Scientific seminar
on information technologies

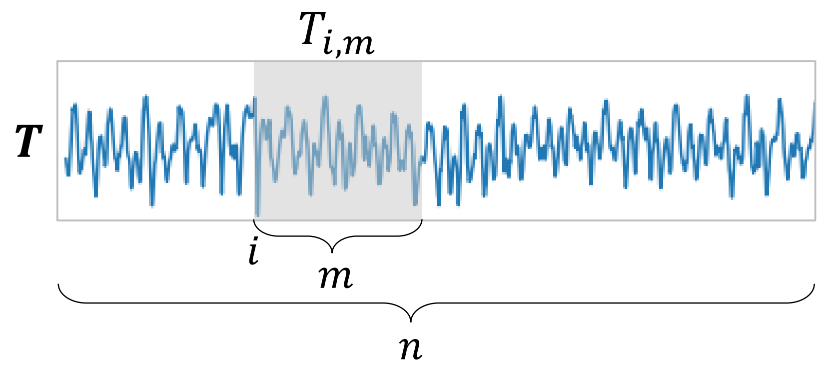
Parallel Algorithm for Discovery Typical Subsequences of a Time Series on Graphical Processor

Andrey Goglachev, Mikhail Zymbler

Chelyabinsk–2021

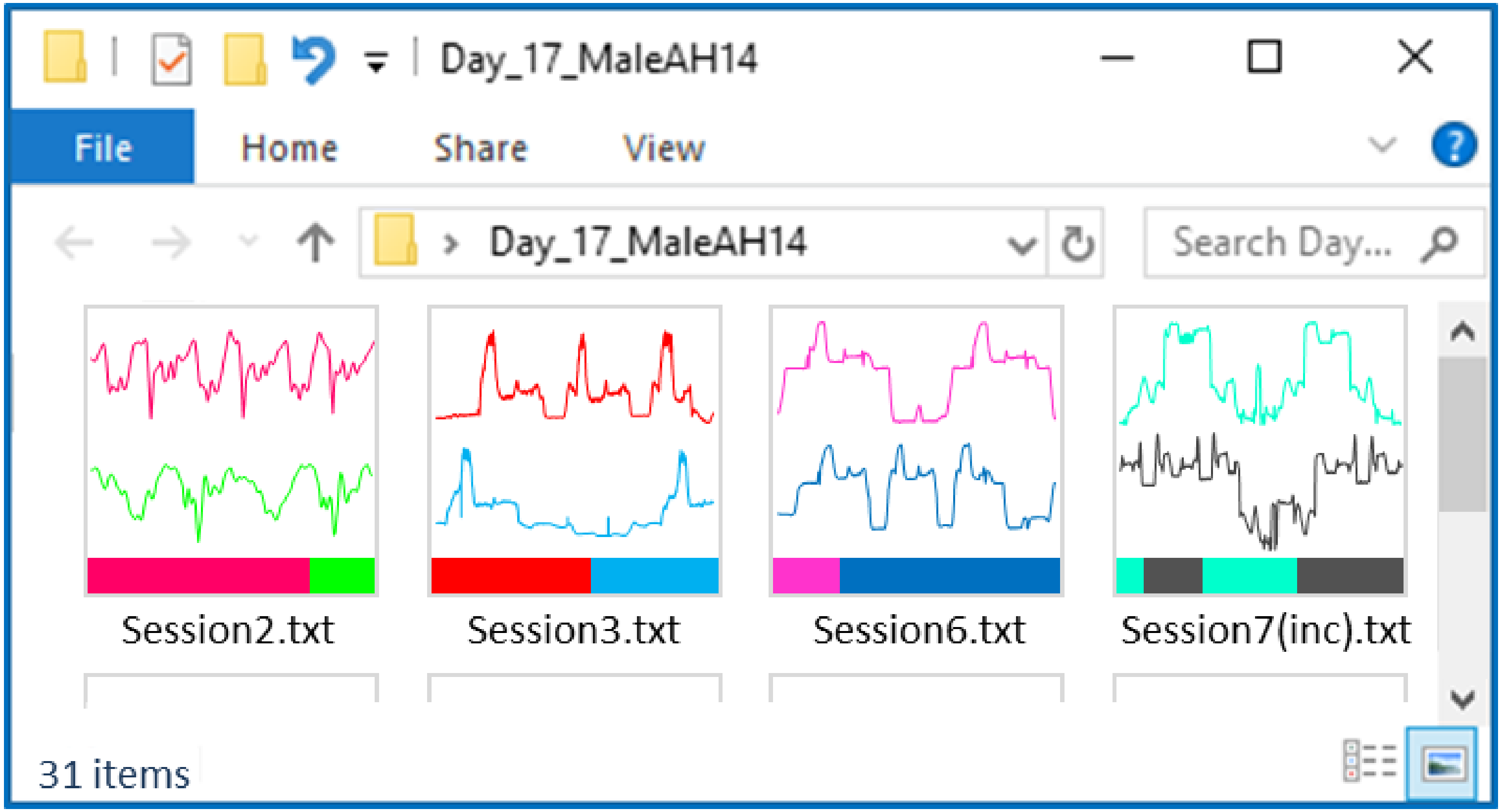
Informal Problem Statement

- We are given:
  - an n-length time series T
  - a subsequence length m
- We must find:
  - a set of subsequences that reflects the respective process/activity



- Application: annotating and visualization of long time series
  - monitoring of human functional diagnostics indicators;
  - monitoring the technical conditions of complex machines and mechanisms;
  - etc.

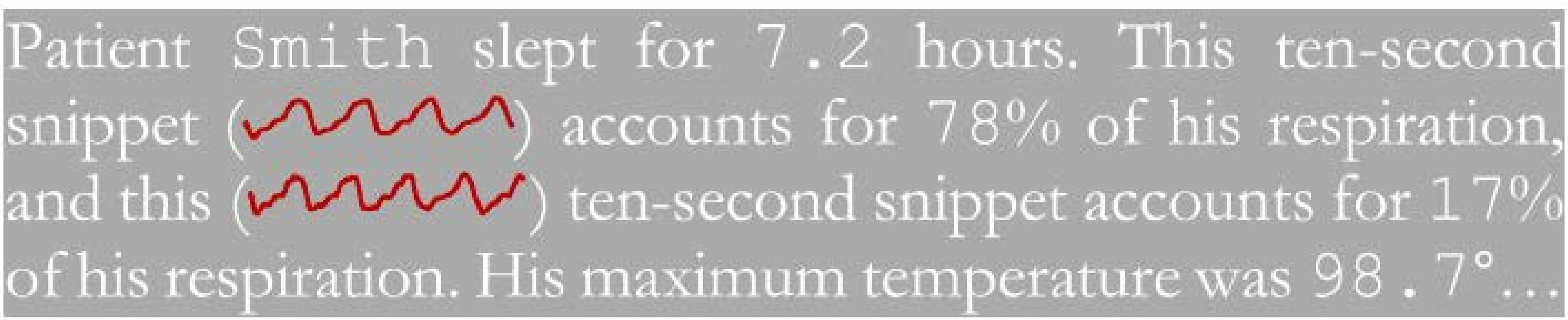
Examples



Summarizing the patient's motor activity
according to the indications of the hip accelerometer

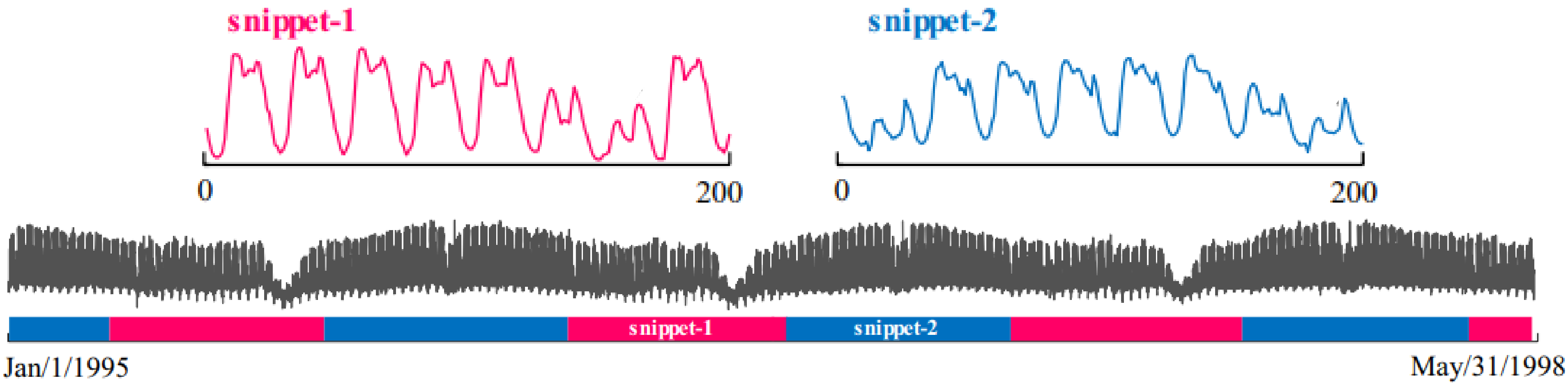
Examples

Summarizing the patient's motor activity
according to the indications of the chest accelerometer



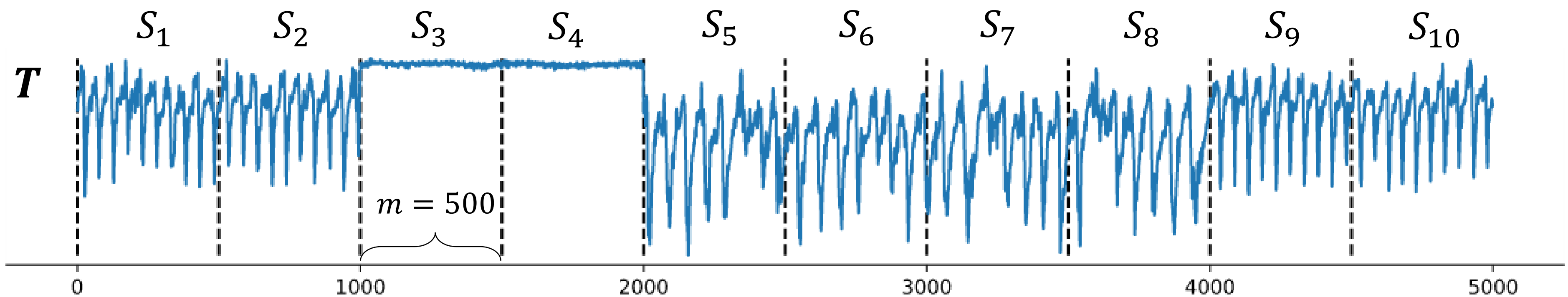
Summarizing the patient's respiratory activity
in studies of apnea syndrome

Examples



Summary of hourly energy consumption in Italy for 3 years.
Typical subsequences are weekly intervals in warm and cold seasons

Formalization: The Snippet Concept*

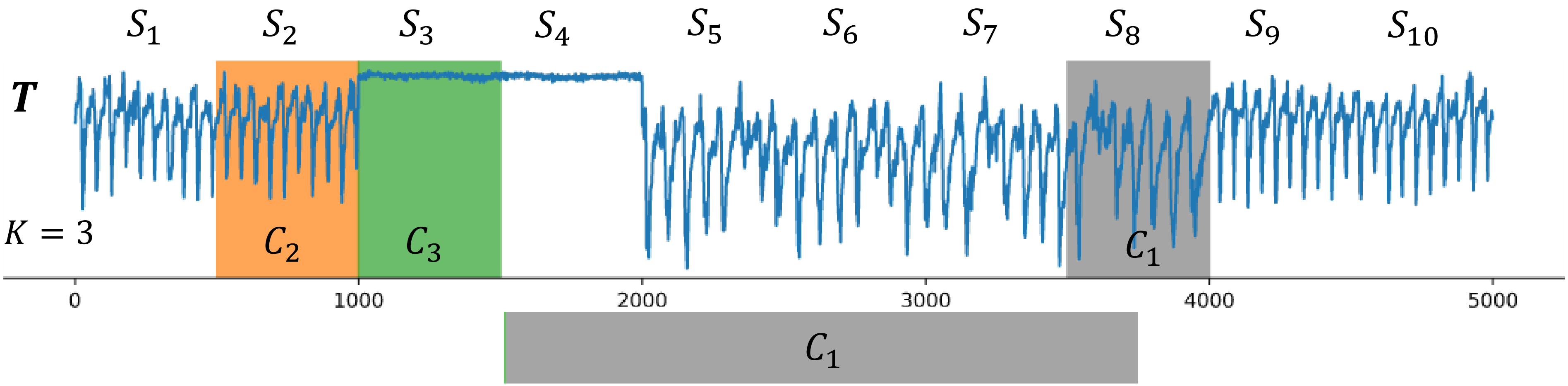


$S_i = T_{m\cdot(i-1)+1,m},\ 1 \le i \le n/m$

1. Let us represent a time series as a set of n/m-length non-overlapped segments
   – if n is not a multiple of m, then pad the time series right by zeroes

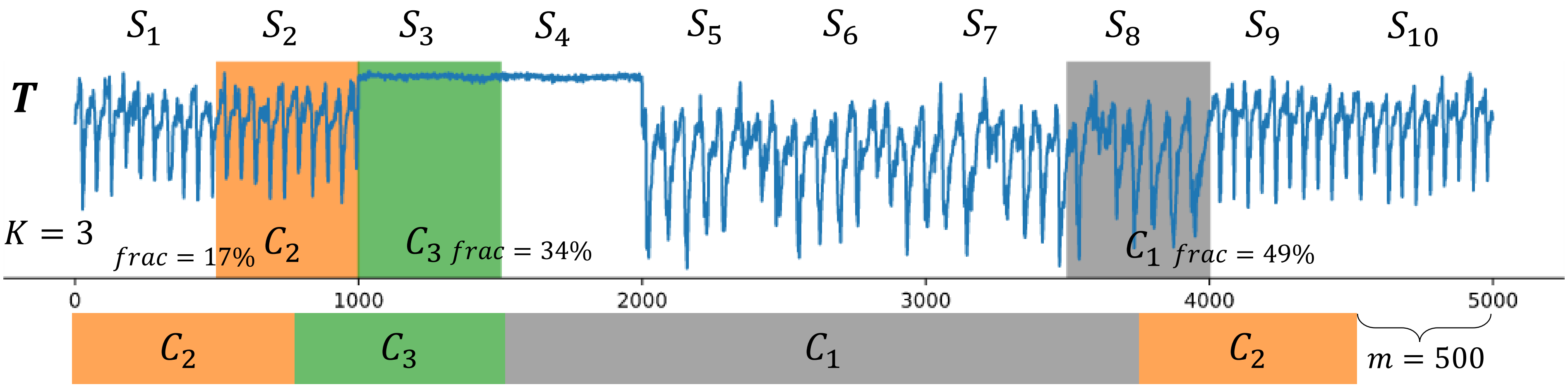
* Imani S., Madrid F., Ding W., Crouter S.E., Keogh E.J. Introducing time series snippets: a new primitive for summarizing long time series. Data Min. Knowl. Discov. 34(6): 1713-1743 (2020). doi: [10.1007/s10618-020-00702-y](https://doi.org/10.1007/s10618-020-00702-y)

Formalization: The Snippet Concept



1. Let us represent a time series as a set of n/m-length non-overlapped segments
2. For each segment, let us find the most similar subsequences (**nearest neighbors**)

Formalization: The Snippet Concept



1. Let us represent a time series as a set of n/m-length non-overlapped segments
2. For each segment, let us find the most similar subsequences (**nearest neighbors**)
3. Let us identify the segment (**snippet**) by its nearest neighbors
4. Let us take the top-$\boldsymbol{K}$ snippets in descending order of the number of their nearest neighbors (**coverage**)

MPdist*: A Subsequence Similarity Measure

Two **m-length time series** are the more similar by the **MPdist measure**,

the more **ℓ-length** $(3 \leq \ell \leq m)$ **normalized subsequences** close to each other by the **Euclidean metric**, are in them

Metric

Measure

1. Identity of indiscernibles: $d(x, y) = 0 \iff x = y$
2. Symmetry: $d(x, y) = d(y, x)$
3. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$

* Gharghabi S., Imani S., Bagnall A.J., Darvishzadeh A., Keogh E.J.: Matrix Profile XII: MPdist: A Novel Time Series Distance Measure to Allow Data Mining in More Challenging Scenarios. ICDM 2018: 965-970. DOI: [10.1109/ICDM.2018.00119](https://doi.org/10.1109/ICDM.2018.00119)

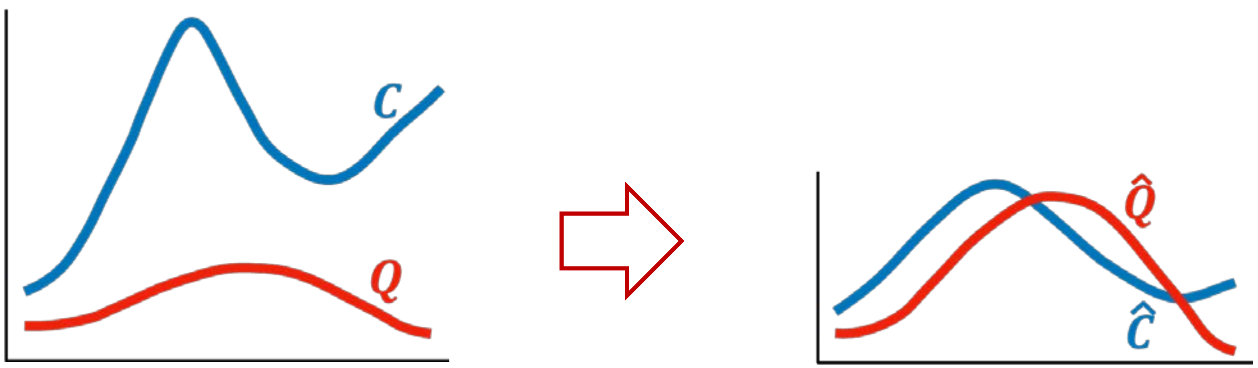
Normalization

- Provides a correct comparison of subsequences with different amplitudes

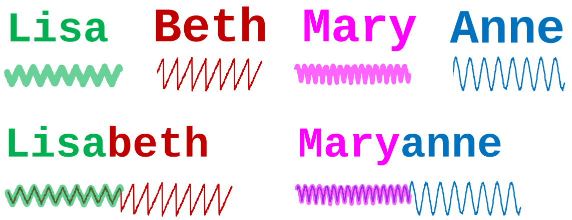
$$\hat{T} = (\hat{t}_1, \dots, \hat{t}_m)$$

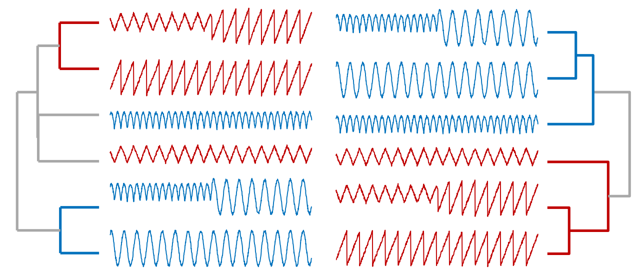
$$\hat{t}_i = \frac{t_i - \mu}{\sigma}$$

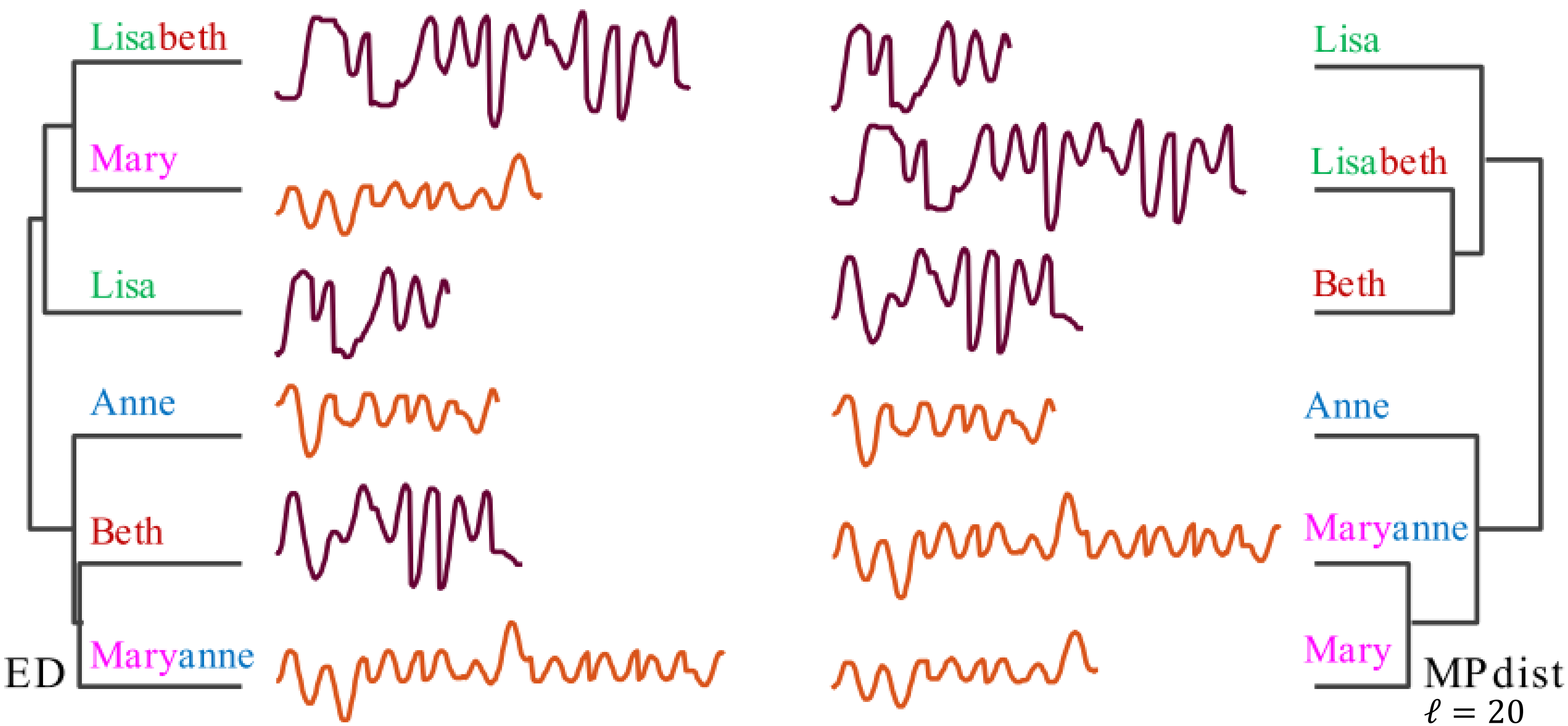
$$\mu = \frac{1}{n} \sum_{i=1}^{m} t_i \qquad \sigma = \sqrt{\frac{1}{m} \sum_{i=1}^{m} t_i^2 - \mu^2}$$



MPdist *vs.* Euclid

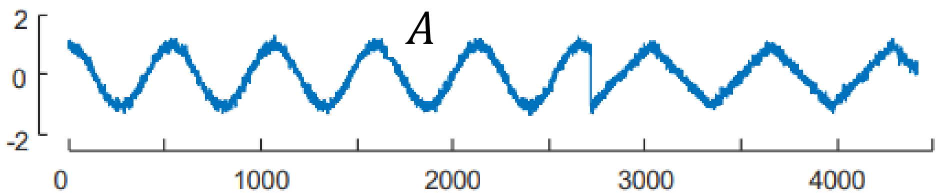


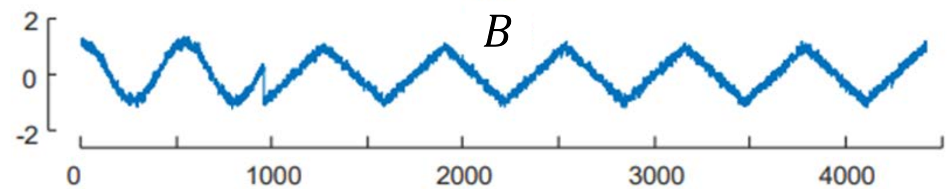




The MPdist Measure

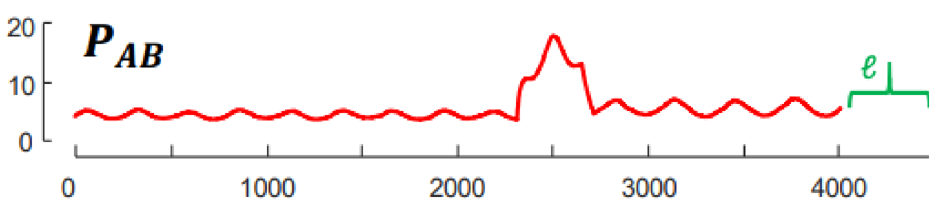
$$|A| = |B| = m$$

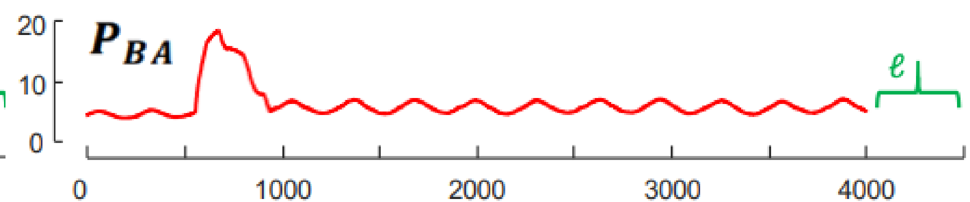




Significant subsequence : $3 \leq \ell \leq m$ (typically, $[0.3m] < \ell \leq [0.8m]$)

Calculation of matrix profiles



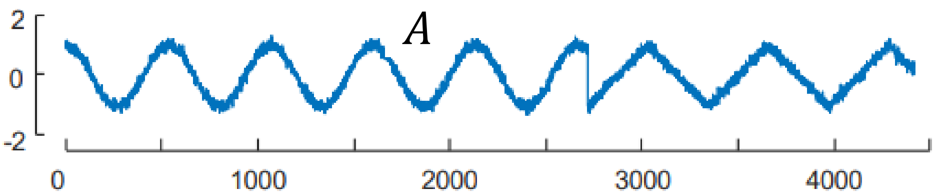


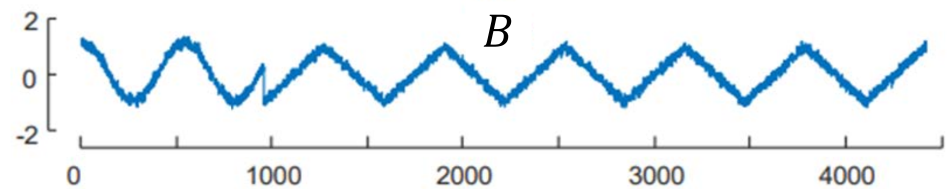
$$\left\{P_{AB}(i) = \mathrm{ED}_{\mathrm{norm}}\left(A_{i,\ell}, B_{j,\ell}\right)\right\}_{i=1}^{m-\ell+1},$$
$$B_{j,\ell} = \arg \min_{1 \leq q \leq m-\ell+1} \mathrm{ED}_{\mathrm{norm}}\left(A_{i,\ell}, B_{q,\ell}\right)$$

$$\left\{P_{BA}(i) = \mathrm{ED}_{\mathrm{norm}}\left(B_{i,\ell}, A_{j,\ell}\right)\right\}_{i=1}^{m-\ell+1},$$
$$A_{j,\ell} = \arg \min_{1 \leq q \leq m-\ell+1} \mathrm{ED}_{\mathrm{norm}}\left(B_{i,\ell}, A_{q,\ell}\right)$$

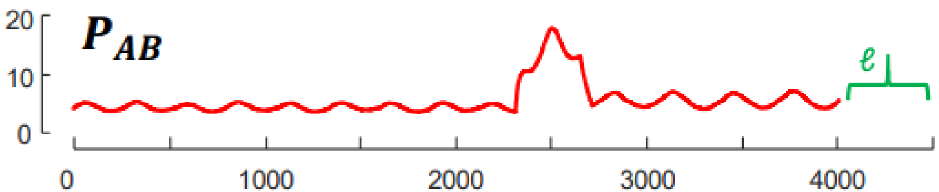
The MPdist Measure

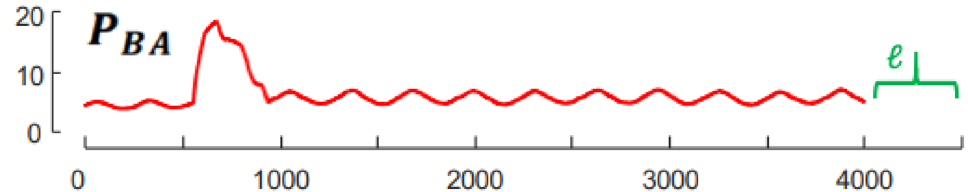
$$|A| = |B| = m$$





$$3 \leq \ell \leq m$$

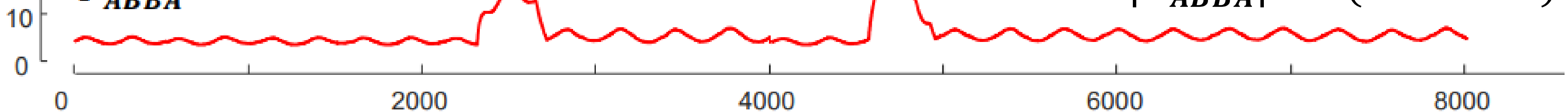




Concatenation of matrix profiles

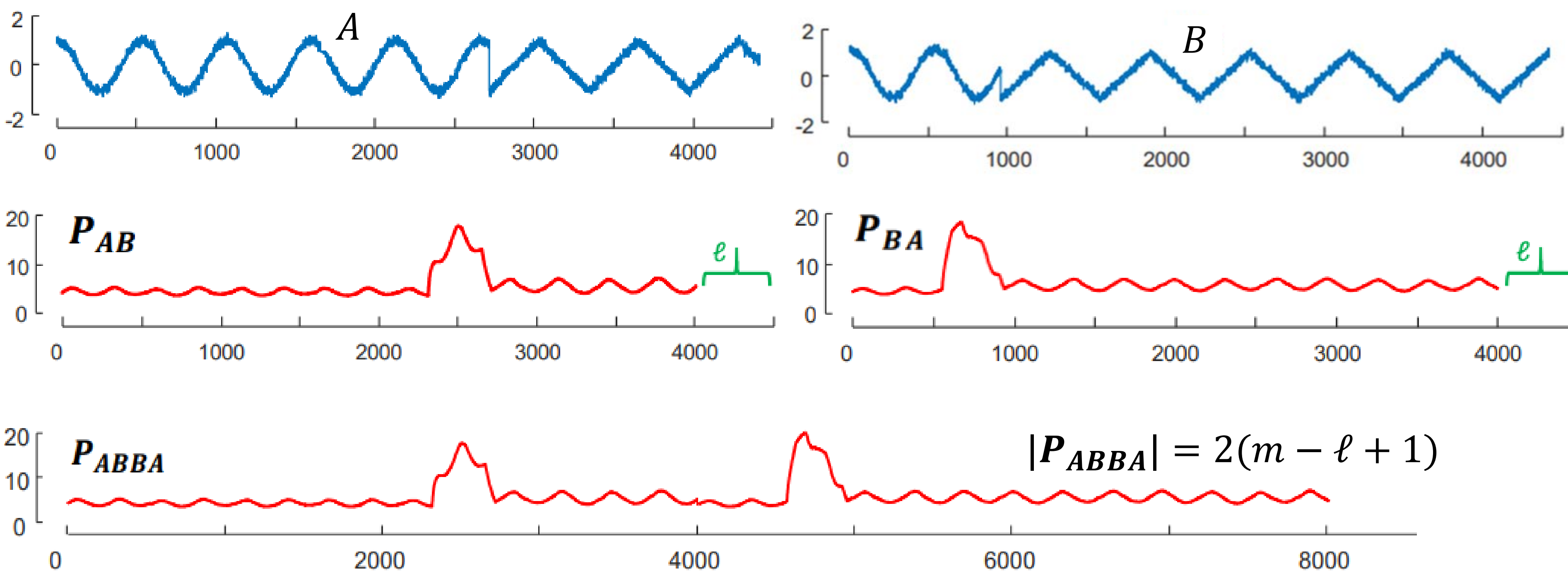
$$P_{ABBA} = P_{AB} \odot P_{BA}$$

$$|P_{ABBA}| = 2(m - \ell + 1)$$



The MPdist Measure

$|A| = |B| = m$ $\qquad$ $3 < \ell \leq m$



$|P_{ABBA}| = 2(m - \ell + 1)$

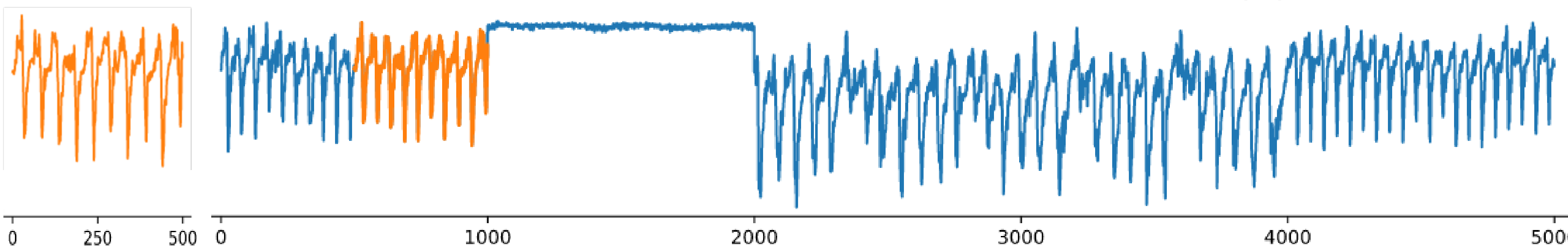
$$\text{MPdist}(A, B, \ell) = \begin{cases} SortedP_{ABBA}(k), & |P_{ABBA}| > k \\ SortedP_{ABBA}(2(m - \ell + 1)), & |P_{ABBA}| \leq k \end{cases}$$

где $k = \lceil 0.05 \cdot 2m \rceil = \lceil 0.1m \rceil$.

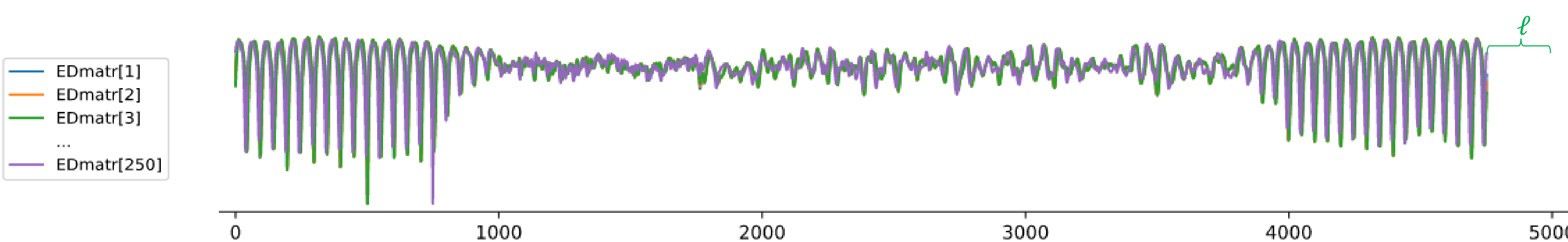
MPdist Profile of a Segment

Segment S
$|S| = m$

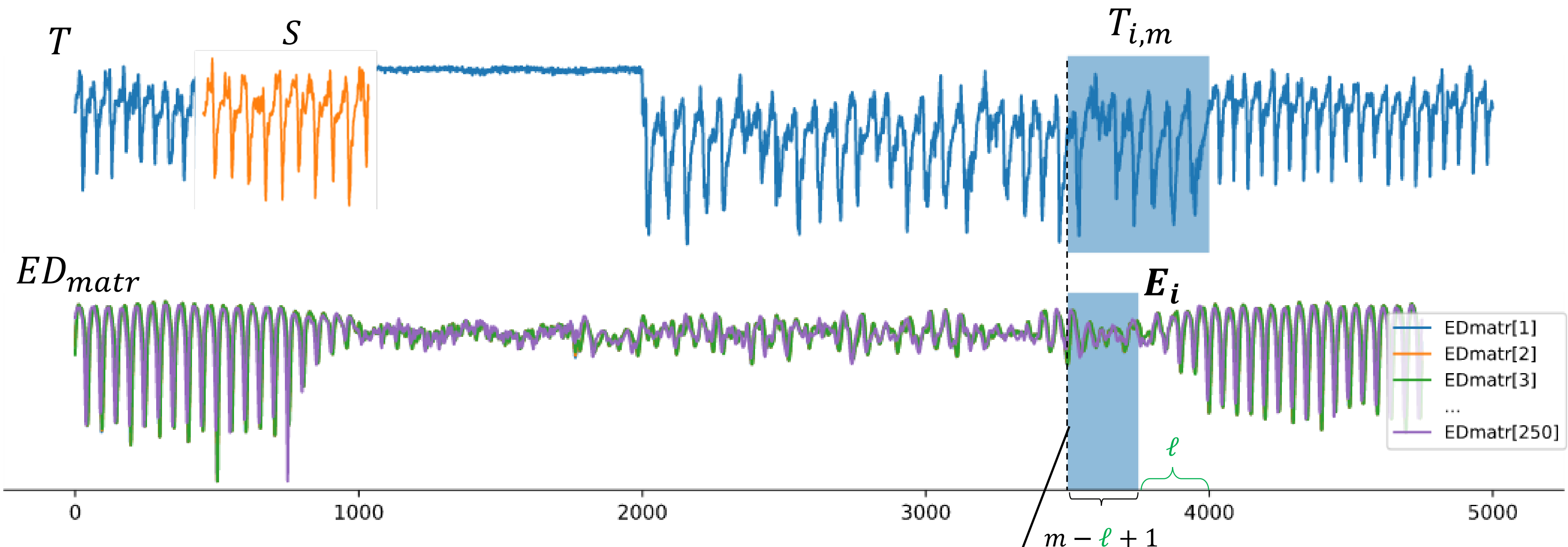
Time series T
$|T| = n \gg m$



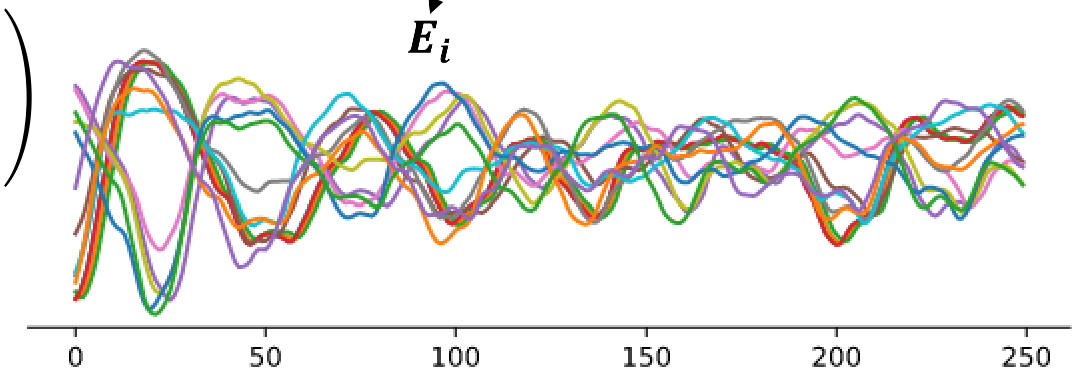
$$\boldsymbol{ED_{matr}}(\boldsymbol{S},\boldsymbol{T},\ell) = \begin{pmatrix} d_{1,1} & \cdots & d_{1,n-\ell+1} \\ \cdots & \ddots & \cdots \\ d_{m-\ell+1,1} & \cdots & d_{m-\ell+1,n-\ell+1} \end{pmatrix}, \qquad d_{i,j} = \mathrm{ED}_{\mathrm{norm}}(S_{i,\ell}, T_{j,\ell})$$



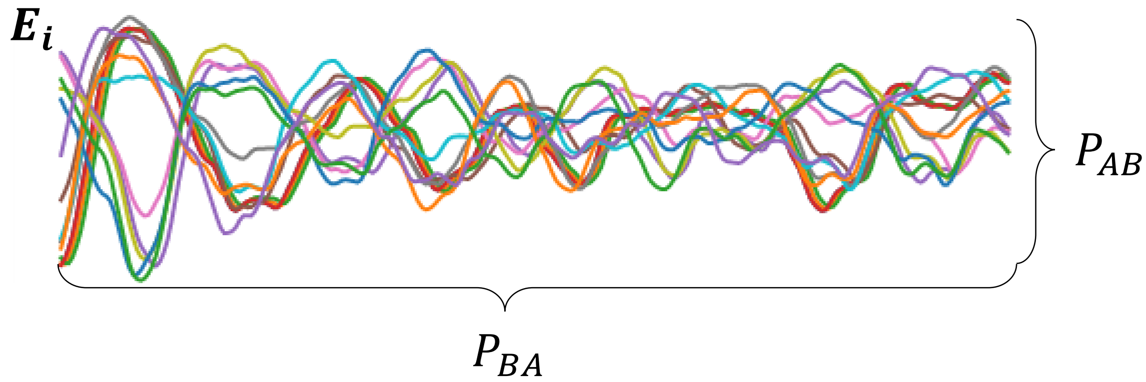
MPdist Profile of a Segment



$$\boldsymbol{E_i} = \begin{pmatrix} d_{1,i} & \cdots & d_{1,i+m-\ell+1} \\ \cdots & \ddots & \cdots \\ d_{m-\ell+1,i} & \cdots & d_{m-\ell+1,i+m-\ell+1} \end{pmatrix}$$



MPdist Profile of a Segment

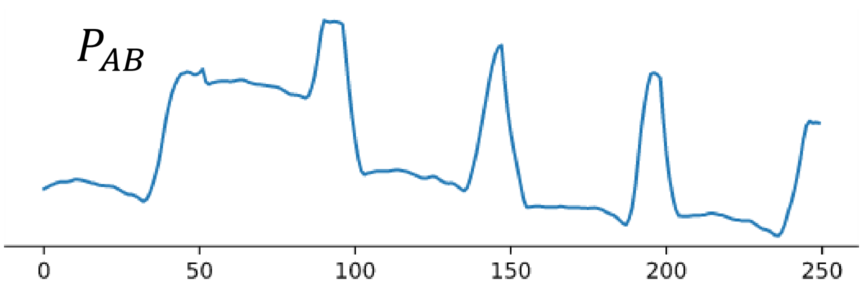


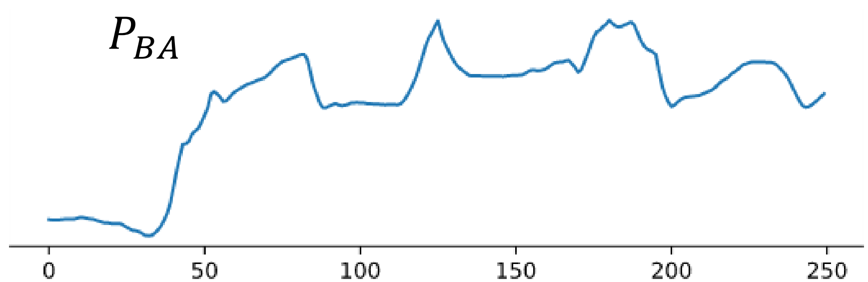
$$P_{AB}(i) = \min_{1 \le j \le m-\ell+1} E(i,j),$$

$$1 \le i \le m - \ell + 1$$

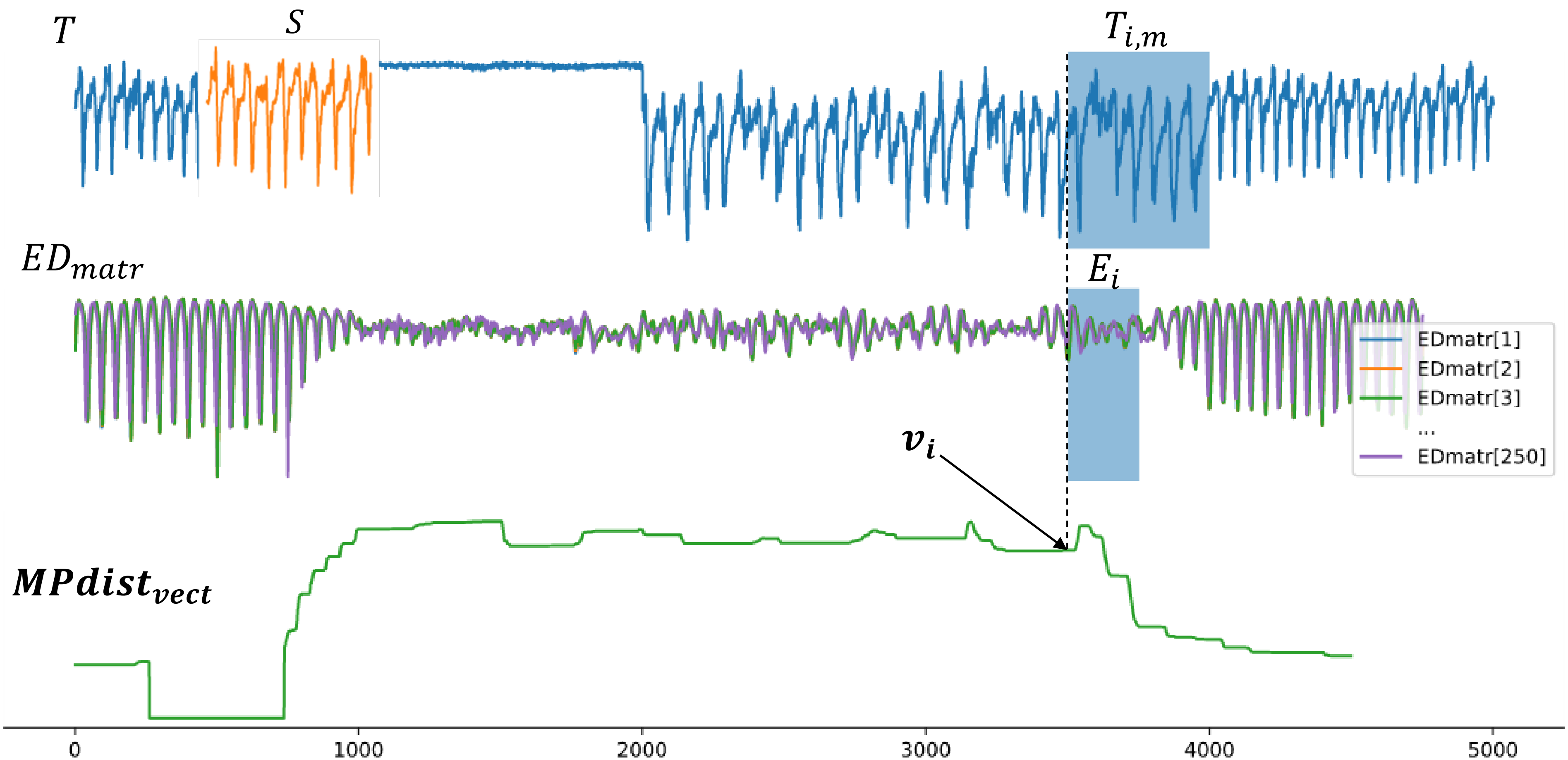
$$P_{BA}(j) = \min_{1 \le i \le m-\ell+1} E(i,j),$$

$$1 \le j \le m - \ell + 1$$





MPdist-profile of a Segment



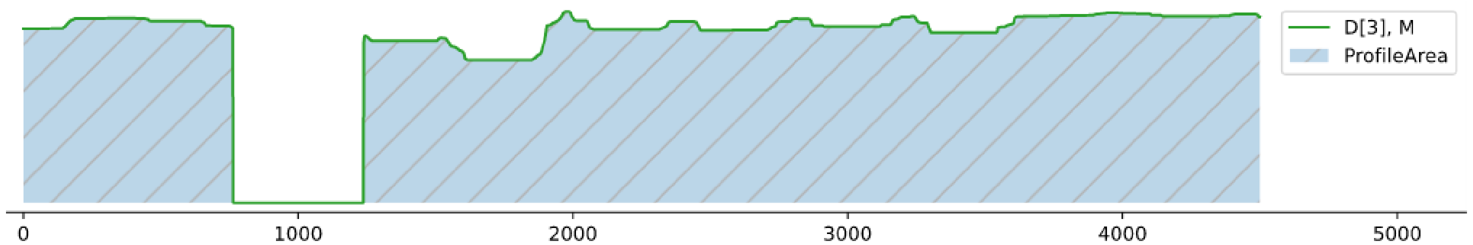
$$\boldsymbol{MPdist_{vect}(Q, T, \ell)} = [v_1, v_2, \dots, v_{n-m+1}], \boldsymbol{v_i} = \mathrm{MPdist}(Q, T_{i,m}, \ell)$$

Discovery of the Top-1 Snippet

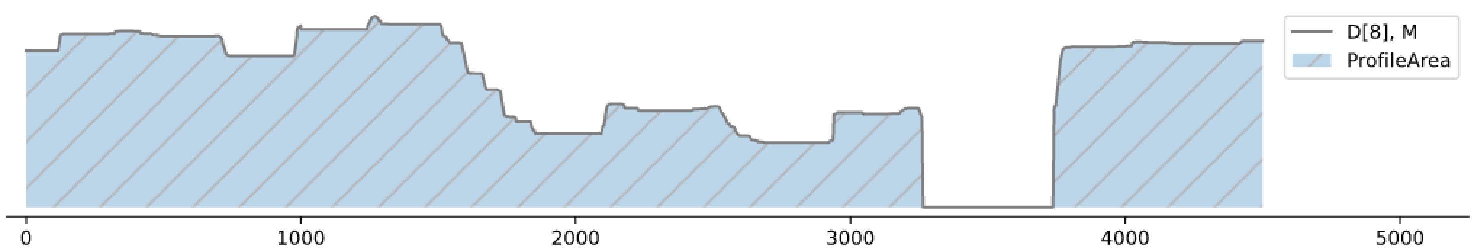
Поиск C_1

i	$ProfileArea$
1	60813
2	60371
3	74451
4	75141
5	56766
6	57729
7	58713
8	53769
9	62127
10	61286

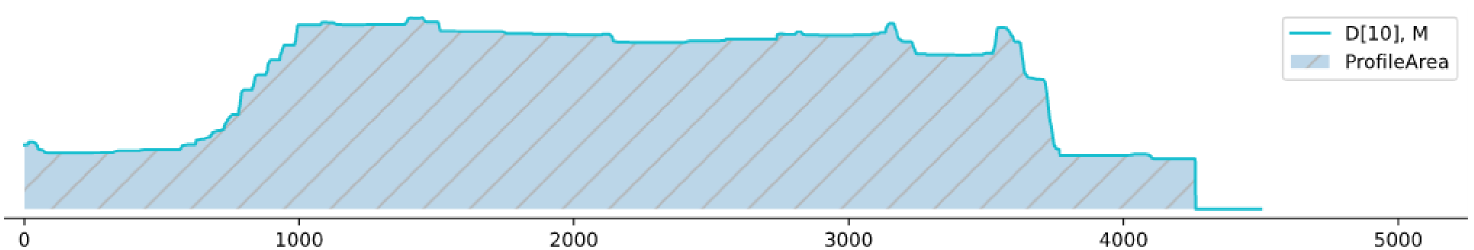
$C_1.index = 8$



$ProfileArea(\{D_3\}) = 74451$

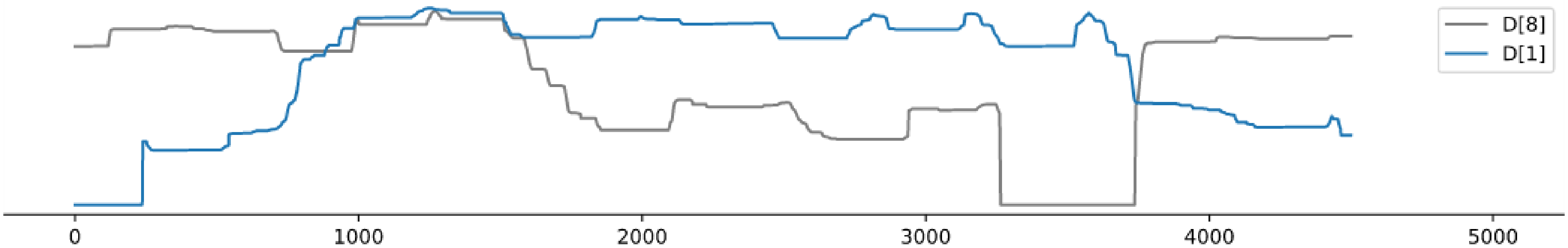


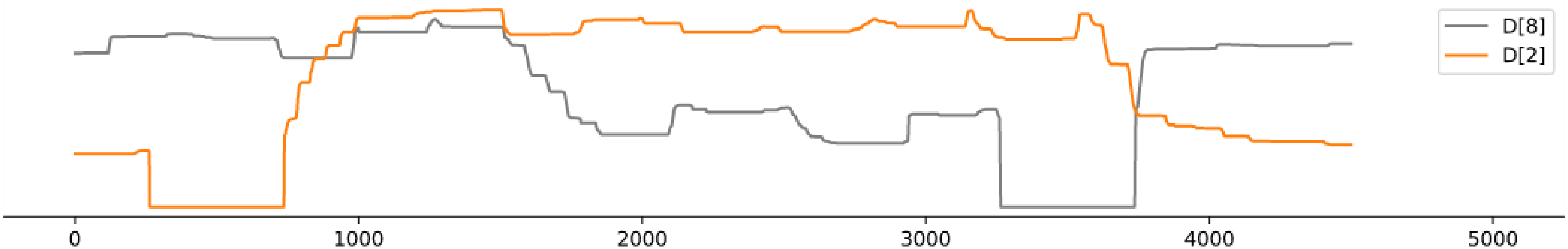
$ProfileArea(\{D_8\}) = 53769$

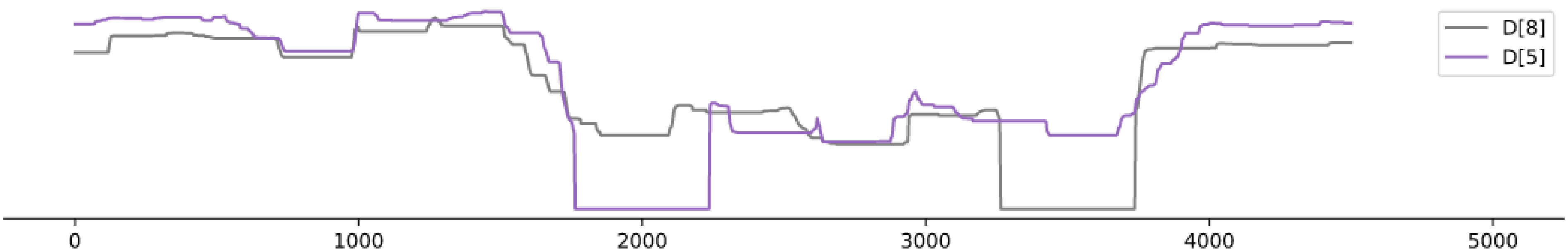


$ProfileArea(\{D_{10}\}) = 61286$

Discovery of the Top-2 Snippet





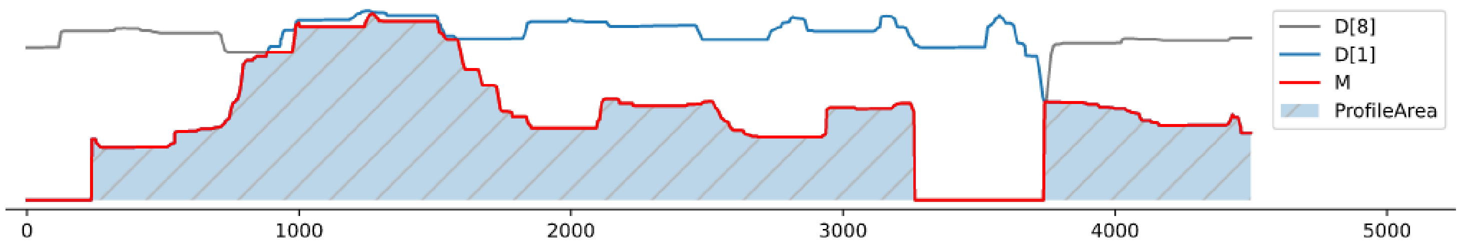


Discovery of the Top-2 Snippet

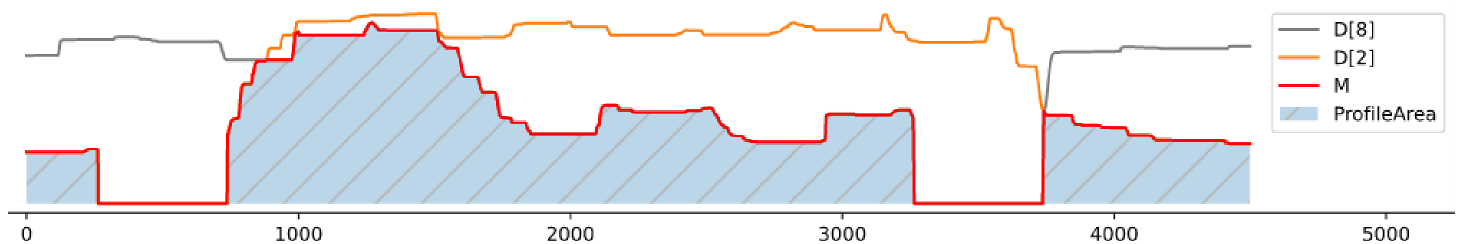
C_2

i	$ProfileArea$
1	38394
2	35769
3	45629
4	45908
5	48857
6	49264
7	48975
9	36684
10	36482

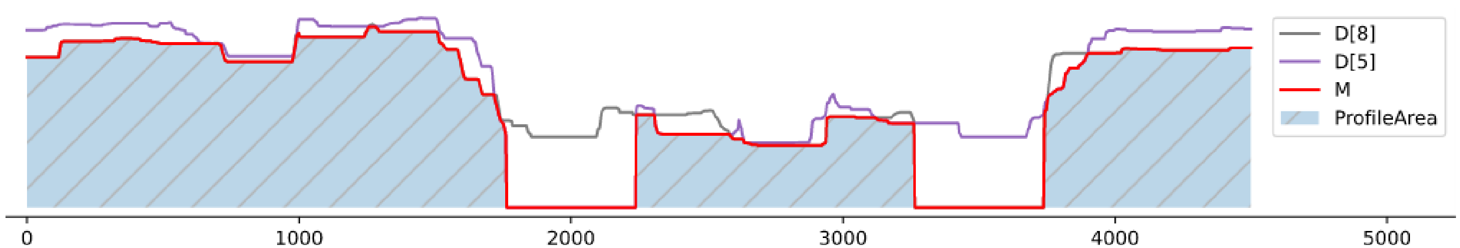
$C_2.index = 2$



$ProfileArea(\{D_8, D_1\}) = 38394$

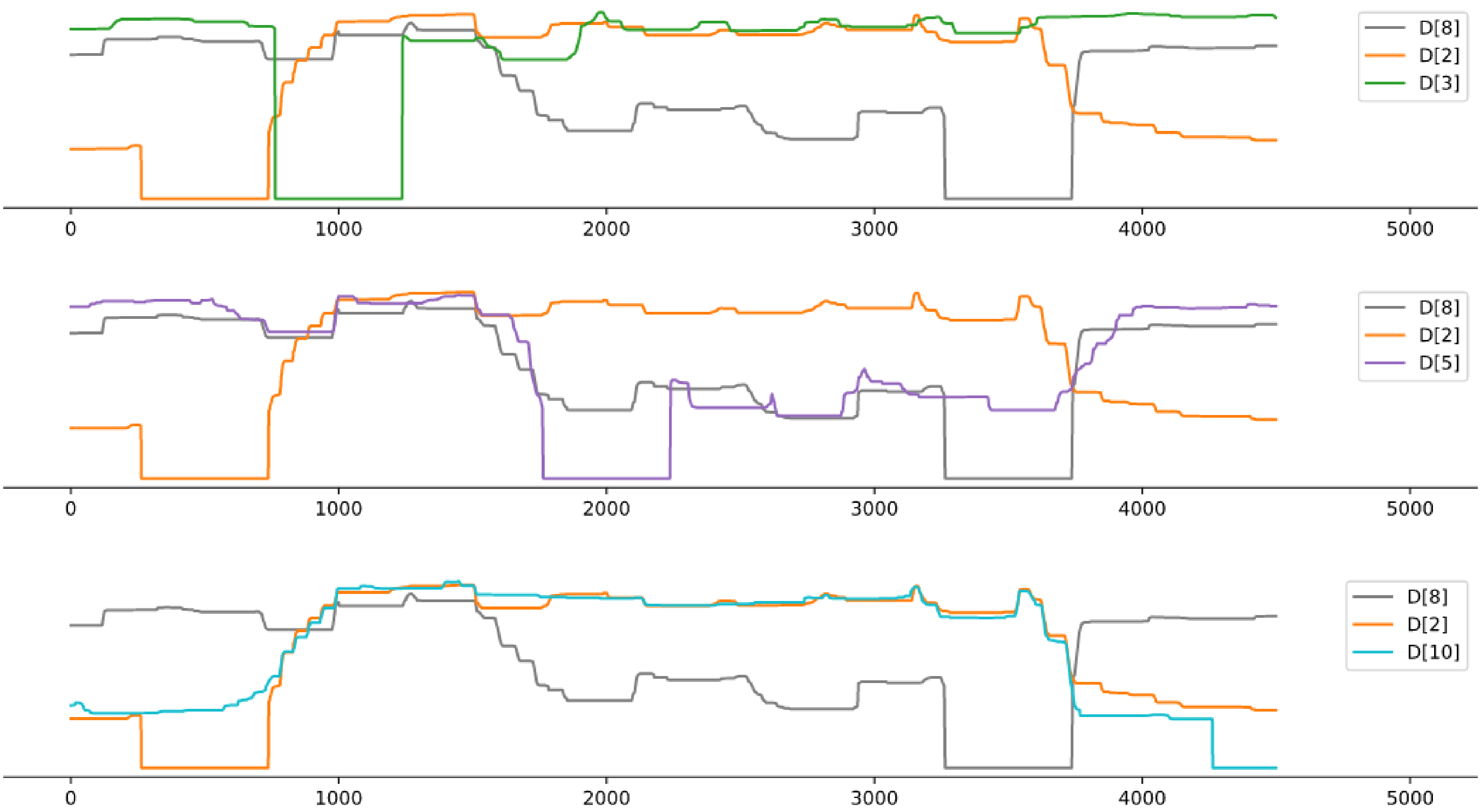


$ProfileArea(\{D_8, D_2\}) = 35769$



$ProfileArea(\{D_8, D_5\}) = 48867$

Discovery of the Top-3 Snippet

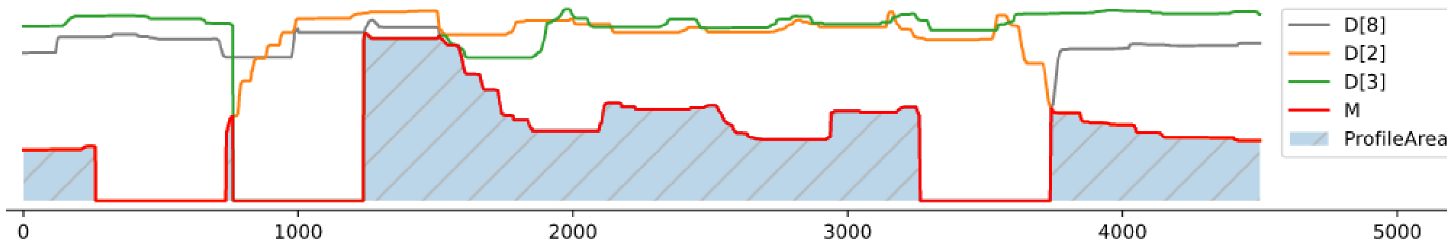


Discovery of the Top-3 Snippet

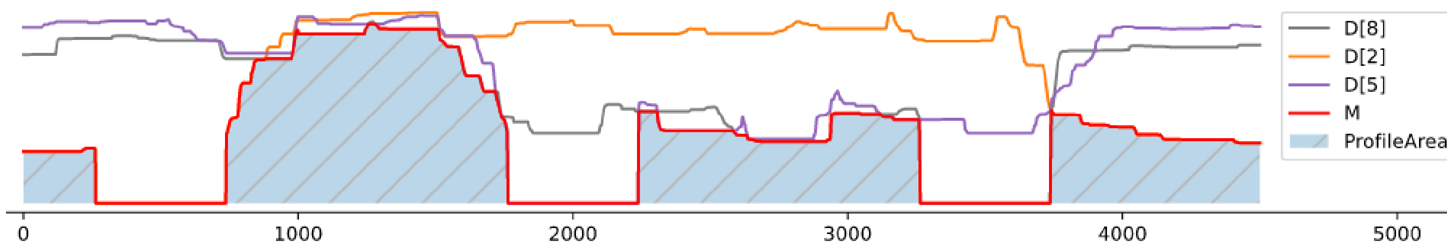
C_3

i	$ProfileArea$
1	34475
3	27899
4	27908
5	31168
6	31532
7	31672
9	31654
10	33044

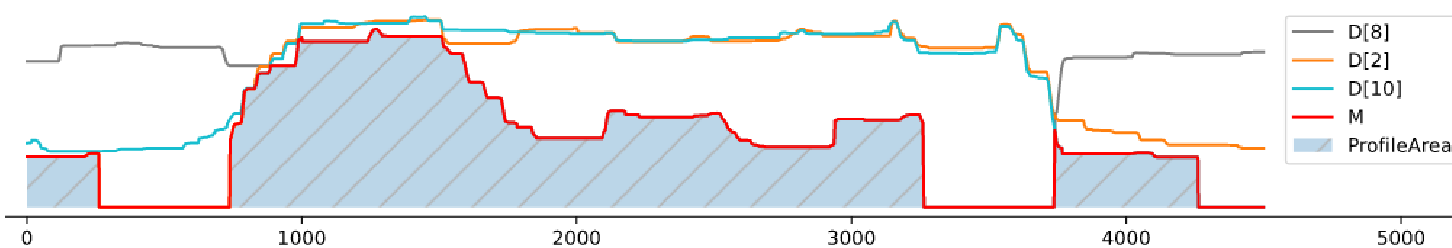
$C_3.index = 3$



$ProfileArea(\{D_8, D_2, D_3\}) = 27899$

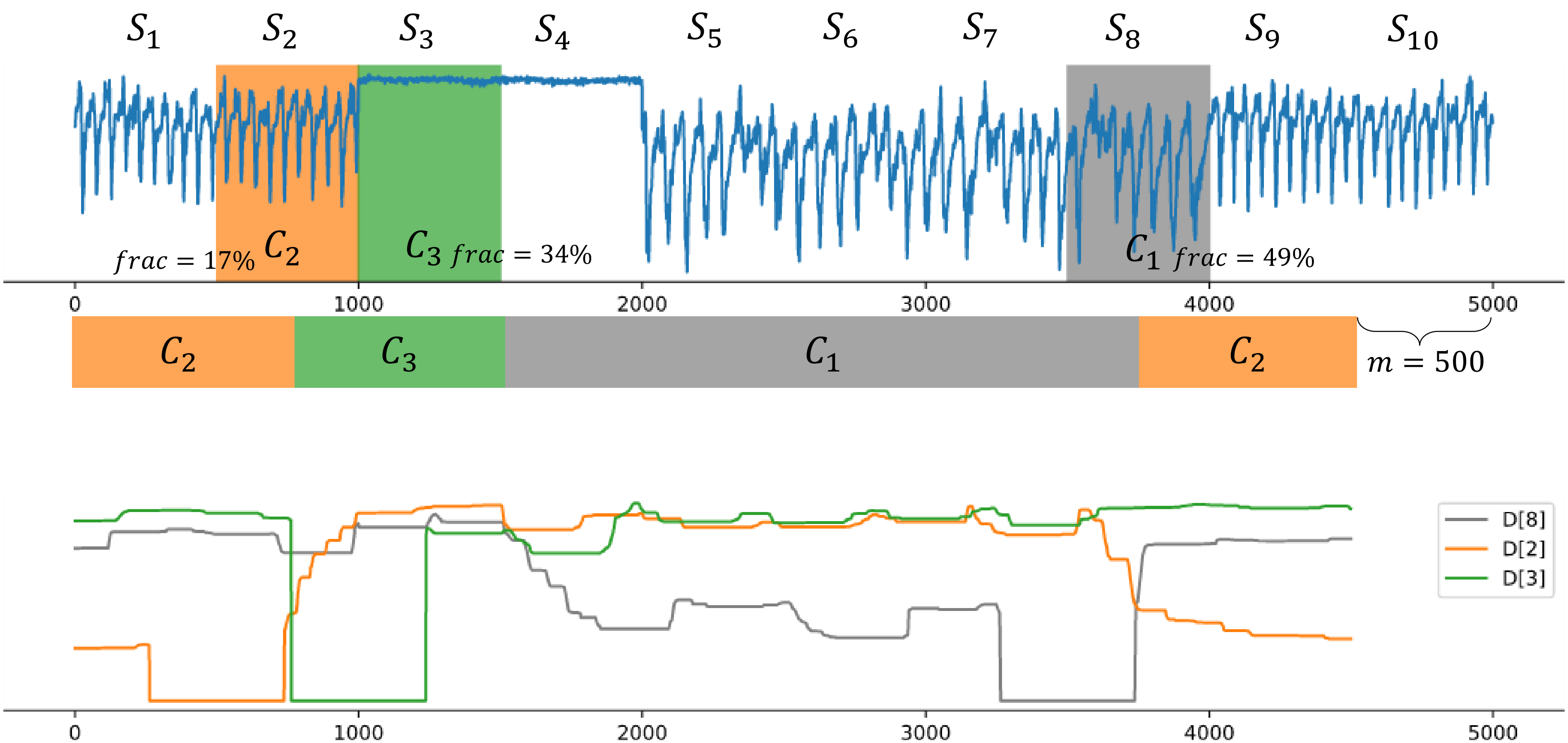


$ProfileArea(\{D_8, D_2, D_5\}) = 31168$

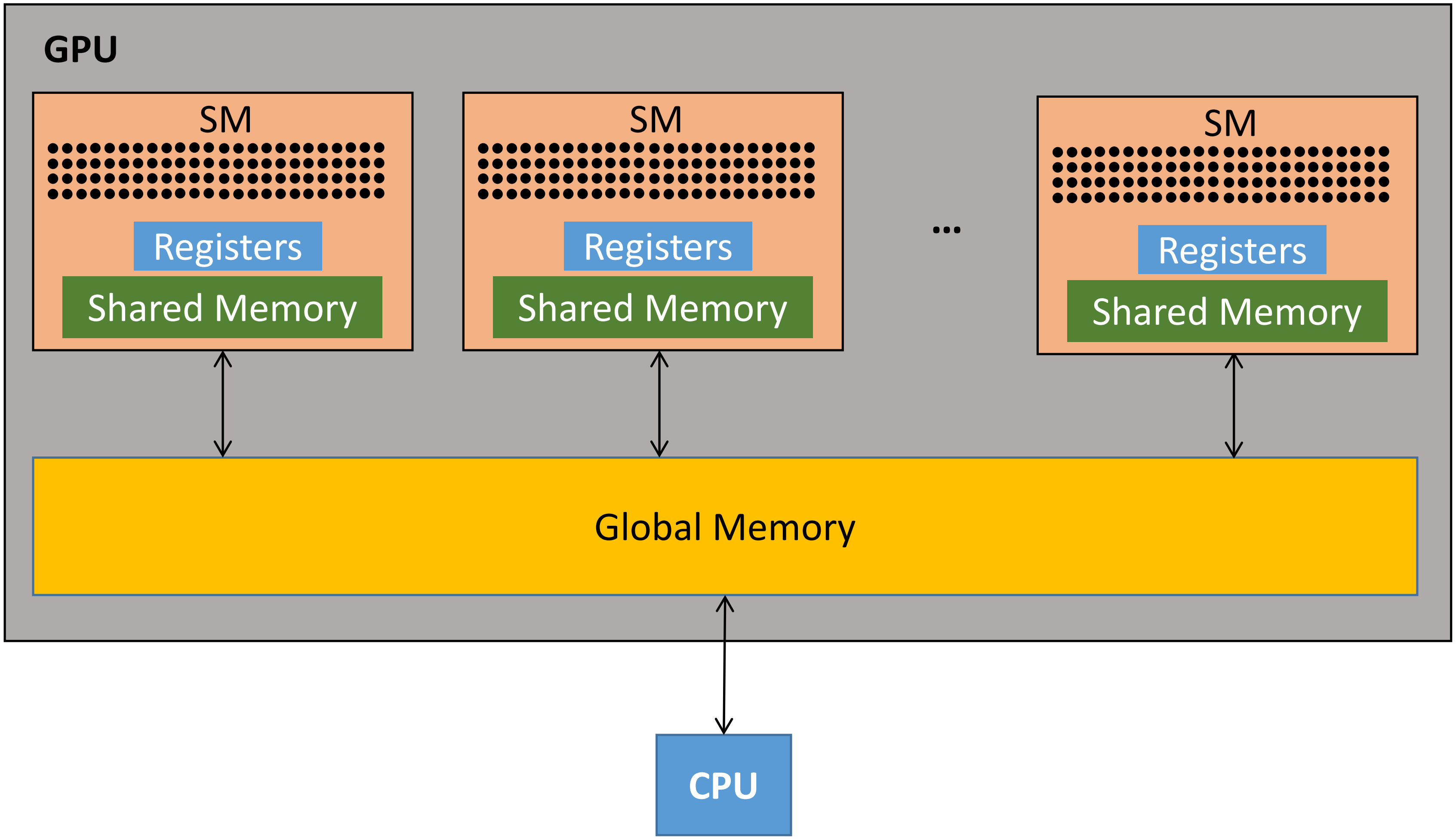


$ProfileArea(\{D_8, D_2, D_{10}\}) = 33044$

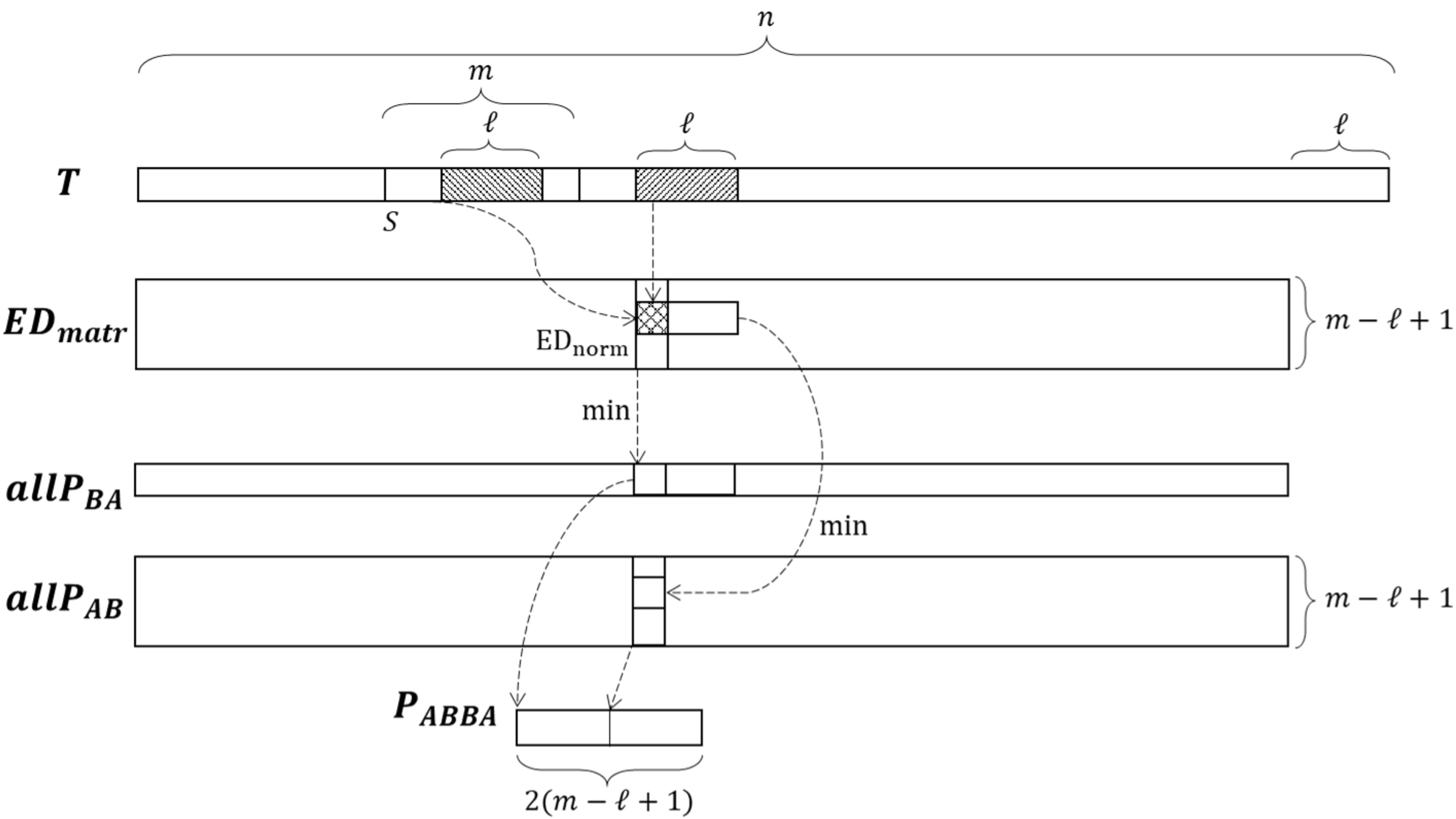
Resulting Snippets



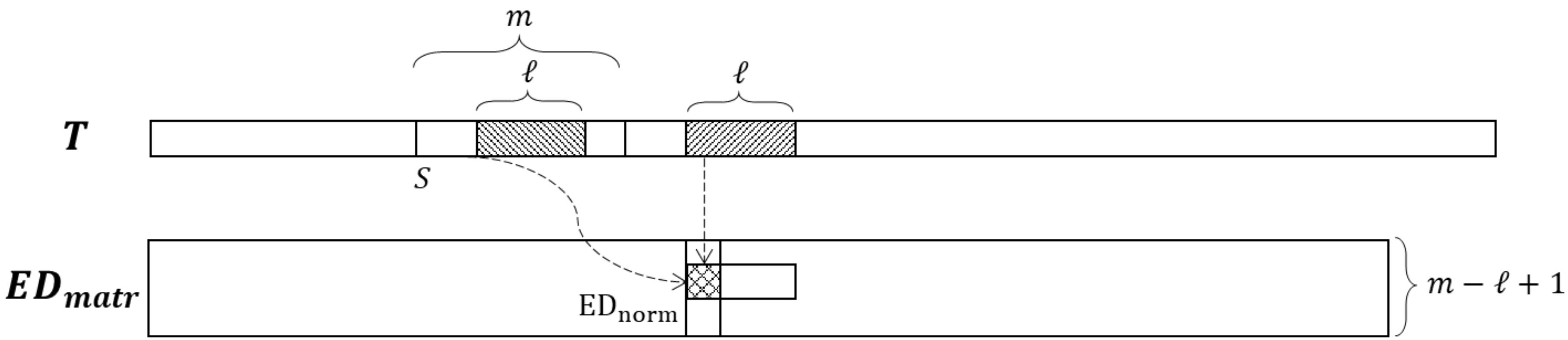
Hardware Architecture



Parallelizing: Data Structures



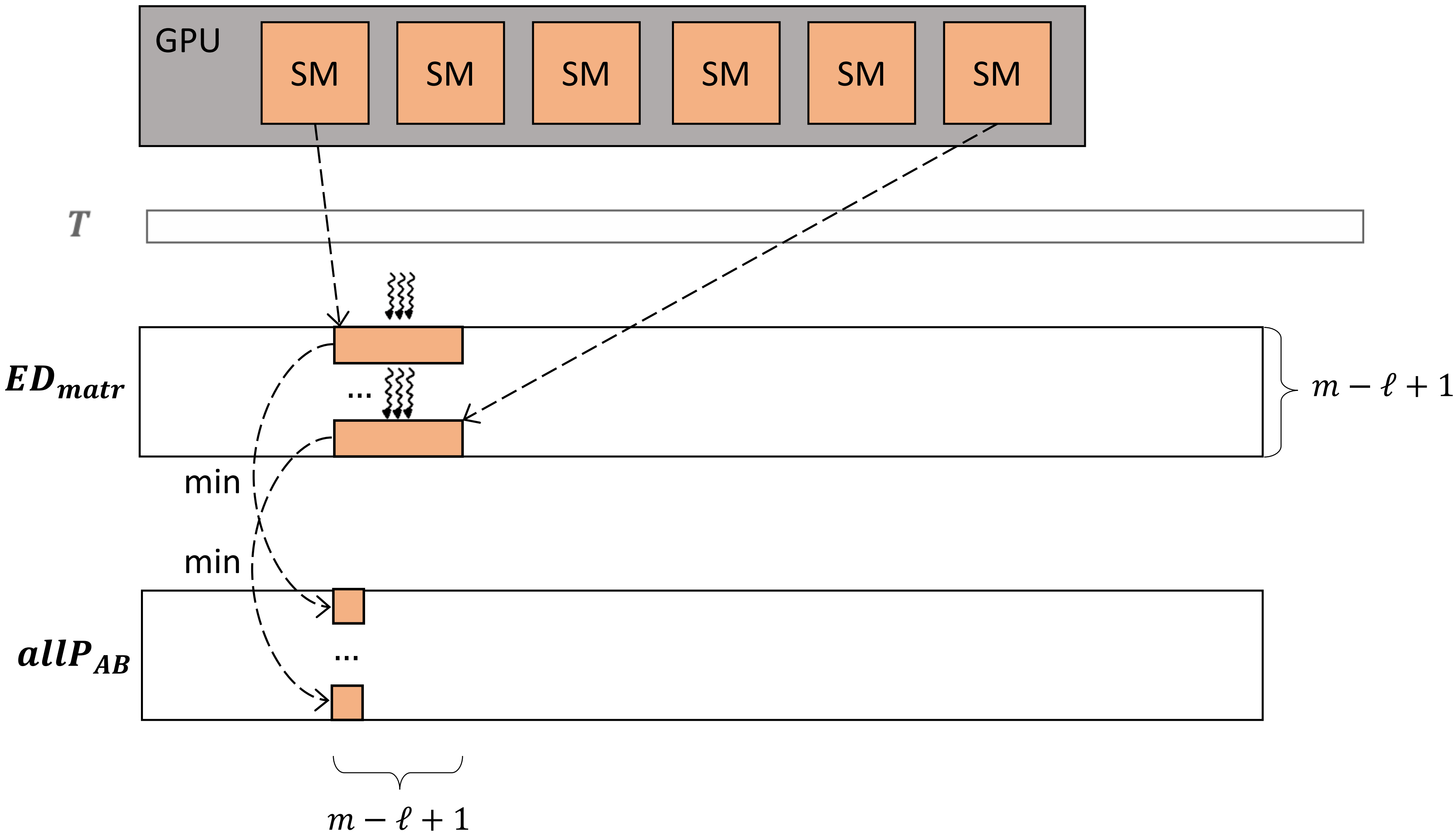
Calculation of $\boldsymbol{ED_{matr}}$



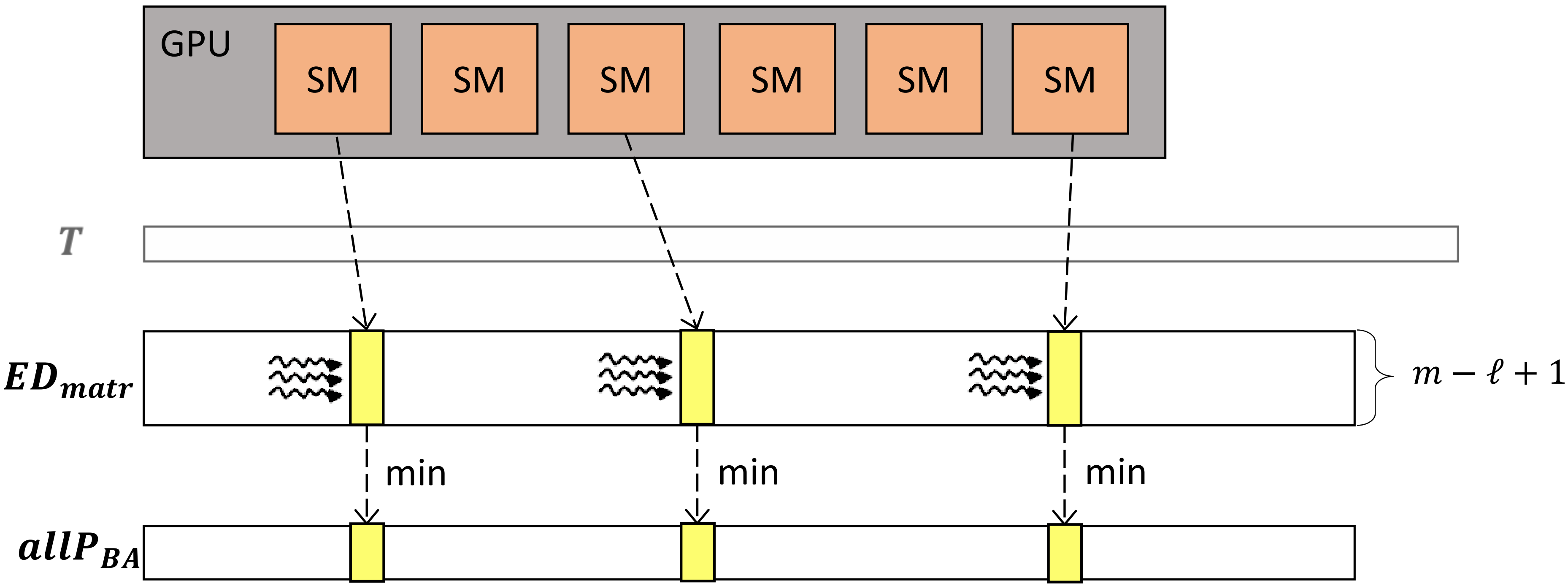
$$\overline{QT}_{i,j} = \overline{QT}_{i-1,j-1} + df_i \cdot dg_j + df_j \cdot dg_i,$$
$$df_0 = 0; df_i = \frac{t_{i+m-1}-t_{i-1}}{2},$$
$$dg_0 = 0; dg_i = (t_{i+m-1} - \mu_i) + (t_{i-1} - \mu_{i-1}),$$
$$\mu_i = \frac{1}{m}\sum_{j=i}^{i+m} t_j,$$
$$T_{i,m} - \mu_i = (t_i - \mu_i, \ldots, t_{i+m-1} - \mu_i),$$
$$P_{i,j} = \overline{QT}_{i,j} \cdot \frac{1}{\|T_{i,m}-\mu_i\|} \cdot \frac{1}{\|T_{j,m}-\mu_j\|},$$
$$\mathrm{ED_{norm}}(T_{i,m}, T_{j,m}) = \sqrt{2m(1 - P_{i,j})}$$

* Zimmerman Z., Kamgar K., Senobari N.S. et al. Matrix Profile XIV: Scaling Time Series Motif Discovery with GPUs to Break a Quintillion Pairwise Comparisons a Day and Beyond. ACM SoCC'2019. DOI: 10.1145/3357223.3-362721.

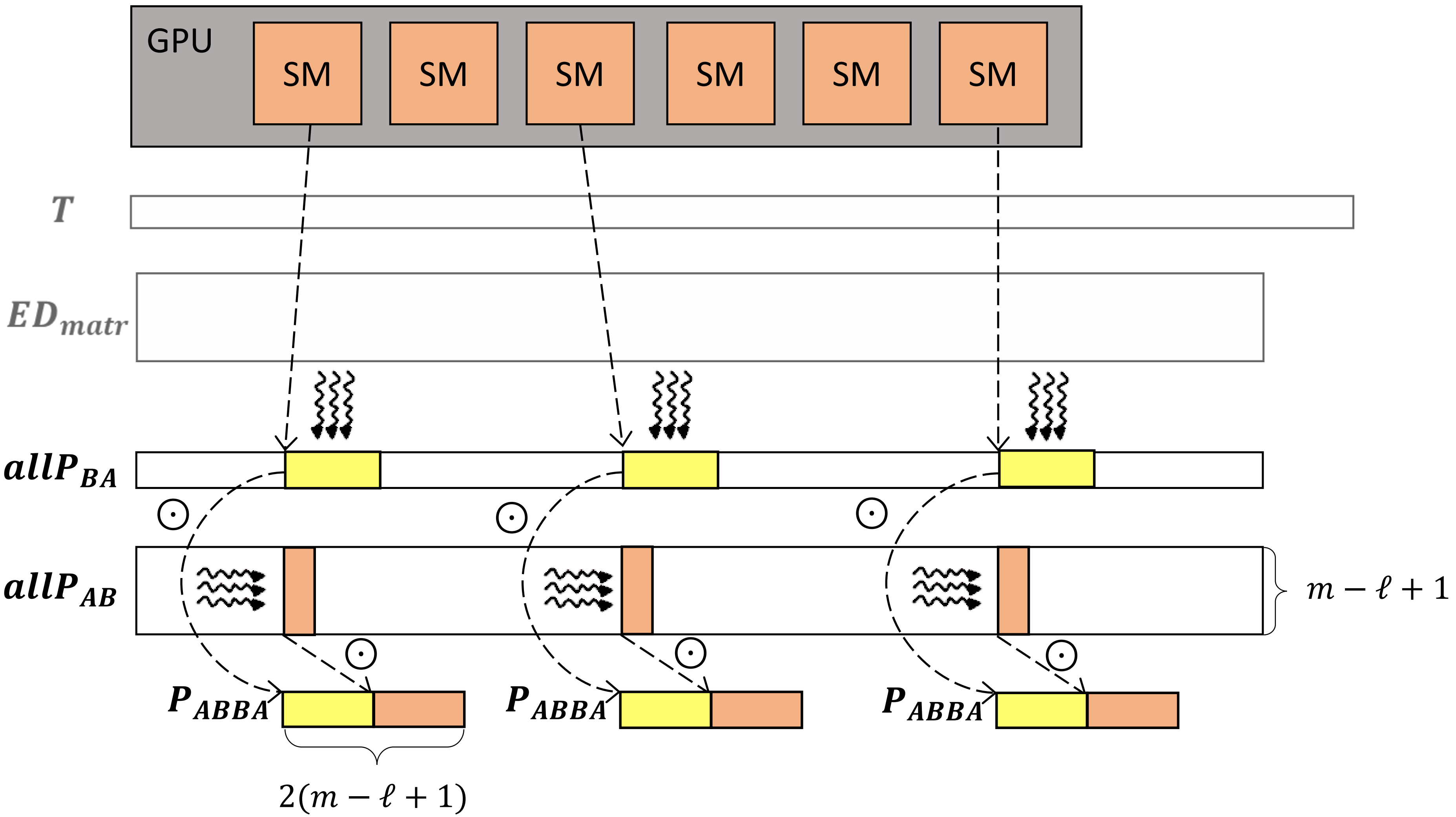
Calculation of $allP_{AB}$



Calculation of $allP_{BA}$



Calculation of P_{ABBA}



Experiments: Hardware

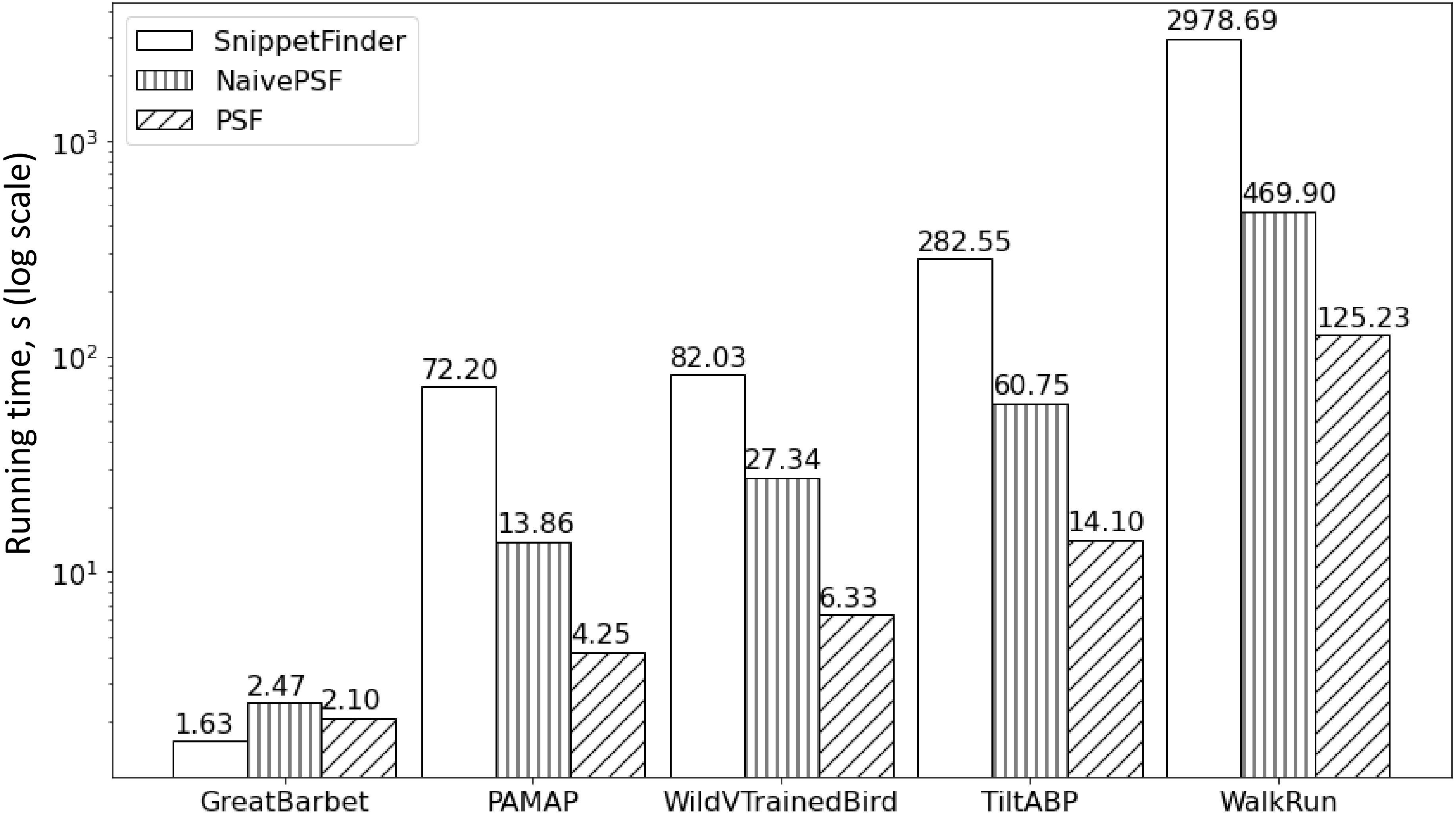
- CPU:
  - Intel Xeon Gold 6254@4 GHz
  - Cores: 18 (but only one was employed)
  - RAM: 64 Gb
  - Peak performance: 1.2 TFLOPS
- GPU:
  - NVIDIA Tesla V100 SXM2
  - Cores: 5120 @1.312 GHz (84 streaming multiprocessors)
  - RAM: 32 Gb
  - Peak performance: 15.7 TFLOPS

Experiments: Data

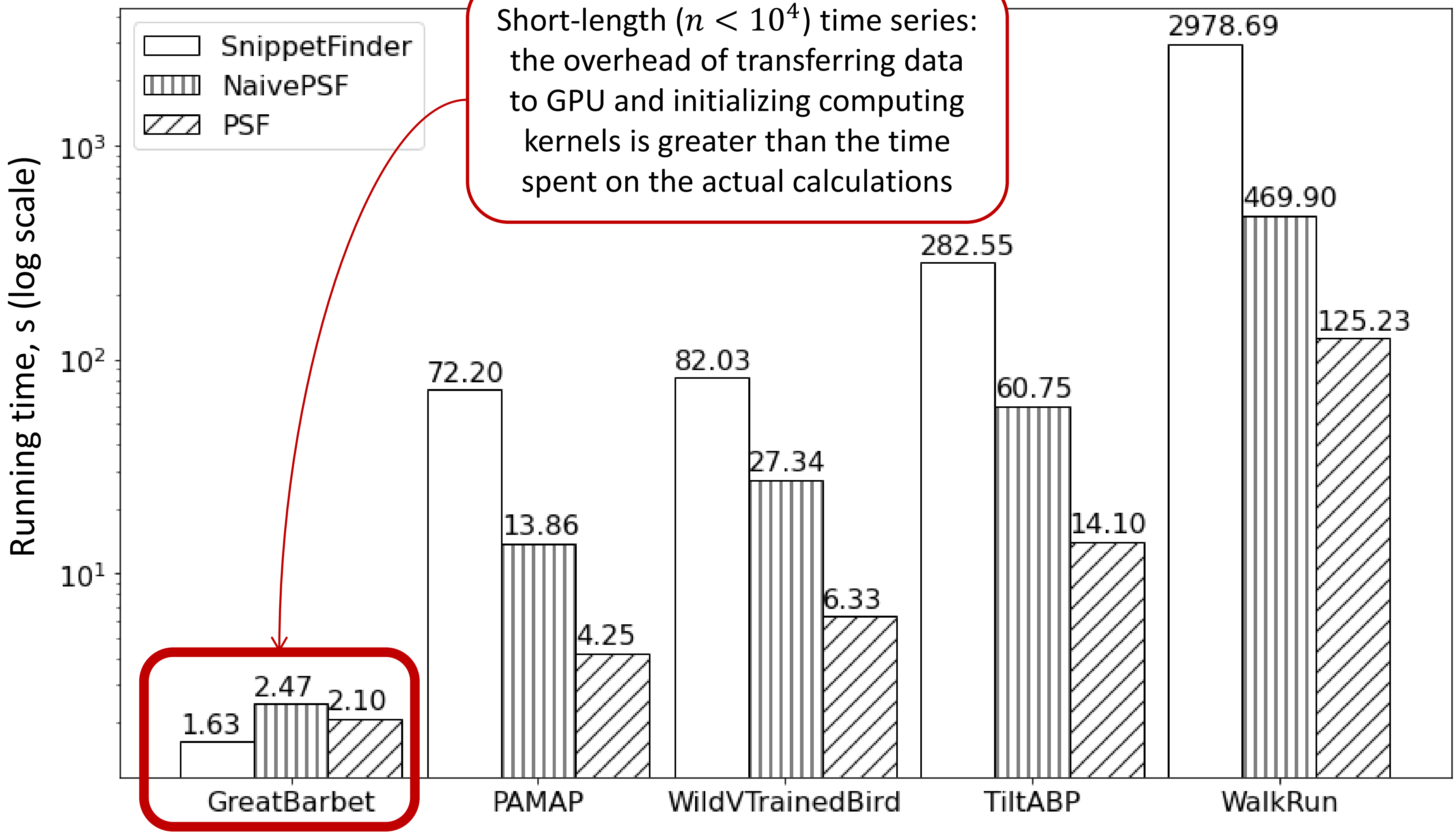
Time series	Length n	Segment length m	Description
GreatBarbet[1]	2 801	150	Physiological indicators of bird vital activity
WildVTrainedBird[1]	20 002	900	
PAMAP[2]	20 002	600	Wearable accelerometer readings during various types of human physical activity
WalkRun[2]	100 000	240	
TiltABP[1]	40 000	630	Human blood pressure readings during rapid tilts

(1) Imani S., Madrid F., Ding W., Crouter S.E., Keogh E.J. Introducing time series snippets: a new primitive for summarizing long time series. Data Min. Knowl. Discov. 34(6): 1713-1743 (2020). doi: 10.1007/s10618-020-00702-y
(2) Reiss A., Stricker D. Introducing a new benchmarked dataset for activity monitoring. ISWC 2012. 108–109. doi: 10.1109/ISWC.2012.13

Experiments: Performance

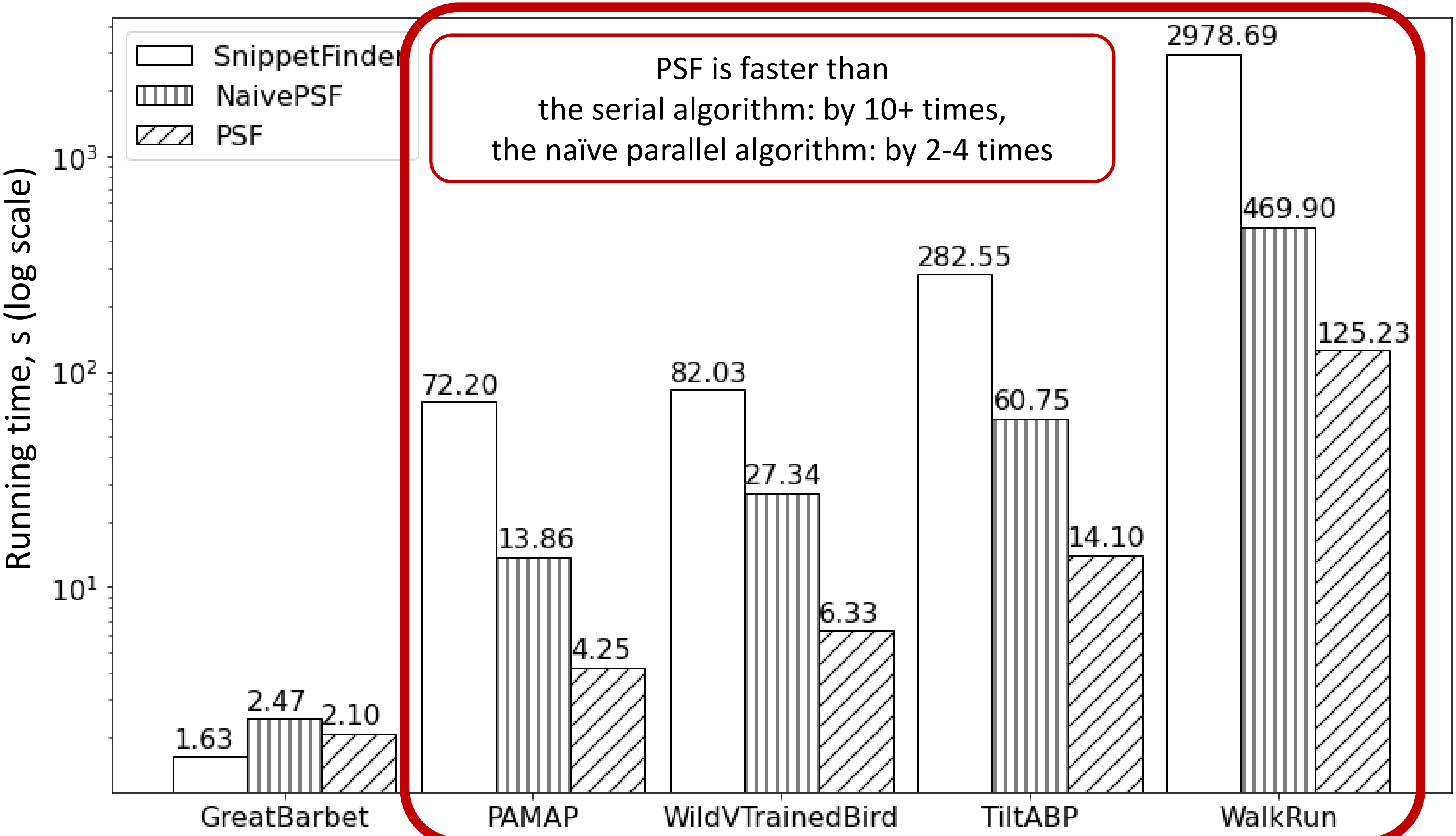


Experiments: Performance



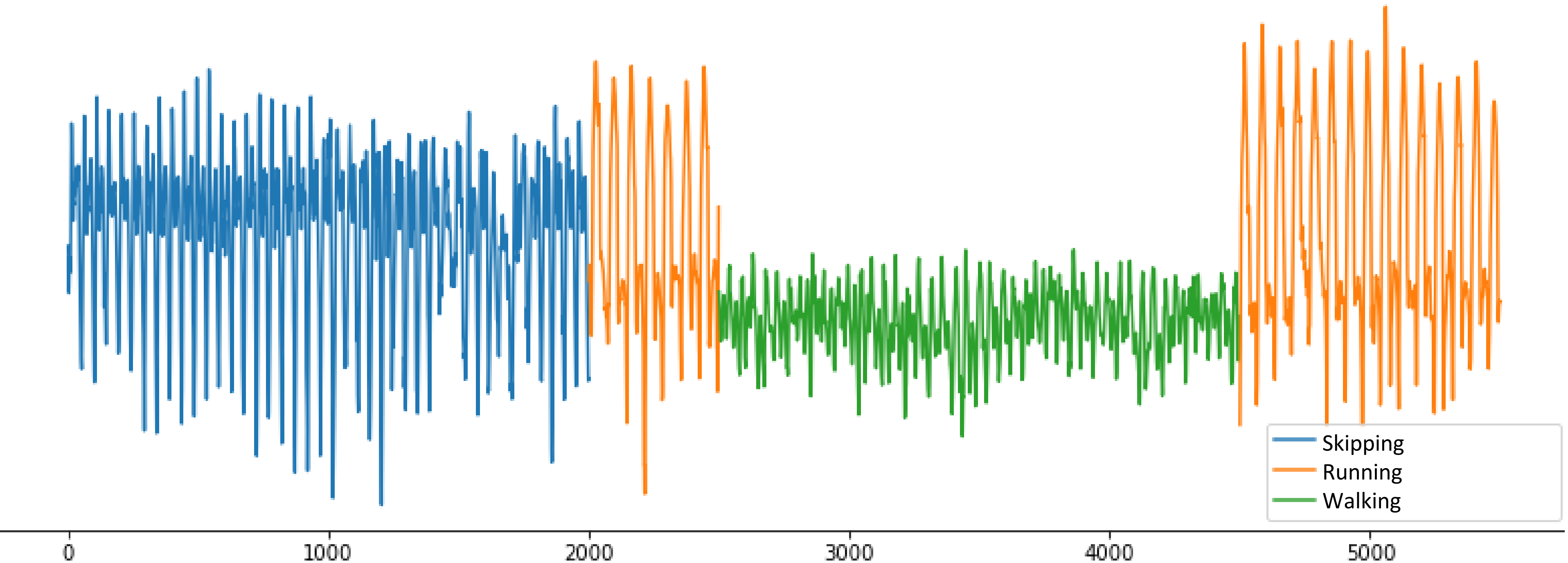
Short-length ($n < 10^4$) time series: the overhead of transferring data to GPU and initializing computing kernels is greater than the time spent on the actual calculations

Experiments: Performance



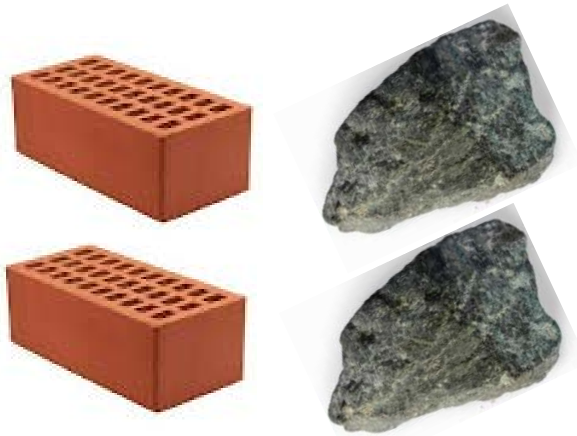
Experiments: Visualization

Time series: PAMAP *

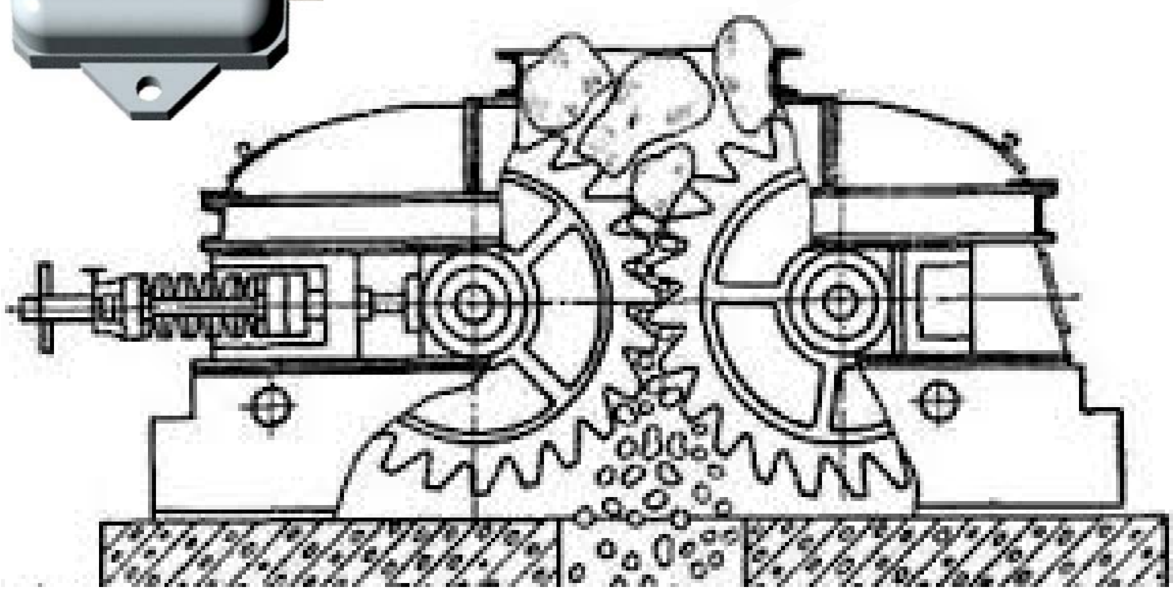


* Reiss A., Stricker D. Introducing a New Benchmarked Dataset for Activity Monitoring, ISWC'2012. DOI: 10.1109/iswc.2012.13

Case Studies: Small-sized Crushing Machine

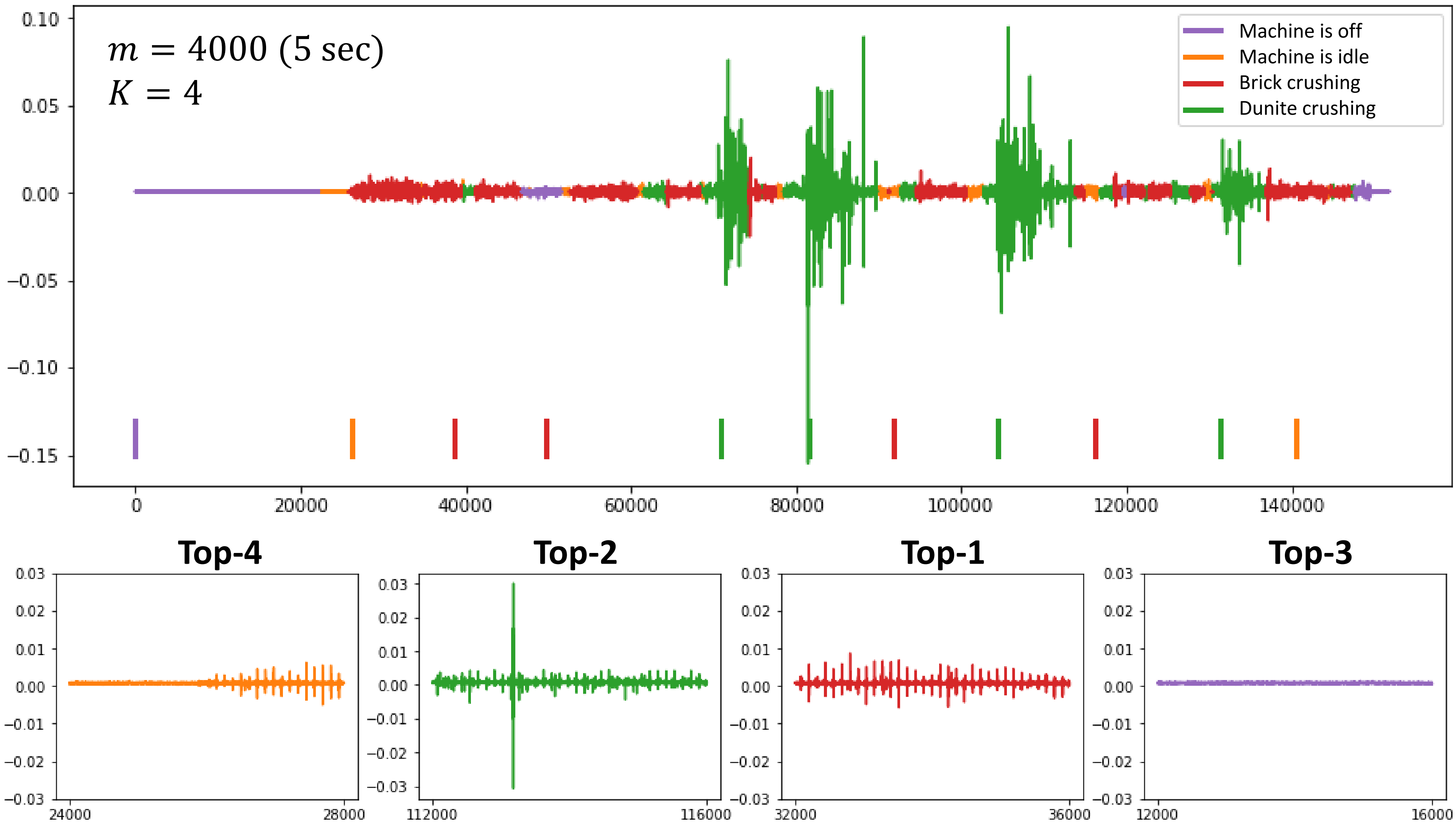


800 Hz



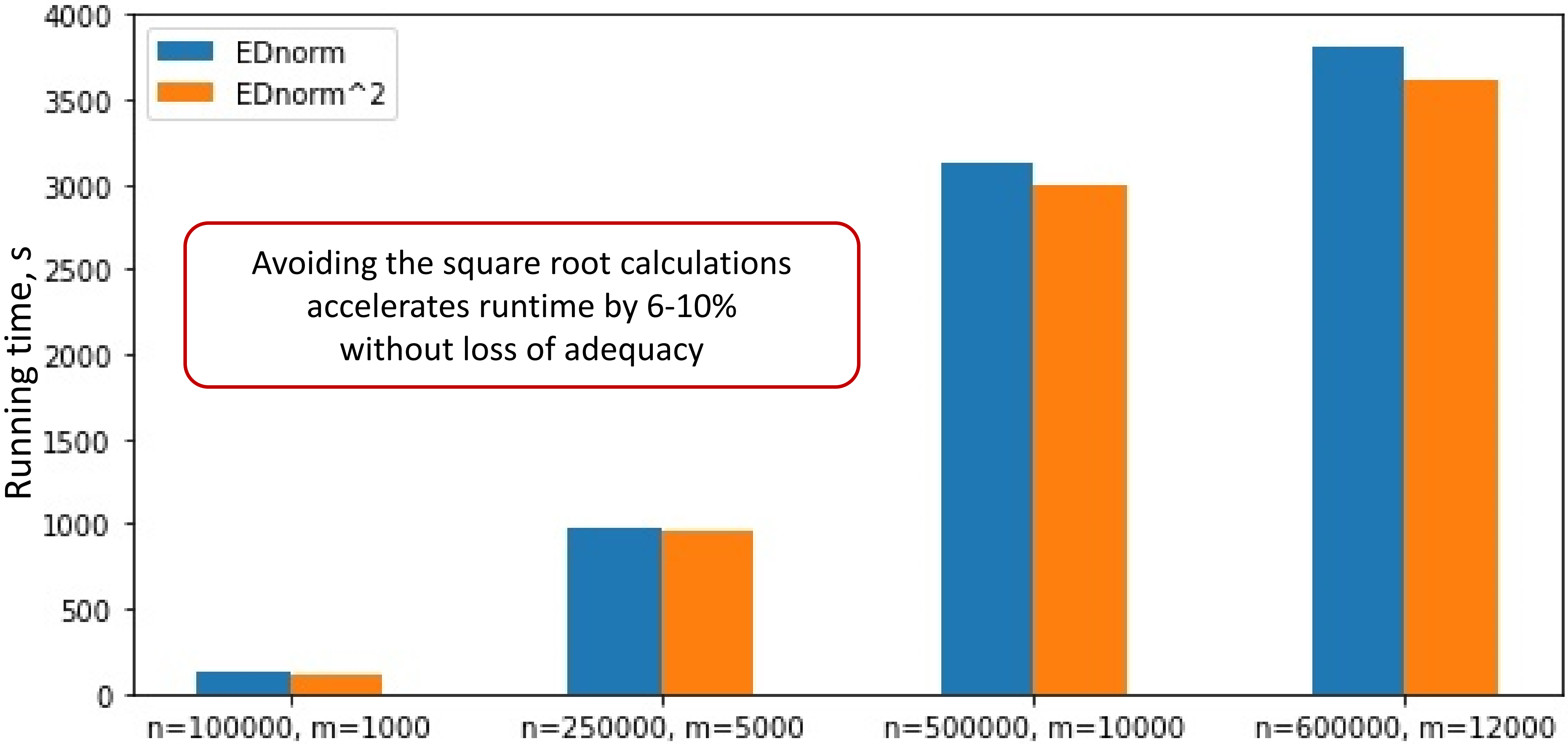
Time	Event
00:00:00	Machine is off
00:32.60	Turning the machine on, machine is idle
00:48.34	Loading with bricks
01:02.09	Loading with bricks
01:28.05	Loading with dunite
01:41.86	Loading with dunite
01:54.68	Loading with bricks
02:10.49	Loading with dunite
02:25.32	Loading with bricks
02:44.20	Loading with dunite
02:55.49	Finishing crushing, machine is idle
03:07.06	Turning the machine off

Case Studies: Small-sized Crushing Machine



Experiments: $\mathrm{ED}^2_{\mathrm{norm}}$ *vs.* $\mathrm{ED}_{\mathrm{norm}}$

Time series: Random Walk *



* Pearson K. The problem of the random walk. Nature. 72(1865), 294 (1905). DOI: 10.1038/072342A0

Conclusions

- PSF (Parallel Snippet Finder) is a novel parallel algorithm to discover snippets of a time series on GPU
- PSF showed high performance in the experiments
- Further study: PSF for HPC-cluster

- Thank you for paying attention! Any questions?
  - Andrey Goglachev, goglachevai@susu.ru