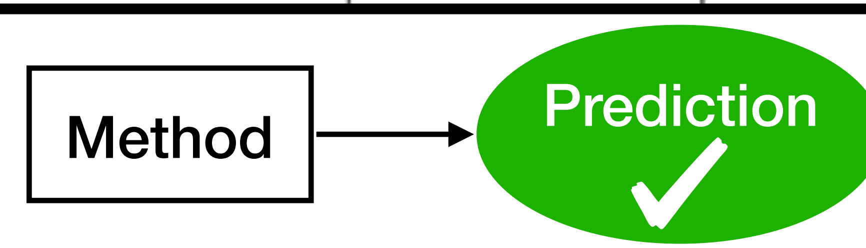
MuseData : *Orchestral pieces of classical music*

- **Frame-level accuracy** measure :

$$Acc = \frac{1}{M+N}\left(\sum_{n=1}^{N}\frac{TP_n}{TP_n + FP_n + FN_n} + \sum_{m=1}^{M}\frac{1}{1+FP_m}\right)$$

- Best model of the literature
- Classic and more sophisticated CNN
- Attention mechanisms
- **Best performances for the complete HAM model**

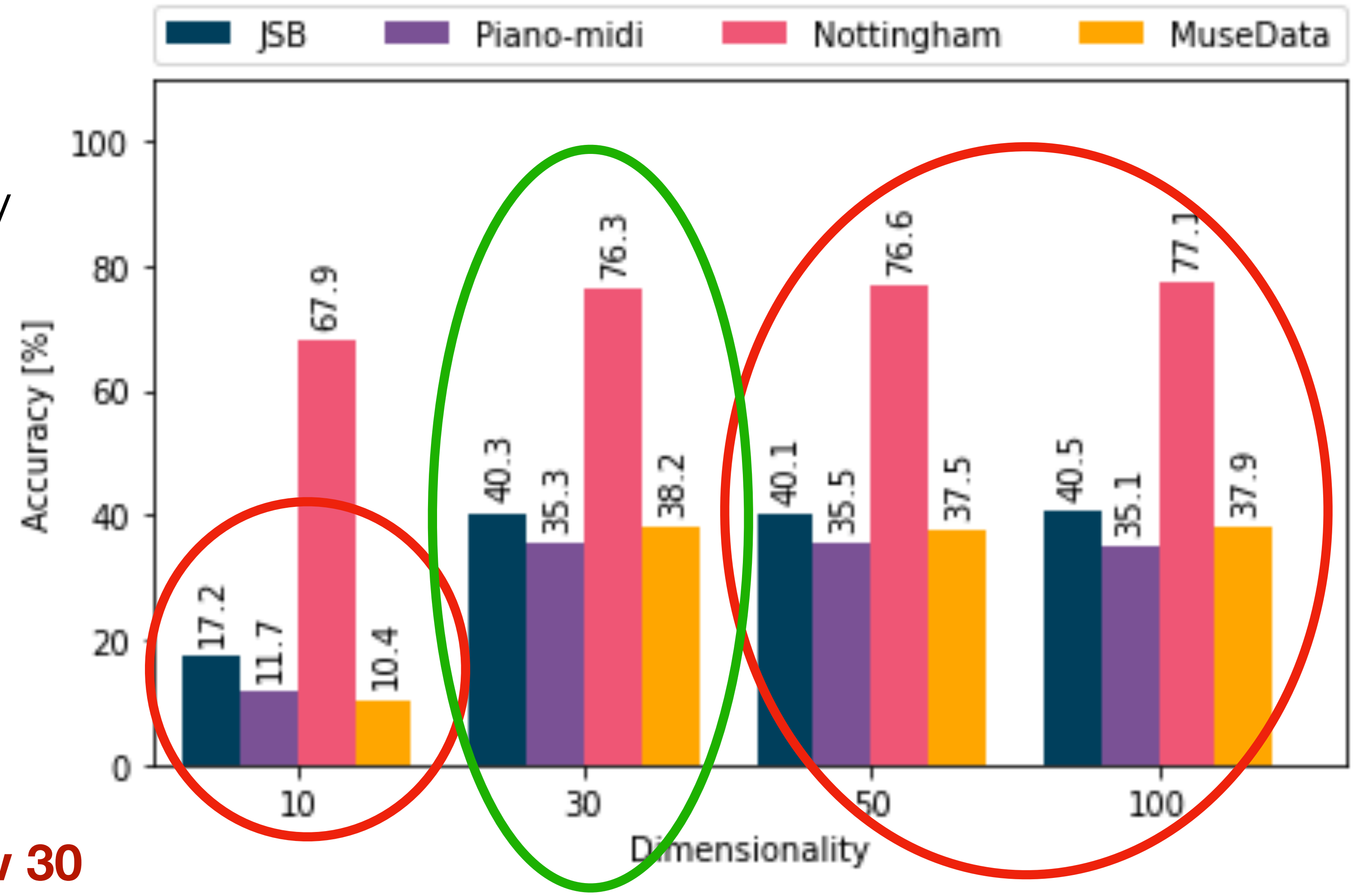
Models	JSB Chorales Acc. (%)	Piano-midi.de Acc. (%)	Nottingham Acc. (%)	MuseData Acc. (%)
RNN-RBM	33.12	28.92	75.40	34.02
RNN-Nade	32.11	20.69	64.95	24.91
Random	4.42	3.35	4.53	3.74
CNN	25.73	22.48	62.31	26.73
Residual	14.85	12.29	53.42	12.30
Dense	15.36	12.74	56.42	16.44
AM-dp	33.61	30.17	64.11	27.17
AM-mh	35.19	32.68	64.25	32.15
HA+	39.07	33.27	76.09	37.84
HAM	**40.25**	**35.28**	**76.25**	**38.15**

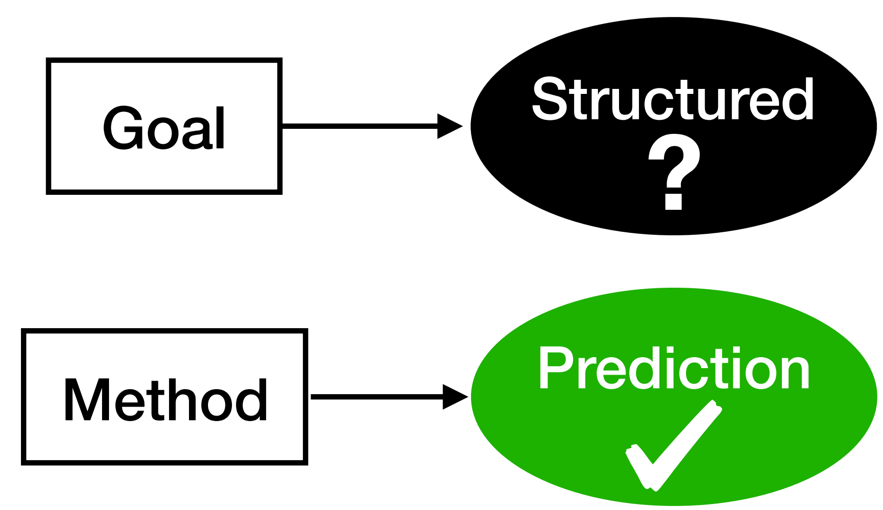


Boulanger-Lewandowski, Nicolas, Yoshua Bengio, and Pascal Vincent. "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription." arXiv preprint arXiv:1206.6392 (2012).

Proposal - Dimensionality

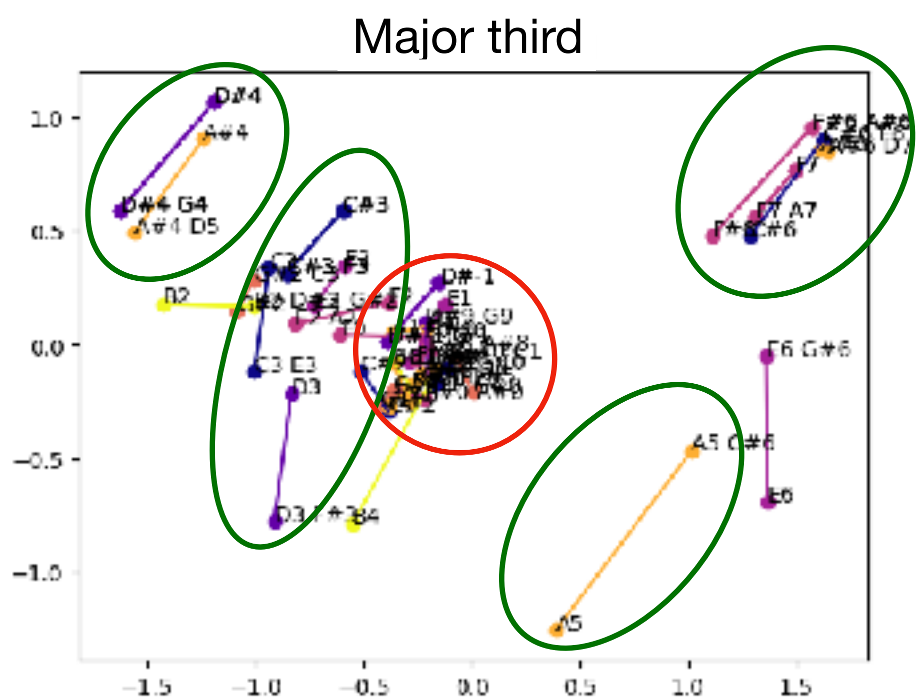
- Embedding **dimensionality**
- **Best trade-off** compression/ amount of information
- Impact on the prediction accuracy
- **10**, **30**, **50** and **100** dimensions
- **Drop in performance below 30**
- **No significant improvement above**
- **Best trade-off for 30 dimensions**

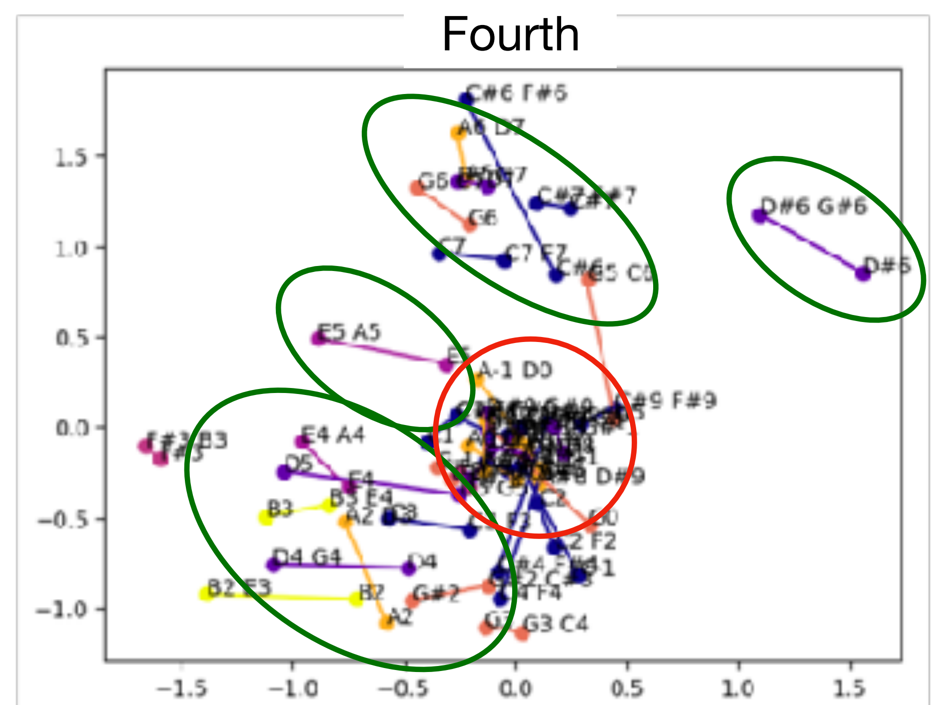




Proposal - Structural results

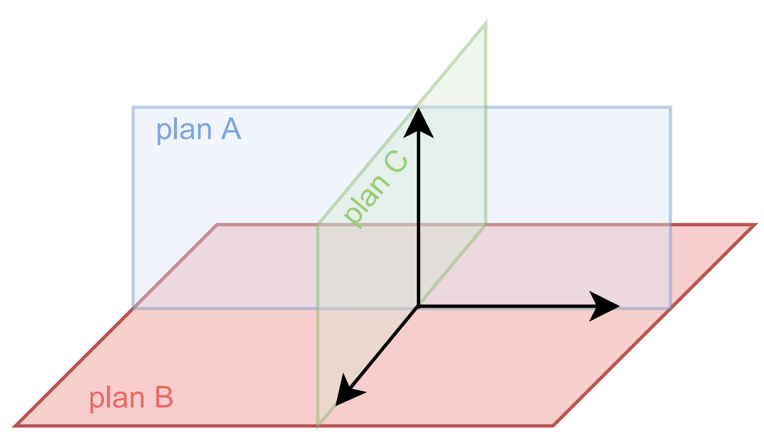
- 2D visualization of the embeddings through **PCA**
- Root notes linked to **musically-related** elements

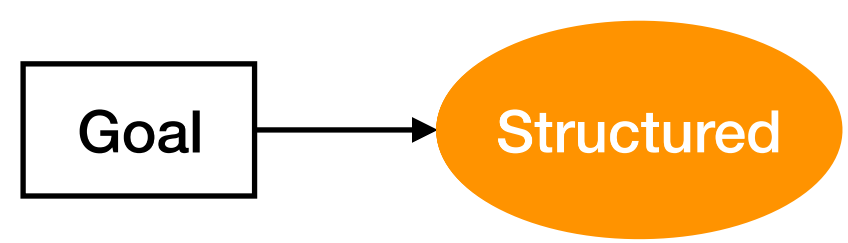




- **Share common geometric properties**
- **Orthogonalization of the embedded data**

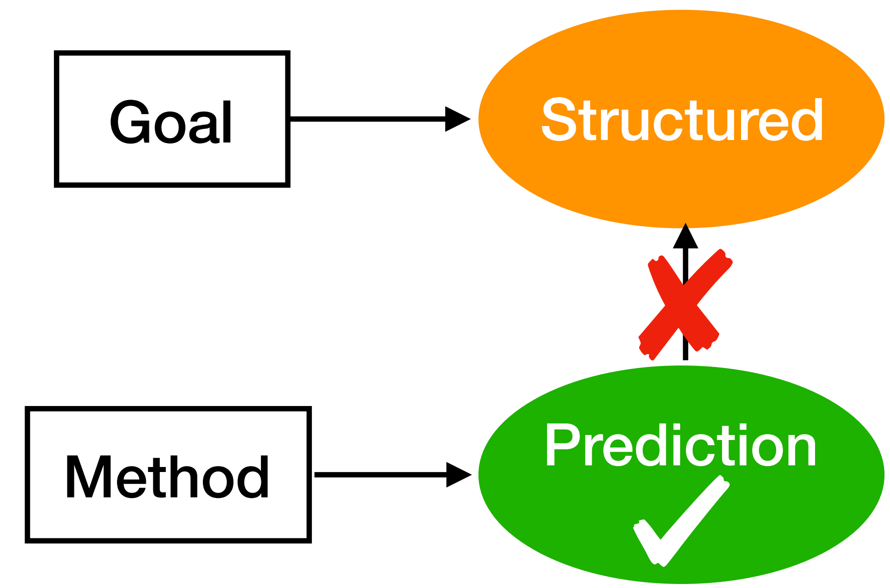
Items embedded in different sub planes





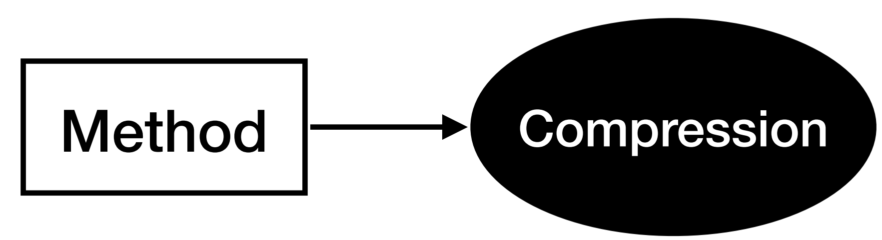
Proposal - Summary

- Explore a method **inspired by the NLP field**
- **Very good performance on the prediction task**
- **Structure weakened by the orthogonalization issue**
- **Assumption not confirmed in our context**



- **Revise the fundamentals** of this approach to develop a new and more effective method

Change the proxy task →



Increase the meaning carried by a single event → *Consider an entire bar as unit to be embedded*

I) First method - Adapt the NLP methods

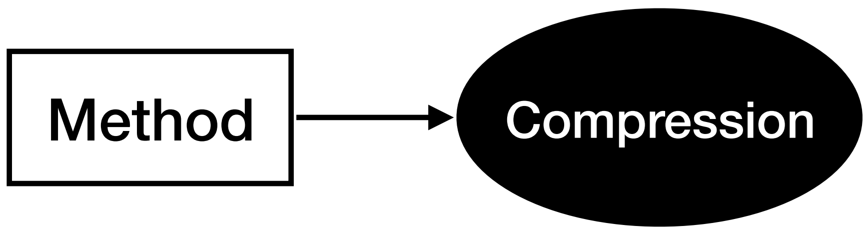
II) Second method - Variational Auto-Encoders

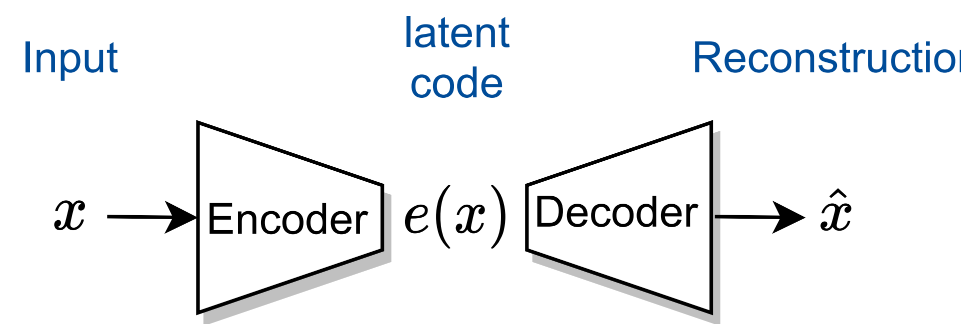
III) Applications

IV) Conclusion

Overview - Variational Auto-Encoder

- Auto-Encoder $\hat{\mathbf{x}} = d(e(\mathbf{x})) \approx \mathbf{x}$



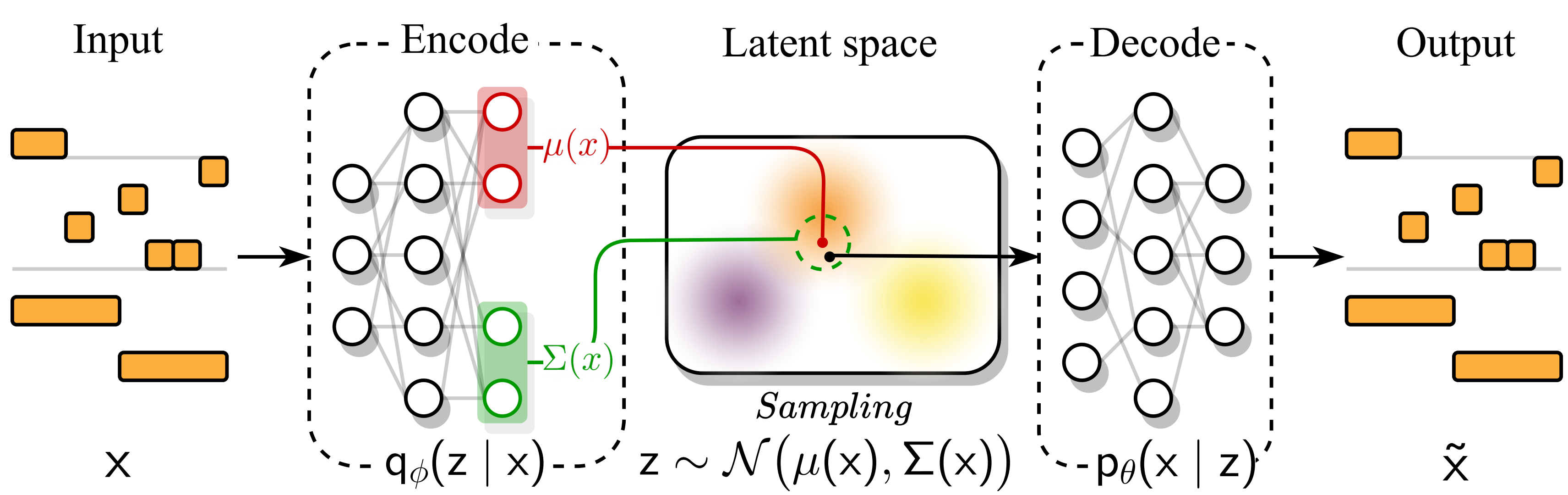


- **Two-term loss** allowing to control the properties of the latent space

*Force the latent space distribution to be **close to the normal distribution***

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right]}_{\text{reconstruction}} - \underbrace{\mathcal{D}_{KL}\left[q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z})\right]}_{\text{regularisation}}$$

- Resulting space **smoothly organized**
- **Continuity** allowing the **generation** of realistic data



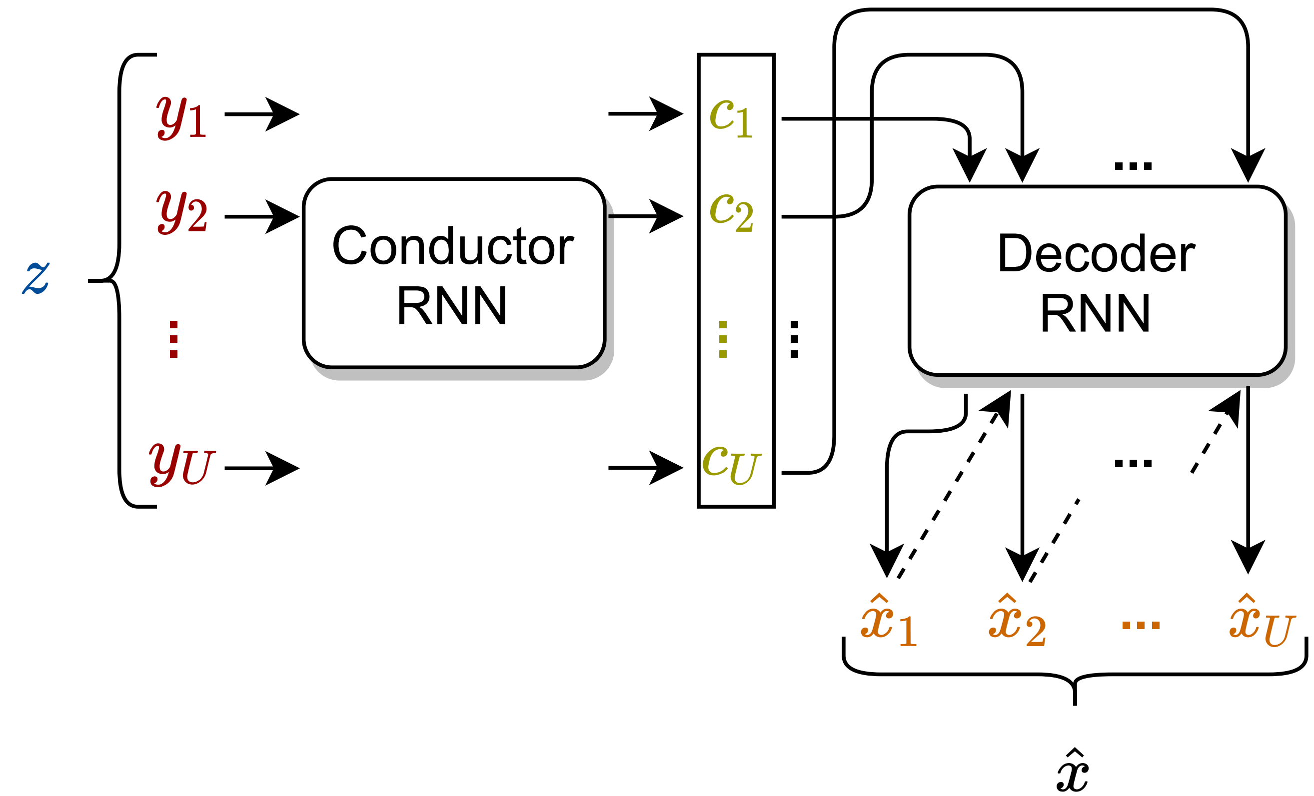
Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

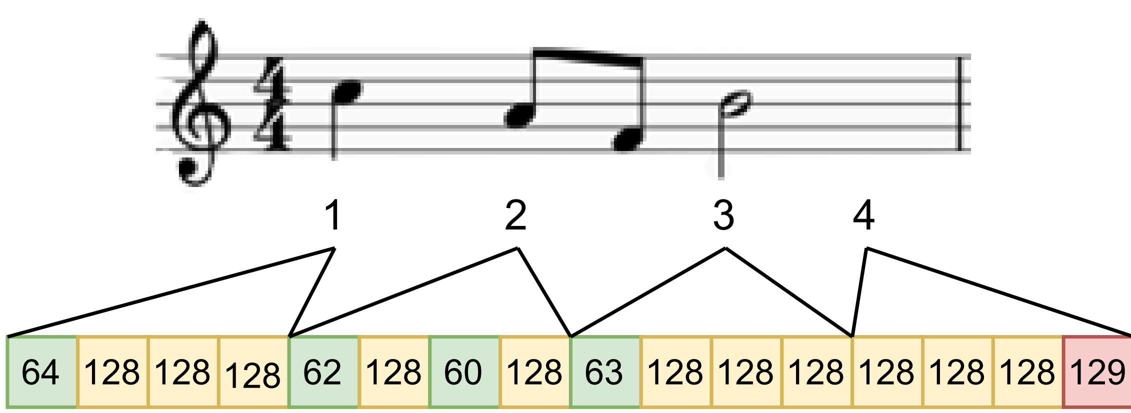
Overview - MusicVAE

- Success of MusicVAE to learn embedding for **monophonic melodies**

- **Two-part decoder** that force the system to rely on the latent space

- Divide the latent code into U non-overlapping subsequences





0-127 : Note ON
128 : Do nothing
129 : Note OFF

- **Efficient input representation**

  Divide the time steps in 16th note

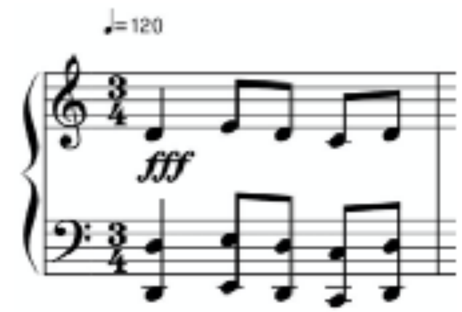
  Vocabulary with 130 different events

- **Only for monophonic data**

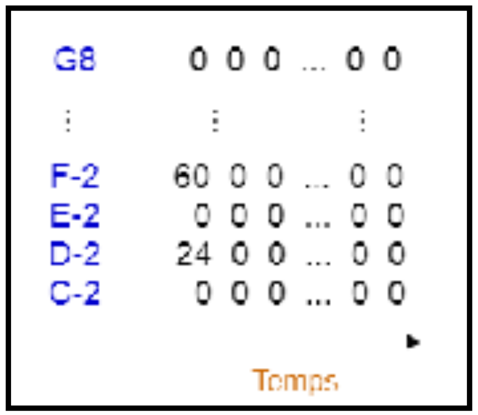
Roberts, Adam, et al. "A hierarchical latent vector model for learning long-term structure in music." International Conference on Machine Learning. PMLR, 2018.

Overview - Symbolic polyphonic representation

- Different existing representation for **polyphonic music**



Piano-roll



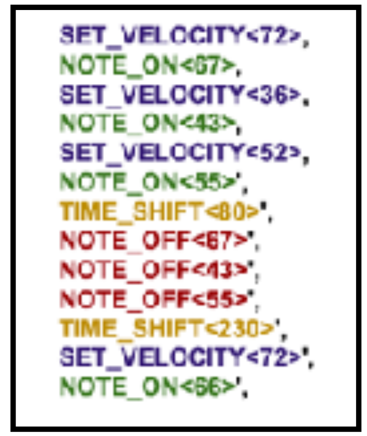
\+ ***Easy*** *to produce*

\+ ***Adaptability***

\- *Very* ***sparse***

\- *A lot of repeated frames,* ***redundancy***

MIDI-like



\+ ***Compact***

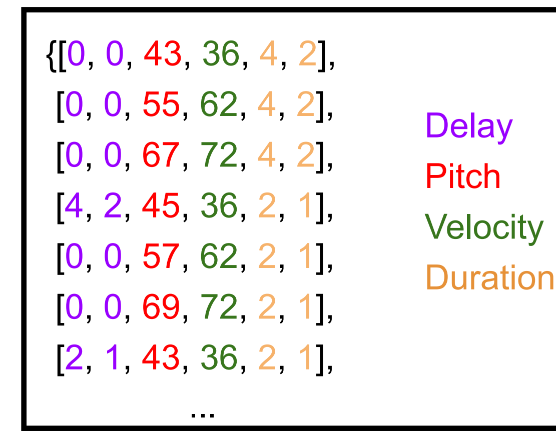
\+ ***Adaptability***

\- *Very* ***large vocabulary***

\- ***Unordered*** *note attributes*

\- *Very* ***sensitive*** *to errors*

NoteTuple



\+ ***Adaptability***

\- ***Many*** *small vocabularies*

- **Find a more effective input representation**

Oore, Sageev, et al. "This time with feeling: Learning expressive musical performance." Neural Computing and Applications 32.4 (2020): 955-967.

Hawthorne, Curtis, et al. "Transformer-NADE for Piano Performances." submission, NIPS Second Workshop on Machine Learning for Creativity and Design. 2018.

Proposal - The signal-like representation

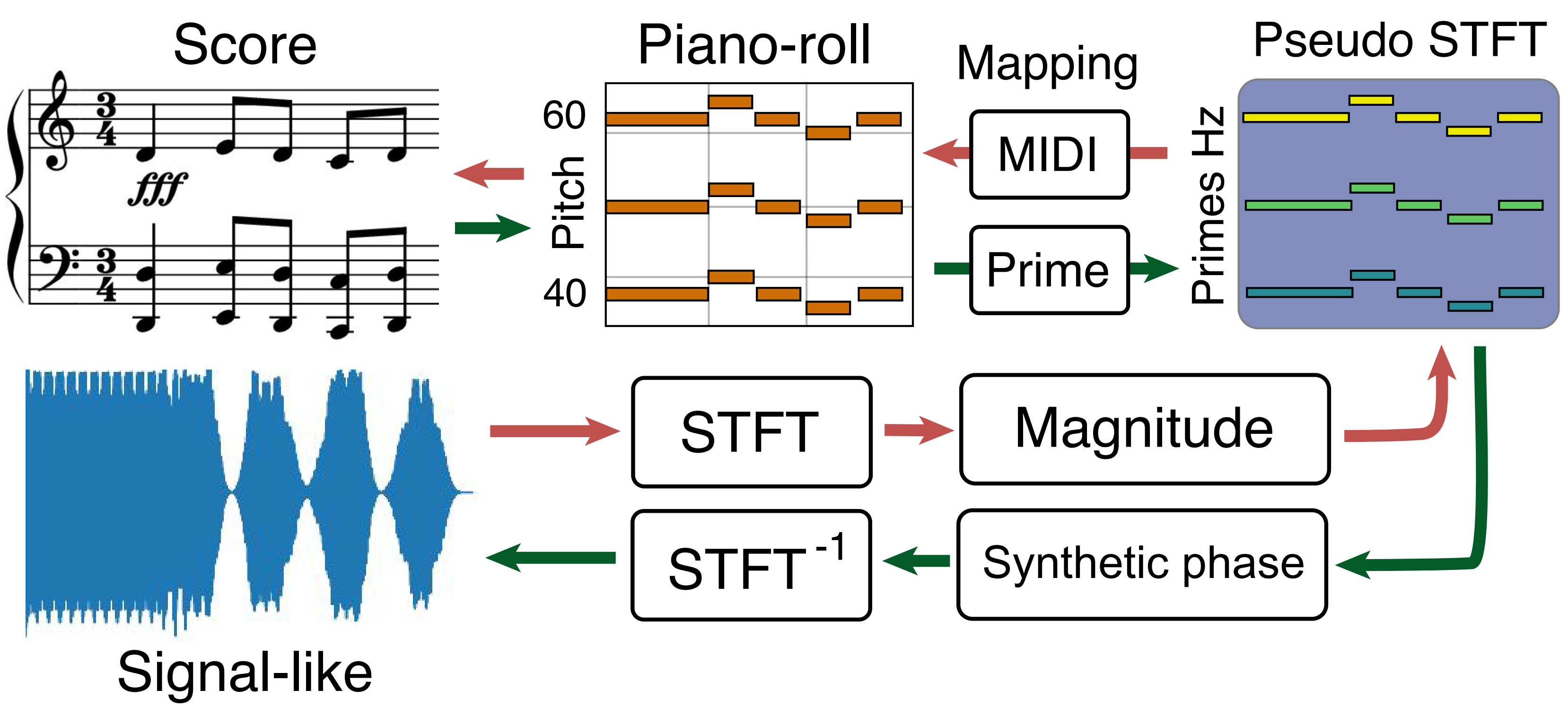
- A new representation for polyphonic music : the **Signal-like**
- In similar fashion to an audio signal
- **Sum of periodic function** oscillating at **different frequencies**

Naturally contains polyphonic information ***Large dimensionality***

Invertibility ***Phase effects in harmonic signal***

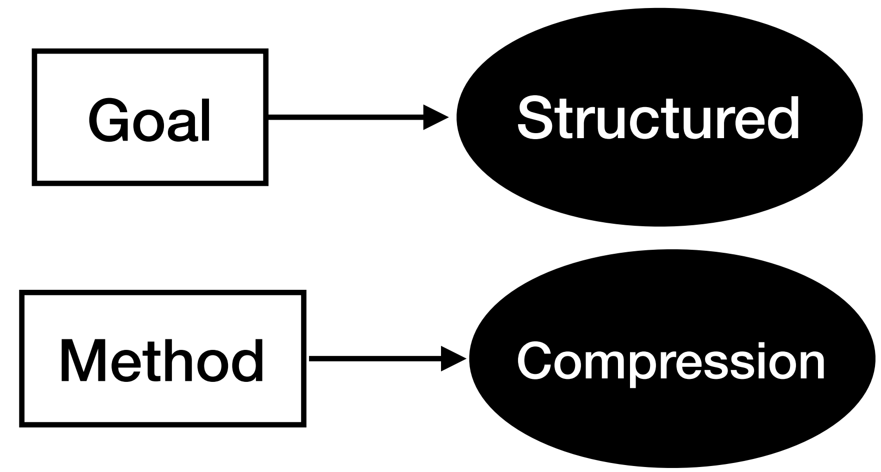
- Construction process

    1. *Compute piano-roll*
    2. *Map pitches to prime numbers*
    3. *Add a complex part to the matrix (artificial phase)*
    4. *Compute the STFT^{-1}*



Proposal - Representation benchmark

- **Benchmark** for learning embedding spaces
  - Testing the **compression results**
  - Testing the **structure** of the resulting embeddings



- Implementation of an architecture similar to MusicVAE
  - Training through the **four representations**
  - On the **JSB Chorales dataset**
  - Very **strict musical rules**
  - Facilitates the **evaluation** of the structure from a **musical point of view**

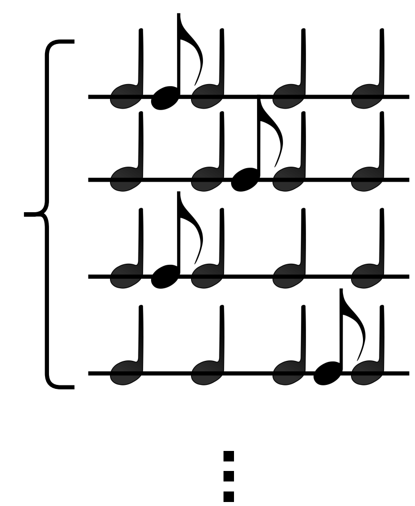
Proposal - Synthetic dataset

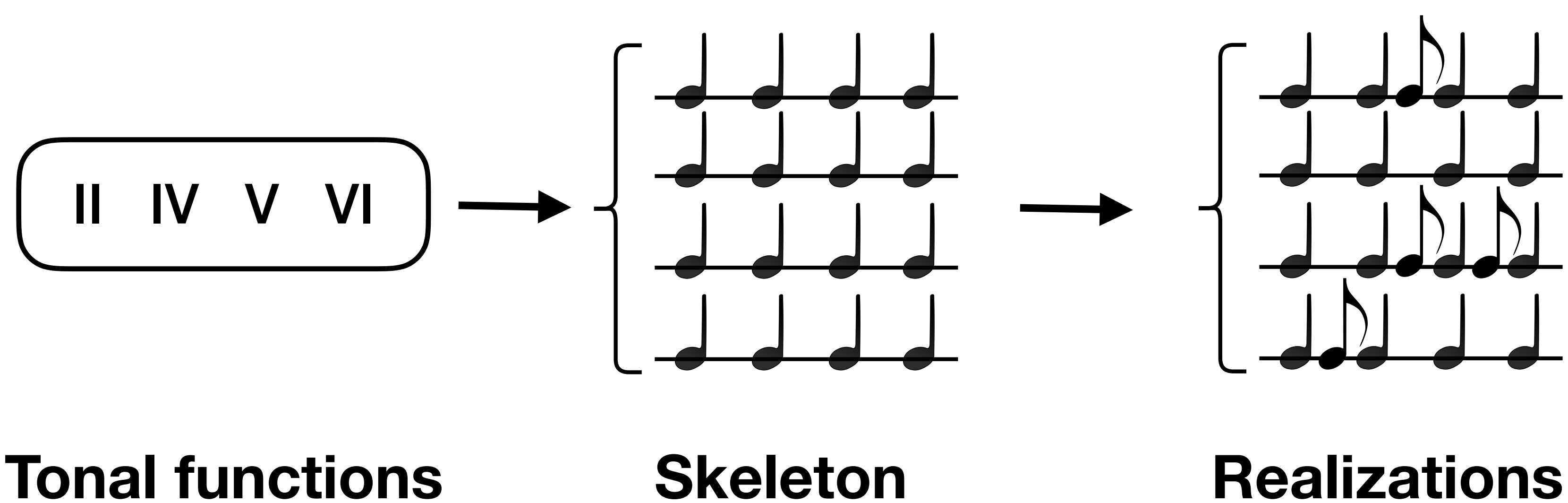
- Synthetic dataset to analyze **the structures** of the embeddings in a **music theory stand point**
- Music theory rules of the Bach chorales

1. Generate sequences of tonal functions
2. Expand to four voices : **skeleton**

   Major triads, minor triads, diminished triads and dominant sevenths
3. Adding non-harmonic tones : **realizations**

   Passing tones, neighboring tones, suspensions and retardations





Proposal - Learning results

- **Reconstruction** and **KL divergence** results

- Monophonic results as reference

- MIDI-like : ill-defined musical sequences

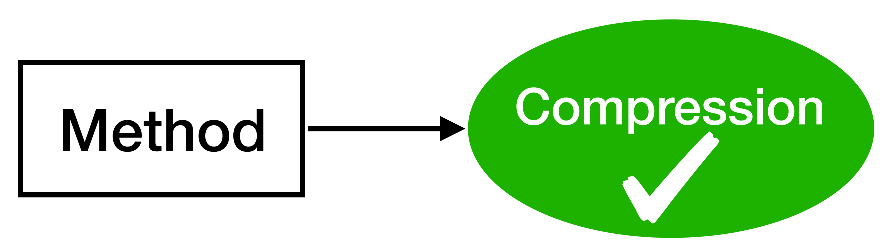
- NoteTuple : **low** reconstruction accuracy, **high** KL div

	Input	Reconstruction accuracy (%)	KL div
Monophonic	Piano-roll	95.8	$2 * 10^3$
	MIDI-like-mono	97.5	$1 * 10^3$
Polyphonic	Piano-roll	94.1	$2 * 10^3$
	MIDI-like	< 1	-
	NoteTuple	17.3	$9 * 10^4$
	Signal-like	**96.5**	$\mathbf{1 * 10^3}$

- Signal-like : **High** reconstruction accuracy, **low** KL div

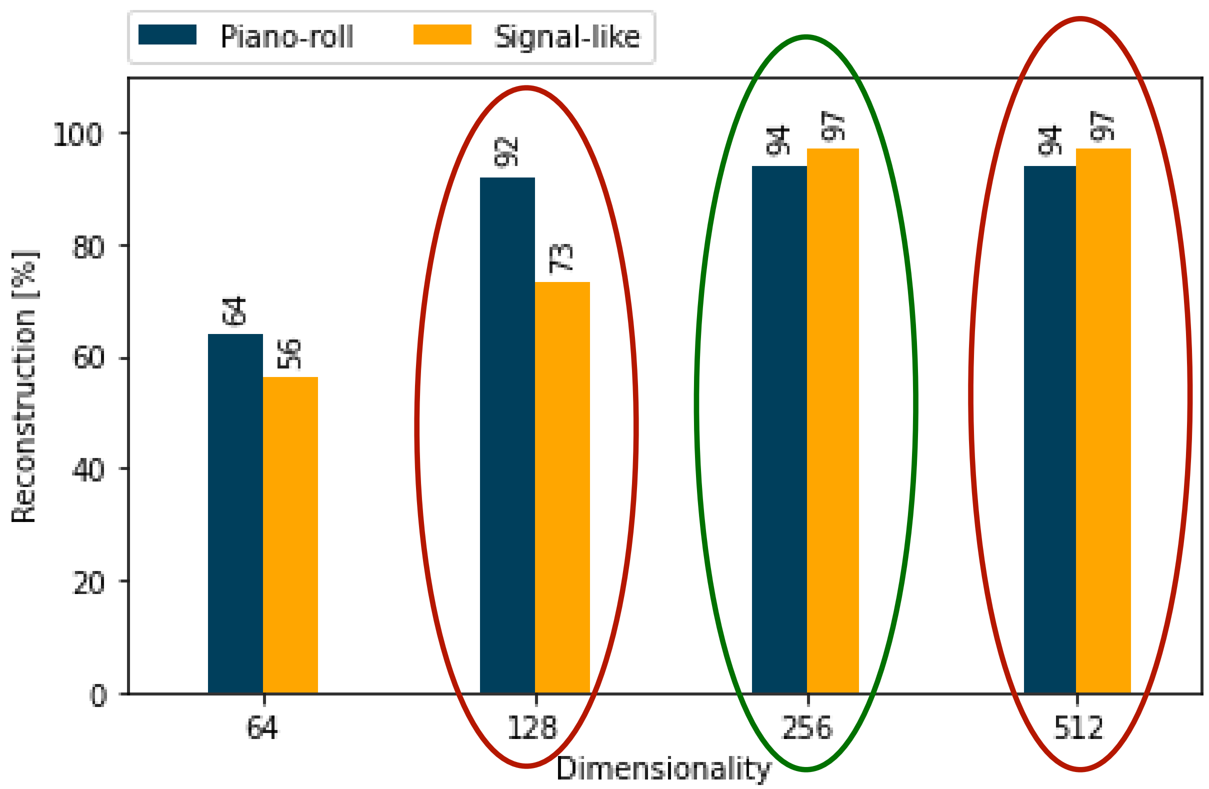
  Improve the learning stability

  Less overfitting

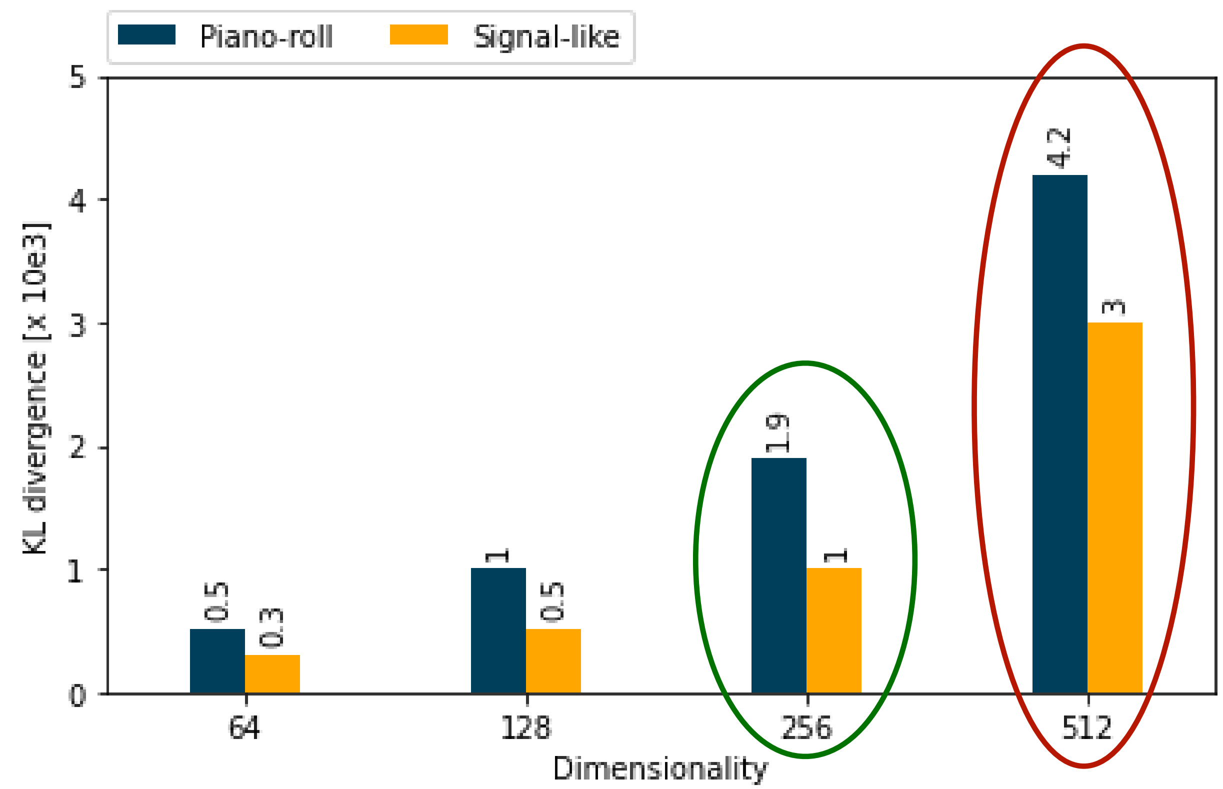


Proposal - Dimensionality

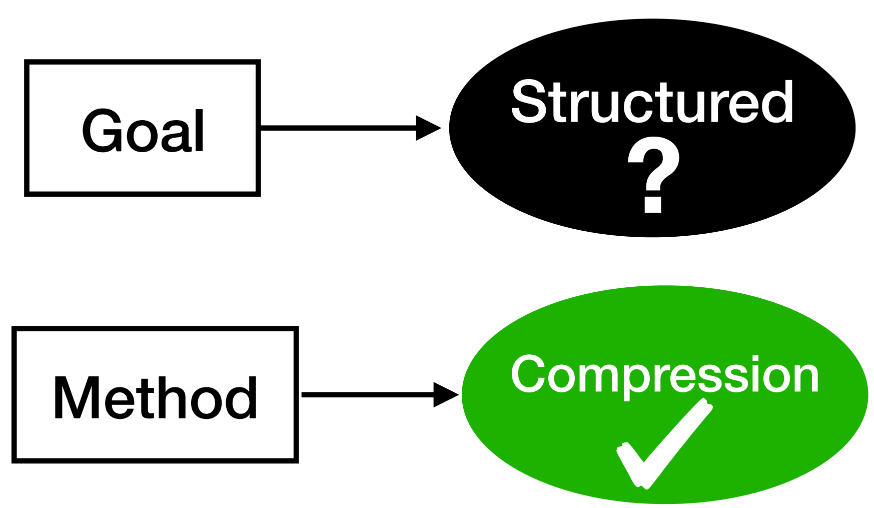
- Impact of the **dimensionality**



- **64**, **128**, **256** and **512** dimensions



- Insufficient reconstruction below 256
- No improvement above
- High KL divergence above 256
- **Best trade-off for 256**

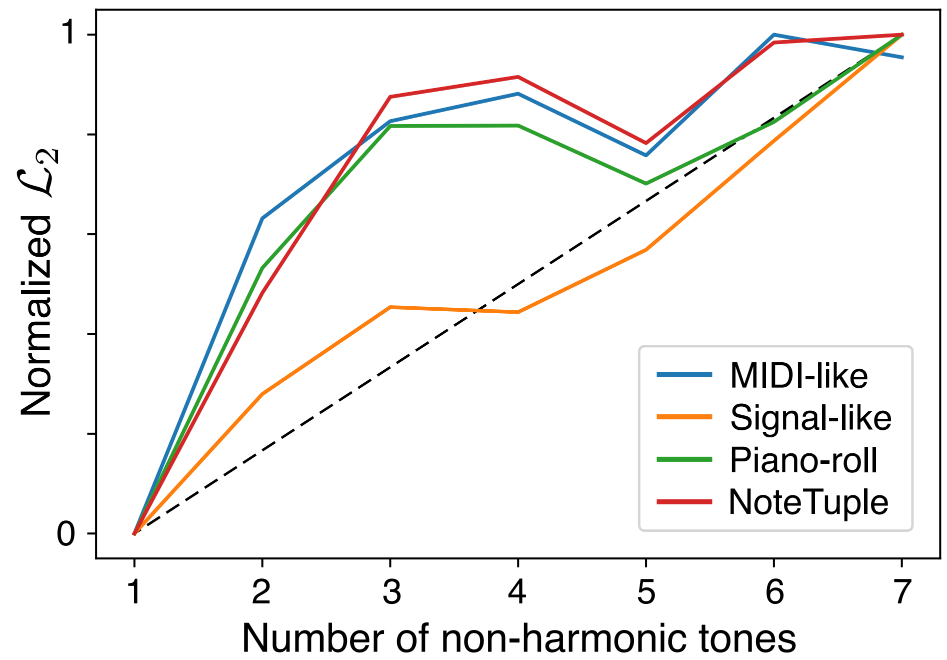


Music theory analysis

- Distances between a skeleton and its realizations according to the **number of non-harmonic tones**

Not even monotonic in the bad cases

Almost linear for the Signal-like



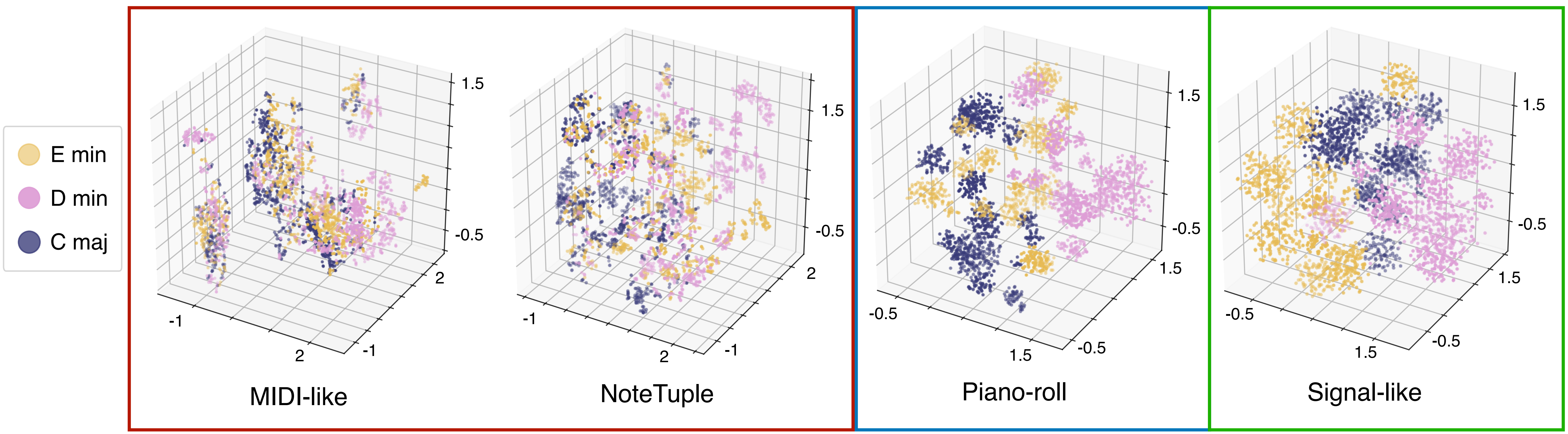
- Distances between **consecutive skeletons (DBCS)** and distances between a **skeleton and its realizations (DBSR)**

A realization will always be closer "musically speaking" to its skeleton than another skeleton

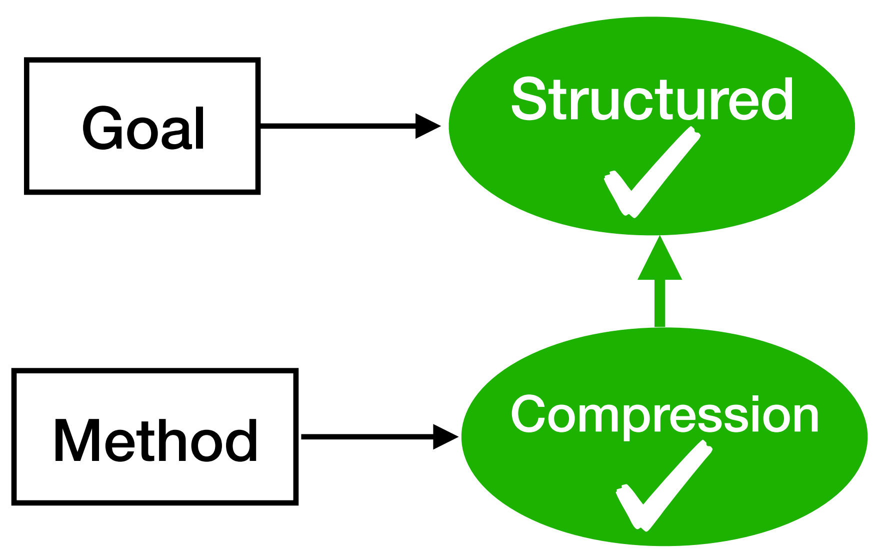
Input	DBSR	DBCS
Piano-roll	229.0 ± 36.0	291.7 ± 20.4
MIDI-like	445.5 ± 169.1	312.8 ± 92.6
NoteTuple	572.5 ± 146.5	292.8 ± 89.4
Signal-like	242.2 ± 34.2	285.5 ± 13.5

Music theory analysis

- Visualization of the bars in the spaces according to their **tonalities**



- Unstructured spaces, no tonality separation
- Structured space, good separation, lack of smoothness between clusters
- Structured space, very good separation, smooth transitions, good continuity



I) First method - Adapt the NLP methods

II) Second method - Variational Auto-Encoders

III) Applications

IV) Conclusion

Composers classification

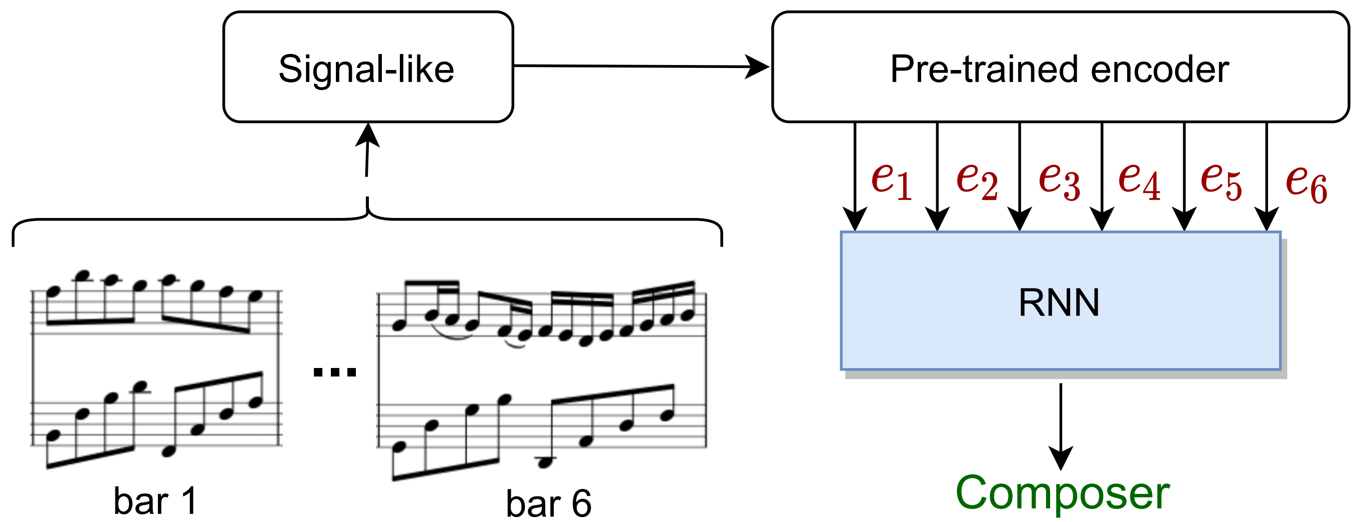
- Train our system on the **MAESTRO** dataset

  Classical music from the 17th to the early 20th century

- **Freeze** the encoder parameters and use the **embedding vectors** as input representation

  Composers classification task

  Train a simple RNN to classify small excerpt of music



- **Results**

  All composers simultaneously

Composers	Train	Test	Accuracy
Bach	3088	553	**84%**
Beethoven	6055	797	**54%**
Schubert	7428	1017	**46%**
Chopin	6027	1367	**46%**
Liszt	5082	485	**77%**
Total	27680	4219	**58%**

  One composer among the others

Composers	Accuracy
Bach	**91%**
Beethoven	**86%**
Schubert	**69%**
Chopin	**61%**
Liszt	**89%**

Hawthorne, Curtis, et al. "Enabling factorized piano music modeling and generation with the MAESTRO dataset." arXiv preprint arXiv:1810.12247 (2018).

Attribute vector arithmetic

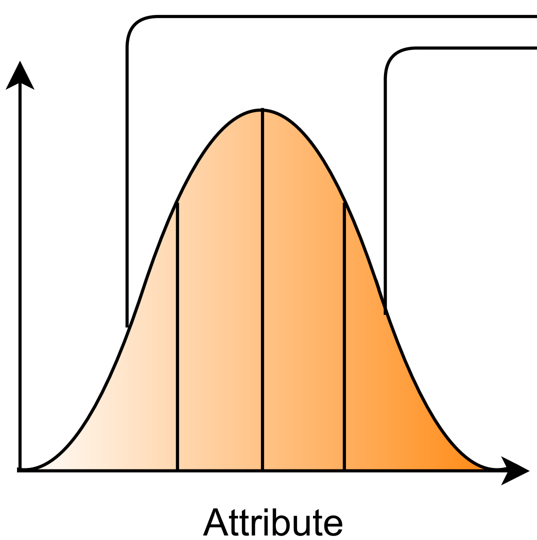
- Aim to **modify musical features** of a given bar intentionally
- Define the **attributes**

> ***C diatonic membership*** *: Amount of note which belong to the C diatonic scale*
>
> ***Note density*** *: Number of note*
>
> ***Average polyphony*** *: Mean number of note played simultaneously*
>
> ***Average note duration*** *: Mean of the notes duration*
>
> ***8th and 16th note syncopation*** *: Syncopated notes proportion*

- Compute the **attributes** for each training sample
- Compute the **attribute vectors**

Split the dataset into ***quartiles*** *according to the attributes*

Compute the ***subtraction*** *between the* ***top*** *and the* ***bottom*** *quartile* ***mean embedding vectors***



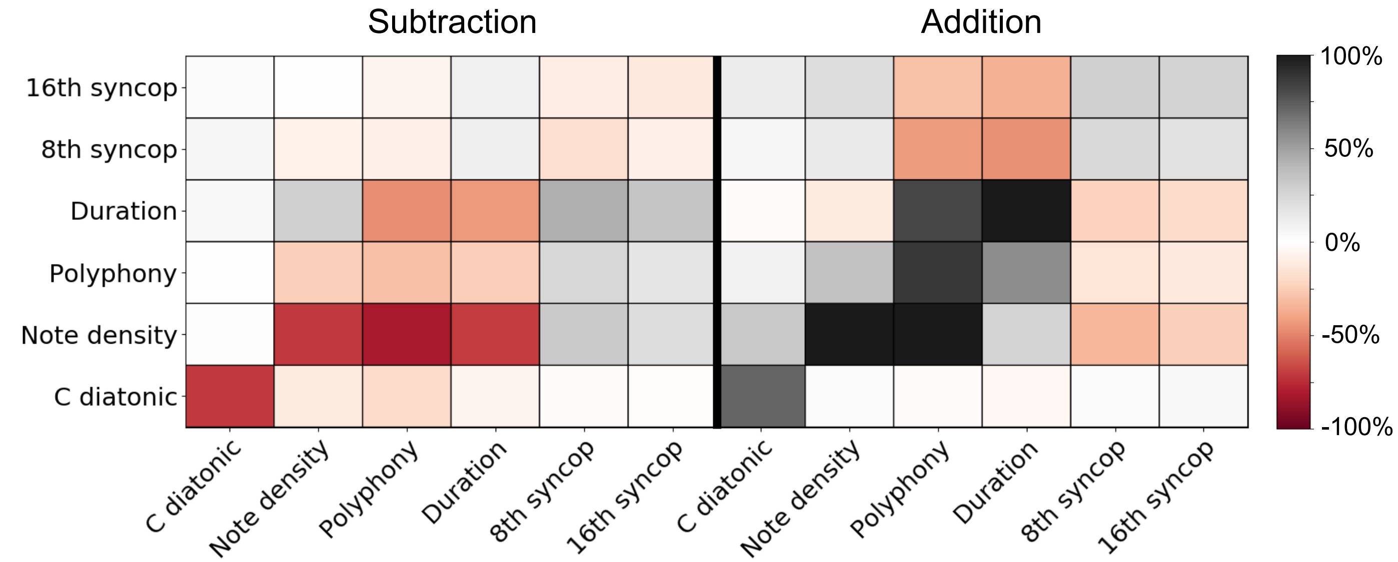
$$\frac{1}{N}\sum_{i=0}^{N}\epsilon_i^4 - \frac{1}{N}\sum_{j=0}^{N}\epsilon_j^1 = A_{vector}$$

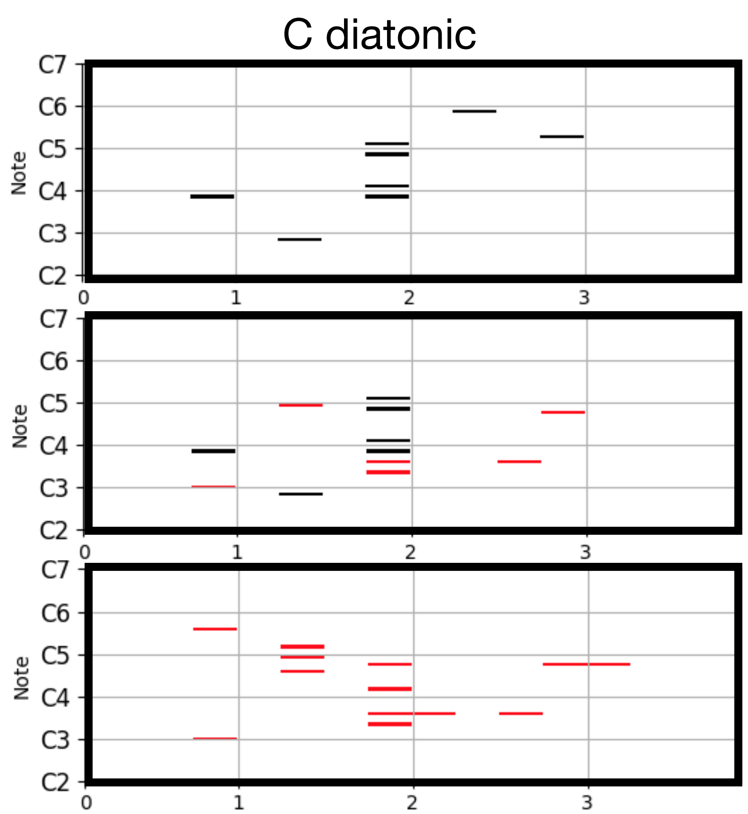
Attribute vector arithmetic

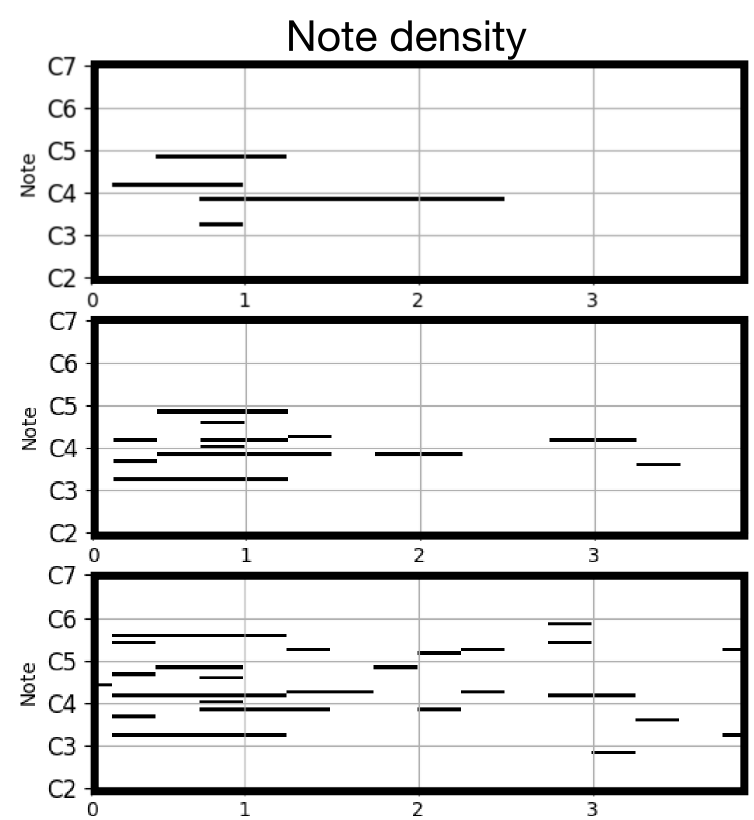
- **Percentage changes** on the attributes of 256 generated bars

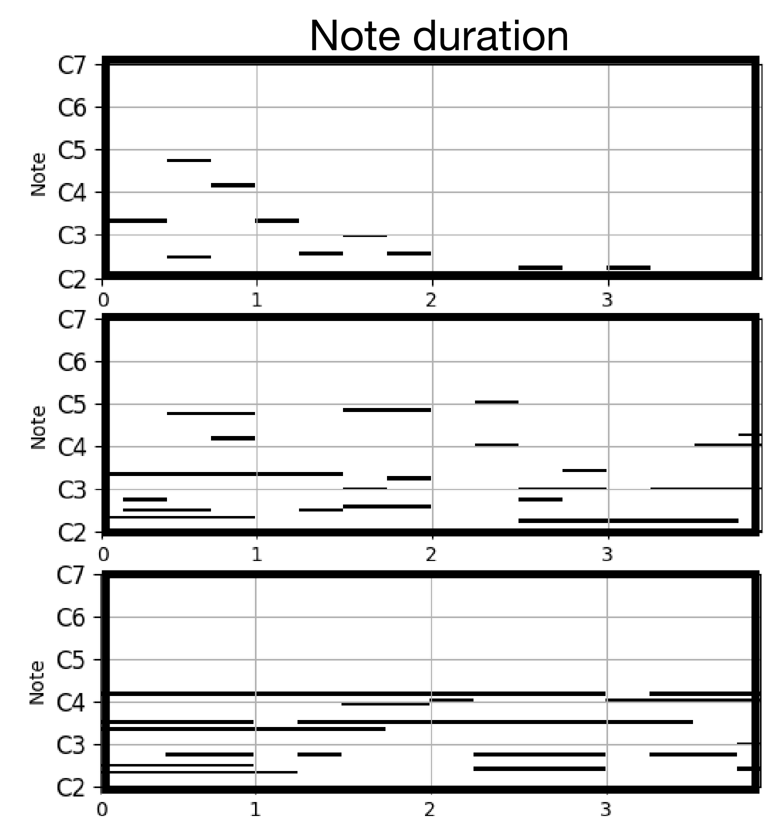
Significant change in the corresponding attributes

With few side effects





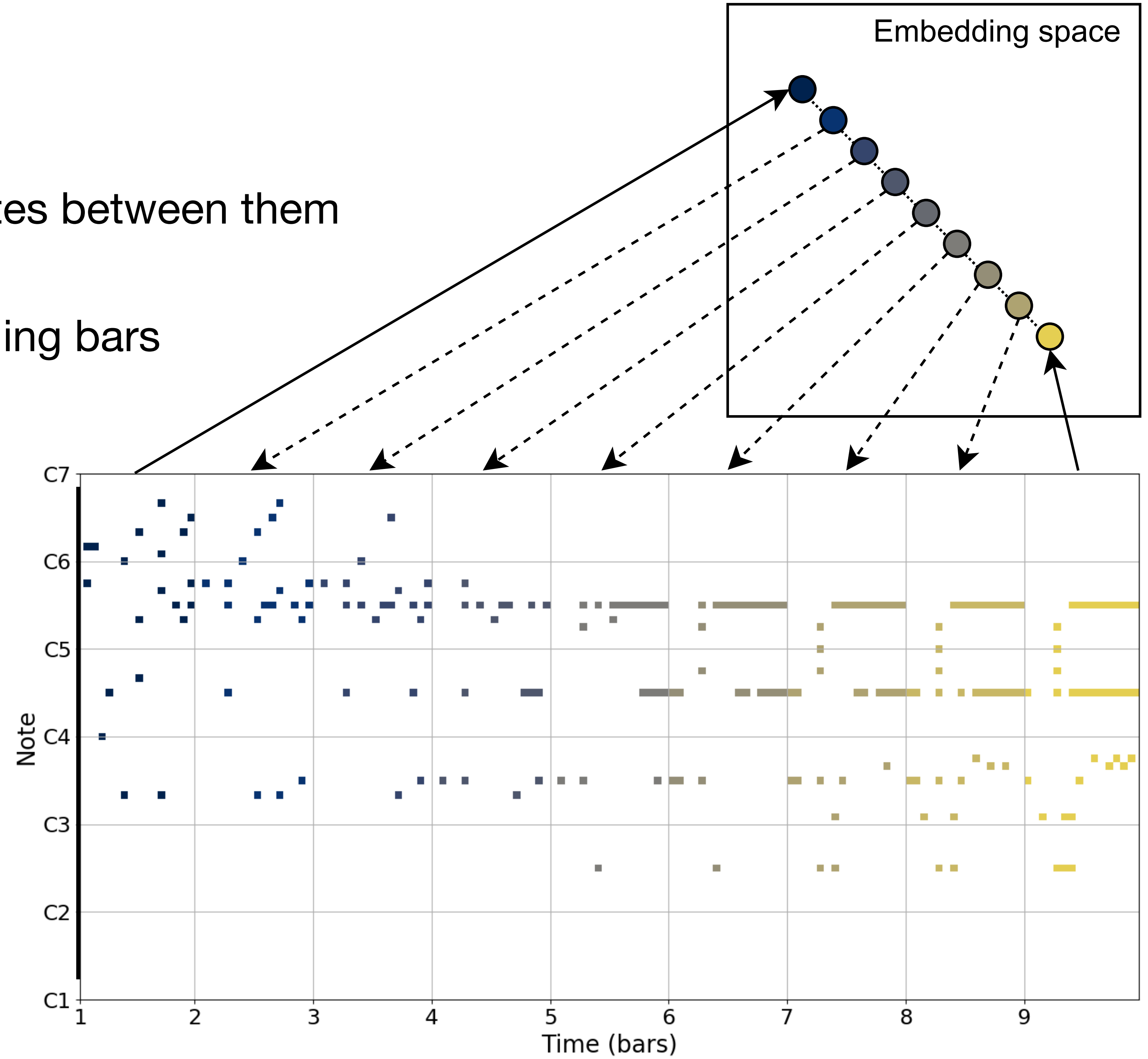




Interpolation

- **Interpolation** between two points in the embedding space
- **Embed** two bars
- **Interpolate** the coordinates between them
- **Generate** the corresponding bars

Can serve as basis for recommendation tools



I) First method - Adapt the NLP methods

II) Second method - Variational Auto-Encoders

III) Applications

IV) Conclusion

Overall summary

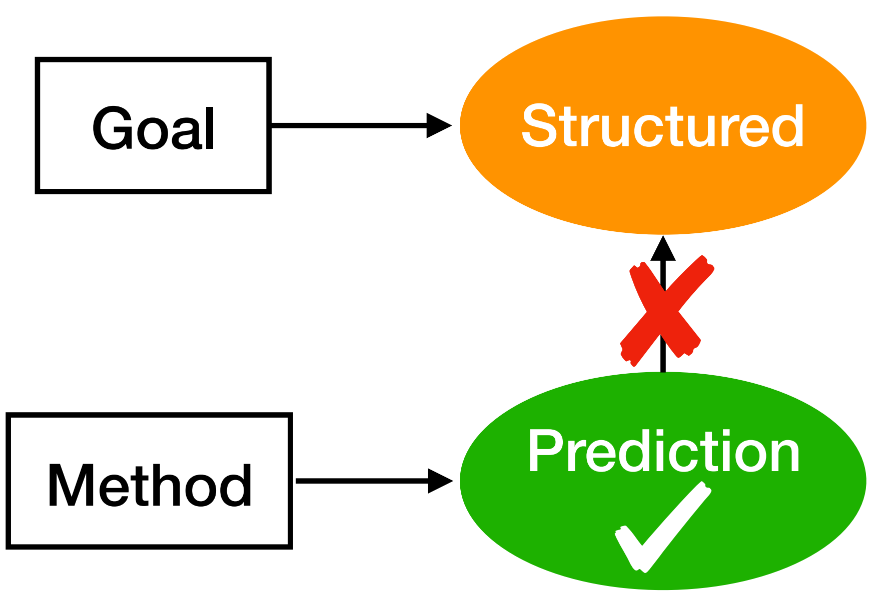
- Explore a method inspired by the NLP field which rely on the prediction task

__CNN-LSTM__ based architecture designed to capture musical and temporal features in piano-rolls

Greatly improve by a __Hierarchical Attention Mechanism__ able to distinguish harmonic salience of elements at all level of abstraction

Very good __prediction accuracy__ score

Lack of __control__ over the latent space properties leading to __orthogonalization__



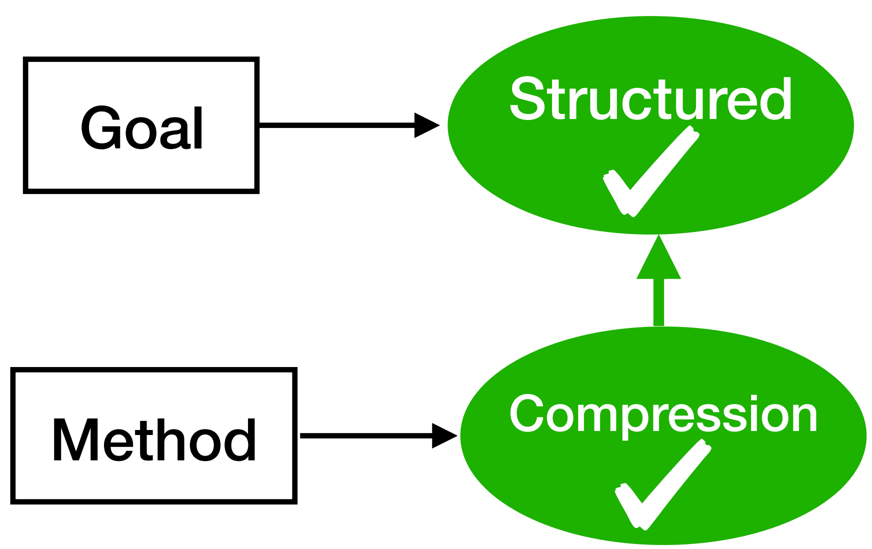
Overall summary

- Define a new approach relying on the Variational Auto-Encoders

  Implementation of an architecture similar to MusicVAE but for ***polyphonic*** *data*

  Introduction of a ***new symbolic representation*** *for small polyphonic excerpts inspired by the audio signal*

  Conduction of an extensive ***benchmark*** *against the main symbolic representation showing the efficiency of our proposal*



- Applications demonstrating the potential of our space for creative or analytical tools

  Composer classification *tool*

  Attribute vector arithmetic *allowing the shift of a given musical attribute in a bar*

  Smooth and realistic ***interpolations*** *showing the benefit of our space in compositional tools*

Future works

- Multimodal embedding framework

  Symbolic, audio, perceptual information

  Powerful tools : audio synthesis from the score, score transcription from the audio signal, perceptual effect predictor and generator

- Identify and discriminate information-carrying dimensions

  Greater control on the generation and modification of embedded bars

  Powerful tool : precisely assessing the compositional process of a given composer or musical trend

Thanks for your attention

Jury members :

- *Reviewers :*

  Frédéric Bimbot

  Anna Jordanous

- *Examiners :*

  Jean-Pierre Briot

  Simon Colton

  Florence Levé

  Geoffroy Peeters