DATA MINING 2

Time Series – Matrix Profile, Motifs & Discords

Riccardo Guidotti

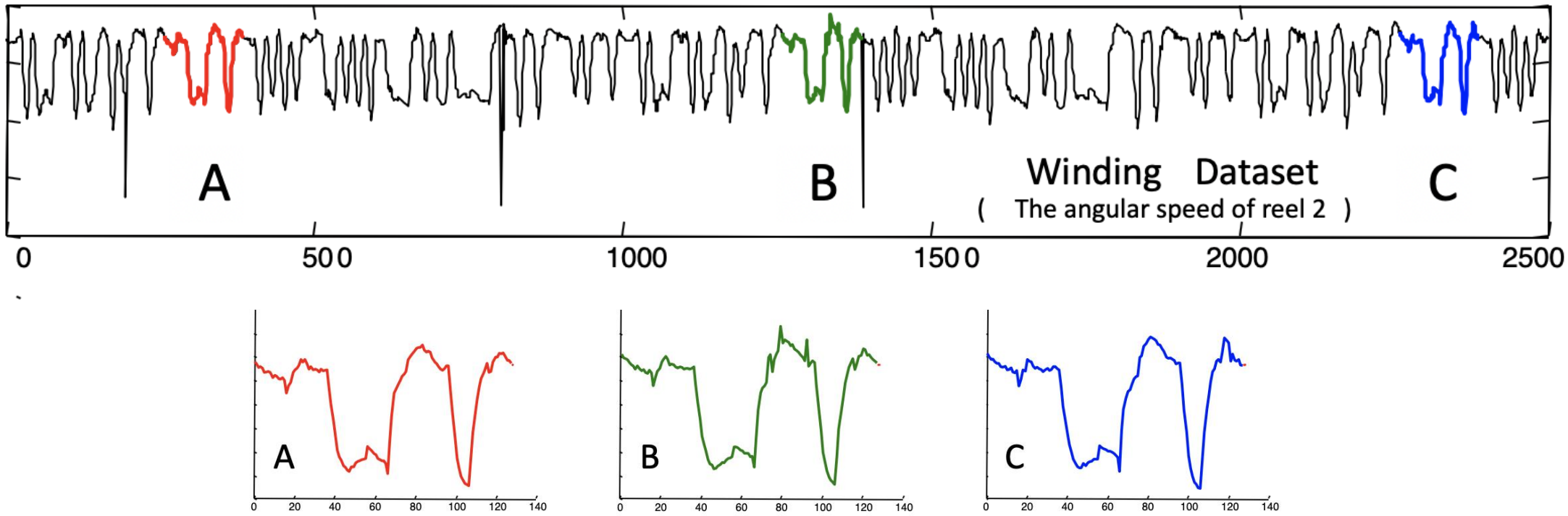
a.a. 2022/2023

Slides edited from Keogh Eamonn's tutorial

UNIVERSITÀ DI PISA

Time Series Motif Discovery

- Finding repeated patterns, i.e., pattern mining.
- Are there any repeated patterns, of length *m* in the TS?



Why Finding Motifs?

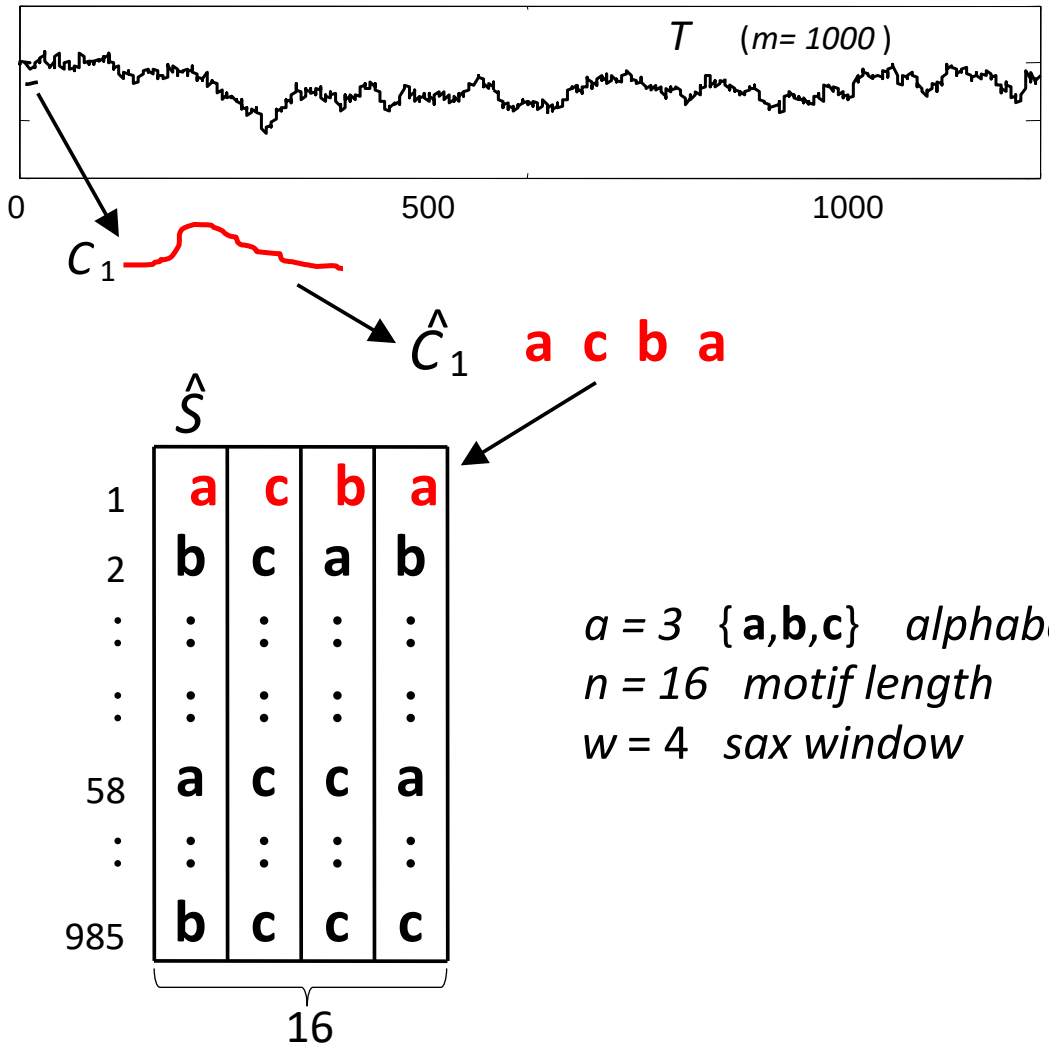
- Mining **association rules** in TS requires the discovery of motifs. These are referred to as primitive shapes and frequent patterns.
- Several **TS classifiers** work by constructing typical prototypes of each class. These prototypes may be considered motifs.
- Many **TS anomaly detection** algorithms consist of modeling normal behavior with a set of typical shapes (which we see as motifs) and detecting future patterns that are dissimilar to all typical shapes.

How do we find Motifs?

- Given a predefined motif length *m*, a brute-force method searches for motifs from all possible comparisons of subsequences.
- It is obviously very slow and computationally expensive.
- The most referenced algorithm is based on a hot idea from bioinformatics, random projection* and the fact that SAX allows use to lower bound discrete representations of TSs.
- J Buhler and M Tompa. Finding motifs using random projections. In RECOMB'01. 2001.

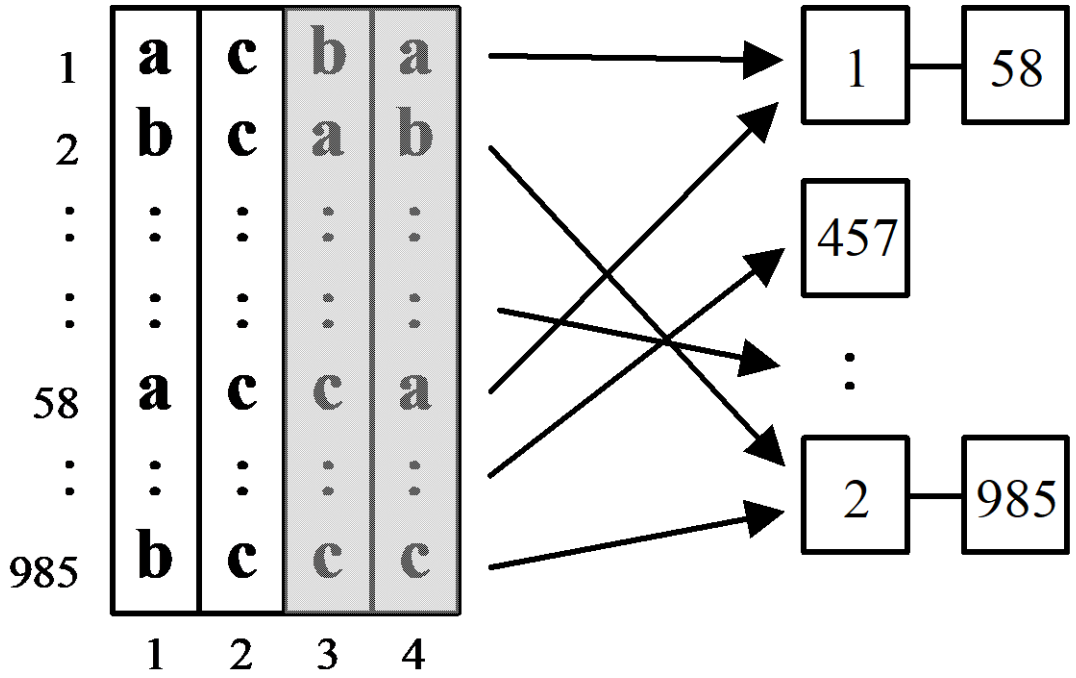
Motif Discovery with Random Projections

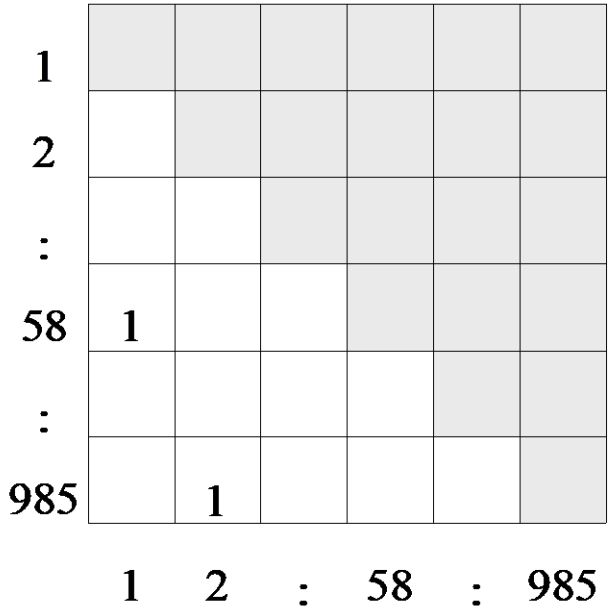
- Assume that we have a time series T of length 1,000, and a motif of length 16, which occurs twice, at time T_1 and time T_{58}.



Motif Discovery with Random Projections

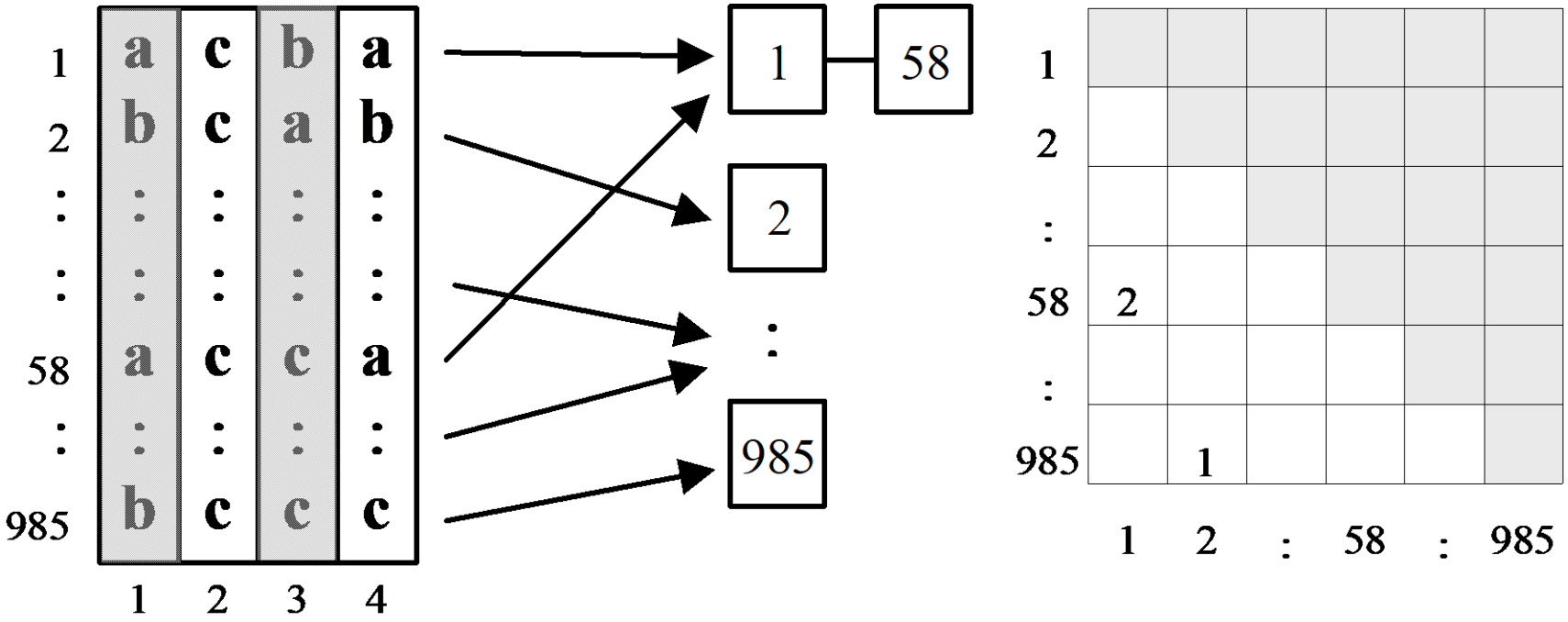
- A mask {1,2} was randomly chosen, so the values in columns {1,2} were used to project matrix into buckets.
- Collisions are recorded by incrementing the appropriate location in the collision matrix.





Motif Discovery with Random Projections

- A mask {2,4} was randomly chosen, so the values in columns {2,4} were used to project matrix into buckets.
- Once again, collisions are recorded by incrementing the appropriate location in the collision matrix.



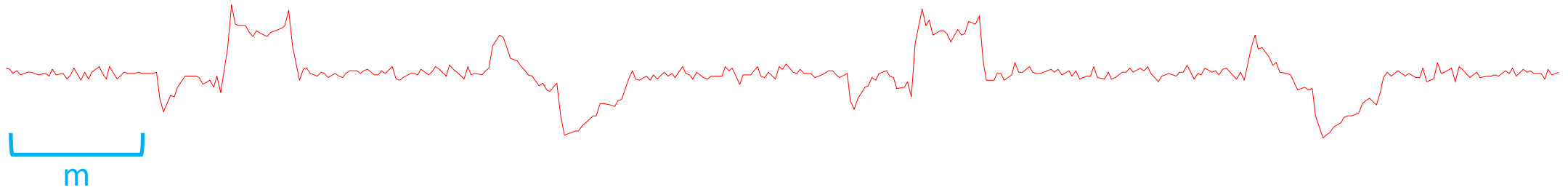
Motif Discovery with Random Projections

- At the end of the random perturbations consider the motifs observing the matrix in decreasing order of occurrences.
- For instance, this matrix indicates a high chance of having a motif staring at positions 1 and 58.
- The problem with this approach is that it is highly dependent from the approximation technique adopted.

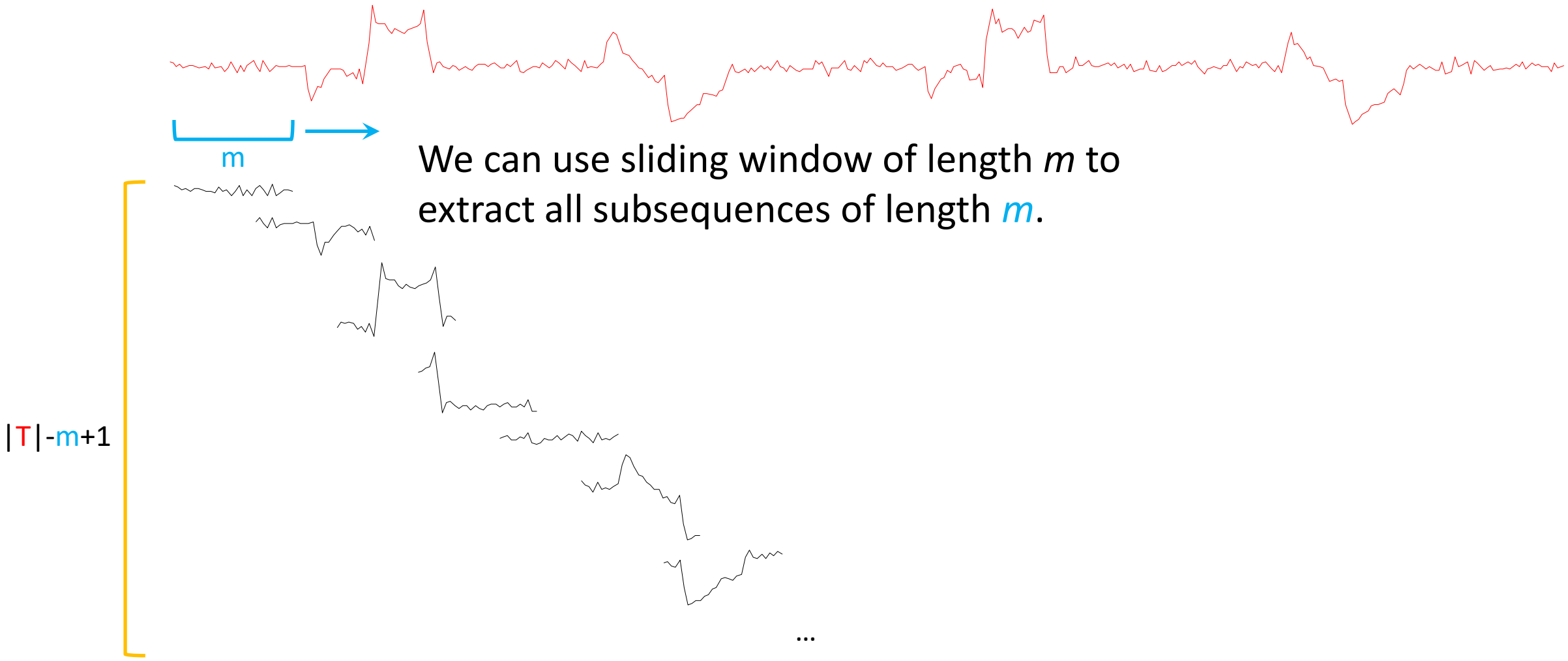
1						
2	2					
:	1	3				
58	**27**	2	1			
:	3	2	2	1		
985	0	1	2	1	3	
	1	2	:	58	:	985

Matrix Profile

- The Matrix Profile (MP) is a data structure that annotates a TS and can be exploited for many purposed: e.g. efficient Motif Discovery.
- Given a time series, T and a desired subsequence length, m.

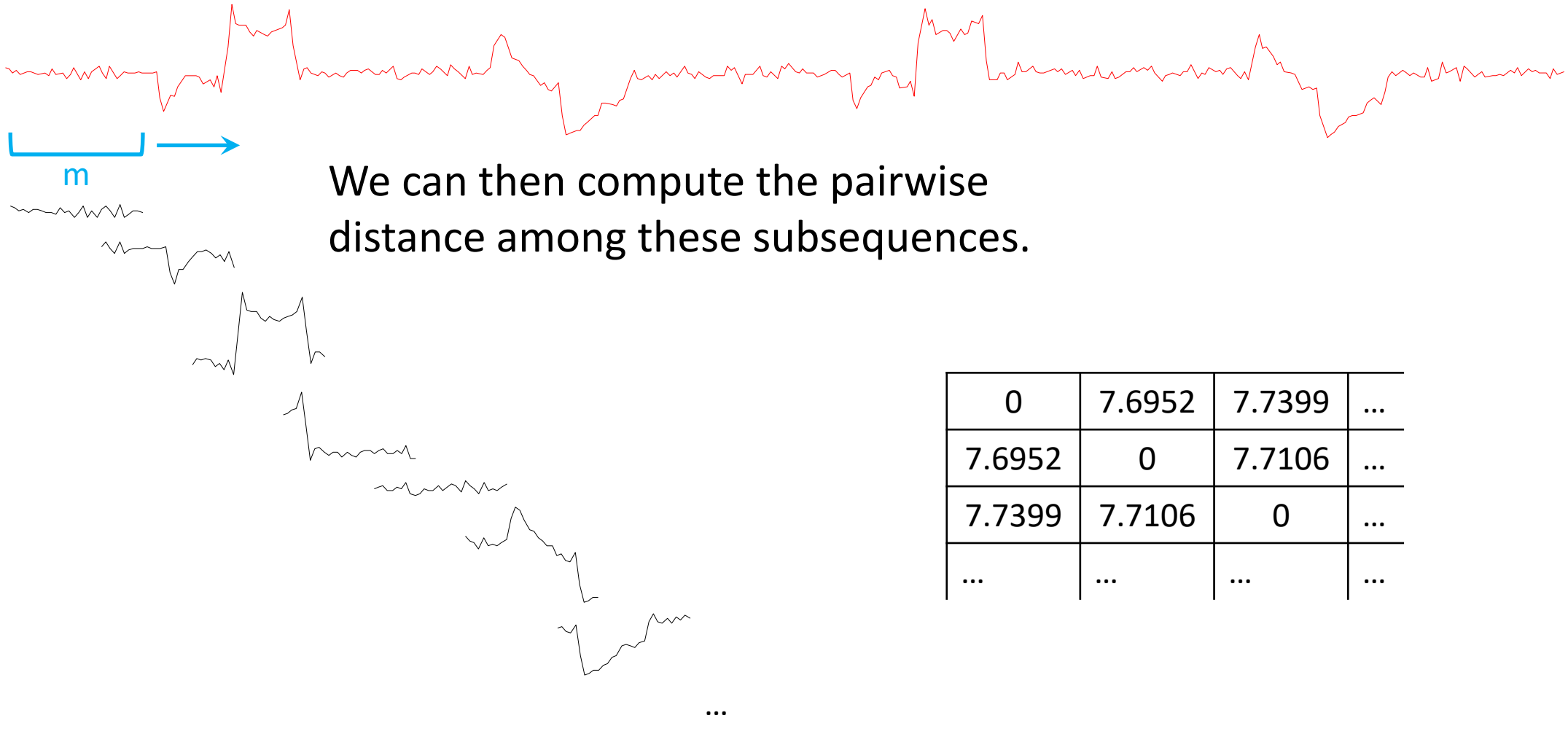


Matrix Profile



We can use sliding window of length *m* to extract all subsequences of length *m*.

Matrix Profile

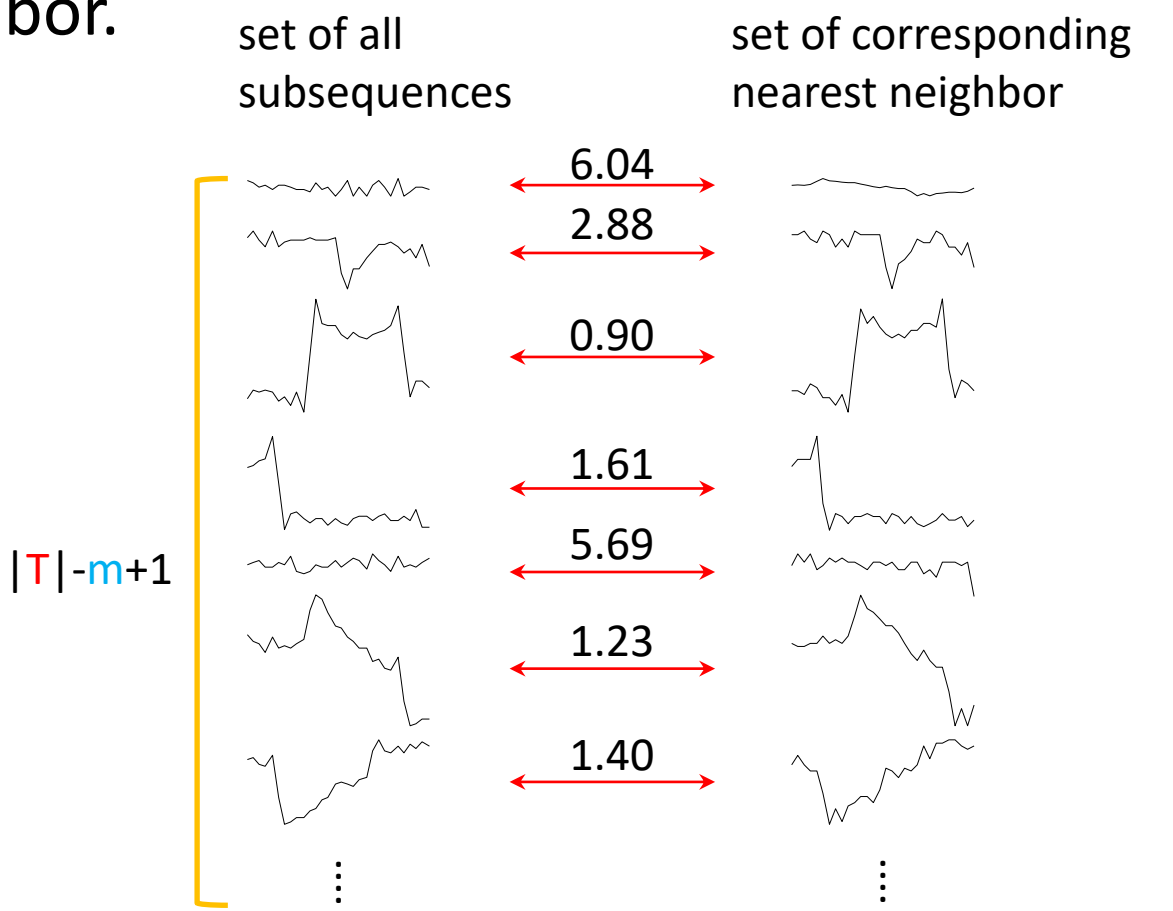


We can then compute the pairwise distance among these subsequences.

0	7.6952	7.7399	…
7.6952	0	7.7106	…
7.7399	7.7106	0	…
…	…	…	…

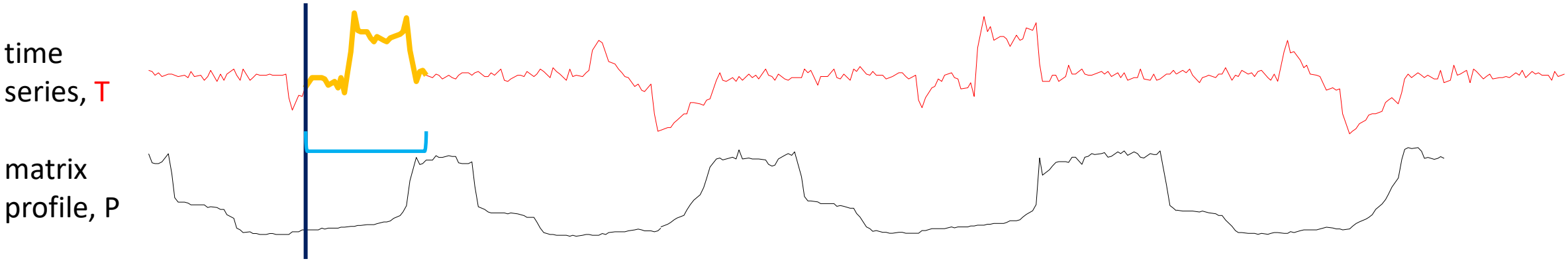
Matrix Profile

- For each subsequence we keep only the distance with the closest nearest neighbor.



Matrix Profile

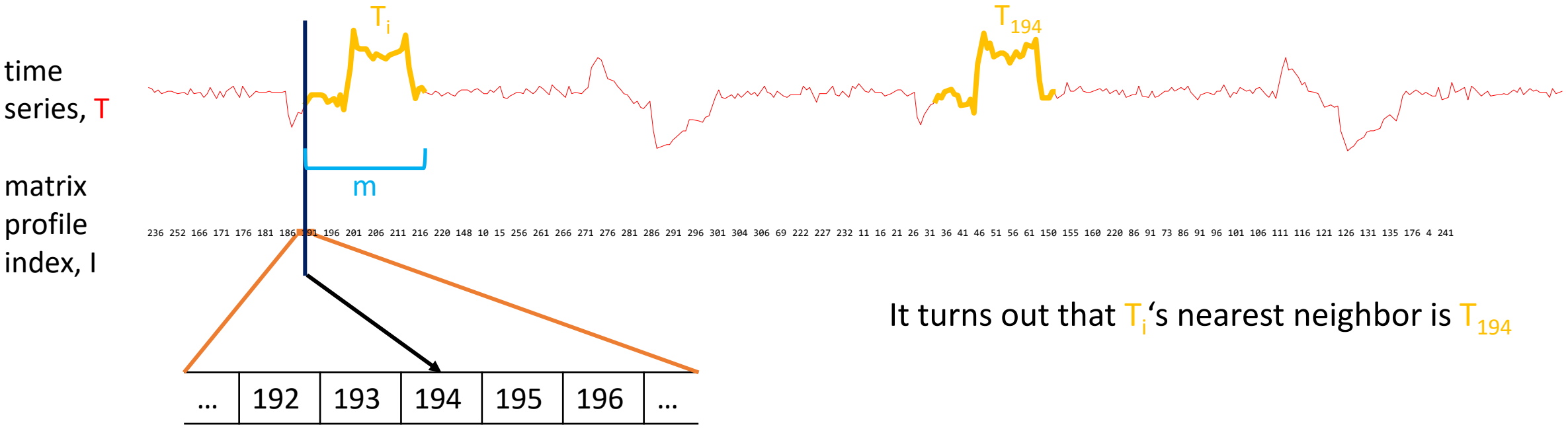
- The distance to the corresponding nearest neighbor of each subsequence can be stored in a vector called **matrix profile P**.



The matrix profile value at location *i* is the distance between T_i and its nearest neighbor

Matrix Profile

- The index of corresponding nearest neighbor of each subsequence is also stored in a vector called matrix profile index.

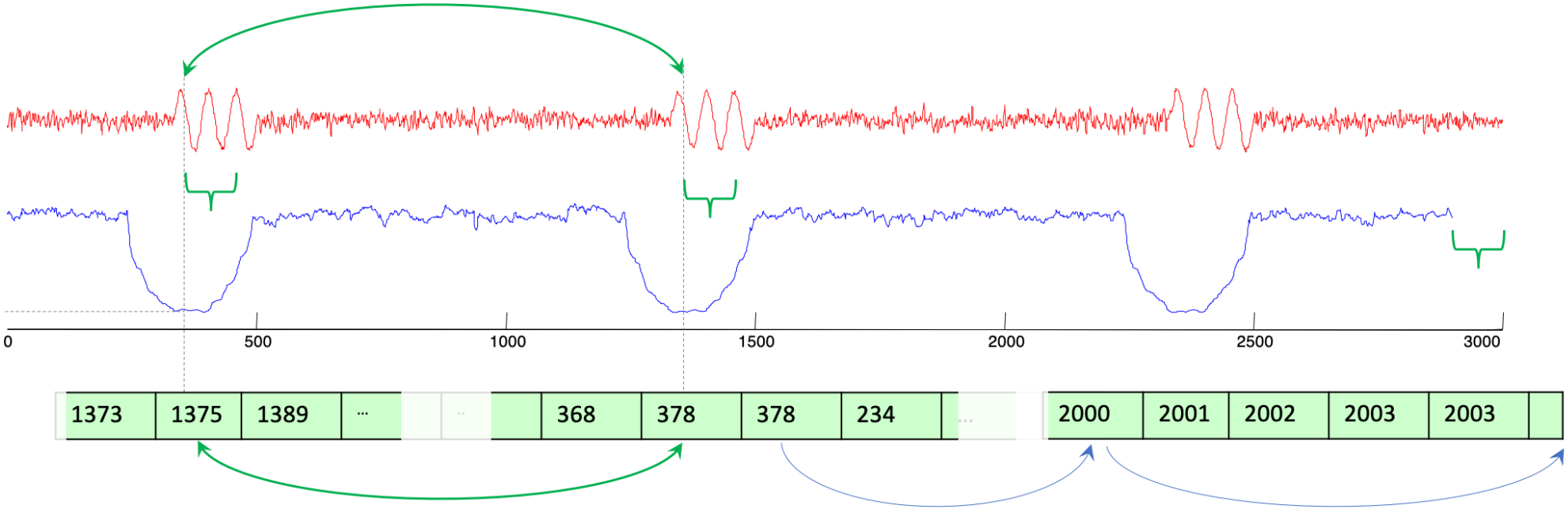


It turns out that T_i's nearest neighbor is T_{194}

The matrix profile value at location *i* is the distance between T_i and its nearest neighbor

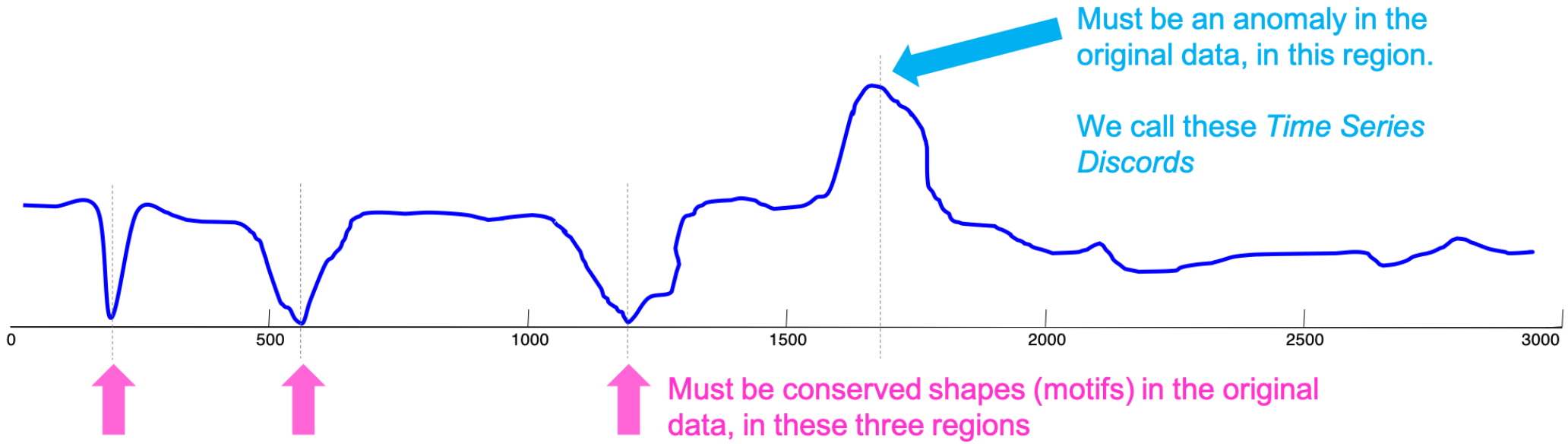
Matrix Profile

- The MP index allows to find the nearest neighbor to any subsequence in constant time.
- Note that the pointers in the matrix profile index are not necessarily symmetric.
- If A points to B, then B may or may not point to A
- The classic TS motif: the two smallest values in the MP must have the same value, and their pointers must be mutual.



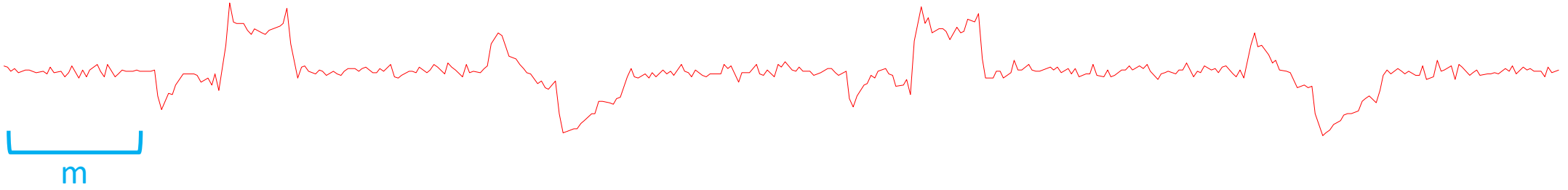
How to “read” a Matrix Profile

- For relatively low values, you know that the subsequence in the original TS must have (at least one) relatively similar subsequence elsewhere in the data (such regions are “motifs”)
- For relatively high values, you know that the subsequence in the original TS must be unique in its shape (such areas are anomalies).



How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.



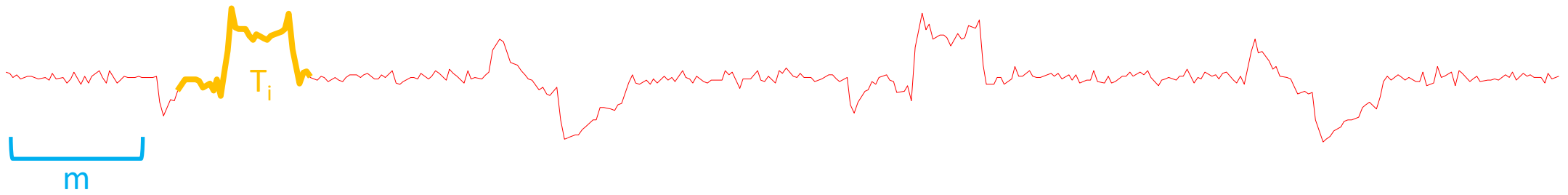
inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf

Matrix profile is initialized as inf vector

This is just a toy example, so the values and the vector length does not fit the time series shown above

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.

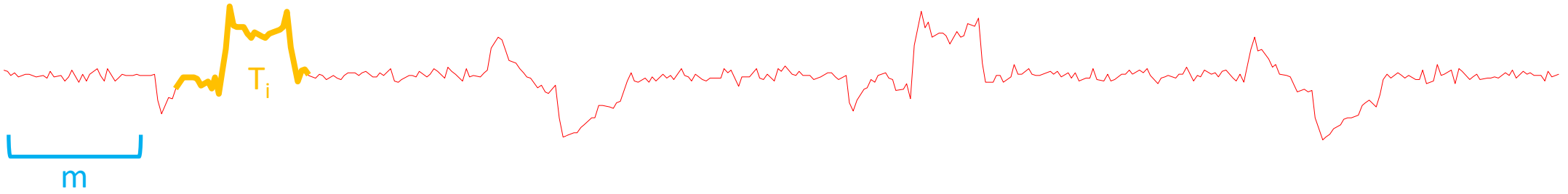


inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf

At the first iteration, a subsequence T_i is randomly selected from T

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.



inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf

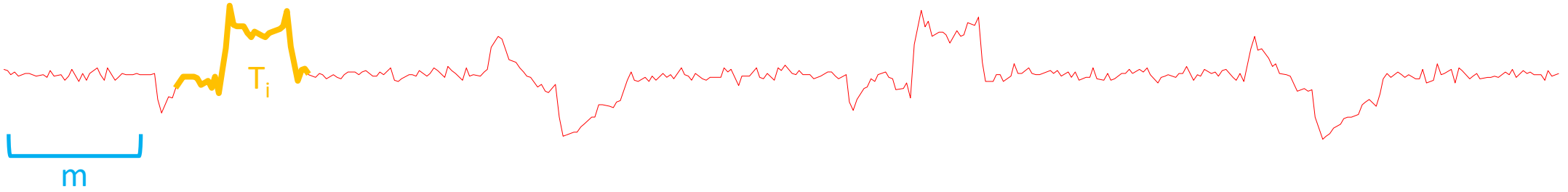
We compute the distances between T_i and every subsequences from T (time complexity = O(|T|log(|T|)))
We then put the distances in a vector based on the position of the subsequences

3	2	0	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

The distance between T_i and T_1 (first subsequence) is 3

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.



inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf

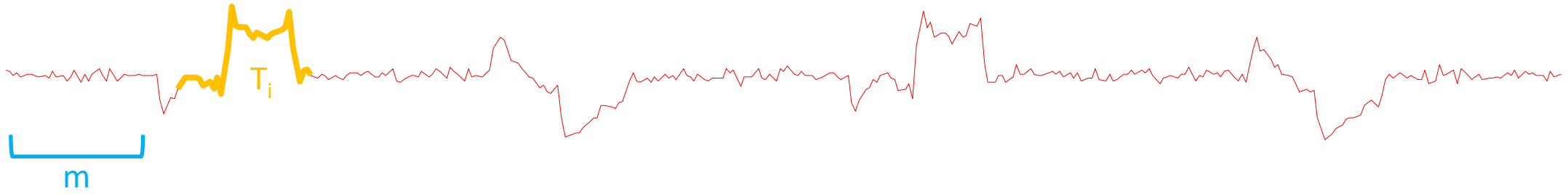
We compute the distances between T_i and every subsequences from T (time complexity = O(|T|log(|T|)))
We them put the distances in a vector based on the position of the subsequences

3	2	0	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

Let say T_i happen to be the third subsequences, therefore the third value in the distance vector is 0

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.



inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf

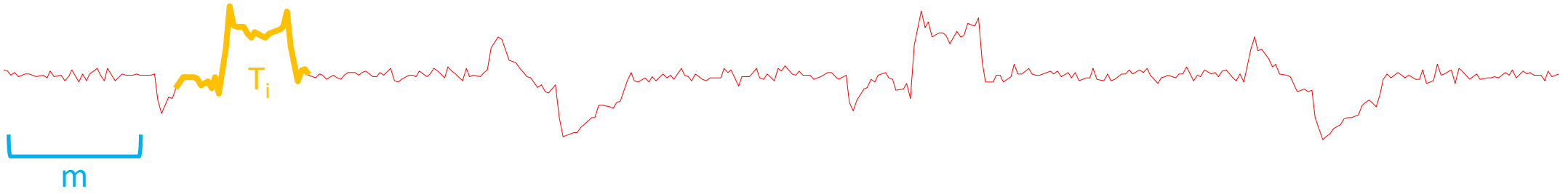
min

Matrix profile is updated by apply elementwise minimum to these two vectors

3	2	0	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.



3	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf	inf

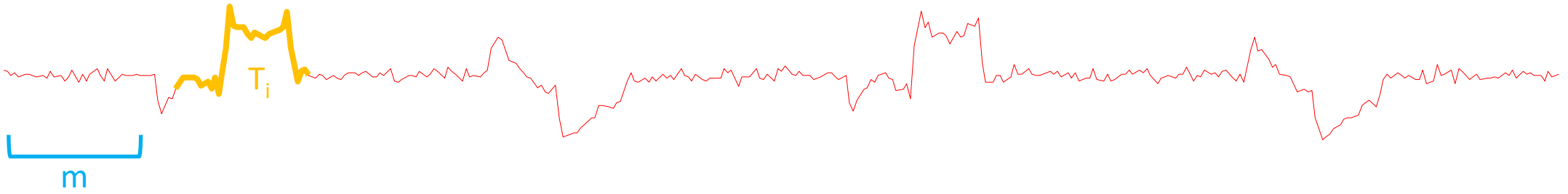
min

Matrix profile is updated by apply elementwise minimum to these two vectors

3	2	0	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.



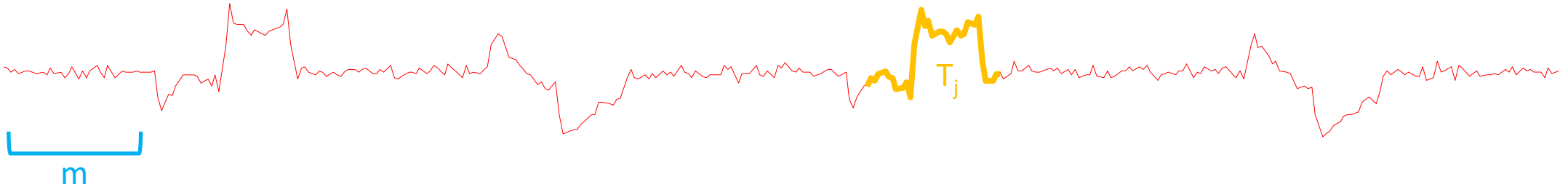
3	2	inf	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

After we finish update matrix profile for the first iteration

3	2	0	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.

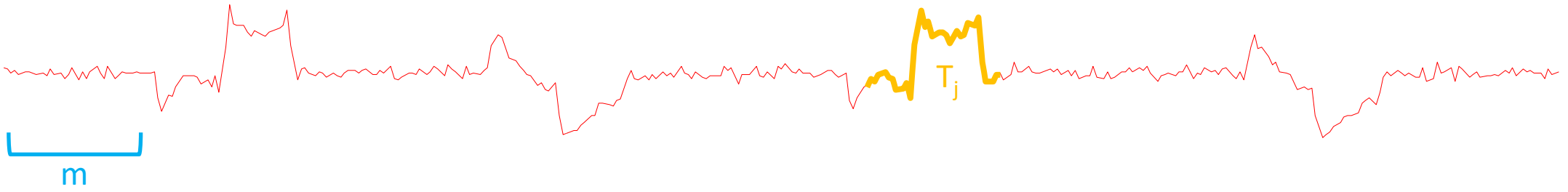


3	2	inf	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

In the second iteration, we randomly select another subsequence T_j and it happens to be the 12^{th} subsequences

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.



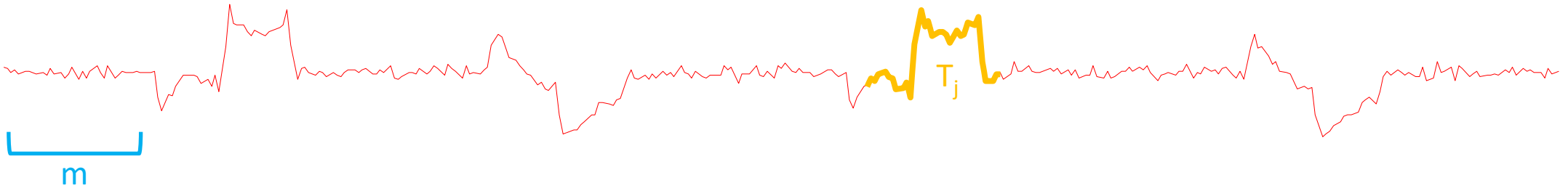
3	2	inf	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

Once again, we compute the distance between T_j and every subsequences of T

2	3	1	4	4	3	6	2	1	5	8	0	2	3	5	9	4	2	2

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.



3	2	inf	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

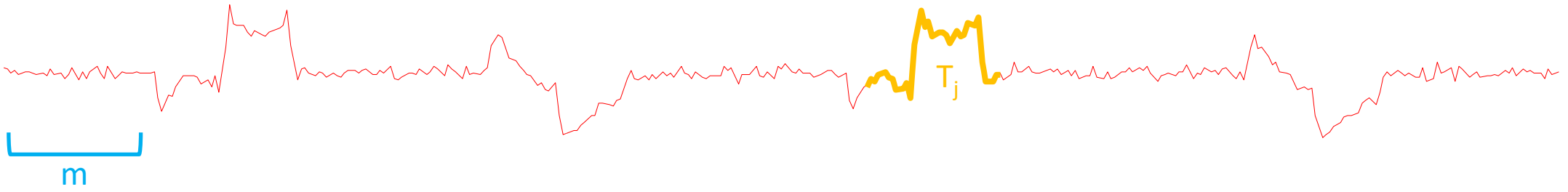
min

The same elementwise minimum

2	3	1	4	4	3	6	2	1	5	8	0	2	3	5	9	4	2	2

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.



2	2	inf	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

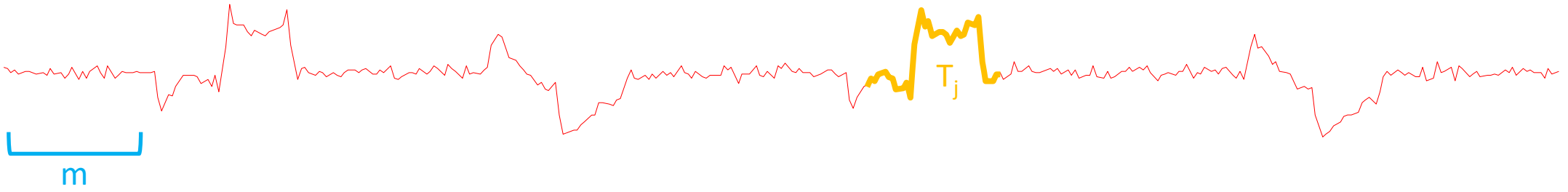
min

The same elementwise minimum

2	3	1	4	4	3	6	2	1	5	8	0	2	3	5	9	4	2	2

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.



2	2	inf	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

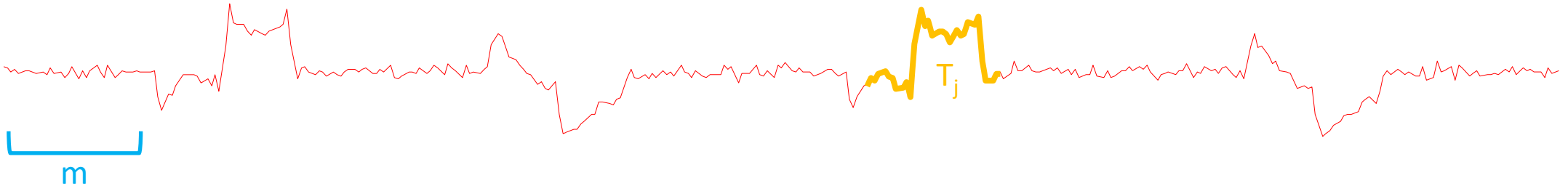
min

The same elementwise minimum

2	3	1	4	4	3	6	2	1	5	8	0	2	3	5	9	4	2	2

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.



2	2	1	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

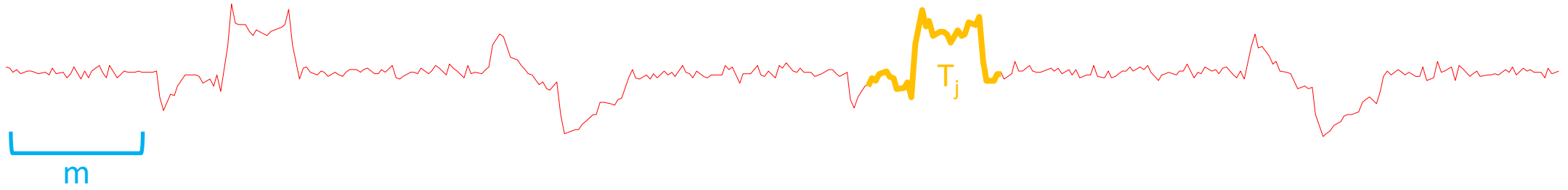
min

The same elementwise minimum

2	3	1	4	4	3	6	2	1	5	8	0	2	3	5	9	4	2	2

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.



2	2	1	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

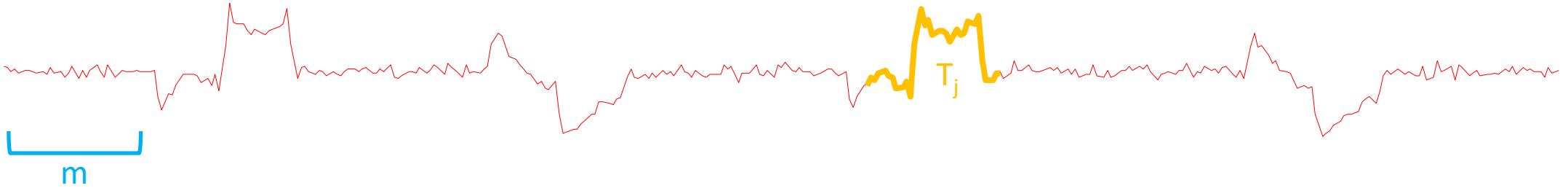
min

2	3	1	4	4	3	6	2	1	5	8	0	2	3	5	9	4	2	2

We repeat the two steps (distance computation and update) until we have used every subsequences. The different indexes are analyzed in parallel and the distance is calculated using the Mueen's Algorithm for Similarity Search (MASS) https://www.cs.unm.edu/~mueen/FastestSimilaritySearch.html

How to Compute Matrix Profile?

- Given a time series, T and a desired subsequence length, m.



2	2	1	5	3	4	5	1	2	9	8	4	2	3	4	8	6	2	1

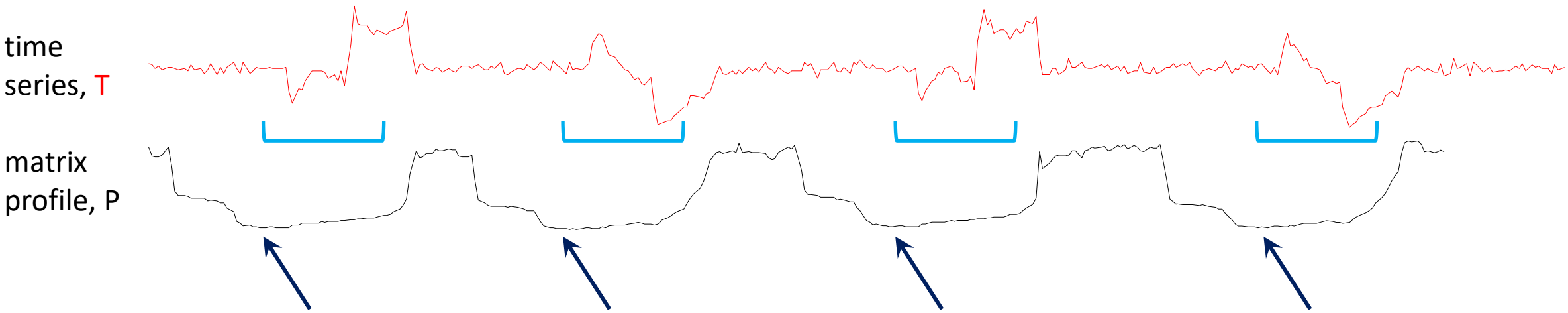
min

2	3	1	4	4	3	6	2	1	5	8	0	2	3	5	9	4	2	2

There are |T| subsequences and the distance computation is O(|T|log(|T|))

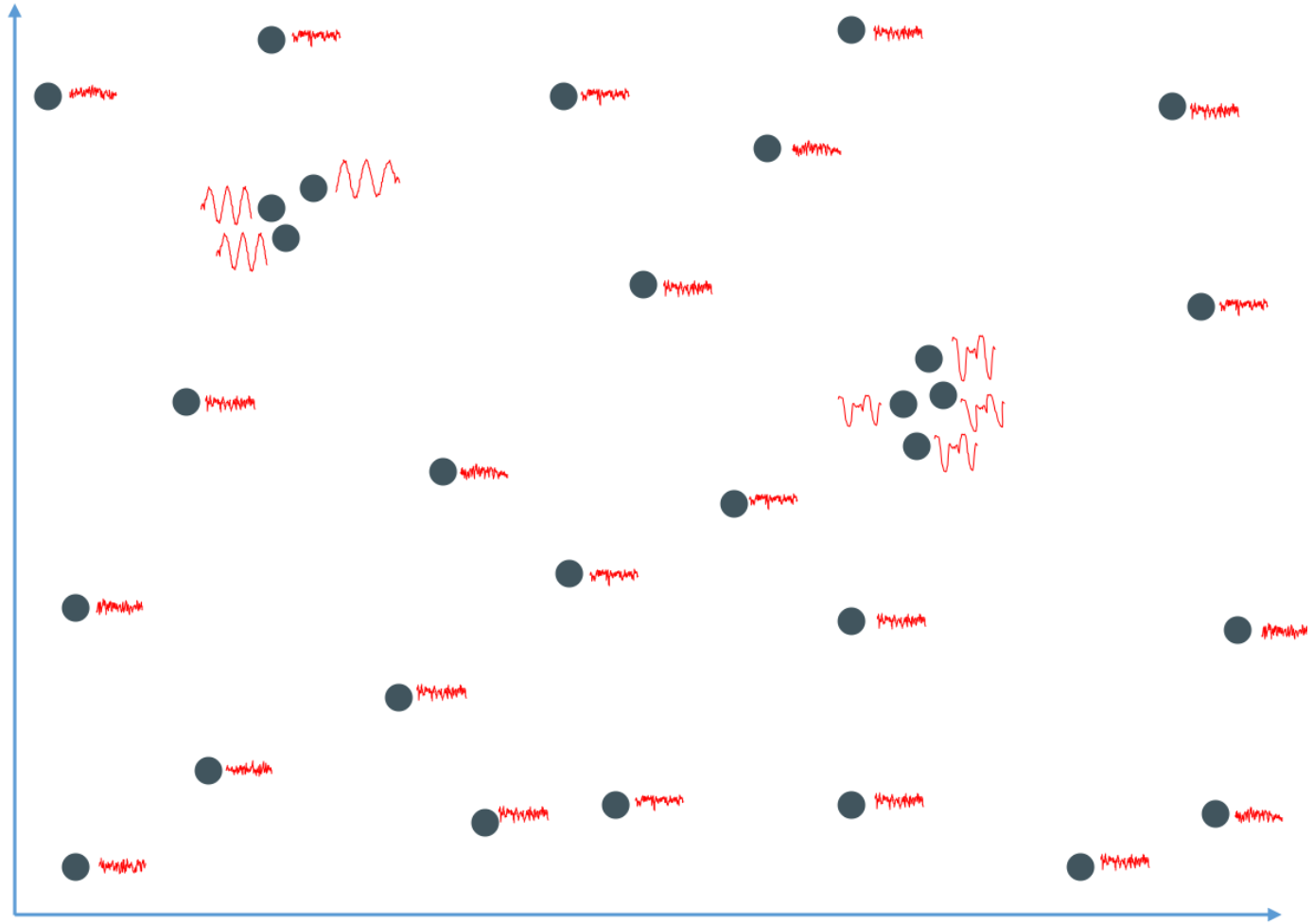
The overall time complexity is $O(|T|^2 log(|T|))$

Motif Discovery From Matrix Profile

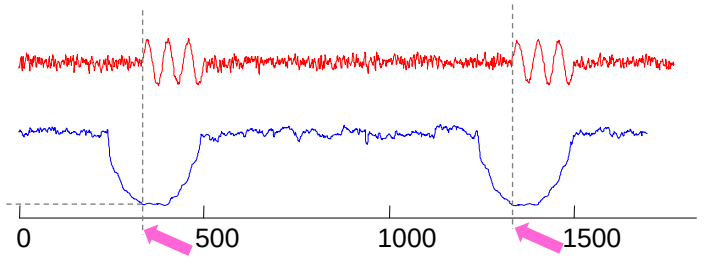


Local minimums are corresponding to motifs

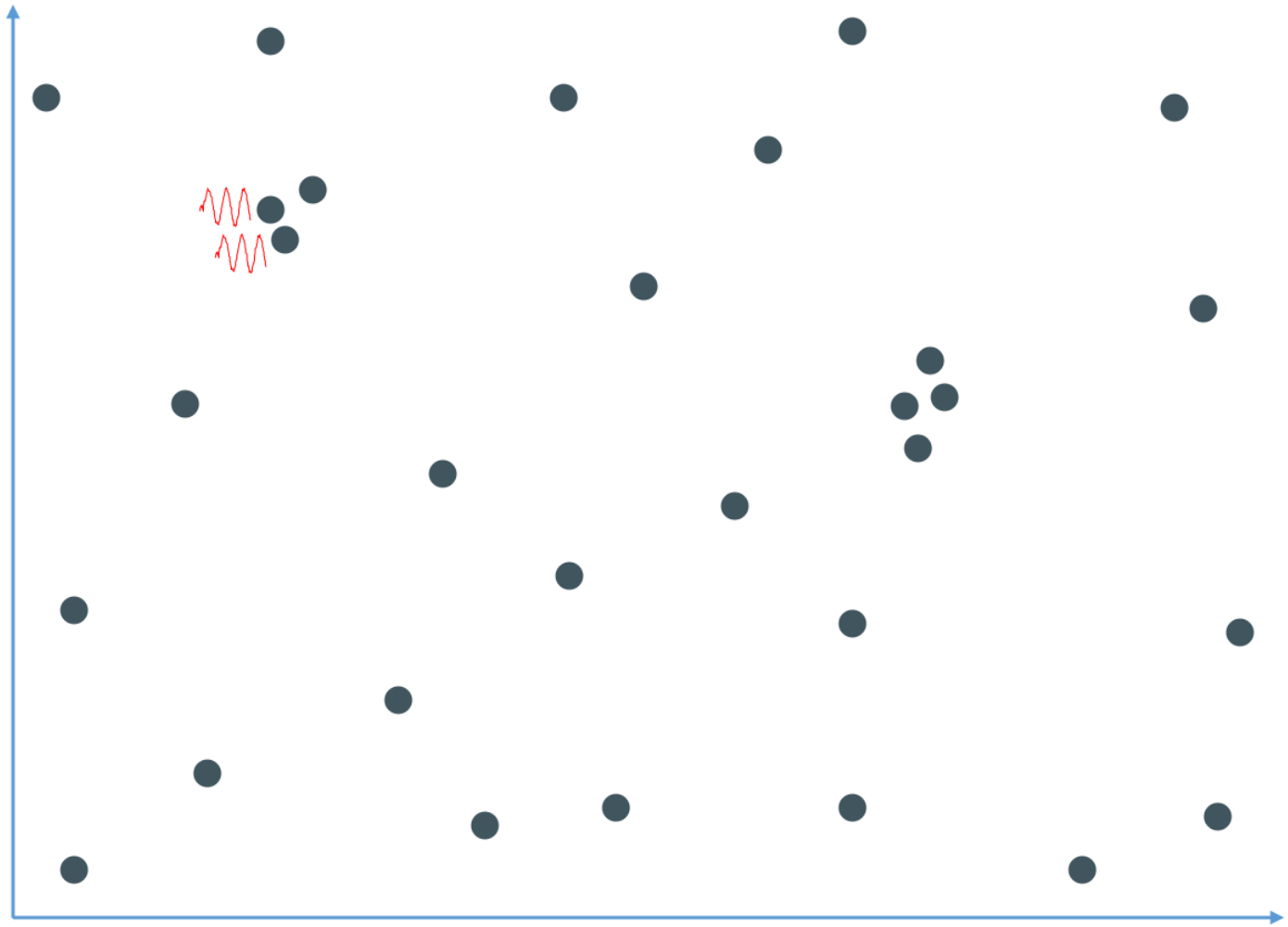
Motif Discovery From Matrix Profile



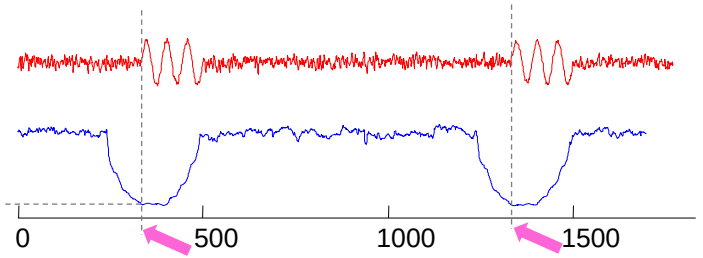
- It is sometime useful to think of time series subsequences as points in m-dimensional space.
- In this view, dense regions in the m-dimensional space correspond to regions of the time series that have a low corresponding MP.



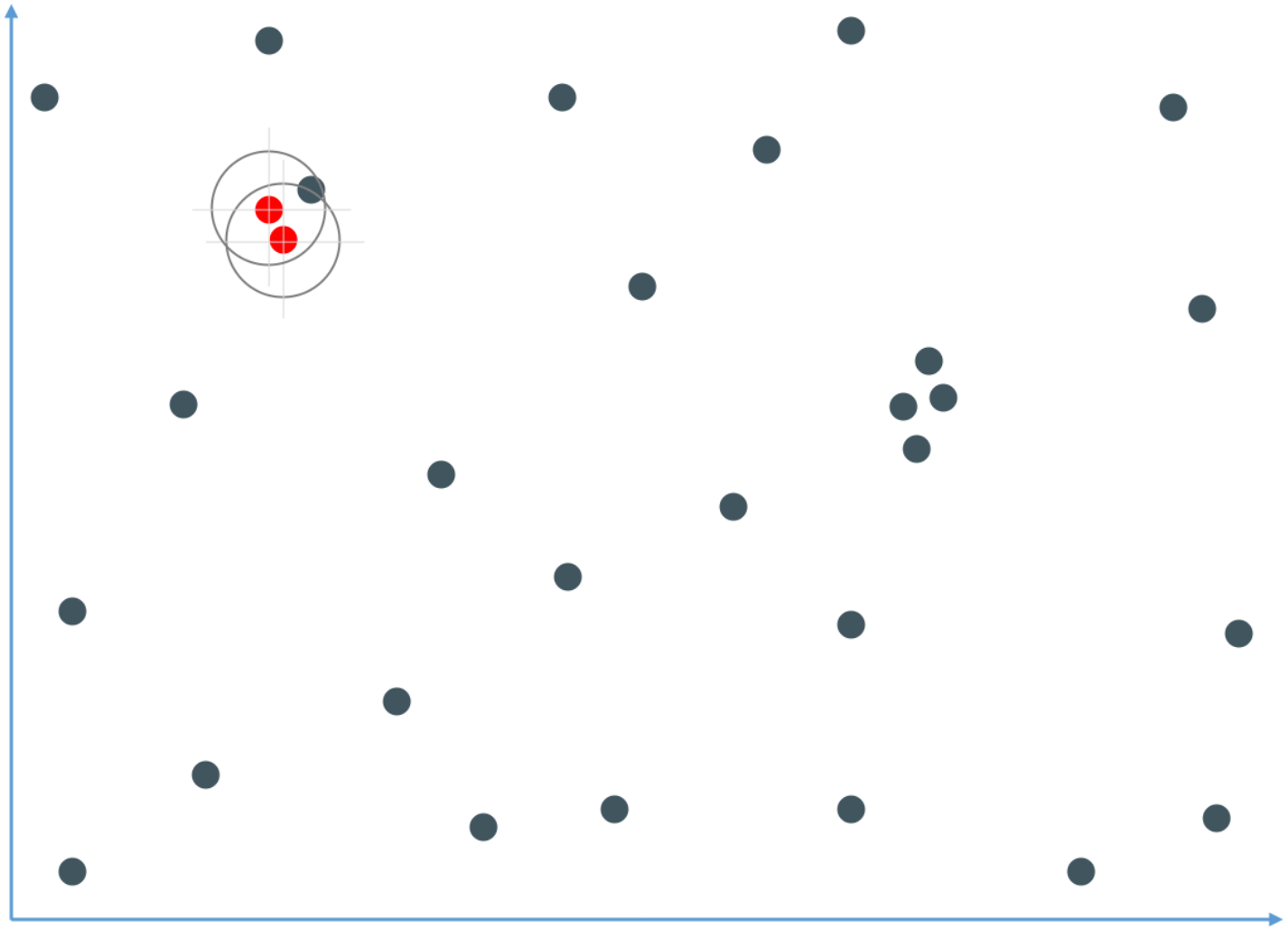
Top-K Motifs



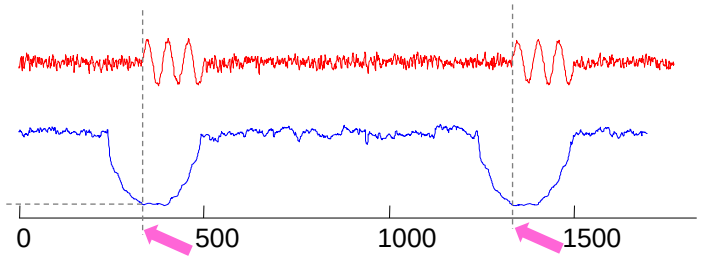
- We need a parameter R.
- 1 < R < (small number, say 3)
- Let's make R = 2 for now.
- We begin by finding the nearest pair of points, the *motif pair*….
- This the pair of subsequences corresponding to lowest pair of values in the MP



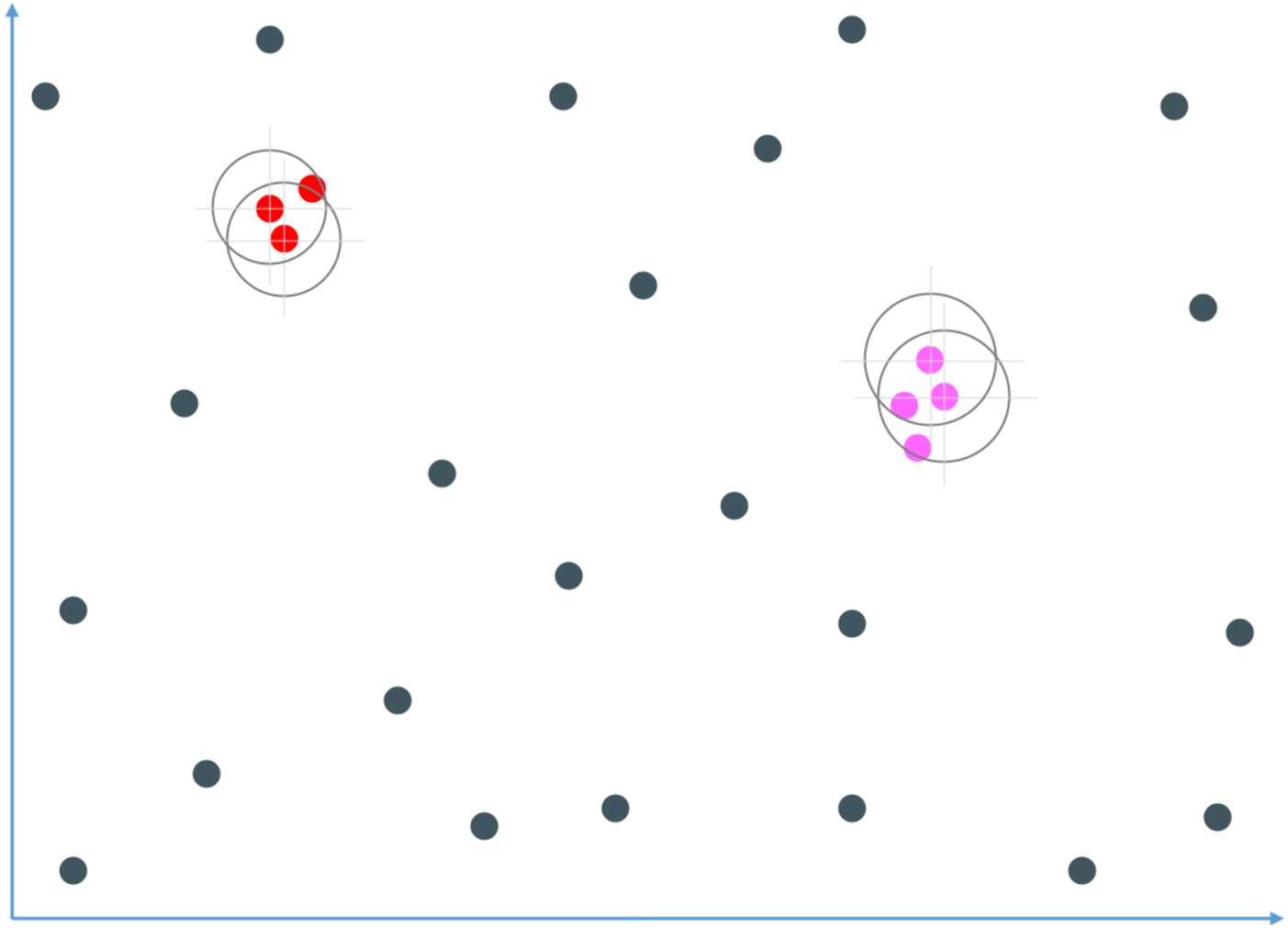
Top-K Motifs



- We find the nearest pair of points are D1 apart.
- Let's draw a circle, D1 times R, around both points.
- Any points that are within either of these circles, are added to this motif, in this case just one.
- The Top-1 motif has three members, it is done.



Top-K Motifs

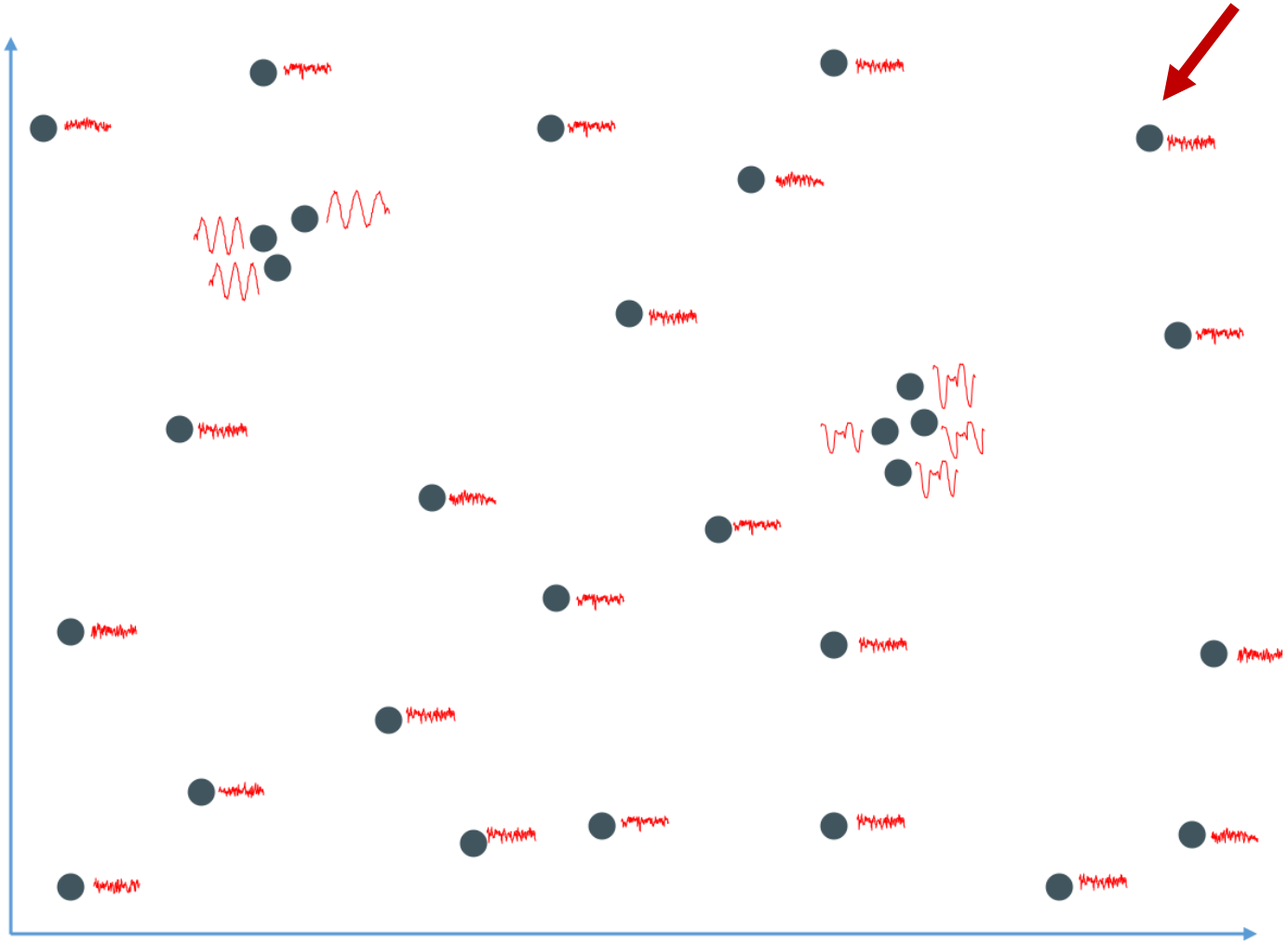


- Now lets find the Top-2 motif. We find the ***nearest pair of points***, excluding anything from the top motif.
- The nearest pair of points are D2 apart.
- Lets draw a circle D2 times R, around both points.
- Any points that are within either of these circles, is added to this motif, in this case there are two for a total of four items in the Top-2 Motif

Top-K Motifs

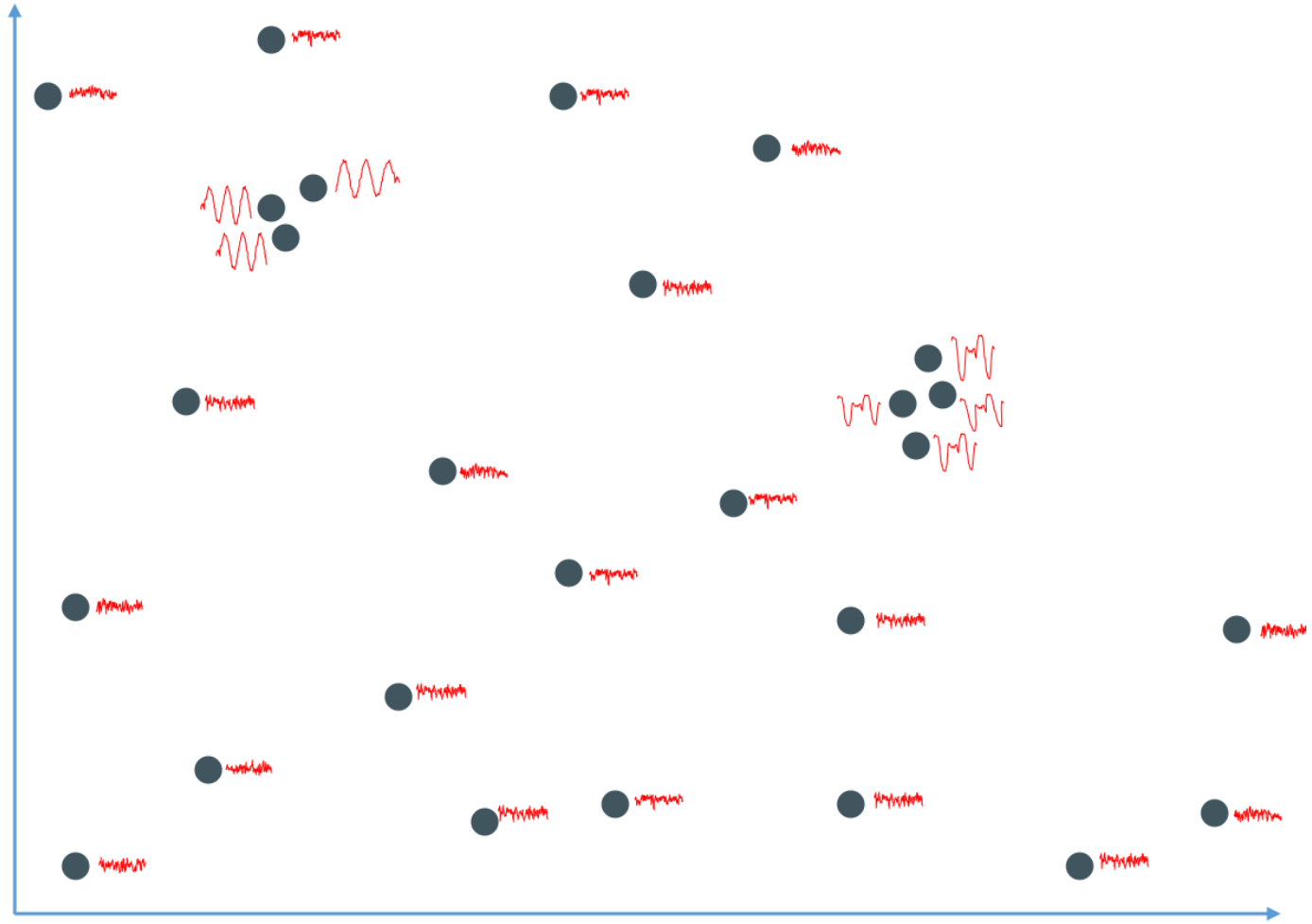
- We are done with the Top-2 Motif
- Note that we will always have:
  - $D_1 < D_2 < D_3 \ldots D_K$
- **When to stop?** (what is K?)
- We could use MDL or a predefined K.

Anomaly Discovery From Matrix Profile



- We need a parameter E of subseqeunces to exclude in the vicinity of the anomaly.
- Lets make E = 2 for now.
- We begin by finding the subsequence with the highest distance in the MP
- This corresponding to biggest anomaly

Top-K Anomaly



- Then we look for the E closest subsequences to the anomaly.
- We remove all of them.
- We can use a predefined K or the MDL to stop.

References

- Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View that Includes Motifs, Discords and Shapelets. Chin-Chia Michael Yeh et al. 1997
- Time Series Shapelets: A New Primitive for Data Mining. Lexiang Ye and Eamonn Keogh. 2016.
- Josif Grabocka, Nicolas Schilling, Martin Wistuba, Lars Schmidt-Thieme (2014): Learning Time-Series Shapelets, in Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2014
- Deep learning for time series classication: a review. Hassan Ismail Fawaz et al. 2019.

Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View that Includes Motifs, Discords and Shapelets

Time Series Shapelets: A New Primitive for Data Mining

Lexiang Ye
Dept. of Computer Science & Engineering
University of California, Riverside, CA 92521
lexiangy@cs.ucr.edu

Eamonn Keogh
Dept. of Computer Science & Engineering
University of California, Riverside, CA 92521
eamonn@cs.ucr.edu

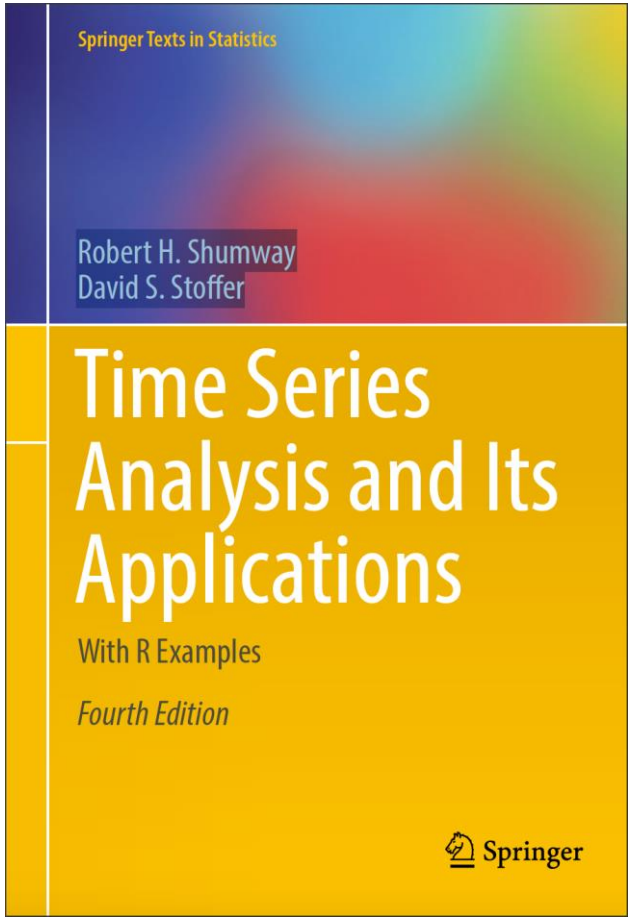
Deep learning for time series classificatio

Hassan Ismail Fawaz[1] · Germain Forestier[1,2] · Jonathan W
Lhassane Idoumghar[1] · Pierre-Alain Muller[1]

arXiv:1809.04356v4 [cs.LG] 14 May 2019

References

- Selective review of offline change point detection methods. Truong, C., Oudre, L., & Vayatis, N. (2020). Signal Processing, 167, 107299.
- Time Series Analysis and Its Applications. Robert H. Shumway and David S. Stoffer. 4th edition.(https://www.stat.pitt.edu/stoffer/tsa4/tsa4.pdf)
- Mining Time Series Data. Chotirat Ann Ratanamahatana et al. 2010. (https://www.researchgate.net/publication/227001229_Mining_Time_Series_Data)
- Dynamic Programming Algorithm Optimization for Spoken Word Recognition. Hiroaki Sakode et al. 1978.
- Experiencing SAX: a Novel Symbolic Representation of Time Series. Jessica Line et al. 2009
- Compression-based data mining of sequential data. Eamonn Keogh et al. 2007.

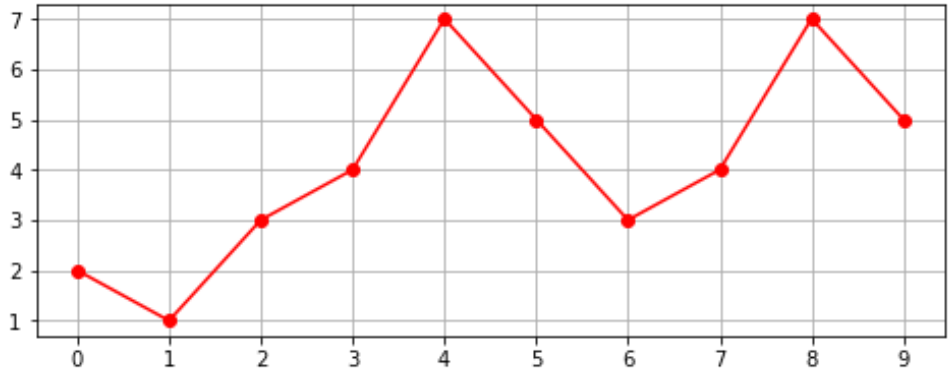


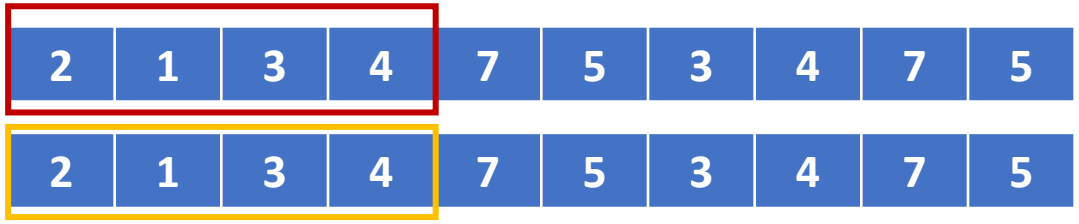
Exercises Matrix Profile

Matrix Profile

Given the TS x = <2,1,3,4,7,5,3,4,7,5>

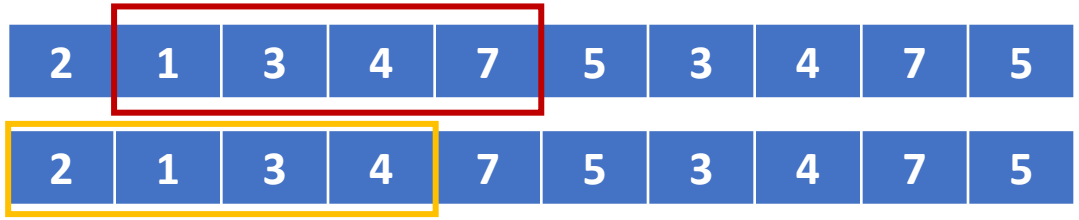
1. Build the Matrix Profile for x with m=4 using the Manhattan distance as distance function between subsequences.
2. Draw the Matrix Profile
3. Identify the motifs with distance equals 0 and length equals to m
4. Which is a correct value for m that would have retrieved more motifs with distance equals to 0?





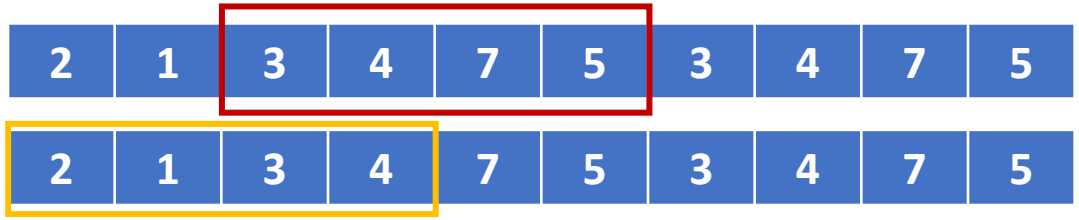
inf						

m = 4



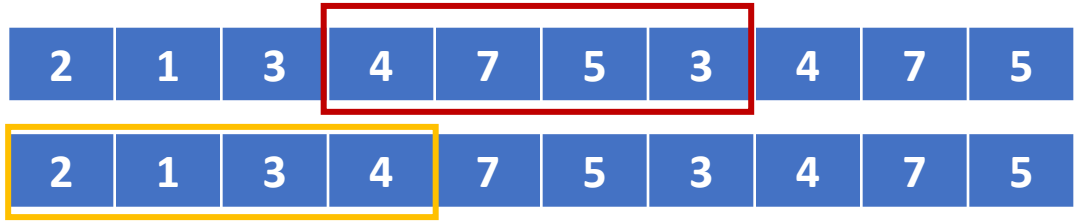
inf	7					

m = 4



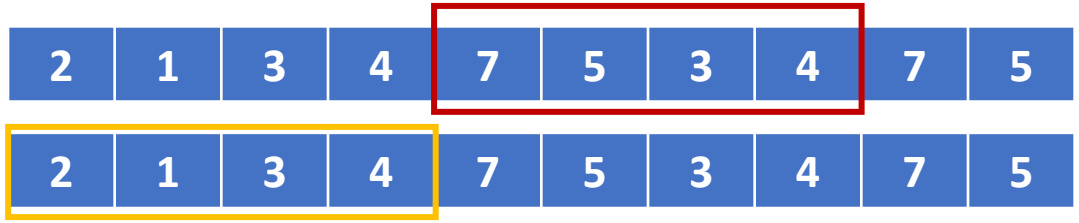
inf	7	9				

m = 4



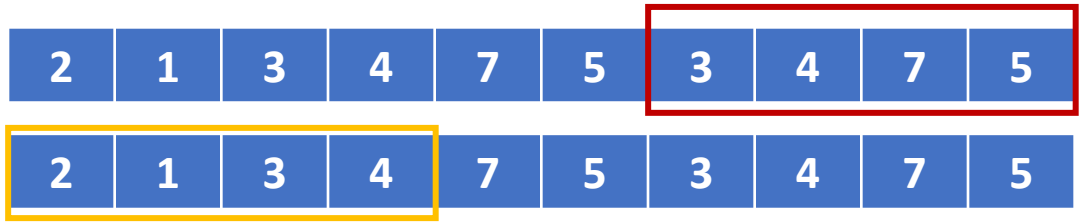
inf	7	9	11			

m = 4



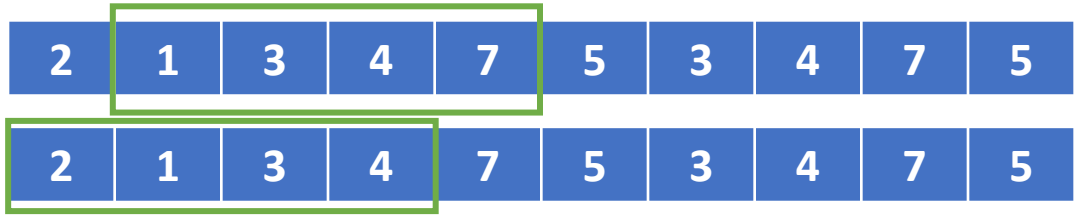
inf	7	9	11	9		

m = 4



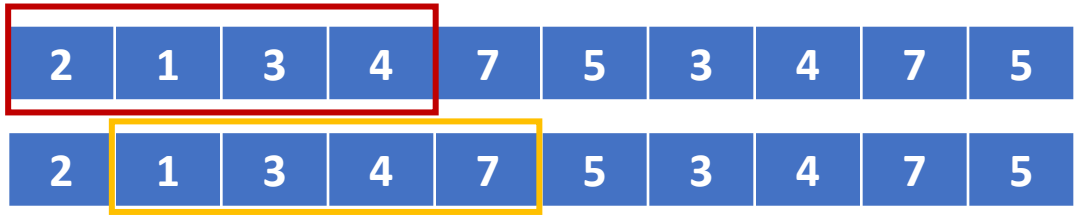
inf	7	9	11	9	9	9

m = 4



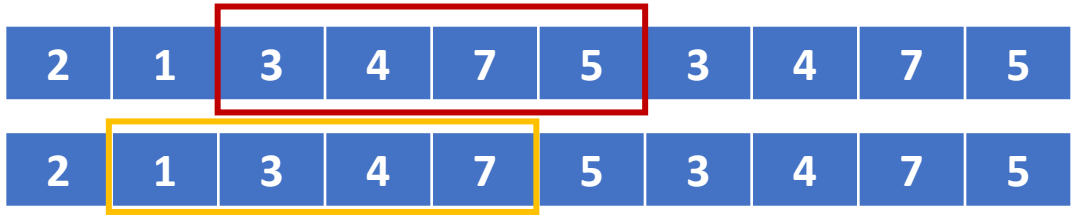
inf	**7**	9	11	9	9	9

m = 4



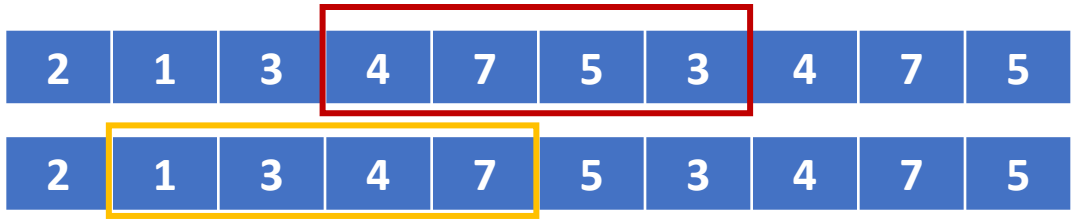
inf	**7**	9	11	9	9	9
7						

m = 4



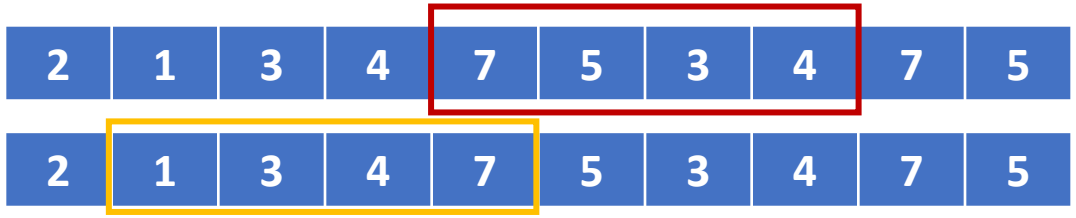
inf	**7**	9	11	9	9	9
7	inf	8				

m = 4



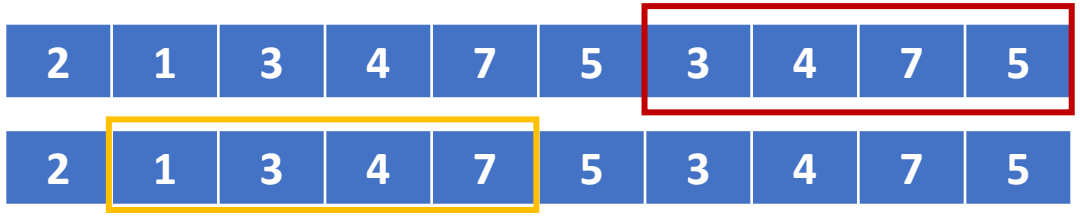
inf	**7**	9	11	9	9	9
7	inf	8	12			

m = 4



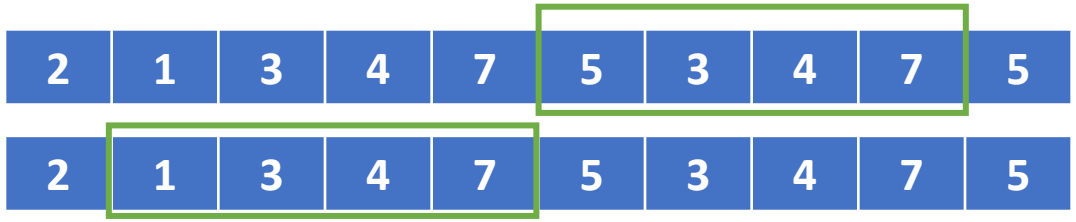
inf	**7**	9	11	9	9	9
7	inf	8	12	12		

m = 4



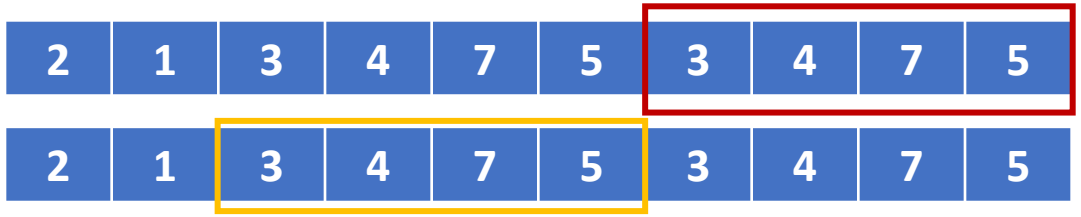
inf	**7**	9	11	9	9	9
7	inf	8	12	12	4	8

m = 4



inf	**7**	9	11	9	9	9
7	inf	8	12	12	**4**	8

m = 4

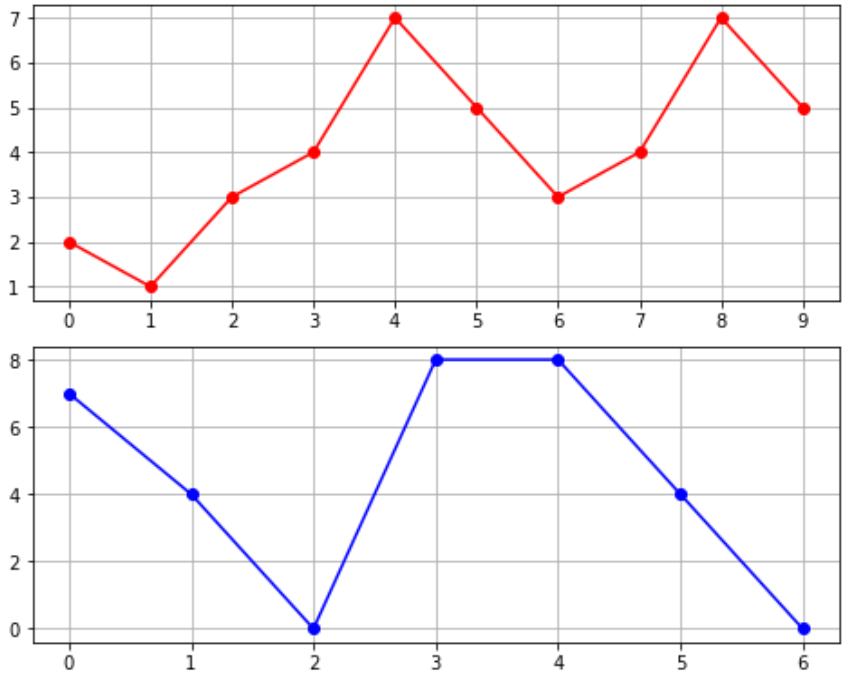


inf	**7**	9	11	9	9	9
7	inf	8	12	12	**4**	8
9	10	inf	8	9	8	**0**

m = 4

Matrix Profile

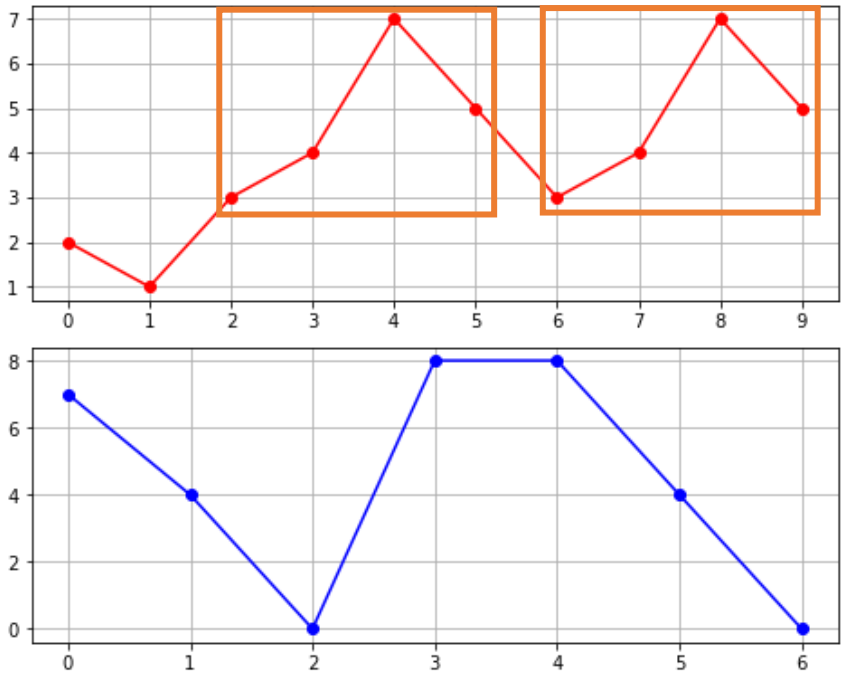
- x = <2, 1, 3, 4, 7, 5, 3, 4, 7, 5>
- mp = < 7, 4, 0, 8, 8, 4, 0 >



Matrix Profile

- x = <2, 1, 3, 4, 7, 5, 3, 4, 7, 5 >
- mp = < 7, 4, 0, 8, 8, 4, 0 >

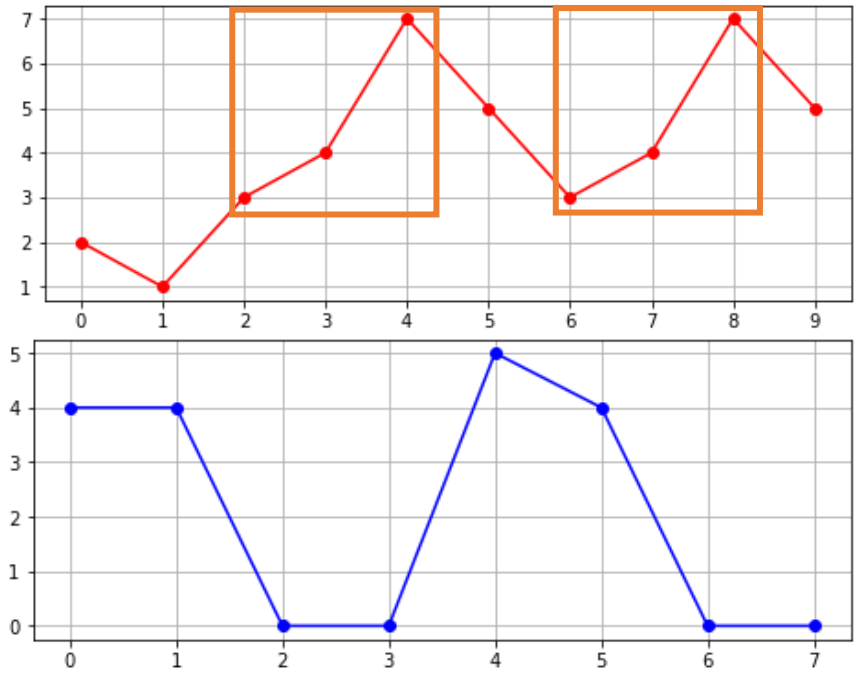
m=4



Matrix Profile

- x = <2, 1, 3, 4, 7, 5, 3, 4, 7, 5 >
- mp = < 4, 4, 0, 0, 5, 4, 0, 0 >

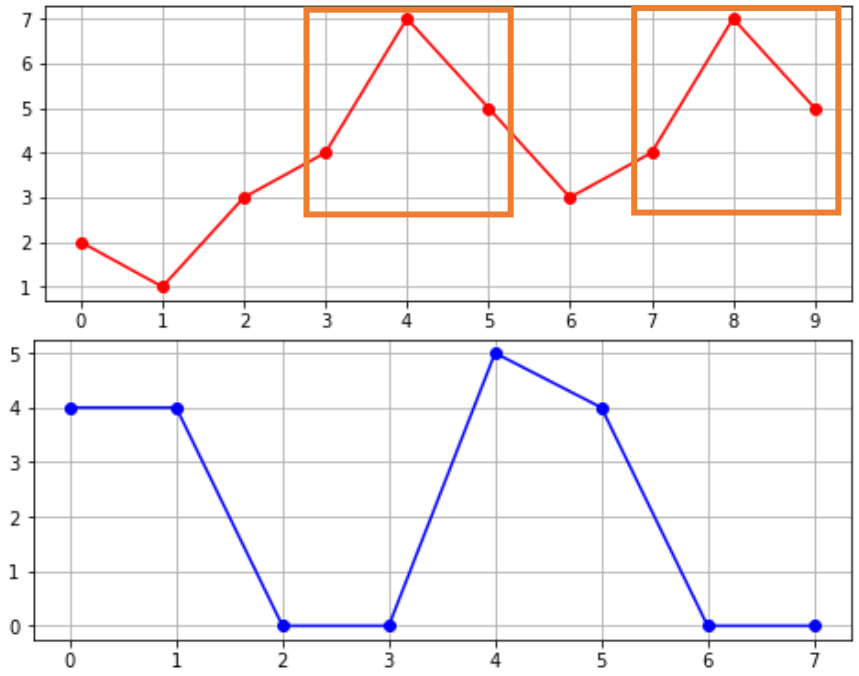
m=3



Matrix Profile

- x = <2, 1, 3, 4, 7, 5, 3, 4, 7, 5 >
- mp = < 4, 4, 0, 0, 5, 4, 0, 0 >

m=3



Matrix Profile

Given the TS x = <5,5,3,5,5,1>

1. Build the Matrix Profile for x with m=2 using the Manahttan distance as distance function between subsequences.
2. Draw the Matrix Profile
3. Identify the motifs with distance equals 0 and length equals to m