

# Поиск типичных подпоследовательностей временного ряда

дела



*Время есть жизнь души, пребывающей  
в переходном движении от одного  
жизненного проявления к другому.*

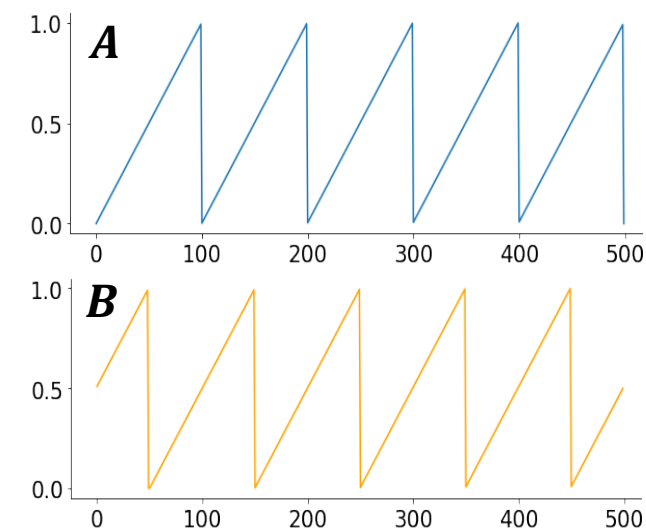
*Плотин*

# Содержание

- Мера MPdist
- Снимпеты и алгоритм Snippet-Finder
- Применение снимпетов

# Мера MPdist (Matrix Profile distance)\*: неформальное определение

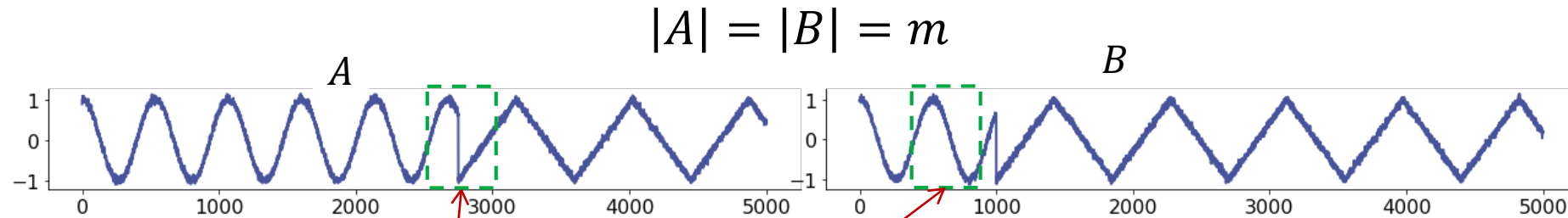
- Схожесть временных рядов  $A$  и  $B$  ( $|A| = |B| = m$ ) по MPdist пропорциональна количеству в них подпоследовательностей длины  $\ell$  ( $3 \leq \ell \leq m$ ), близких в смысле нормализованного евклидова расстояния
- MPdist не удовлетворяет аксиоме треугольника
- MPdist устойчива к шумам, инвариантна к фазе и амплитуде рядов



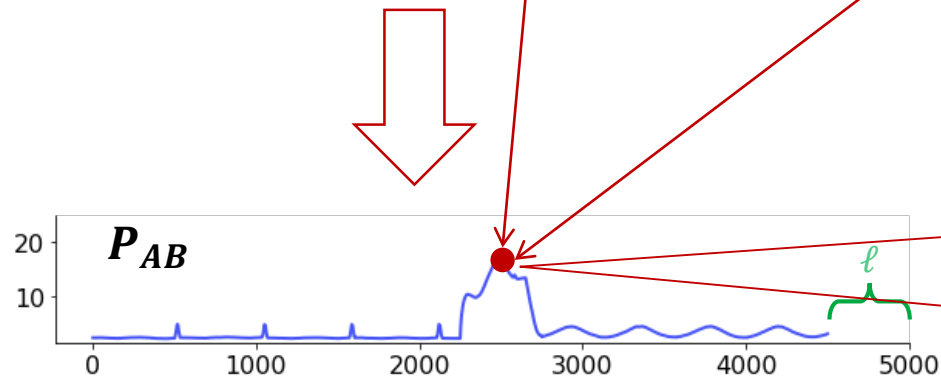
|                          |          |
|--------------------------|----------|
| ED( $A, B$ )             | 11.2     |
| <b>MPdist</b> ( $A, B$ ) | <b>0</b> |

\* Gharghabi S. *et al.* An ultra-fast time series distance measure to allow data mining in more complex real-world deployments. Data Min. Knowl. Discov. 2020. Vol. 34. P. 1104–1135. DOI: [10.1007/s10618-020-00695-8](https://doi.org/10.1007/s10618-020-00695-8)

# Мера MPdist, формальное определение: $P_{AB}$



Значимая длина подпоследовательности:  $3 \leq \ell \leq m$  (обычно  $[0.3m] < \ell \leq [0.8m]$ )

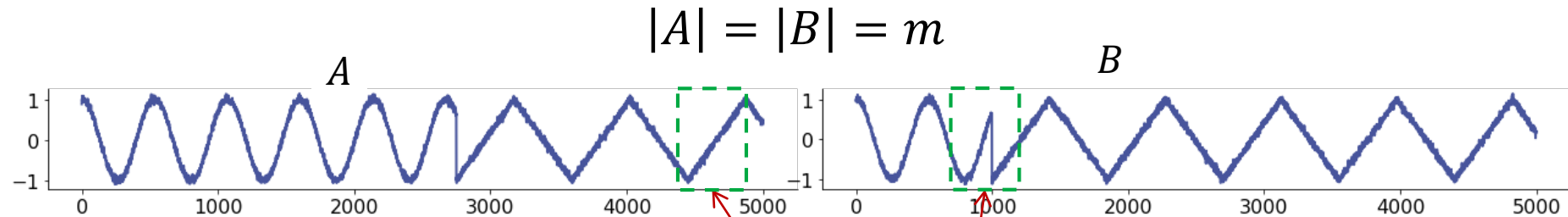


$$P_{AB}(i) = \text{ED}_{\text{norm}}(A_{i,\ell}, B_{\text{nn}})$$

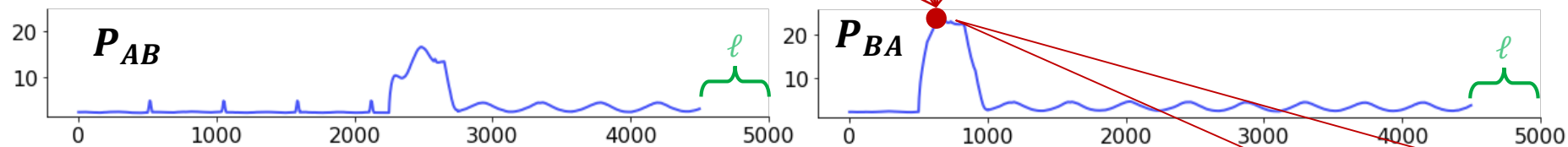
$$B_{\text{nn}} = \arg \min_{B_{j,\ell} \in S_B^\ell} \text{ED}_{\text{norm}}(A_{i,\ell}, B_{j,\ell})$$

Нормализованное евклидово расстояние от  $i$ -й подпоследовательности длины  $\ell$  из  $A$  до ее ближайшего соседа из  $B$

# Мера MPdist, формальное определение: $P_{BA}$



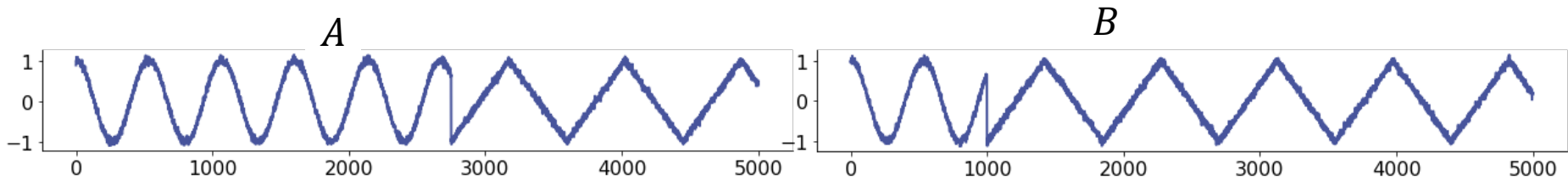
Значимая длина подпоследовательности:  $3 \leq \ell \leq m$  (обычно  $[0.3m] < \ell \leq [0.8m]$ )



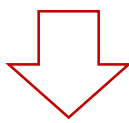
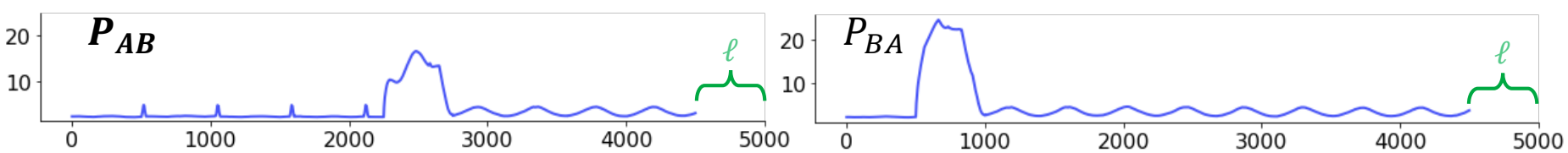
$$P_{BA}(j) = \text{ED}_{\text{norm}}(A_{\text{nn}}, B_{j,\ell})$$

$$A_{\text{nn}} = \arg \min_{A_{i,\ell} \in S_A^\ell} \text{ED}_{\text{norm}}(A_{i,\ell}, B_{j,\ell})$$

# Мера MPdist, формальное определение: $P_{ABBA}$

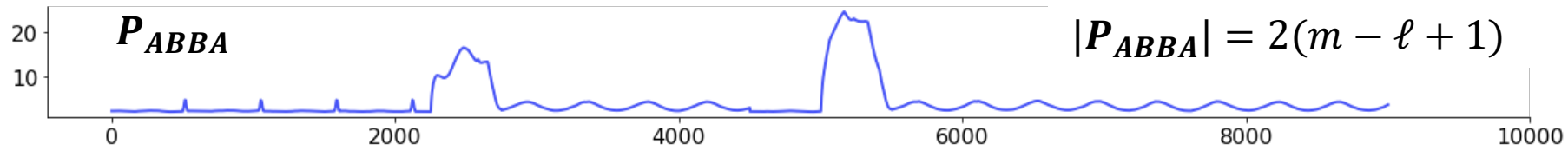


$3 \leq \ell \leq m$

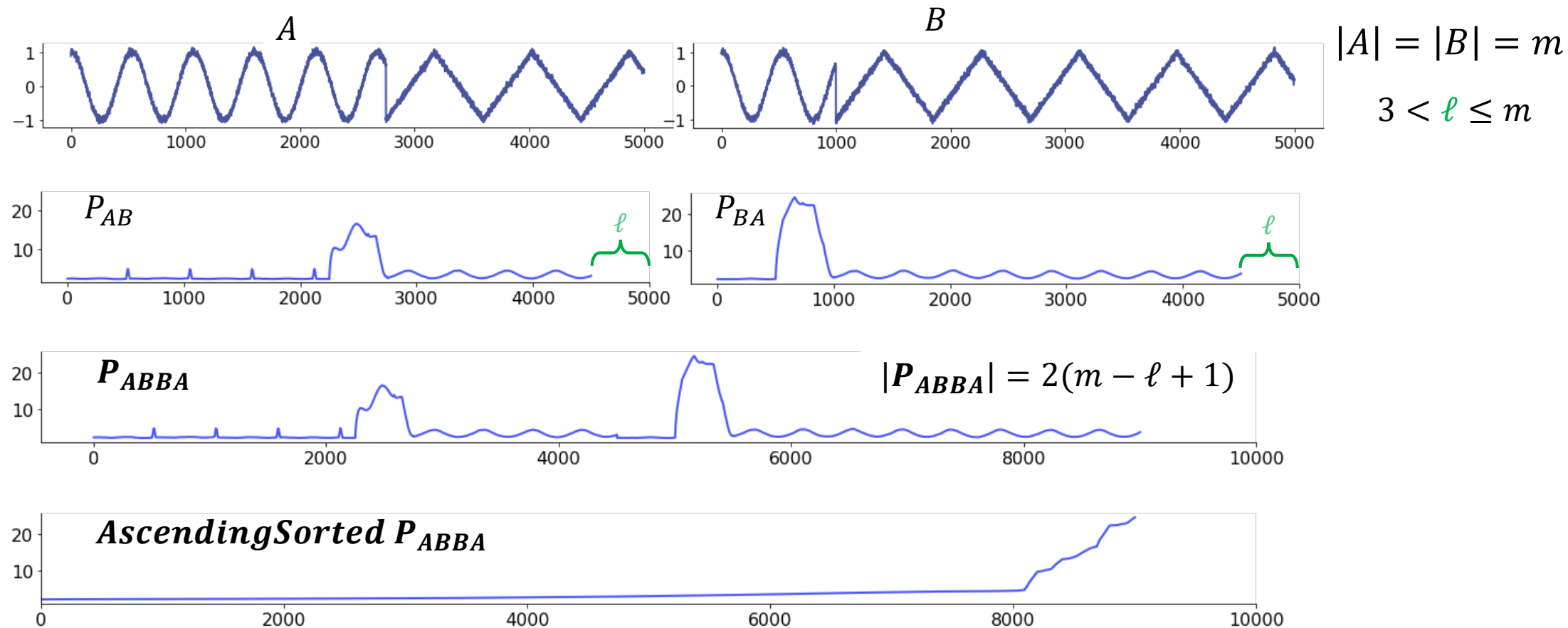


Конкатенация  
(склейка)

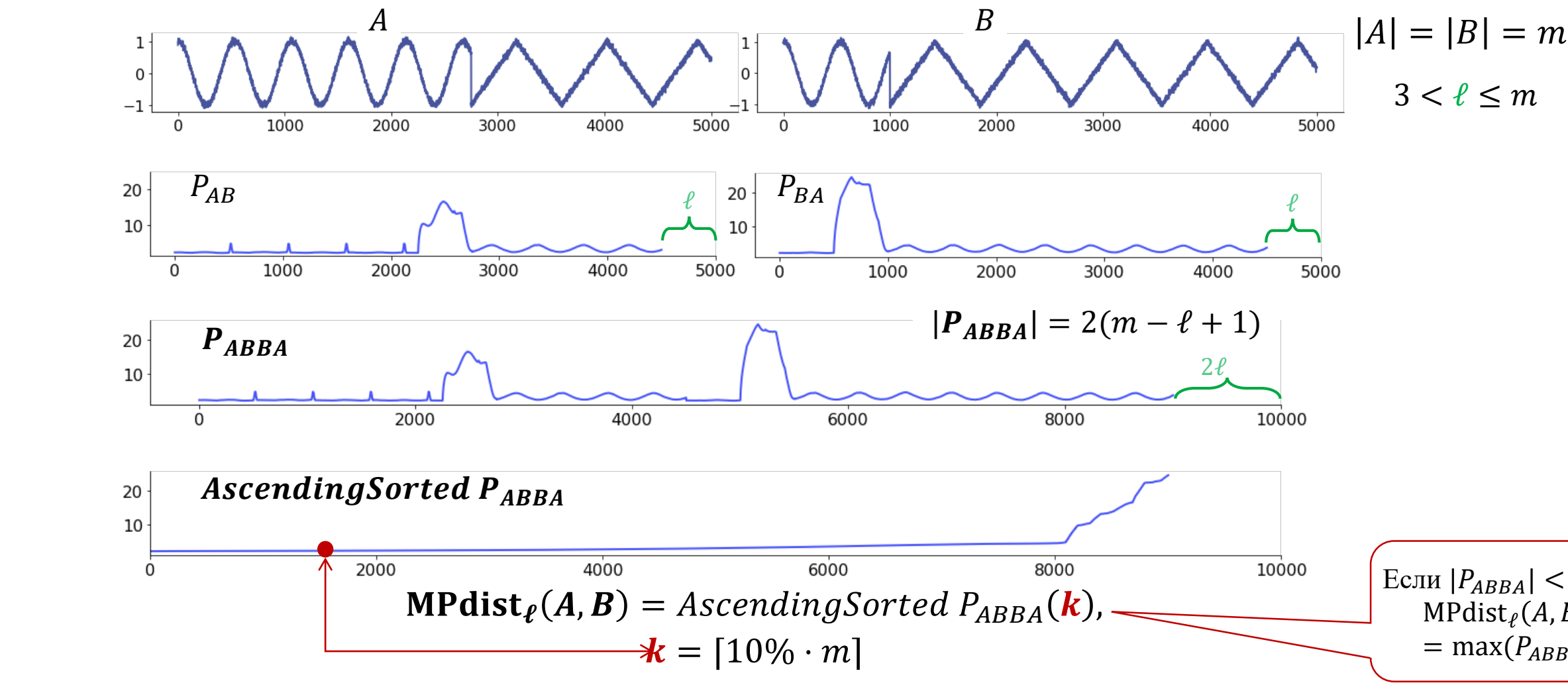
$P_{ABBA} = P_{AB} \bullet P_{BA}$



# Мера MPdist, формальное определение: $Sorted P_{ABBA}$

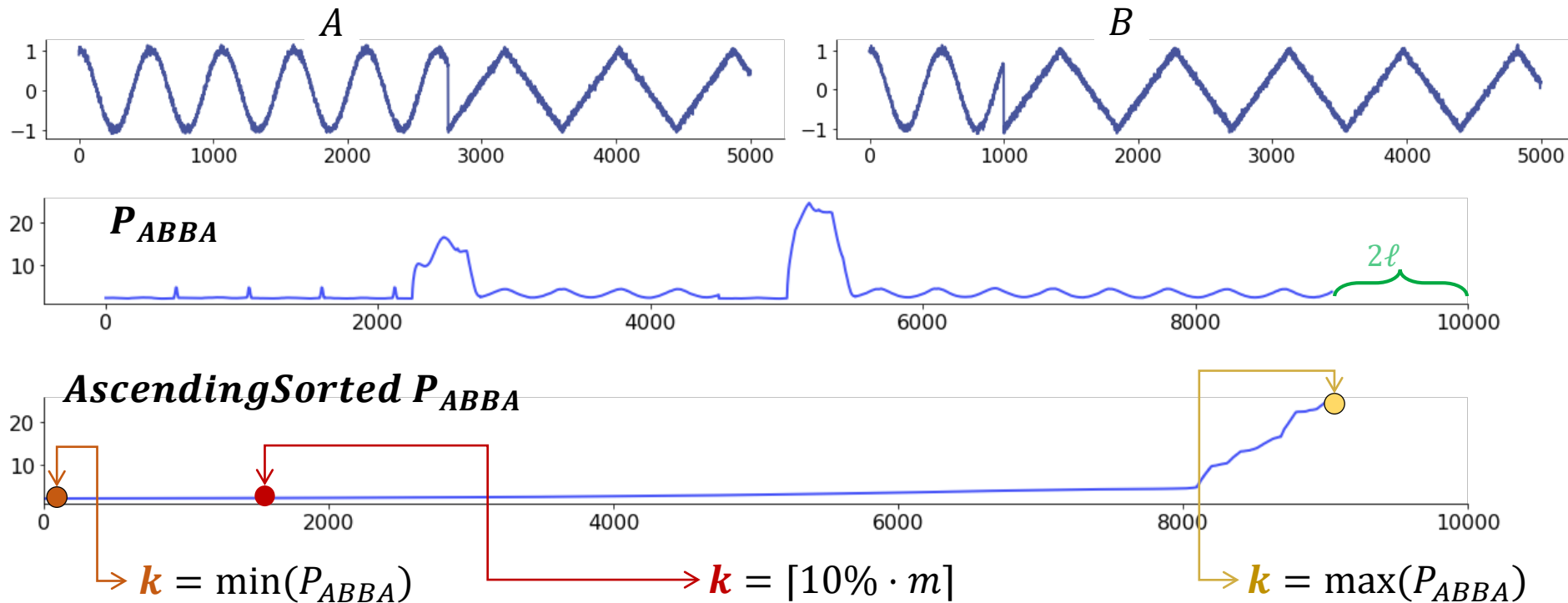


# Мера MPdist, формальное определение: $Sorted P_{ABBA}$





# MPdist: Параметр $k$



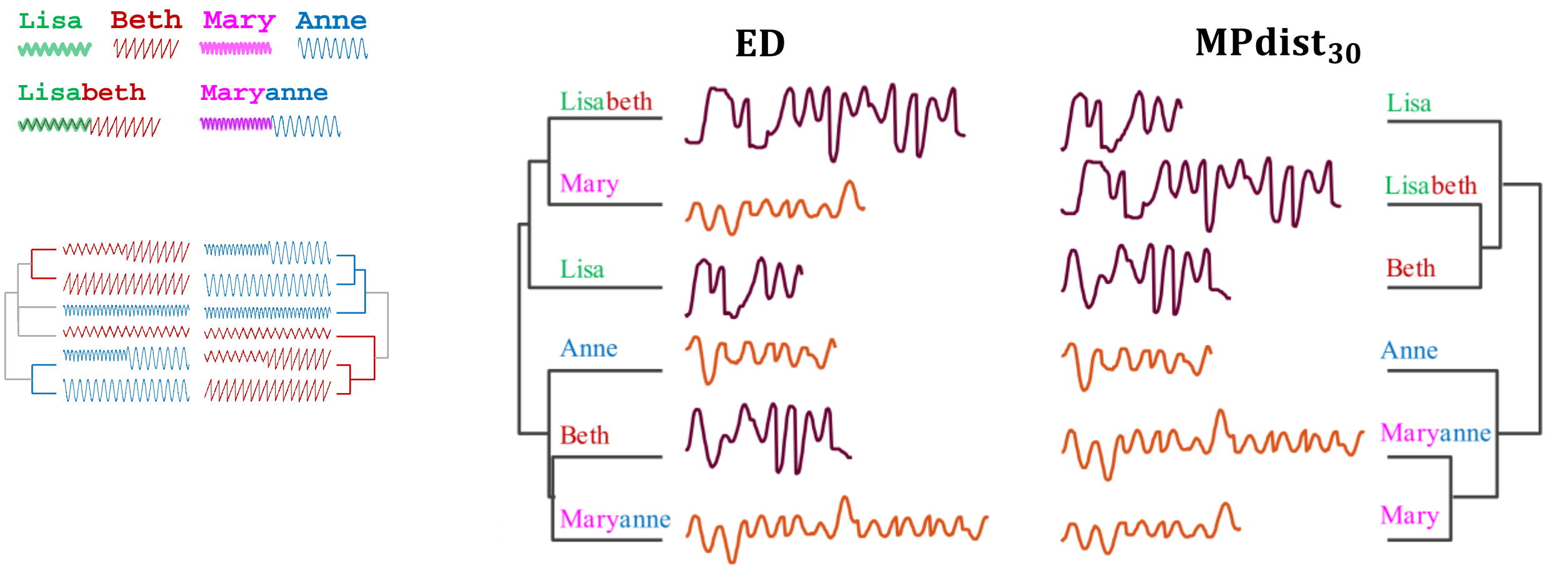
$$|A| = |B| = m$$

$$3 < \ell \leq m$$

$$MPdist_{\ell}(A, B) = AscendingSorted P_{ABBA}(k)$$

- $k = \min(P_{ABBA})$  – сходство многих рядов (ср. “a”, “the” в англ. тексте)
- $k = \max(P_{ABBA})$  – влияние шумов/выбросов во многих
- $k = [5\% \cdot 2m] = [10\% \cdot m]$  – подобрано эмпирически

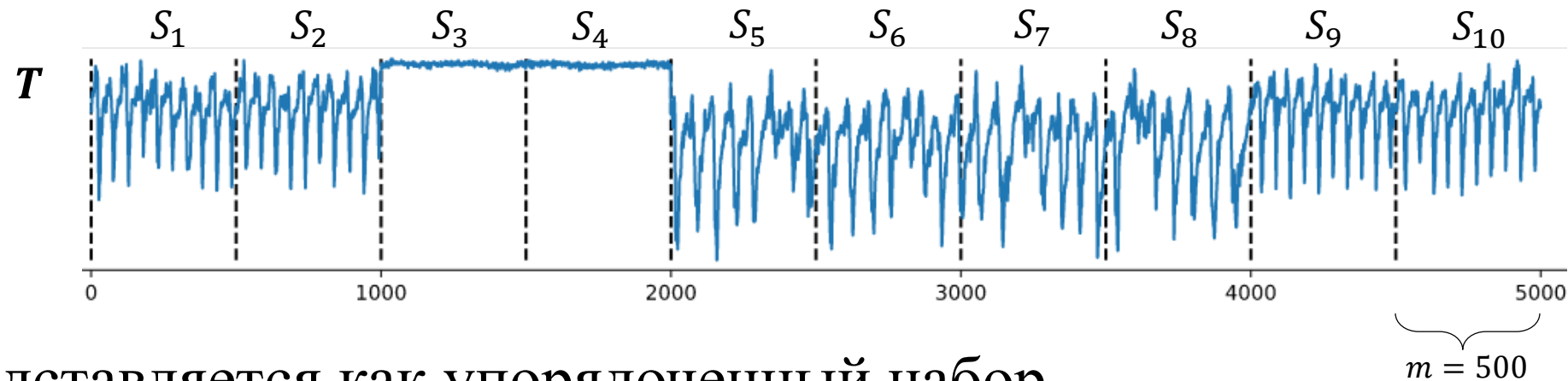
# Применение MPdist



# Содержание

- Мера MPdist
- **Сниппеты и алгоритм Snippet-Finder**
- Применение сниппетов

# Сниппеты (snippets) – наиболее типичные подпоследовательности



1. Ряд представляется как упорядоченный набор непересекающихся сегментов заданной длины:

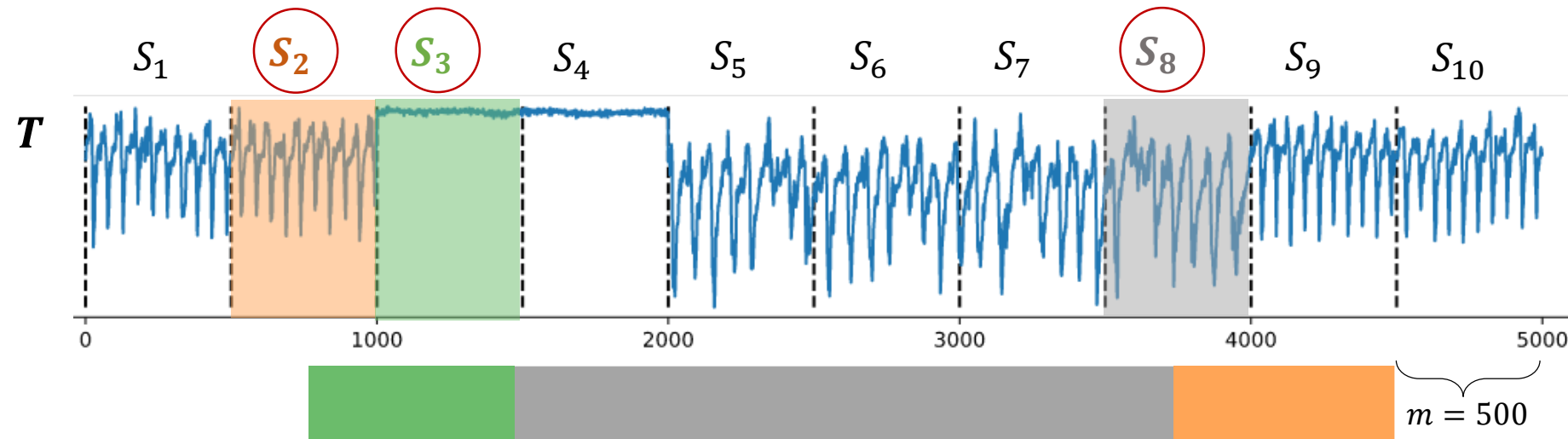
$$S = \{S_i\}_{i=1}^{n/m}, S_i = T_{m(i-1)+1, m}$$

Если  $n$  не кратно  $m$ , то ряд дополняется нулями справа

Сниппетами будут избранные сегменты: те, на которые более похожи многие другие подпоследовательности ряда

Imani S. *et al.* Introducing time series snippets: a new primitive for summarizing long time series. *Data Min. Knowl. Discov.* 2020. 34(6). 1713–1743. DOI: [10.1007/s10618-020-00702-y](https://doi.org/10.1007/s10618-020-00702-y)

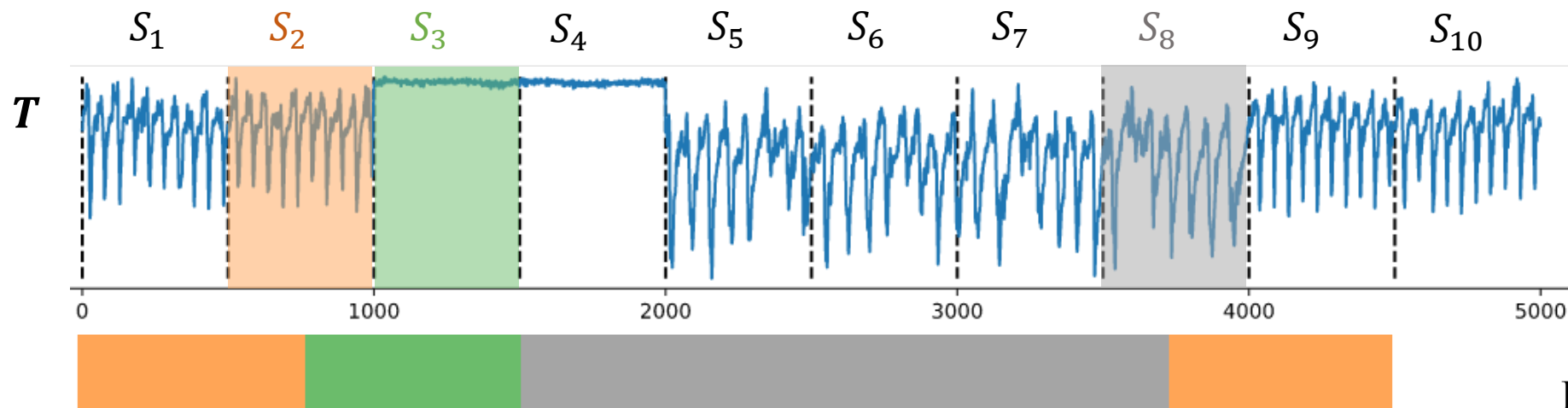
# Сниппеты (snippets) – наиболее типичные подпоследовательности



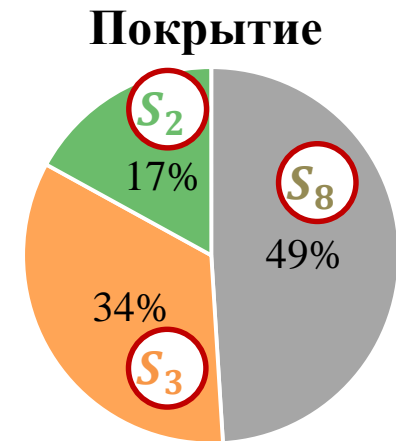
1. Ряд представляется как упорядоченный набор непересекающихся сегментов заданной длины
2. Для каждого сегмента найдем его ближайших соседей в смысле MPdist

Imani S. *et al.* Introducing time series snippets: a new primitive for summarizing long time series. *Data Min. Knowl. Discov.* 2020. 34(6). 1713–1743. DOI: [10.1007/s10618-020-00702-y](https://doi.org/10.1007/s10618-020-00702-y)

# Сниппеты (snippets) – наиболее типичные подпоследовательности

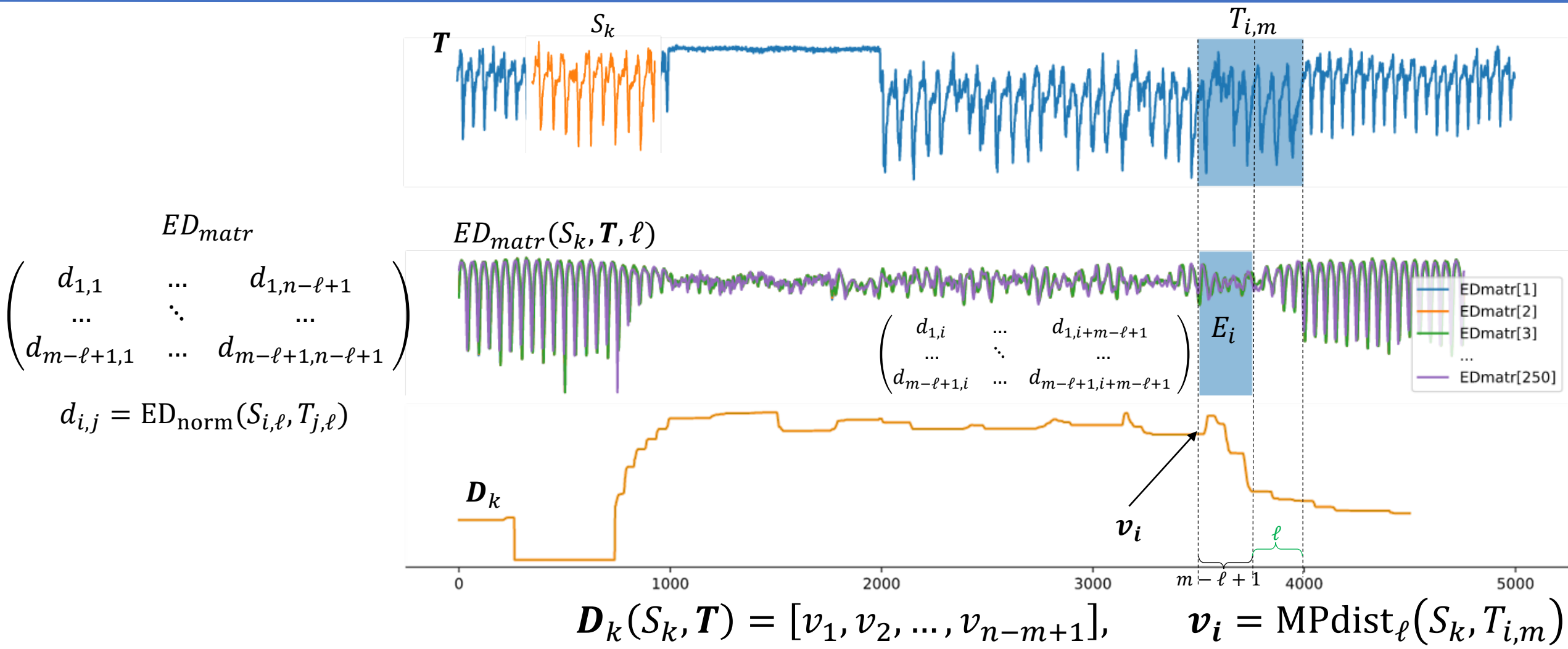
 $K = 3$ 

1. Ряд представляется как упорядоченный набор непересекающихся сегментов заданной длины
2. Для каждого сегмента найдем его ближайших соседей в смысле MPdist
3. Для каждого сегмента найдем его покрытие (долю от  $|S_T^m| = n - m + 1$ ) и возьмем в качестве сниппетов top- $K$  сегментов по покрытию



Imani S. *et al.* Introducing time series snippets: a new primitive for summarizing long time series. Data Min. Knowl. Discov. 2020. 34(6). 1713–1743. DOI: [10.1007/s10618-020-00702-y](https://doi.org/10.1007/s10618-020-00702-y)

# Алгоритм Snippet-Finder: MPdist профили



# Алгоритм Snippet-Finder

**Algorithm 1** SNIPPETFINDER (IN  $T, m, K$ ; OUT  $C_T^m$ )

```

1:  $C_T^m \leftarrow \emptyset; M \leftarrow +\infty$ 
2:  $D \leftarrow \text{GETALLPROFILES}(T, m)$ 
3: while  $|C_T^m| \neq K$  do
4:    $\text{minArea} \leftarrow +\infty$ 
5:   for  $i \leftarrow 1$  to  $n/m$  do
6:      $\text{ProfileArea} \leftarrow \sum_{j=1}^{n-m} \min(D_i(j), M_j)$ 
7:     if  $\text{ProfileArea} < \text{minArea}$  then
8:        $\text{minArea} \leftarrow \text{ProfileArea}; \text{idx} \leftarrow i$ 
9:    $M \leftarrow \{\min(D_{\text{idx}}(i), M_i)\}_{i=1}^{n-m}$ 
10:   $C \leftarrow T_{m \cdot (\text{idx}-1)+1, m}; C.\text{index} \leftarrow \text{idx}$ 
11:   $C_T^m \leftarrow C_T^m \cup C$ 
12: for  $i \leftarrow 1$  to  $K$  do
13:    $f \leftarrow |\{t \in D_{C_i.\text{index}} \mid t = M_i\}|$ 
14:    $C_i.\text{frac} \leftarrow f / (n - m + 1)$ 
15: return  $C_T^m$ 

```

$i = 1:$   $C_1 = \arg \min_{1 \leq j \leq n/m} \text{ProfileArea}(\{D_j\})$   
 $i = 2:$   $C_2 = \arg \min_{1 \leq j \leq n/m} \text{ProfileArea}(\{D_{C_1}, D_j\})$   
 $3 \leq i \leq K:$   $C_i = \arg \min_{1 \leq j \leq n/m} \text{ProfileArea}(\{D_{C_1}, \dots, D_{C_{i-1}}, D_j\})$

**Algorithm 2** GETALLPROFILES (IN  $T, m$ ; OUT  $D$ )

```

1:  $D \leftarrow \emptyset$ 
2: for  $i \leftarrow 1$  to  $n/m$  do
3:    $D_i \leftarrow \text{MPDISTPROFILE}(T, T_{m \cdot (i-1)+1, m})$ 
4:    $D \leftarrow D \cup D_i$ 
5: return  $D$ 

```

**Algorithm 3** MPDISTPROFILE (IN  $T, Q$ ; OUT  $\text{MPD}$ )

```

1:  $\text{MPD} \leftarrow \emptyset$ 
2: for  $i \leftarrow 1$  to  $n - \ell$  do
3:    $d_i \leftarrow \text{MPdist}_\ell(T_{i, m}, Q)$ 
4:    $\text{MPD} \leftarrow \text{MPD} \cup d_i$ 
5: return  $\text{MPD}$ 

```

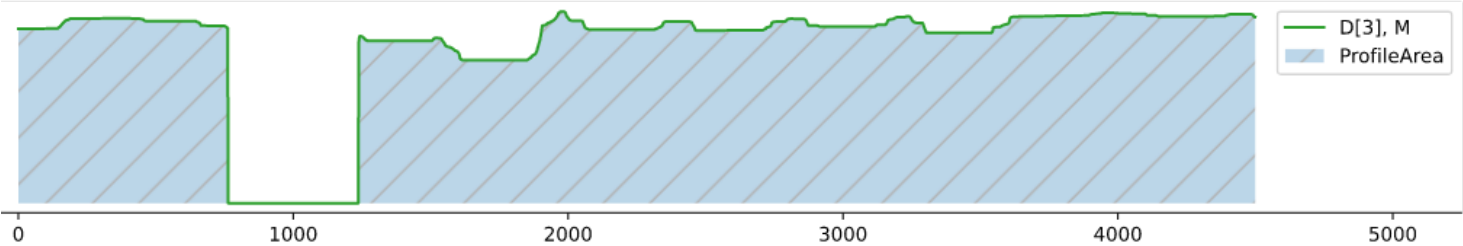


# Алгоритм Snippet-Finder: Поиск снippetsа top-1

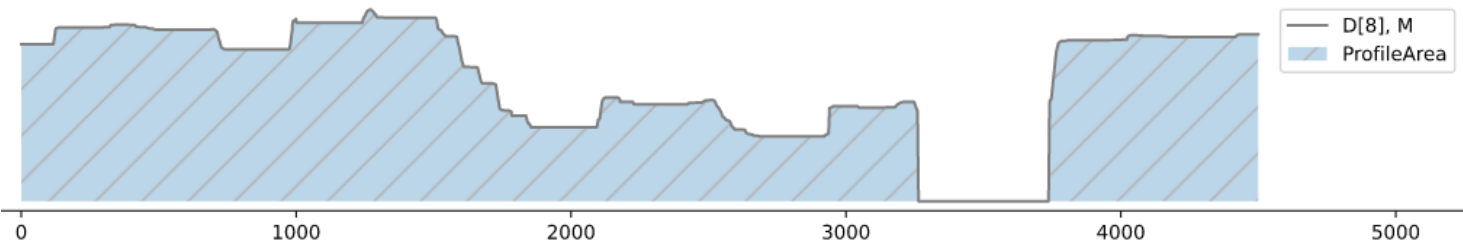
Поиск  $C_1$

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

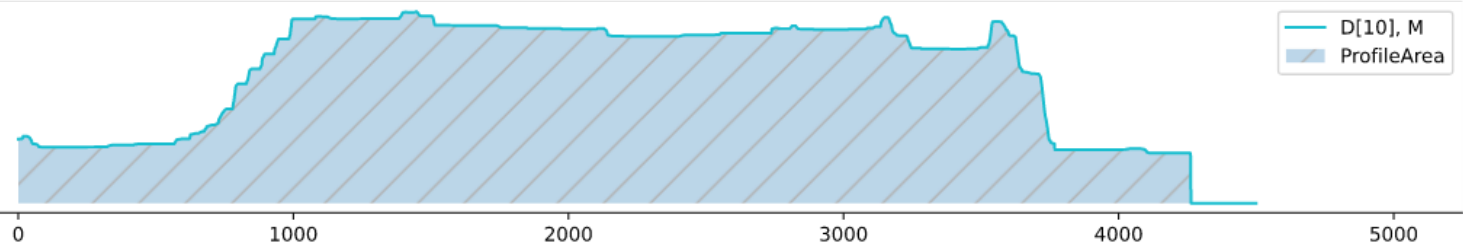
$C_1.index = 8$



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

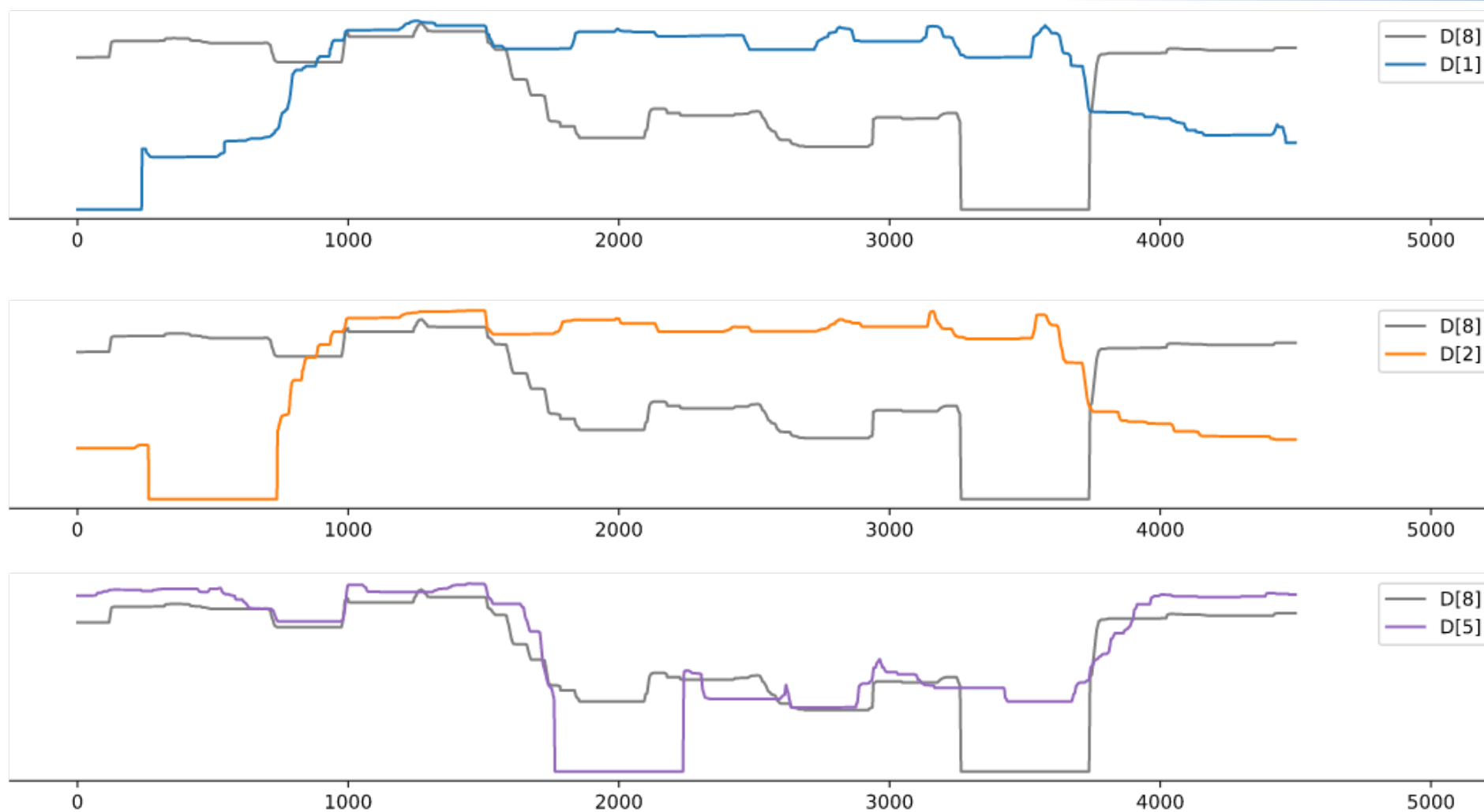


$ProfileArea(\{D_8\}) = 53769$



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

# Алгоритм Snippet-Finder: Поиск снippetsа top-2

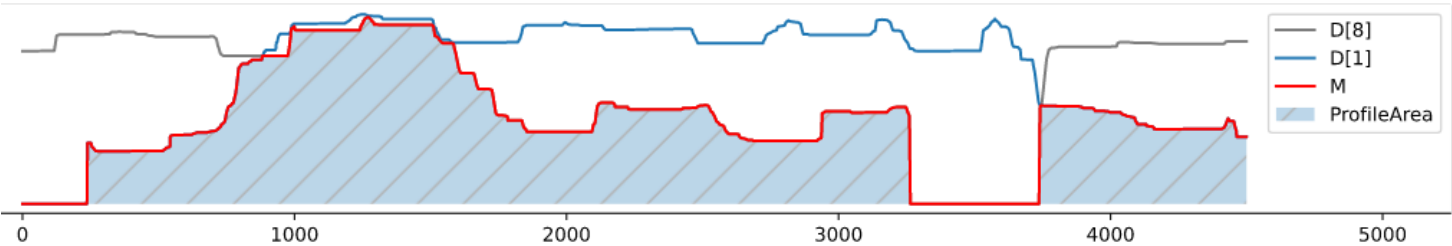


# Алгоритм Snippet-Finder: Поиск снippetsа top-2

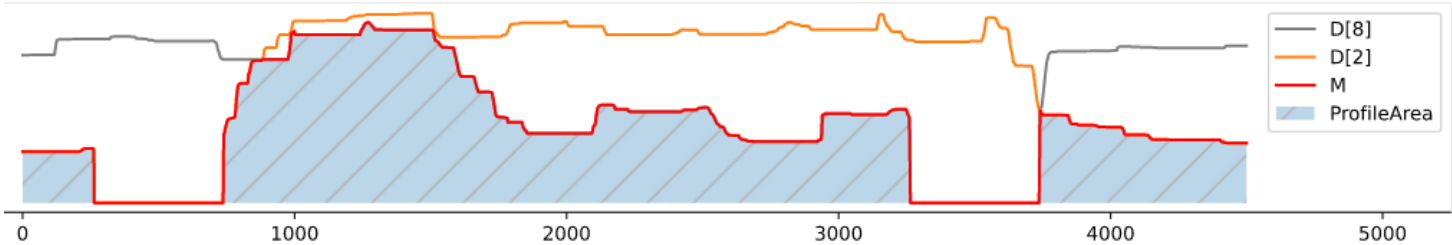
Поиск  $C_2$

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

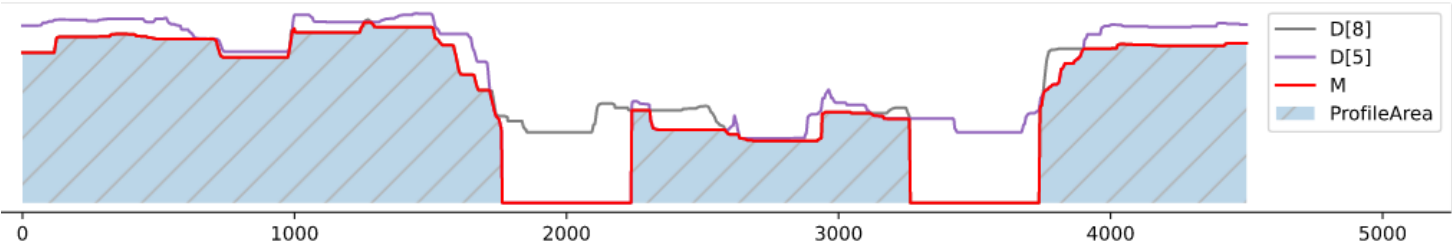
$C_2.index = 2$



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

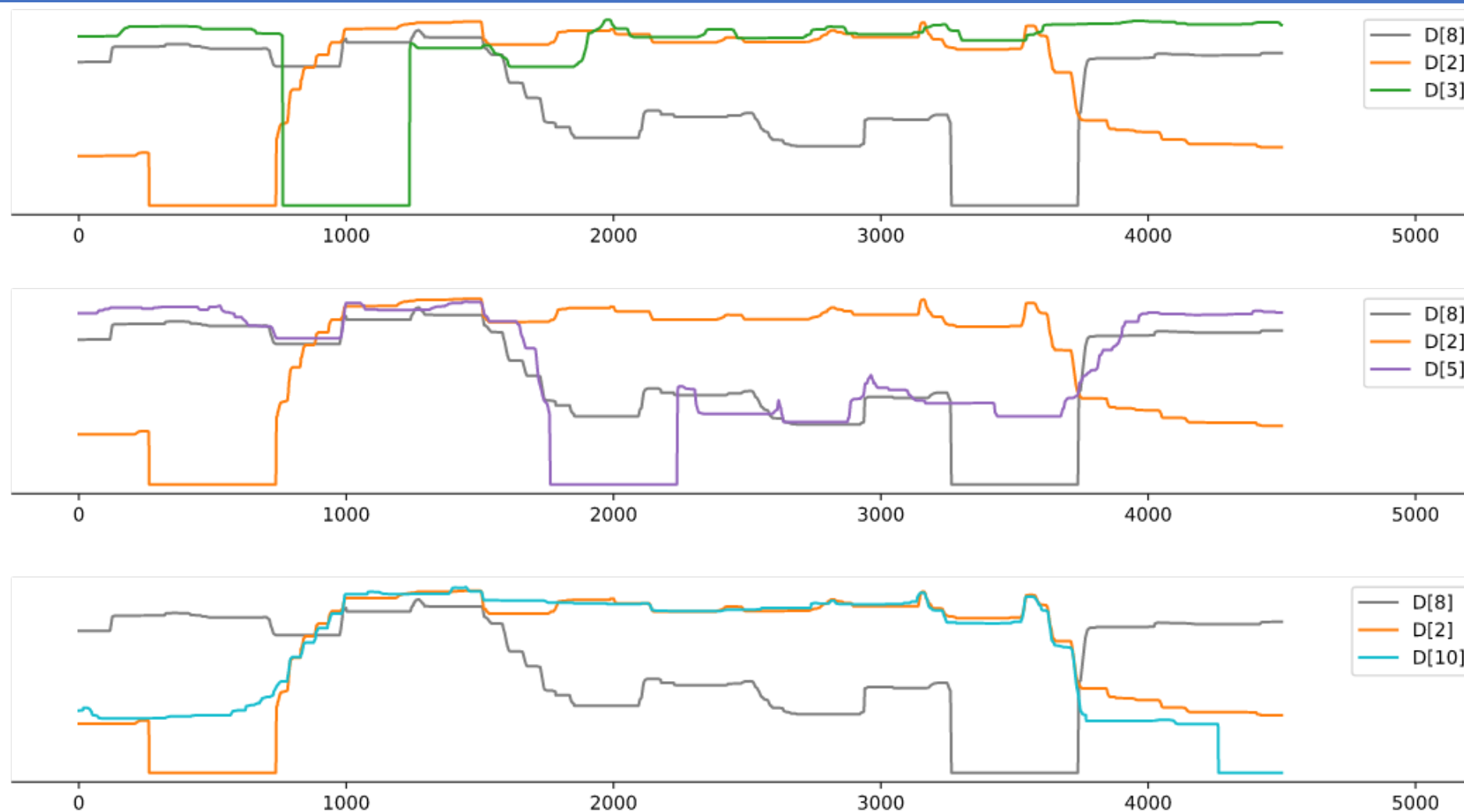


$ProfileArea(\{D_8, D_2\}) = 35769$



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

# Алгоритм Snippet-Finder: Поиск снippetsа top-3

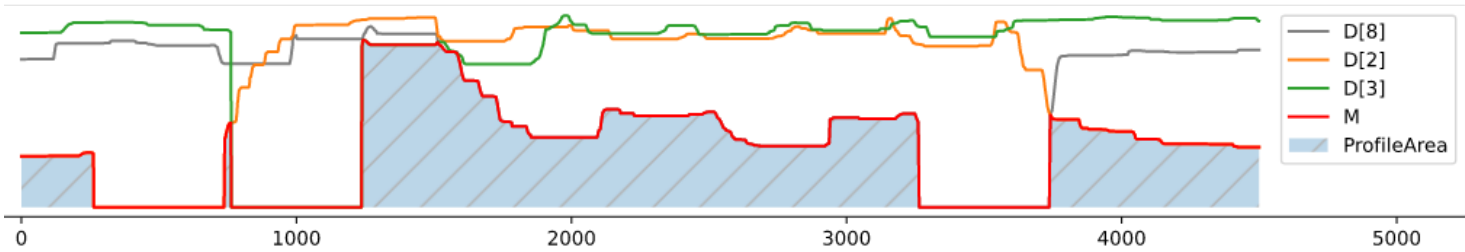


# Алгоритм Snippet-Finder: Поиск снippetsа top-3

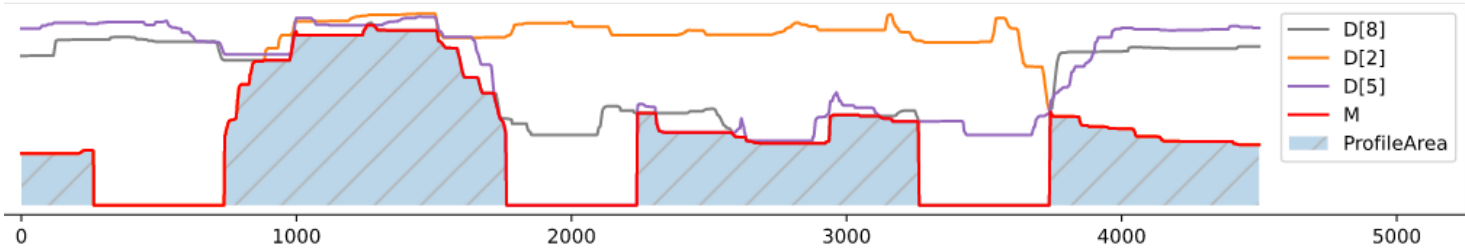
Поиск  $C_3$

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

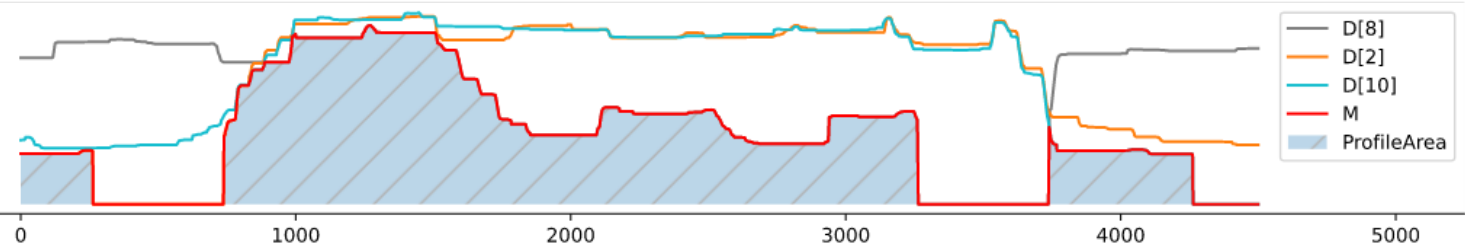
$C_3.index = 3$



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



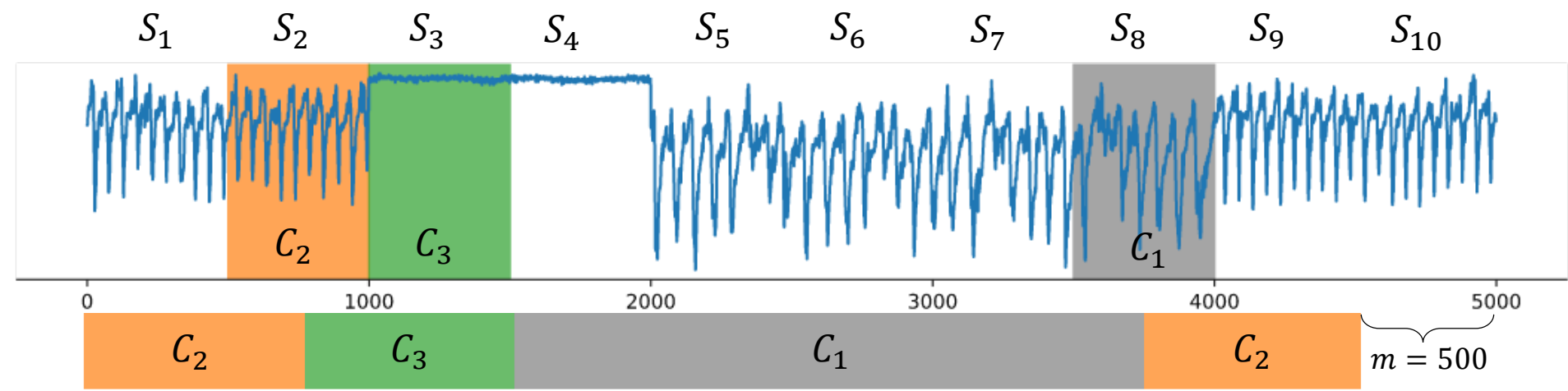
$ProfileArea(\{D_8, D_2, D_5\}) = 31168$



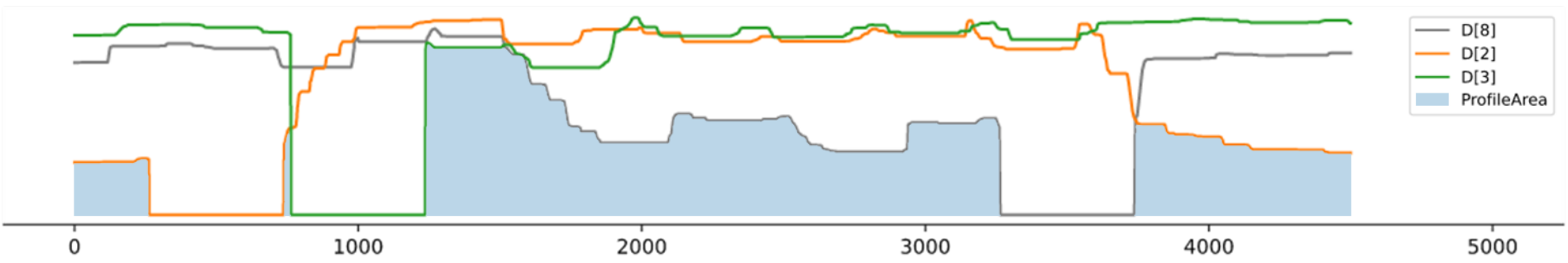
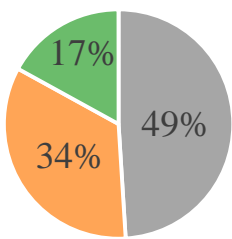
$ProfileArea(\{D_8, D_2, D_{10}\}) = 33044$

# Алгоритм Snippet-Finder: итог

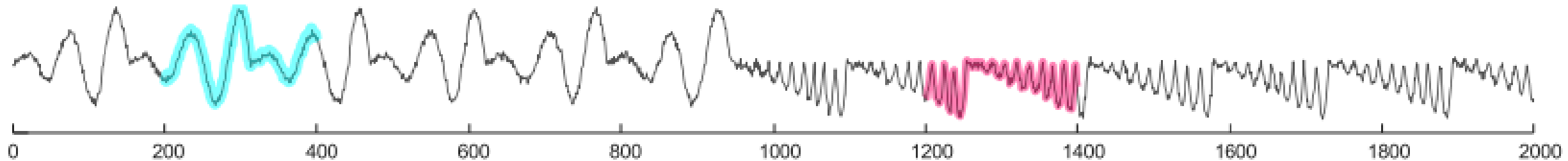
$K = 3$



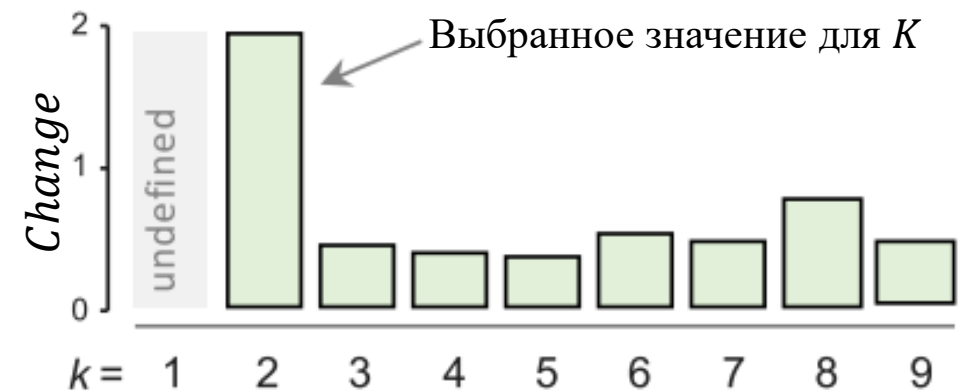
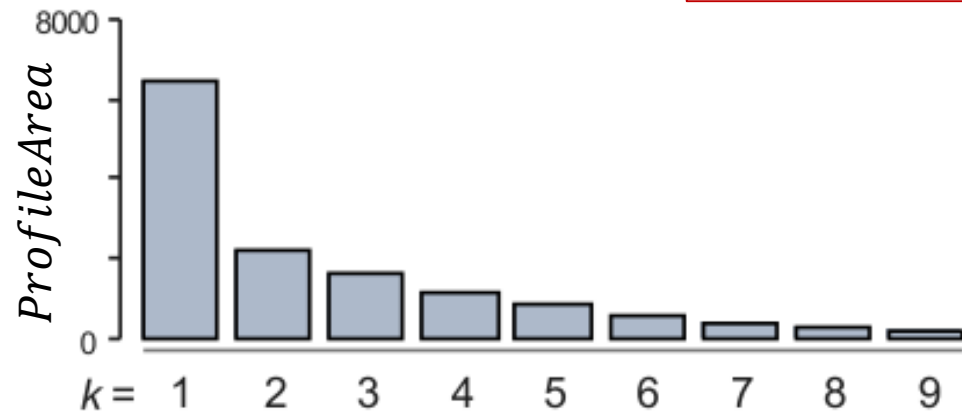
Покрытие



# Подбор количества снippetов $K$



$$Change_k = \frac{ProfileArea_{k-1}}{ProfileArea_k} - 1$$



Большой пик на  $k$ -м значении показывает предпочтительное  $K$

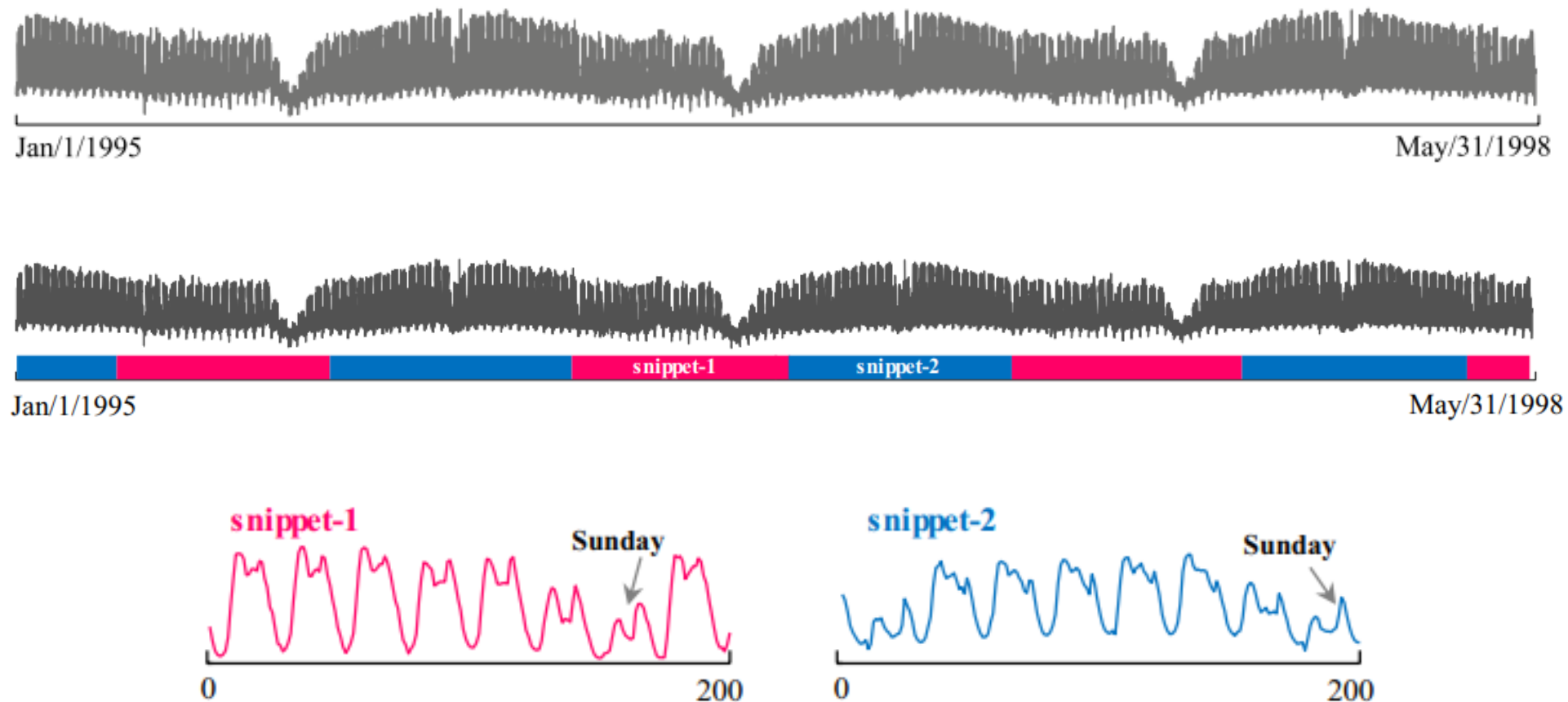
# Содержание

- Мера MPdist
- Снимпеты и алгоритм Snippet-Finder
- **Применение снимпетов**



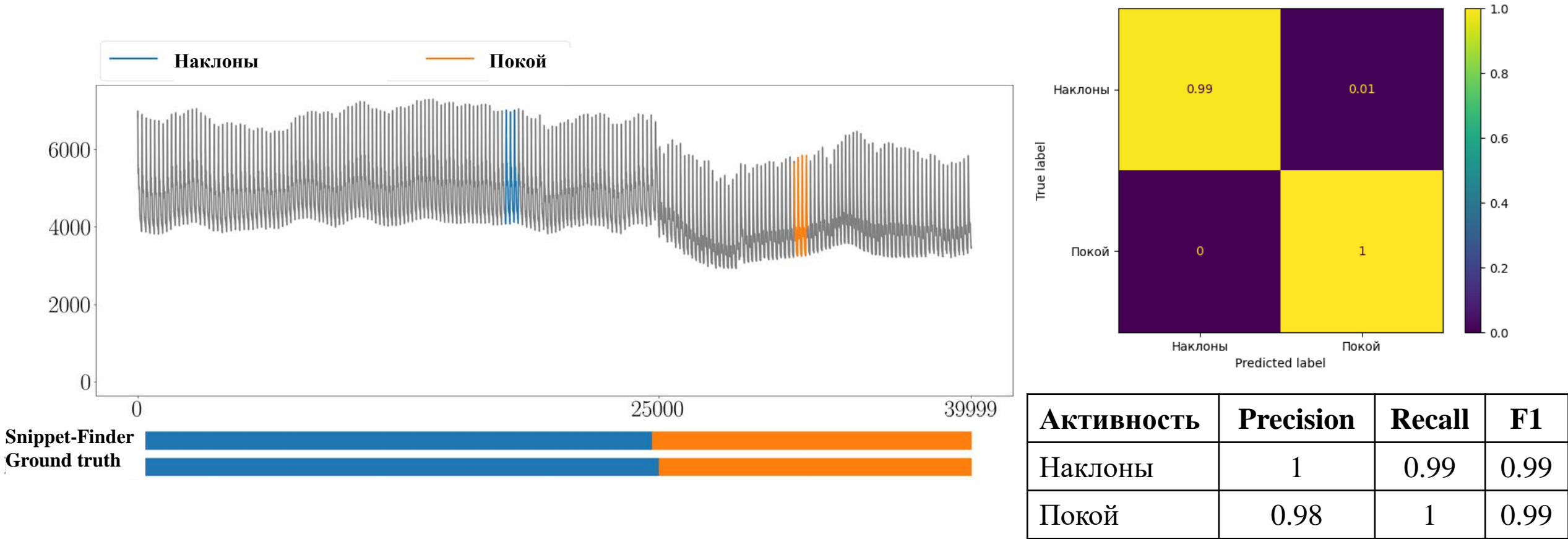
# Применение сниппетов

## Энергопотребление города в Италии за 3 года



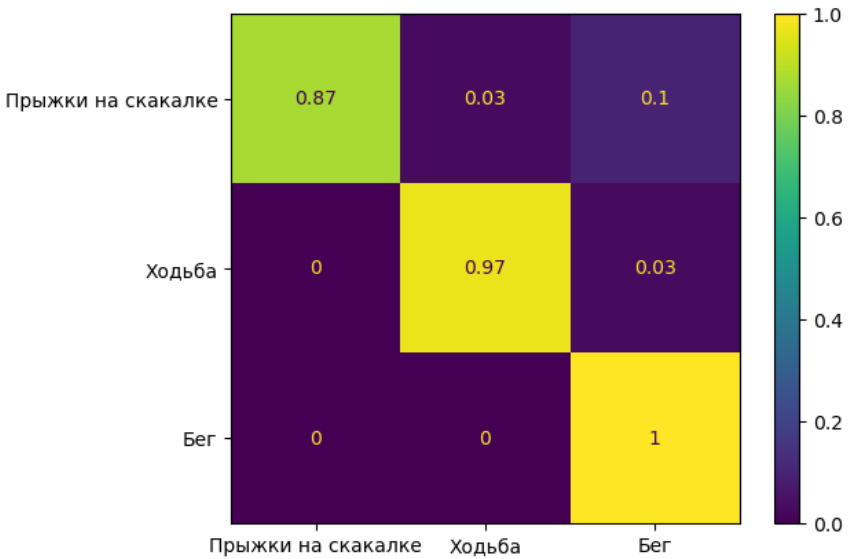
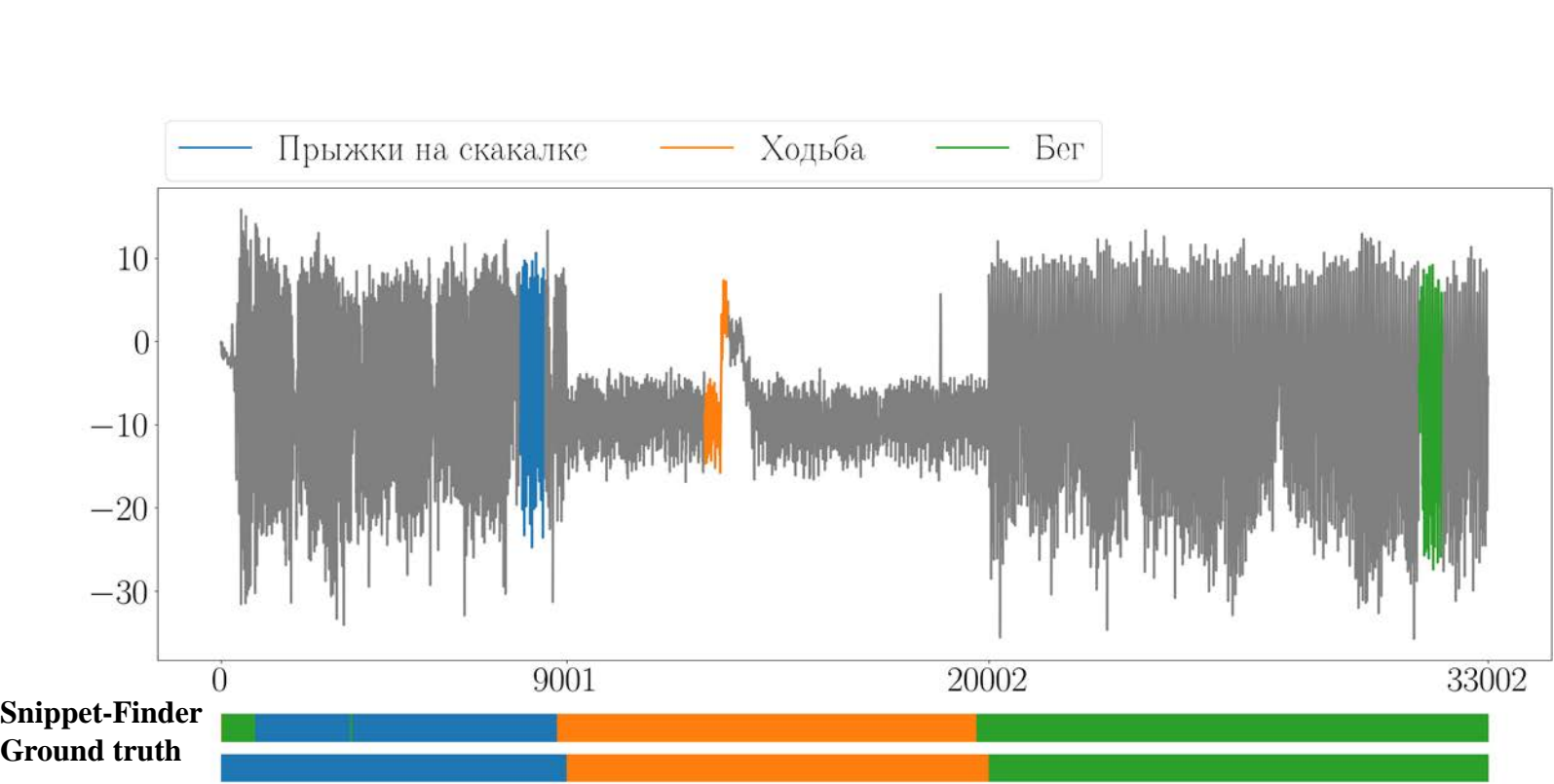
Сниппеты показывают типичное поведение в теплые и холодные сезоны

# Применение сниппетов: распознавание активностей



\* Imani S., et al. 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)

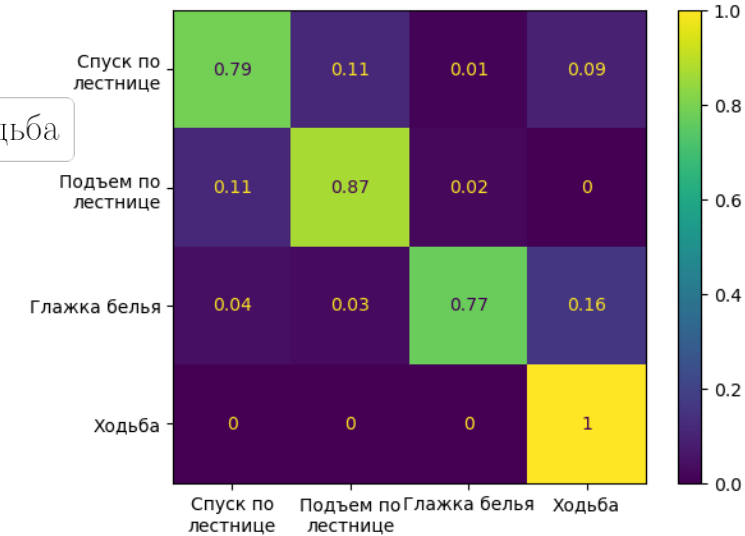
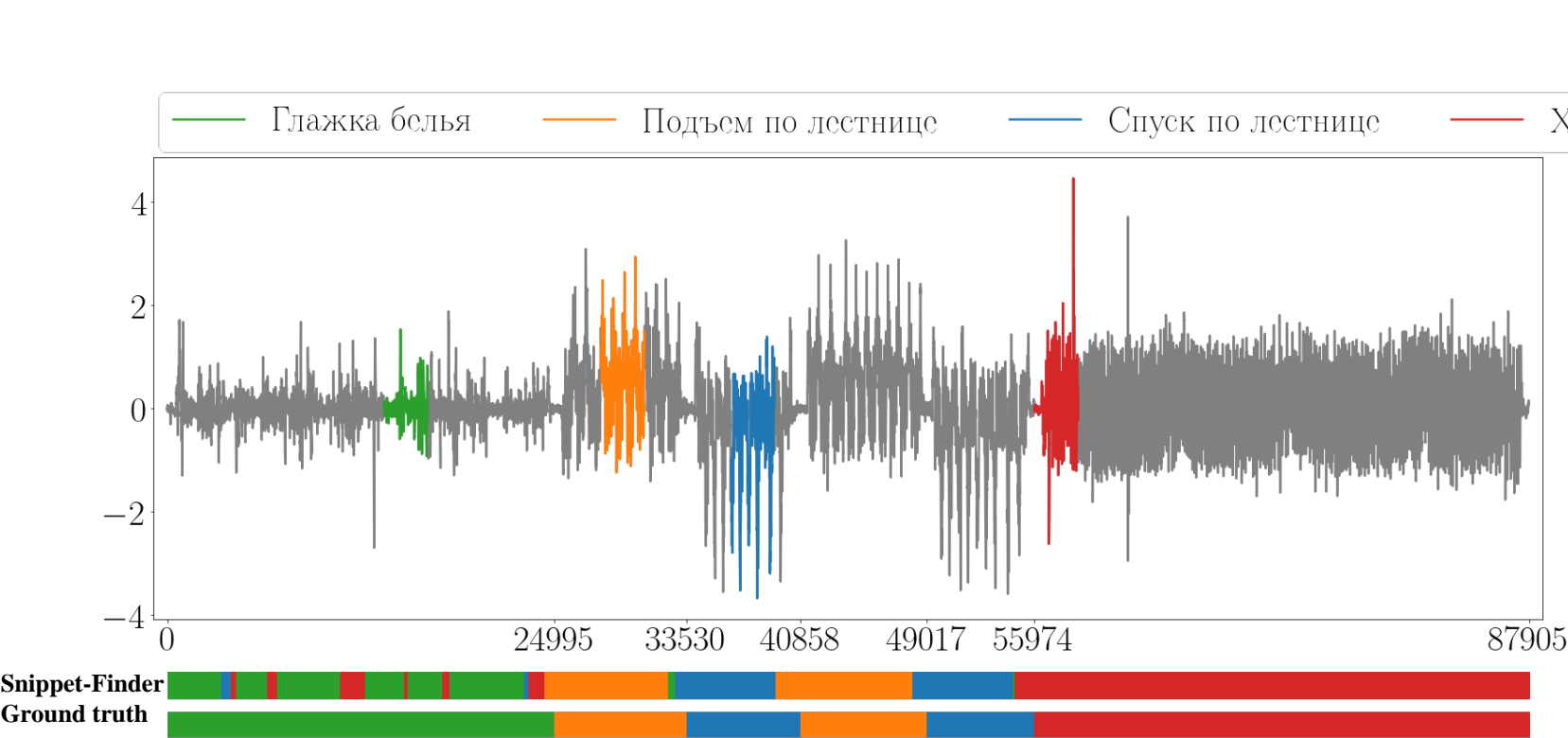
# Применение сниппетов: распознавание активностей



| Активность         | Precision | Recall | F1   |
|--------------------|-----------|--------|------|
| 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](https://doi.org/10.1109/ISWC.2012.13)

# Применение сниппетов: распознавание активностей



| Активность         | Precision | Recall | F1   |
|--------------------|-----------|--------|------|
| 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](https://doi.org/10.1109/ISWC.2012.13)

# Литература

1. Gharghabi S., Imani S., Bagnall A.J., Darvishzadeh A., Keogh E.J. An ultra-fast time series distance measure to allow data mining in more complex real-world deployments. *Data Min. Knowl. Discov.* 2020. 34, pp. 1104–1135. <https://doi.org/10.1007/s10618-020-00695-8>.
2. Imani S., Madrid F., Ding W., Crouter S.E., Keogh E.J. Matrix Profile XIII: Time series snippets: A new primitive for time series data mining. *Proc. of the 2018 IEEE Int. Conf. on Big Knowledge, ICBK 2018, Singapore, 17–18 November 2018.* pp. 382–389. <https://doi.org/10.1109/ICBK.2018.00058>.
3. 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.* 2020. 34. pp. 1713–1743. <https://doi.org/10.1007/s10618-020-00702-y>.