Symbolization for time series : towards a unified framework for interpretable analysis of multivariate and multimodal time series

Laurent Oudre
laurent.oudre@ens-paris-saclay.fr
Séminaire LPSM 2026

Centre Borelli

- Applied mathematics lab at ENS Paris-Saclay
- Fusion of CMLA (Mathematics - ENS Cachan) and COGNAC G (Neurosciences - Université Paris Descartes)
- Identity : interaction with various disciplines (physics, biology, medicine, humanities, industry…)
- No formal “teams”, but a collective vision for addressing concrete problems and interacting with experts from other fields

Time series at Centre Borelli

- 3 main application fields

  - **Medicine :** physiological signals of different forms (accelerometry, 2D eye-tracking trajectory, multi sensor 3D motion capture, plethysmography, EEG, ECG…)
  - **Biology :** video and tracking devices
  - **Industry :** data collected on production lines or during maintenance

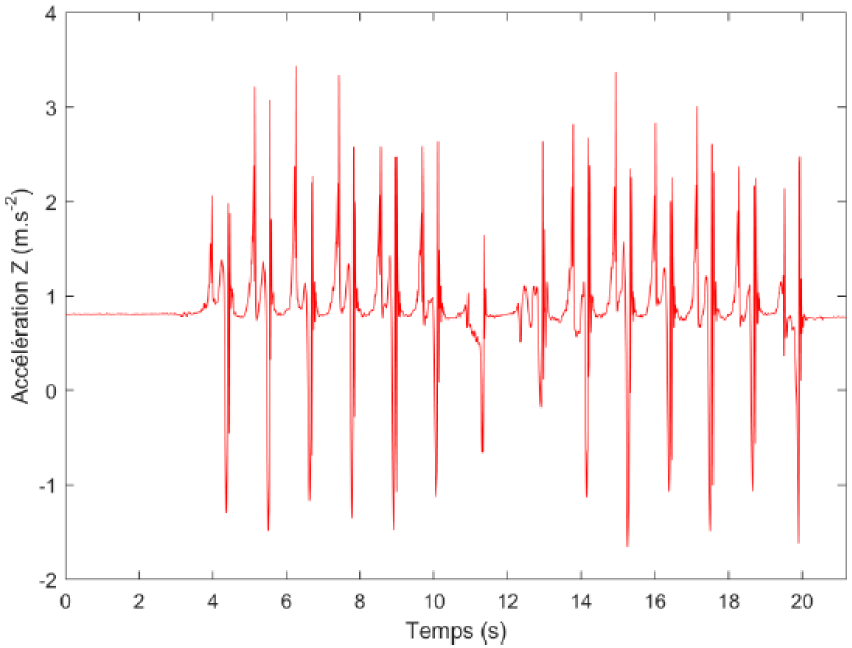
- All the projects involve real practitioners : medical doctors, biologists, industrial operators, neuroscientists

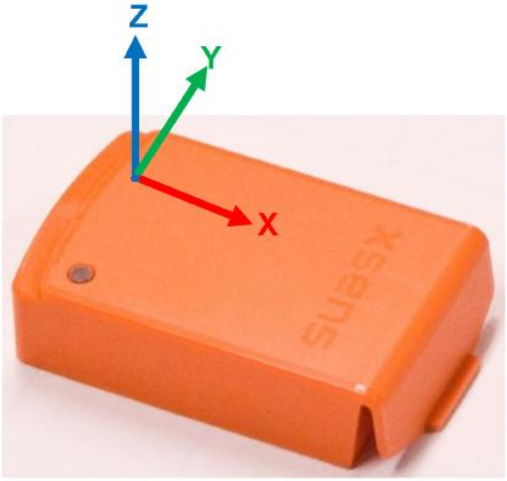
- Our aim is to provide “useful” solutions for them, while deriving some theory on the considered signals/algorithms : no “off the shelf” tasks

The main scientific questions

- Quantification of human and animal behavior
- Implementation of measurement chains "*pipelines*", platforms and intelligent tools but also of procedures for analysis, measurement and processing of data
- Creation of tools for diagnostic assistance, inter-individual comparison and longitudinal follow-up
- Integration into a clinical environment and interaction between algorithms and medical/neuroscience experts
- Publication of open-source annotated datasets

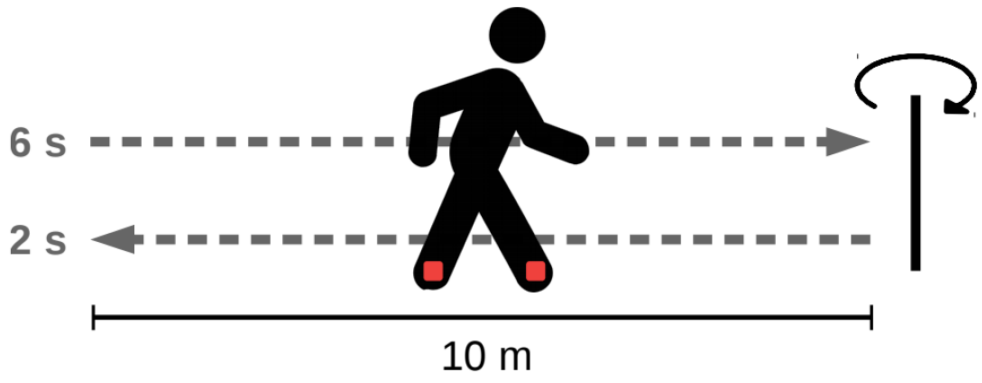
First usecase : gait analysis





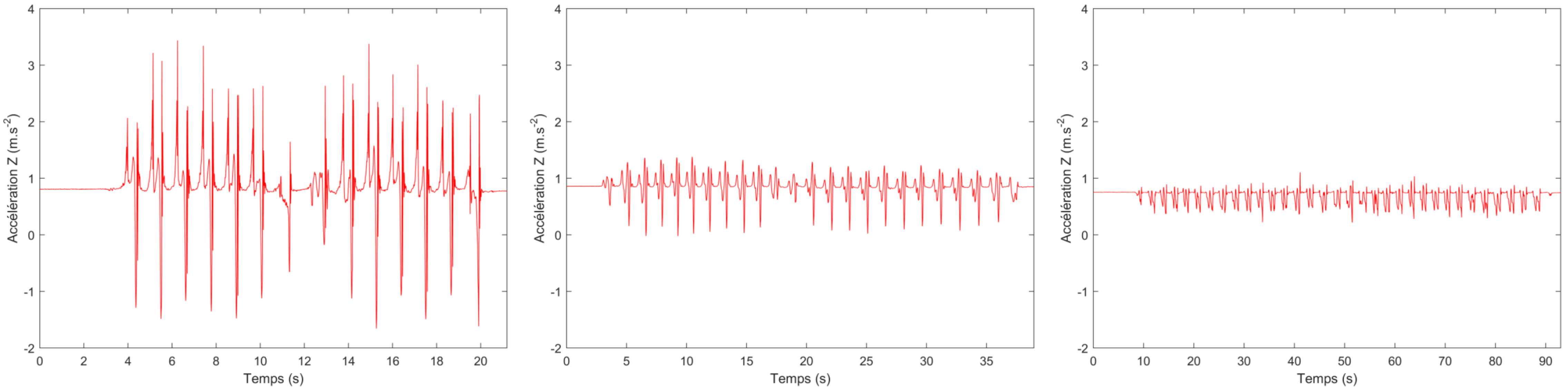
- Objective quantification of locomotion using inertial measurement units
- Longitudinal follow-up for neurodegenerative diseases
- Field testing: small wireless sensors and fully automatic device for routine use in medical consultations

First usecase : gait analysis



- Comfort speed protocol: stop (6 sec), walk forward (10 m), turn around, return, stop
- Four wireless inertial units: left foot, right foot, lower back, head
- Nine signals per sensor: linear acceleration (3D), angular velocity (3D), magnetic field (3D)

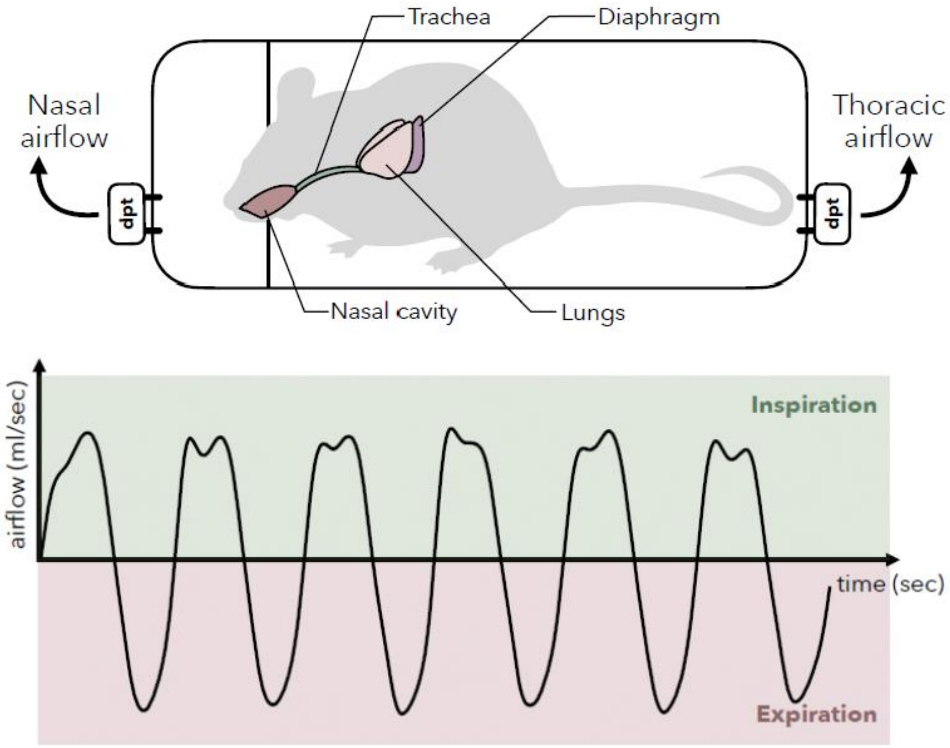
First usecase : gait analysis



Scientific questions and challenges

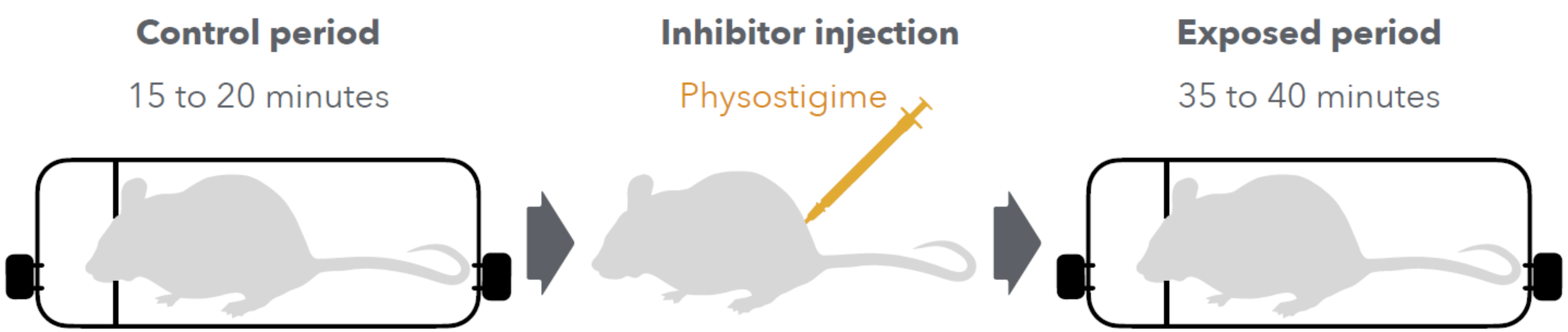
- Non-stationary signals, notion of repetitive patterns
- Aim : interindividual comparison + longitudinal follow-up
- How could we define a notion of « distance » for such signals ?

Second usecase : plethysmography



- Objective quantification of respiratory activity of mouse
- Study the impact of some genotype/pathogen (acetylcholine, a neurotransmitter controlling ventilation) on the respiratory function
- Double chamber plethysmography : access to the nasal flow and volume

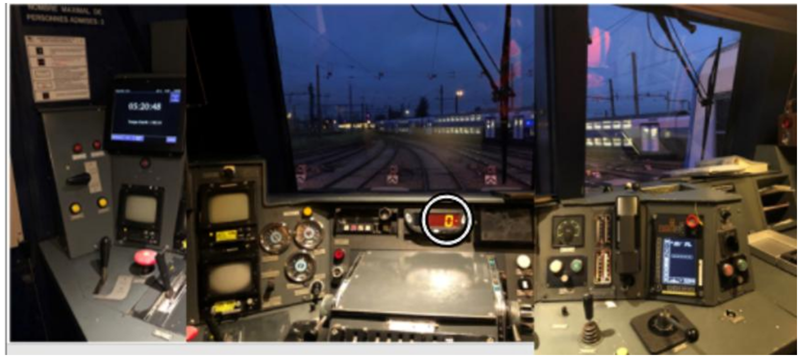
Second usecase : plethysmography



Scientific questions and challenges

- One recording : 30000+ breathing cycles
- Existing pipelines : features before/after and statistical tests
- Aim : study the evolution and the deformation of these cycles along time
- How could we compare the respiratory behaviors of different animals and experimental conditions ?

Third usecase : cognitive workload

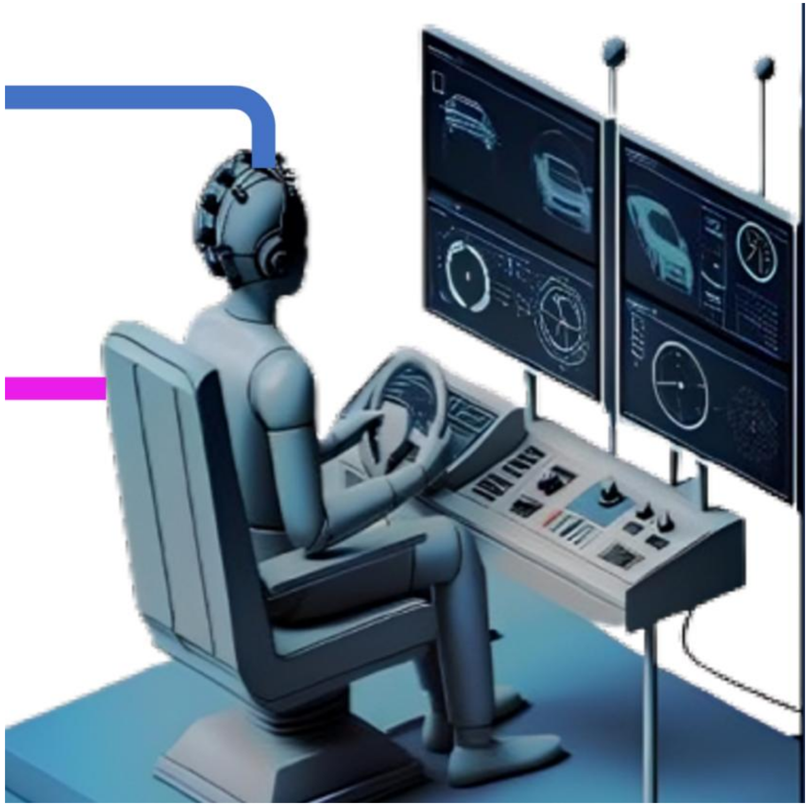


- Objective quantification of cognitive workload

- Eye-tracking (2D trajectories), electrocardiogram (ECG), electrodermal activity (EDA)

- Multimodal setting : different sampling frequencies, different physical properties

Third usecase : cognitive workload

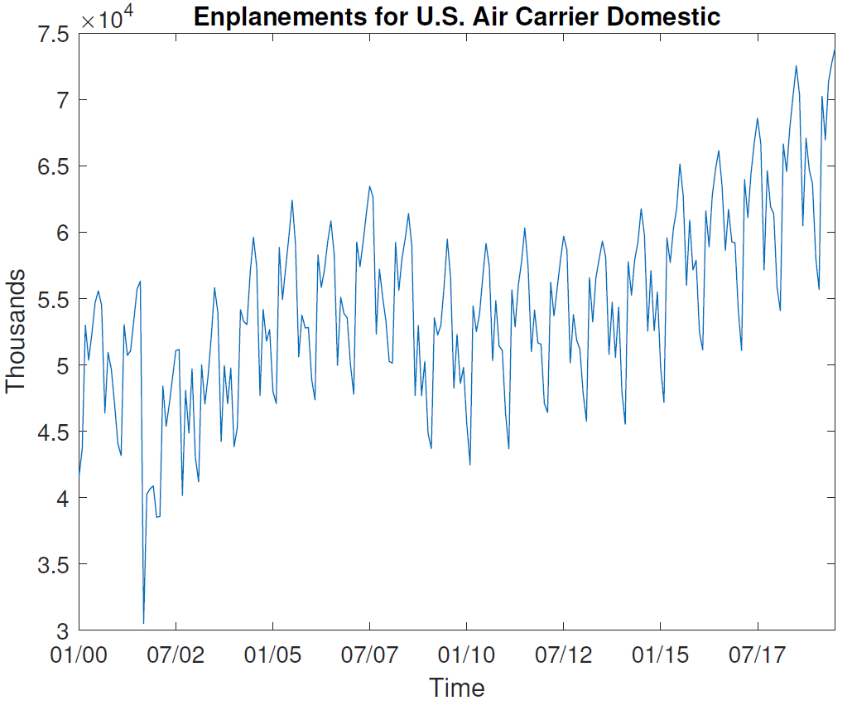


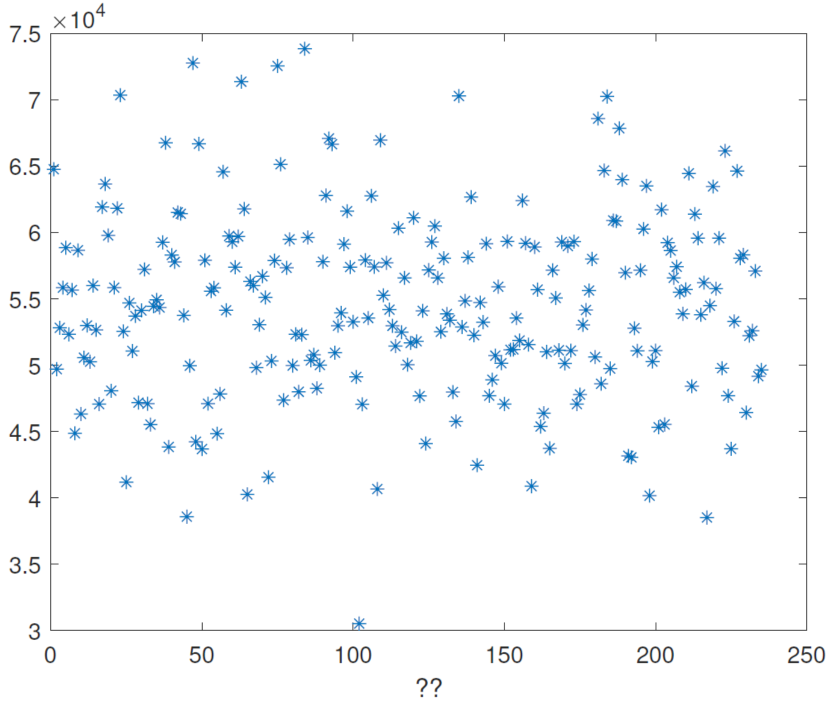
- CL-Drive dataset : 9 graded difficulty levels, 3 minutes recordings
- Scientific questions and challenges
  - Multimodal setting : fusion of various sensors
  - Notion of distance for multimodal data streams to perform classification

The main challenges

- Real data
  - Noisy and/or missing data
  - Small amount of data
  - No or only a few annotations
  - Multivariate, multimodal, heterogenous
- Importance of the expertise
  - Confusing and hyper-specialized annotations (very rare to have proper labels)
  - Difficulty to translate expertise into ML-compatible annotations
- Importance of the “good relationships” with the experts
  - They expect to understand everything : no black boxes
- How do you interact with time series ? (as opposed to images, sound, and text)

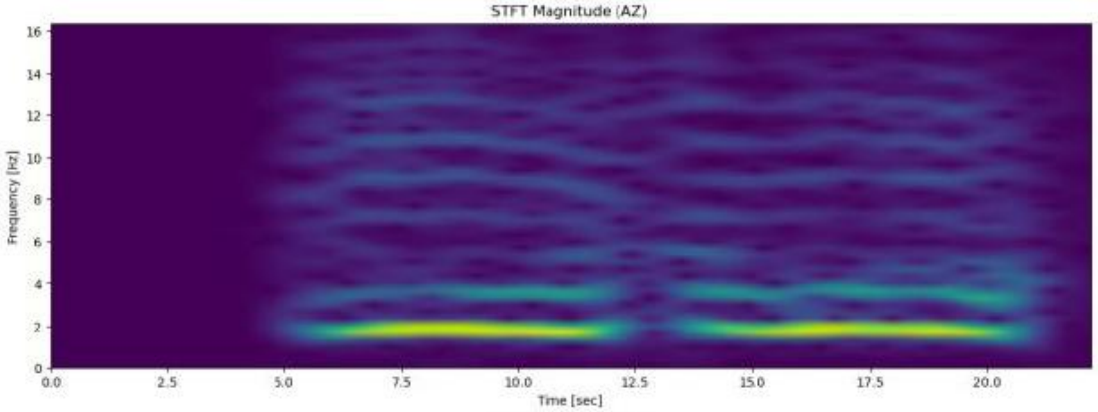
What about time ?

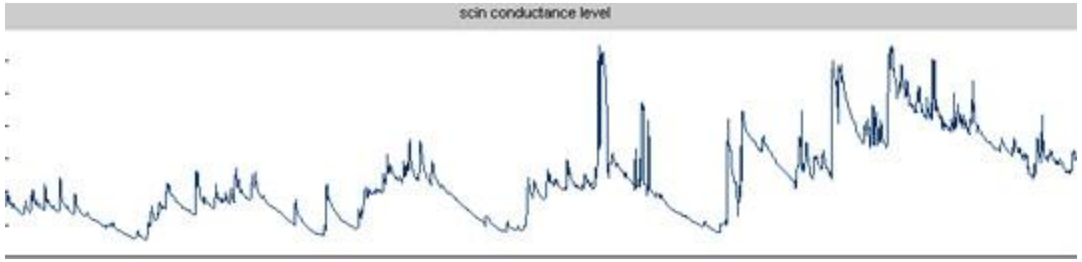




- What is the difference between regular data and time series ? Notion of sequence and chronology
- Time allows to study the evolution of the phenomenon and should be taken into account for processing the data

Structured time series data





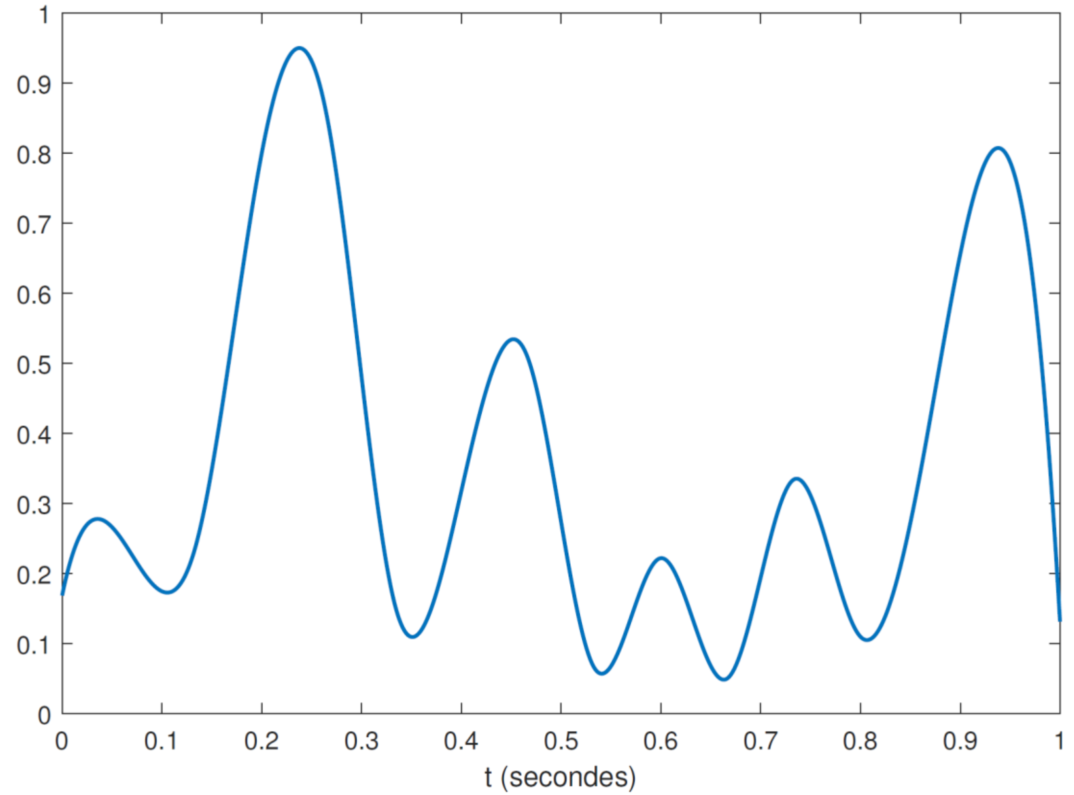
- Each physiological signal has its own “inner structure” : patterns, regimes
- Importance of the notion of events
- Analysis should be performed with a relevant analysis grid : importance of the time representation

The dream

- Summarize the content of potentially large time series in a low dimensional space
- This low-dimension representation should be simplified, semantic, and interpretable for the practitioner
- Robustness to noise and focus on the meaningful events
- Time information should be preserved as much as possible
- Longitudinal follow-up and interindividual comparaison: this low-dimensional space should be equipped with a metric

General principle of time series symbolization

General principle



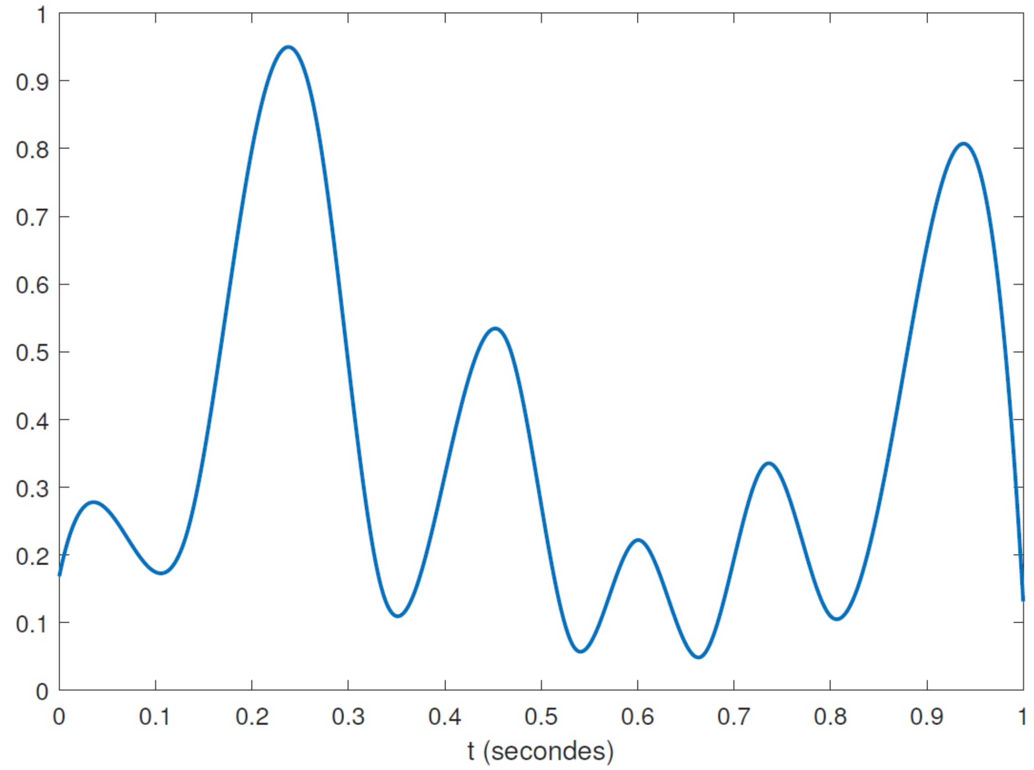
How would you describe such as signal?

small up, small down, big up, big down, up, down…

CCCFJHCBEEBBBACCBEHF

- By describing the whole time series as a succession of local behaviors, it is possible to fully encode the time series
- This encoding process is referred to as symbolization : it takes a time series as input and outputs a sequence of discrete symbols
- Each symbol represents a local behavior (amplitude, slope…) and statistical features can be extracted from that (number of occurrence of each symbol, number of transitions, etc…)

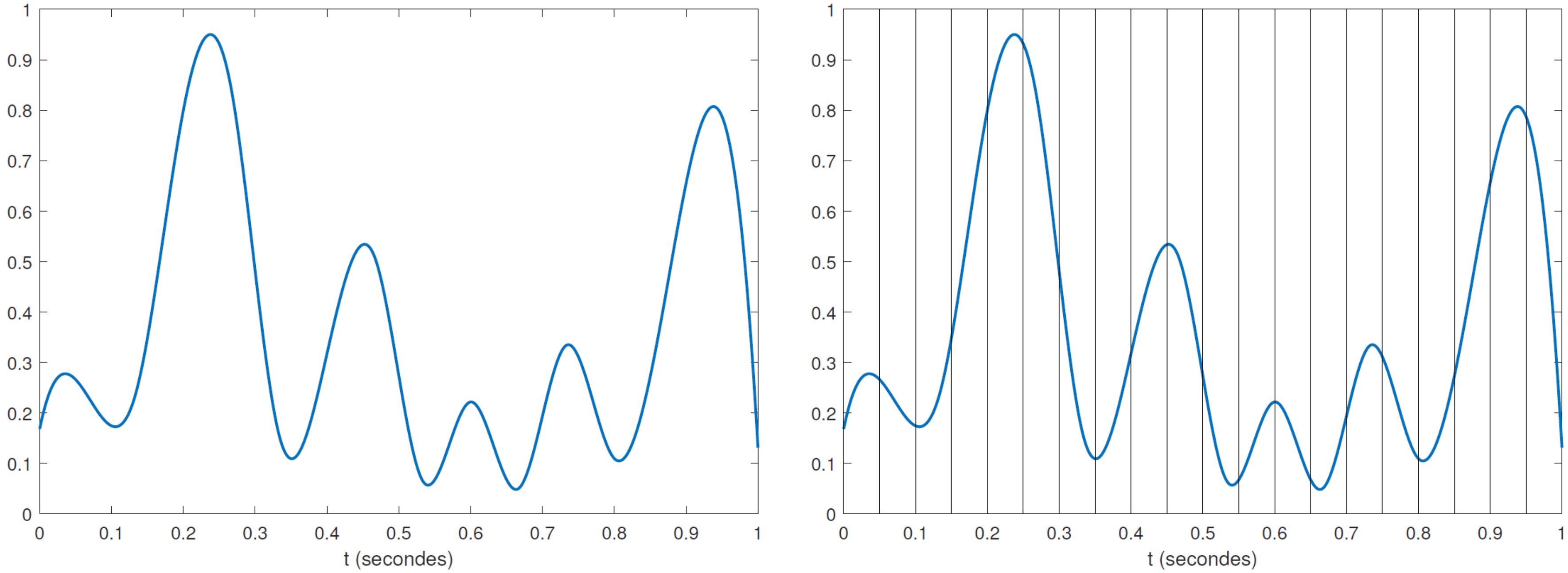
General principle



CCCFJHCBEEBBBACCBEHF

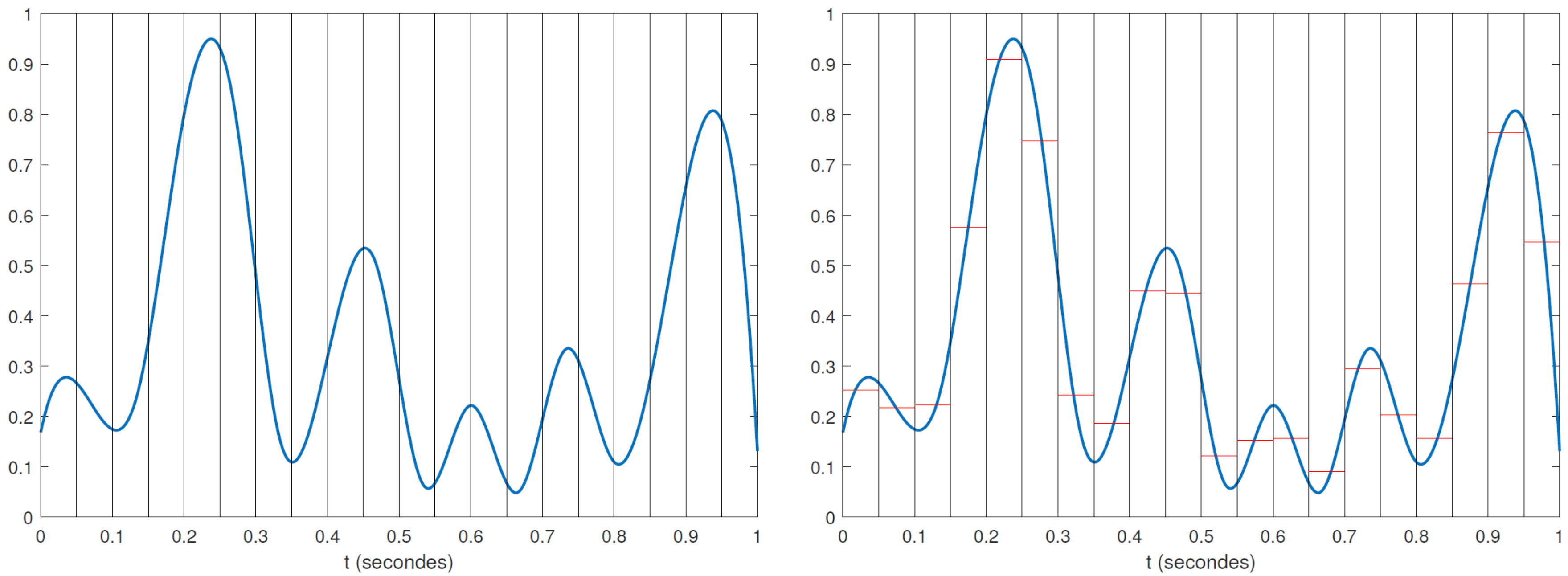
- Aim : transform a (possibly multivariate or multimodal) real-valued time series into a sequence of symbols
- N : original number of samples
- w : final length of the symbolic sequence
- A : alphabet size

Example



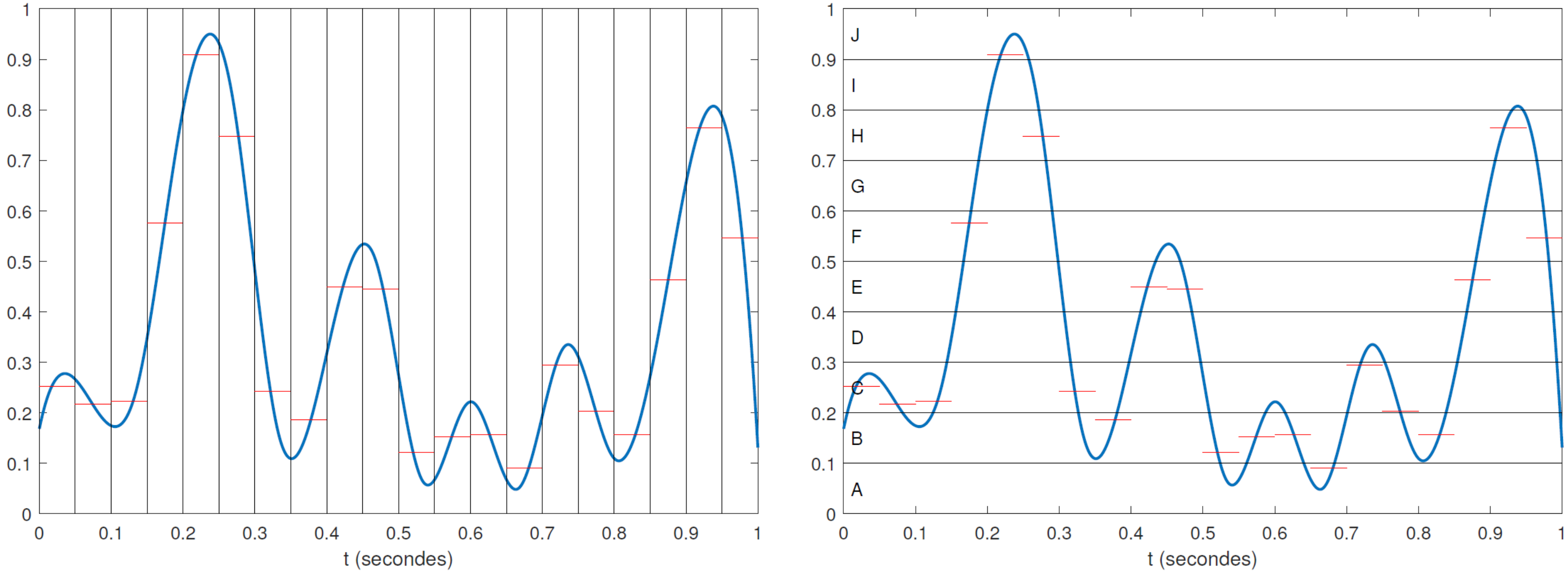
Division into non-overlapping frames of 500 samples

Example



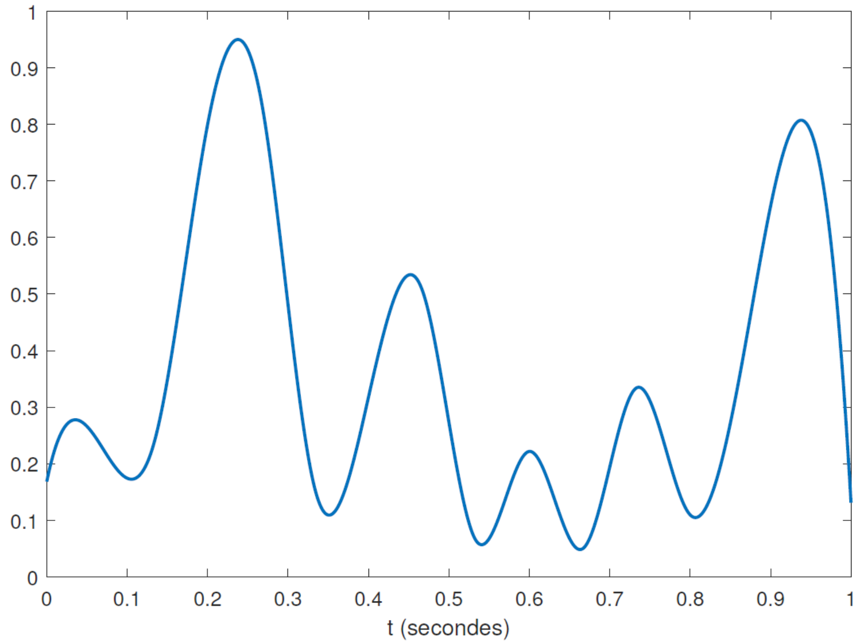
Average value on each segment

Example



Quantization of the amplitudes

Example

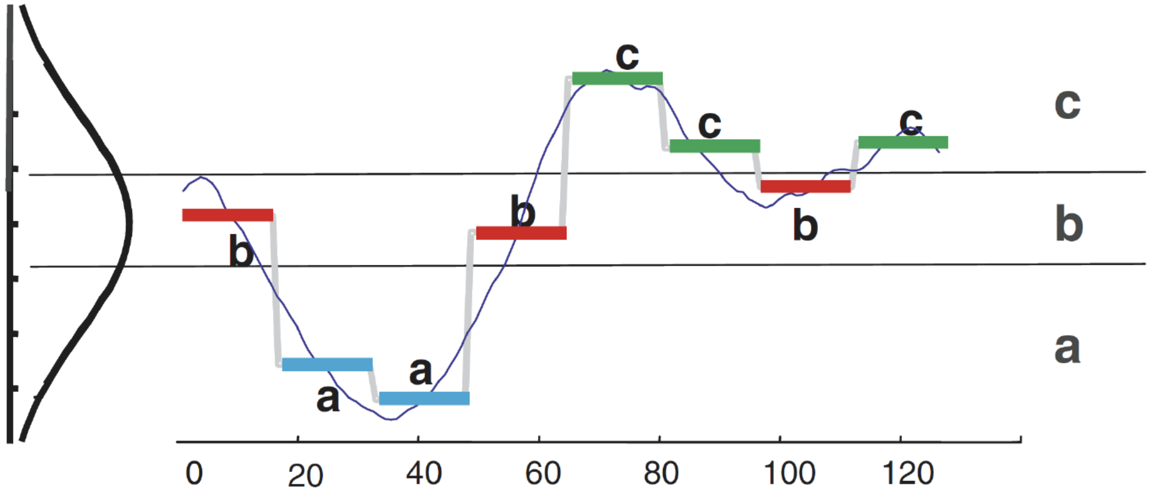


CCCFJHCBEEBBBACCBEHF

Symbolization of the time series

640 kbits vs. 0.1 kbit !

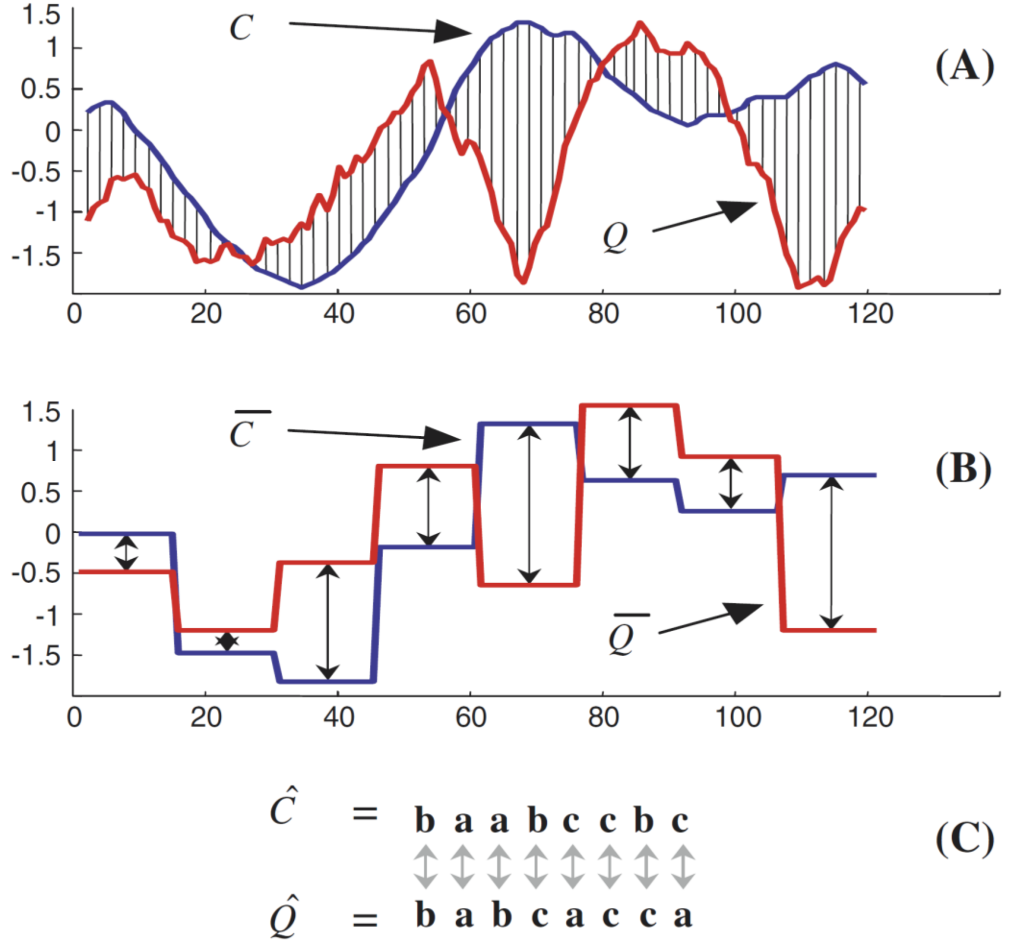
The SAX representation



- SAX (Symbolic Aggregate ApproXimation) with fixed *w* (length of output sequence) and alphabet size *A*
- Allows for fast indexing of the waveform : constant piecewise approximation
- z-normalization of the time series so that each symbol is equiprobable on the dataset

- Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing SAX: a novel symbolic representation of time series. Data Mining and knowledge discovery, 15(2), 107-144.

The SAX representation



- Definition of a “metric” such that :

$$\mathrm{MINDIST}\left(\mathrm{SAX}(\mathbf{x}), \mathrm{SAX}(\mathbf{y})\right) \leqslant d_{EUC}(\mathbf{x}, \mathbf{y})$$

	a	**b**	**c**	**d**
a	0	0	0.67	1.34
b	0	0	0	0.67
c	0.67	0	0	0
d	1.34	0.67	0	0

This table is for an alphabet of cardinality of 4, i.e. $a = 4$. The distance between two symbols can be read off by examining the corresponding row and column. For example, *dist*(**a**,**b**) = 0 and *dist*(**a**,**c**) = 0.67.

- “Rigid” distance : assumes a perfect time alignment between the two time series

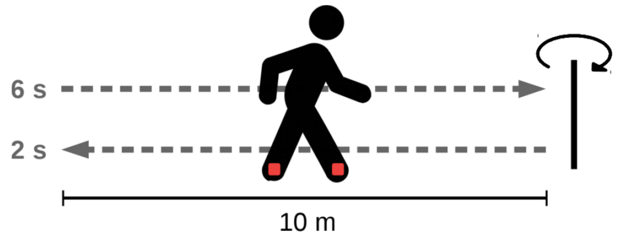
- Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing SAX: a novel symbolic representation of time series. Data Mining and knowledge discovery, 15(2), 107-144.

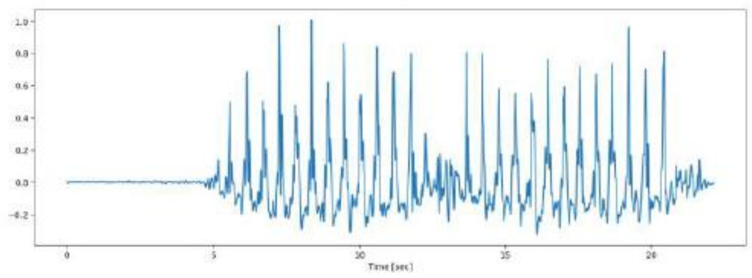
The limitations of these approaches

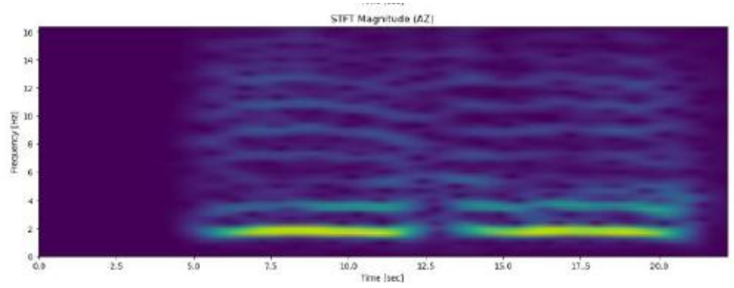
- Numerous variants : piecewise linear, min/max, local Fourier coefficients… The aim is to compress the waveform
- No semantics, no adaptability
- Only usable for fixed-length time series, no possibility of handling deformations of the timeline
- How can we adapt this to the introduced usecases ?

Symbolization based on regimes

First usecase : gait analysis

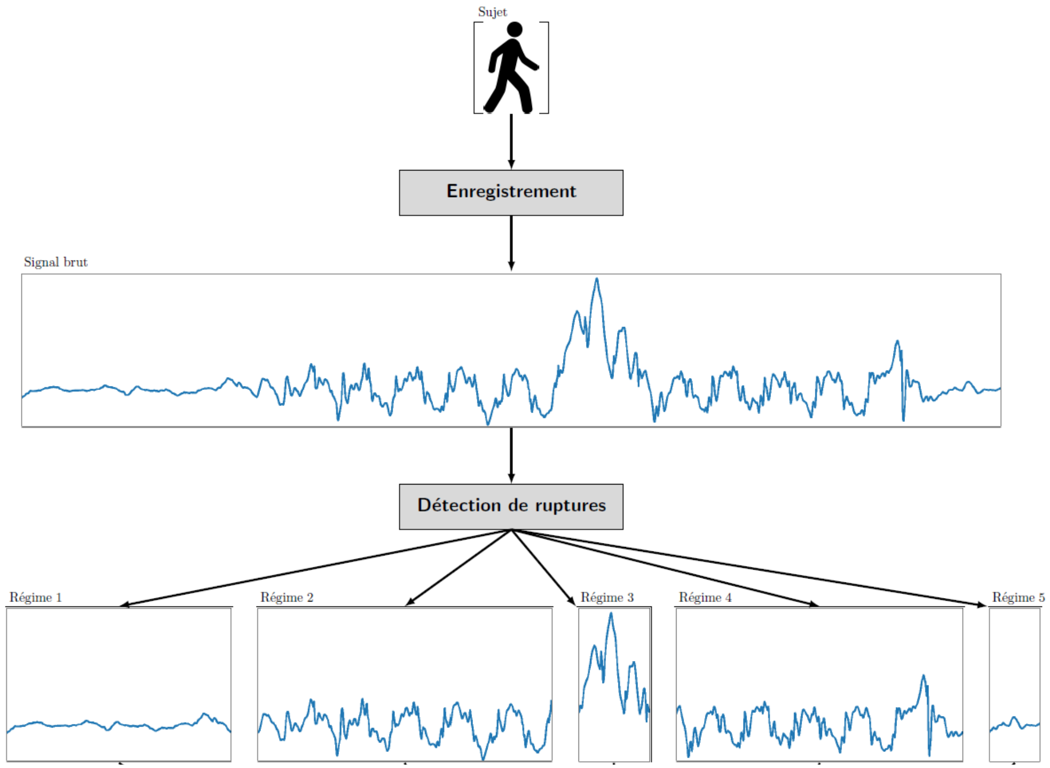






- Analysis grid : the different regimes of the protocol
- Periodic activity : transformation into the time-frequency domain
- Each regime is characterized by a spectral signature
- Aim : compare two spectrograms (seen as a multivariate time series) with a dedicated measure of fit

Change-point detection



Input :

- Multivariate time series $y = \{y_t\}_{t=1}^{T}$

Assumption :

- The signal undergoes some *changes* at times

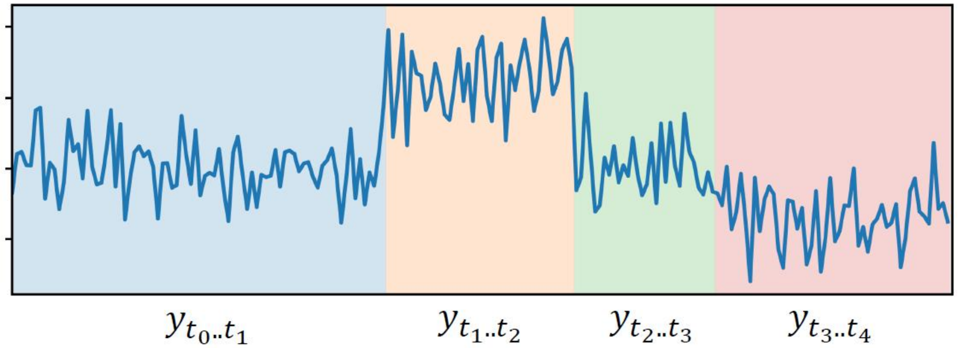
$$\mathcal{T}^* = (t_1^*, \ldots, t_{K^*}^*)$$

Problem :

- Estimate the number of changes and their temporal positions

Change-point detection

$$(\hat{t}_1, \ldots, \hat{t}_K) = \underset{(t_1,\ldots,t_K)}{\text{argmin}} \sum_{k=0}^{K} c(y_{t_k..t_{k+1}})$$



Notion of cost function

- Defines the notion of homogeneity and thus the type of detected breakpoints
- Often linked to a probabilistic model

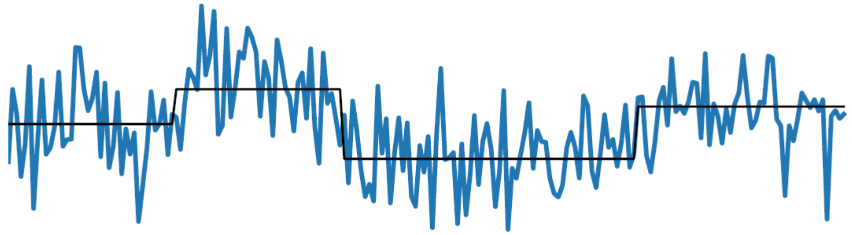
Resolution method

- Dynamic programming for the optimal resolution
- Approximate resolution with sliding windows

Change-point detection

Cost function : L2 norm in the time-frequency domain

- Gaussian model with fixed variance
- Detects changes in spectral content: change in the continuous component + change in instantaneous frequency



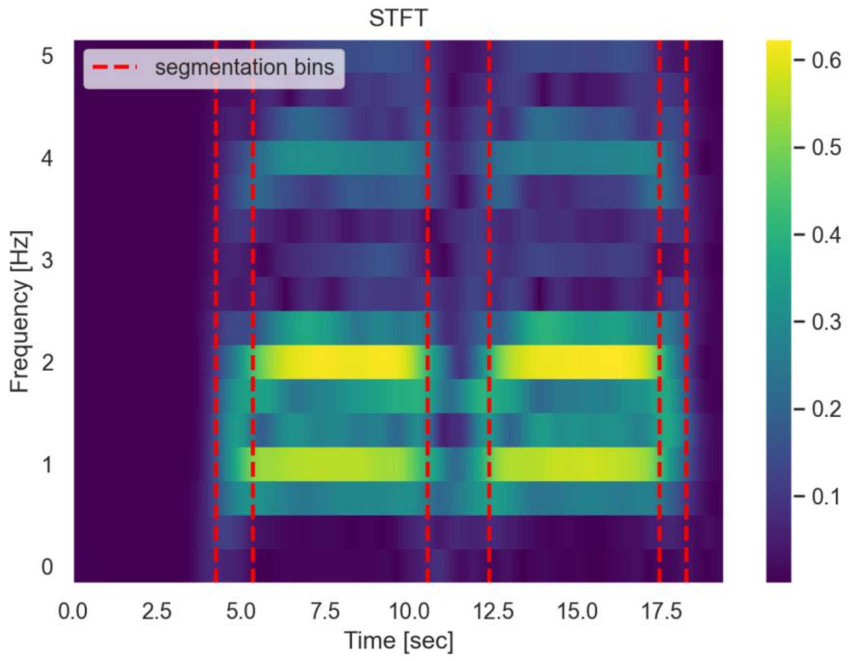
$$c_{L_2}(x[a:b]) = \sum_{n=a+1}^{b} \|x[n] - \mu_{a:b}\|_2^2$$

Estimation of the number of change points

- Penalized resolution of the problem : compromise between data fitting and complexity
- PELT : Pruned Exact Linear Time algorithm
- Fixed level of granularity with parameter β

$$(\hat{t}_1, \ldots, \hat{t}_{\hat{K}}) = \underset{(t_1,\ldots,t_K),K}{\text{argmin}} \sum_{k=0}^{K} c(x[t_k : t_{k+1}]) + \beta K$$

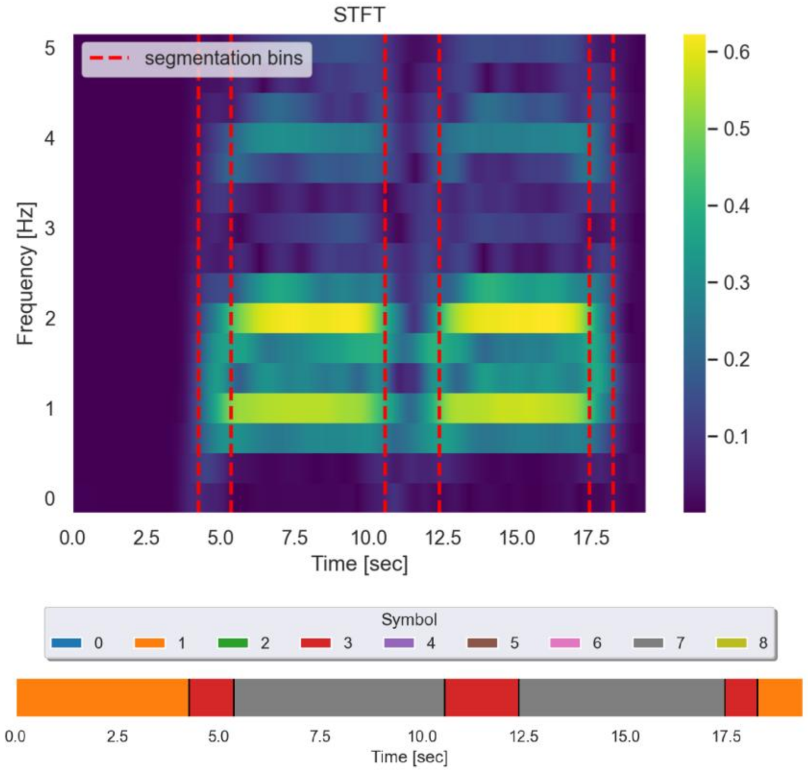
Step 1 : change-point detection in the time-frequency domain

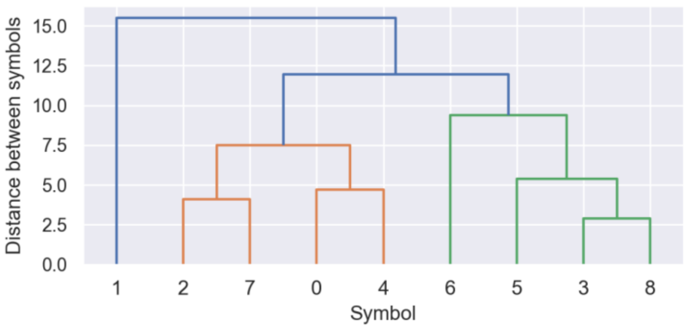


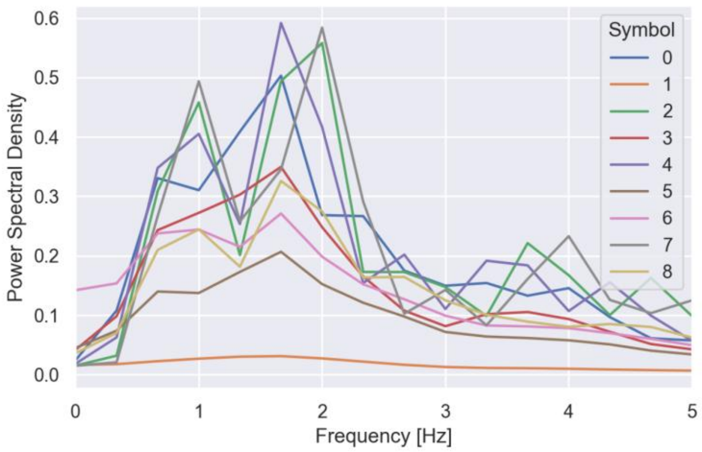
- Automatic identification of the different regimes in the signal
- Parameter **β** allows to define the adequate granularity
- Compared to SAX : one window = one regime
- All frames belonging to the same regime will be assigned the same symbol : adaptative symbolization

Step 2 : clustering to define the symbols

- Each regime can be summarized by the mean vector in $\mathbb{R}^d$
- Instead of using fixed boundaries between symbols, use of adaptative clustering (K-means or other)
- Each symbol represents a spectral behavior + a notion of distance can be constructed between symbols







Step 3 : construction of the symbolic sequence

- The vocabulary (clusters) are estimated from all the segments of the dataset
- Encoding of a whole dataset into symbolic sequences

Step 4 : construction of the distance

Algorithm 2: General Levenshtein distance, with $\mathcal{A}$ the alphabet of size A.

Data: Strings $\mathbf{x} \in \mathcal{A}^m$ and $\mathbf{y} \in \mathcal{A}^n$; deletion costs $W_{del} \in \mathbb{R}^A$, insertion costs $W_{ins} \in \mathbb{R}^A$, substitution costs $W_{sub} \in \mathbb{R}^{A \times A}$

Result: General Levenshtein distance $D_{\mathsf{gLev}}(\mathbf{x}, \mathbf{y})$

$C \leftarrow 0_{m \times n}$;

for $i \leftarrow 1$ **to** m **do**

  $C(i, 1) = i$;

for $j \leftarrow 1$ **to** n **do**

  $C(1, j) = j$;

for $i \leftarrow 2$ **to** m **do**

  for $j \leftarrow 2$ **to** n **do**

$$C(i,j) = \min \begin{cases} C(i, j-1) + W_{ins}(j) \\ C(i-1, j) + W_{del}(i) \\ C(i-1, j-1) + W_{sub}(i,j) \end{cases} ;$$

return $C(m, n)$;

- Construction of an *elastic* distance d_{symb} between symbolic sequences
- General Levensthein distance :
  - Insertion of a character
  - Deletion of a character
  - Substitution of characters
- Each operation is associated to a cost
  - Substitution : distance between the centroids of the two symbols
  - Insertion/deletion : maximum distance between two centroids
- Resolution with dynamic programming

Results

- Clustering (3 clusters : healthy, neurological, orthopedic), 442 signals
  - Working in the symbolic space better captures the differences between the gaits of the patients
  - Two signals that have common semantic properties will be « close » according to the metric

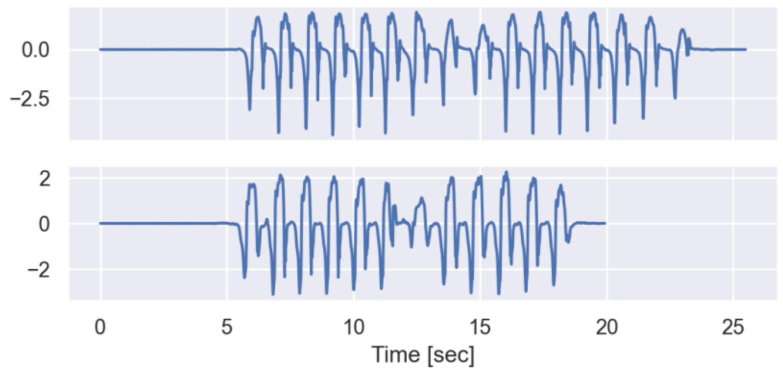
TABLE I
SUMMARY OF THE SILHOUETTE SCORES FOR EACH DISTANCE MEASURE, AVERAGED OVER ALL SIGNALS.

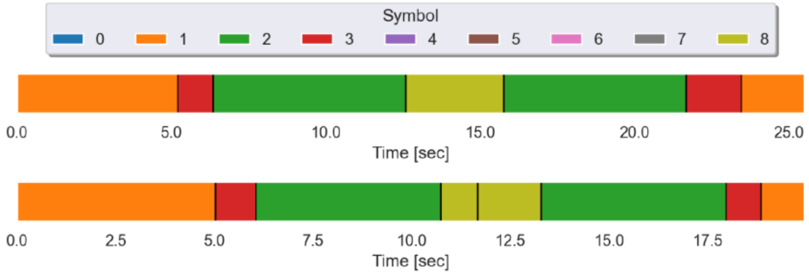
Distance measure	Mean Silhouette score	Median Silhouette score
DTW-D	0.15	0.18
DTW-I	0.15	0.19
d_{symb}	0.33	0.40

- Query by content : find the opposite foot in the dataset

TABLE II
MEAN RANK OF QUERYING A SIGNAL'S OPPOSITE LATERALITY.

	Pathology group		
Distance measure	**Healthy**	**Neurological**	**Orthopedic**
DTW-D	15.8	19.9	6.4
DTW-I	6.7	12.6	3.9
d_{symb}	9.3	5.1	2.0

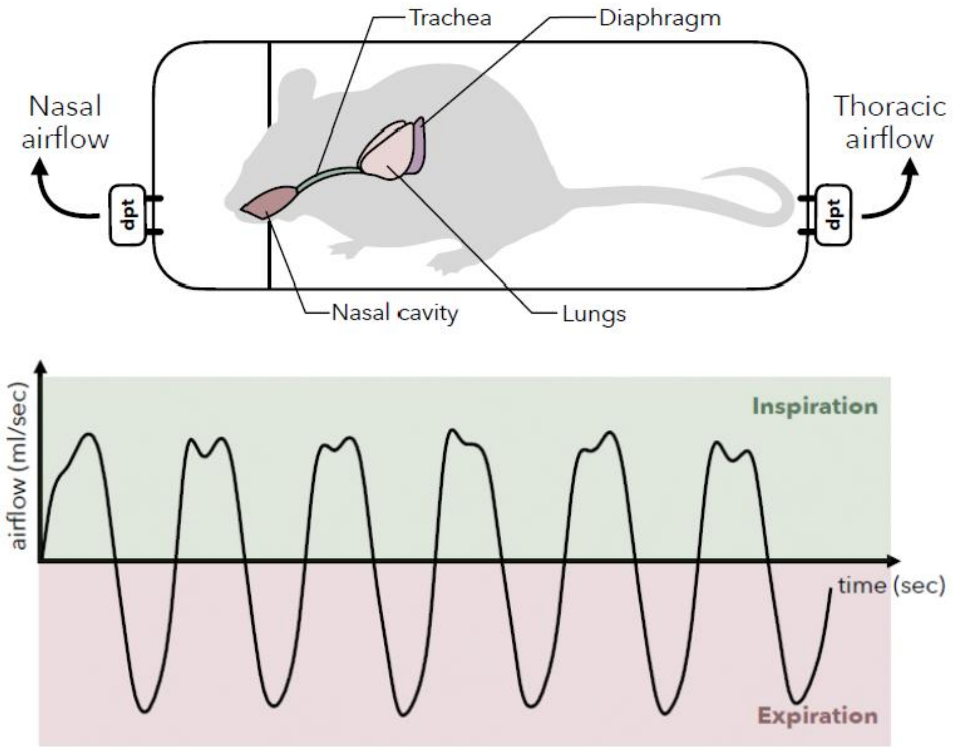




- S. W. Combettes, C. Truong, and L. Oudre. An Interpretable Distance Measure for Multivariate Non-Stationary Physiological Signals. In Proceedings of the International Conference on Data Mining Workshops (ICDMW), pages 533-539, Shanghai, China, 2023.
- S. W. Combettes, C. Truong, and L. Oudre. Symbolic representations for time series. In Proceedings of the European Signal Processing Conference (EUSIPCO), pages 1962-1966, Lyon, France, 2024.
- S. Perochon, and L. Oudre. Unsupervised Action segmentation of Untrimmed Egocentric Videos. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pages 1-5, Rhodes, Greece, 2023.

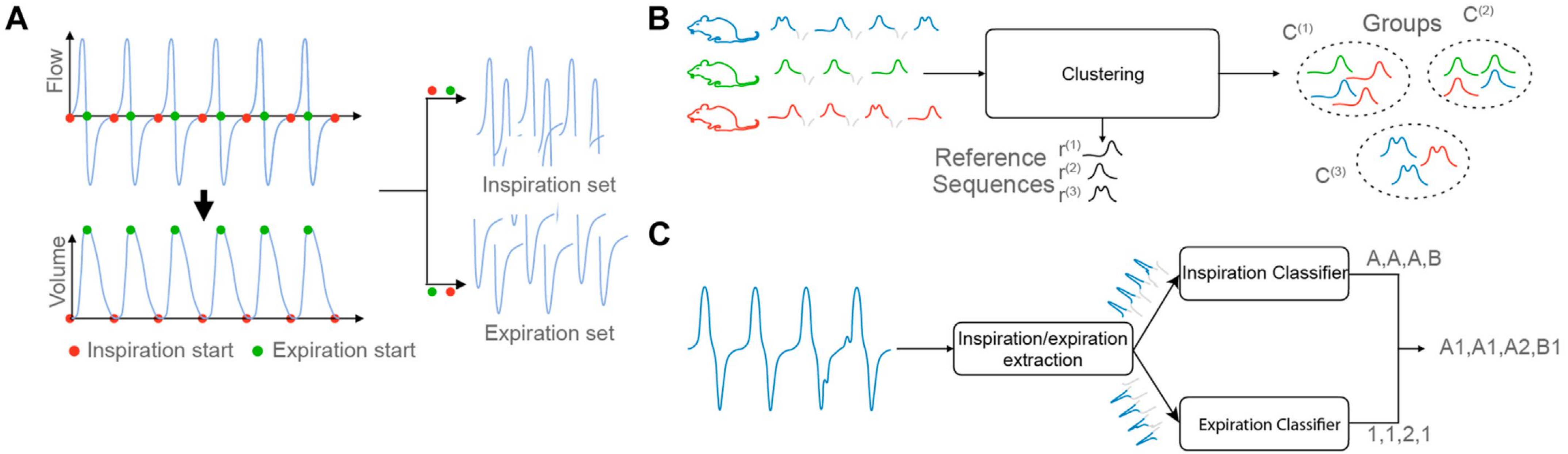
Symbolization based on patterns

Second usecase: plethysmography



- Analysis grid : the respiratory cycles (patterns)
- One symbol = one type of cycle
- Biologist’s input : separate the inspiration and expiration

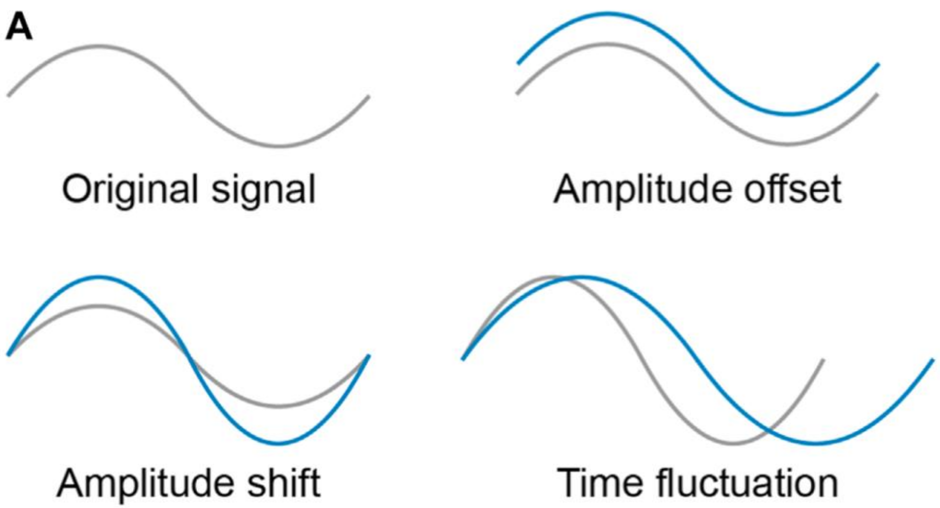
General pipeline



- Step 1 : Segmentation of the respiratory cycles (inspiration/expiration)
- Step 2 : Clustering and construction of the reference sequences
- Step 3 : Construction of the symbolic sequence

- Germain, T., Truong, C., Oudre, L., & Krejci, E. (2023). Unsupervised classification of plethysmography signals with advanced visual representations. Frontiers in Physiology, 14, 1154328.

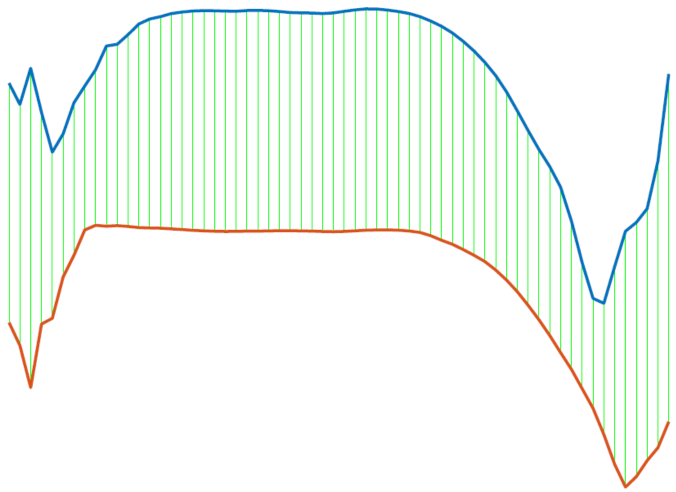
Clustering and creation of the symbols



- Input : set of inspiration/expiration cycles from different mice + different experimental conditions
- Aim : construct clusters + centroids from the set with K-means algorithm
- Concept of invariance: how to compare and calculate a barycenter from time series with deformations (amplitude and/or time line)?

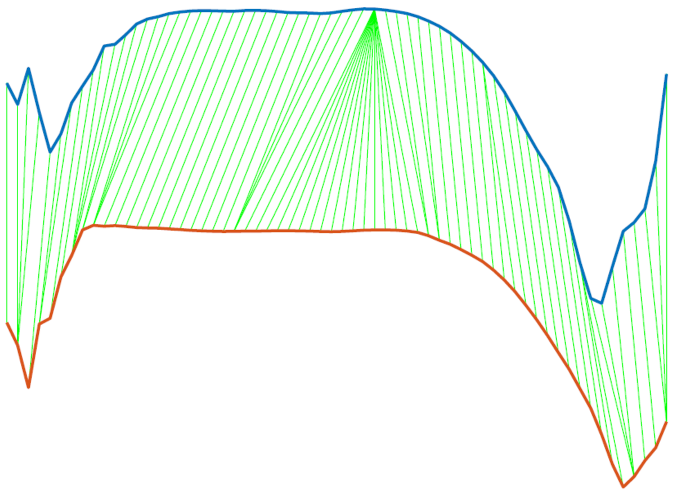
Elastic distances for time series

Euclidean distance



Sample $x[n]$ is associated to sample $y[n]$

DTW

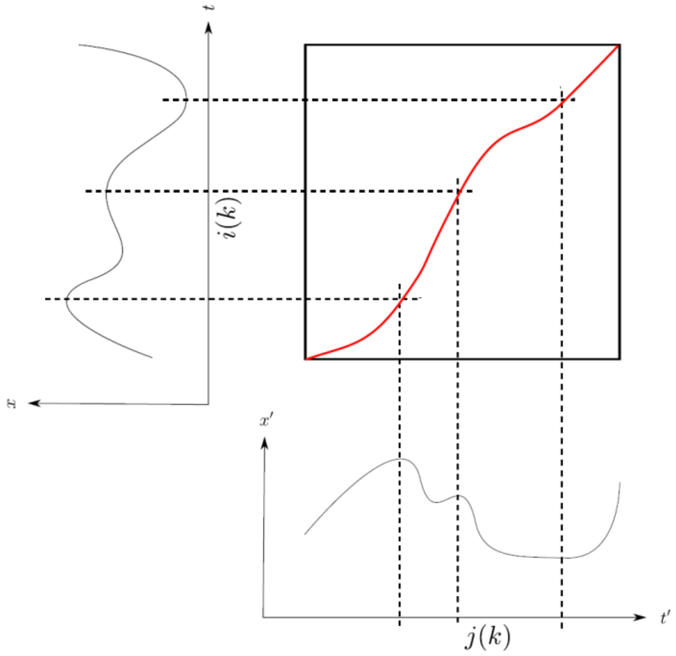


Sample $x[i_k]$ is associated to sample $y[j_k]$

DTW : Dynamic Time Warping

- Sakoe, H., & Chiba, S. (2003). Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing, 26(1), 43-49.

DTW : notion of path



- DTW computes a correspondence between the elements of $\mathbf{x}$ and those of $\mathbf{y}$.
- Mapping function called **path** :

$$P = ((i_1, j_1)), \cdots, (i_{K_P}, j_{K_P})) \in (\mathbb{N} \times \mathbb{N})^{K_P}, K_P \in \mathbb{N}$$

$$(i_k, j_k) \in P \Leftrightarrow y[j_k] \text{ is matched with } x[i_k]$$

DTW : choice of the optimal path

- The path P is evaluated through a cost function

$$w(P) = \sum_{k=1}^{K_P} (x[i_k] - y[j_k])^2$$

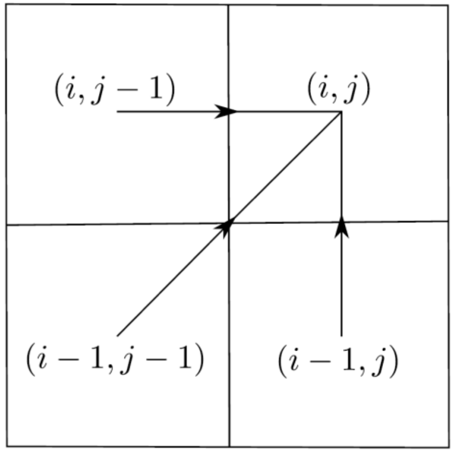
- The final DTW distance is computed as the minimal value for the cost function

$$\boxed{d_{DTW}(\mathbf{x}, \mathbf{y}) = \sqrt{\min_{P \in \mathcal{P}} w(P)}}$$

- Note that if $K_P = N$ and $i_k = j_k = k$, DTW is exactly equal to the Euclidean distance

DTW is not a proper distance : no triangle inequality, no separability

Set of acceptable paths and algorithm



Limited set of indexes to get to (i_k, j_k)

$$(i_{k-1}, j_{k-1}) = \begin{cases} & (i_k - 1, j_k) \\ \text{or} & (i_k, j_k - 1) \\ \text{or} & (i_k - 1, j_k - 1). \end{cases}$$

Algorithm 1: Dynamic Time Warping

Inputs : Time series $\mathbf{x} \in \mathbb{R}^M$ and $\mathbf{y} \in \mathbb{R}^N$
Output: DTW $d_{DTW}(\mathbf{x}, \mathbf{y})$
$\mathbf{D} = \mathbf{0}_{M \times N}$;
$\mathbf{C} = \mathbf{0}_{M \times N}$;
for $1 \leq i \leq M, 1 \leq j \leq N$ **do**
  $D(i, j) = (x[i] - y[j])^2$;
end
for $1 \leq i \leq M, 1 \leq j \leq N$ **do**
  $C(1, j) = \sum_{\ell}^{j} D(1, \ell)$;
  $C(i, 1) = \sum_{\ell}^{i} D(\ell, 1)$;
end
for $2 \leq i \leq M, 2 \leq j \leq N$ **do**
  $C(i, j) = D(i, j) + \min\{C(i-1, j-1), C(i-1, j), C(i, j-1)\}$;
end
$d_{DTW}(\mathbf{x}, \mathbf{y}) = \sqrt{C(M, N)}$;

Recursive formulation with dynamic programming

Construction of the centroids : S-DBA

S-DBA : Stochastic Subgradient DTW Barycenter Averaging

$$F_{\mathsf{X}'} : \mathbf{y} \in \mathbb{R}^l \mapsto \sum_{\mathbf{x}' \in \mathsf{X}'} \mathrm{DTW}_\alpha^2(\mathbf{y}, \mathbf{x}') \in \mathbb{R}$$

- L : median length of the signals

$$\nabla F_{\mathsf{X}'}(\mathbf{y}) = \frac{2}{M} \sum_{u=1}^{M} \left(V^{(u)} \mathbf{y} - W^{(u)} \mathbf{x}'^{(u)} \right)$$

- Subgradient at point y
- $W^{(u)}$: optimal warping matrix between
- $\boldsymbol{y}$ and $\boldsymbol{x}'^{(\boldsymbol{u})}$
- $V^{(u)}$: diagonal matrix such that

$$V_{i,i}^{(u)} = \sum_{j=1}^{n} W_{ij}^{(u)}$$

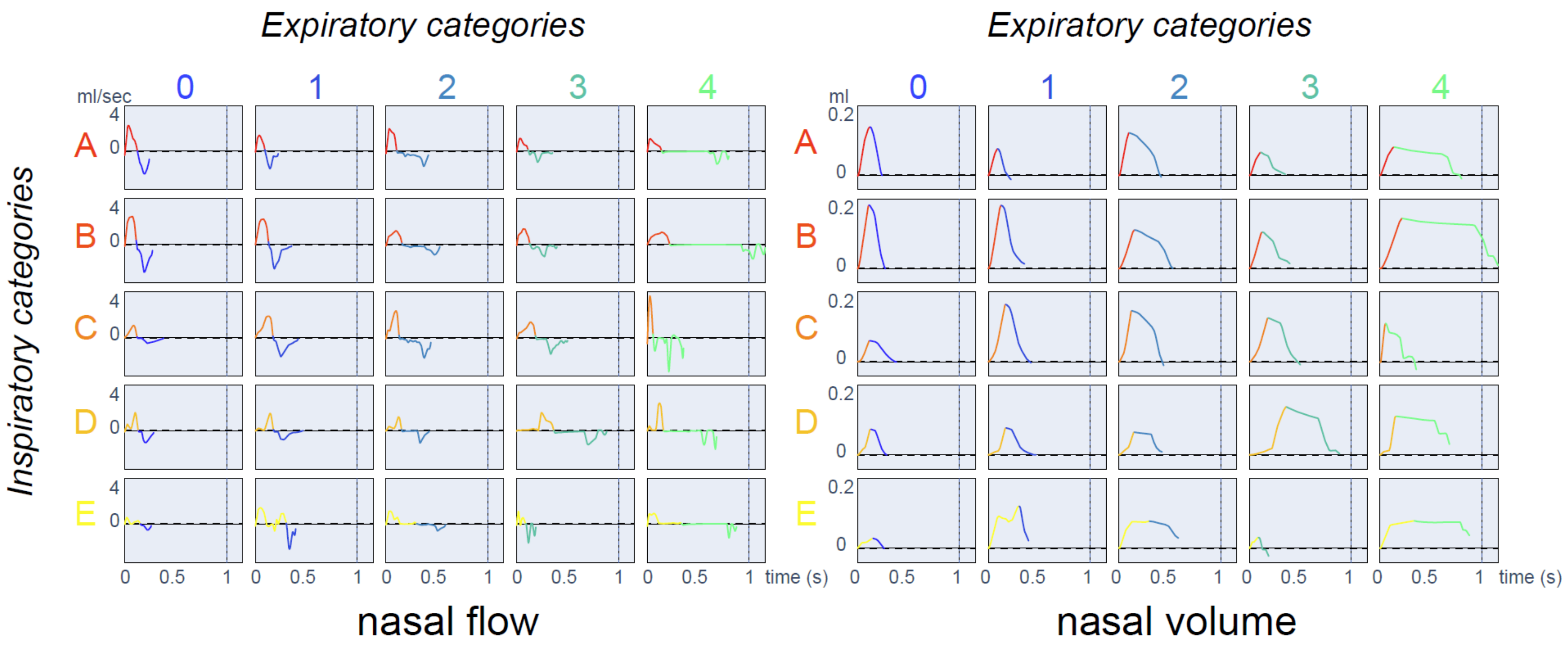
Iterative procedure :

1. Initialization of the centroid
2. Computation of the warping matrices
3. Gradient descent
4. Update of the centroid + repeat 2, 3, 4

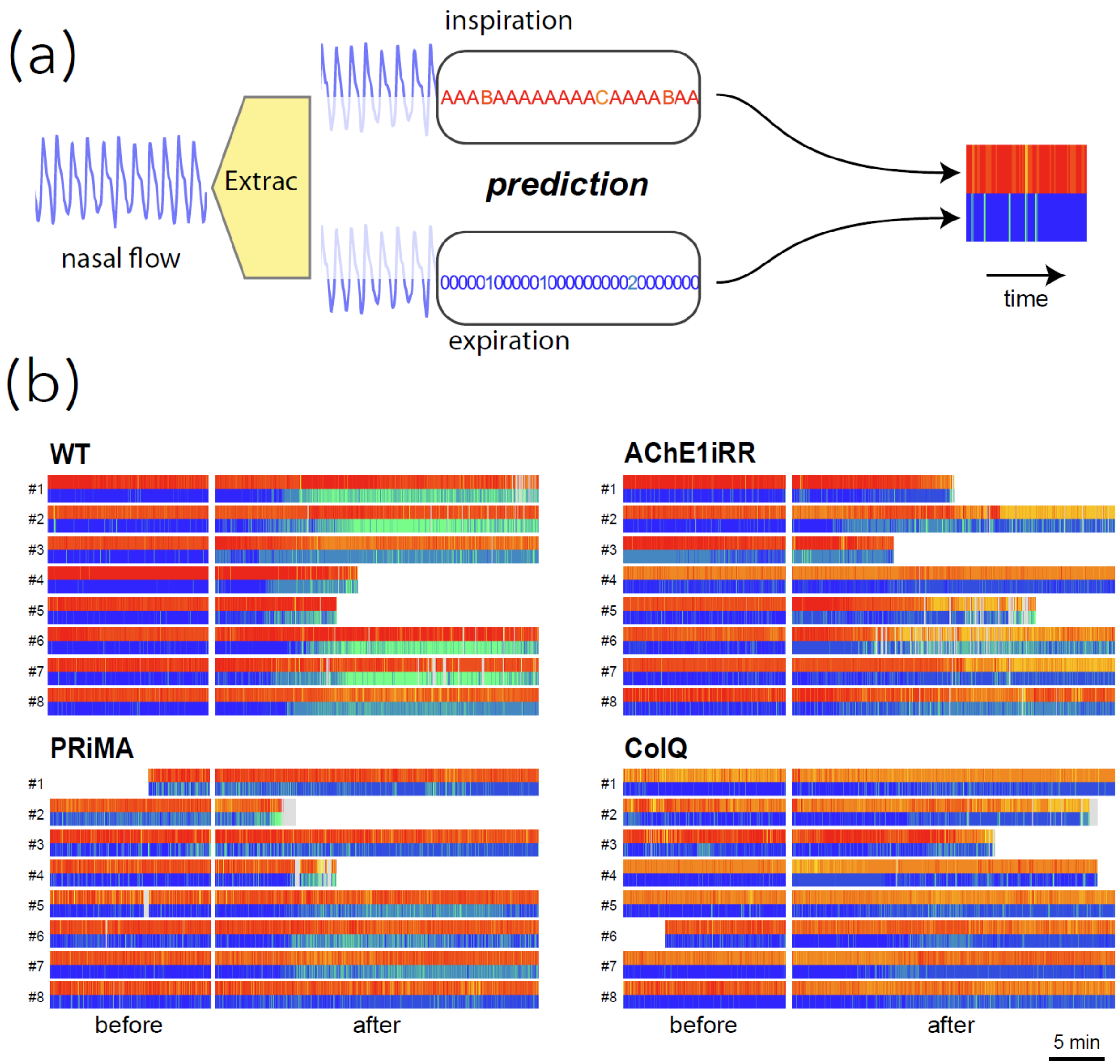
- Schultz, D., & Jain, B. (2018). Nonsmooth analysis and subgradient methods for averaging in dynamic time warping spaces. Pattern recognition, 74, 340-358.

Results

- 32 signals (250 Hz), 8 recordings per genotype (WT, PRiMA, AChE1iRR, ColQ)



Results

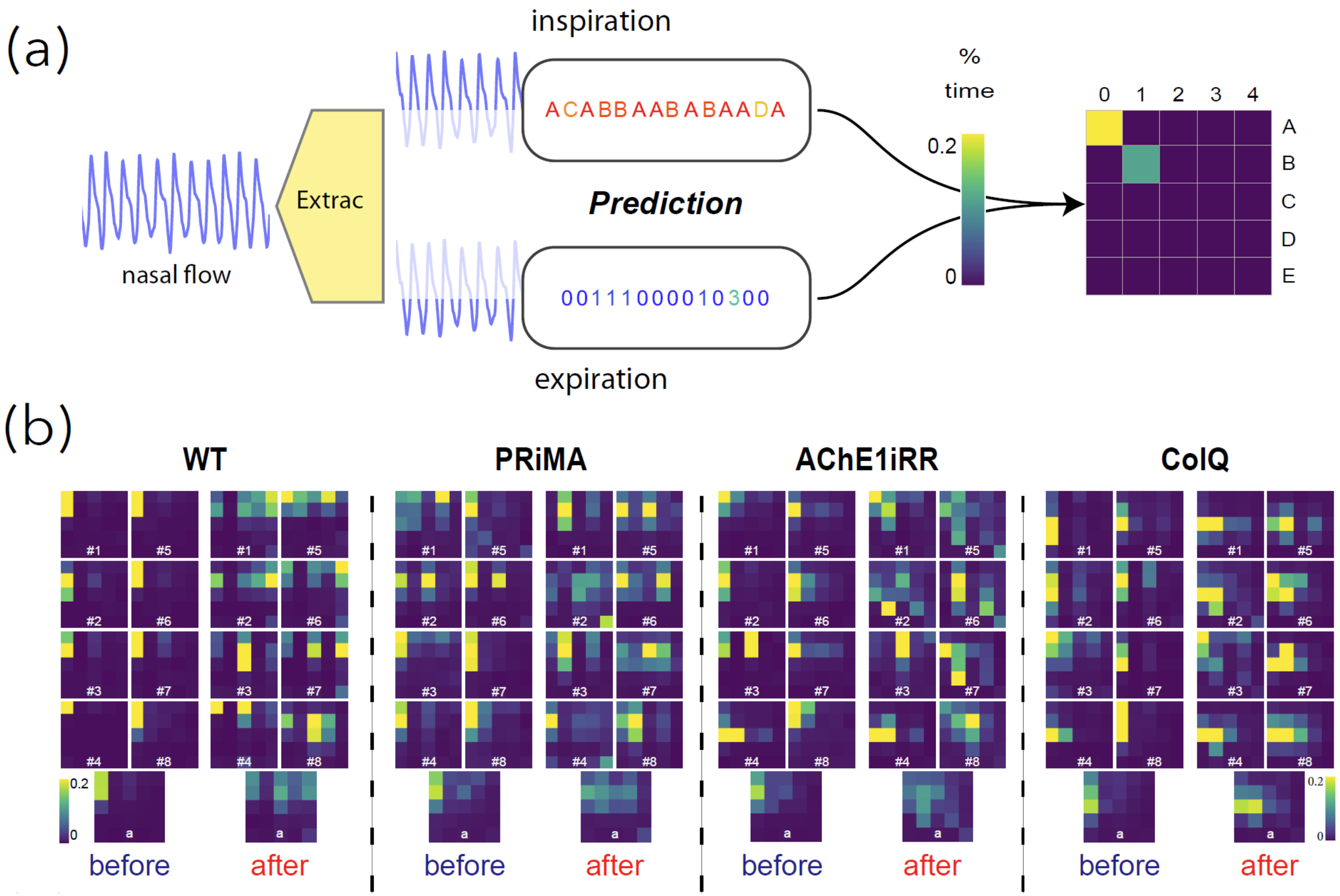


- WT : before injection, mostly A0 and B0. After injection : no changes in inspiration, but expiration use more and more classes 2, 3, 4. Post-inspiratory pauses

- ColQ mice : C0 and D0 before injection : inspiration in two phases. Relatively nchanged afer injection

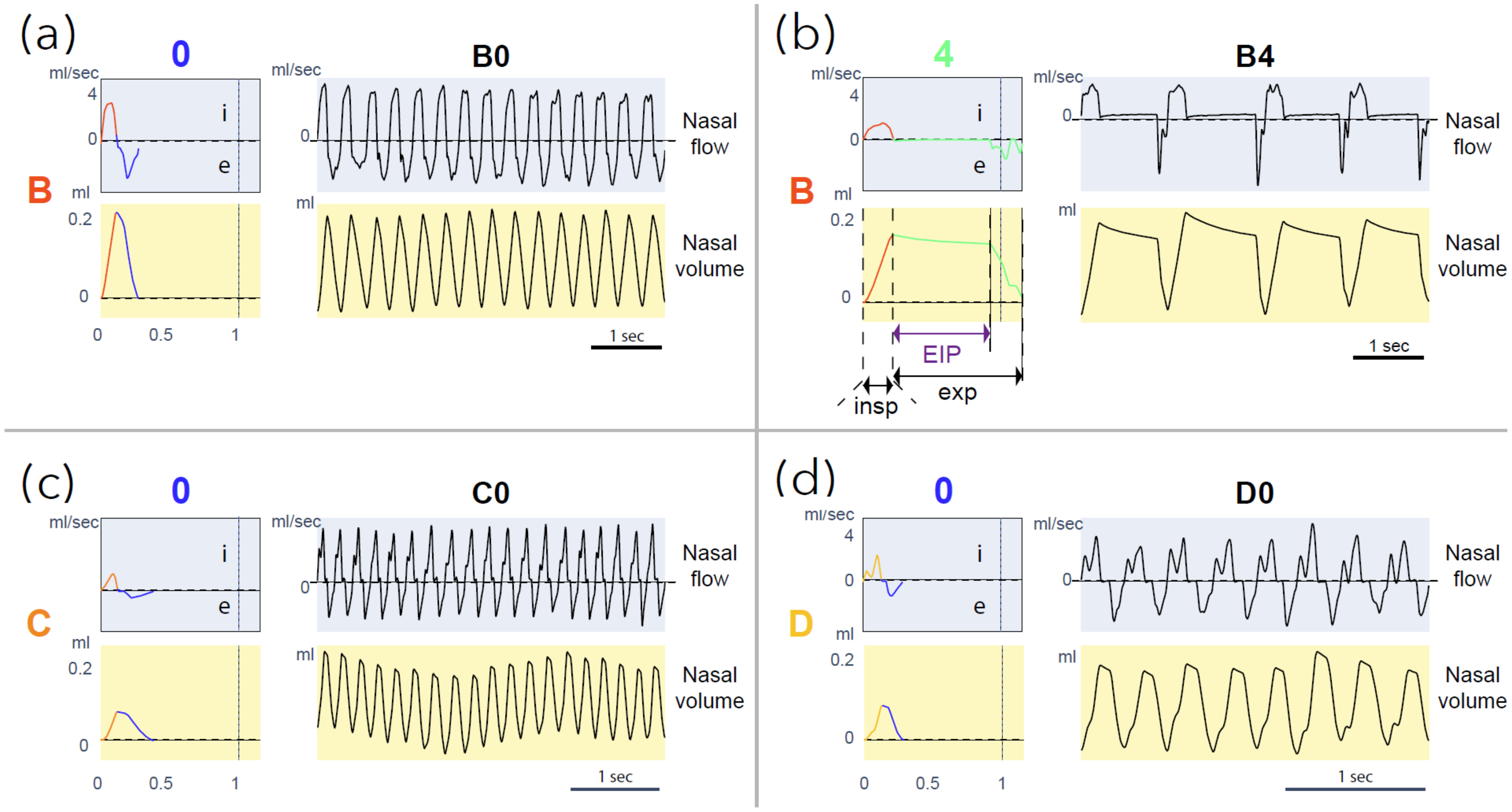
- Direct and intuitive visualisation of complex temporal data

Results



% of time assigned to a particular symbol (inspiration + expiration)

Results



Clusters/symbols can model complex respiratory behaviors

Perspectives and open questions

Current works

- Construction of distances for time series with templates : adaptation of $\boldsymbol{d_{sym}}$ with proper invariants
- Fixed protocol and « known » sequences of actions : creation of a barycenter or « ideal » symbolic sequence + distance to this « ideal »
- Biomarkers and outputs from ethograms for the study of animal/human behavior : frequent pattern mining, Markov models ?
- Hybrid/multimodal data : one sequence per modality (potentially some with the « regime »-based method and some with the « pattern »-based method) : how can we merge the sequences ? Rule association mining to understand the relationships between the modalities
- Semantic summary of complex signals : patterns, change-point, anomalys through event-based symbolization

Thank you !
Questions ?