

Поиск диссонансов временного ряда



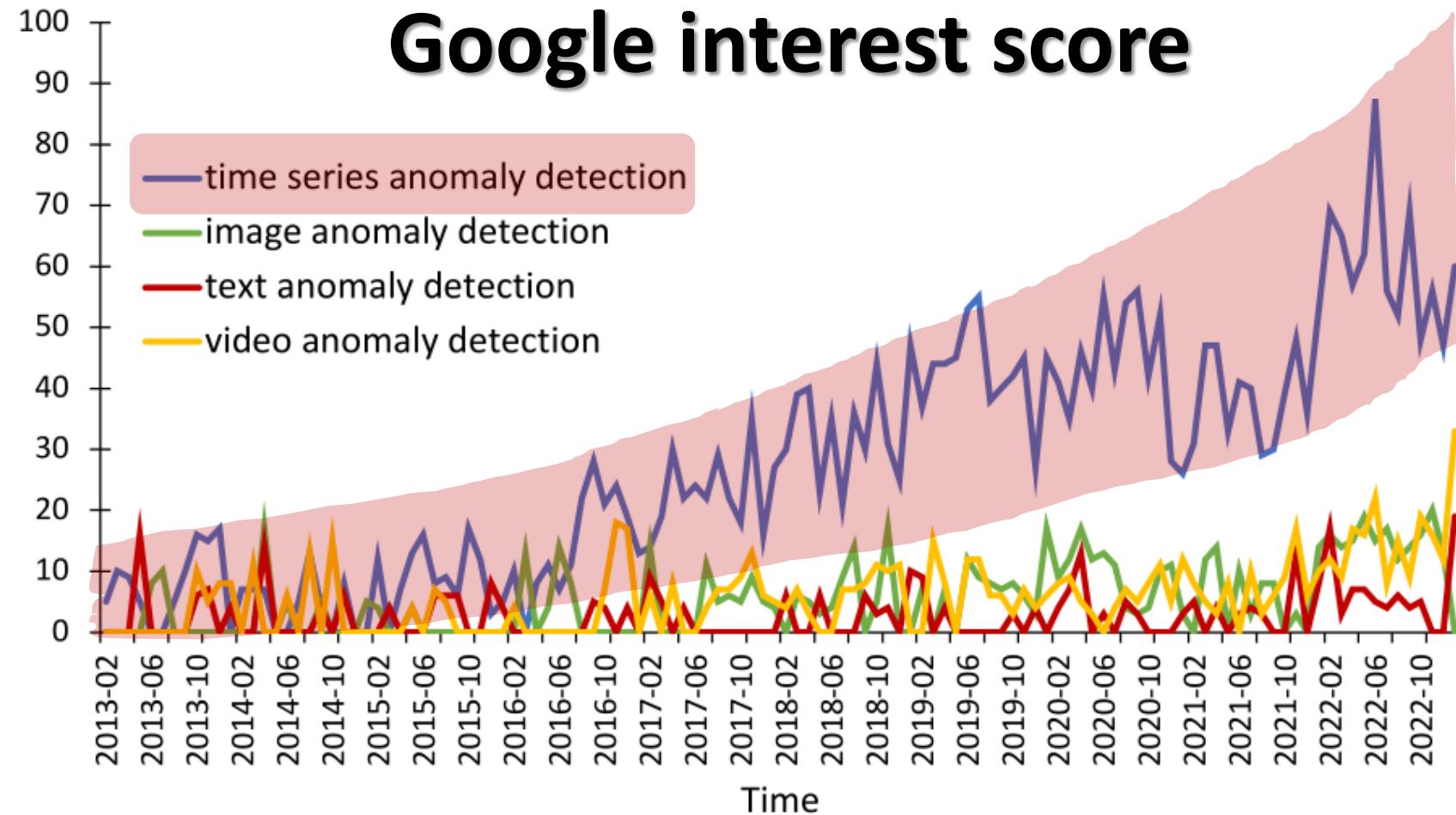
*Мир всегда приходит в норму.
Важно лишь, чья она.*

C.E. Лец

Содержание

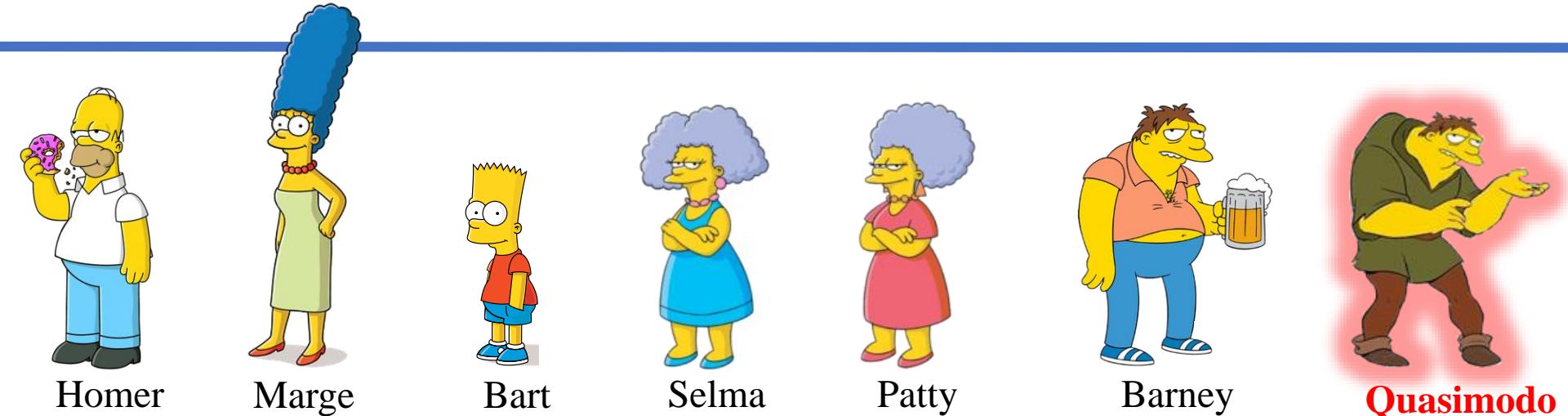
- Понятия аномалии и диссонанса
- Алгоритм HOTSAX
- Алгоритм DRAG
- Алгоритм MERLIN
- Распараллеливание поиска диссонансов

Аномалии временных рядов вызывают аномально высокий интерес исследователей*



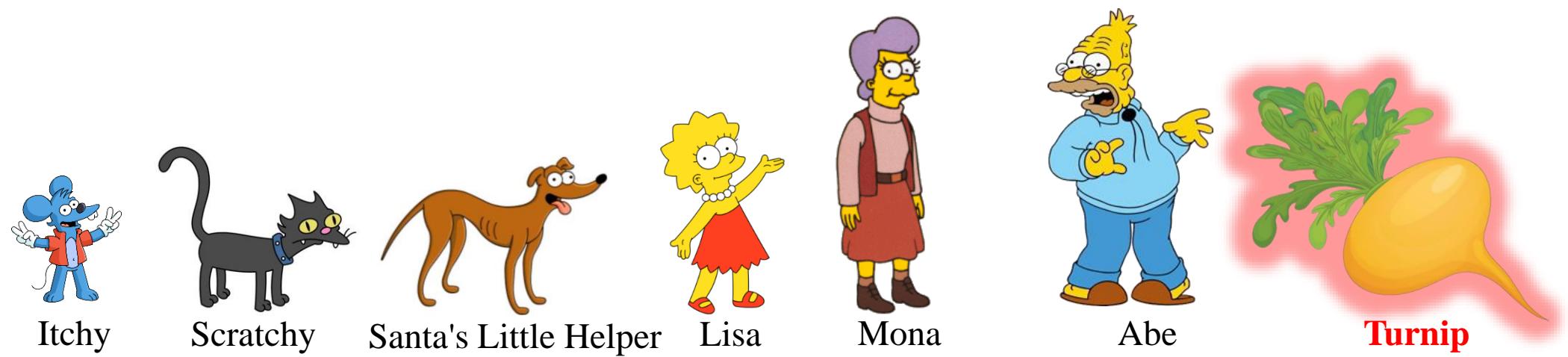
* Boniol P. et al. New trends in time-series anomaly detection. EDBT'2023. pp. 847-850. DOI: [10.48786/edbt.2023.80](https://doi.org/10.48786/edbt.2023.80)

Аномалия – неформальное понятие, зависящее от предметной области

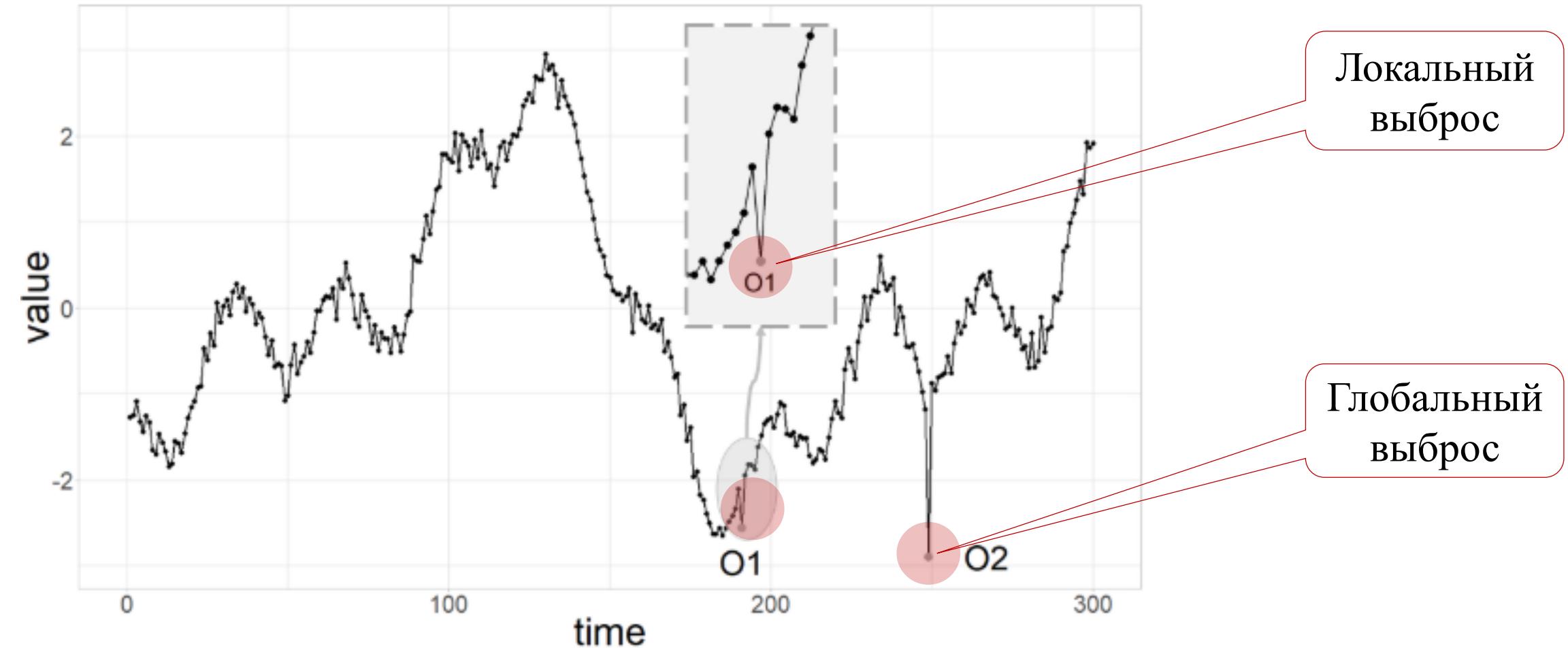


Аномалия – наблюдение, которое настолько сильно отличается от других наблюдений, что вызывает подозрения в том, что оно было создано иным механизмом.

Hawkins D.M. Identification of outliers. Monographs on applied probability and statistics. Springer, 1980. DOI: [10.1007/978-94-015-3994-4](https://doi.org/10.1007/978-94-015-3994-4).

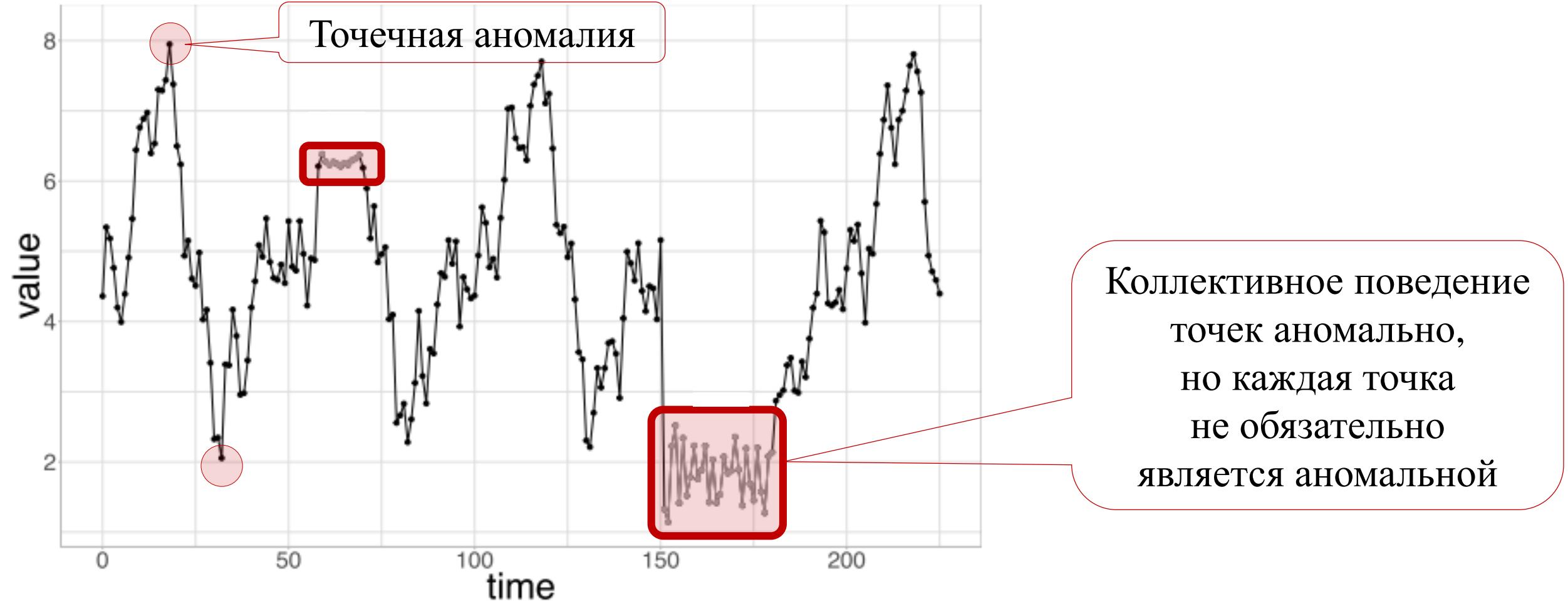


Точечные аномалии: локальные и глобальные



Blazquez-Garca A., et al. A review on outlier/anomaly detection in time series data. ACM Comput. Surv. 54(3), 56:1-56:33 (2021). <https://doi.org/10.1145/3444690>

Аномальная подпоследовательность



Blazquez-Garca A., et al. A review on outlier/anomaly detection in time series data. ACM Comput. Surv. 54(3), 56:1-56:33 (2021). <https://doi.org/10.1145/3444690>

Формализация аномальной подпоследовательности ряда

If the removal of a point P from the time sequence results in a sequence that can be represented more succinctly than the original one (by more than the increment required for explicitly keeping track of P separately), then the point P is a **deviant**.

Jagadish H. *et al.* Mining deviants in a time series database. VLDB 1999. pp. 102-113. [URL](#)



Формализация аномальной подпоследовательности ряда

- **Outliers** are the data points for that there are fewer than p other data points within distance d .

Knorr E., Ng N. Finding intensional knowledge of distance-based outliers. VLDB 1999. pp. 211-222.

[URL](#)

- **Outliers** are the top n data points whose distance to their k -th nearest neighbor is greatest.

Ramaswamy S. et al. Efficient algorithms for mining outliers from large dataset. SIGMOD 2000. pp. 427-438. DOI: [10.1145/342009.335437](https://doi.org/10.1145/342009.335437)

- **Outliers** are the top n data points whose average distance to their k nearest neighbors is greatest.

Angiulli F., Pizzuti C. Fast outlier detection in high dimensional spaces. PKDD 2002. pp. 15-26. DOI: [10.1007/3-540-45681-3_2](https://doi.org/10.1007/3-540-45681-3_2)



Формализация аномальной подпоследовательности ряда

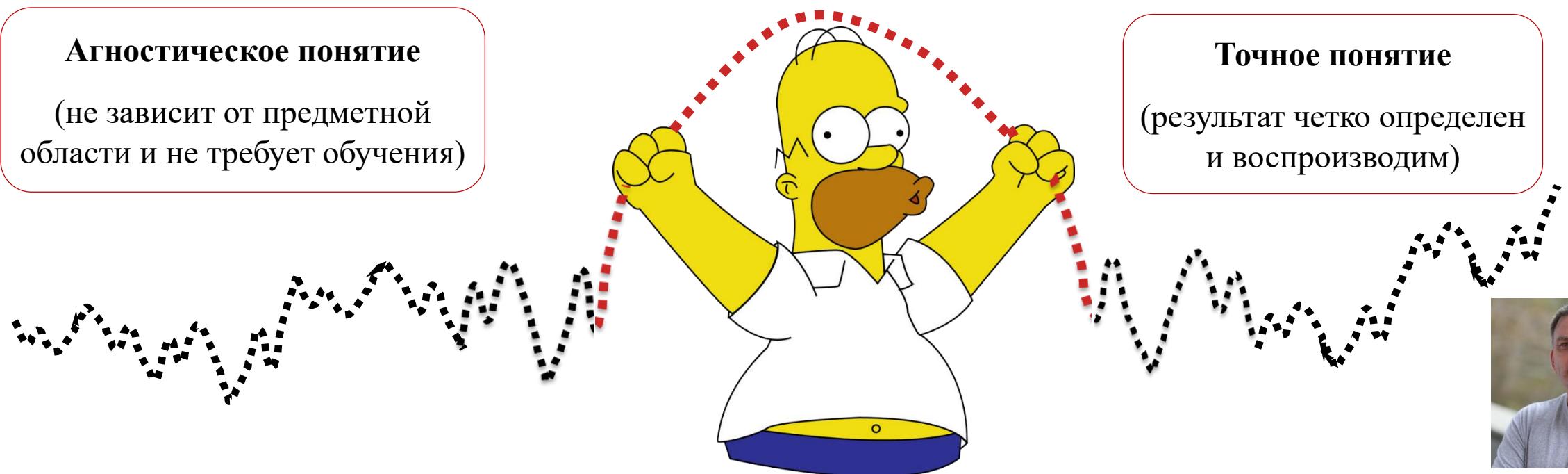
Discord* is a subsequence of the given length whose distance to its nearest neighbor is greatest

Агностическое понятие

(не зависит от предметной области и не требует обучения)

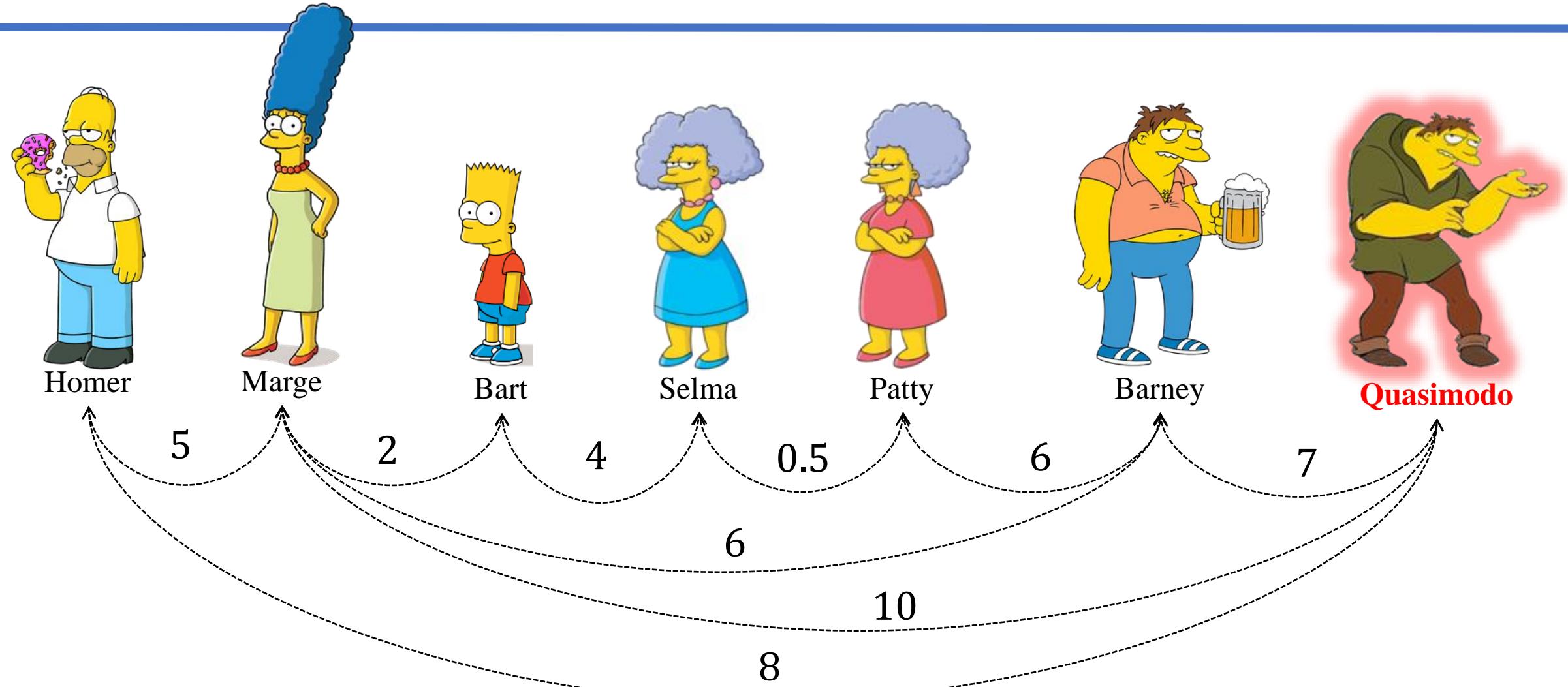
Точное понятие

(результат четко определен и воспроизводим)

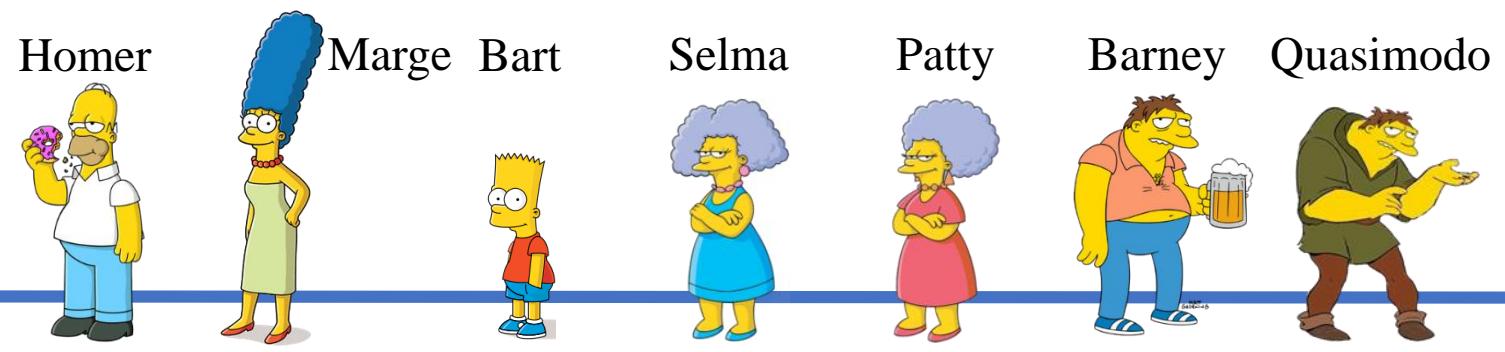


* Keogh E. et al. HOT SAX: Efficiently finding the most unusual time series subsequence. ICDM 2005. pp. 226-233. DOI: [10.1109/ICDM.2005.79](https://doi.org/10.1109/ICDM.2005.79)

Диссонанс: Расстояние отражает схожесть



Диссонанс

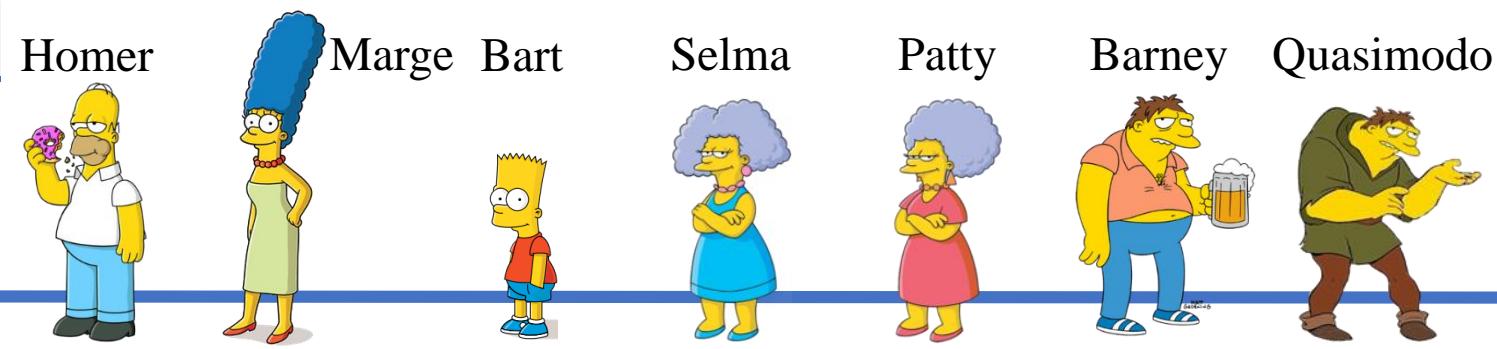


Матрица расстояний:
чем ближе соседи,
тем более они схожи



Homer	Marge	Bart	Selma	Patty	Barney	Quasimodo
0						
	0					
		0				
			0			
				0		
					0	
						0

Диссонанс



Матрица расстояний
с вычисленными
расстояниями



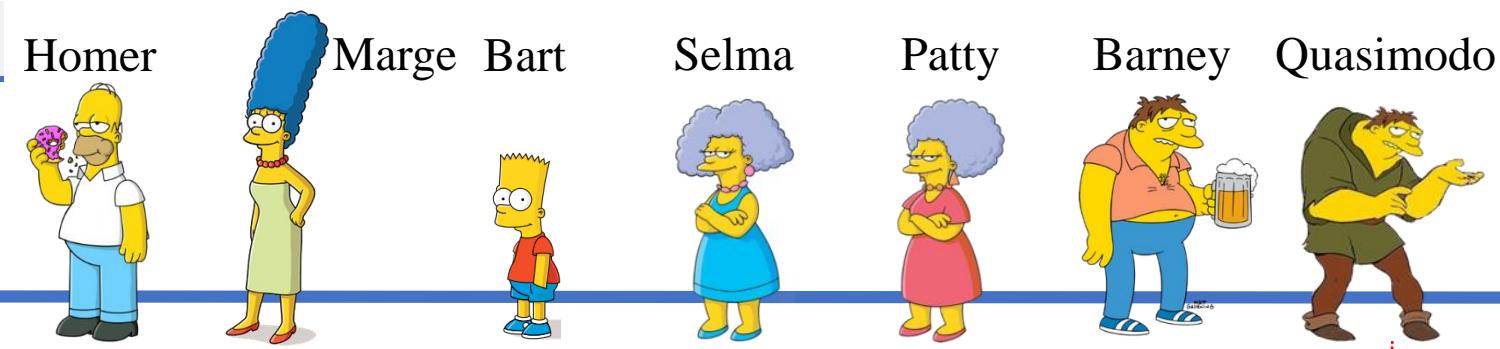
	Homer	Marge	Bart	Selma	Patty	Barney	Quasimodo
0	5	2	4	4	6	8	
5	0	2.5	3	3	6	10	
2	2.5	0	4	4	6	9	
4	3	4	0	0.5	5	8	
4	3	4	0.5	0	5	8	
6	6	6	5	5	0	7	
8	10	9	8	8	7	0	

Диссонансы

Матрица расстояний
с **расстояниями**
до ближайших соседей
(т.е. минимумы
по столбцам)

Homer	Marge	Bart	Selma	Patty	Barney	Quasimodo
0	5	2	4	4	6	8
5	0	2.5	3	3	6	10
2	2.5	0	4	4	6	9
4	3	4	0	0.5	5	8
4	3	4	0.5	0	5	8
6	6	6	5	5	0	7
8	10	9	8	8	7	0

Диссонанс

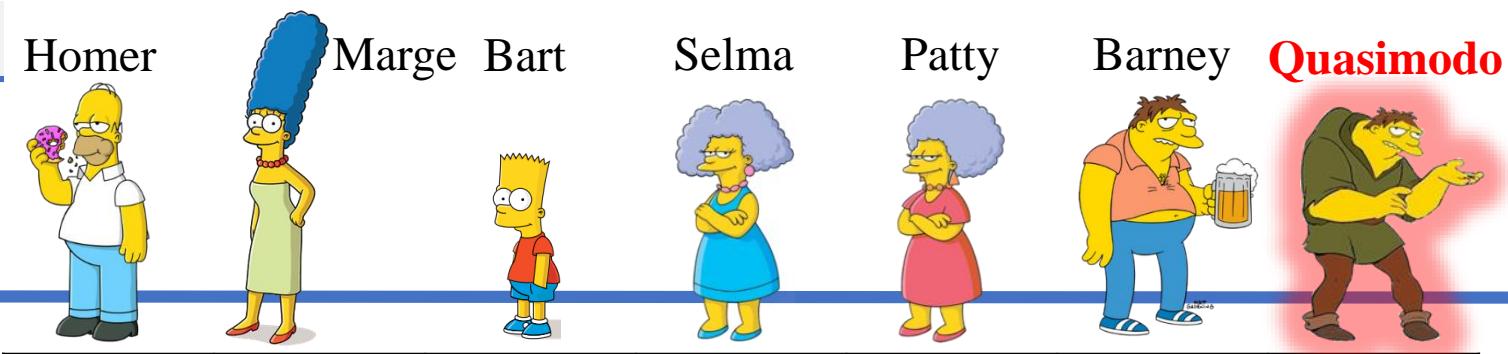


Матрица расстояний
с **наибольшим**
расстоянием
до ближайшего соседа
(т.е. максимум
минимумов по столбцам)



	Homer	Marge	Bart	Selma	Patty	Barney	Quasimodo
0	5	2	4	4	6	8	
5	0	2.5	3	3	6	10	
2	2.5	0	4	4	6	9	
4	3	4	0	0.5	5	8	
4	3	4	0.5	0	5	8	
6	6	6	5	5	0	7	
8	10	9	8	8	7	0	

Диссонанс



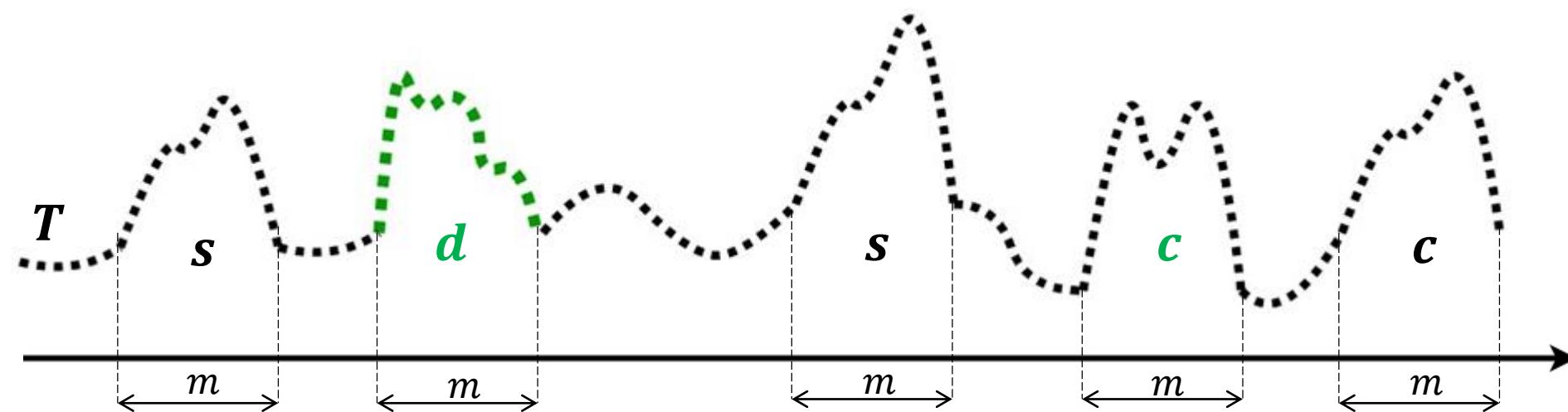
Диссонанс объект
с самым далеким
ближайшим соседом
(т.е. аргумент
максимума минимумов
по столбцам)



	0	5	2	4	4	6	8
5	0	2.5	3	3	6	10	
2	2.5	0	4	4	6	9	
4	3	4	0	0.5	5	8	
4	3	4	0.5	0	5	8	
6	6	6	5	5	0	7	
8	10	9	8	8	7	0	

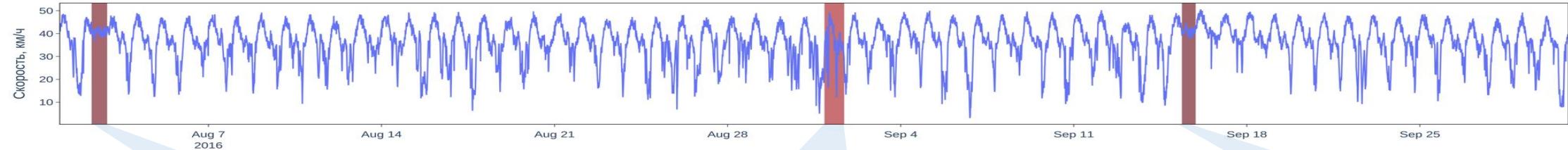
Диссонанс (discord)

- *Диссонанс* – подпоследовательность ряда, расстояние от которой до ее ближайшего соседа максимально
- Дано: ряд T , длина диссонанса m , функция расстояния $\text{Dist}(\cdot, \cdot)$
- Найти: $d = \arg \max_{s \in S_T^m} \min_{\{c \in S_T^m \mid s \cap c = \emptyset\}} \text{Dist}(c, s)$

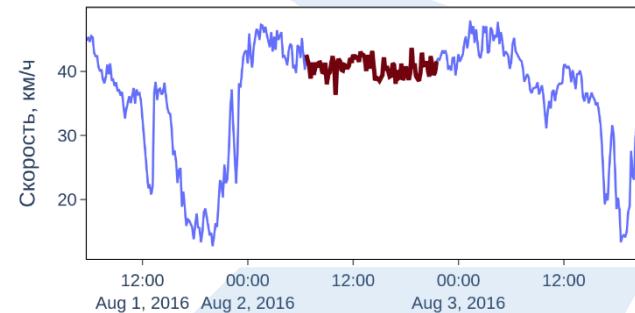


Диссонанс отражает аномалию в реальной жизни, ...

Средняя скорость городского трафика в Гуанчжоу, Китай*



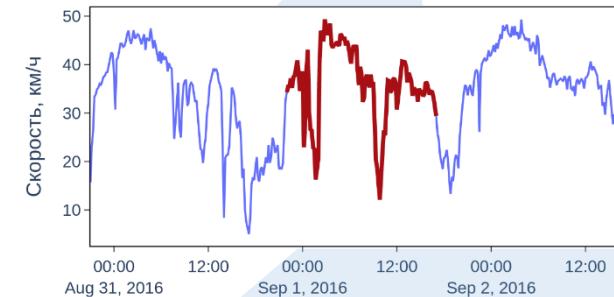
Top-2 диссонанс



Тайфун Ниди



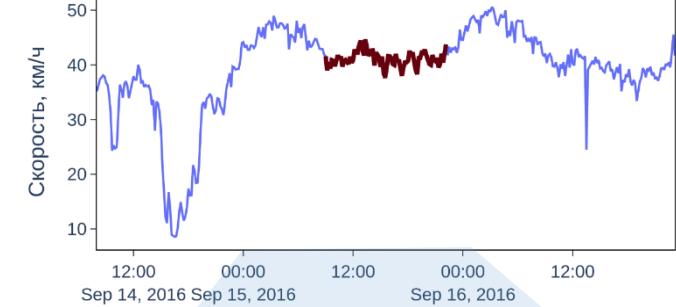
Top-3 диссонанс



День Победы над Японией



Top-1 диссонанс

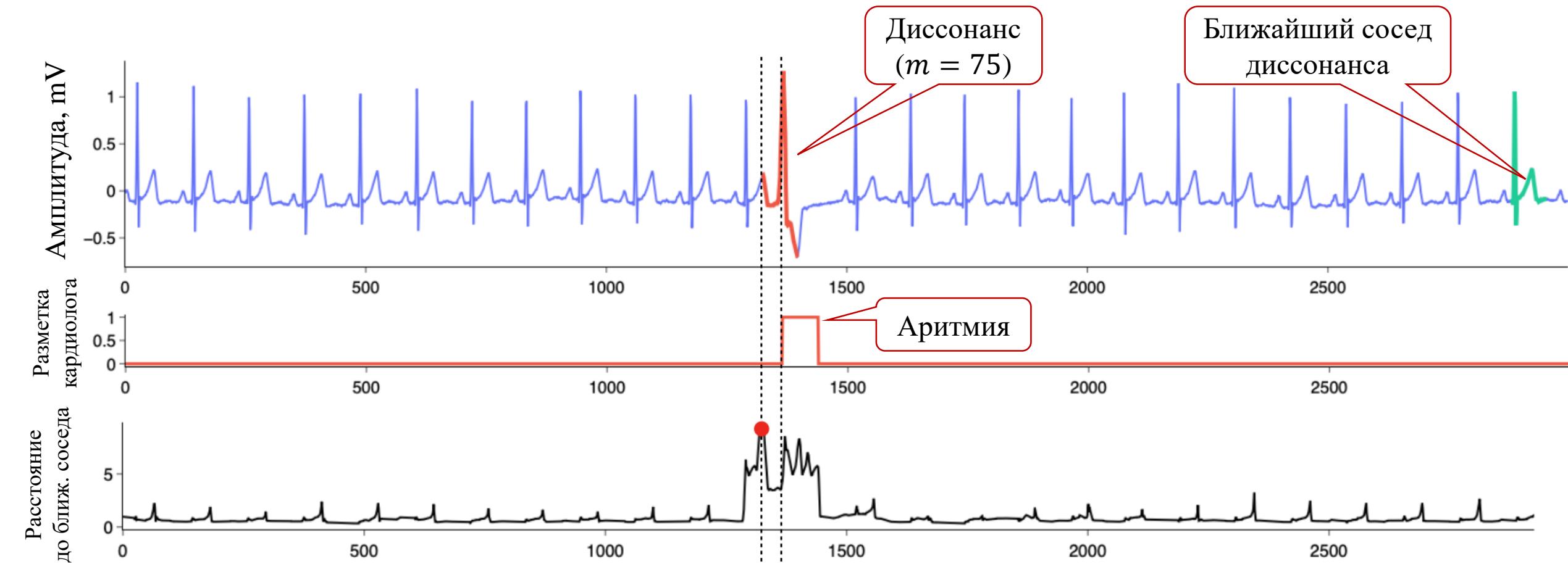


Фестиваль Луны

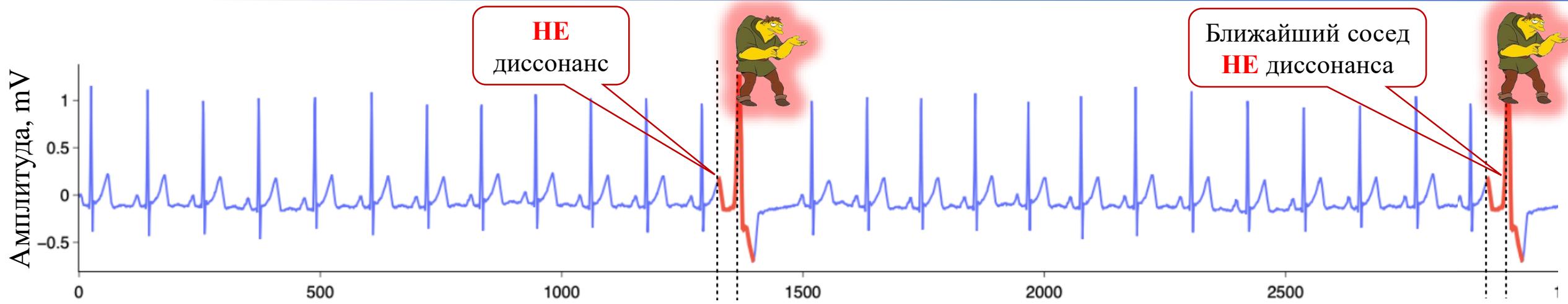


* Chen X, Chen Y, He Z. Urban traffic speed dataset of Guangzhou, China. 2018. DOI: [10.5281/zenodo.1205229](https://doi.org/10.5281/zenodo.1205229).

..., но диссонанс не идентичен аномалии



Проблема уродливых близнецов (twin freaks)



- Все счастливые семьи похожи друг на друга, каждая несчастливая семья несчастлива по-своему¹⁾
- **Повторяющиеся аномалии, связанные с изменением формы, редки²⁾**
- Для экспериментов с методом поиска аномалий, решающим проблему уродливых близнецов, использованы синтетические данные, полученные копированием-вставкой аномалии³⁾

¹⁾ Толстой Л.Н. Анна Каренина.

²⁾ Nakamura T. et al. MERLIN++: parameter-free discovery of time series anomalies. Data Min. Knowl. Disc. 2023. 37(2). pp. 670-709. DOI: [10.1007/s10618-022-00876-7](https://doi.org/10.1007/s10618-022-00876-7)

³⁾ Bu Y. et al. Efficient anomaly monitoring over moving object trajectory streams. ACM SIGKDD KDD 2009. pp. 159-168. DOI: [10.1145/1557019.1557043](https://doi.org/10.1145/1557019.1557043)

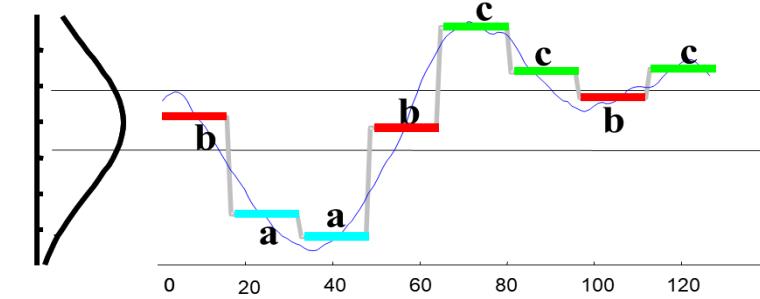
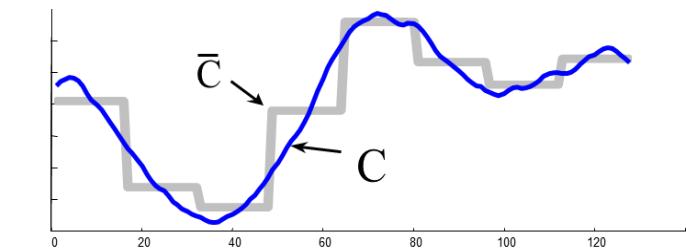
Содержание

- Понятия аномалии и диссонанса
- **Алгоритм HOT SAX**
- Алгоритм DRAG
- Алгоритм MERLIN
- Распараллеливание поиска диссонансов

Алгоритм HOT SAX

(Heuristically Ordered Time series using Symbolic Aggregate ApproXimation)

- Особенности
 - Ряд может быть размещен в оперативной памяти
 - Ответ не является точным
- Ключевые идеи
 - Сжатие подпоследовательностей исходного ряда
 - Кодирование сжатых подпоследовательностей ряда
 - Перебор кодированных подпоследовательностей ряда с отбрасыванием

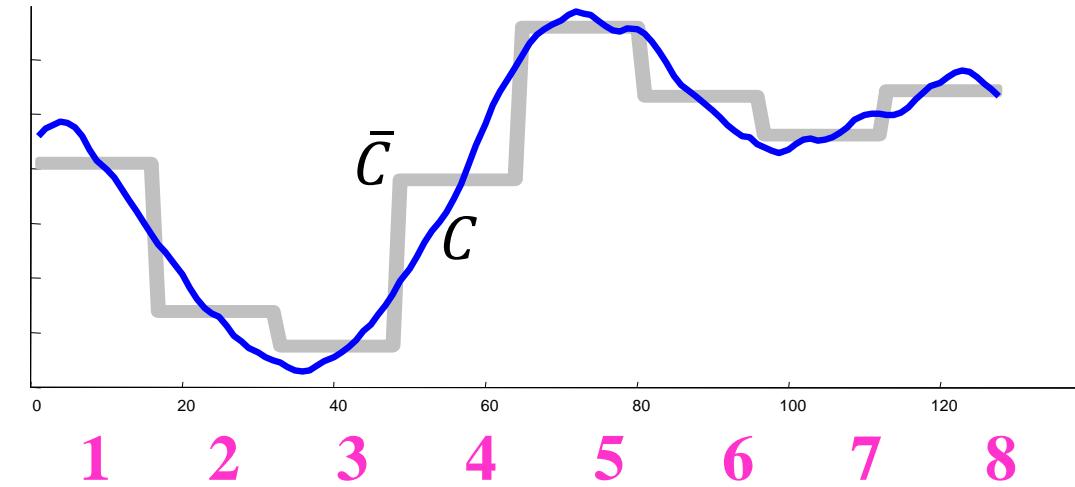


РАА (Piecewise Aggregate Approximation)

- Сжатие (аппроксимация) подпоследовательностей ряда с уменьшением их длины до $w \ll m$

$$\bar{c}_i = \frac{w}{m} \sum_{j=\frac{m}{w}(i-1)+1}^m c_j$$

Разбить временной ряд длины m на w промежутков равной длины, заменить значения в каждом из них на среднее арифметическое элементов ряда в промежутке

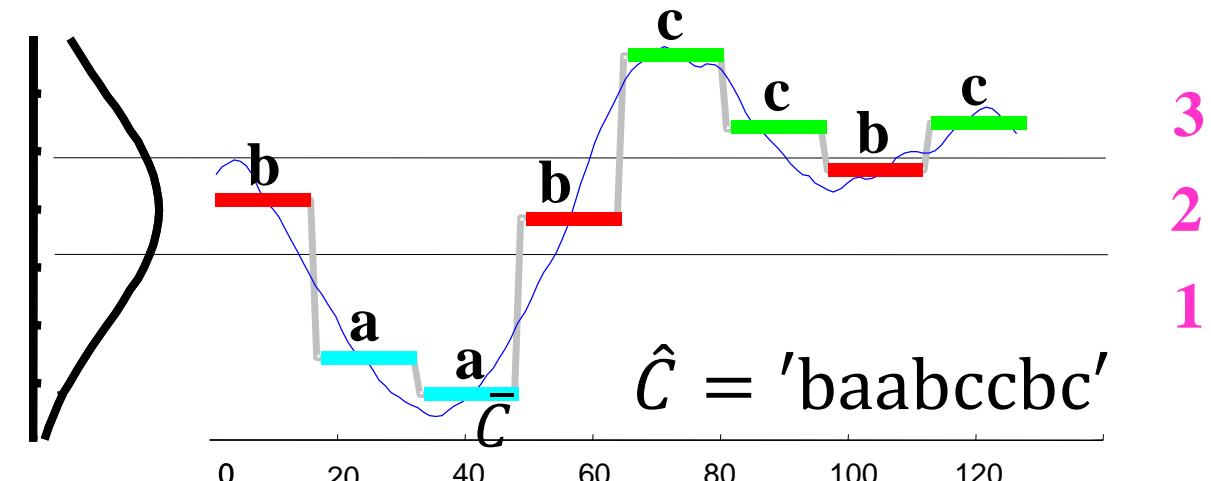


Lin J., Keogh E.J., Lonardi S., Chiu B.Y. A symbolic representation of time series, with implications for streaming algorithms. Proc. of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, DMKD 2003, San Diego, California, USA, June 13, 2003. 2003. P. 2–11. URL: <https://doi.org/10.1145/882082.882086>

SAX (SAX, Symbolic Aggregate AppRoXimation)

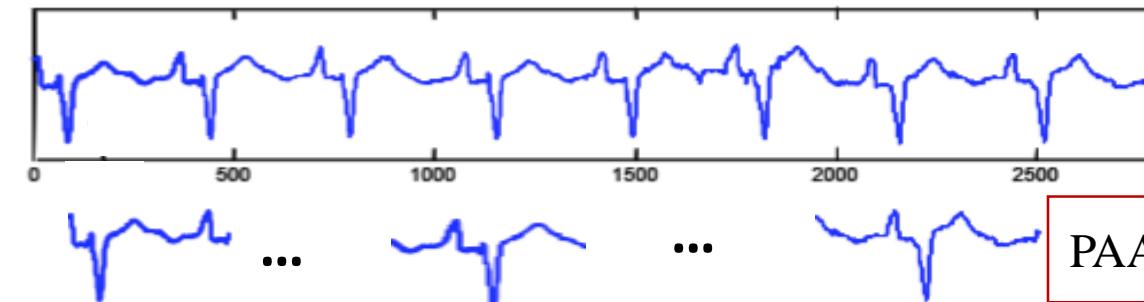
- Кодирование сжатых подпоследовательностей в слова алфавита $\mathcal{A} = (\alpha_1, \dots, \alpha_{|\mathcal{A}|})$, $\alpha_1 = 'a'$, $\alpha_2 = 'b'$ и т.д. ($|\mathcal{A}| > 2$)
- Таблица кодирования:
 - $\hat{c}_i = \alpha_i \Leftrightarrow \beta_{j-1} \leq c_i < \beta_j$
 - $\beta_0 = -\infty, \beta_{|\mathcal{A}|} = +\infty$
 - Площадь под кривой $N(0,1)$ между β_{j-1} и β_j равна $\frac{1}{|\mathcal{A}|}$

$(-\infty; -0.67)$	$[-0.67; 0]$	$[0; 0.67)$	$[0.67; \infty)$
a	b	c	d



$\beta_i \setminus a$	3	4	5	6	7	8
β_1	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15
β_2	0.43	0	-0.25	-0.43	-0.57	-0.67
β_3		0.67	0.25	0	-0.18	-0.32
β_4			0.84	0.43	0.18	0
β_5				0.97	0.57	0.32
β_6					1.07	0.67
β_7						1.15

Построение префиксного дерева частот слов



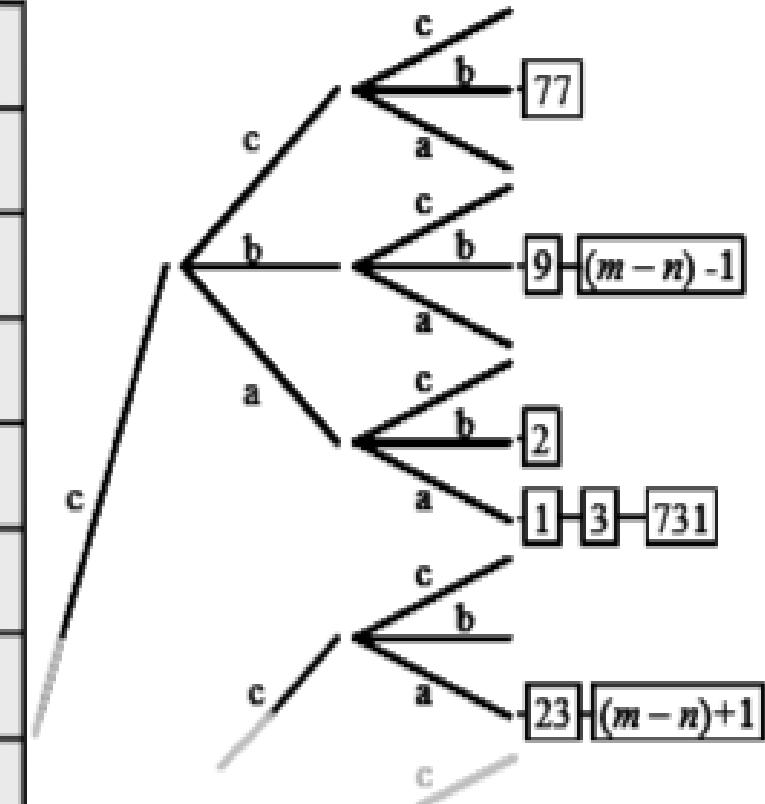
PAA, SAX

Частотный индекс слов

1	c	a	a	3
2	c	a	b	1
3	c	a	a	3
...
$(m-n)-1$	c	b	b	2
$(m-n)$	a	c	b	1
$(m-n)+1$	b	c	a	2

- Каждое ребро дерева помечено символом алфавита. Ребра, соединяющие узел с его сыновьями, помечены разными символами
- SAX-код – конкатенация пометок ребер на пути от корня до листа
- В листе – упорядоченный по возрастанию список индексов подпоследовательностей исходного ряда с соответствующим кодом

Префиксное дерево



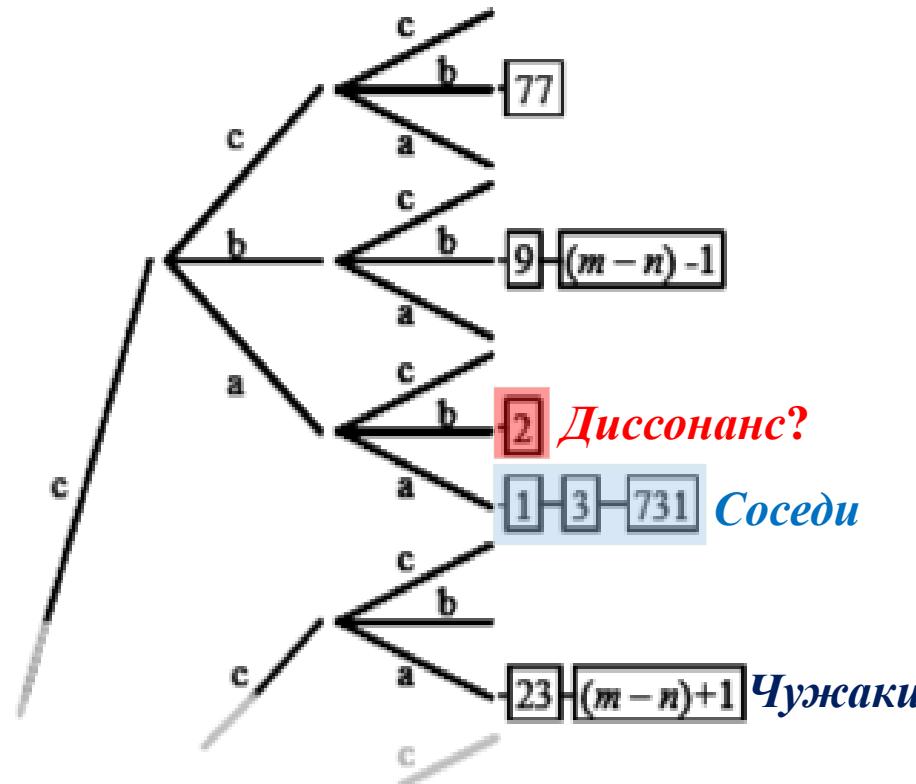
Перебор слов

Частотный индекс слов

	1	c	a	a	3
	2	c	a	b	1
	3	c	a	a	3

$(m - n) - 1$		c	b	b	2
$(m - n)$		a	c	b	1
$(m - n) + 1$		b	c	a	2

Префиксное дерево



Алгоритм HOTSAK

```

 $dist_{bsf} \leftarrow 0; dist_{min} \leftarrow \infty$ 
for  $C_i \in \text{Диссонансы?}$  • Остальные
    for  $C_j \in \text{Соседи}$  • Чужаки
         $d \leftarrow \text{Dist}(C_i, C_j)$ 
        if  $d < dist_{bsf}$ 
            break
         $dist_{min} \leftarrow \min(d, dist_{min})$ 
         $dist_{bsf} \leftarrow \max(dist_{min}, dist_{bsf})$ 
         $pos_{bsf} \leftarrow i$ 
return  $\{pos_{bsf}, dist_{bsf}\}$ 

```

Проблемы HOTSAX

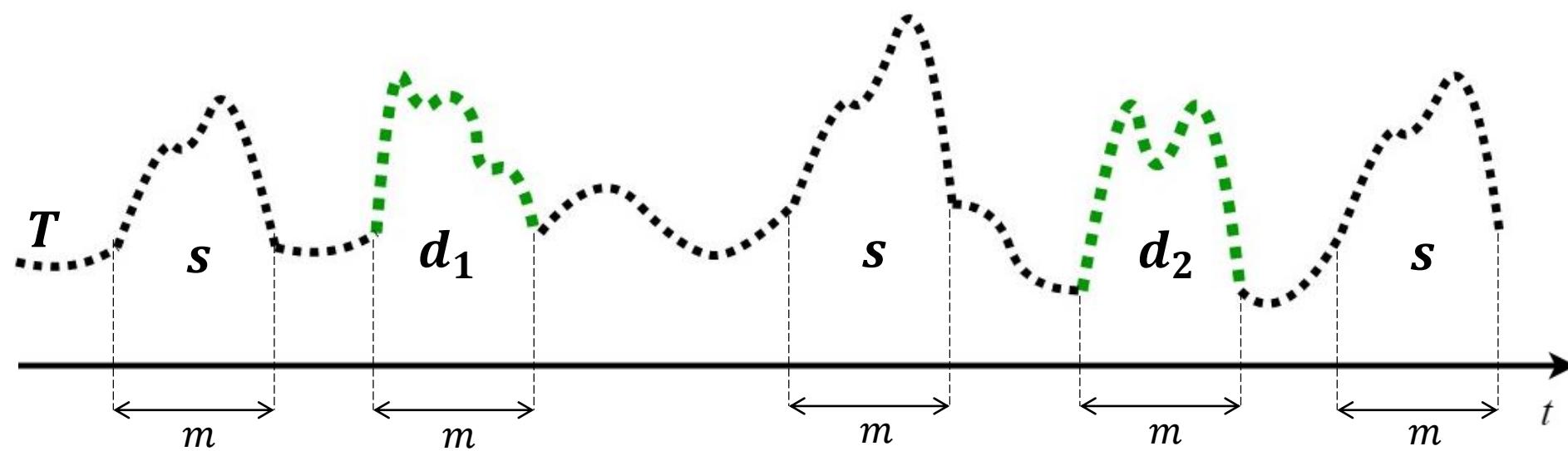
- Не подходит для рядов, которые не могут быть целиком размещены в оперативной памяти
- Находит приближенные диссонансы, т.к. РАА и SAX сжимают и кодируют подпоследовательности ряда

Содержание

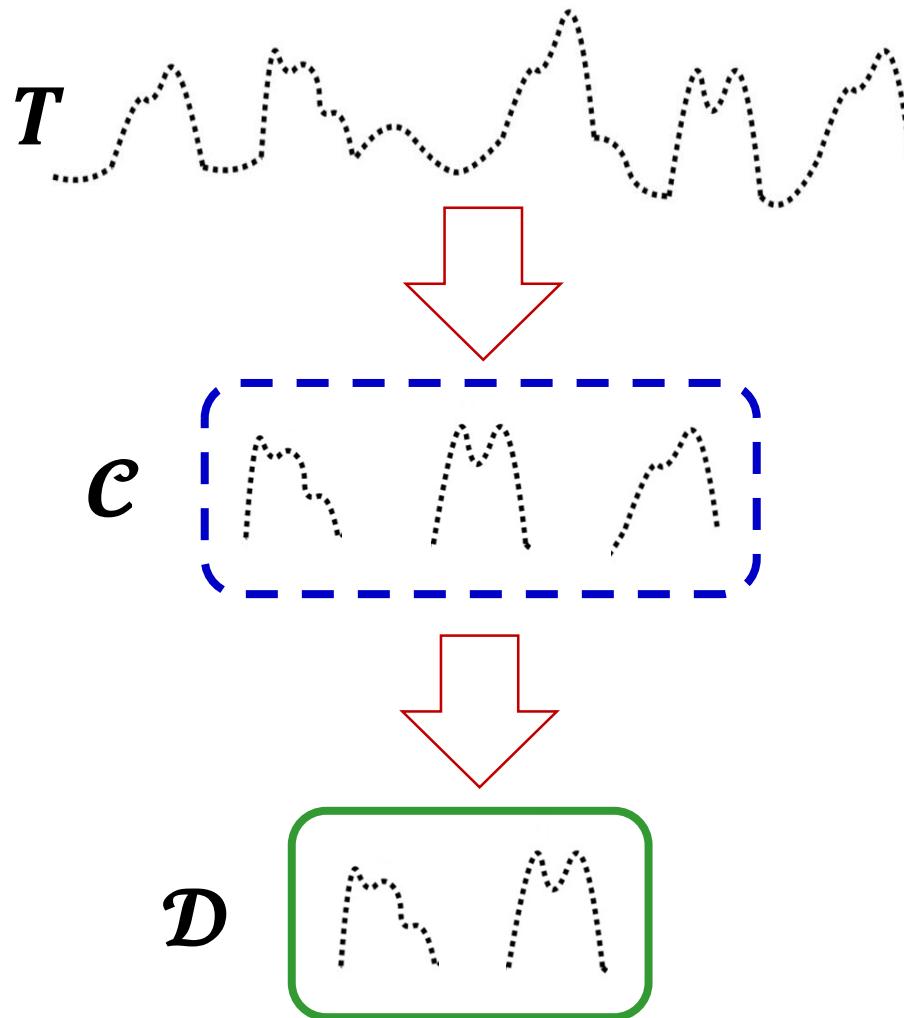
- Понятия аномалии и диссонанса
- Алгоритм HOTSAX
- **Алгоритм DRAG**
- Алгоритм MERLIN
- Распараллеливание поиска диссонансов

Диапазонный диссонанс (range discord)

- Диапазонный диссонанс – подпоследовательность ряда, расстояние от которой до ее ближайшего соседа не ниже заданного порога
- Дано: ряд T , длина диссонанса m , порог r
- Найти: $\mathcal{D} = \{d_1, d_2, \dots\}$ $d_i \in \mathcal{D} \Leftrightarrow \min_{\{s \in S_T^m \mid s \cap d_i = \emptyset\}} \text{Dist}(d_i, s) \geq r$



Алгоритм DRAG (Discord Range Aware Gathering)



1. Отбор

За одно сканирование ряда сформировать **множество кандидатов** в диссонансы

2. Очистка

За одно сканирование ряда **отбросить кандидатов**, которые являются ложными диссонансами

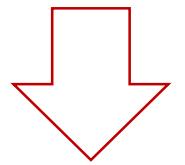
DRAG: Отбор кандидатов

```

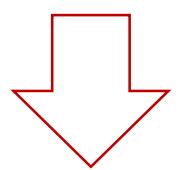
 $\mathcal{C} := \{T_{1,m}\}$ 
while not end of  $T$ 
    get next subsequence  $s$ 
    isCandidate := TRUE
    for each  $c_i \in \mathcal{C}$  and  $s \cap c_i = \emptyset$ 
        if  $\text{Dist}(s, c_i) < r$  then
             $\mathcal{C} := \mathcal{C} \setminus c_i$ ; isCandidate := FALSE
    if isCandidate = TRUE then  $\mathcal{C} := \mathcal{C} \cup s$ 

```

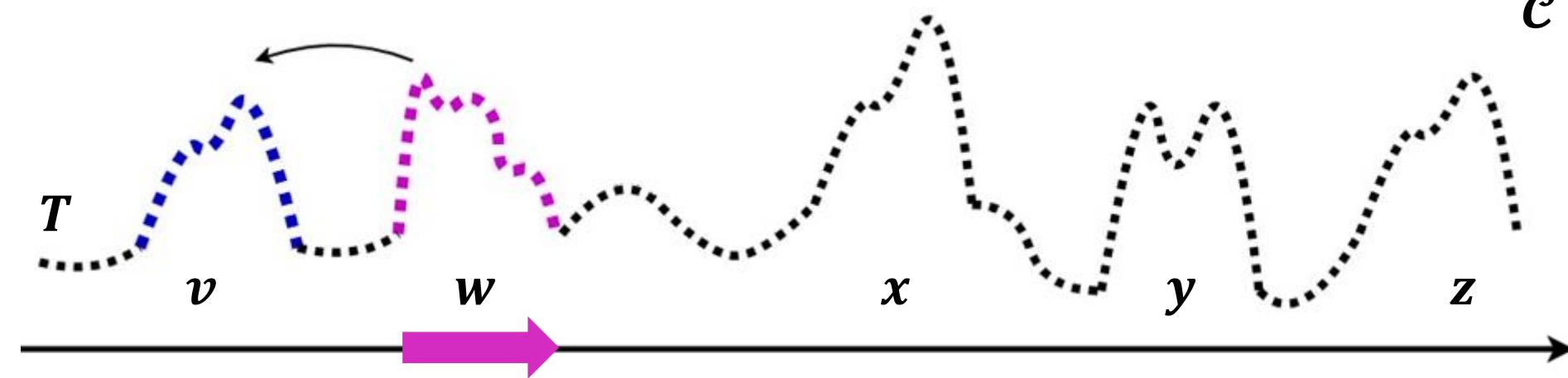
$$\mathcal{C} = \{\nu\}$$



$$\text{Dist}(w, \nu) \geq r$$



$$\mathcal{C} = \{\nu, w\}$$



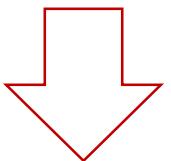
DRAG: Отбор кандидатов

```

 $\mathcal{C} := \{T_{1,m}\}$ 
while not end of  $T$ 
    get next subsequence  $s$ 
     $isCandidate := \text{TRUE}$ 
    for each  $c_i \in \mathcal{C}$  and  $s \cap c_i = \emptyset$ 
        if  $\text{Dist}(s, c_i) < r$  then
             $\mathcal{C} := \mathcal{C} \setminus c_i$ ;  $isCandidate := \text{FALSE}$ 
    if  $isCandidate = \text{TRUE}$  then  $\mathcal{C} := \mathcal{C} \cup s$ 

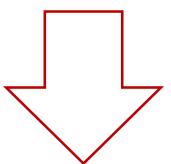
```

$$\mathcal{C} = \{\mathbf{v}, \mathbf{w}\}$$

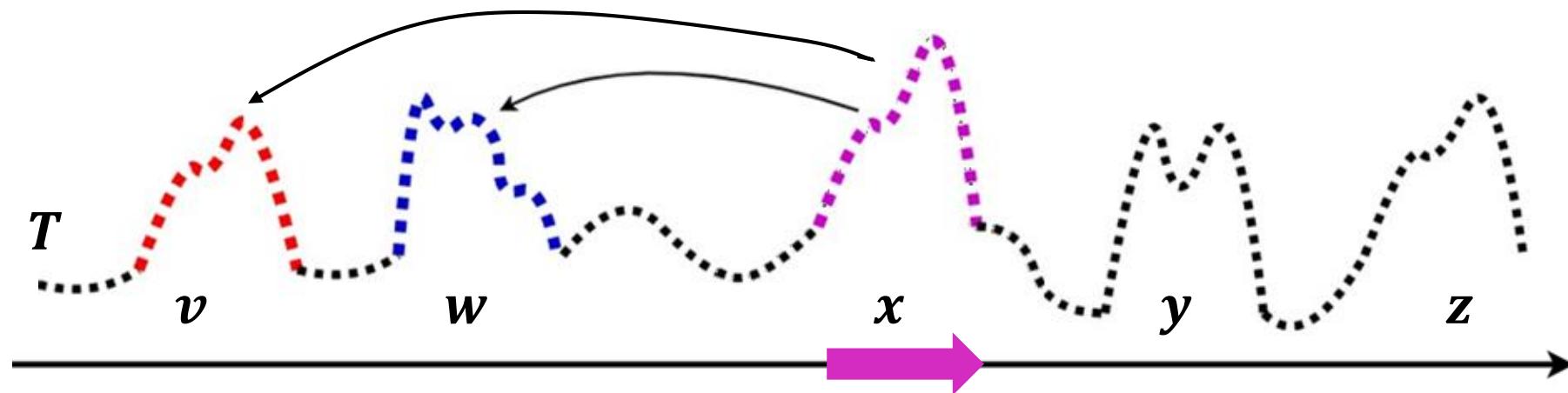


$$\text{Dist}(\mathbf{x}, \mathbf{v}) < r$$

$$\text{Dist}(\mathbf{x}, \mathbf{w}) \geq r$$



$$\mathcal{C} = \{\mathbf{w}\}$$

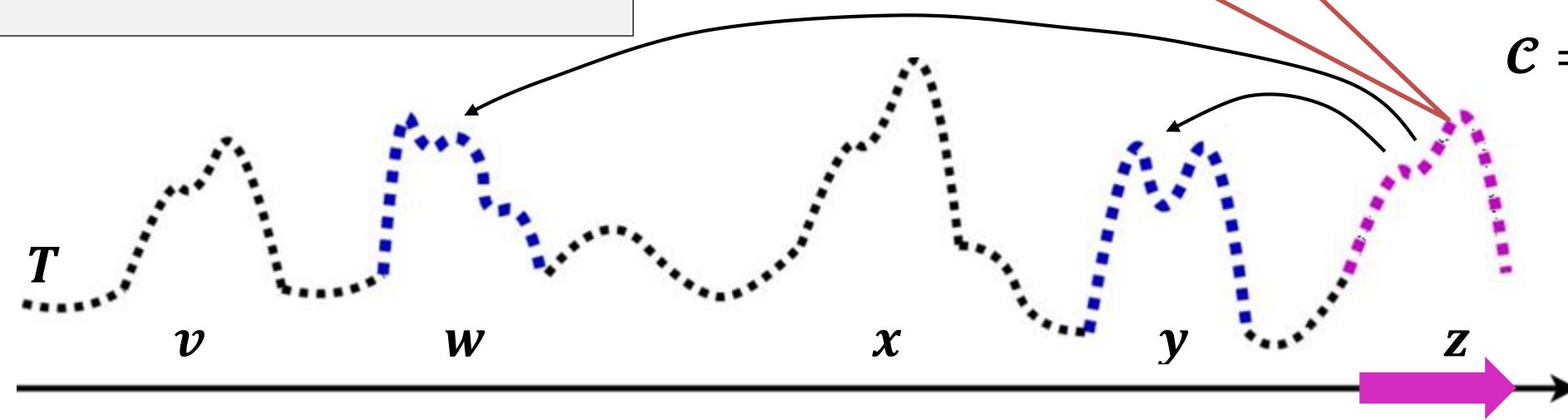


DRAG: Отбор кандидатов

```

 $\mathcal{C} := \{T_{1,m}\}$ 
while not end of  $T$ 
    get next subsequence  $s$ 
     $isCandidate := \text{TRUE}$ 
    for each  $c_i \in \mathcal{C}$  and  $s \cap c_i = \emptyset$ 
        if  $\text{Dist}(s, c_i) < r$  then
             $\mathcal{C} := \mathcal{C} \setminus c_i$ ;  $isCandidate := \text{FALSE}$ 
    if  $isCandidate = \text{TRUE}$  then  $\mathcal{C} := \mathcal{C} \cup s$ 

```



**z – ложный диссонанс, т.к.
 $\text{Dist}(z, v) < r$
 $\text{Dist}(z, x) < r$.
 Но v и x были удалены!**

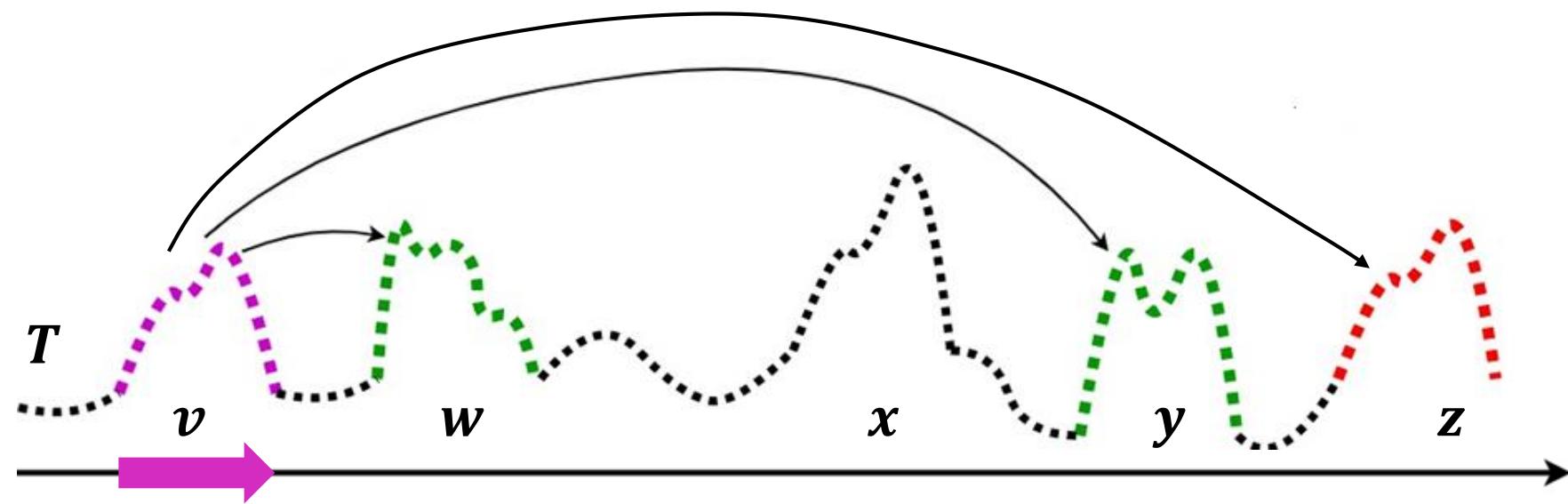
$\mathcal{C} = \{w, y\}$
 ↓
 $\text{Dist}(z, w) \geq r$
 $\text{Dist}(z, y) \geq r$
 ↓
 $\mathcal{C} = \{w, y, z\}$

DRAG: Очистка кандидатов

```

 $\mathcal{D} := \mathcal{C}$ 
while not end of  $T$ 
  get next subsequence  $s$ 
  for each  $d_i \in \mathcal{D}$  and  $s \cap d_i = \emptyset$ 
    if  $\text{Dist}(s, d_i) < r$  then
       $\mathcal{D} := \mathcal{D} \setminus d_i$ 

```



$$\mathcal{D} = \{\textcolor{violet}{w}, \textcolor{green}{y}, z\}$$


$$\mathcal{D} = \{\mathbf{w}, \mathbf{y}\}$$

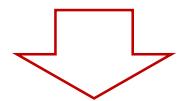
DRAG: Очистка кандидатов

```

 $\mathcal{D} := \mathcal{C}$ 
while not end of  $T$ 
    get next subsequence  $s$ 
    for each  $d_i \in \mathcal{D}$  and  $s \cap d_i = \emptyset$ 
        if  $\text{Dist}(s, d_i) < r$  then
             $\mathcal{D} := \mathcal{D} \setminus d_i$ 

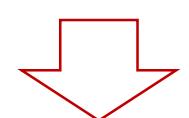
```

$$\mathcal{D} = \{w, y\}$$

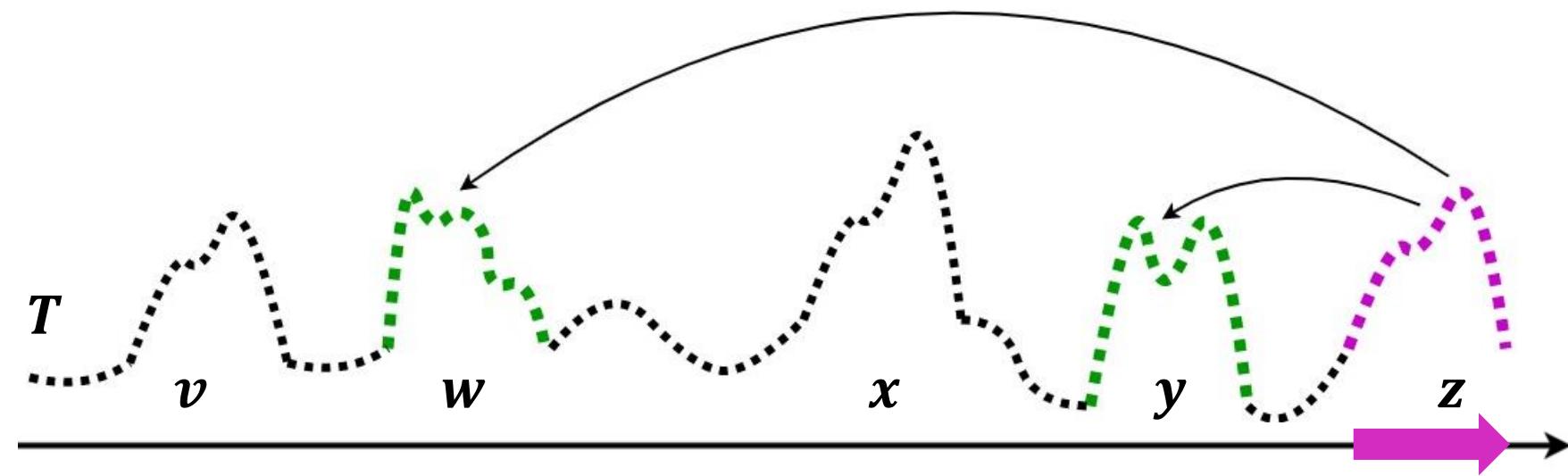


$$\text{Dist}(z, w) \geq r$$

$$\text{Dist}(z, y) \geq r$$



$$\mathcal{D} = \{w, y\}$$



Эвристический подбор параметра r

1. Выбрать случайный сегмент ряда максимальной длины, который может быть размещен в памяти
2. Найти в выбранном сегменте диссонанс с помощью алгоритма HOT SAX
3. Взять в качестве порога r расстояние от найденного диссонанса до его ближайшего соседа

Проблемы DRAG

1. Ручной подбор длины диссонанса t

- Не всегда заранее известна длина аномалии
- Запуск DRAG для всех возможных длин вычислительно неосуществим

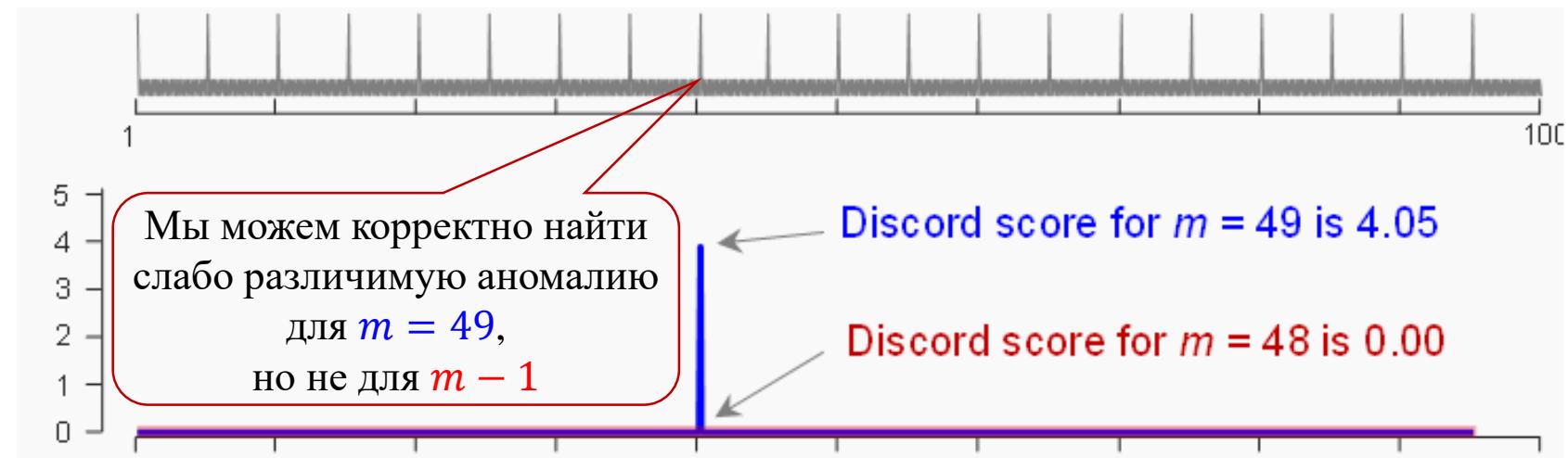
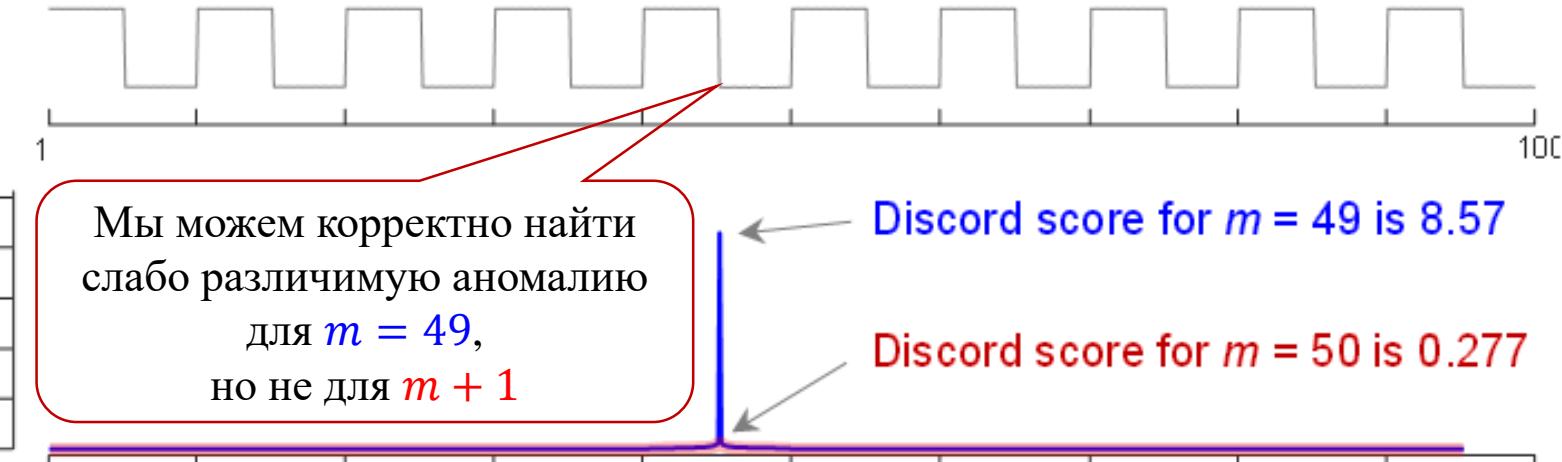
2. Ручной подбор порога r

- Слишком большой порог – нет диссонансов, слишком маленький порог – много ложных диссонансов

Чтобы найти *все* аномалии, нужно проверить *все* значения m

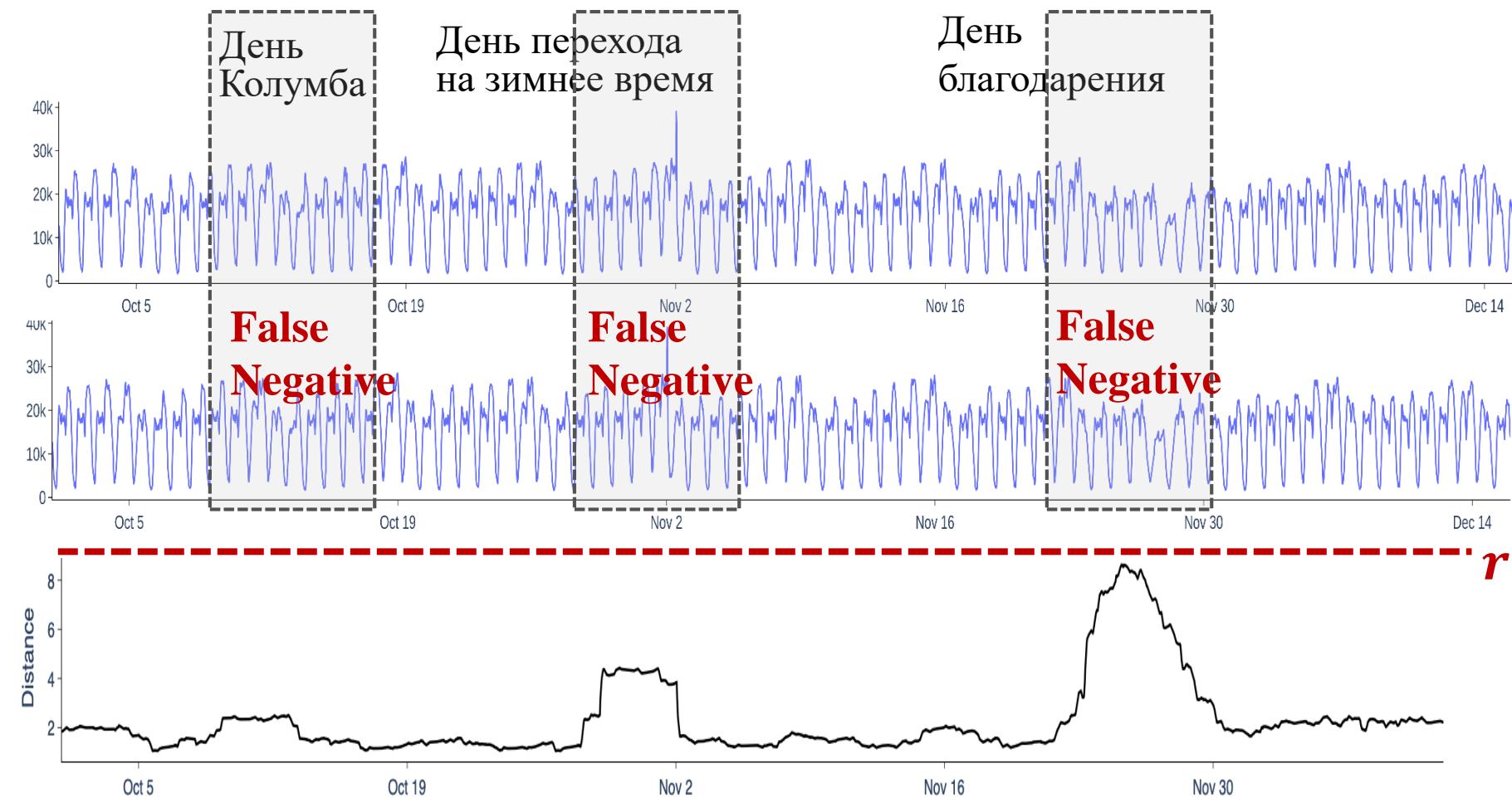
```
// Создание временного ряда
for i ← 1 to 1000 do
     $T_i \leftarrow i \bmod 2$ 
end for
for i ← 1 to 1000 step 100 do
     $T_{i:i+50} \leftarrow T_{i:i+50} + 100$ 
end for
// Создание диссонанса
 $T_{453:499} \leftarrow \text{rand}() / 20$ 

// Создание временного ряда
for i ← 1 to 1000 do
    if  $i \bmod 50 \neq 0$  then
         $T_i \leftarrow i \bmod 2$ 
    else
         $T_i \leftarrow 5$ 
    end if
end for
// Создание диссонанса
 $T_{430:431} \leftarrow 0$ 
```

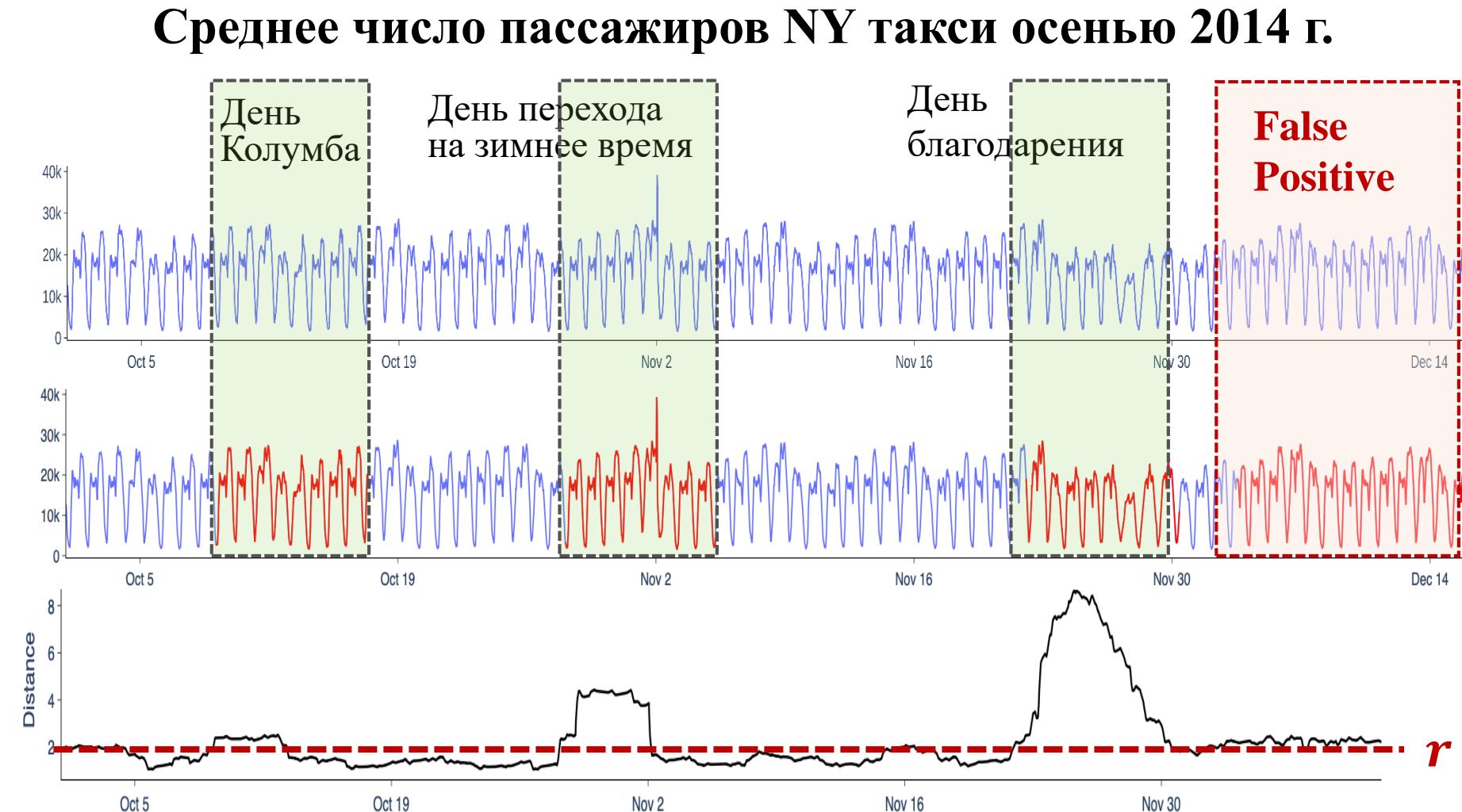


Ручной подбор порога: $r \rightarrow +\infty \Rightarrow$ нет диссонансов

Среднее число пассажиров NY такси осенью 2014 г.



Ручной подбор порога: $r \rightarrow 0 \Rightarrow$ ложные аномалии

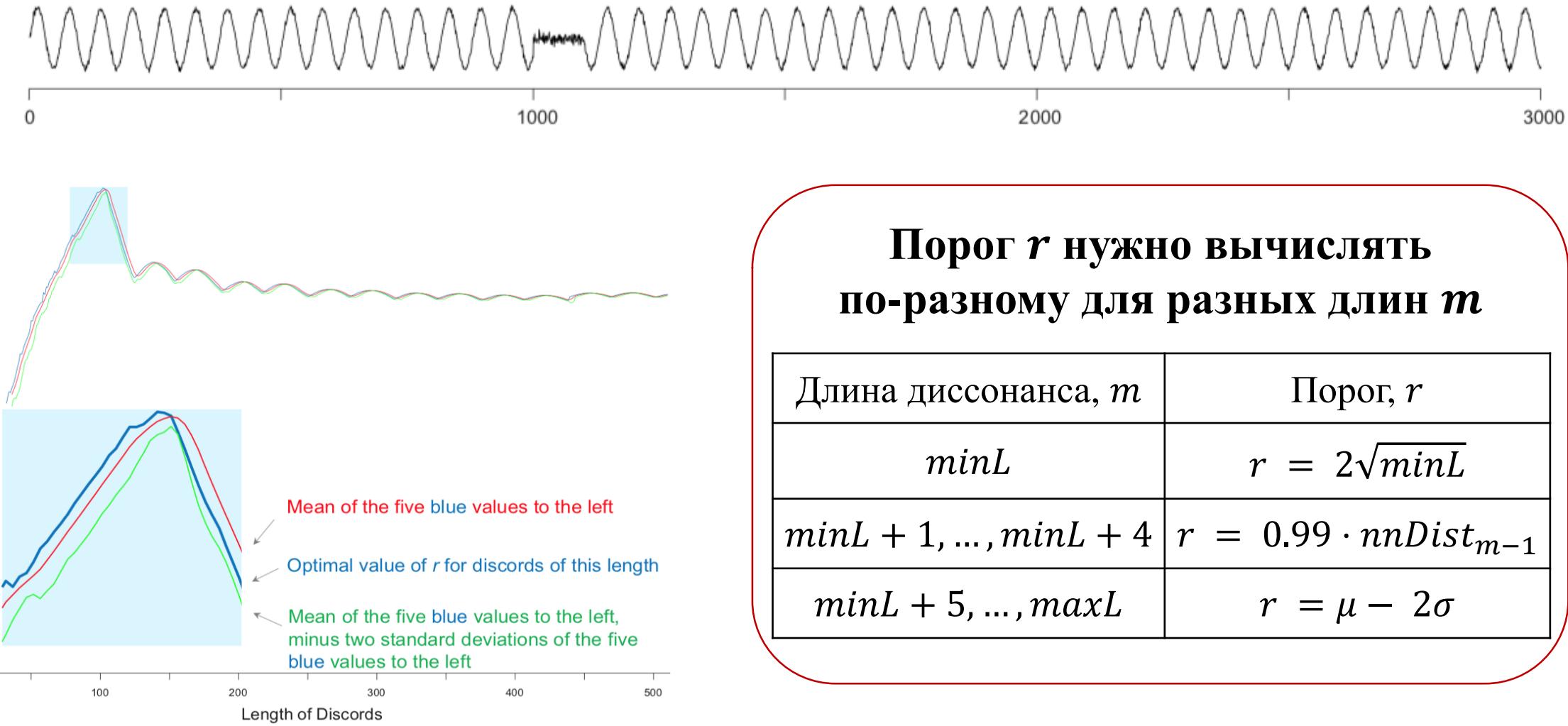


Содержание

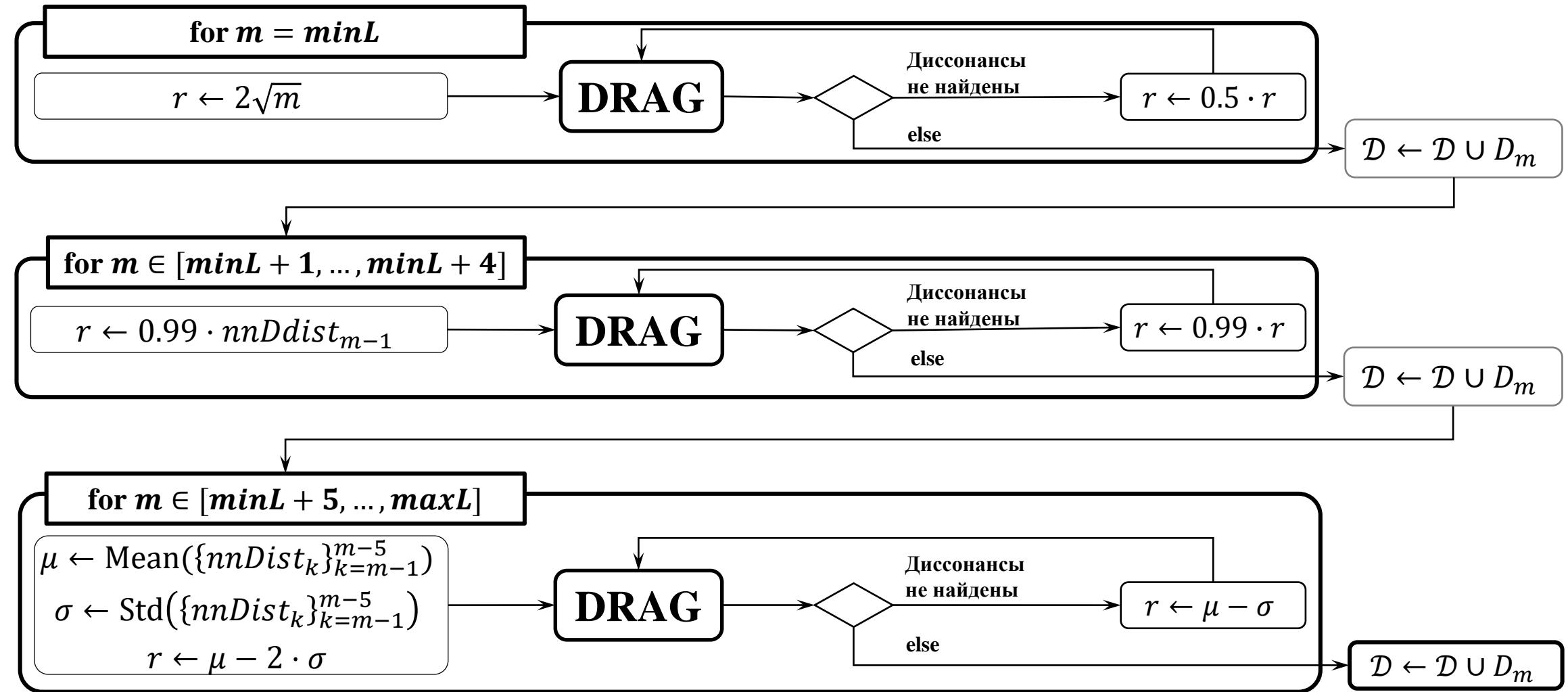
- Понятия аномалии и диссонанса
- Алгоритм HOTSAX
- Алгоритм DRAG
- **Алгоритм MERLIN**
- Распараллеливание поиска диссонансов

MERLIN : адаптивное вычисление порога r

Поиск аномалий временного ряда



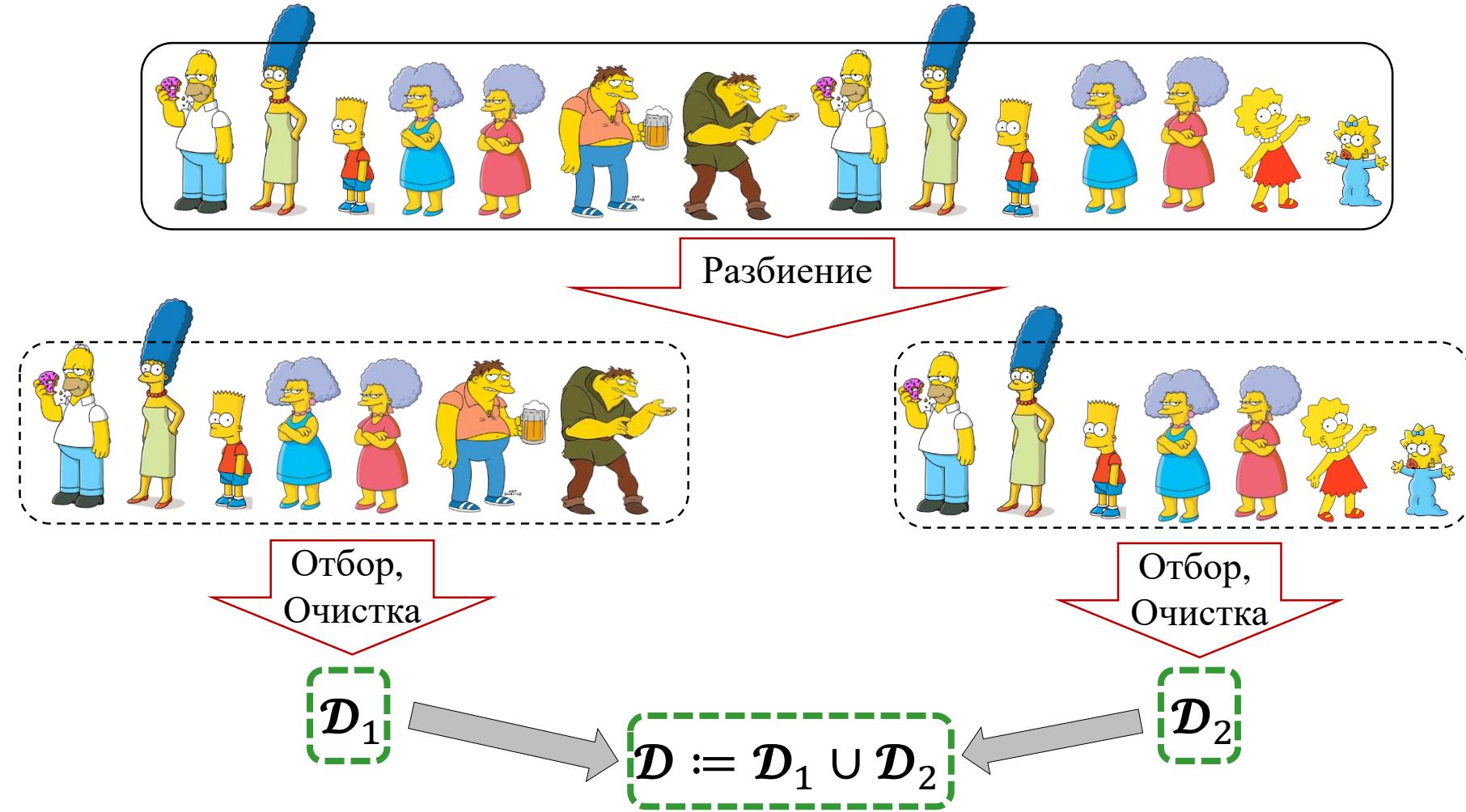
MERLIN



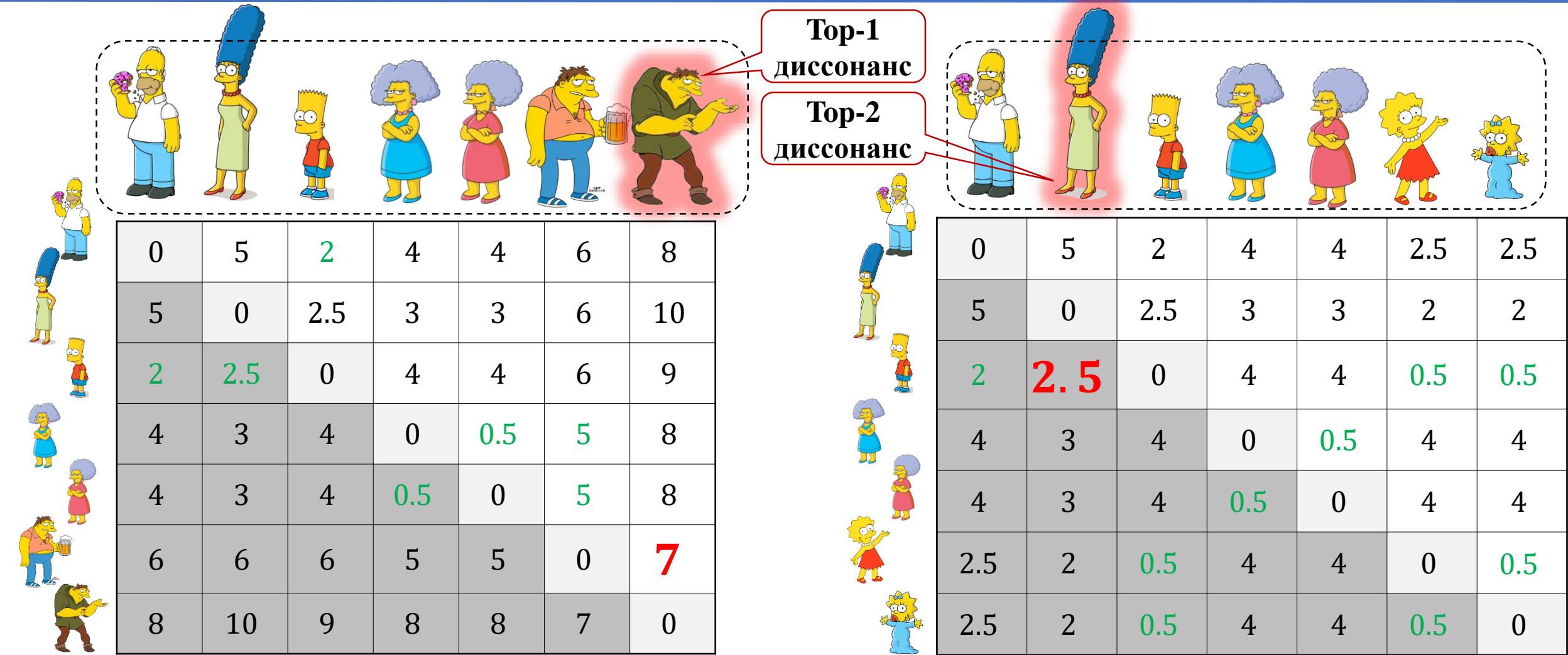
Содержание

- Понятия аномалии и диссонанса
- Алгоритм HOTSAX
- Алгоритм DRAG
- Алгоритм MERLIN
- **Распараллеливание поиска диссонансов**

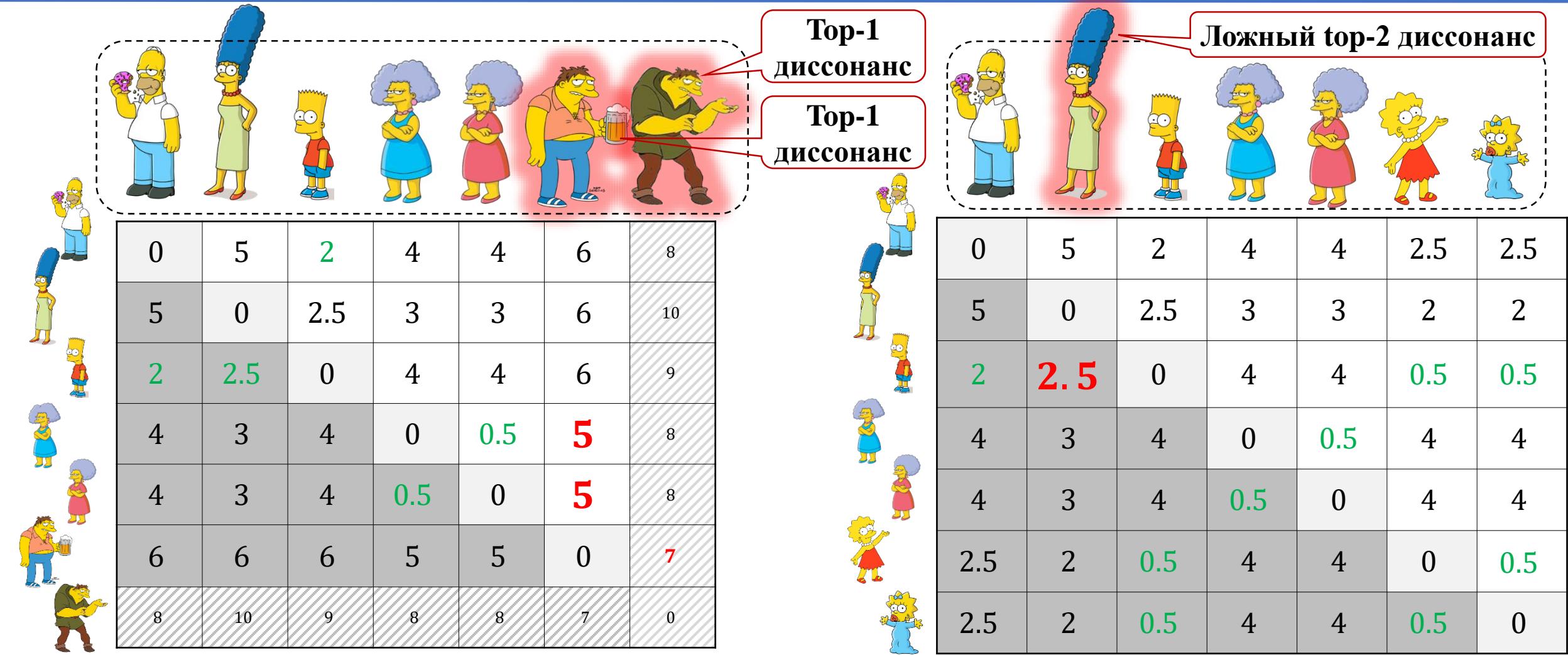
Наивное распараллеливание ...



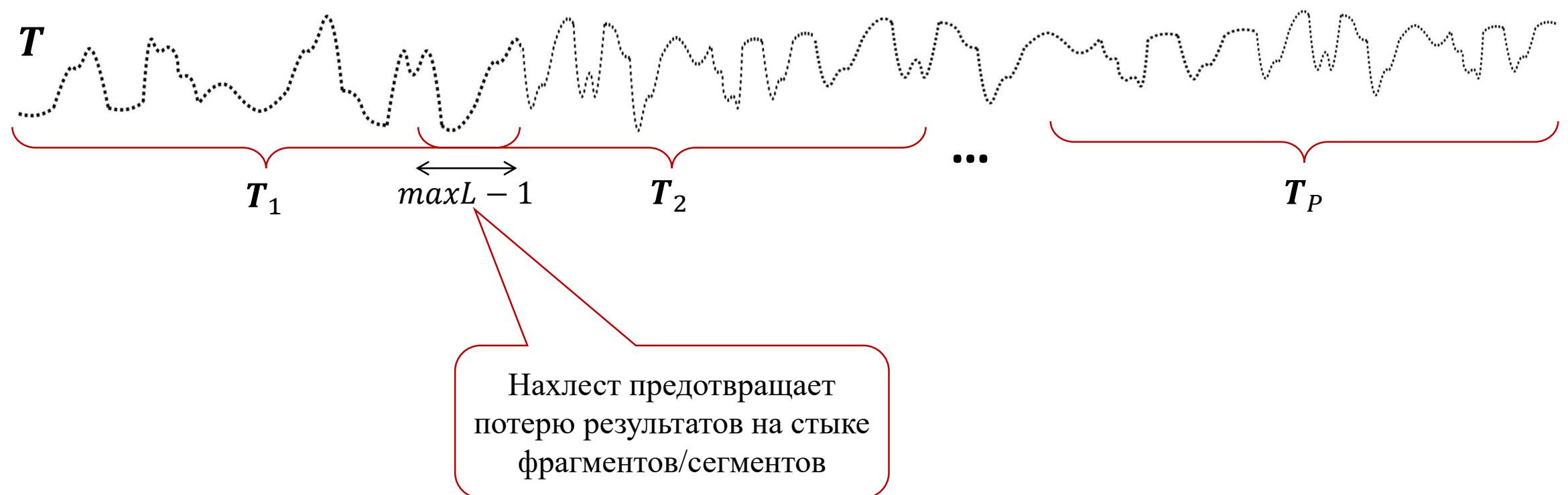
Наивное распараллеливание ...



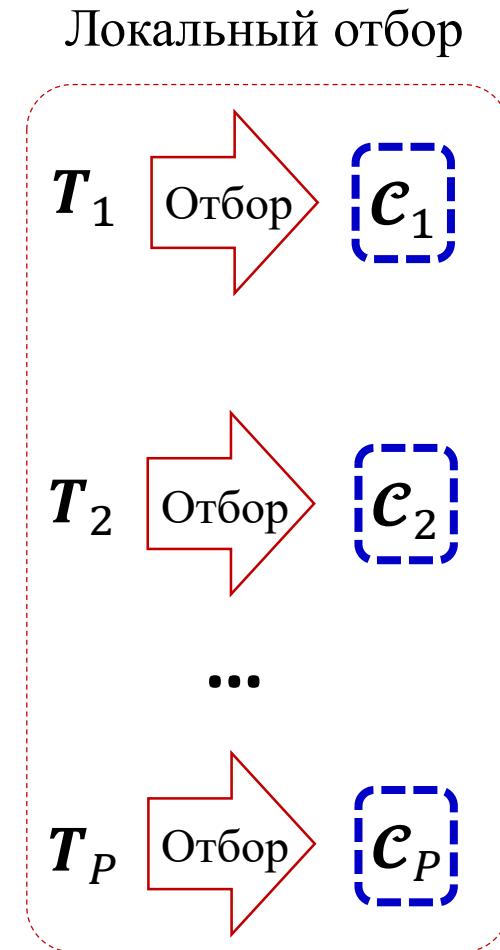
... не работает



Фрагментация/сегментация ряда

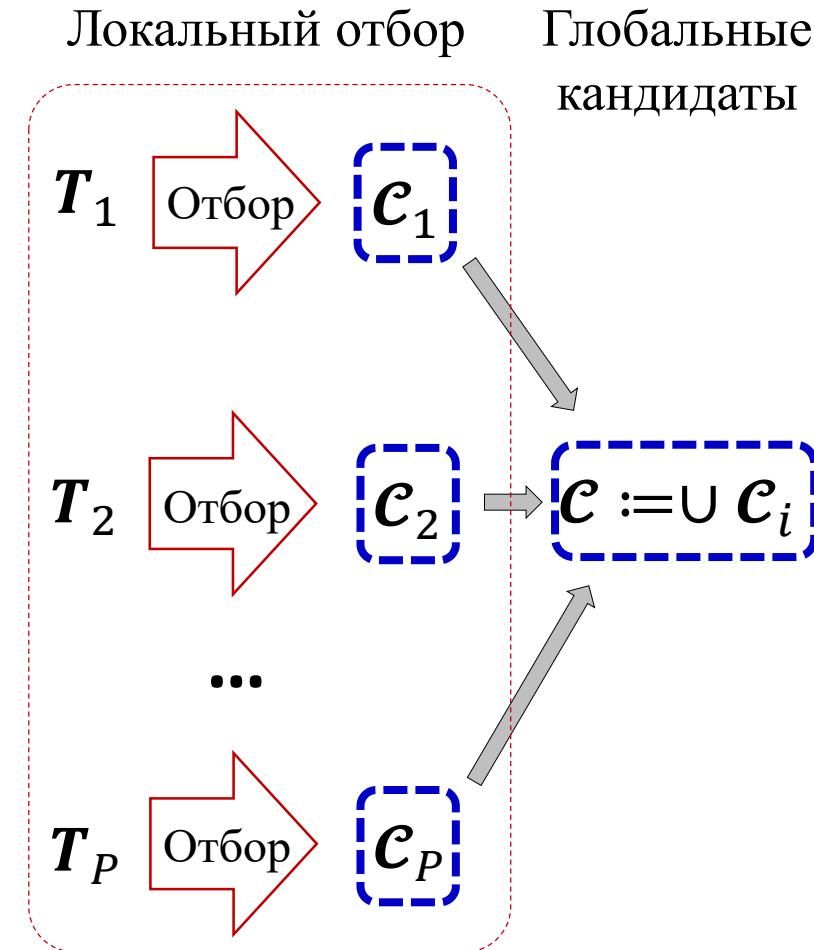


Параллельный поиск диссонансов: Схема 1*



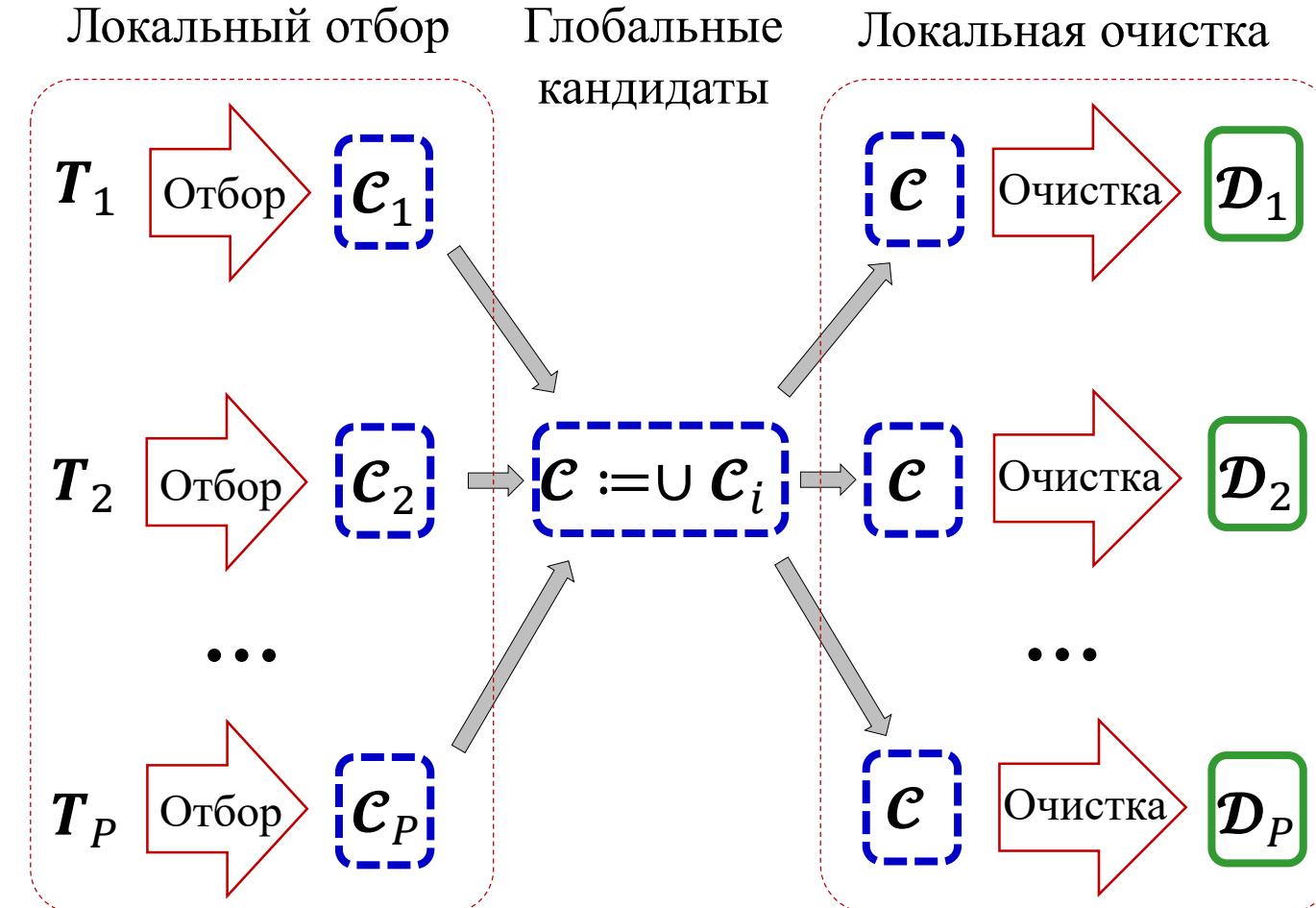
* Yankov D. et al. Disk aware discord discovery: finding unusual time series in terabyte sized datasets. ICDM 2007. pp. 381-390. DOI: [10.1109/ICDM.2007.61](https://doi.org/10.1109/ICDM.2007.61).

Параллельный поиск диссонансов: Схема 1*



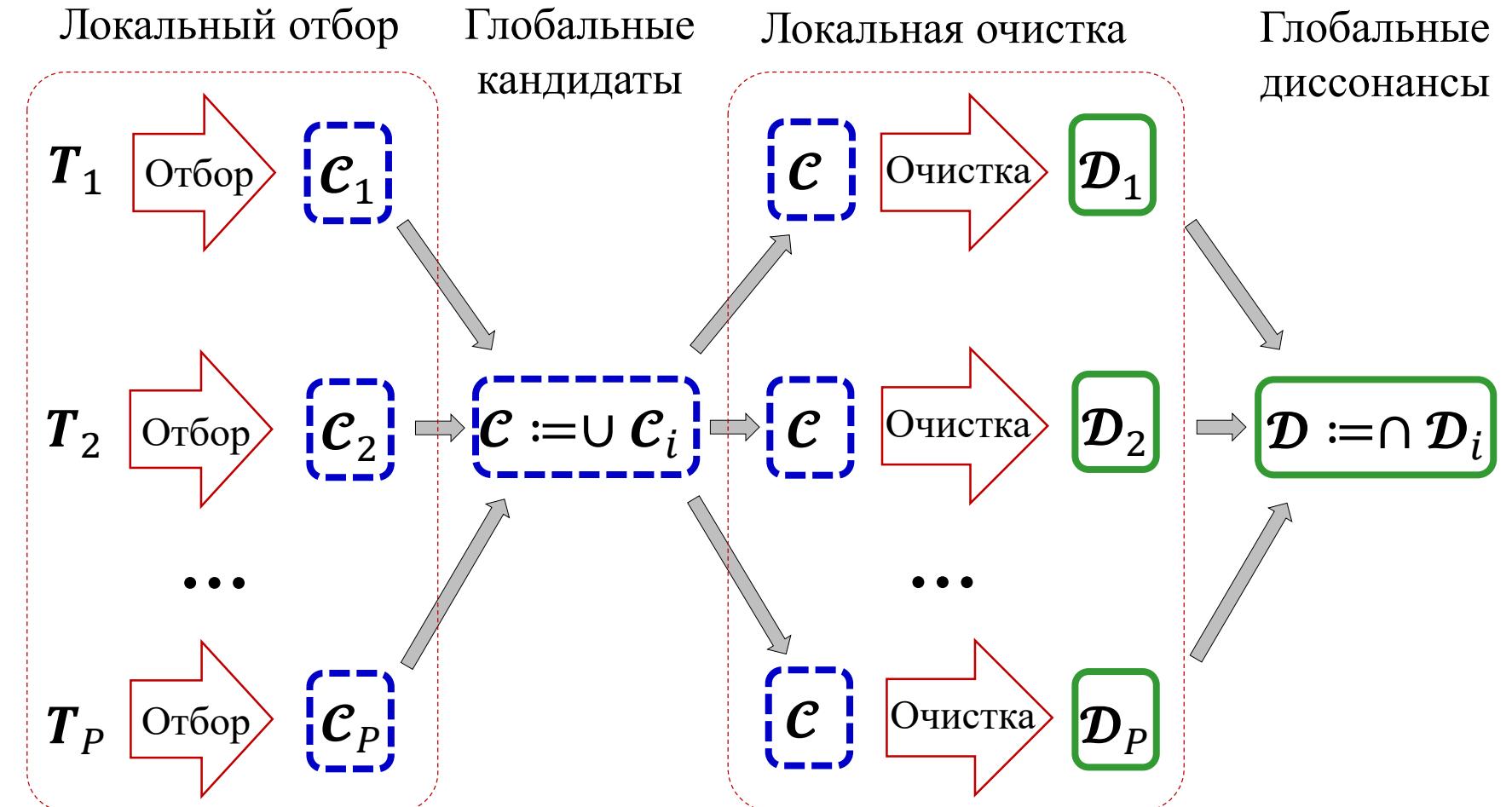
* Yankov D. et al. Disk aware discord discovery: finding unusual time series in terabyte sized datasets. ICDM 2007. pp. 381-390. DOI: [10.1109/ICDM.2007.61](https://doi.org/10.1109/ICDM.2007.61).

Параллельный поиск диссонансов: Схема 1*



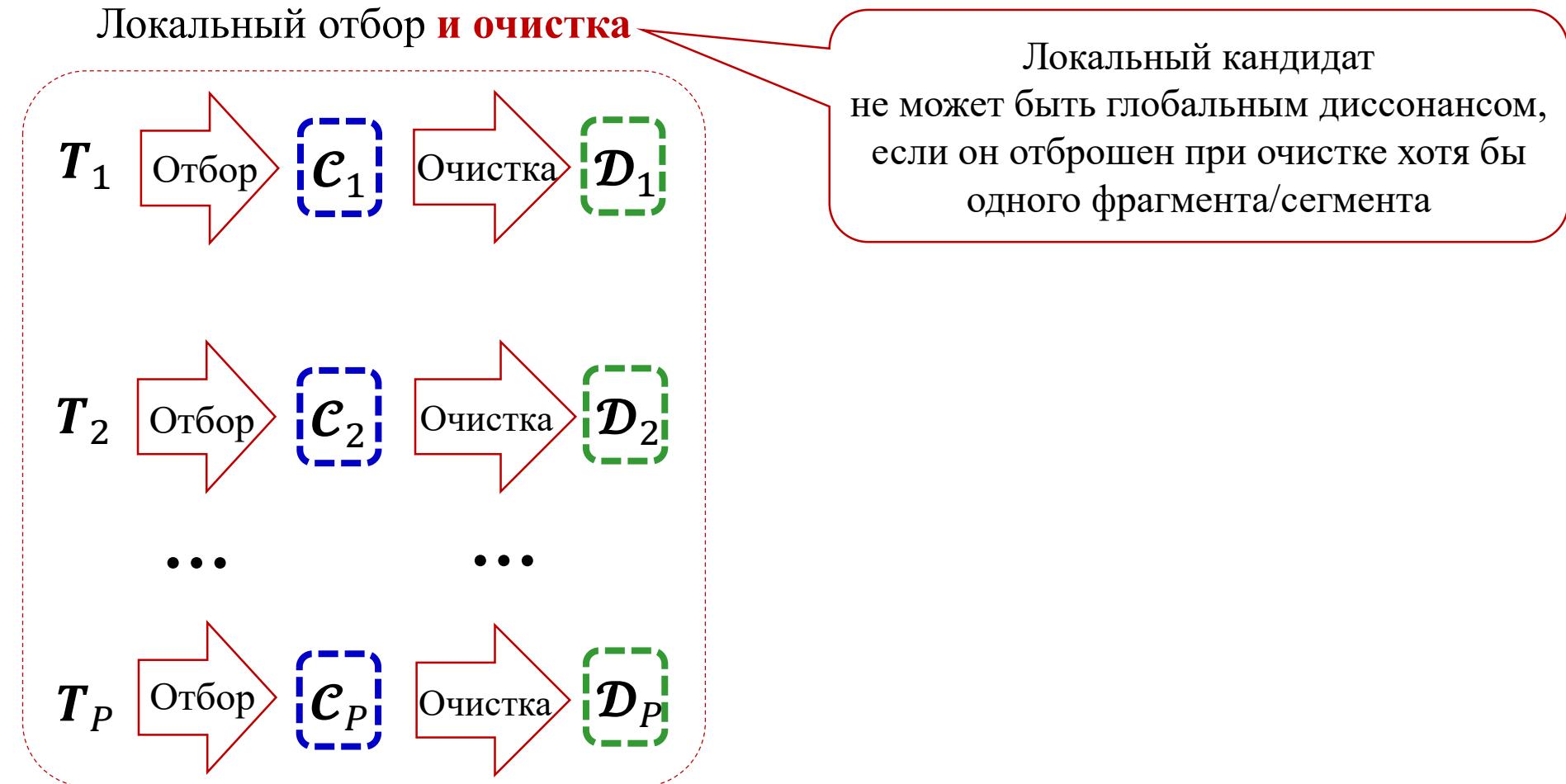
* Yankov D. et al. Disk aware discord discovery: finding unusual time series in terabyte sized datasets. ICDM 2007. pp. 381-390. DOI: [10.1109/ICDM.2007.61](https://doi.org/10.1109/ICDM.2007.61).

Параллельный поиск диссонансов: Схема 1*



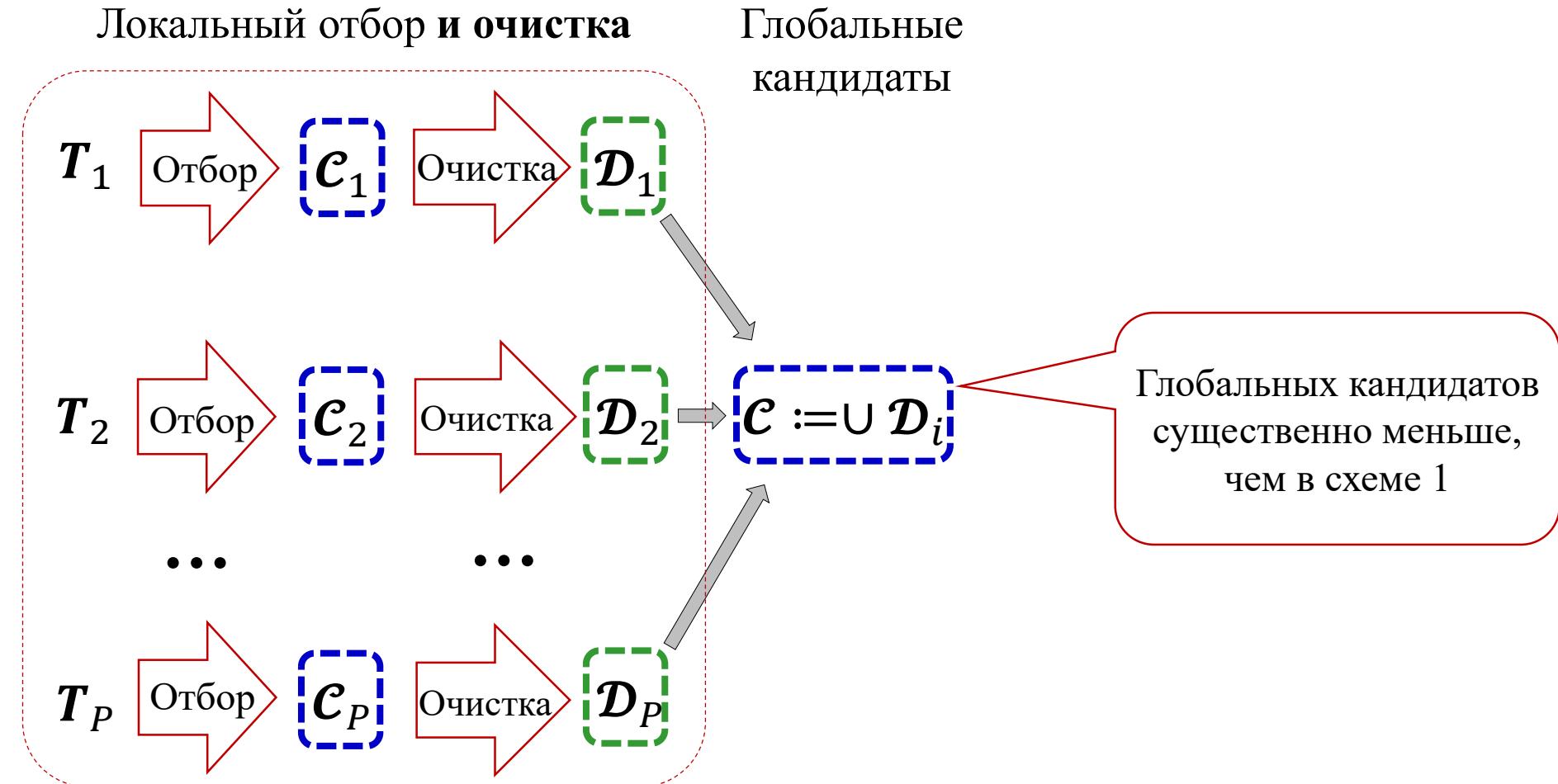
* Yankov D. et al. Disk aware discord discovery: finding unusual time series in terabyte sized datasets. ICDM 2007. pp. 381-390. DOI: [10.1109/ICDM.2007.61](https://doi.org/10.1109/ICDM.2007.61).

Параллельный поиск диссонансов: Схема 2*



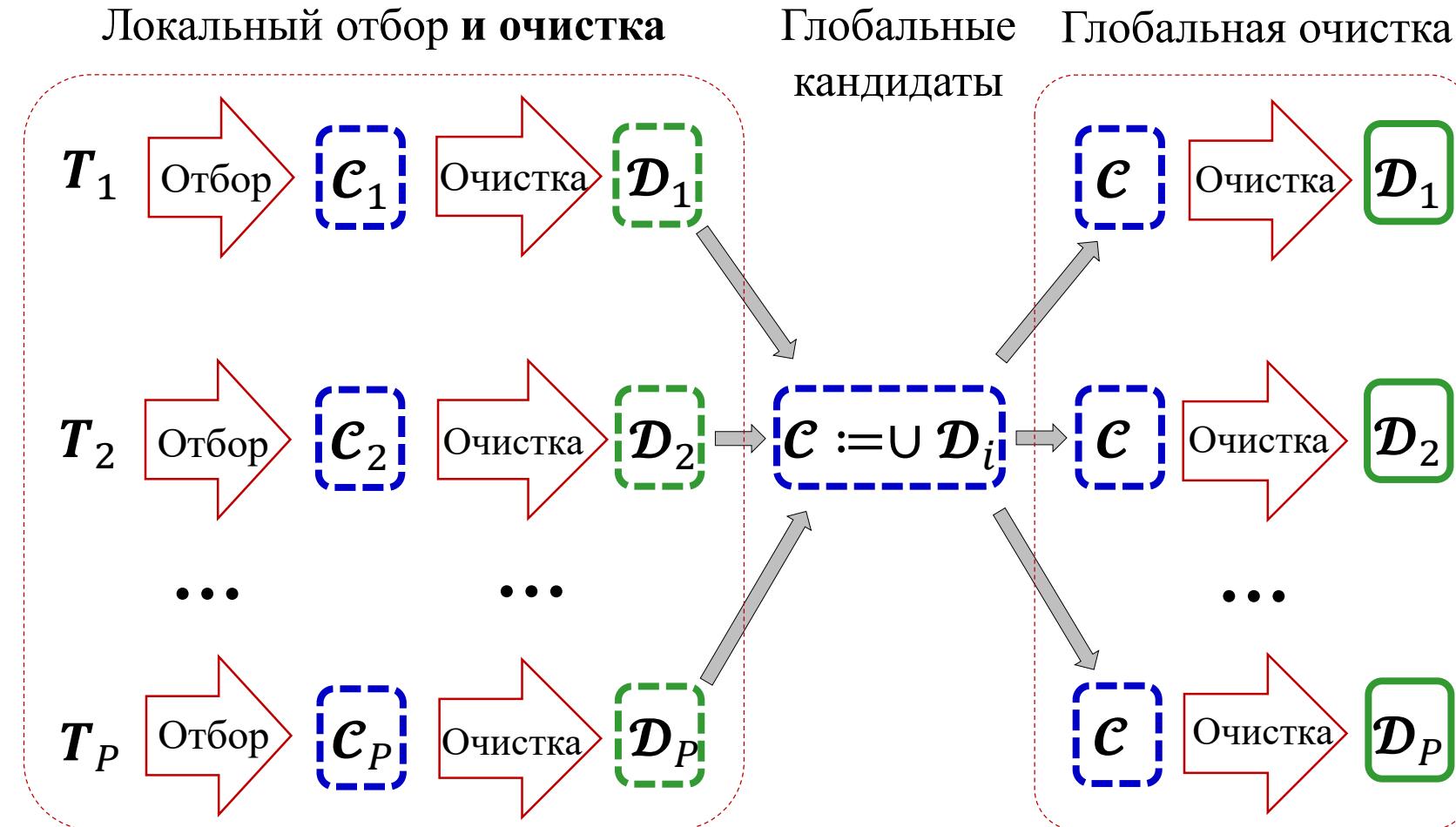
* Zymbler M. et al. A parallel approach to discords discovery in massive time series data. Computers, Materials & Continua. 2021. 66(2). pp. 1867–1876. DOI: [10.32604/cmc.2020.014232](https://doi.org/10.32604/cmc.2020.014232)

Параллельный поиск диссонансов: Схема 2*



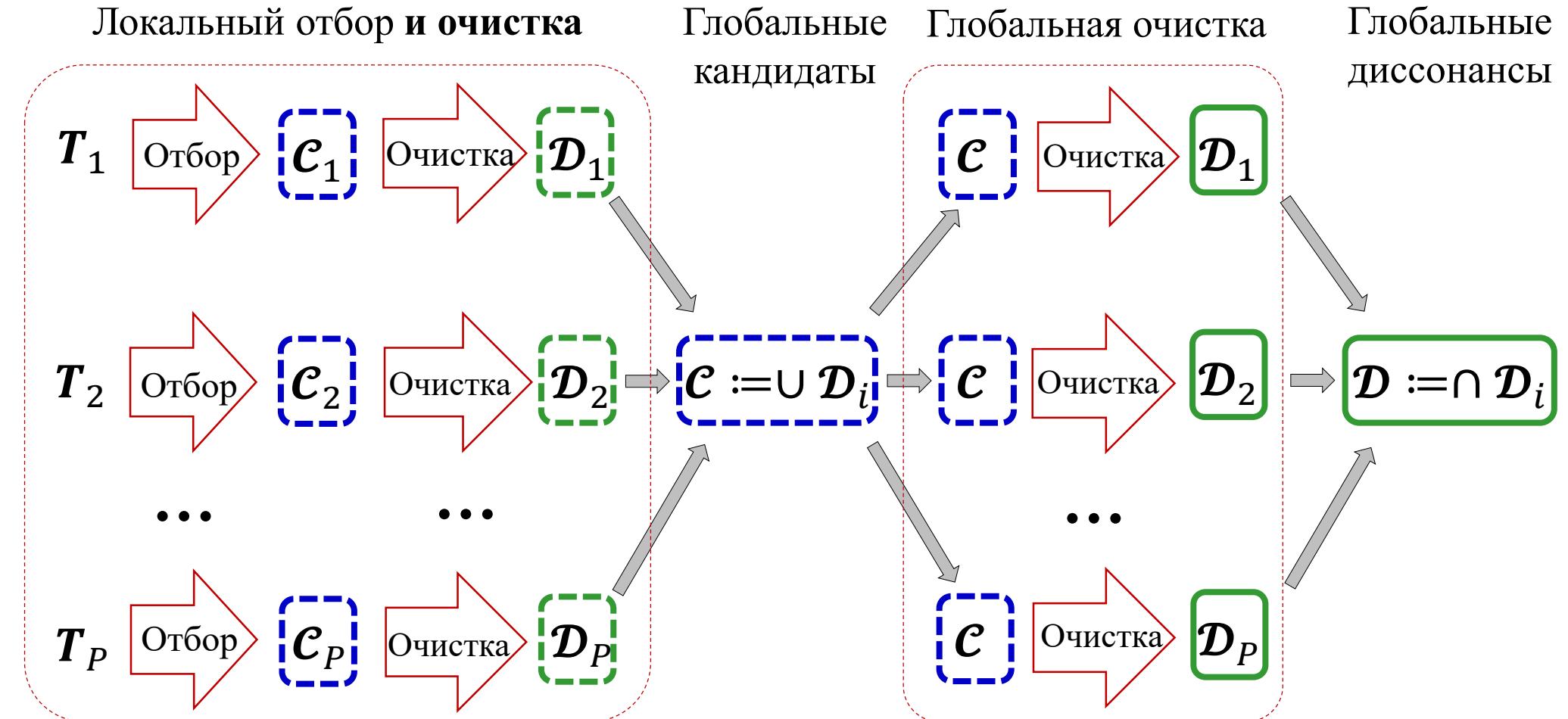
* Zymbler M. et al. A parallel approach to discords discovery in massive time series data. Computers, Materials & Continua. 2021. 66(2). pp. 1867–1876. DOI: [10.32604/cmc.2020.014232](https://doi.org/10.32604/cmc.2020.014232)

Параллельный поиск диссонансов: Схема 2*



* Zymbler M. et al. A parallel approach to discords discovery in massive time series data. Computers, Materials & Continua. 2021. 66(2). pp. 1867–1876. DOI: [10.32604/cmc.2020.014232](https://doi.org/10.32604/cmc.2020.014232)

Параллельный поиск диссонансов: Схема 2*



* Zymbler M. et al. A parallel approach to discords discovery in massive time series data. Computers, Materials & Continua. 2021. 66(2). pp. 1867–1876. DOI: [10.32604/cmc.2020.014232](https://doi.org/10.32604/cmc.2020.014232)

Литература

1. Lin J., Keogh E.J., Fu A.W., Herle H.V. Approximations to magic: Finding unusual medical time series. 18th IEEE Symp. on Computer-Based Med. Syst. (CBMS 2005), 23-24 June 2005, Dublin, Ireland. pp. 329-334. IEEE (2005). <https://doi.org/10.1109/CBMS.2005.34>
2. Yankov D., Keogh E.J., Rebbapragada U. Disk aware discord discovery: Finding unusual time series in terabyte sized datasets. Proc. of the 7th IEEE Int. Conf. on Data Mining (ICDM 2007), October 28-31, 2007, Omaha, Nebraska, USA. pp. 381-390. IEEE (2007).
<https://doi.org/10.1109/ICDM.2007.61>
3. Nakamura T., Imamura M., Mercer R., Keogh E.J. MERLIN: parameter-free discovery of arbitrary length anomalies in massive time series archives. 20th IEEE Int. Conf. on Data Mining, ICDM 2020, Sorrento, Italy, November 17-20, 2020. pp. 1190-1195. IEEE (2020).
<https://doi.org/10.1109/ICDM50108.2020.00147>