Научный семинар

лаборатории больших данных и машинного обучения

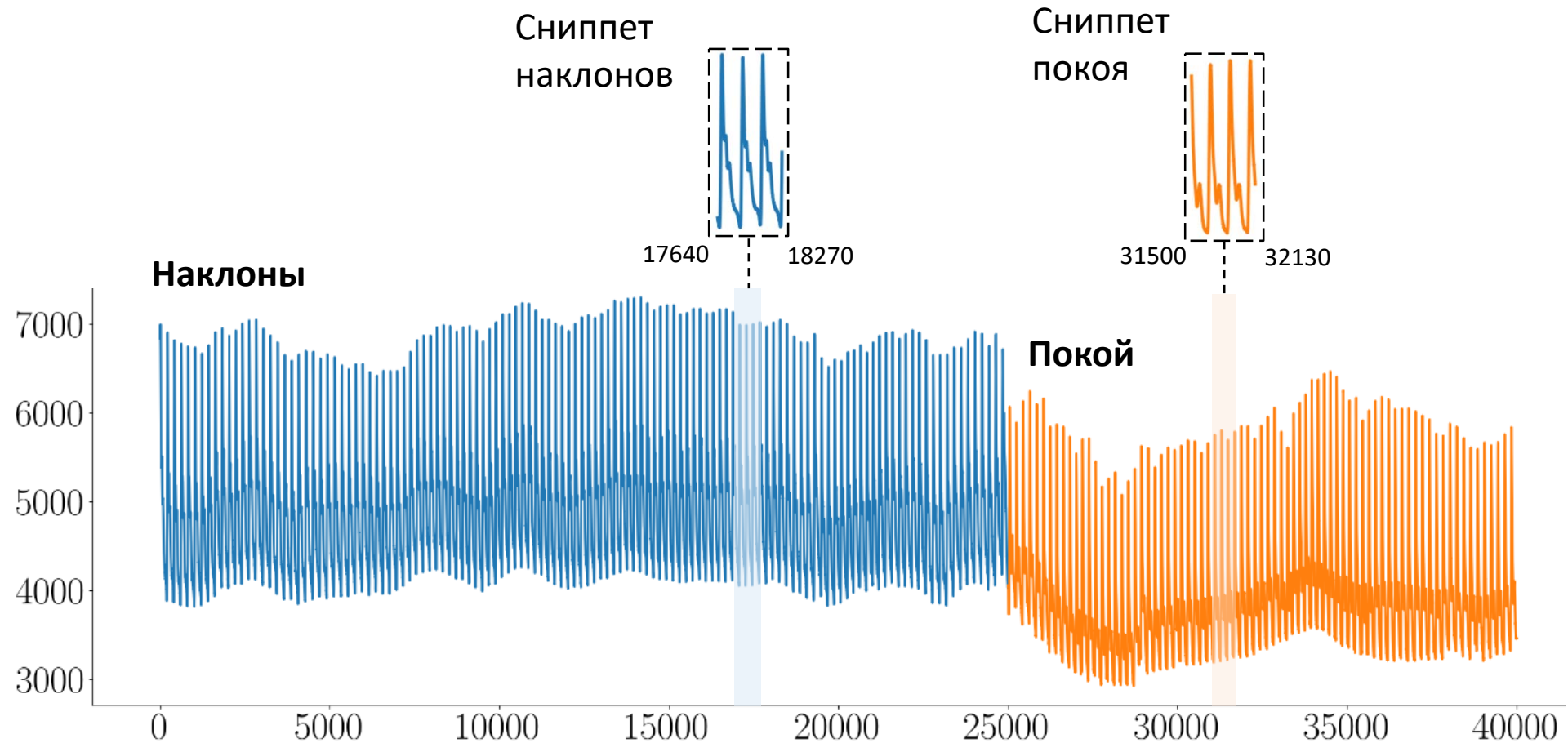
Параллельный алгоритм поиска сниппетов временного ряда для графического процессора

А.И. Гоглачев, М.Л. Цымблер

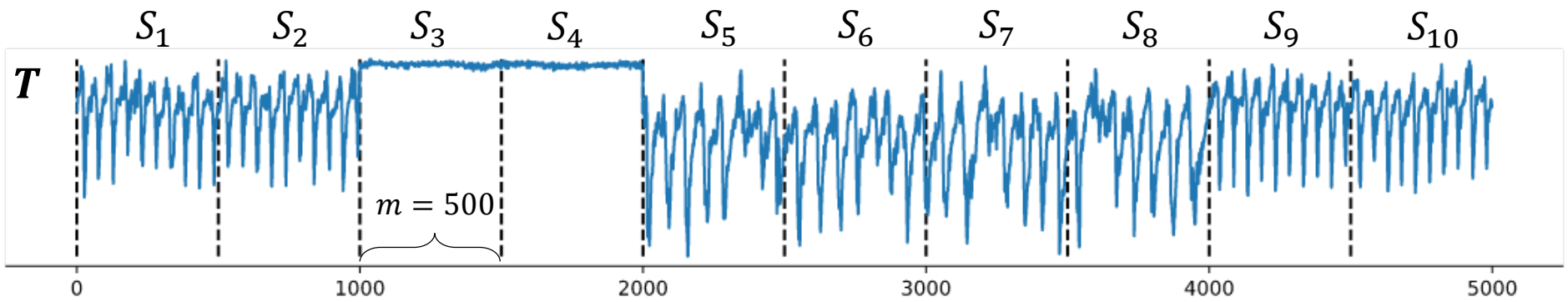
Челябинск–2022

Сниппет: неформальное определение

- подпоследовательность ряда, на которую похожи многие другие подпоследовательности этого ряда



Сниппет: формальное определение*

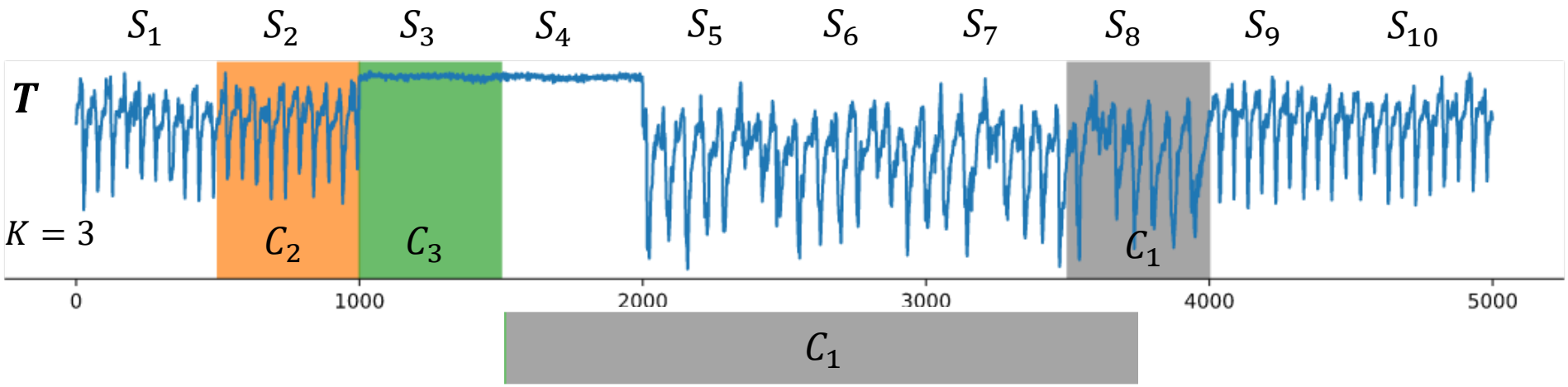


$S_i = T_{m\cdot(i-1)+1,m}, 1 \leq i \leq n/m$

1. Представим ряд как набор непересекающихся сегментов длины n/m
   - если n не кратно m, дополним ряд нулями справа

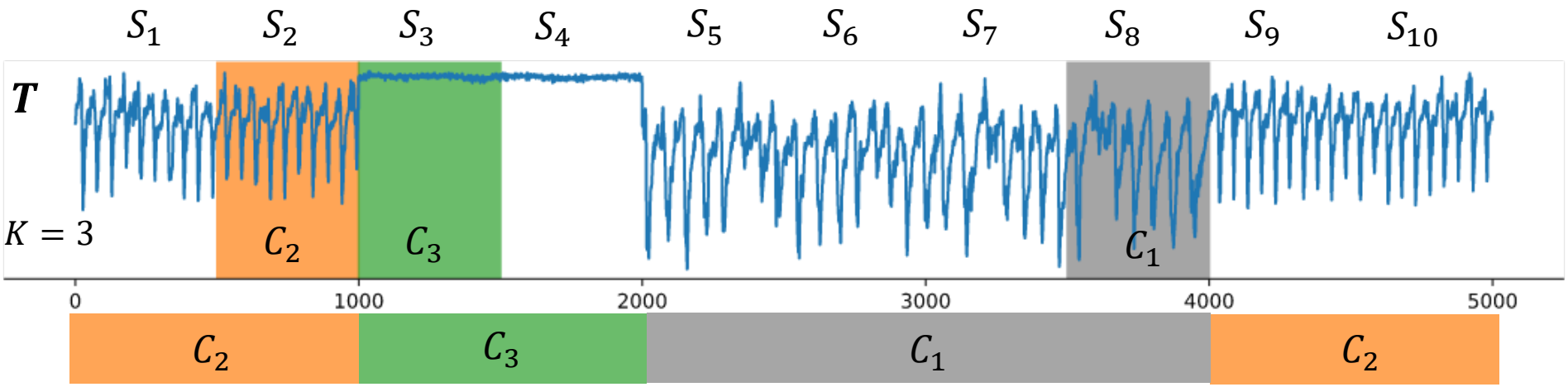
* Imani S., Madrid F., Ding W., Crouter S.E., Keogh E.J. Introducing time series snippets: a new primitive for summarizing long time series. Data Min. Knowl. Discov. 34(6): 1713-1743 (2020). doi: 10.1007/s10618-020-00702-y

Сниппет: формальное определение

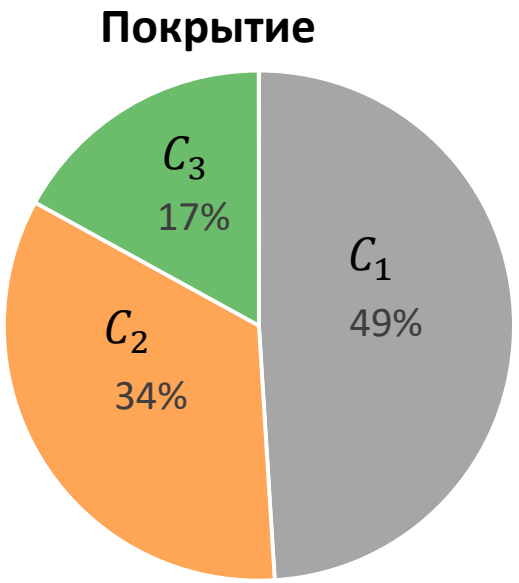


1. Представим ряд как набор непересекающихся сегментов длины n/m
2. Для каждого сегмента найдем наиболее похожие на него подпоследовательности (**ближайшие соседи**)

Сниппет: формальное определение

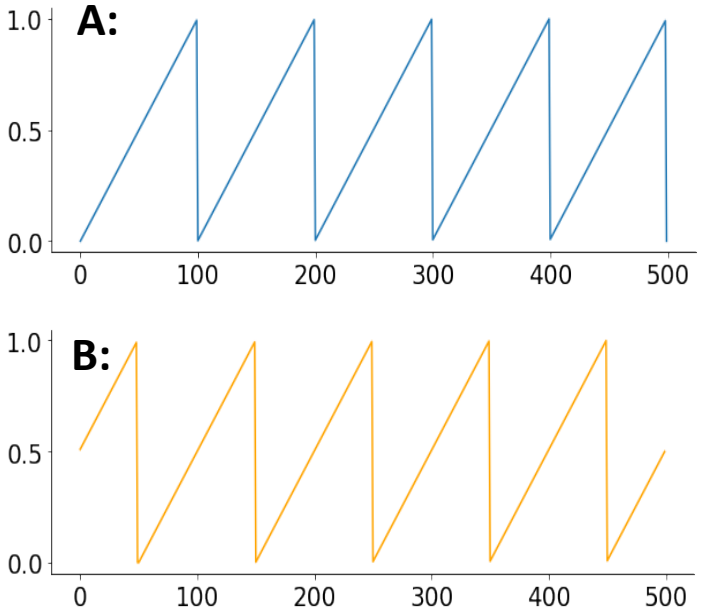


1. Представим ряд как набор непересекающихся сегментов
2. Для каждого сегмента найдем его ближайших соседей
3. Возьмем **top-K** сниппетов по убыванию их **покрытия** (доля мощности ближайших соседей)



Сниппет: мера схожести MPdist*

Два ряда длины m тем более похожи друг на друга в смысле **меры MPdist**, чем больше в них имеется подпоследовательностей меньшей длины ℓ, близких друг к другу в смысле **нормализованного евклидова расстояния**

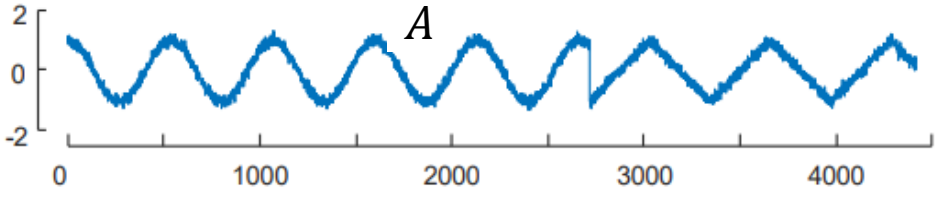


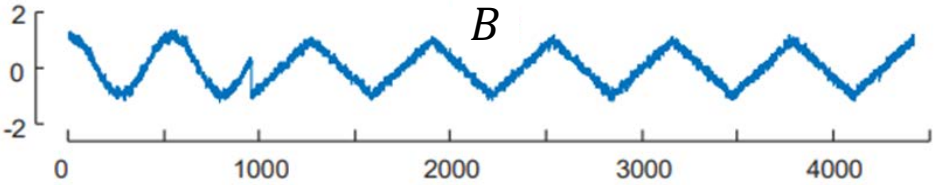
$\mathrm{ED}(A, B)$	**MPdist**(A, B)
11.2	**0**

* Gharghabi S., Imani S., Bagnall A.J. et al. An ultra-fast time series distance measure to allow data mining in more complex real-world deployments // Data Mining and Knowledge Discovery. 2020. Vol. 34. P. 1104–1135. DOI: 10.1007/s10618-020-00695-8

MPdist: формальное определение

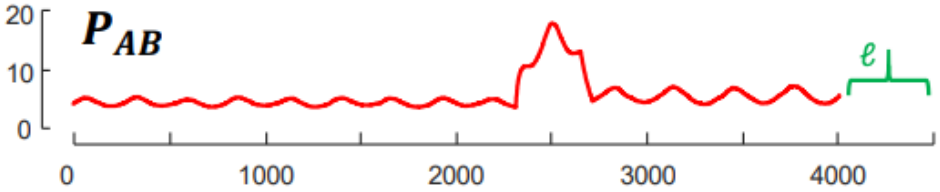
$$|A| = |B| = m$$





Значимая подпоследовательность : $3 \leq \ell \leq m$ (обычно: $\lceil 0.3m \rceil < \ell \leq \lceil 0.8m \rceil$)

Вычисление матричного профиля рядов



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

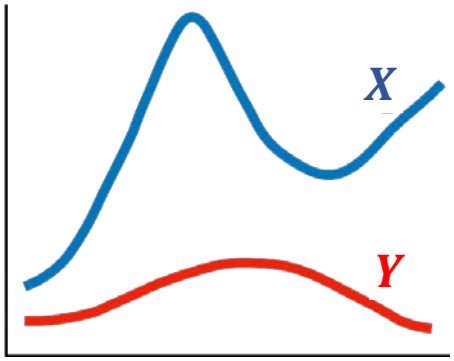
$$B_{j,\ell} = \arg \min_{1 \leq q \leq m-\ell+1} \mathrm{ED}_{\mathrm{norm}}(A_{i,\ell}, B_{q,\ell})$$

Нормализованное евклидово расстояние

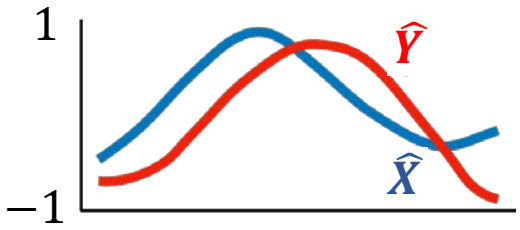
- Обеспечивает корректное сравнение подпоследовательностей с различной амплитудой

$$\mathrm{ED}_{\mathrm{norm}}(X,Y) = \mathrm{ED}(\hat{X},\hat{Y}) = \sqrt{\sum_{i=1}^{\ell}(\hat{x}_i - \hat{y}_i)^2},$$

$$\hat{x}_i = \frac{x_i - \mu_x}{\sigma_x}, \qquad \mu_x = \frac{1}{\ell}\sum_{i=1}^{\ell} x_i, \qquad \sigma_x^2 = \frac{1}{\ell}\sum_{i=1}^{\ell} x_i^2 - \mu_x^2$$

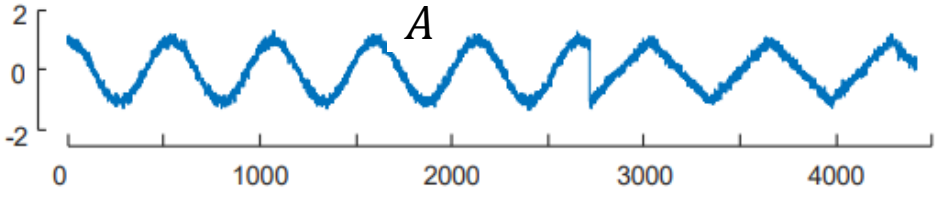


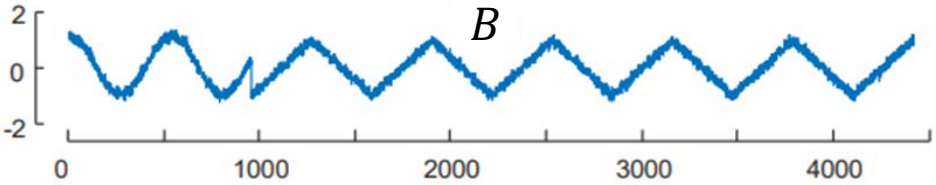
Z-нормализация



MPdist: формальное определение

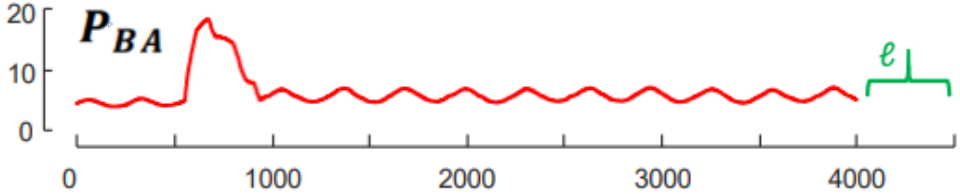
$$|A| = |B| = m$$





Значимая подпоследовательность : $3 \le \ell \le m$ (обычно: $\lceil 0.3m \rceil < \ell \le \lceil 0.8m \rceil$)

Вычисление матричного профиля рядов

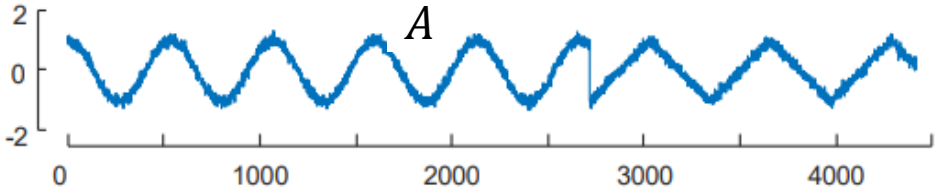


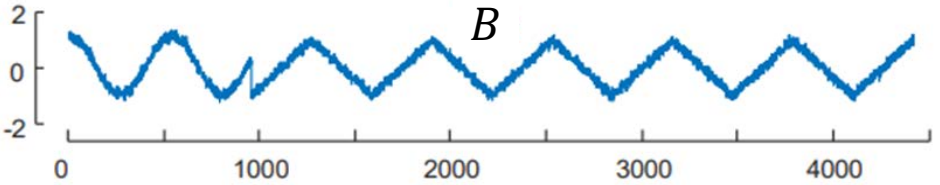
$$\{P_{AB}(i) = \mathrm{ED}_{\mathrm{norm}}(A_{i,\ell}, B_{j,\ell})\}_{i=1}^{m-\ell+1},$$
$$B_{j,\ell} = \arg \min_{1 \le q \le m-\ell+1} \mathrm{ED}_{\mathrm{norm}}(A_{i,\ell}, B_{q,\ell})$$

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

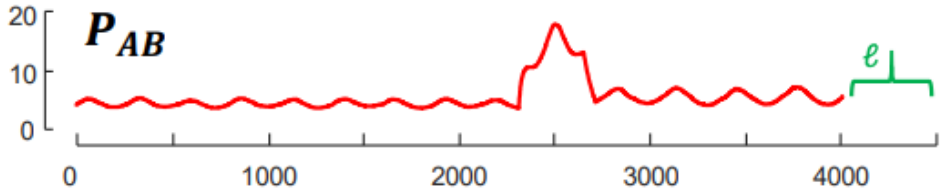
MPdist: формальное определение

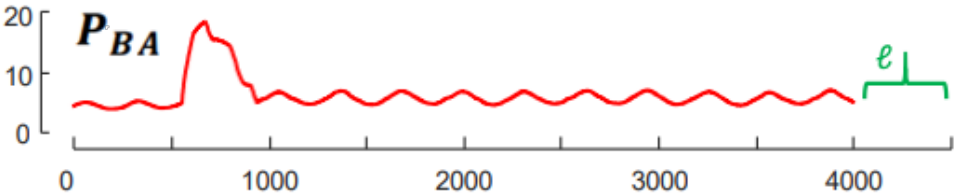
$$|A| = |B| = m$$





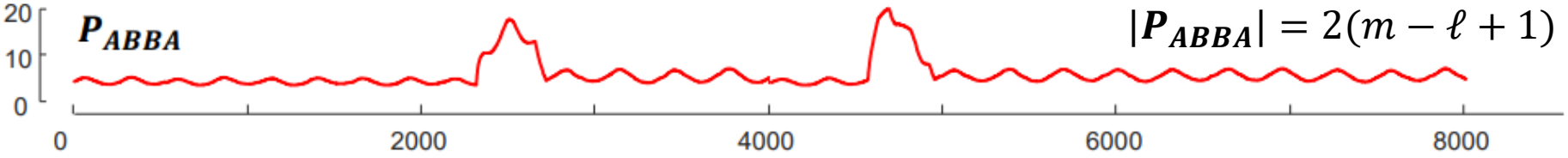
$$3 \leq \ell \leq m$$





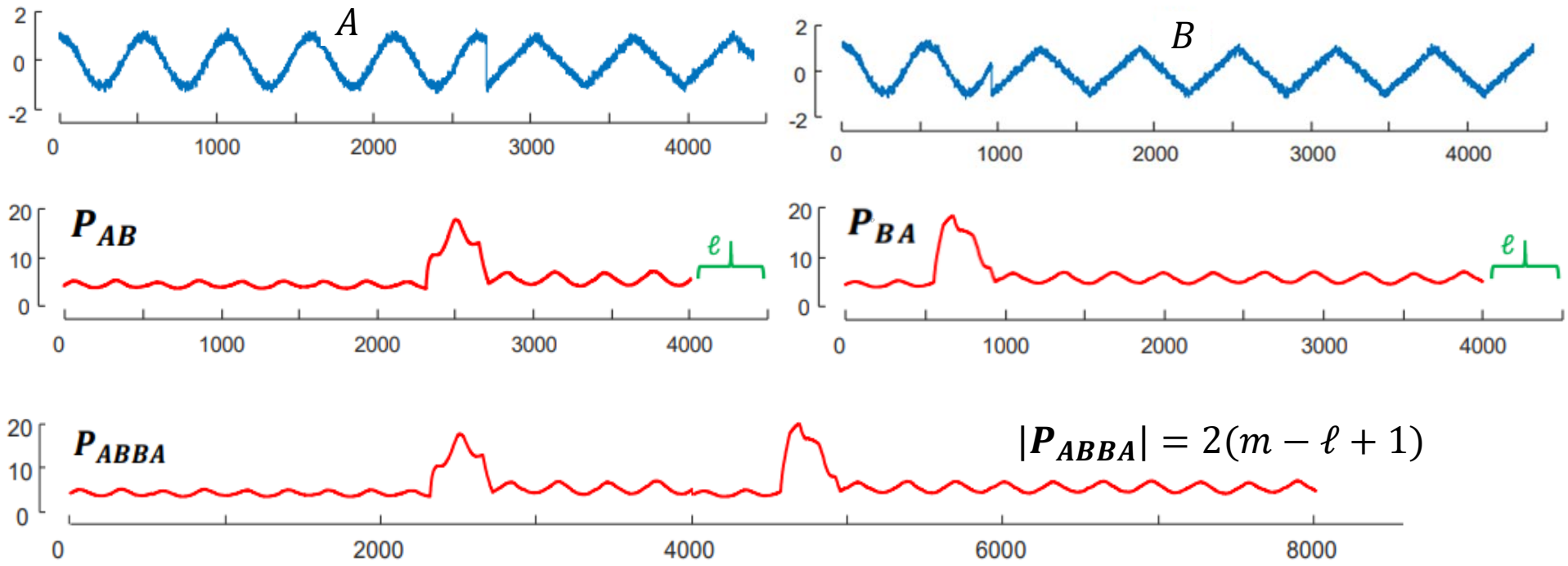
Конкатенация матричных профилей рядов

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



MPdist: формальное определение

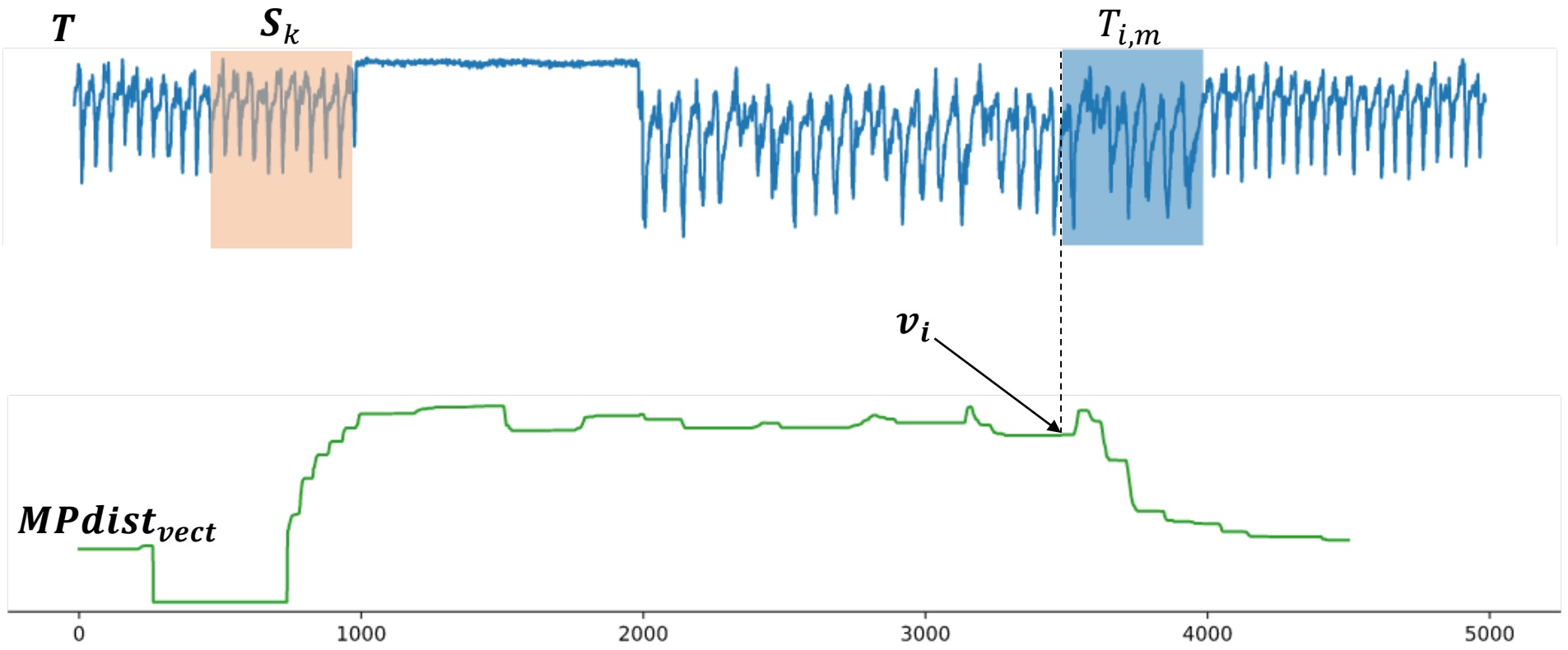
$|A| = |B| = m$ $3 < \ell \leq m$



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

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

MPdist профиль сегмента



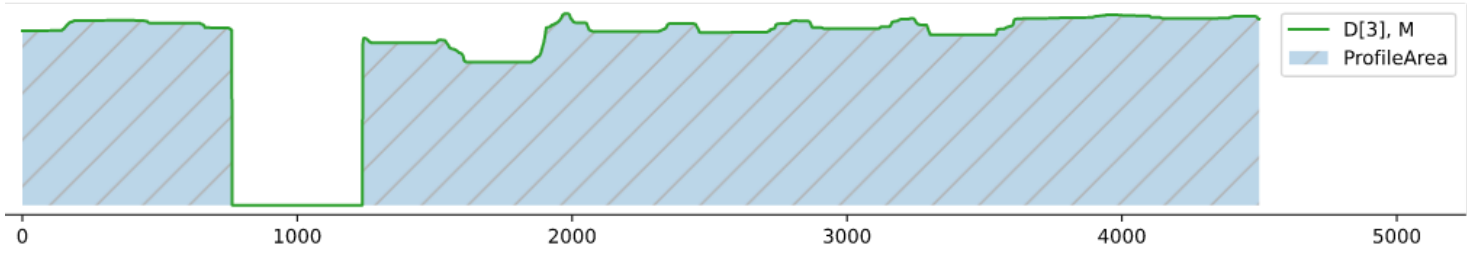
$$MPdist_{vect}(S_k, T, \ell) = [v_1, v_2, \ldots, v_{n-m+1}], v_i = \mathrm{MPdist}(S_k, T_{i,m}, \ell)$$

Поиск сниппета top-1

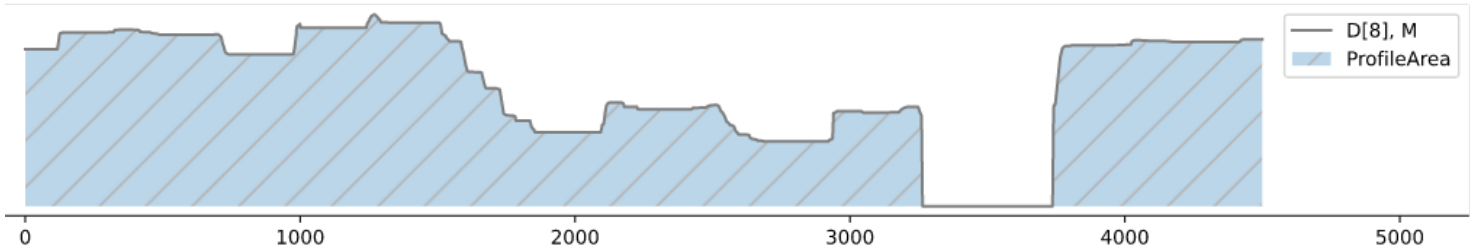
Поиск C_1

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

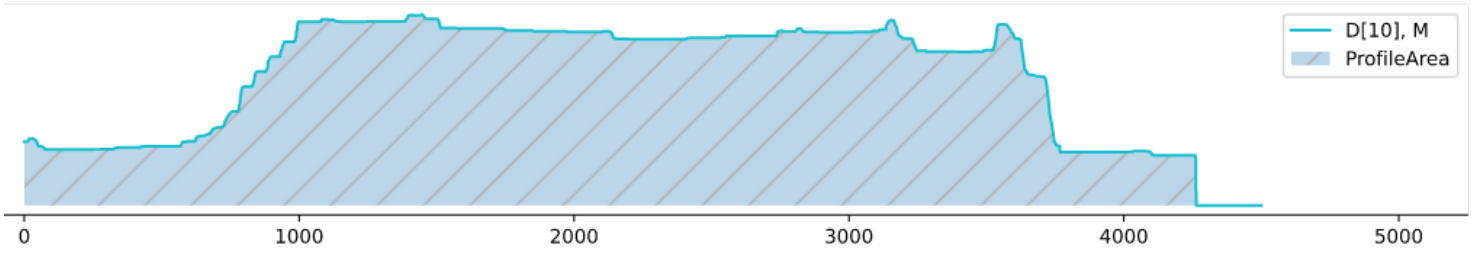
$C_1.index = 8$



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

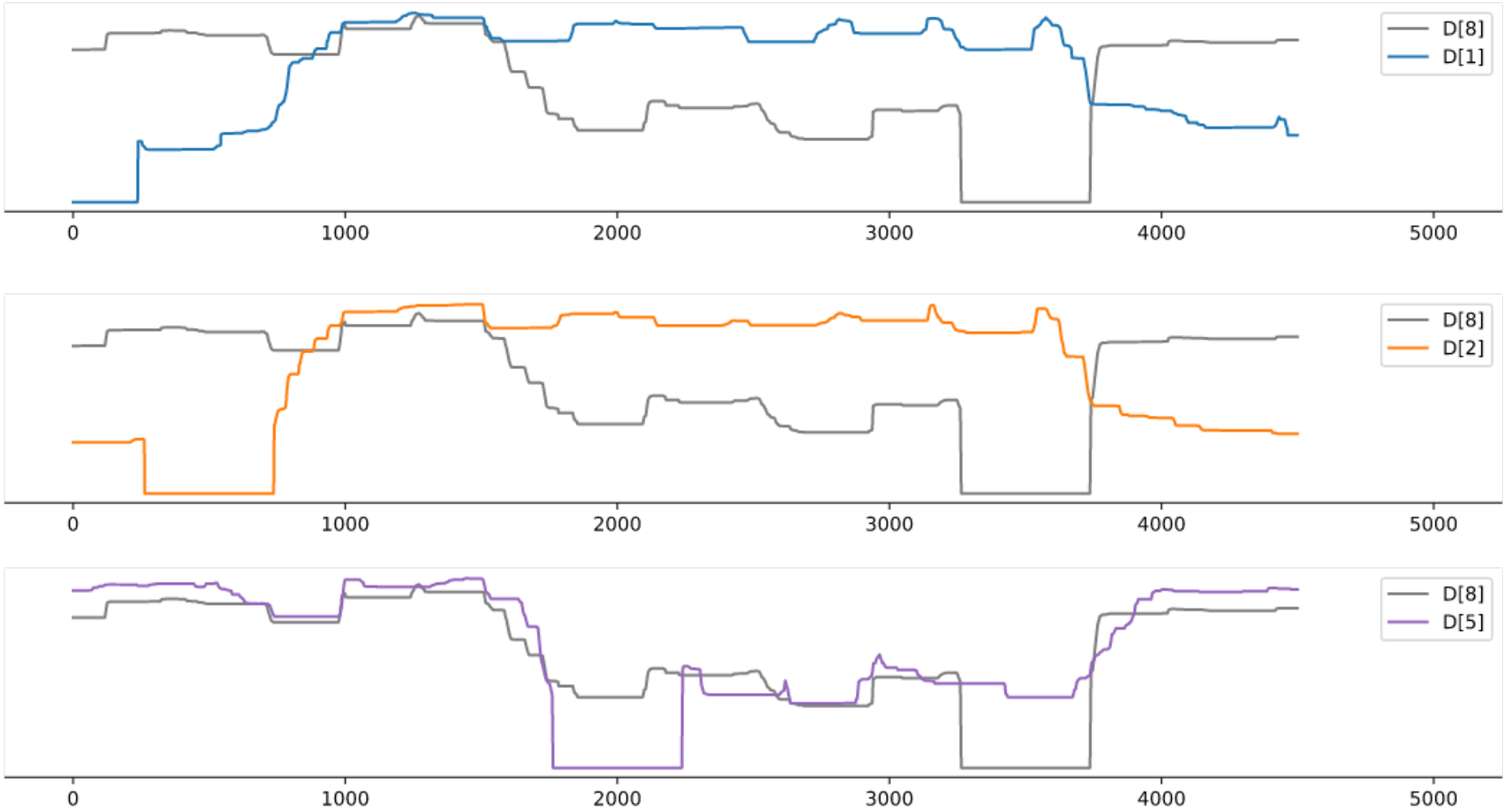


$ProfileArea(\{D_8\}) = 53769$



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

Поиск сниппета top-2

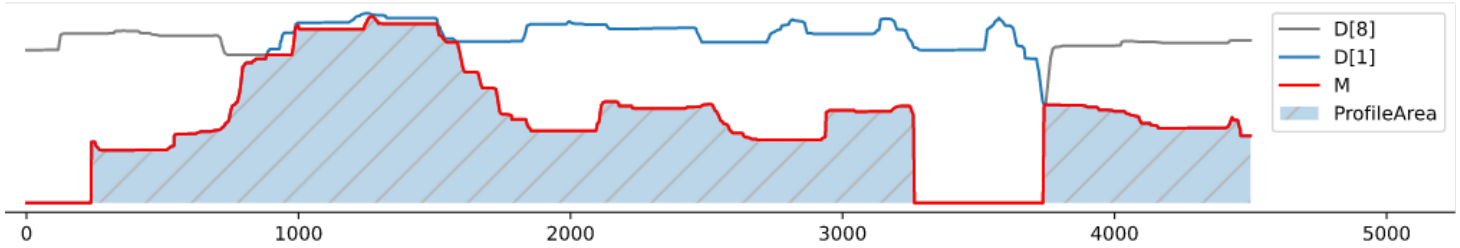


Поиск сниппета top-2

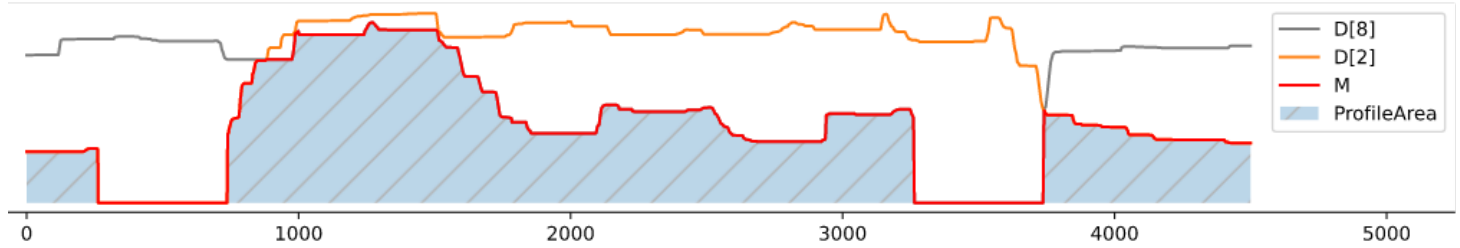
Поиск C_2

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

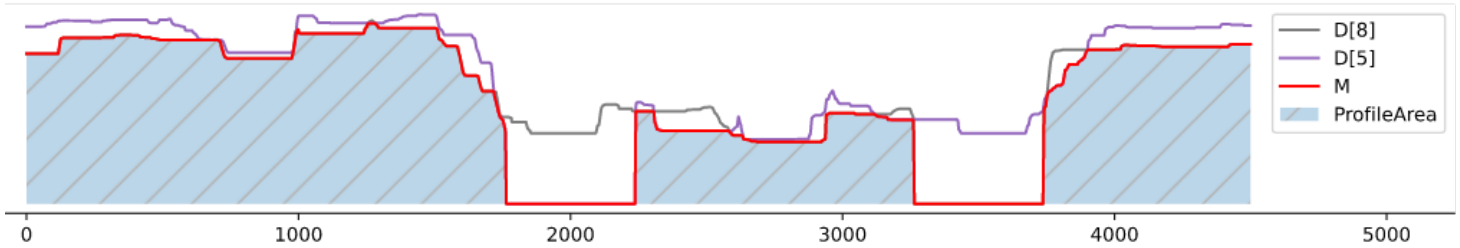
$C_2.index = 2$



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

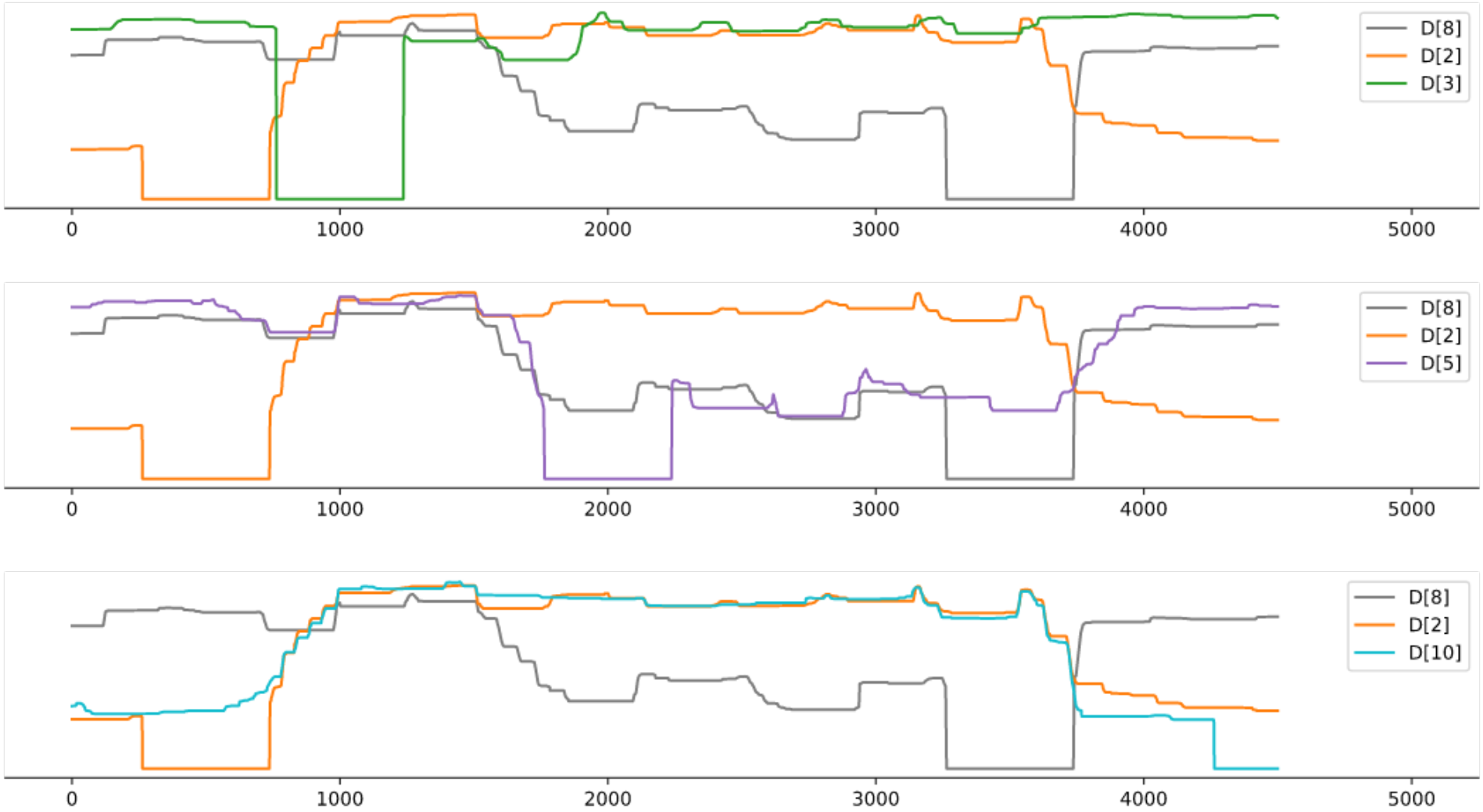


$ProfileArea(\{D_8, D_2\}) = 35769$



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

Поиск сниппета top-3

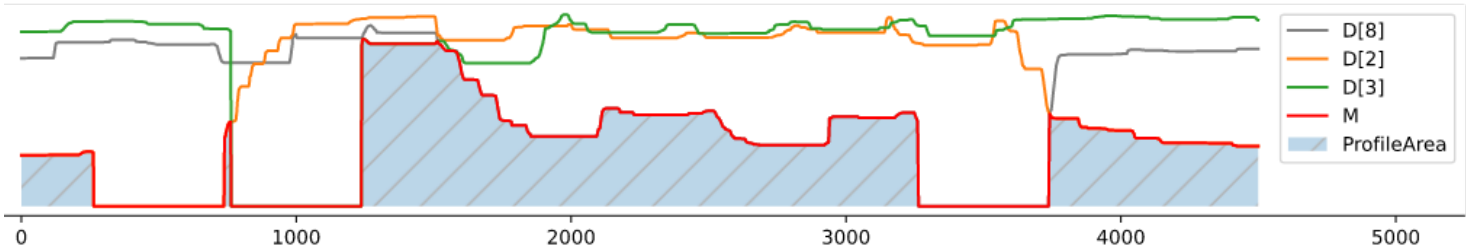


Поиск сниппета top-3

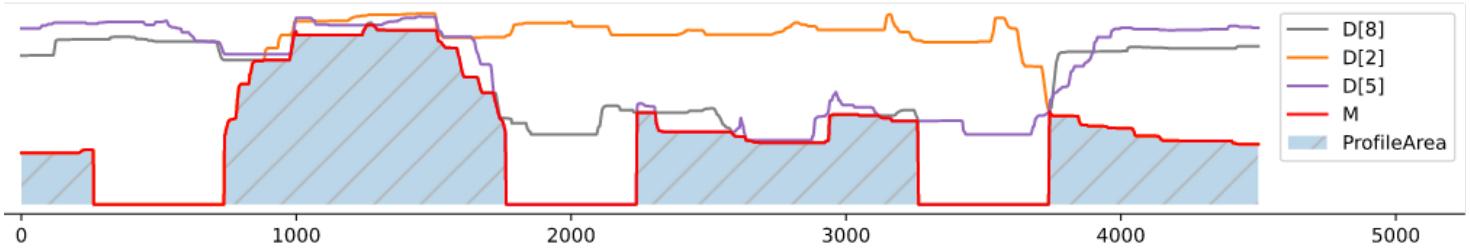
Поиск C_3

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

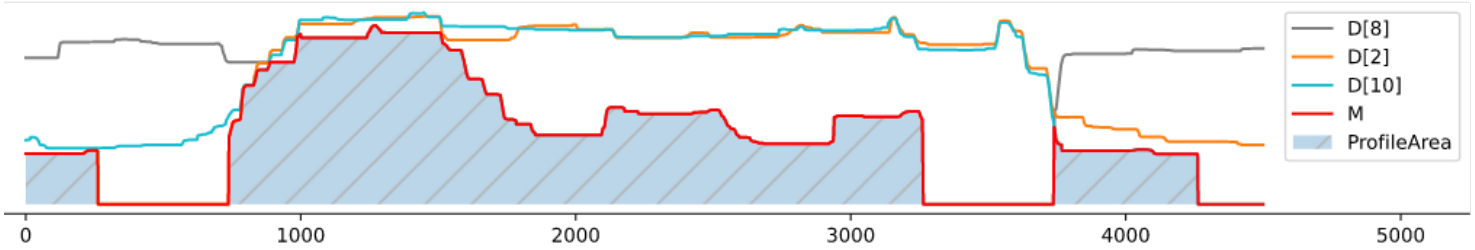
$C_3.index = 3$



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

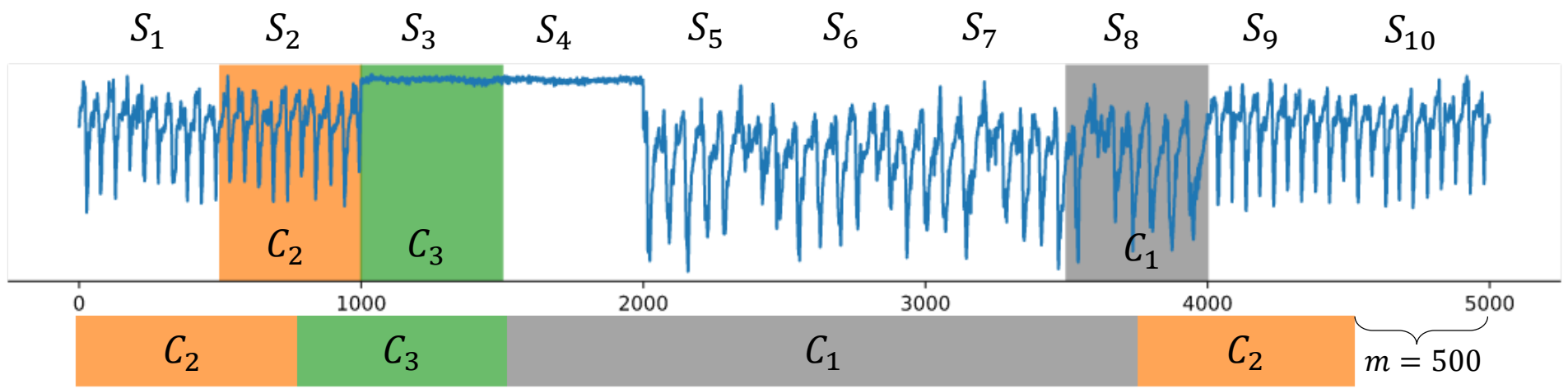


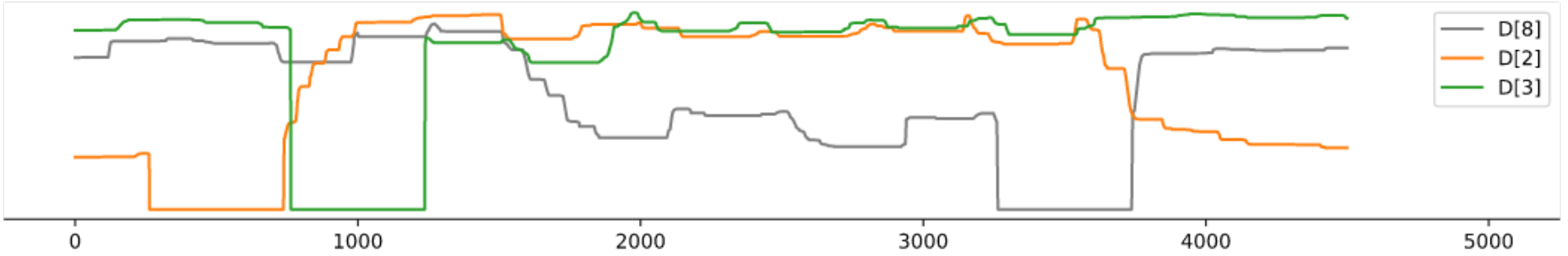
$ProfileArea(\{D_8, D_2, D_5\}) = 31168$

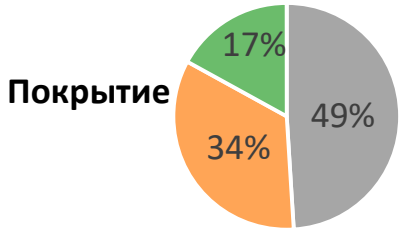


$ProfileArea(\{D_8, D_2, D_{10}\}) = 33044$

Найденные сниппеты







PSF: Parallel Snippet-Finder для GPU*

Шаг	**Snippet-Finder**, сложность $O(n^2 \cdot \frac{n-m}{m})$	**PSF**
1. Вычисление матричного профиля P_{AB}	$\{P_{AB}(i) = \mathrm{ED}_{\mathrm{norm}}(A_{i,\ell}, B_{j,\ell})\}_{i=1}^{m-\ell+1}$, $B_{j,\ell} = \arg \min_{1 \le q \le m-\ell+1} \mathrm{ED}_{\mathrm{norm}}(A_{i,\ell}, B_{q,\ell})$	Вычисление матрицы нормализованных евклидовых расстояний ED_{matr} $allP_{AB}(i,j) = \min_{j \le c \le j+m-\ell+1} ED_{matr}(i,c)$
2. Вычисление матричного профиля P_{BA}	$\{P_{BA}(i) = \mathrm{ED}_{\mathrm{norm}}(B_{i,\ell}, A_{j,\ell})\}_{i=1}^{m-\ell+1}$, $A_{j,\ell} = \arg \min_{1 \le q \le m-\ell+1} \mathrm{ED}_{\mathrm{norm}}(B_{i,\ell}, A_{q,\ell})$	$allP_{BA}(j) = \min_{1 \le i \le m-\ell+1} ED_{matr}(i,j)$
3. Вычисление матричного профиля P_{ABBA}	$P_{ABBA} = P_{AB} \odot P_{BA}$	$P_{ABBA} = allP_{AB}(i, m-\ell) \odot allP_{BA}(i)$
4. Вычисление MPdist профиля	$\mathrm{MPdist}(A,B,\ell) = \begin{cases} SortedP_{ABBA}(k), & \lvert P_{ABBA}\rvert > k \\ SortedP_{ABBA}(2(m-\ell+1)), & \lvert P_{ABBA}\rvert \le k \end{cases}$	

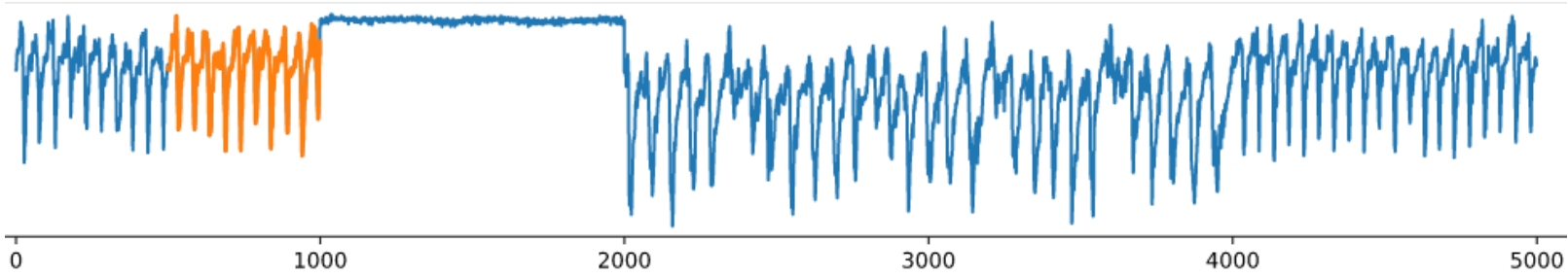
*Zymbler M., Goglachev A. Fast Summarization of Long Time Series with Graphics Processor. Mathematics. 2022. Vol. 10, No. 10. Article 1781. DOI: 10.3390/math10101781

Гоглачев А.И., Цымблер М.Л. Свидетельство Роспатента о государственной регистрации программы для ЭВМ «PSF: программа для автоматического аннотирования временного ряда на графическом процессоре» № 2022619627 от 24.05.2022

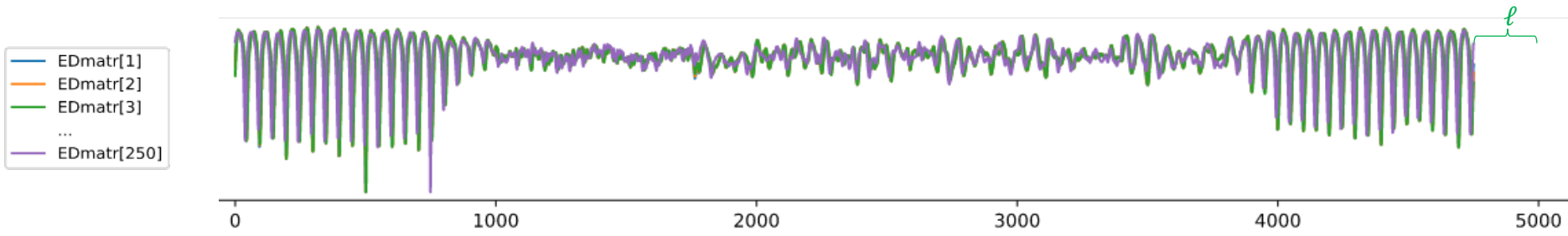
PSF: вычисление матрицы расстояний

Сегмент S
$|S| = m$

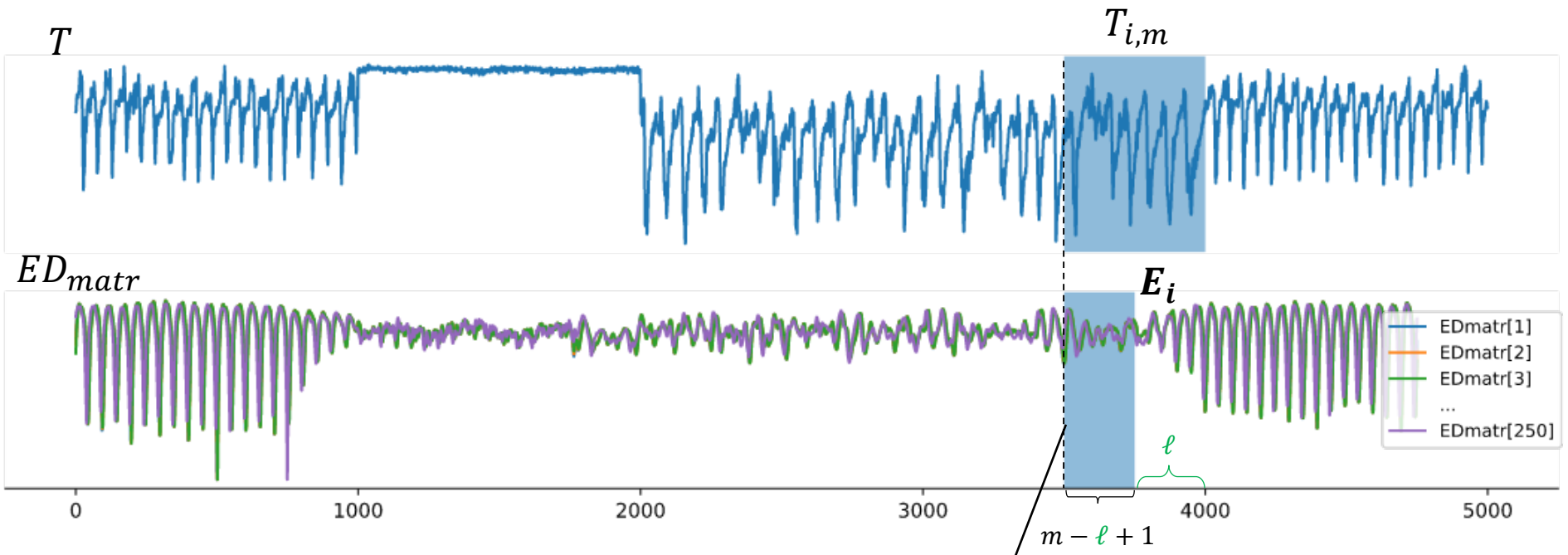
Временной ряд T
$|T| = n \gg m$



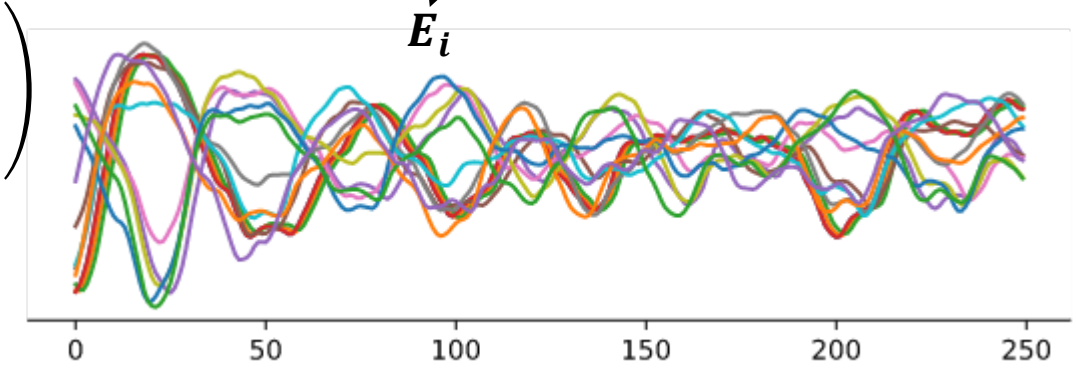
$$\boldsymbol{ED_{matr}}(\boldsymbol{S}, \boldsymbol{T}, \ell) = \begin{pmatrix} d_{1,1} & \cdots & d_{1,n-\ell+1} \\ \cdots & \ddots & \cdots \\ d_{m-\ell+1,1} & \cdots & d_{m-\ell+1,n-\ell+1} \end{pmatrix}, \qquad d_{i,j} = \mathrm{ED}_{\mathrm{norm}}(S_{i,\ell}, T_{j,\ell})$$



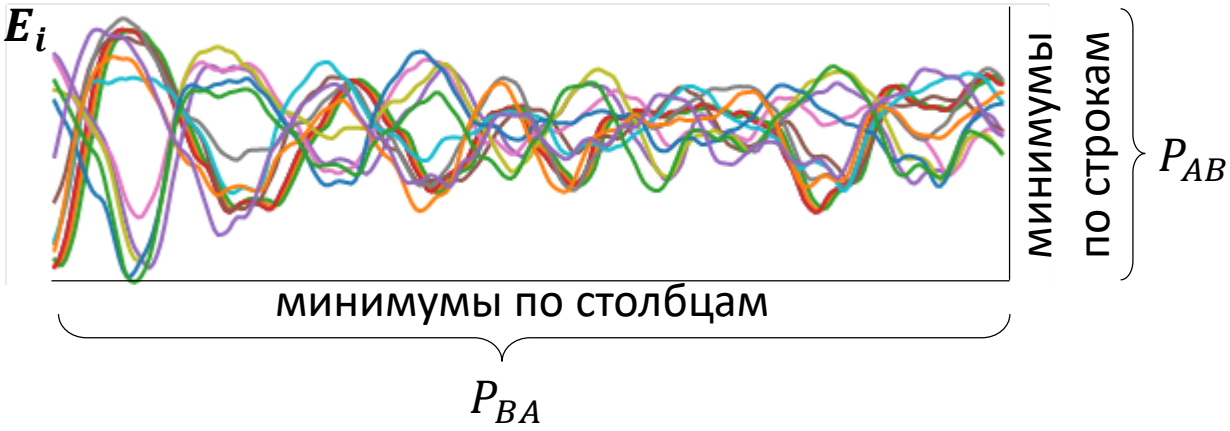
PSF: вычисление матричного профиля



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



PSF: вычисление матричного профиля

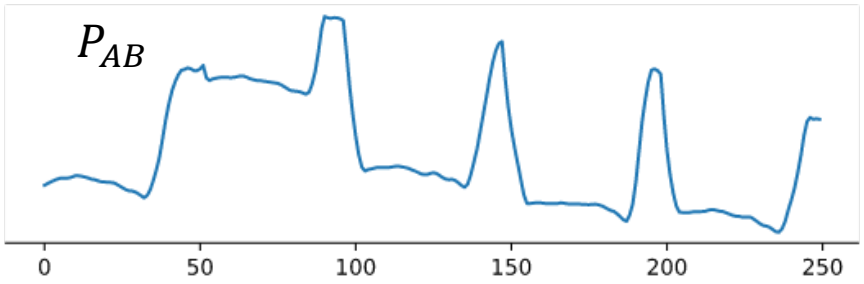


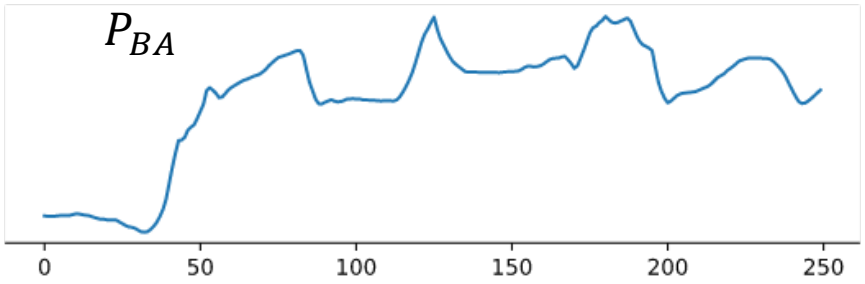
$$P_{AB}(i) = \min_{1 \le j \le m-\ell+1} E(i,j),$$

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

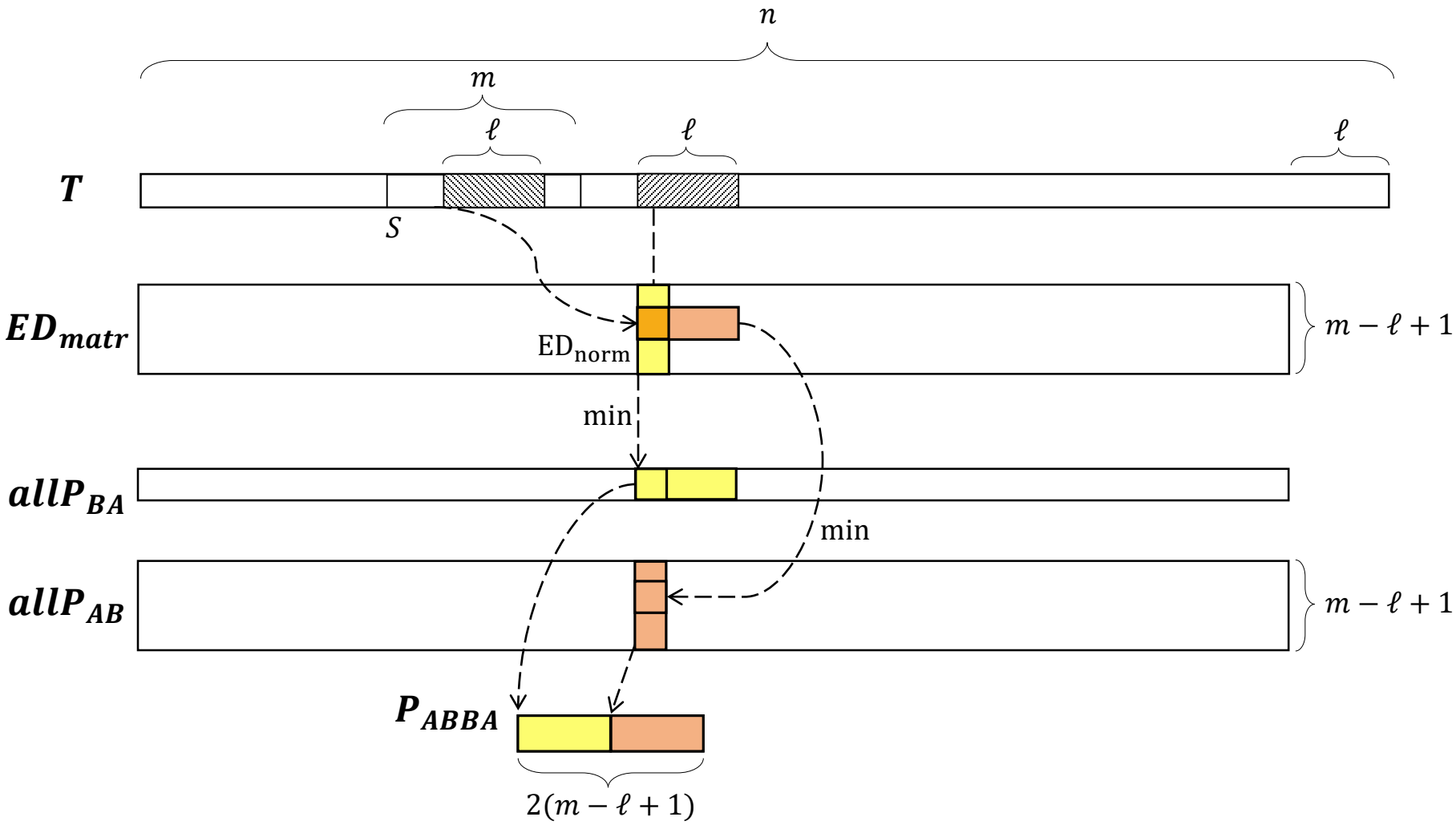
$$P_{BA}(j) = \min_{1 \le i \le m-\ell+1} E(i,j),$$

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

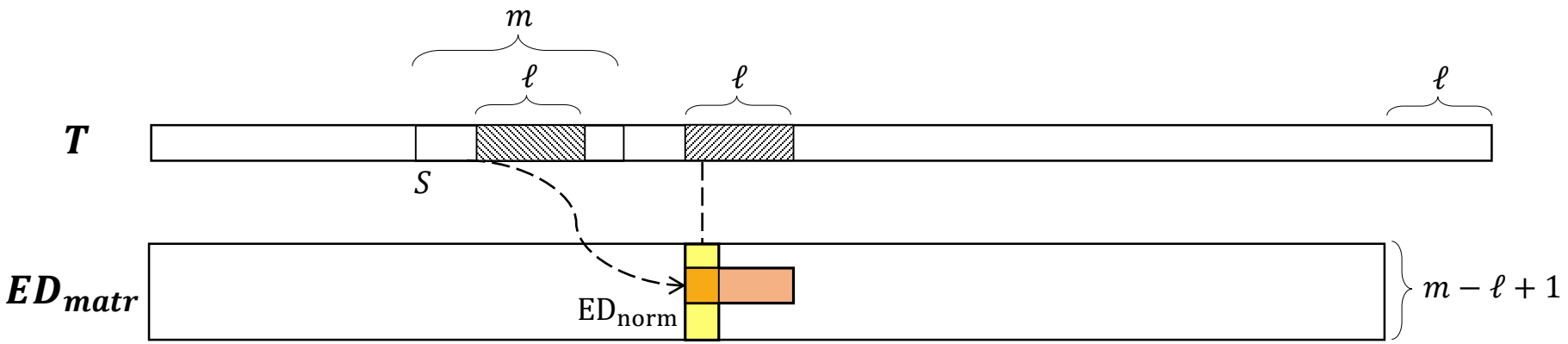




PSF: структуры данных



PSF: вычислительное ядро $\boldsymbol{ED_{matr}}$



$$\mathrm{ED}_{\mathrm{norm}}(T_{i,m}, T_{j,m}) = \sqrt{2m(1 - P_{i,j})}$$

$$P_{i,j} = \overline{QT}_{i,j} \cdot \frac{1}{\|T_{i,m} - \mu_i\|} \cdot \frac{1}{\|T_{j,m} - \mu_j\|},$$

$$T_{i,m} - \mu_i = (t_i - \mu_i, \ldots, t_{i+m-1} - \mu_i),$$

$$\mu_i = \frac{1}{m} \sum_{j=i}^{i+m} t_j,$$

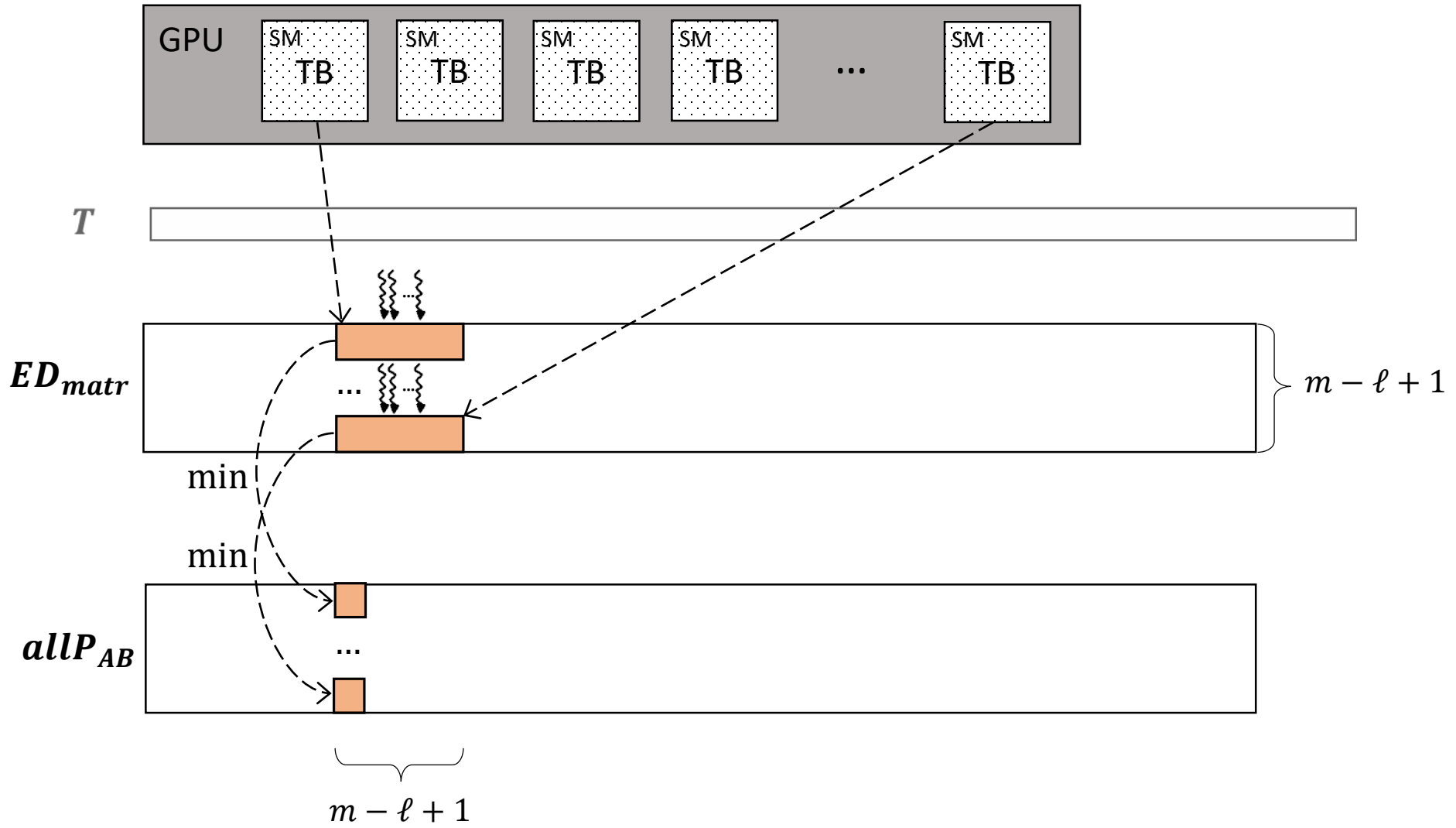
$$dg_0 = 0; dg_i = (t_{i+m-1} - \mu_i) + (t_{i-1} - \mu_{i-1}),$$

$$df_0 = 0; df_i = \frac{t_{i+m-1} - t_{i-1}}{2},$$

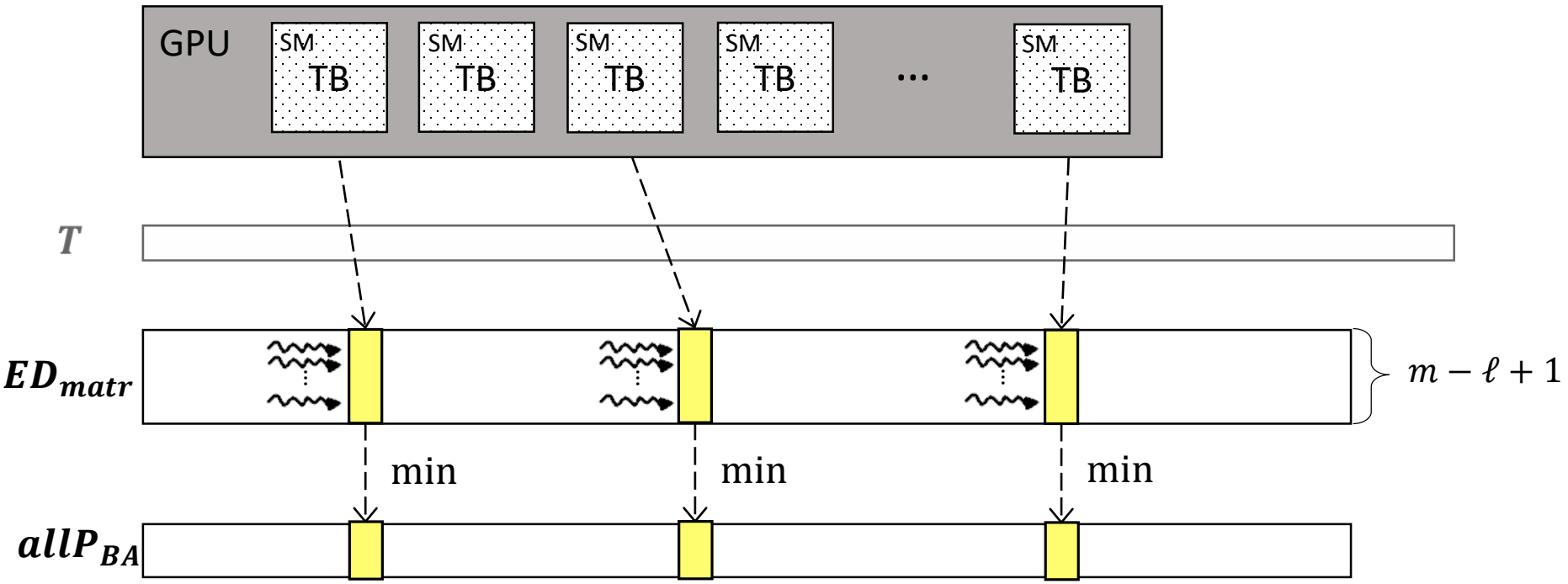
$$\overline{QT}_{i,j} = \overline{QT}_{i-1,j-1} + df_i \cdot dg_j + df_j \cdot dg_i,$$

* 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.

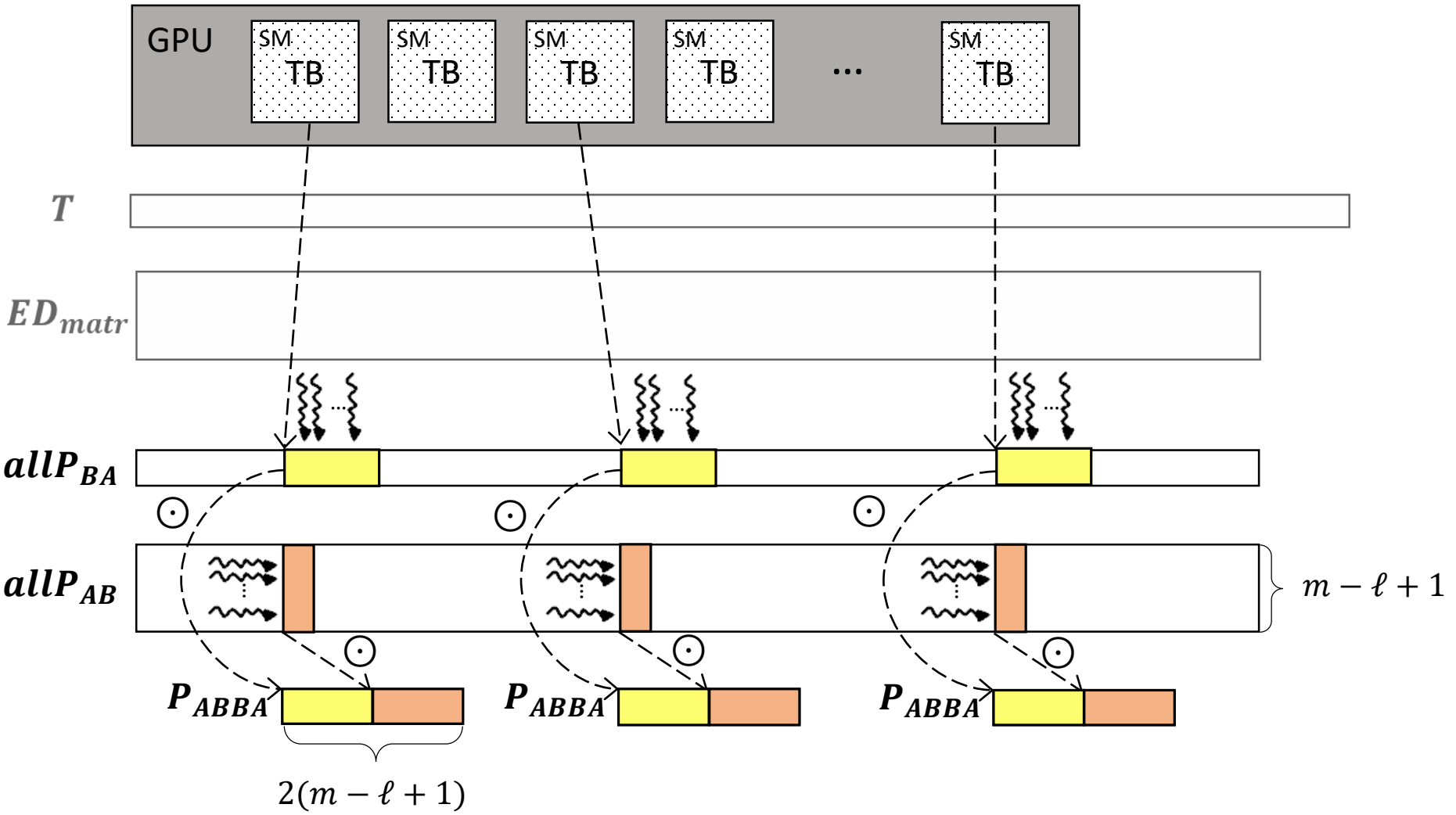
PSF: вычислительное ядро $\boldsymbol{allP_{AB}}$



PSF: вычислительное ядро $\boldsymbol{allP_{BA}}$



PSF: вычислительное ядро $\boldsymbol{P_{ABBA}}$



Эксперименты

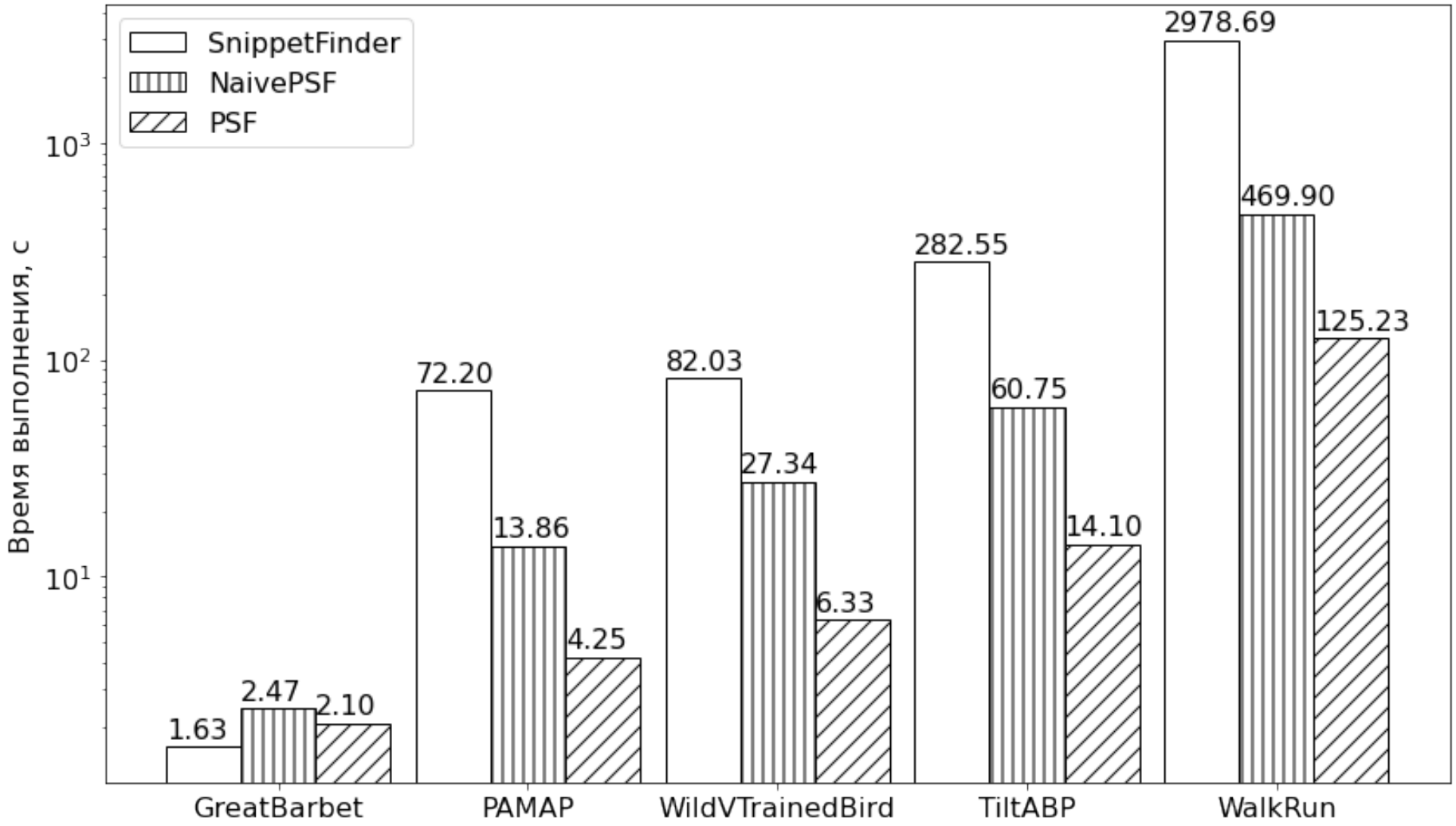
- Платформа: NVIDIA Tesla V100 SXM2 (5120 ядер @1.3 GHz, 15.7 TFLOPS)
- Конкуренты:
  - Snippet-Finder (оригинальный последовательный алгоритм)
  - NaivePSF (параллельное вычисление матричного профиля между всеми парами сегментов и подпоследовательностей)
- Данные:

Название	Длина ряда n	Длина сегмента m	Описание
GreatBarbet[1]	2 801	150	Показатели физиологической активности птиц
WildVTrainedBird[1]	20 002	900	
PAMAP[2]	20 002	600	Показания носимого акселерометра во время различных видов физической активности человека
WalkRun[2]	100 000	240	
TiltABP[1]	40 000	630	Показания кровяного давления человека во время быстрых наклонов

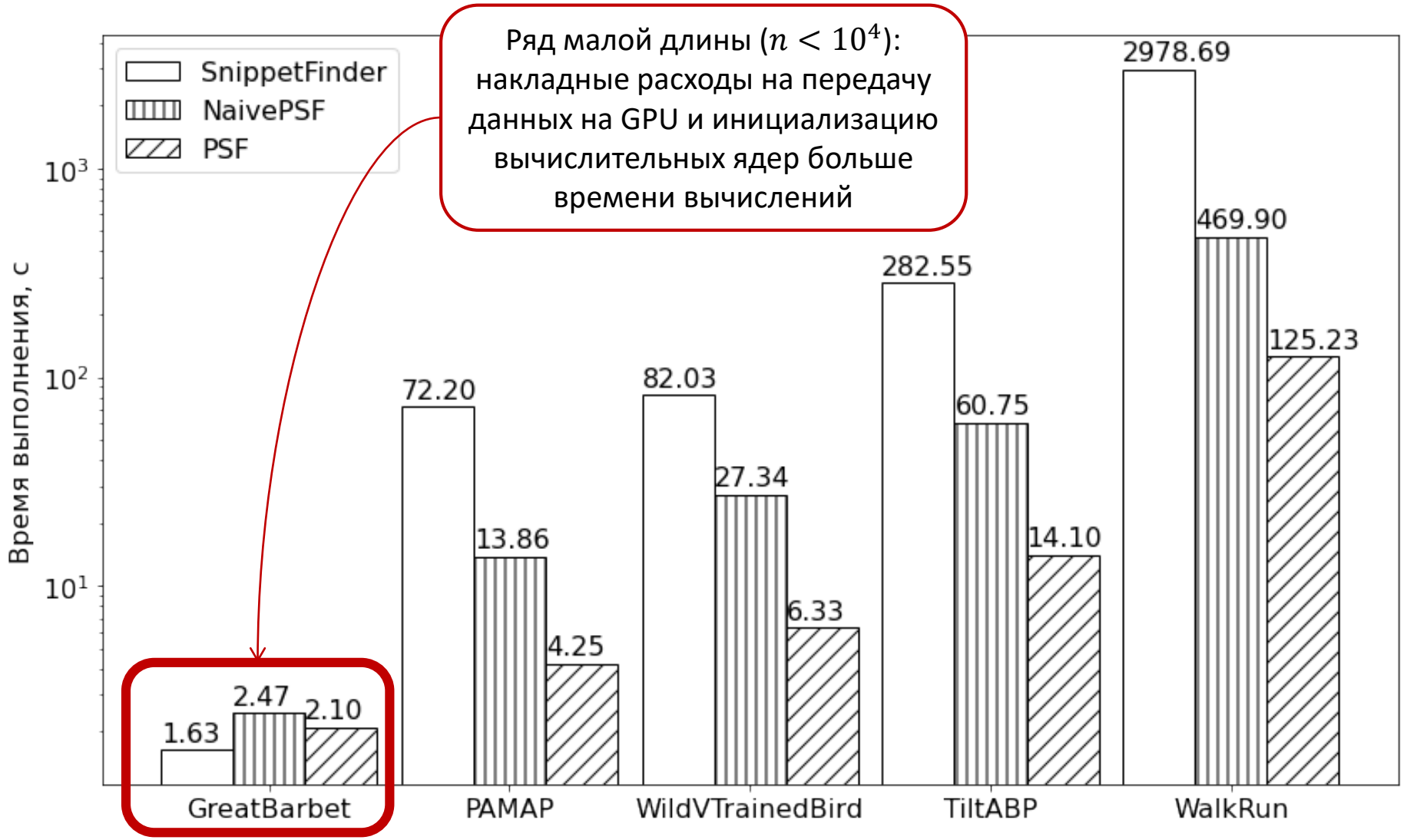
[1] Imani S., Madrid F., Ding W., Crouter S.E., Keogh E.J. Introducing time series snippets: a new primitive for summarizing long time series. Data Min. Knowl. Discov. 34(6): 1713-1743 (2020). doi: 10.1007/s10618-020-00702-y

[2] Reiss A., Stricker D. Introducing a new benchmarked dataset for activity monitoring. ISWC 2012, Newcastle, UK, June 18-22, 2012. 108–109. IEEE (2012). doi: 10.1109/ISWC.2012.13

Эксперименты: производительность



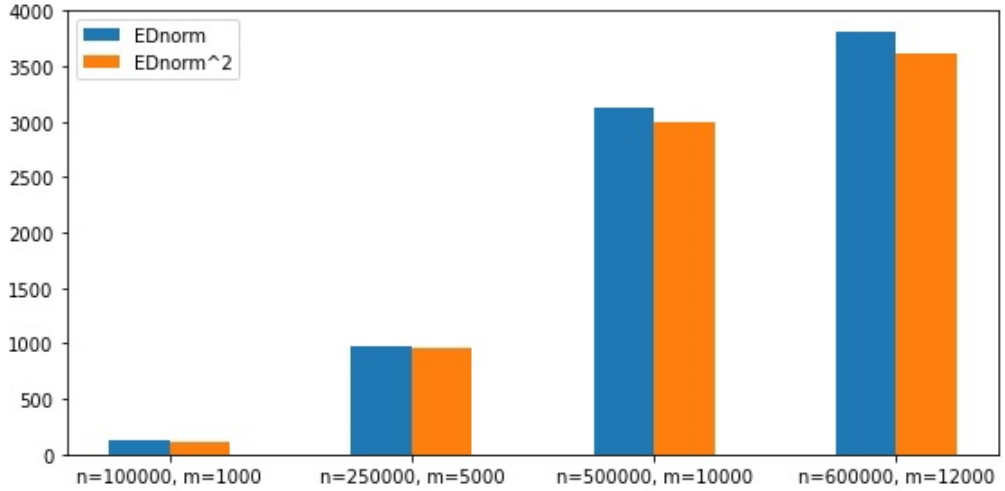
Эксперименты: производительность



Эксперименты: производительность

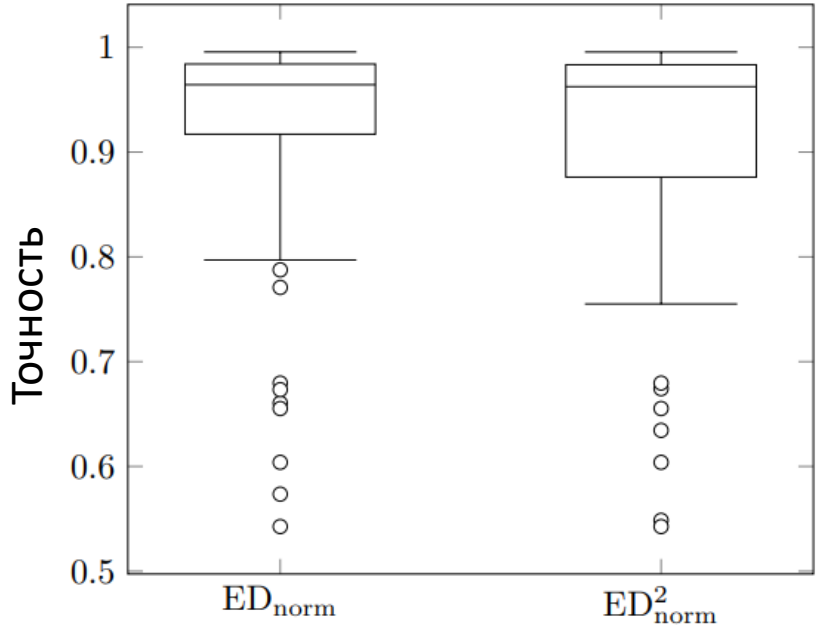
Эксперименты: $\mathrm{ED}^2_{\mathrm{norm}}$ *vs.* $\mathrm{ED}_{\mathrm{norm}}$

Временной ряд Random Walk*



$$\text{Точность} = \frac{\text{кол-во корректно размеченных точек}}{n}$$

Набор из 100 рядов MixedBag**



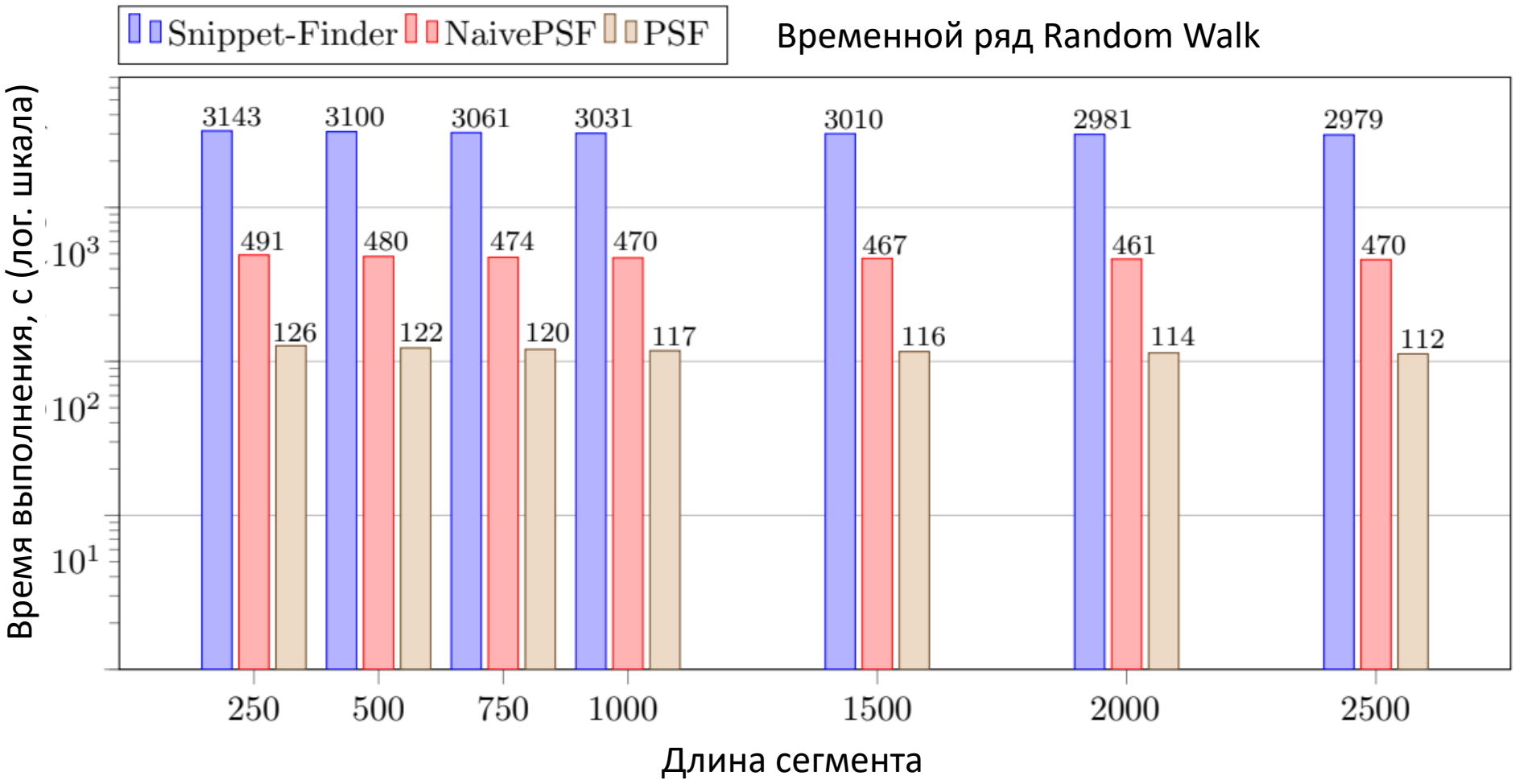
Метрика в основе MPdist

Отказ от вычисления квадратного корня ускоряет вычисления на 6-10% без потери адекватности

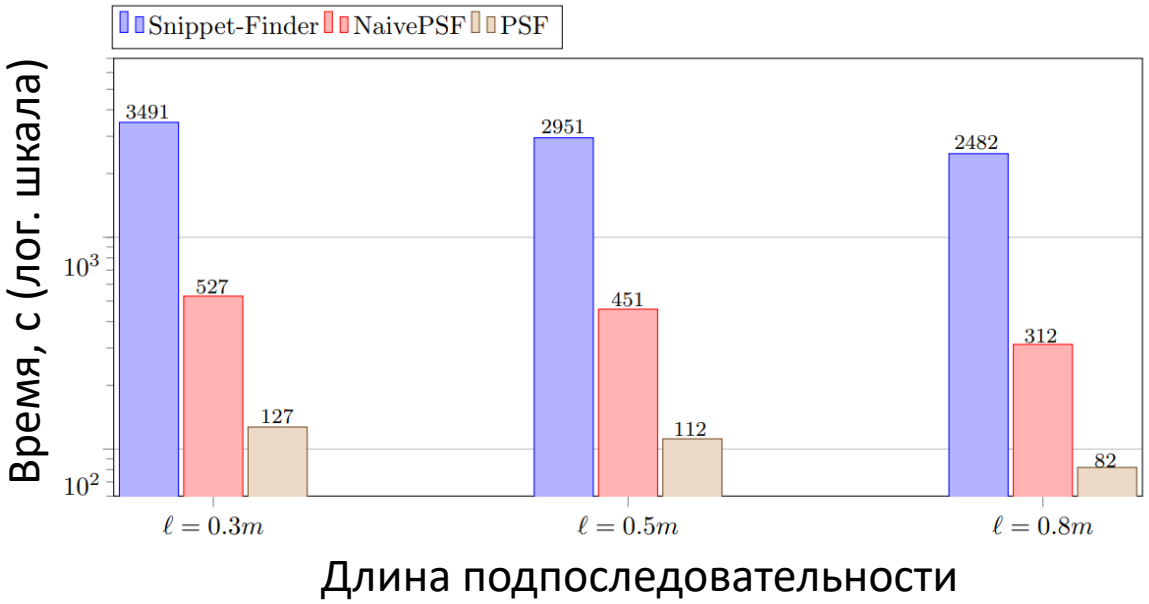
* Pearson K. The problem of the random walk. Nature. 72(1865), 294 (1905). DOI: 10.1038/072342A0

** 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

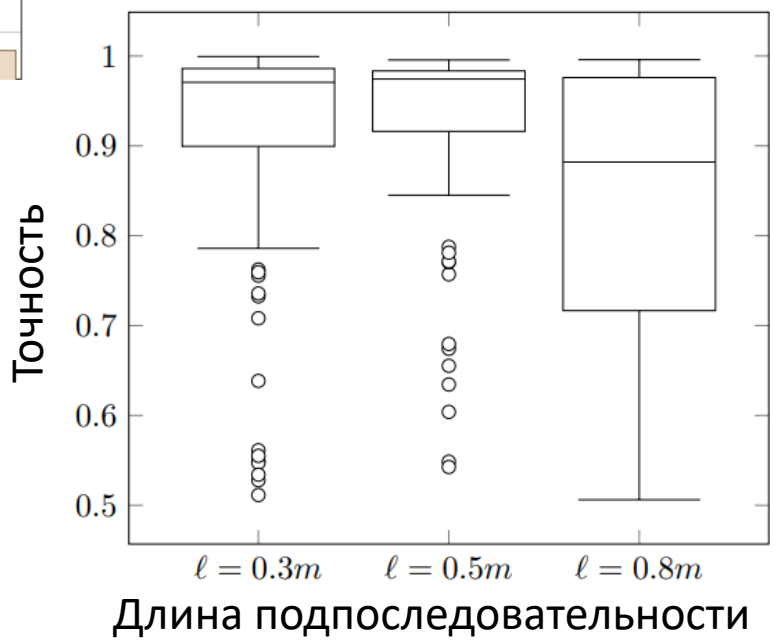
Эксперименты: длина сегмента



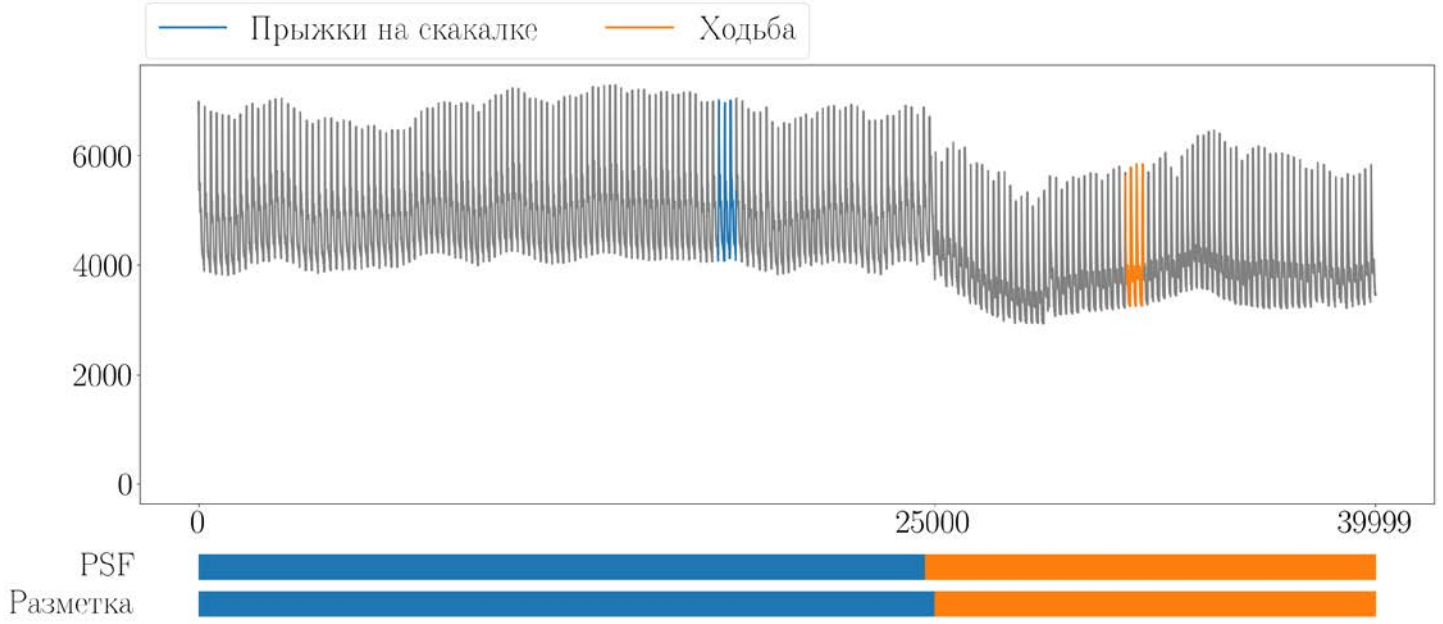
Эксперименты: длина подпоследовательности

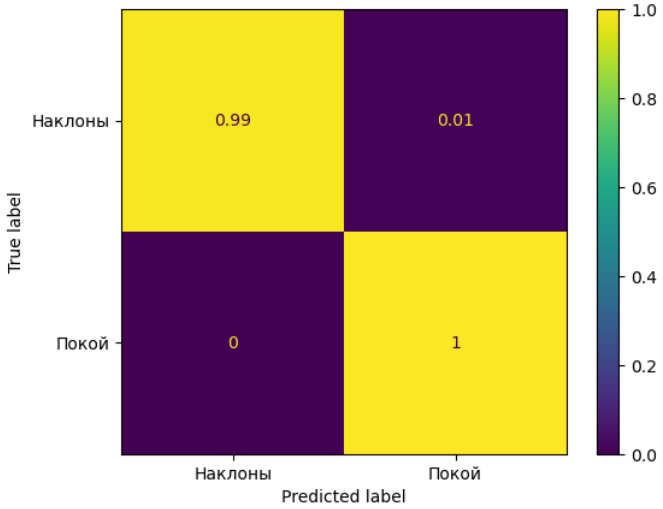


Набор из 100 рядов MixedBag



Case studies: артериальное давление*

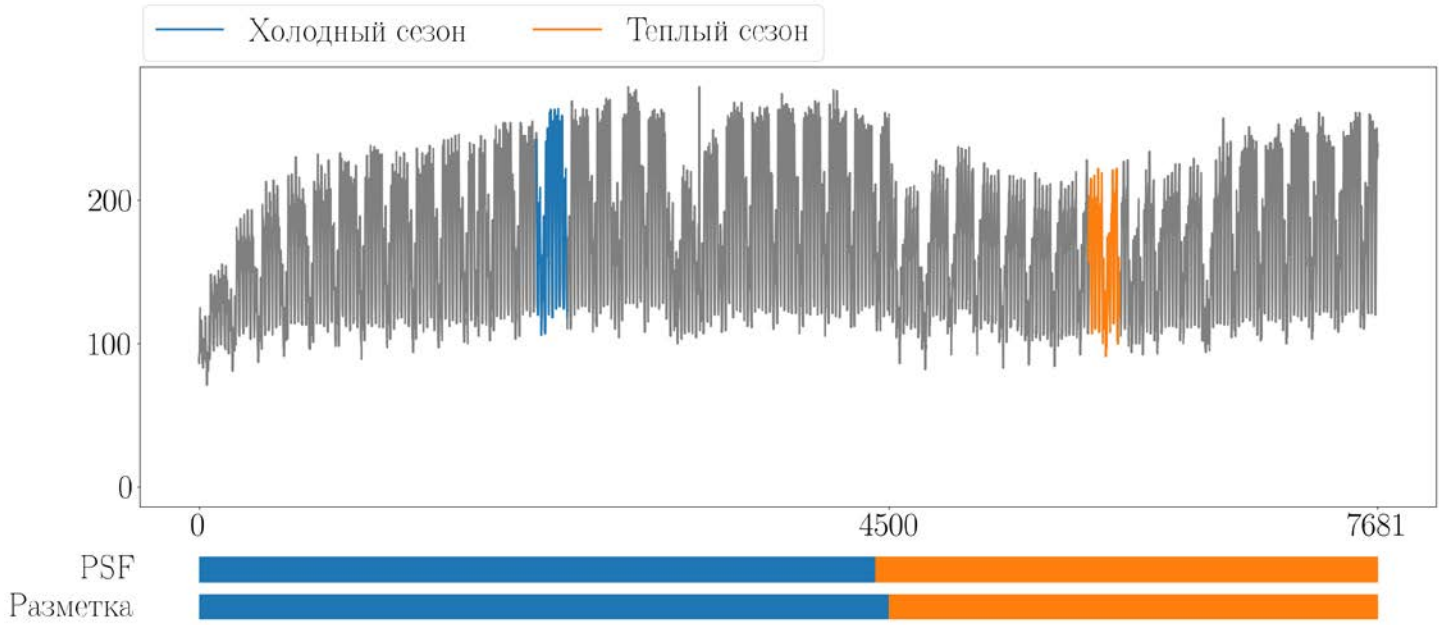


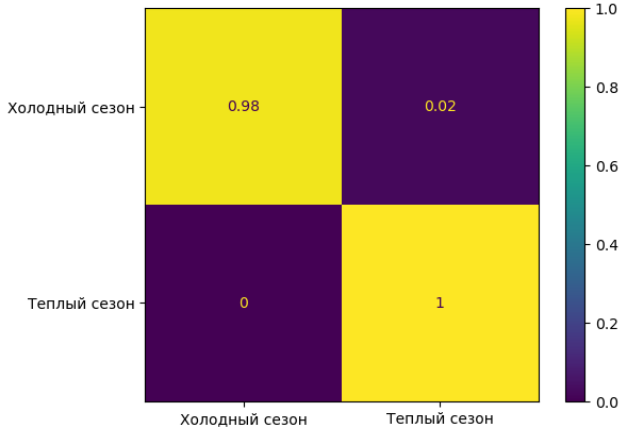


Активность	Точность	Полнота	F1-мера
Наклоны	1	0.99	0.99
Покой	0.98	1	0.99

* 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

Case studies: энергопотребление*

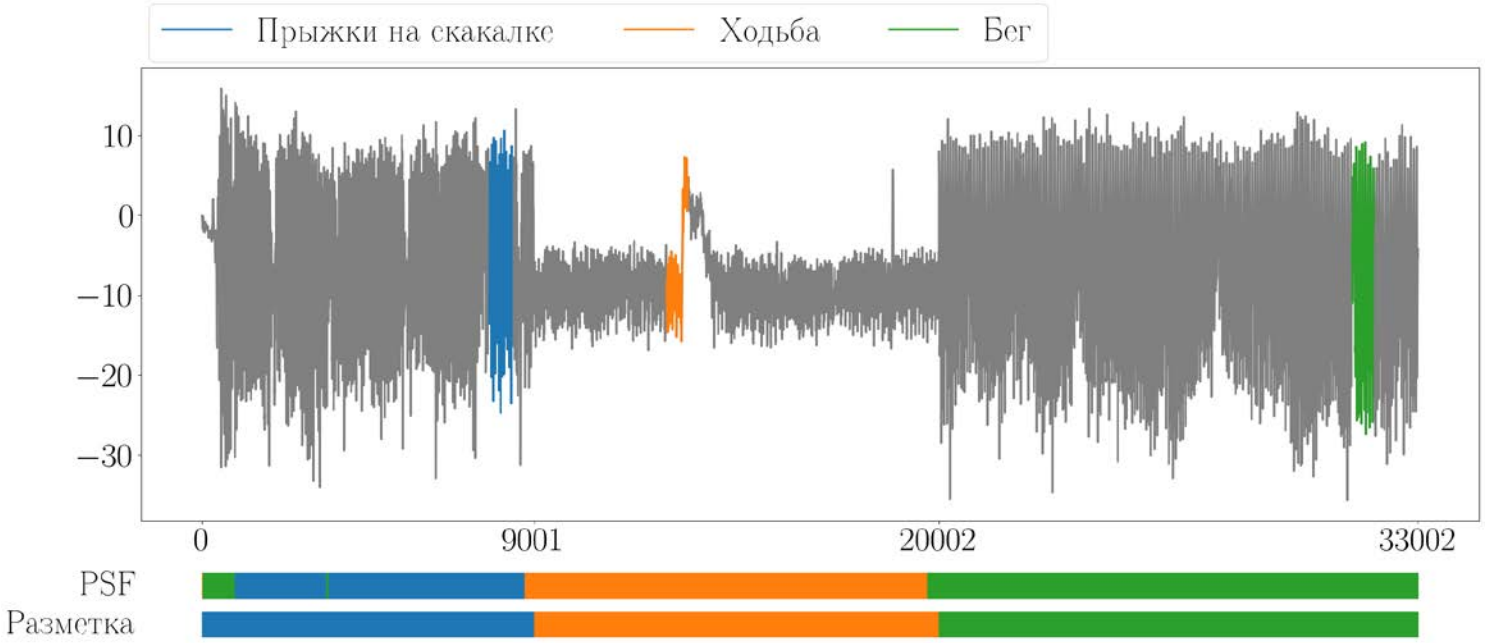


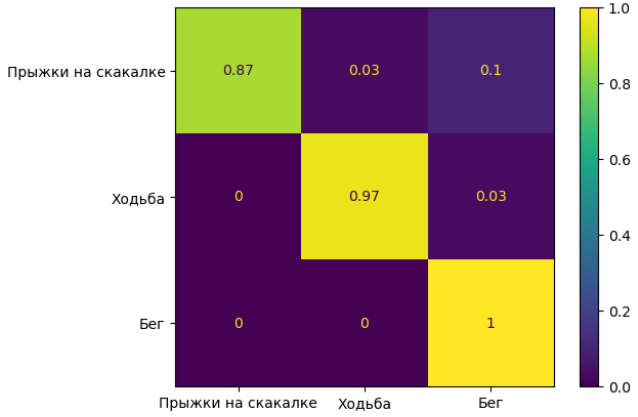


Активность	Точность	Полнота	F1-мера
Холодный сезон	1	0.98	0.99
Теплый сезон	0.97	1	0.99

* 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

Case studies: носимый акселерометр*

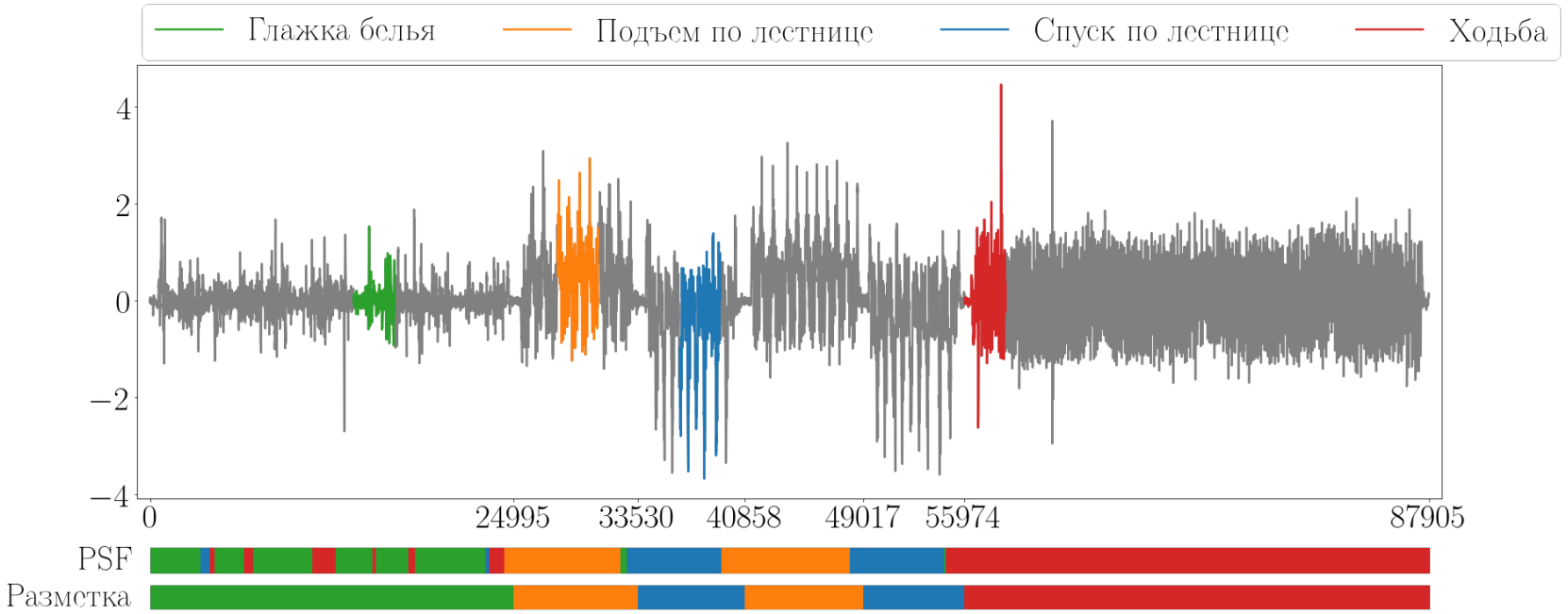




Активность	Точность	Полнота	F1-мера
Прыжки на скакалке	1	0.87	0.93
Ходьба	0.98	0.97	0.97
Бег	0.77	1	0.87

* Reiss A., Stricker D. Introducing a new benchmarked dataset for activity monitoring. ISWC 2012, Newcastle, UK, June 18-22, 2012. 108–109. IEEE (2012). doi: 10.1109/ISWC.2012.13

Case studies: носимый акселерометр*



Активность	Точность	Полнота	F1-мера
Спуск по лестнице	0.80	0.79	0.80
Подъем по лестнице	0.87	0.87	0.87
Глажка белья	0.97	0.77	0.86
Ходьба	0.86	1	0.92

* Reiss A., Stricker D. Introducing a new benchmarked dataset for activity monitoring. ISWC 2012, Newcastle, UK, June 18-22, 2012. 108–109. IEEE (2012). doi: 10.1109/ISWC.2012.13

Продолжение исследований

1. Разработка версии PSF для HPC-кластера с GPU узлами
2. Применение PSF в разработке нейросетевой модели для поведенческой классификации многомерного временного ряда