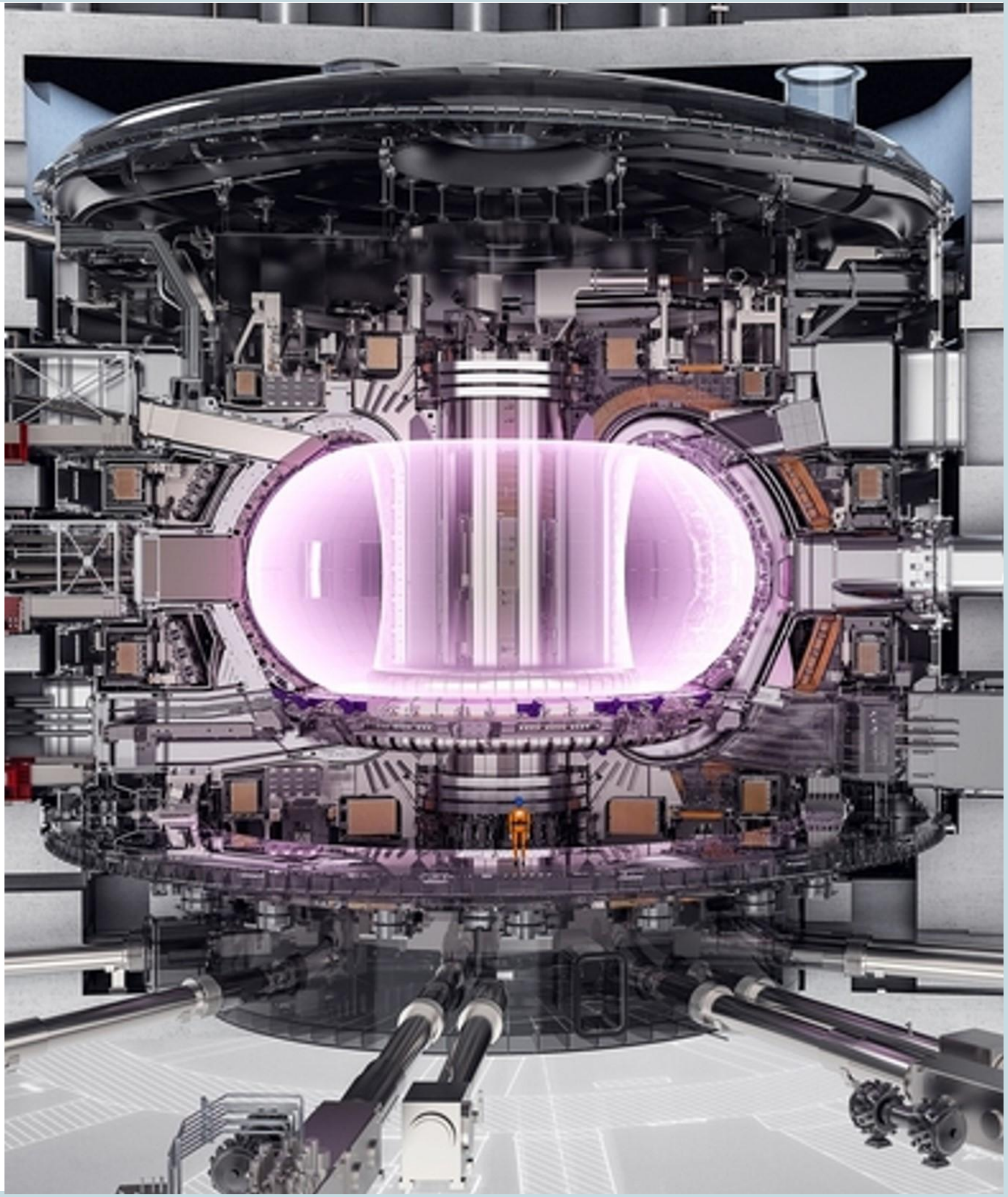
EFFICIENT PATTERN SEARCH IN LARGE-SCALE TIME SERIES

J. Hostalet[1], R. Castro[2], J. Vega[2], S. Dormido-Canto[1]

[1]Dpto. De Informática y Automática, UNED, Madrid, Spain
[2]Laboratorio Nacional de Fusión – CIEMAT, Madrid, Spain

Second Visualizing Offline and Live Data with AI (VOLDA) Workshop - 2025



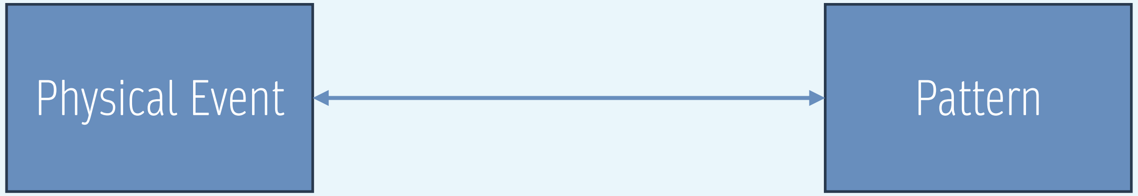
This work was partially funded by the Spanish Ministry of Science and Innovation under the Project Nos. PID2022-137680OB-C31 and PID2022-137680OB-C32

OVERVIEW

- CONTEXT AND MOTIVATION
- GOAL
- FIRST APPROACHES AND INITIAL TESTS
- VECTOR APPROACH
- IMPLEMENTATION
- RESULTS AND ANALYSIS
- CONCLUSIONS
- NEXT STEPS

CONTEXT AND MOTIVATION
PATTERNS IN FUSION

- Diagnostics produce similar signals for reproducible behaviour. Physical properties are transformed into patterns
- There is a direct correspondence between physical behaviour and the structural shape of signals. The concept of structural pattern is connected with the structural shape of signals



CONTEXT AND MOTIVATION STATISTICAL ANALYSIS

- Structural patterns are the entry point to data analysis
- Data analysis with statistical relevance requires to select a large number of similar structural shapes from different discharges
- This selection is a manual and tedious procedure in which the signals need to be examined individually
  - Visual data analysis
- Issues with visual data analysis
  - Manual searches
  - Massive databases
  - Typical scale of the pattern
  - Long pulse conditions

CONTEXT AND MOTIVATION
DATA RETRIEVAL MODELS

- Classical Model
  - INPUT: Shot number
  - OUTPUT: Samples to be visually analyzed
- New model
  - INPUT: Pattern
  - OUTPUT: Shot numbers and locations where those patterns occur

See: J. Vega et al. Fus. Eng. Des. 83 (2008) 132-139

CONTEXT AND MOTIVATION EFFICIENCY

- The searching process of patterns has to be efficient
- Efficiency means
  - not to traverse the entire database when searching for similar patterns
  - Developing intelligent mechanisms to reduce the searching space just to the most likely signals of containing a similar pattern
- Unsupervised classification systems are required
  - It allows the indexation of patterns in such a way that group similar objects into clusters
  - Each cluster represents a reduced searching space in which the more likely objects to be similar can be found

CONTEXT AND MOTIVATION ADDITIONAL NOTES

- Signals are grouped into collections for pattern oriented data retrieval
  - The searching process of similar patterns is carried out inside collections
- A signal collection is the complete set of recorded signals for all discharges of interest
  - Plasma current in a Tokamak between shots 1 – 68741
  - Soft x-ray chord emission at a specific plasma radius between shots 30114 - 56246
- Classification systems for pattern recognition are created from signal collections

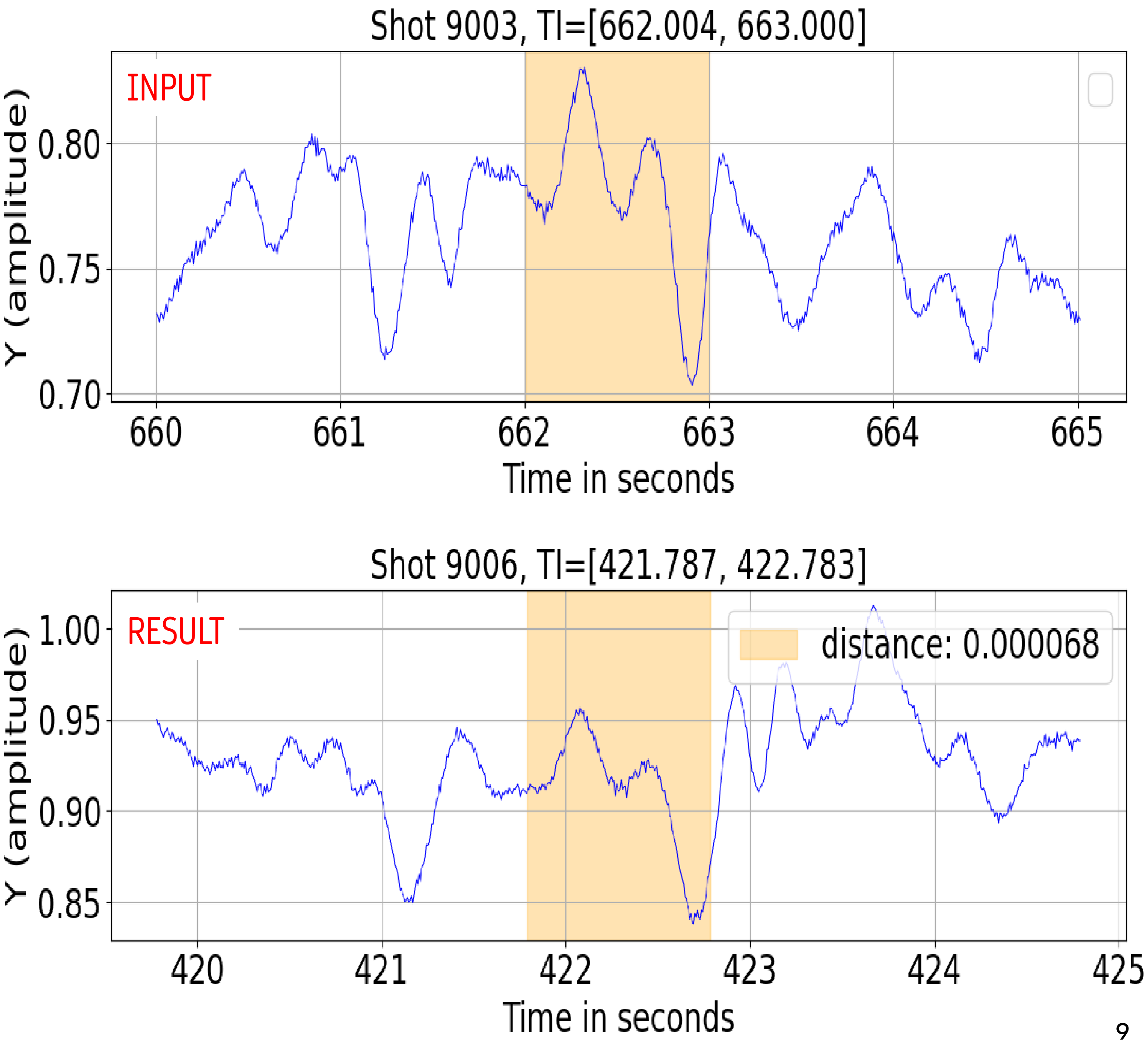
CONTEXT AND MOTIVATION

- Long-pulse fusion devices can produce data from thousands of sensors.
- New fusion devices will generate much more data than the previous ones.
- In ITER we expect a single discharge may last more than 30 minutes, typically recorded at millisecond or even higher resolution.
- The source data to perform data analysis will be much bigger than previous experiments.
- Within this enormous dataset, the goal is to **search quickly** for patterns (physical events) that are similar to a given one.
- Here, *similarity* refers to shape-based comparison between patterns.

GOAL

Use case: given a reference pattern, determine in which **discharges** and at what **times** similar patterns can be found.

Goal: enhance the tools available to analysts to improve the offline analysis process performed after a discharge.



FIRST APPROACHES

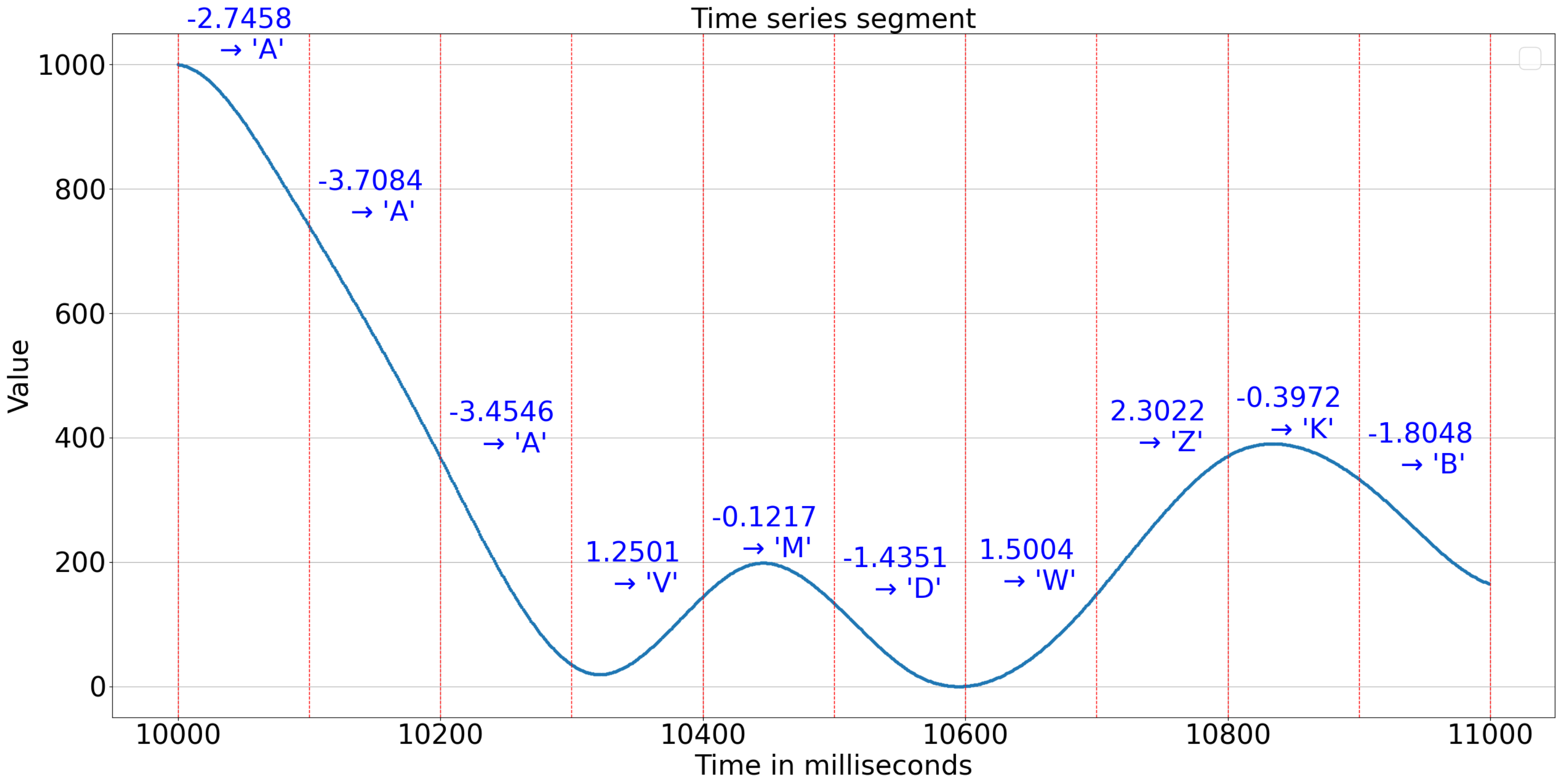
SAX (Symbolic Aggregate approXimation): converts a continuous-valued time series into a **sequence of symbols**, simplifying comparison.

TEST1: The sample is divided in 10 regions, for each region, slopes are mapped to letters.

TEST2: Straight lines are calculated between peaks, for each line, slopes are mapped to letters.

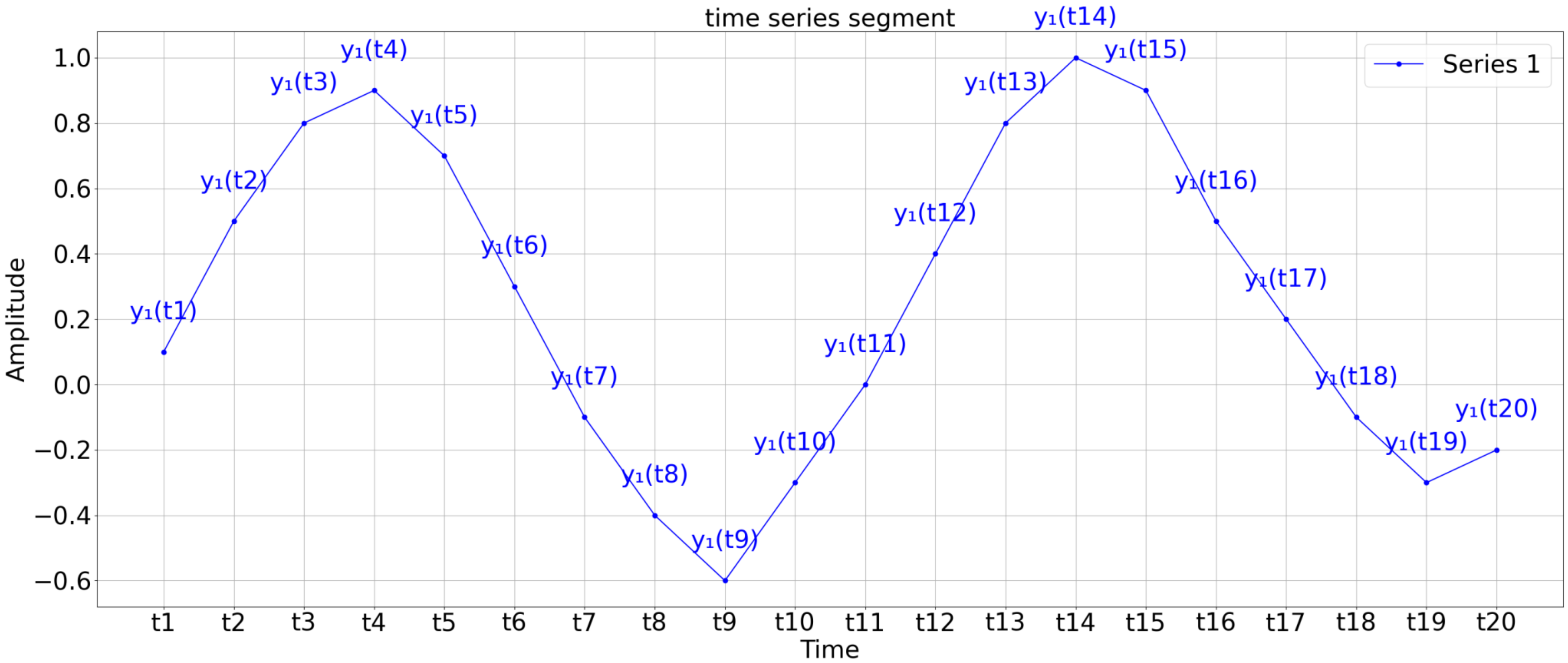
CONS: the method is highly sensitive to noise.

See: S. Dormido-Canto et al. Rev. Sci. Ins. [DOI: 10.1063/1.2219409]



VECTOR APPROACH (1/3)

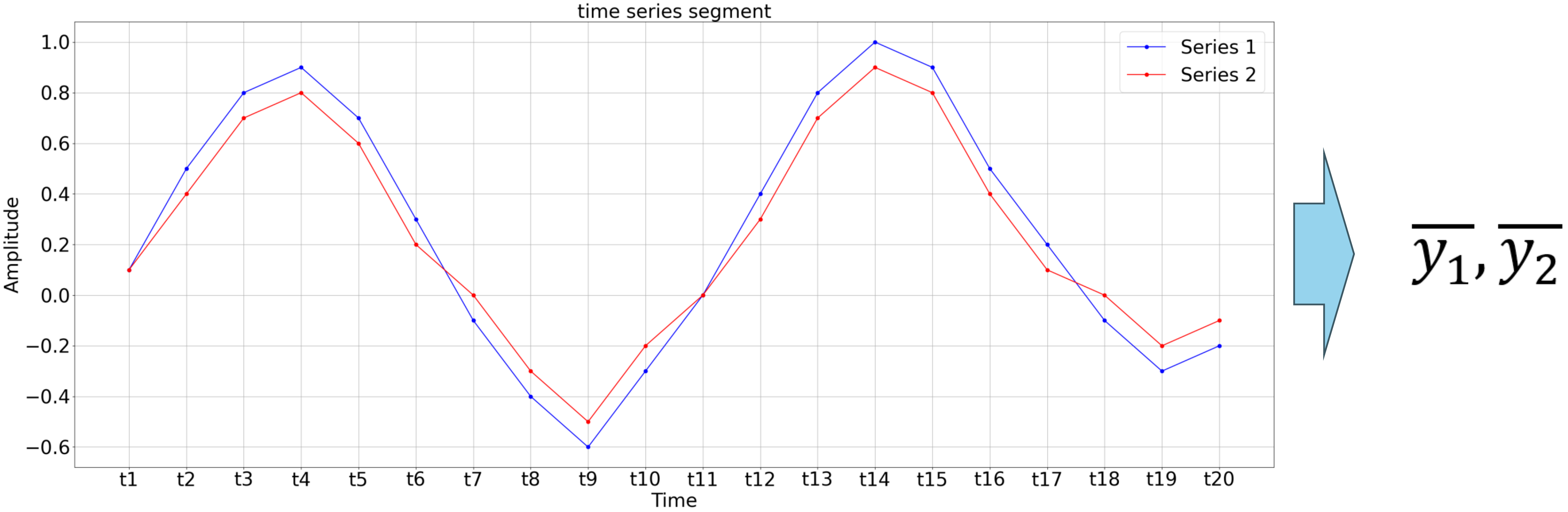
A vector can be constructed using the amplitudes of all measurements. This vector defines the position of a point in a 'D'-dimensional hyperplane. 'D' is the dimension of the vector (the number of samples in a time interval).



$$\bar{y} = \begin{bmatrix} y_{t1} \\ y_{t2} \\ y_{t3} \\ y_{t4} \\ y_{t5} \\ \dots \\ y_{tD} \end{bmatrix}$$

VECTOR APPROACH (2/3)

Example of two vectors pretty similar but not identical.



VECTOR APPROACH (3/3)

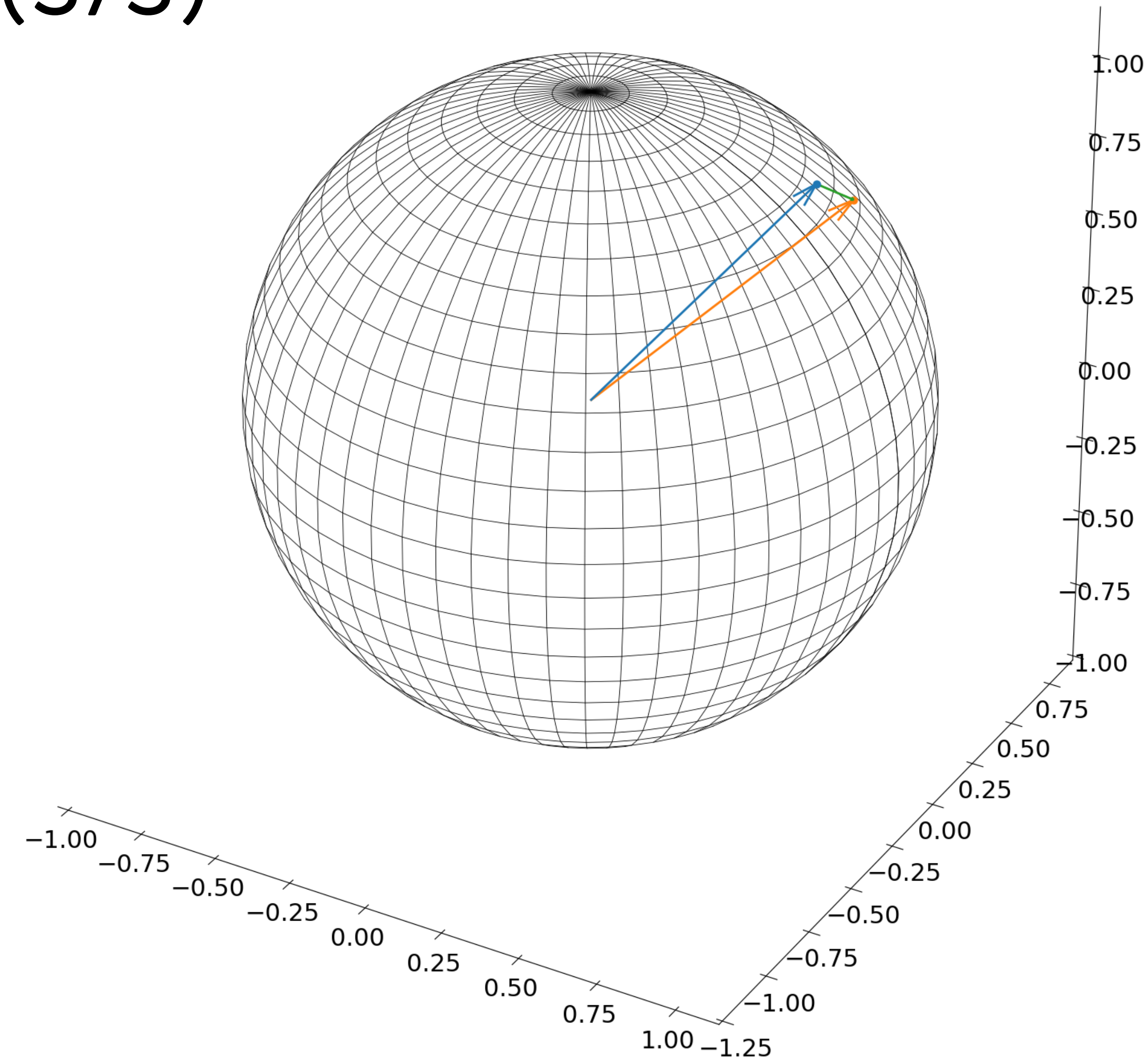
The **cosine distance** measures how different the *directions* of two vectors are, regardless of their magnitude.

If d = 0 vectors are identical

If d = 1 vectors are orthogonal

$$d(\overline{y}_1, \overline{y}_2) = 1 - \frac{\overline{y}_1 \cdot \overline{y}_2}{\|\overline{y}_1\| \, \|\overline{y}_2\|}$$

$$0 \le d(\overline{y}_1, \overline{y}_2) \le 1$$



IMPLEMENTATION

IMPLEMENTATION

RAW data → **Dimension reduction** → Vector extraction → Database creation → Similarity search

A **Haar wavelet filter** is applied to reduce the problem dimensionality (number of data points).

This filter can significantly decrease the data size while preserving the overall shape of the pattern.

Only the **approximation coefficients** are kept, which represent the *low-frequency* components of the signal capturing the general trend or shape.

Time Series	Number of samples	Reduction
Original	1,800,000	--
Filtered at level 2	524,288	71%
Filtered at level 3	262,144	85%
Filtered at level 4	131,072	93%
Filtered at level 5	65,536	96%

$$W' = \frac{W}{2^L}$$

IMPLEMENTATION

RAW data → Dimension reduction → Vector extraction → Database creation → Similarity search

Question: given a time series segment with W samples and a required minimum dimension D_{min}, which is the optimal Haar level L to apply ?

$$Given \begin{cases} W' = \frac{W}{2^L} \\ W' \geq D_{min} \end{cases}$$

$$\frac{W}{2^L} \geq D_{min} \quad \rightarrow \quad \boxed{L = \left\lfloor Log_2\left(\frac{W}{D_{min}}\right)\right\rfloor}$$

IMPLEMENTATION

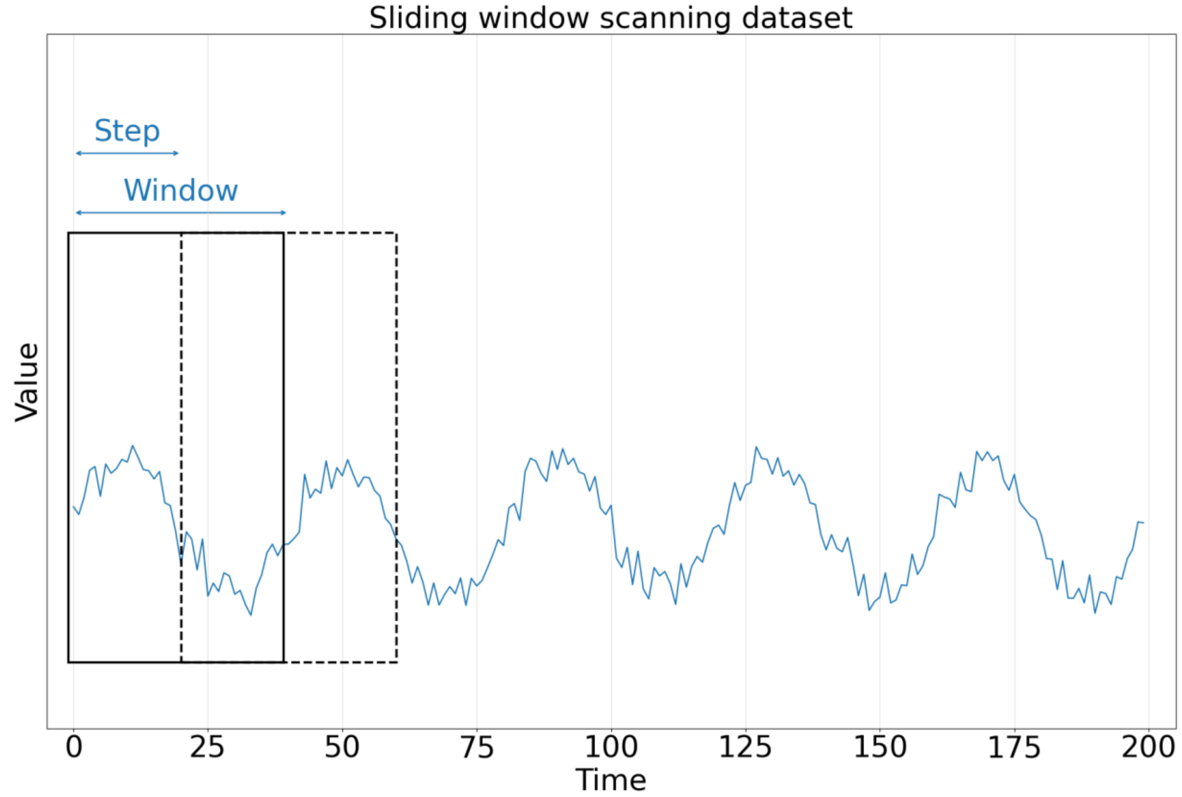
RAW data → Dimension reduction → **Vector extraction** → Database creation → Similarity search

A **sliding-window technique** is applied to the time series. The window has a fixed size, equal to the vector dimension, and moves forward one sample at a time (step) until the entire series is scanned.

Each vector is **normalized to unit length.**

Window size is fixed to one of the following values (milliseconds): 500, 1000, 2000, 4000

System is window size specific.



IMPLEMENTATION

RAW data → Dimension reduction → Vector extraction → **Database creation** → Similarity search

- **FAISS** (Facebook AI Similarity Search): an **open-source** library developed by Meta (formerly Facebook) for performing efficient **similarity search** and clustering **on large collections of high-dimensional vectors.** (see: https://github.com/facebookresearch/faiss)
- **FAISS** is designed to solve this problem efficiently: "Given a query vector, find the closest vectors (neighbors) in a large dataset"
- Optimized in C++ with Python bindings. Also supports **GPU** acceleration.
- Objectives: **Maximum accuracy with the highest possible speed**

FAISS MODES

Index Type	Memory use	Speed	Accuracy	Notes
Flat	High	Slow	100%	Exact search
IVF	Medium	Fast	~95%	Cluster based
PQ	Low	Very fast	~90%	Compressed
HNSW	Medium	Very fast	~98%	Graph-based

The Flat index in FAISS stores all vectors in memory without any compression or clustering.

It performs an exhaustive (brute-force) search, but it’s heavily optimized at a low level in C++ and can use multi-threading and GPU acceleration.

It can compute many dot products simultaneously

IMPLEMENTATION

RAW data → Dimension reduction → Vector extraction → Database creation → **Similarity search**

Test were executed with 690 synthetic signals (1,800,000 samples each) generated to simulate real signals.

The result of the similarity search is a subset of patterns whose distance is lower than a given maximum distance.

$$d(\bar{y}_1, \bar{y}_2) = 1 - \frac{\bar{y}_1 \cdot \bar{y}_2}{\|\bar{y}_1\| \, \|\bar{y}_2\|}$$

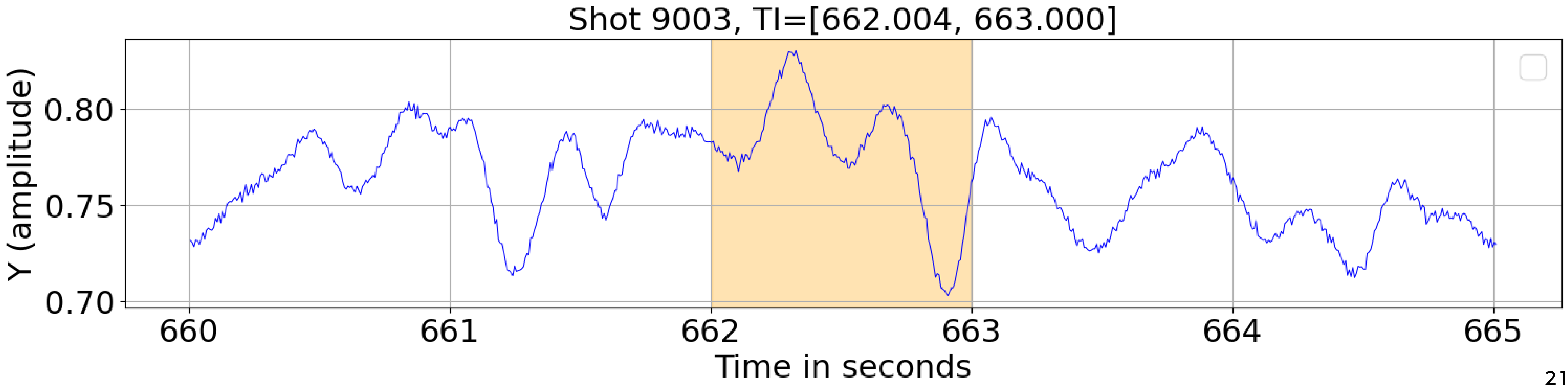
$$0 \leq d(\bar{y}_1, \bar{y}_2) \leq 1$$

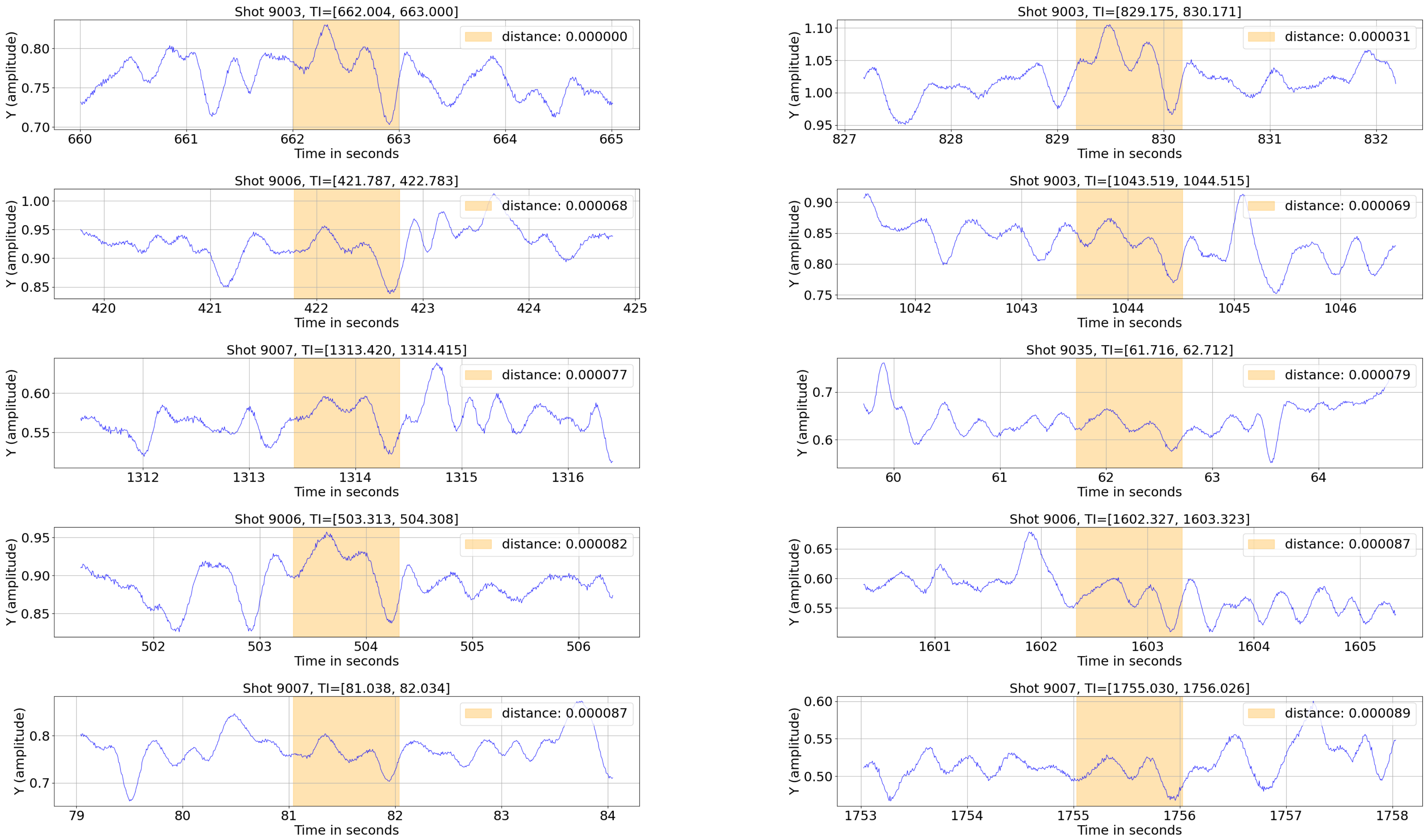
$$S = \left\{ \bar{y}_i \mid d(\bar{y}_q, \bar{y}_i) < D_{max} \right\}$$

SIMILARITY SEARCH EXAMPLE

RAW data → Dimension reduction → Vector extraction → Database creation → Similarity search

The pattern segment at time interval [662.0;663.0] is searched for similarities




Shot 9003, TI=[662.004, 663.000]
distance: 0.000000
Shot 9003, TI=[829.175, 830.171]
distance: 0.000031
Shot 9006, TI=[421.787, 422.783]
distance: 0.000068
Shot 9003, TI=[1043.519, 1044.515]
distance: 0.000069
Shot 9007, TI=[1313.420, 1314.415]
distance: 0.000077
Shot 9035, TI=[61.716, 62.712]
distance: 0.000079
Shot 9006, TI=[503.313, 504.308]
distance: 0.000082
Shot 9006, TI=[1602.327, 1603.323]
distance: 0.000087
Shot 9007, TI=[81.038, 82.034]
distance: 0.000087
Shot 9007, TI=[1755.030, 1756.026]
distance: 0.000089
Y (amplitude)
Time in seconds

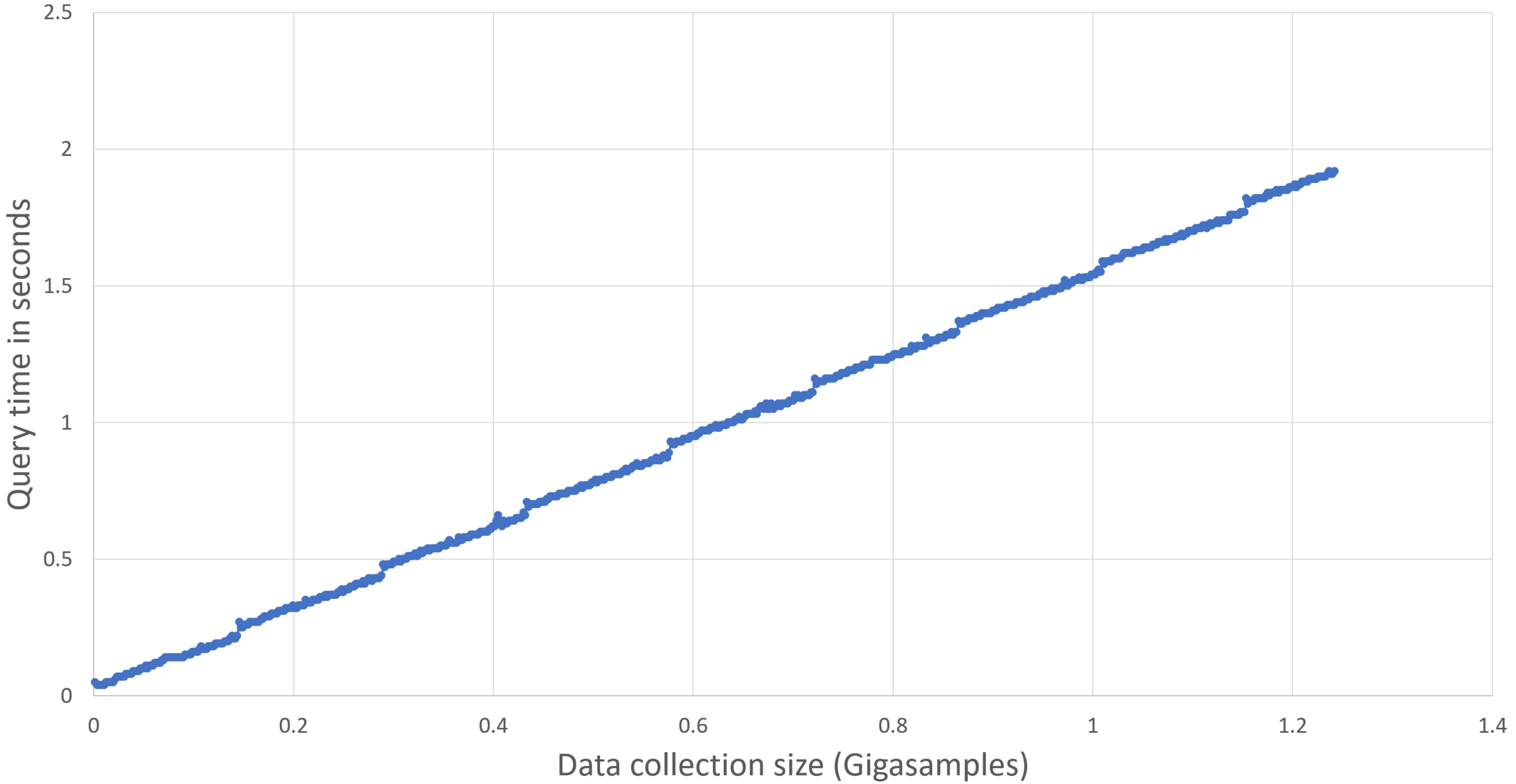
RESULTS AND ANALYSIS

Query time (seconds) vs Data samples.

Test executed with 690 synthetic signals generated.

Each signal contains 1.8M samples before filtering and 262K after filtering with the Haar Wavelet transform (level 3).

Level 3 Haar reduced data size by 85%.



CONCLUSIONS

- SAX-based methods are highly sensitive to small variations, making them less suitable for noisy signals.
- The vector-based approach performs very well when waveform similarity is the main criterion.
- Haar wavelet filtering significantly reduces the number of points while preserving the overall waveform shape.
- Using the cosine distance provides a solid framework for measuring similarity.
- We developed a method to select the optimal Haar level, considering the minimum required dimension.
- FAISS is an excellent tool for efficient similarity searches.
- Initial tests have shown very fast response times.

NEXT STEPS

- **Scalability and Distributed Search**: Implement sharding to handle datasets beyond a single machine's memory; Experiment with multi-GPU and multi-node FAISS configurations.
- **Real Data Validation**: Apply the system to actual fusion discharges.
- **User Interface and Integration**: Develop a visual analysis tool to display search results directly on waveform plots. Develop interfaces and libraries to allow further integration with existing tools.
- **Tests other FAISS modes other than Flat**
- **Hybrid and Hierarchical Indexing**: Investigate multi-phase similarity searches (reduce the target shards to be scanned)

THANK YOU

J. Hostalet (Dpto. Informática y Automática, UNED, Madrid, Spain)
R. Castro (Laboratorio Nacional de Fusión, Madrid, Spain)
J. Vega (Laboratorio Nacional de Fusión, Madrid, Spain)
S. Dormido-Canto (Dpto. Informática y Automática, UNED, Madrid, Spain)

UNED

GOBIERNO DE ESPAÑA

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

Ciemat
Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

This work was partially funded by the Spanish Ministry of Science and Innovation under the Project Nos. PID2022-137680OB-C31 and PID2022-137680OB-C32