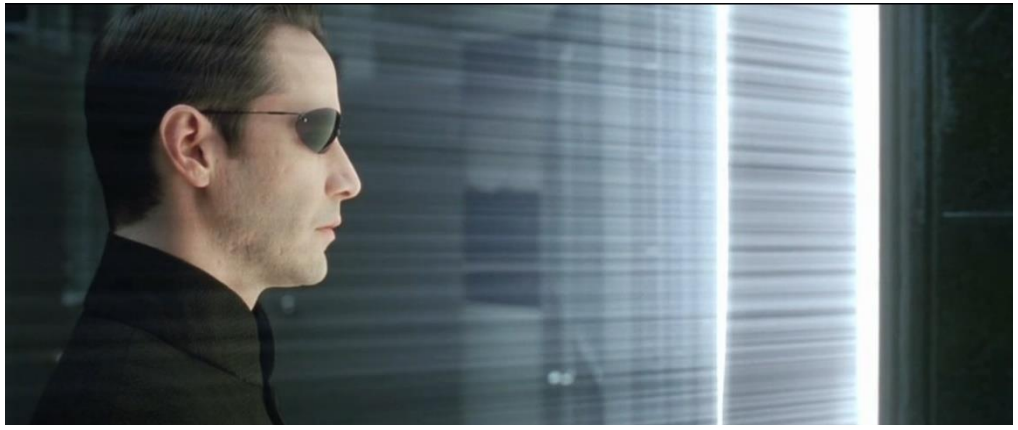
Матричный профиль временного ряда



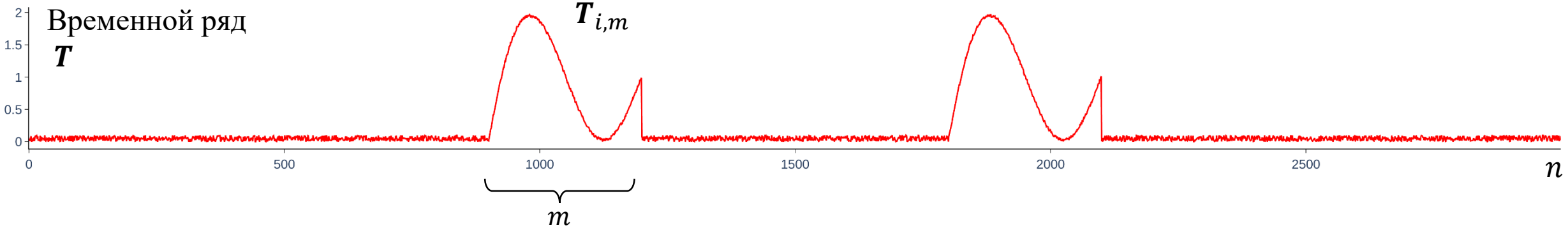
What is the Matrix? Control.

Morpheus

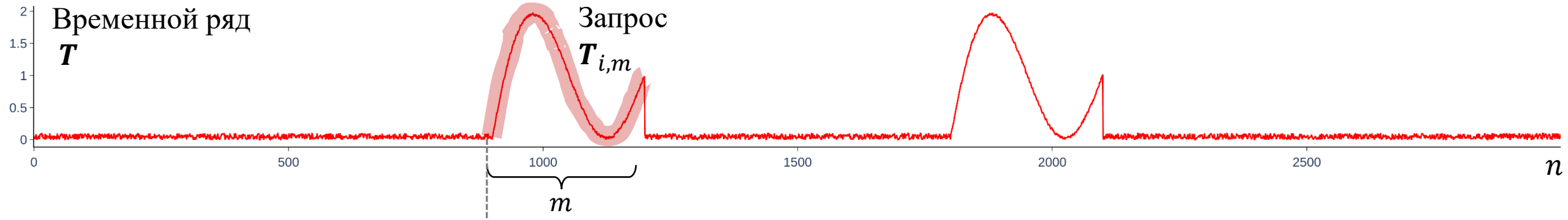
Содержание

- Понятие матричного профиля
- Примеры задач, решаемых на основе матричного профиля
- Алгоритмы вычисления матричного профиля

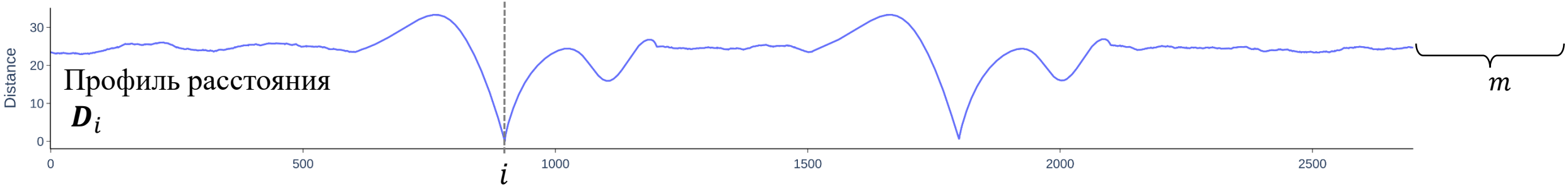
Матричный профиль: Временной ряд



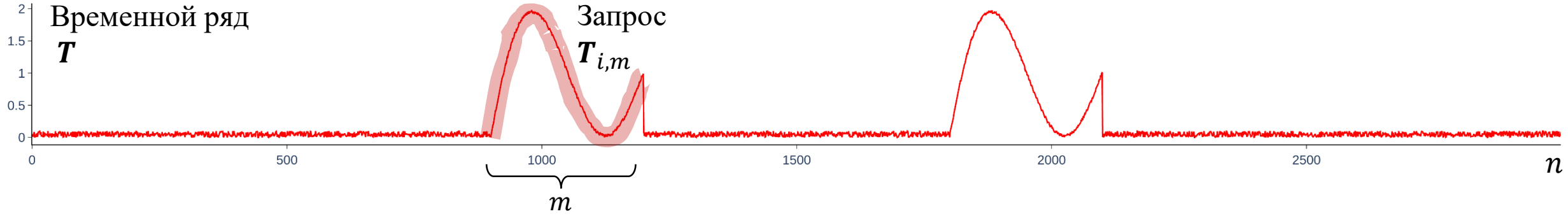
Матричный профиль: Профиль расстояния



$$D_i = \left\{\mathrm{Dist}\left(Q, T_{j,m}\right)\right\}_{j=1}^{n-m+1}, \quad Q = T_{i,m}, \quad 1 \le i \le n-m+1$$

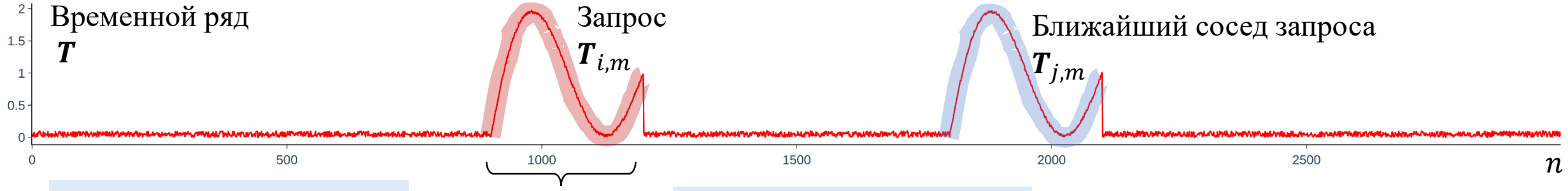


Матричный профиль: Матрица профилей расстояния



	$T_{1,m}$	$T_{2,m}$	…	$T_{i,m}$	…	$T_{n-m+1,m}$
D_1	0	$\text{Dist}(T_{1,m}, T_{2,m})$				
D_2	$\text{Dist}(T_{2,m}, T_{1,m})$	0				
…	…		0	…		
D_i	$\text{Dist}(T_{i,m}, T_{1,m})$	$\text{Dist}(T_{i,m}, T_{2,m})$	…	0		
…	…				0	…
D_{n-m+1}	$\text{Dist}(T_{n-m+1,m}, T_{1,m})$	$\text{Dist}(T_{n-m+1,m}, T_{2,m})$	…	$\text{Dist}(T_{n-m+1,m}, T_{i,m})$	…	0

Матричный профиль: Поиск ближайших соседей



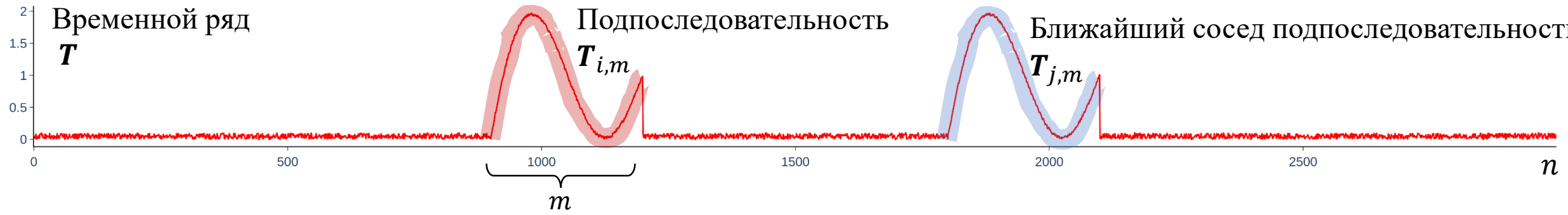
$$\min_{\substack{1\le i<n-m+1 \\ |i-1|\ge m}} \mathrm{Dist}(T_{i,m}, T_{1,m})$$

$$\min_{\substack{1\le i<n-m+1 \\ |i-j|\ge m}} \mathrm{Dist}(T_{i,m}, T_{j,m})$$

Соседи не пересекаются: $|i-j| \ge m$

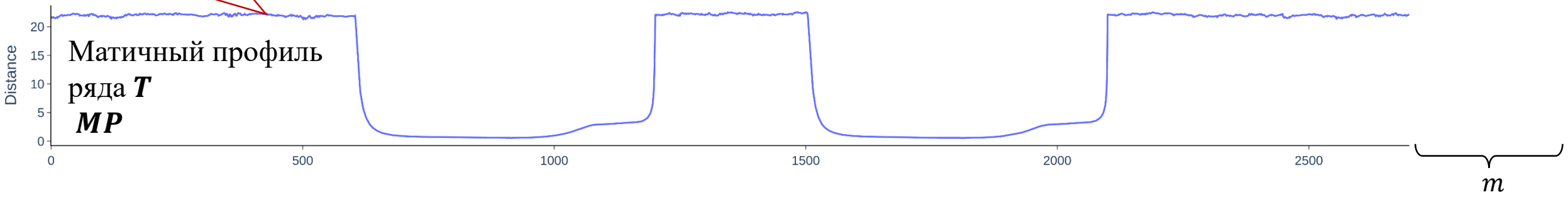
	$T_{1,m}$	$T_{2,m}$	…	$T_{i,m}$	…	$T_{n-m+1,m}$
D_1	0	$\mathrm{Dist}(T_{1,m}, T_{2,m})$				
D_2	$\mathrm{Dist}(T_{2,m}, T_{1,m})$	0				
…	…	…	0	…		
D_i	$\mathrm{Dist}(T_{i,m}, T_{1,m})$	$\mathrm{Dist}(T_{i,m}, T_{2,m})$	…	0		
…	…	…			0	…
D_{n-m+1}	$\mathrm{Dist}(T_{n-m+1,m}, T_{1,m})$	$\mathrm{Dist}(T_{n-m+1,m}, T_{2,m})$	…	$\mathrm{Dist}(T_{n-m+1,m}, T_{i,m})$	…	0

Матричный профиль

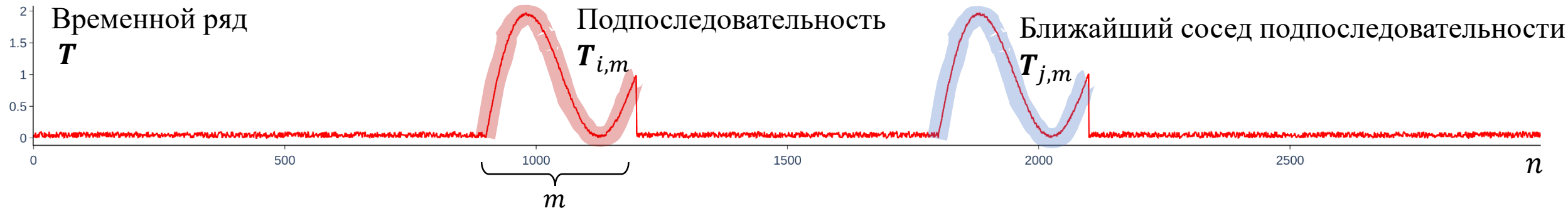


Расстояния до ближайшего соседа подпоследовательностей ряда

$$MP_T^m(i) = \min_{\substack{1 \le j \le n-m+1 \\ |i-j| \ge m}} \mathrm{Dist}\left(T_{i,m}, T_{j,m}\right)$$

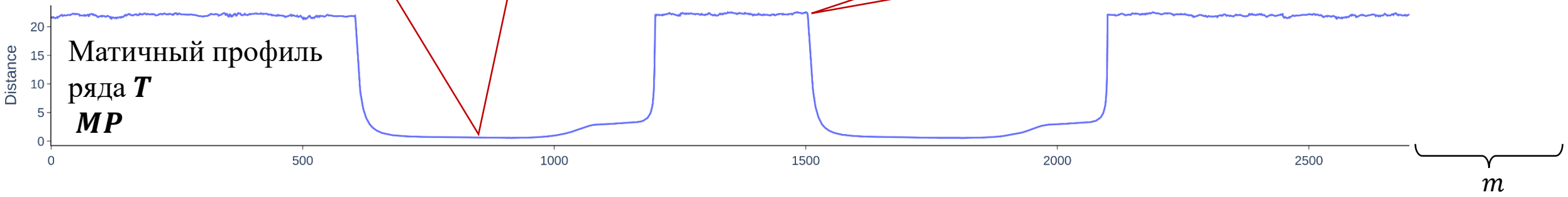


Простое понимание матричного профиля

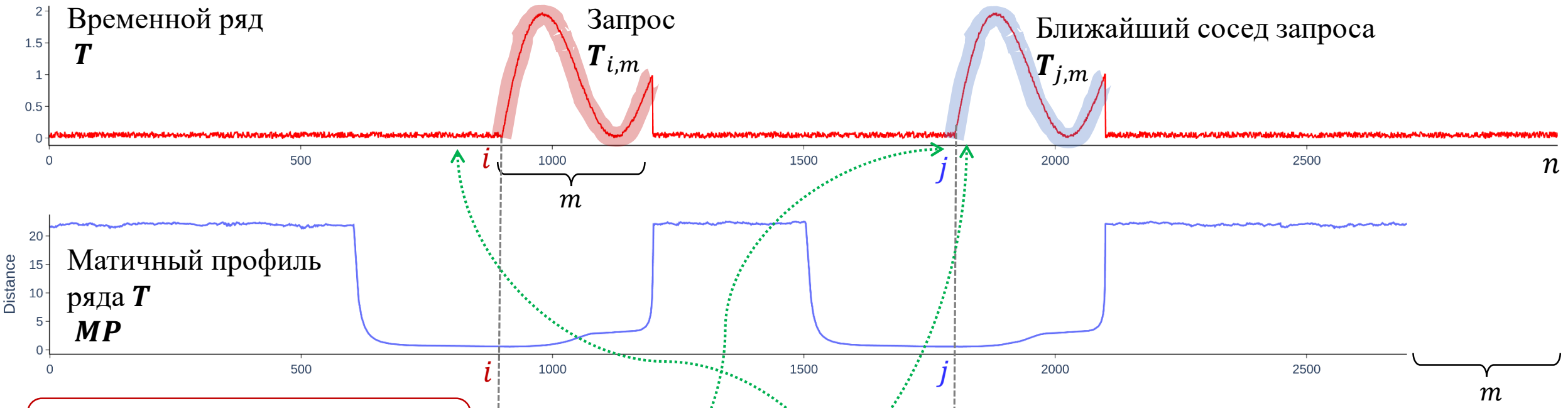


Локальные минимумы МП соответствуют мотивам (шаблонам) ряда

Локальные максимумы МП соответствуют диссонансам (аномалиям) ряда



Индекс матричного профиля



Индекс ближайшего соседа подпоследовательностей ряда

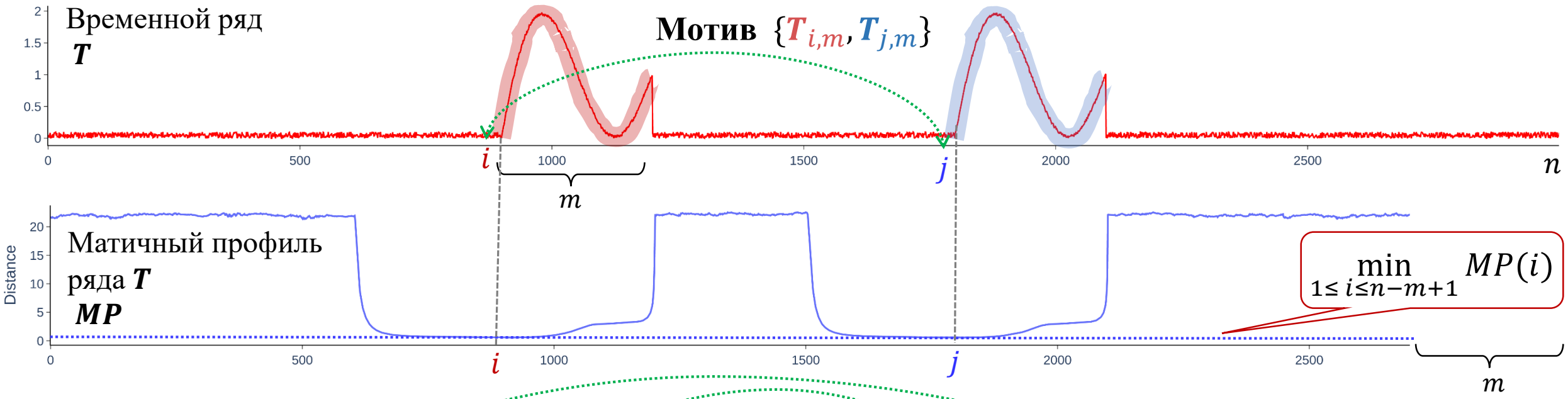
$$MPI_T^m(i) = \arg \min_{\substack{1 \le j \le n-m+1 \\ |i-j| \ge m}} \mathrm{Dist}\left(T_{i,m}, T_{j,m}\right)$$

Индекс матричного профиля ряда T

MPI

...	1785	1789	1793	...	885	890	890	...

Индекс МП *не* симметричен в общем случае



Индекс МП ряда T

	92		879	890			1788	1789	1790		2571	
MPI	...	2571	...	1785	**1789**	1791	...	885	**890**	890	...	314	...

$MP(i) = j$ и $MP(j) = i$ только в случае мотива

Функция $\mathbf{Dist}(\cdot,\cdot)$ для матричного профиля

- ED
- ED^2
- $\mathbf{ED_{norm}}$
- $\mathbf{ED^2_{norm}}$
- DTW
- Hamming
- …

Можно использовать любую функцию расстояния (метрику или не-метрику), это вопрос двух факторов:

- релевантность функции предметной области
- сложность (быстрота) вычисления МП

Обобщение: Матричный профиль соединения (Join MP)

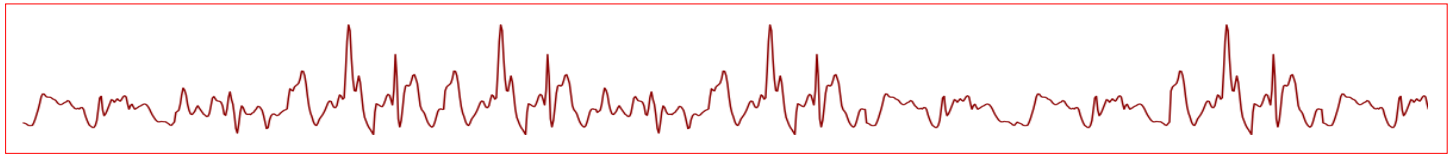
- Ряды
  $A, |A| = n_A, B, |B| = n_B$
- Профиль расстояния
  $D_i = \{\text{Dist}(A_{i,m}, B_{j,m})\}_{j=1}^{n_B-m+1}$,
  $1 \le i \le n_A - m + 1$
- МП соединения
  $MPjoin_{AB}^{m}(i) = \min_{1 \le j \le n_B-m+1} \text{Dist}(A_{i,m}, B_{j,m})$,
  $1 \le i \le n_A - m + 1$
- Индекс МП соединения
  $MPIjoin_{AB}^{m}(i) = \arg\min_{1 \le j \le n_B-m+1} \text{Dist}(A_{i,m}, B_{j,m})$,
  $1 \le i \le n_A - m + 1$

Нет условия недопустимости тривиального совпадения

В общем случае построение МП соединения – **не** коммутативная операция:

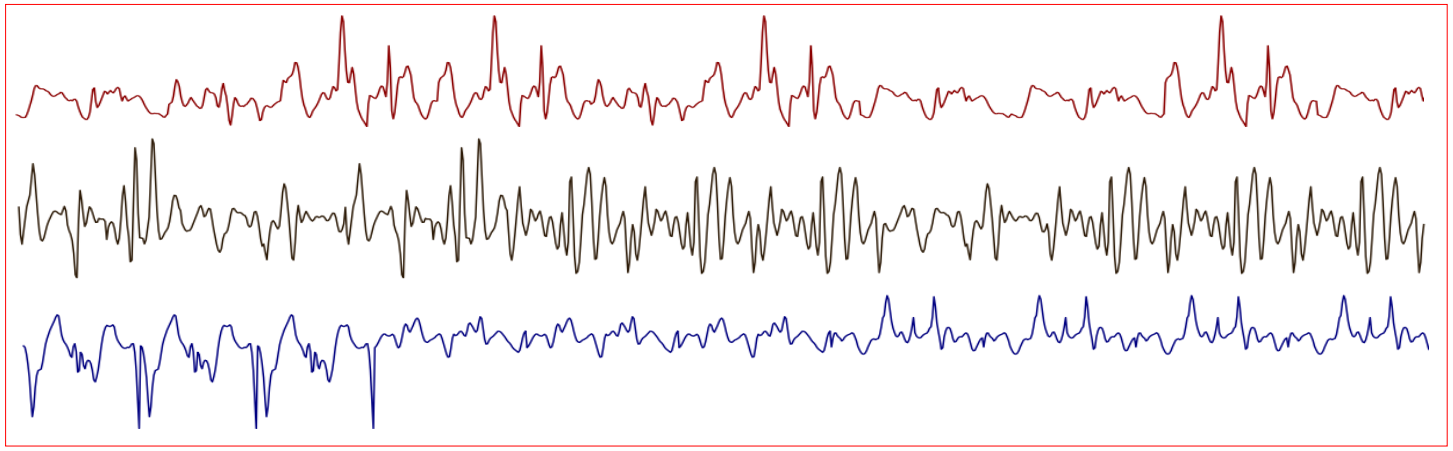
$$MPjoin_{AB}^{m} \neq MPjoin_{BA}^{m}$$

Матричный профиль многомерного ряда



Dist(,)

$ED, ED^2,$
$ED_{norm}, ED^2_{norm},$
DTW, …



Dist(,)

Агрегация

Dist(,)

Dist(,)

Dist(,)

$$\text{median}(\{ED^2_{norm}(\cdot,\cdot)\}_{i=1}^{d})$$

Содержание

- Понятие матричного профиля
- **Примеры задач, решаемых на основе матричного профиля**
- Алгоритмы вычисления матричного профиля

Матричный профиль – базис для решения большинства задач интеллектуального анализа временных рядов*

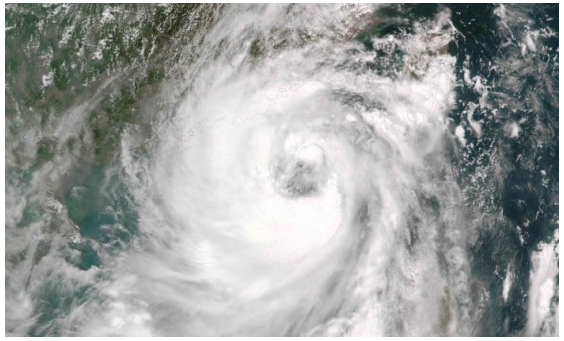
- Диссонансы
- Мотивы
- Шейплеты
- Сниппеты
- Цепочки
- Сравнение рядов
- …

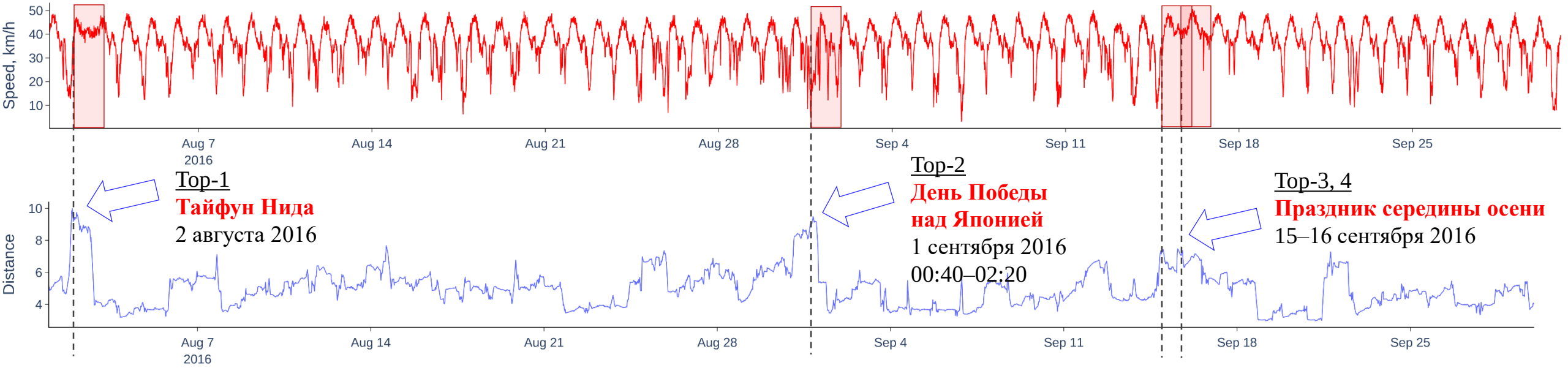
Имонн Кеог
(Калифорнийский университет в Риверсайде, США)
Eamonn Keogh
(University of California, Riverside, USA)

* Zhu Y. et al. The Swiss army knife of time series data mining: Ten useful things you can do with the matrix profile and ten lines of code. Data Min. Knowl. Discov. 34(4): 949-979 (2020). DOI: 10.1007/s10618-019-00668-6.

Поиск аномалий

Скорость городского трафика Гуанчжоу*

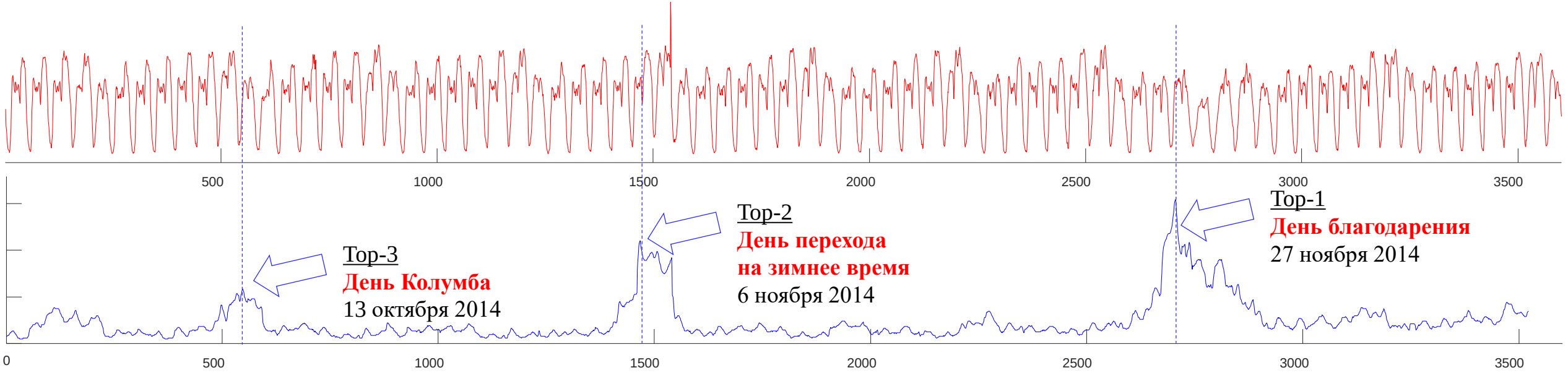




* Chen X., Chen Y., He Z. Urban traffic speed dataset of Guangzhou, China. 2018. DOI: 10.5281/zenodo.1205229.

Поиск аномалий

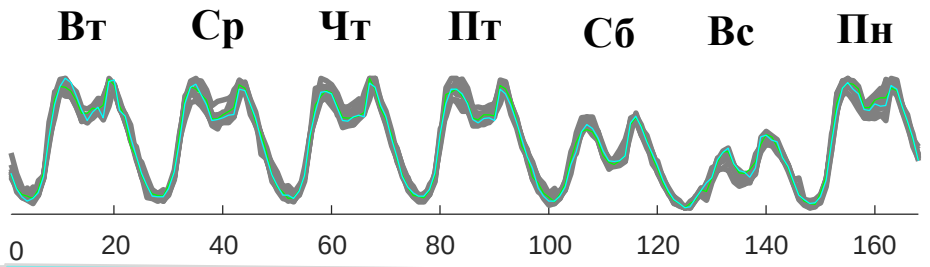
Среднее число пассажиров нью-йоркского такси (осень 2014 г., каждые полчаса)*

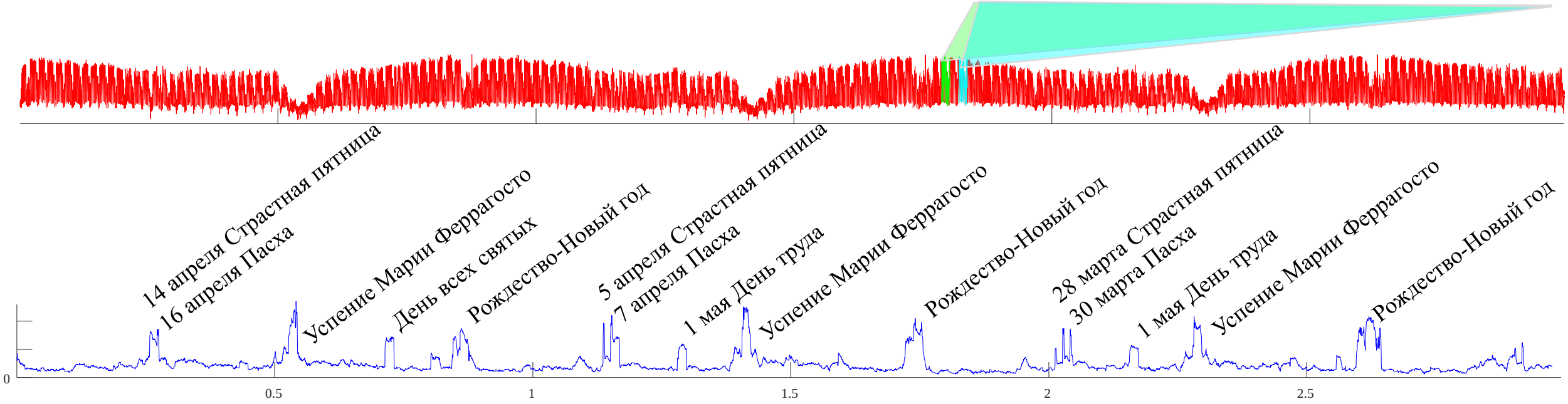


* 2014 New York City Taxi Trips. URL: https://www.kaggle.com/datasets/kentonnlp/2014-new-york-city-taxi-trips.

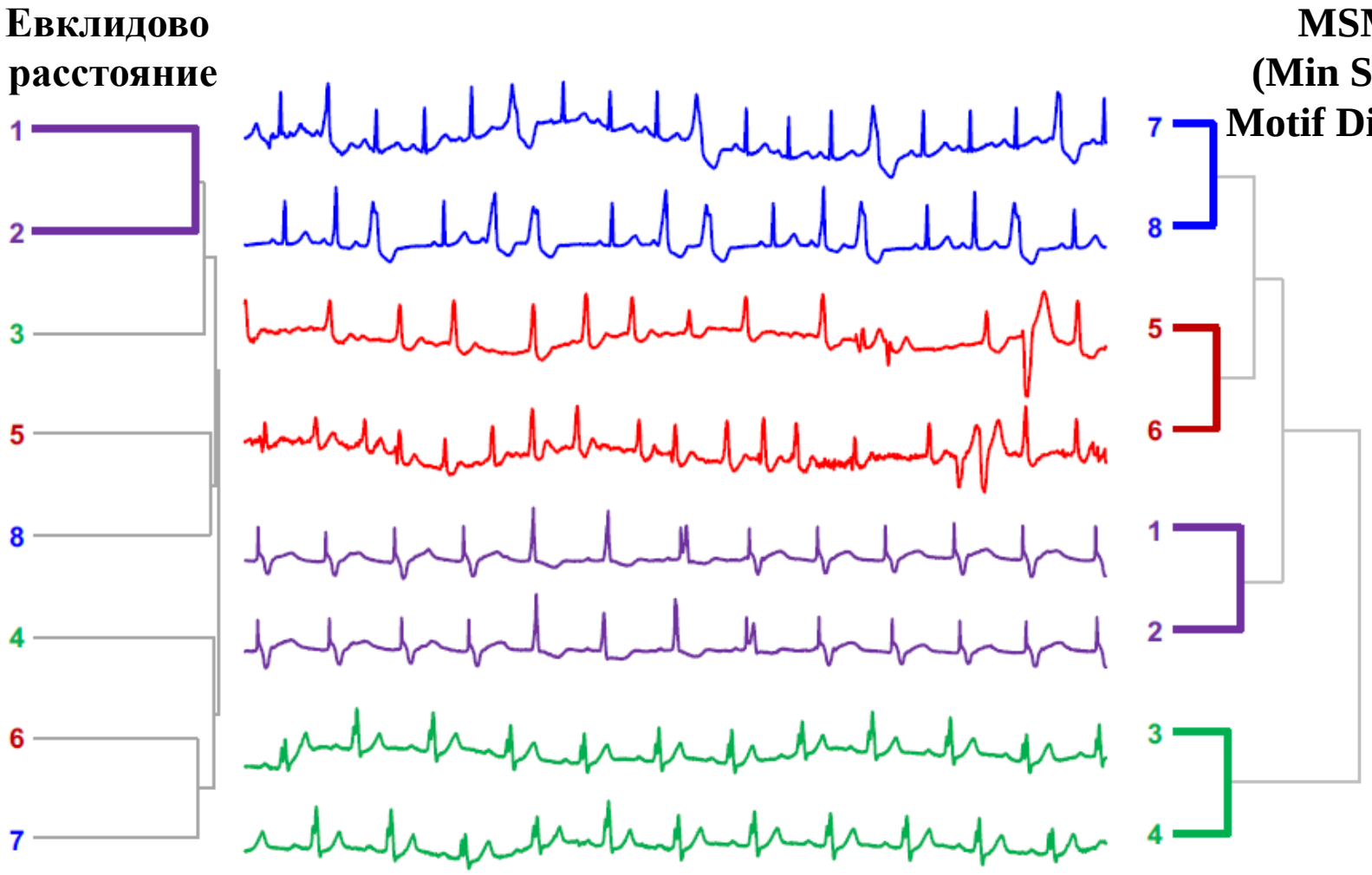
Поиск мотивов

Энергопотребление в Италии 1995-1998 гг.





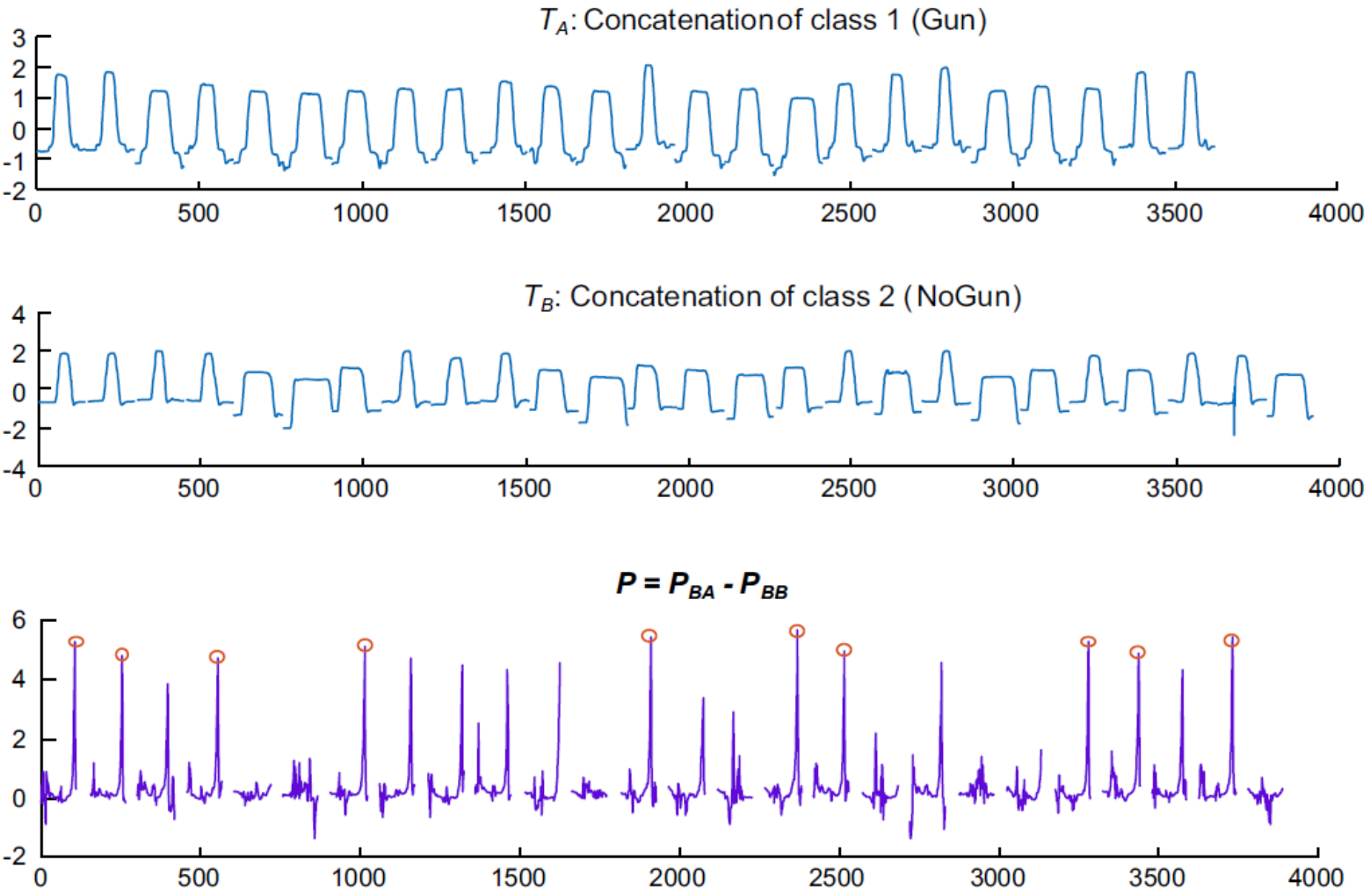
Сравнение временных рядов



- Сравнение 10 с фрагментов ЭКГ четырех пациентов
- Кластеризация на основе евклидова расстояния дает не адекватные итоги, в отличие от применения матричного профиля

$$\{P, I\} := MPjoin(A, B, m)$$
$$MSMD := \min(P)$$

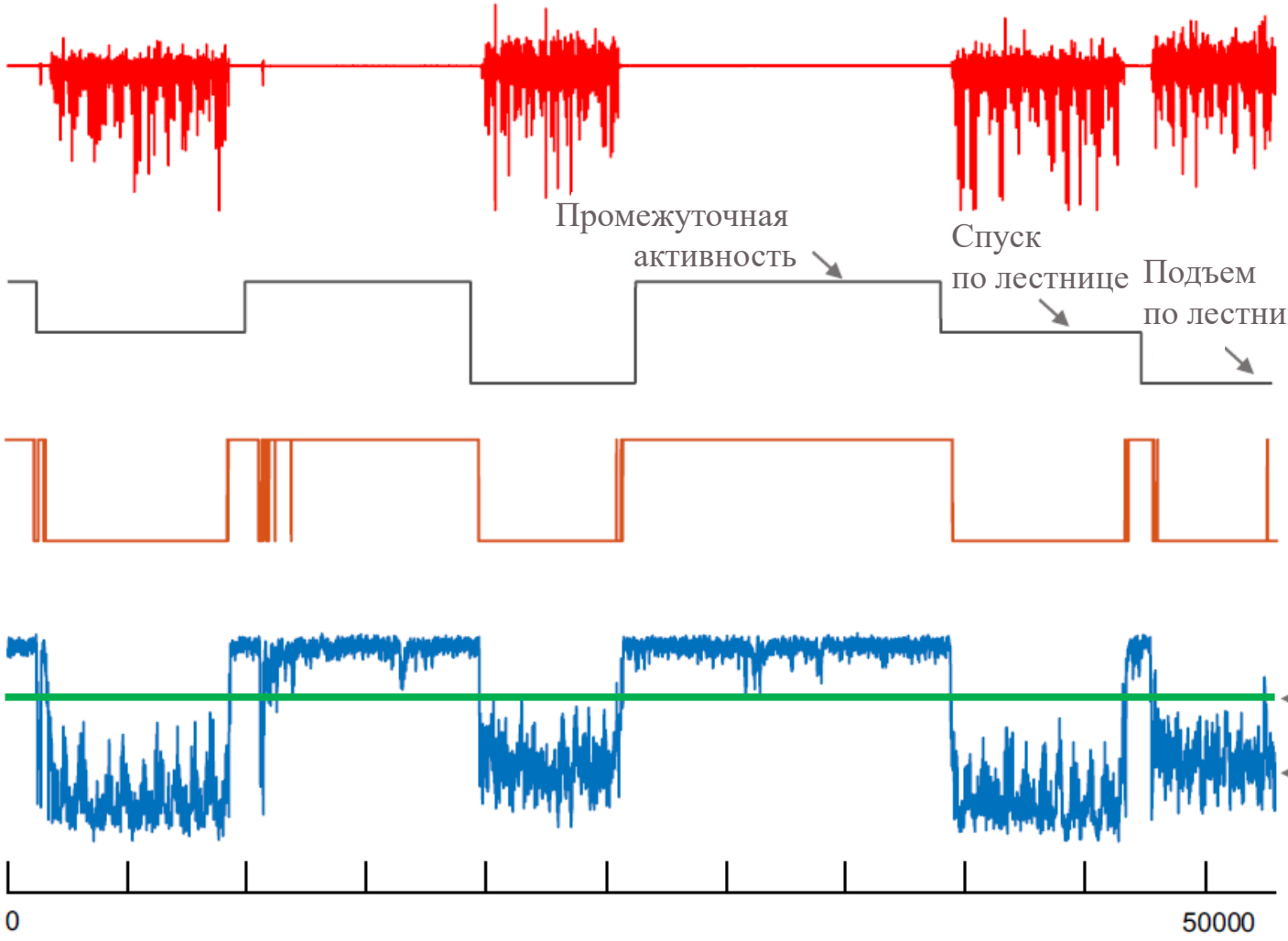
Поиск шейплетов (shapelet)



- *Шейплет* – лучший (наиболее репрезентативный) представитель класса подпоследовательностей (рядов)
- A и B – классы подпоследовательностей, T_A и T_B – ряды, полученные склейкой подпоследовательностей из своих классов (NaN разделяет каждую пару)
- Шейплеты – подпоследовательности, дающие топ-k максимумы в матричном профиле разницы $P = P_{BA} - P_{BB}$

$$\{PBB, Ibb\} := MPjoin(B, B, m)$$
$$\{PBA, Iba\} := MPjoin(B, A, m)$$
$$P := PBA - PBB$$
$$TopKshapelets := TopMax(P, k)$$

Сегментация повторяющихся активностей



Данные носимого акселерометра*

Истинная разметка

Сегментация на основе матричного профиля

Порог

Матричный профиль

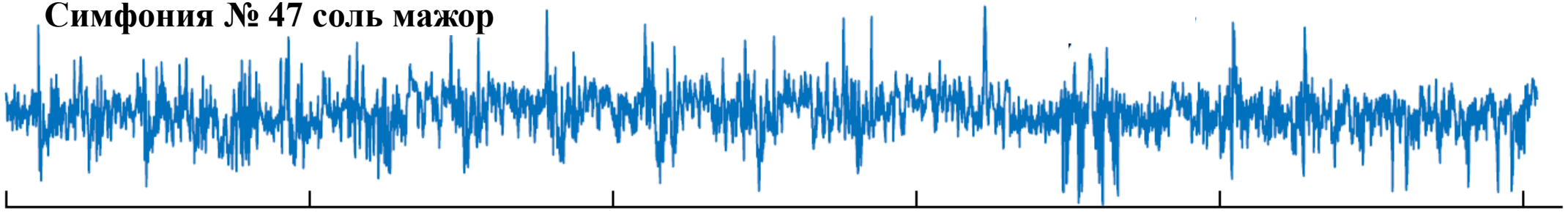
$$\{P, I\} := MP(A, m)$$
$$RA := \{i \mid P(i) < \alpha \cdot (\min(P) + \max(P))\}$$

* Reiss A., Stricker D. Introducing a new benchmarked dataset for activity monitoring. Proc. of the 16th Int. Symp. on Wearable Computers (ISWC). 2012. DOI: 10.1109/ISWC.2012.13.

Поиск перевертышей (semordnilap: god↔dog, lived↔devil, …)

Йозеф Гайдн
(Joseph Haydn)
1732-1809

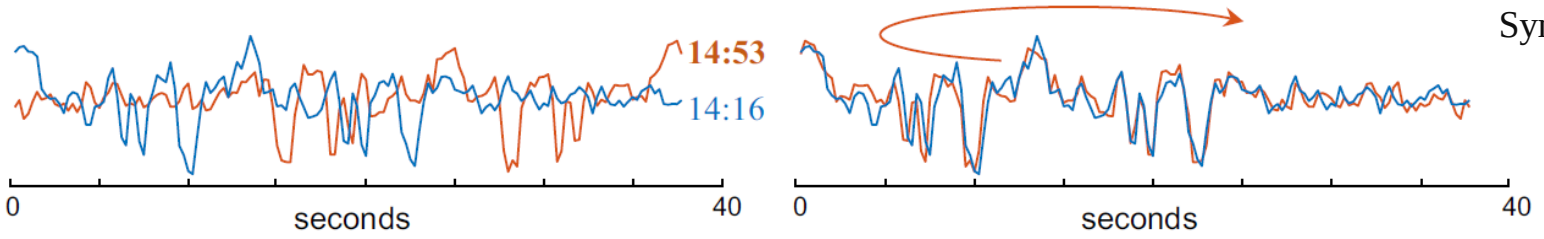
Симфония № 47 соль мажор



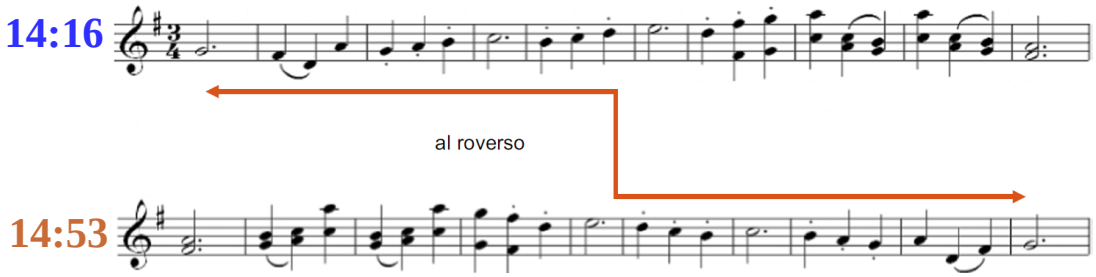
MFCC (Mel-frequency cepstral coefficients):
окно 0.5 с, перекрытие 50%

Joseph Haydn,
Symphony No. 47 in G major "Palindrome",
directed by Bruno Weil

Найденный мотив



Ноты мотива



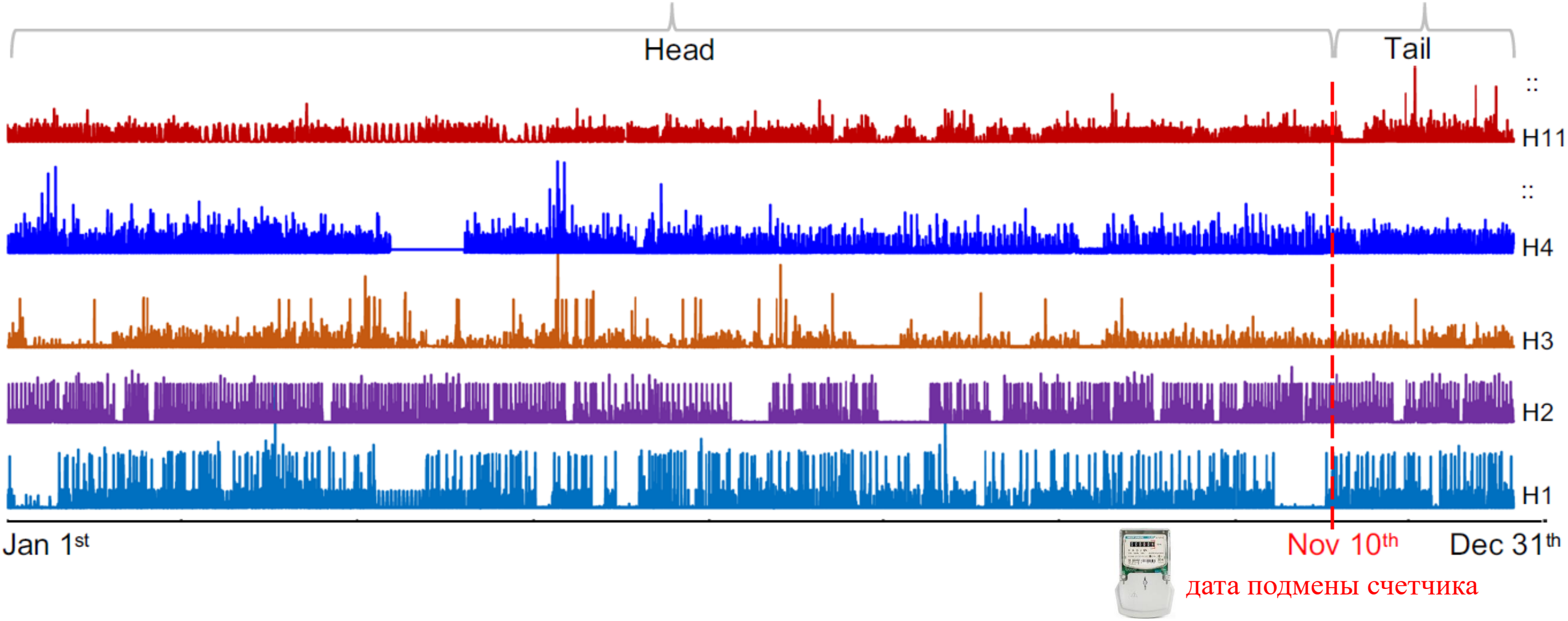
$$cisum := Reverse(music)$$

$$\{P, I\} := MPjoin(music, cisum, 150)$$ // 37.5 с

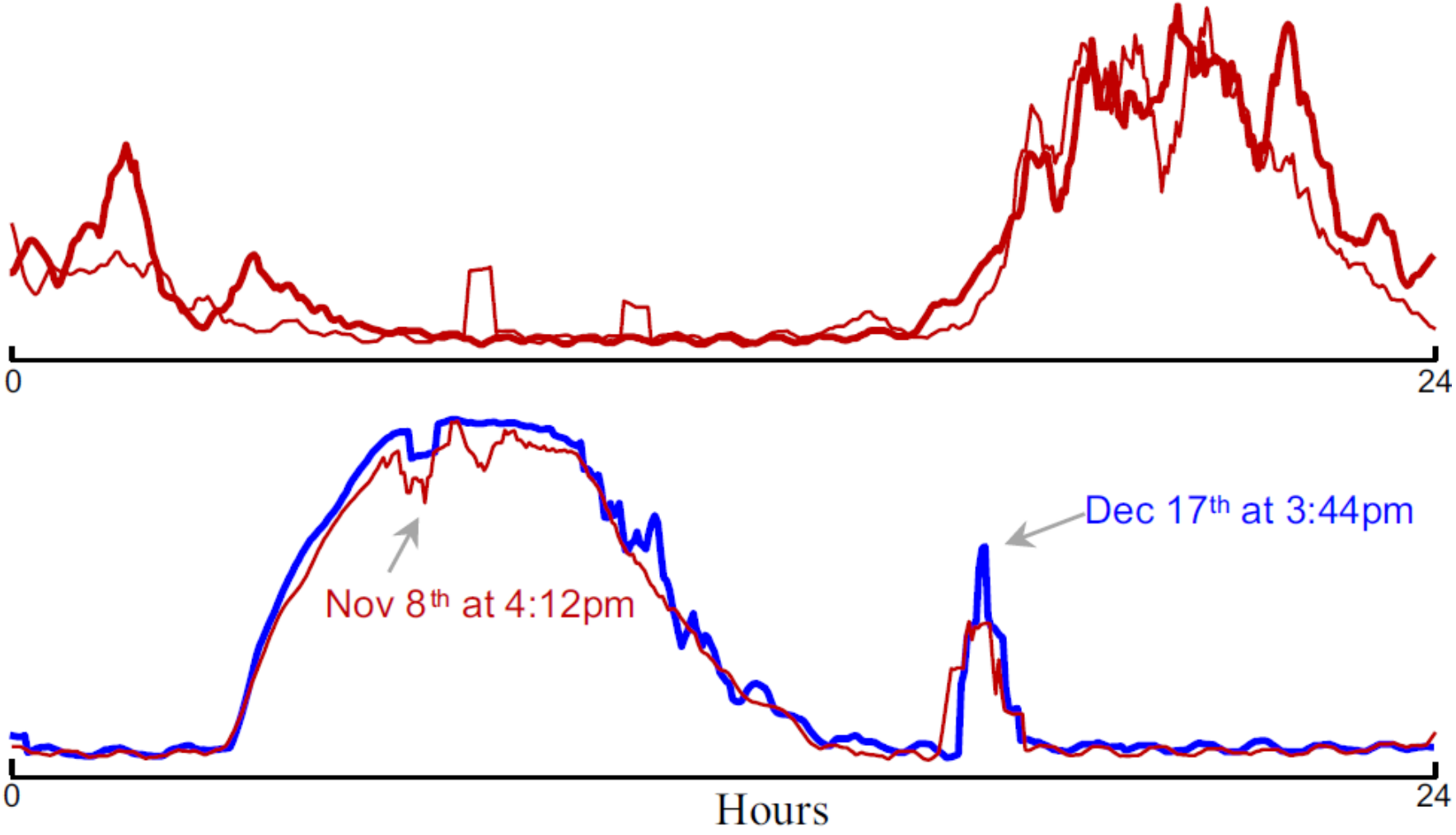
Раскрытие краж электричества подменой счетчиков (meter-swapping)

дата подмены счетчика

Раскрытие краж электричества подменой счетчиков (meter-swapping)



Раскрытие краж электричества подменой счетчиков (meter-swapping)



Top-1 мотив
$MPjoin(Head(H_{11}), Tail(H_{11}))$

$\mathrm{ED}(left, right) = 9.56$
(менее среднего расстояния
по всем домам)

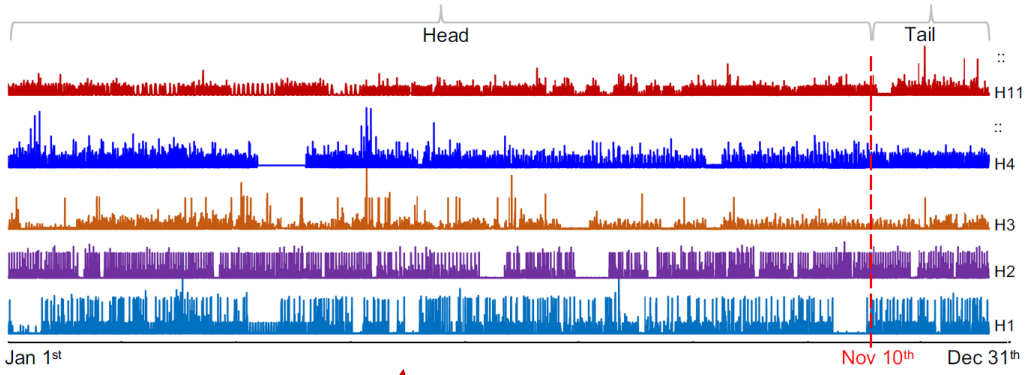
Top-1 мотив
$MPjoin(Head(H_{11}), Tail(H_4))$

$\mathrm{ED}(left, right) = 2.85$

Раскрытие краж электричества подменой счетчиков (meter-swapping)

$$SwapScore(H_i, H_j) = \frac{\min MPjoin(Head(H_i), Tail(H_j))}{\min MPjoin(Head(H_i), Tail(H_i)) + \varepsilon}$$

$minScore := +\infty$
for $i := 1$ **to** $NumHouse$ **do**
  $\{P, I\} := MPjoin(Head(H_i), Tail(H_i), m)$
  $minP := \min(P)$
  for $j := i + 1$ **to** $NumHouse$ **do**
    $\{J, JI\} := MPjoin(Head(H_i), Tail(H_j), m)$
    $SwapScore := \min(J)/(minP + \varepsilon)$
    if $SwapScore < minScore$ **then**
      $minScore := SwapScore$
      $suspect := \{H_i, H_j\}$



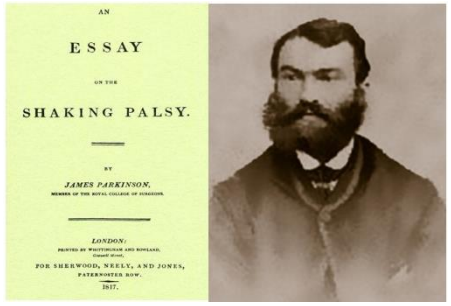
дата подмены счетчика

$suspect = \{H_{11}, H_4\}$

top-1 мотив
$Head(H_{11}), Tail(H_4)$

Кража со стороны H4

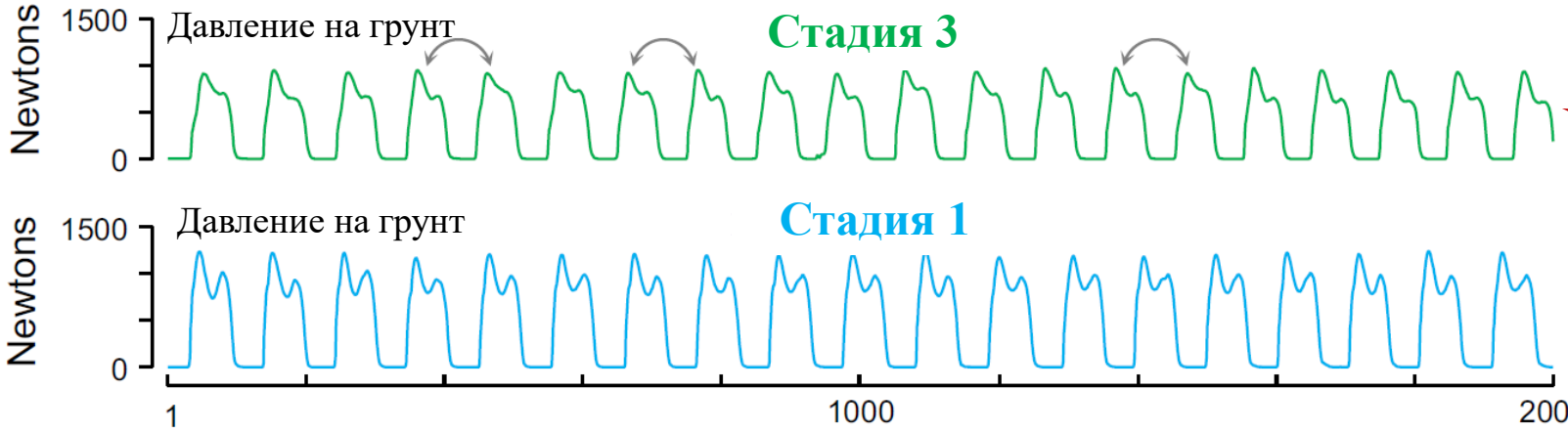
Оценка тяжести болезни Паркинсона



Джеймс Паркинсон
(James Parkinson)
1755-1824

Шкала Хён—Яра
(Hoehn M.M., Yahr M.D.)

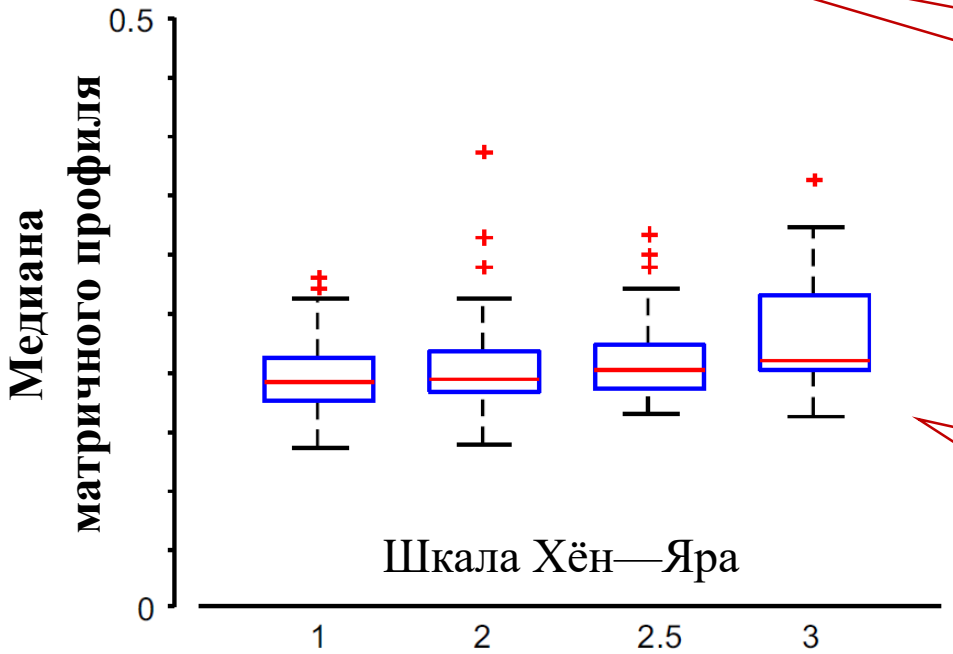
Стадия	Симптоматика
0	Нет признаков заболевания
1	Проявления на одной из конечностей
1.5	Проявления на одной из конечностей и туловище
2	Двусторонние проявления без постуральной неустойчивости
2.5	Двусторонние проявления с постуральной неустойчивостью
3	Двусторонние проявления. Постуральная неустойчивость. Способность к самообслуживанию
4	Обездвиженность, потребность в посторонней помощи. Способность ходить и/или стоять без поддержки
5	Обездвиженность, инвалидизация

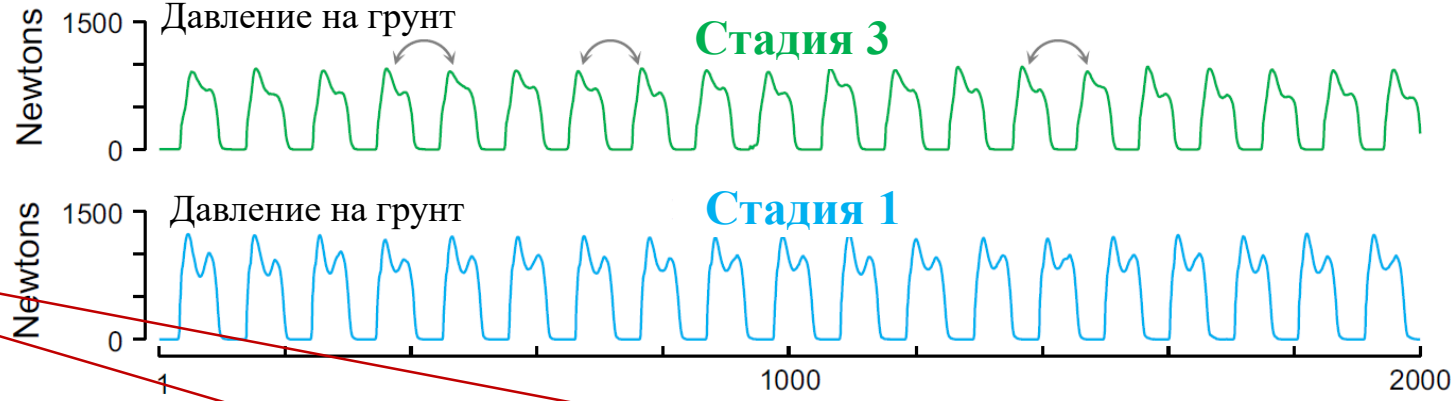


При нарушениях двигательной активности циклы походки повторяются не идеально

Оценка тяжести болезни Паркинсона

for $i \coloneqq 1$ **to** $NumPatients$ **do**
$\{P, I\} \coloneqq MP(PatientGait(i), m)$
$HoehnYahr(i) \coloneqq \text{median}(P)$





Нет отличий между соседями на ранних стадиях болезни,
сильные отличия на поздних стадиях.
Можно взять медиану матричного профиля в качестве индикатора

Эксперимент на наборе PhysioBank*
(93 ряда: 73 – стадия 1, 20 – стадии 2, 2.5, 3)
Медиана матричного профиля увеличивается на поздних стадиях

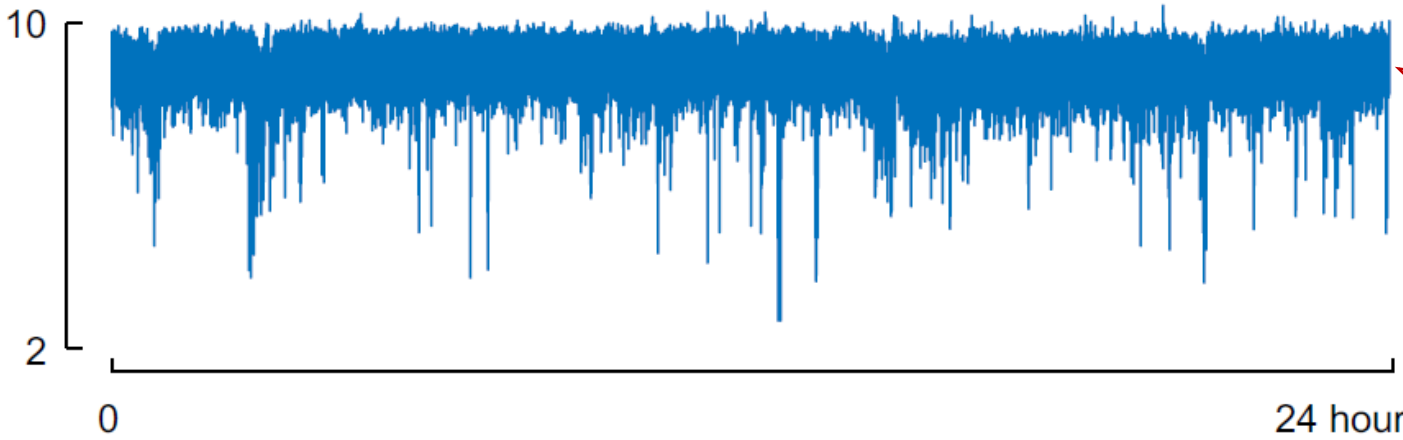
* Goldberger A.L., *et al.* PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation. 2000. 101(23), e215–e220. DOI: 10.1161/01.cir.101.23.e215

Обнаружение низкочастотных землетрясений (LFE, low-frequency earthquake)


Разлом Сан-Андреас, Калифорния (США)

- Станции FROB и JCNB в 10 км друг от друга снимают показания сейсмографа возле разлома Сан-Андреас (частота 20 Гц, 1.728 млн. точек за сутки)

- **Как автоматически фильтровать ложные LFE?**

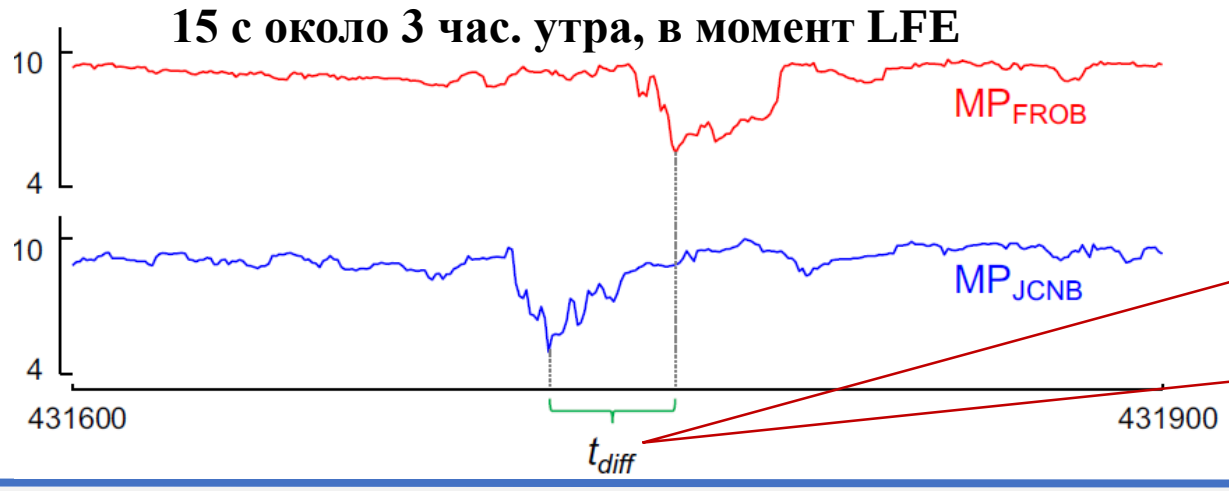


Матричный профиль суточной записи сейсмографа 9 октября 2007 на станции FROB

Лишь 10% «впадин» – истинные землетрясения

Обнаружение низкочастотных землетрясений (LFE, low-frequency earthquake)

- Шумы в сейсмографе локальны, но LFE обнаруживается им в близкие (но не идентичные) моменты времени
- В момент истинного LFE матричные профили разных станций показывают низкие значения. Наоборот, при ложном LFE *один* профиль покажет низкие значения, остальные – высокие
- Для фильтрации возьмем поэлементный максимум матричного профиля (?)

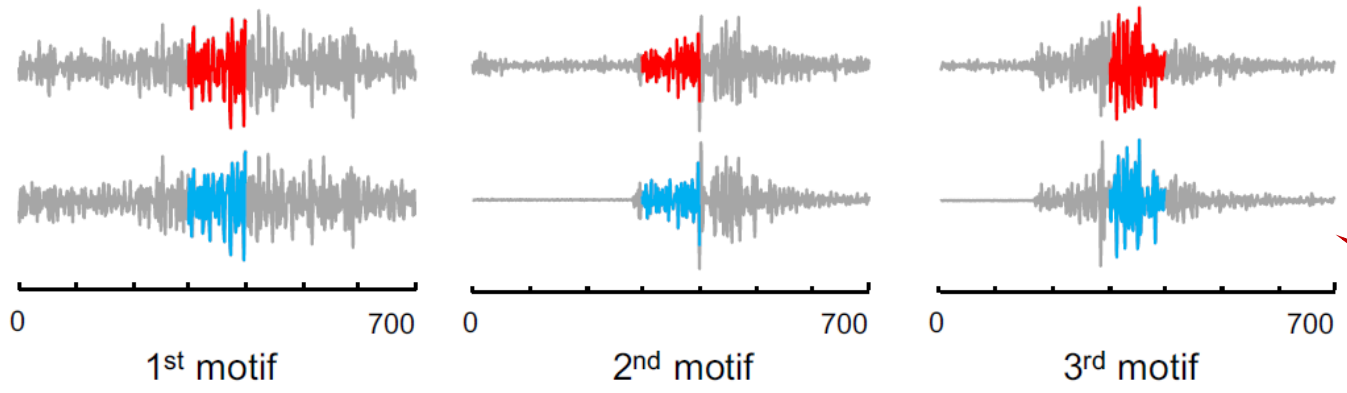


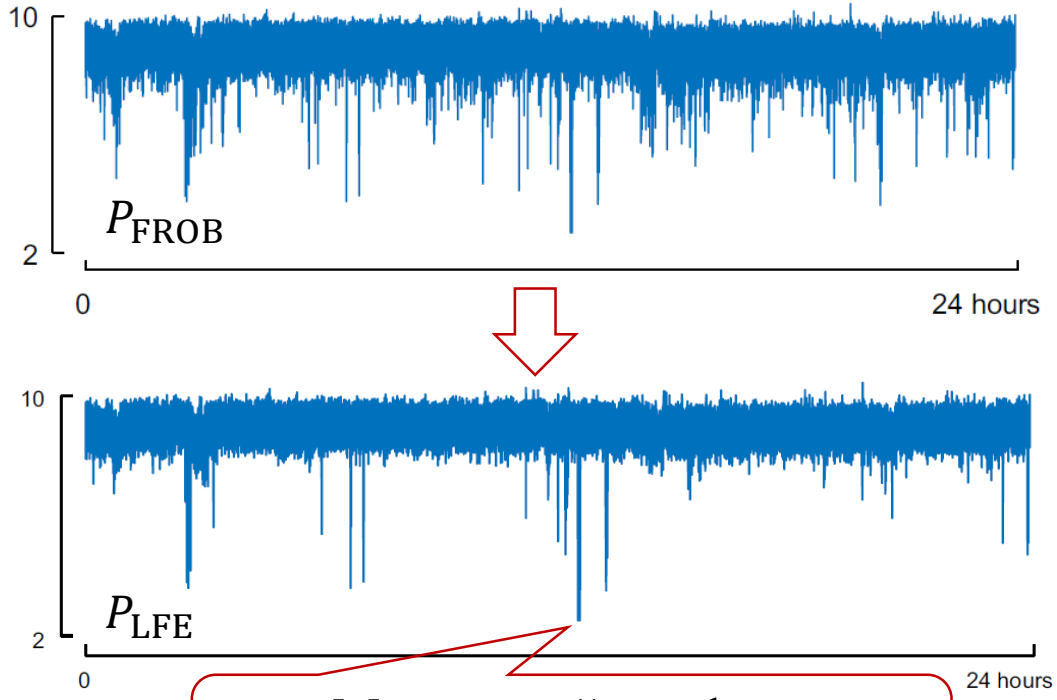
Поэлементный максимум не подходит

Эпицентр землетрясения ближе к JCNB, чем к FROB, поэтому имеется запаздывание t_{diff}. Скорость распространения волны 3-4 км/с, поэтому $t_{diff} \leq 5$ с (100 точек)

Обнаружение низкочастотных землетрясений (LFE, low-frequency earthquake)

$\{P_{\text{FROB}}, I_{\text{FROB}}\} := MP(FROB, m)$
$\{P_{\text{JCNB}}, I_{\text{JCNB}}\} := MP(JCNB, m)$
for $i := 1$ **to** $|FROB|$ **do**
$minVal := \min_{\max(i-100,1) \le k \le \min(i+100,|JCNB|)} P_{\text{JCNB}}(k)$
$minIdx := \arg\min_{\max(i-100,1) \le k \le \min(i+100,|JCNB|)} P_{\text{JCNB}}(k)$
$P_{\text{LFE}}(i) := \max(P_{\text{FROB}}(i), minVal)$
$I_{\text{LFE}}(i) := minIdx$





Матричный профиль с истинными землетрясениями

Top-3 мотива, найденных по очищенному матричному профилю (истинные LFE)

Содержание

- Понятие матричного профиля
- Примеры задач, решаемых на основе матричного профиля
- **Алгоритмы вычисления матричного профиля**

Алгоритмы вычисления матричного профиля

Алгоритм	Длина ряда	Источник
STAMP	10^5	Yeh C.-C.M., Zhu Y., Ulanova L., Begum N., Ding Y., Dau H.A., Silva D.F., Mueen A., Keogh E.J. Matrix Profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. Proc. of the IEEE 16th Int. Conf. on Data Mining, ICDM 2016, Barcelona, Spain, 12–15 December, 2016. pp. 1317–1322. https://doi.org/10.1109/ICDM.2016.0179.
STOMP	10^5	Zhu Y., Zimmerman Z., Senobari N.S., Yeh C.-C.M., Funning G., Mueen A., Brisk P., Keogh E.J. Matrix profile II: Exploiting a novel algorithm and GPUs to break the one hundred million barrier for time series motifs and joins. Proc. of the IEEE 16th Int. Conf. on Data Mining, ICDM 2016, Barcelona, Spain, 12–15 December, 2016. pp. 739–748. https://doi.org/10.1109/ICDM.2016.0085
SCRIMP++	10^6	Zhu Y., Yeh C.-C.M., Zimmerman Z., Kamgar K., Keogh E. Matrix profile XI: SCRIMP++: time series motif discovery at interactive speeds. Proc. of the IEEE 18th Int. Conf. on Data Mining, ICDM 2018, Singapore, November 17-20, 2018. pp. 837–846. https://doi.org/10.1109/ICDM.2018.00099
SCAMP	10^7	Zimmerman Z., Kamgar K., Senobari N.S., Crites B., Funning G.J., Brisk P., Keogh E.J. Matrix Profile XIV: Scaling Time Series Motif Discovery with GPUs to Break a Quintillion Pairwise Comparisons a Day and Beyond. Proc. of the ACM Symposium on Cloud Computing, SoCC 2019, Santa Cruz, CA, USA, November 20-23, 2019. pp. 74–86. https://doi.org/10.1145/3357223.3362721

Литература

1. Yeh C.M., Zhu Y., Ulanova L., Begum N., Ding Y., Dau H.A., Silva D.F., Mueen A., Keogh E.J. Matrix Profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. Proc. of the IEEE 16th Int. Conf. on Data Mining, ICDM 2016, Barcelona, Spain, 12–15 December 2016. pp. 1317–1322. https://doi.org/10.1109/ICDM.2016.0179.
2. Zhu Y., Gharghabi S., Silva D.F., Dau H.A., Yeh C.-C.M., Senobari N.S., Almaslukh A., Kamgar K., Zimmerman Z., Funning G., Mueen A., Keogh E. The Swiss army knife of time series data mining: Ten useful things you can do with the matrix profile and ten lines of code. Data Min. Knowl. Discov. 34(4): 949-979 (2020). https://doi.org/10.1007/s10618-019-00668-6.