Поиск диссонансов временного ряда

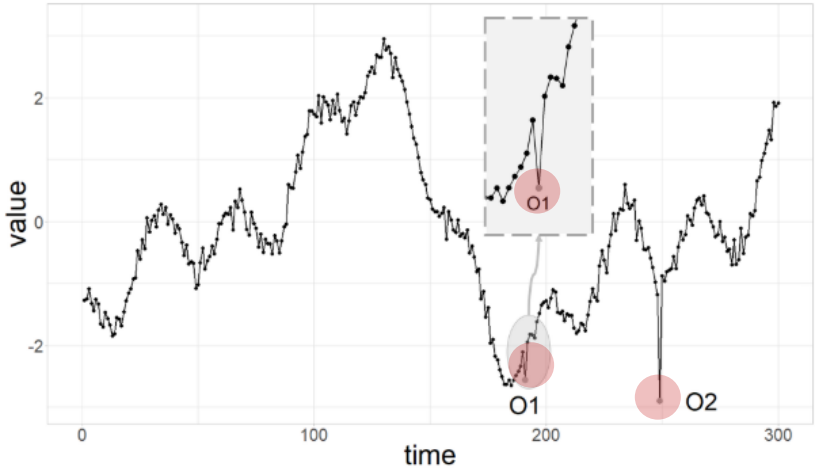
Мир всегда приходит в норму.
Важно лишь, чья она.

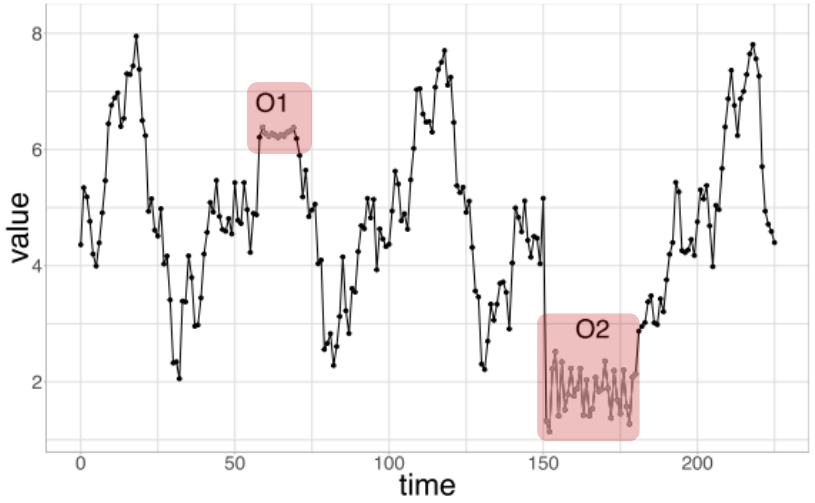
С. Е. Лец

Содержание

- Понятия аномалии и диссонанса
- Алгоритм HOTSAX
- Алгоритм DRAG
- Алгоритм MERLIN

Аномалия – отклонение от нормы





- *Точечная аномалия* – элемент ряда, который является необычным в сравнении со всеми элементами *(глобальный выброс)* или с соседними элементами *(локальный выброс)*
- *Аномальная подпоследовательность* – промежуток элементов ряда, совместное поведение которых необычно, хотя каждый элемент не обязательно является точечной аномалией

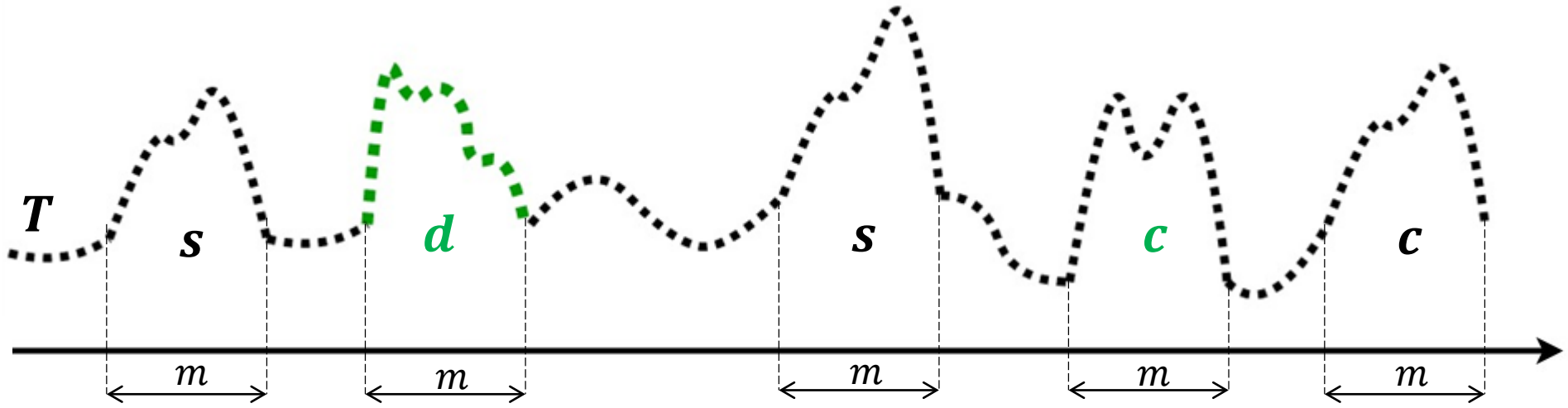
Blazquez-Garca A., *et al.* A review on outlier/anomaly detection in time series data. ACM Comput. Surv. 54(3), 56:1-56:33 (2021). https://doi.org/10.1145/3444690

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.
- **Outliers** are the data points for which 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.
- **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.
- **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.
- **Discord in a time series** is a subsequence of length n whose distance to its nearest non-self match neighbor is greatest.
  Keogh E. *et al.* HOT SAX: Efficiently finding the most unusual time series subsequence. ICDM 2005. pp. 226-233.

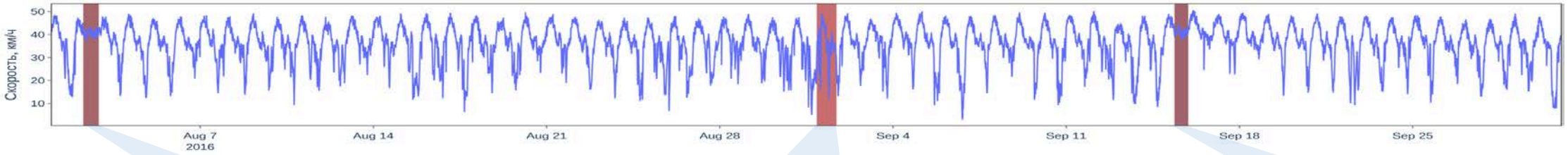
Диссонанс (discord)

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

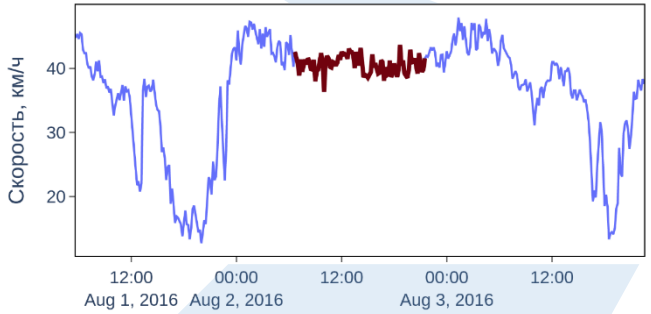


Диссонанс отражает аномалию реальной жизни

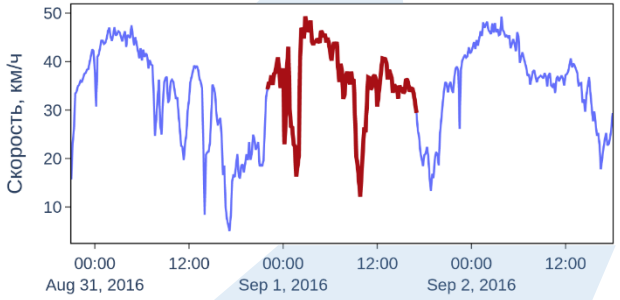
Средняя скорость городского трафика в Гуанчжоу, Китай



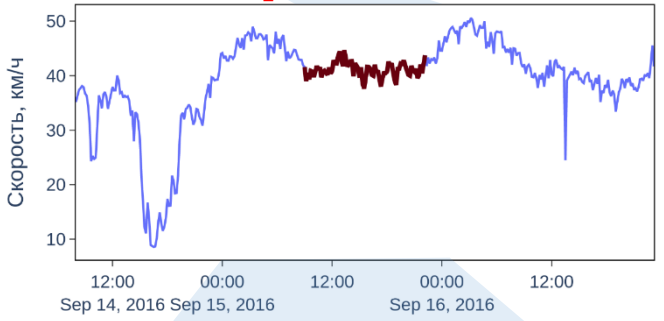
Top-2 диссонанс



Top-3 диссонанс



Top-1 диссонанс

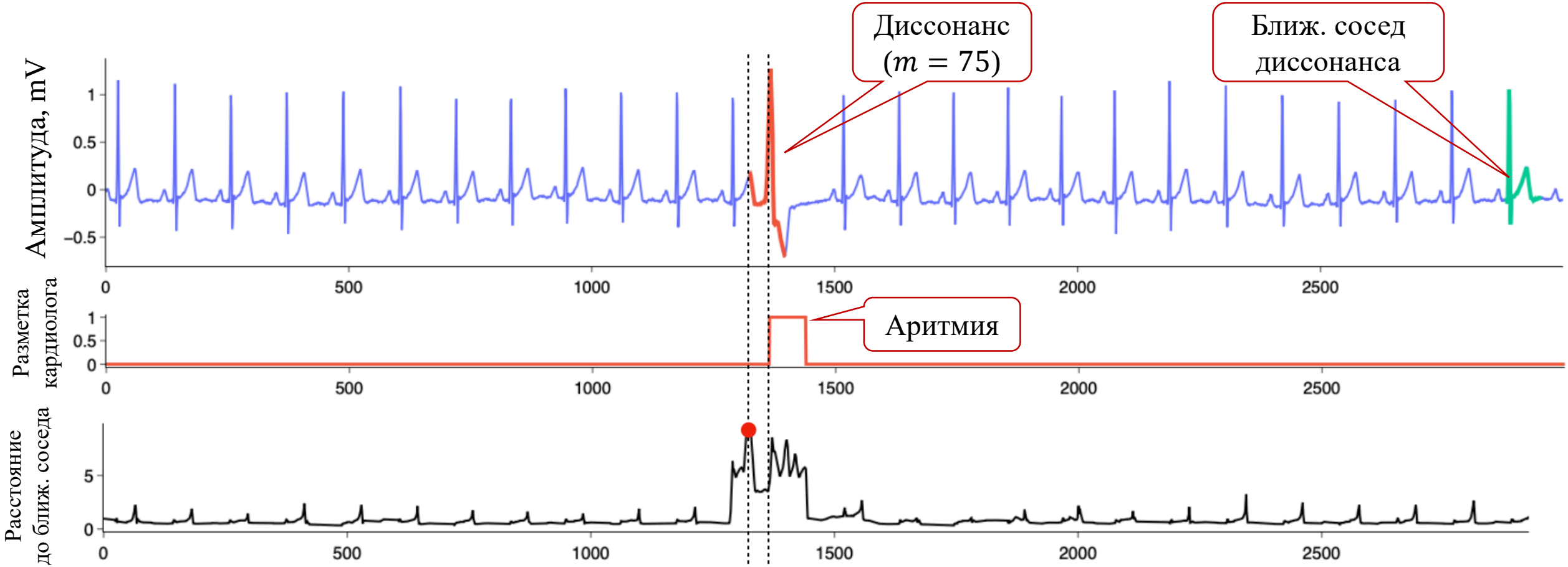


Тайфун Нида

День Победы над Японией

Фестиваль Луны

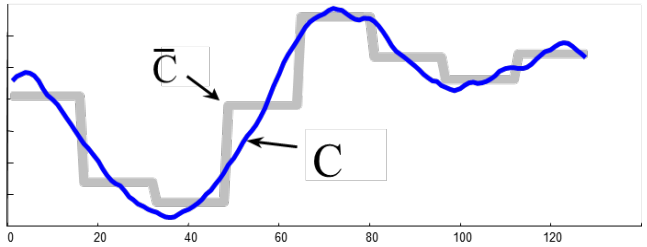
…, но диссонанс не идентичен аномалии

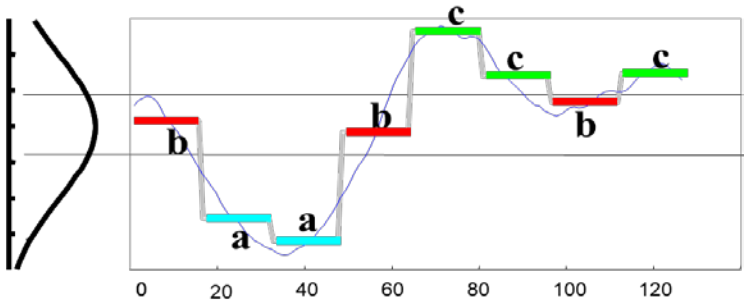


Содержание

Алгоритм HOTSAX (Heuristically Ordered Time series using Symbolic Aggregate ApproXimation)

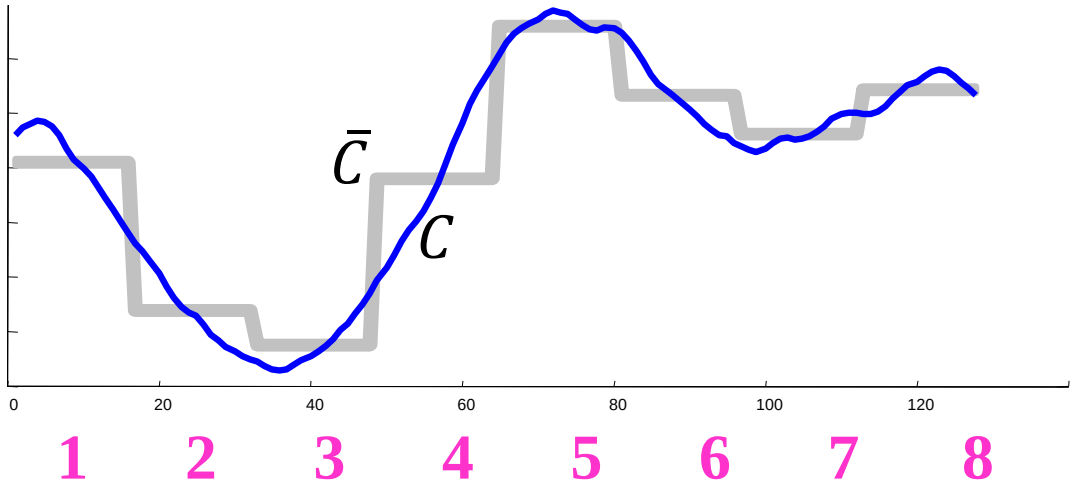
- Особенности
  - Ряд может быть размещен в оперативной памяти
  - Ответ не является точным
- Ключевые идеи
  - Сжатие подпоследовательностей исходного ряда
  - Кодирование сжатых подпоследовательностей ряда
  - Перебор кодированных подпоследовательностей ряда с отбрасыванием





PAA (Piecewise Aggregate Approximation)

- Сжатие (аппроксимация) подпоследовательностей ряда с уменьшением их длины до $w \ll m$
- $\bar{c}_i = \frac{w}{m} \sum_{j=\left(\frac{m}{w}\right)(i-1)+1}^{\left(\frac{m}{w}\right)i} c_j$

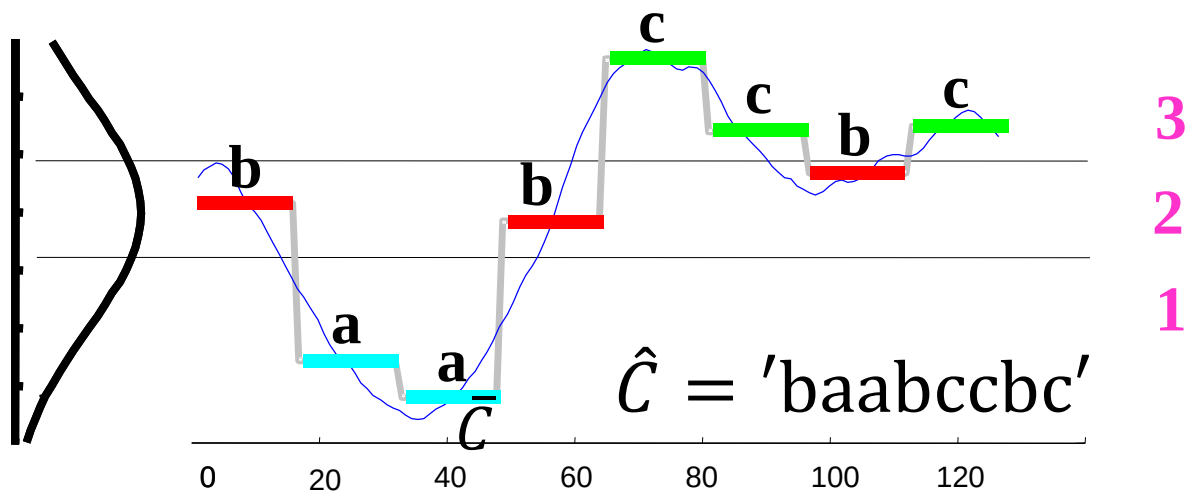


Lin J., Keogh E.J., Lonardi S., Chiu B.Y. A symbolic representation of time series, with implications for streaming algorithms. Proc. of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery, DMKD 2003, San Diego, California, USA, June 13, 2003. 2003. P. 2–11. URL: https://doi.org/10.1145/882082.882086

SAX (SAX, Symbolic Aggregate ApproXimation)

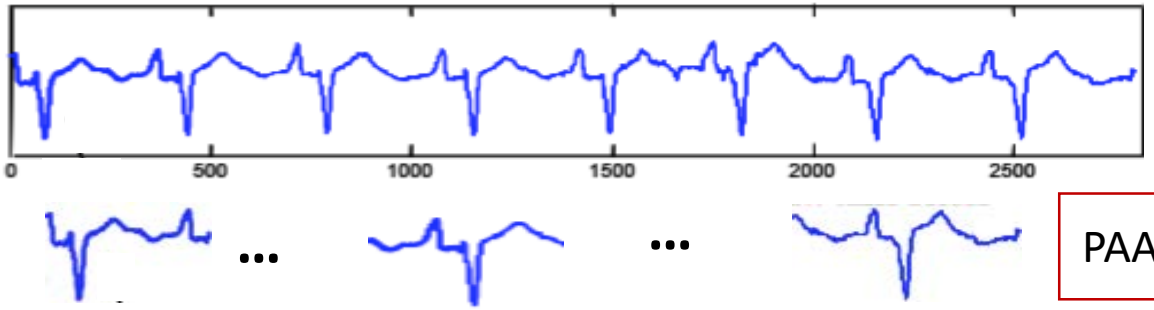
- Кодирование сжатых подпоследовательностей в слова алфавита $\mathcal{A} = (\alpha_1, \dots, \alpha_{|\mathcal{A}|})$, $\alpha_1 = 'a'$, $\alpha_2 = 'b'$ и т.д.
- Таблица кодирования:
  - $\hat{c}_i = \alpha_i \Leftrightarrow \beta_{j-1} \le c_i < \beta_j$
  - $\beta_0 = -\infty$, $\beta_{|\mathcal{A}|} = +\infty$
  - 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



β_i \ 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

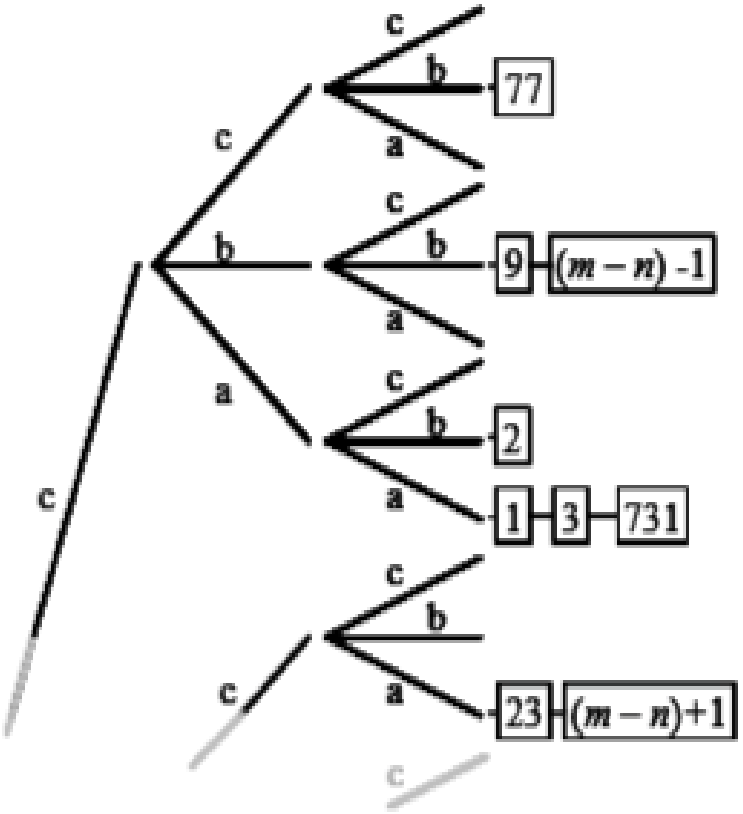


PAA, SAX

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

Префиксное дерево



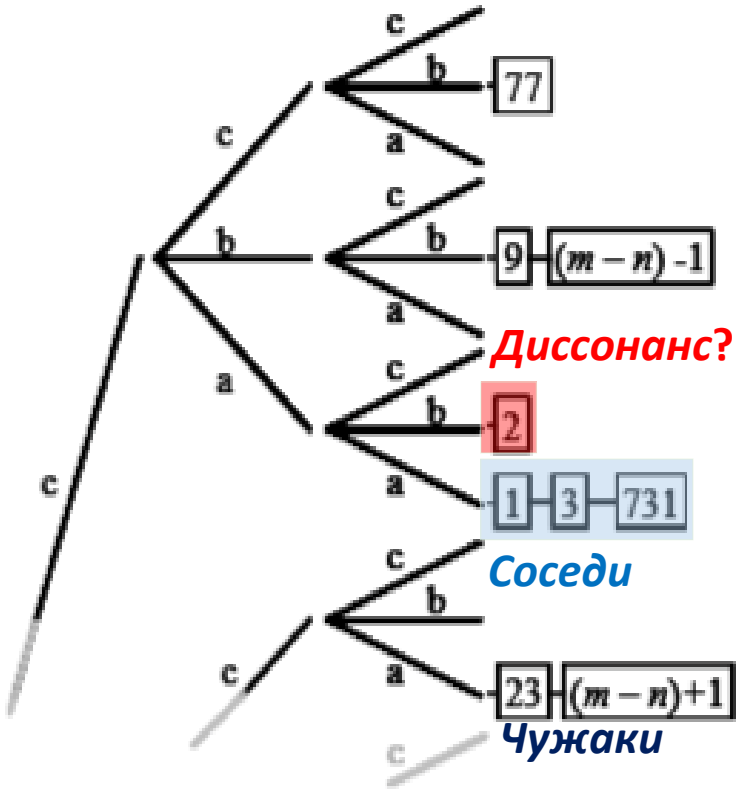
- Каждое ребро дерева помечено символом алфавита. Ребра, соединяющие узел с его сыновьями, помечены разными символами
- SAX-код – конкатенация пометок ребер на пути от корня до листа
- В листе – упорядоченный по возрастанию список индексов подпоследовательностей исходного ряда с соответствующим кодом

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 HOTSAX

$dist_{bsf} \leftarrow 0$; $dist_{min} \leftarrow \infty$

for $C_i \in$ ***Диссонансы?*** • ***Остальные***

  for $C_j \in$ ***Соседи*** • ***Чужаки***

    $d \leftarrow \mathrm{Dist}(C_i, C_j)$

    if $d < dist_{bsf}$

      break

    $dist_{min} \leftarrow \min(d, dist_{min})$

  $dist_{bsf} \leftarrow \max(dist_{min}, dist_{bsf})$

  $pos_{bsf} \leftarrow i$

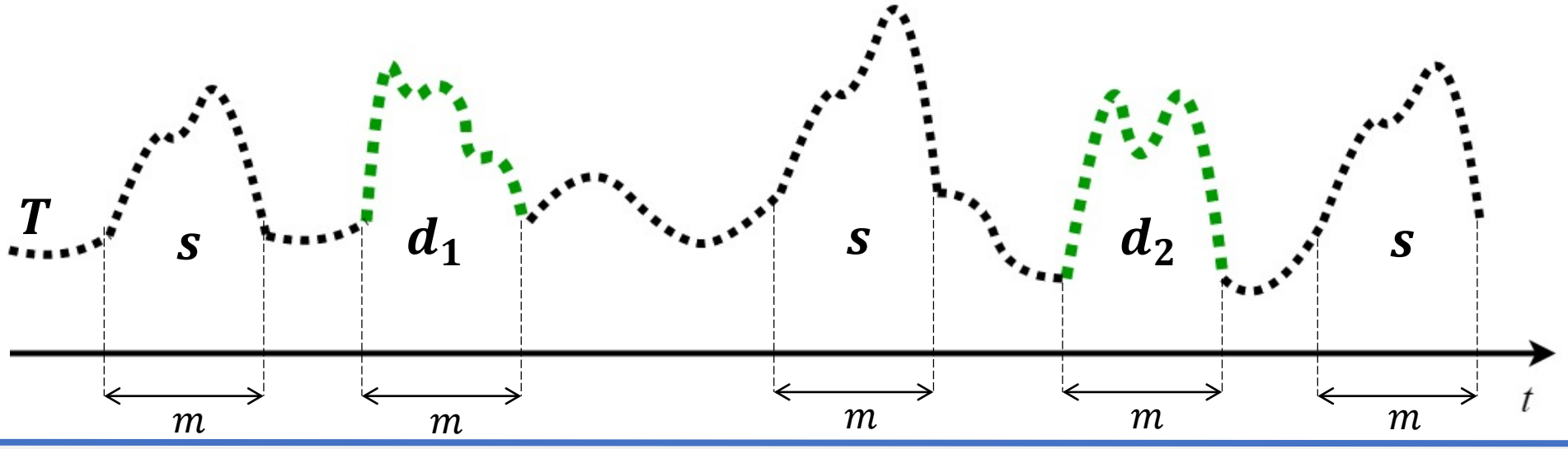
return $\{pos_{bsf}, dist_{bsf}\}$

Содержание

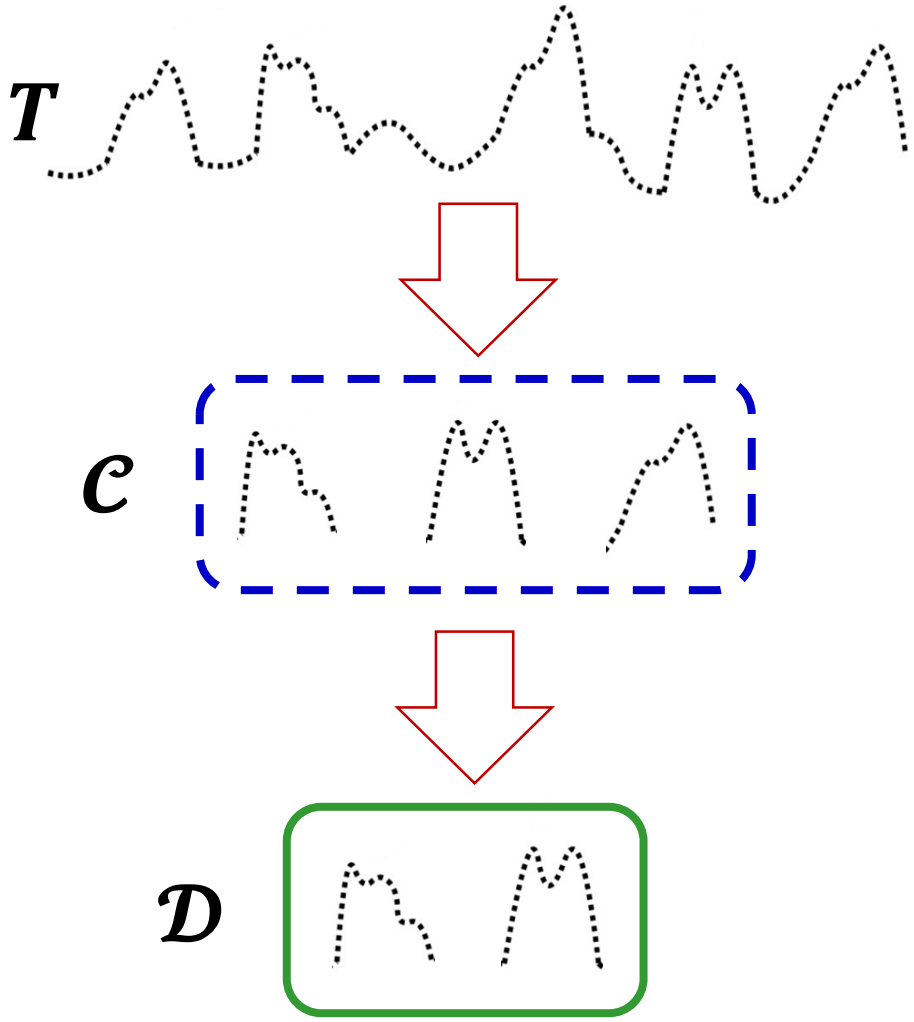
- Понятия аномалии и диссонанса
- Алгоритм HOTSAX
- **Алгоритм DRAG**
- Алгоритм MERLIN

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: $\mathcal{D} = \{d_1, d_2, \dots\}\ d_i \in \mathcal{D} \Leftrightarrow \forall\, s \in T \min_{s \cap d_i = \emptyset} \mathrm{Dist}(d_i, s) \geq r$



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

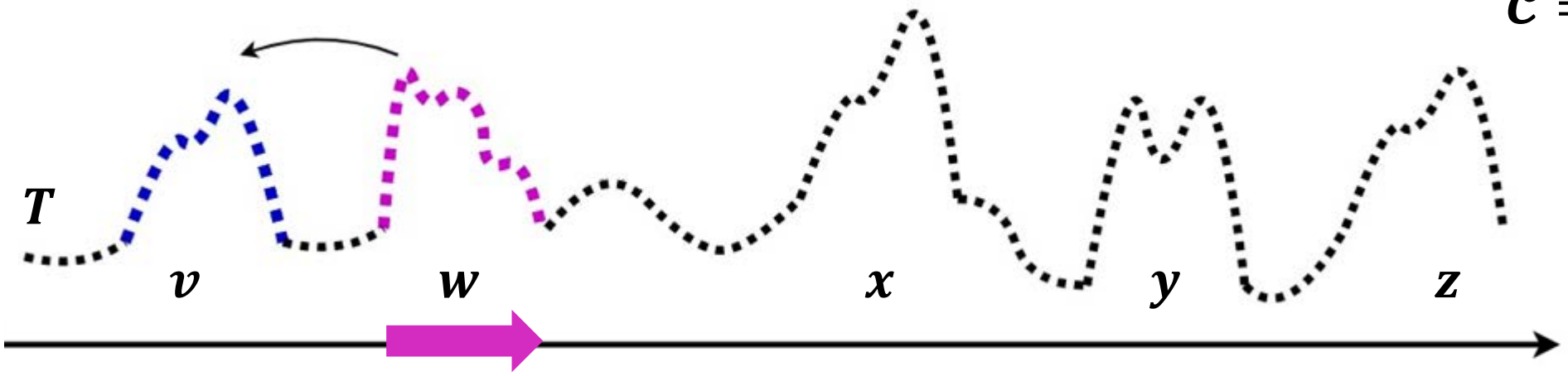
DRAG: Отбор кандидатов

пока не конец ряда T:
    текущая подпоследовательность s
    Кандидат := TRUE
    для всех $c_i \in \mathcal{C}$ и $s \cap c_i = \emptyset$
        если $\text{Dist}(s, c_i) < r$ то
            $\mathcal{C} := \mathcal{C} \setminus c_i$; Кандидат := FALSE
    если Кандидат = TRUE то $\mathcal{C} := \mathcal{C} \cup s$

$\mathcal{C} = \{v\}$

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

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



DRAG: Отбор кандидатов

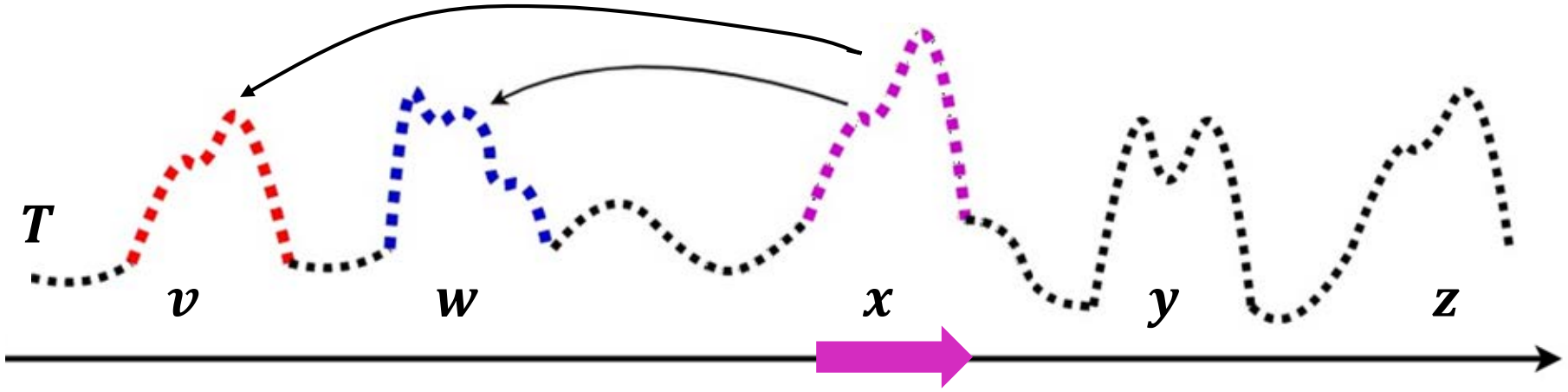
пока не конец ряда T:
  текущая подпоследовательность s
  Кандидат := TRUE
  для всех $c_i \in \mathcal{C}$ и $s \cap c_i = \emptyset$
    если $\text{Dist}(s, c_i) < r$ то
      $\mathcal{C} := \mathcal{C} \setminus c_i$; Кандидат := FALSE
  если Кандидат = TRUE то $\mathcal{C} := \mathcal{C} \cup s$

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

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

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

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



DRAG: Отбор кандидатов

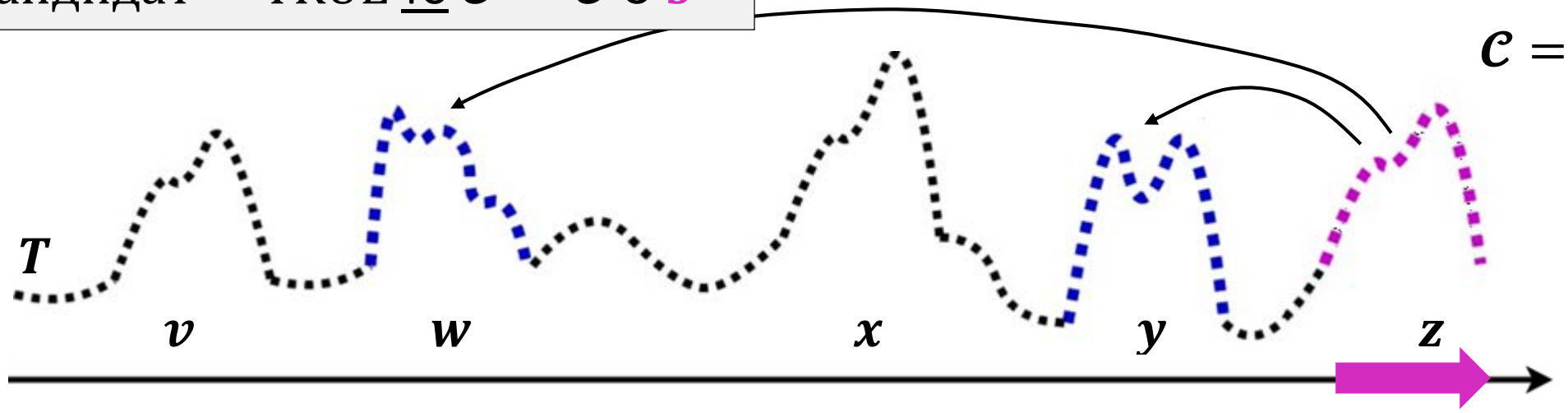
пока не конец ряда T:
    текущая подпоследовательность s
    Кандидат := TRUE
    для всех $c_i \in \mathcal{C}$ и $s \cap c_i = \emptyset$
        если $\text{Dist}(s, c_i) < r$ то
            $\mathcal{C} := \mathcal{C} \setminus c_i$; Кандидат := FALSE
    если Кандидат = TRUE то $\mathcal{C} := \mathcal{C} \cup s$

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

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

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

$\mathcal{C} = \{w, y, z\}$



DRAG: Очистка кандидатов

$\mathcal{D} := \mathcal{C}$

пока не конец ряда T:

текущая подпоследовательность s

для всех $d_i \in \mathcal{D}$ и $s \cap d_i = \emptyset$

если $\text{Dist}(s, d_i) < r$ то

$\mathcal{D} := \mathcal{D} \setminus d_i$

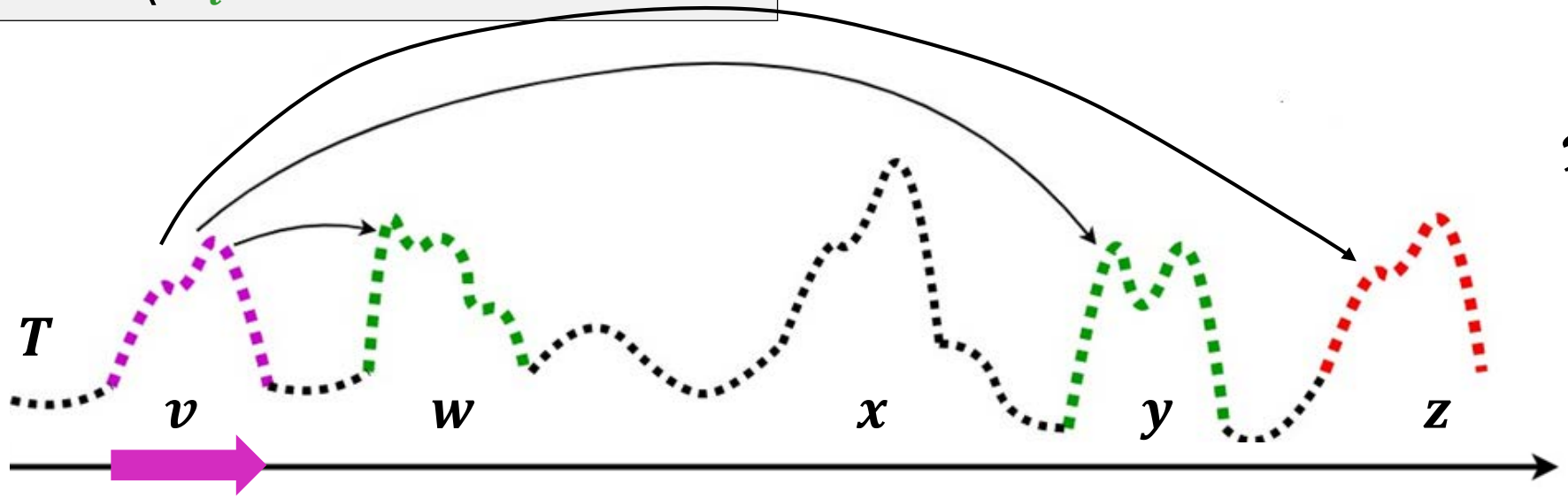
$\mathcal{D} = \{w, y, z\}$

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

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

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

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



DRAG: Очистка кандидатов

$\mathcal{D} := \mathcal{C}$

пока не конец ряда T:

текущая подпоследовательность s

для всех $d_i \in \mathcal{D}$ и $s \cap d_i = \emptyset$

если $\text{Dist}(s, d_i) < r$ то

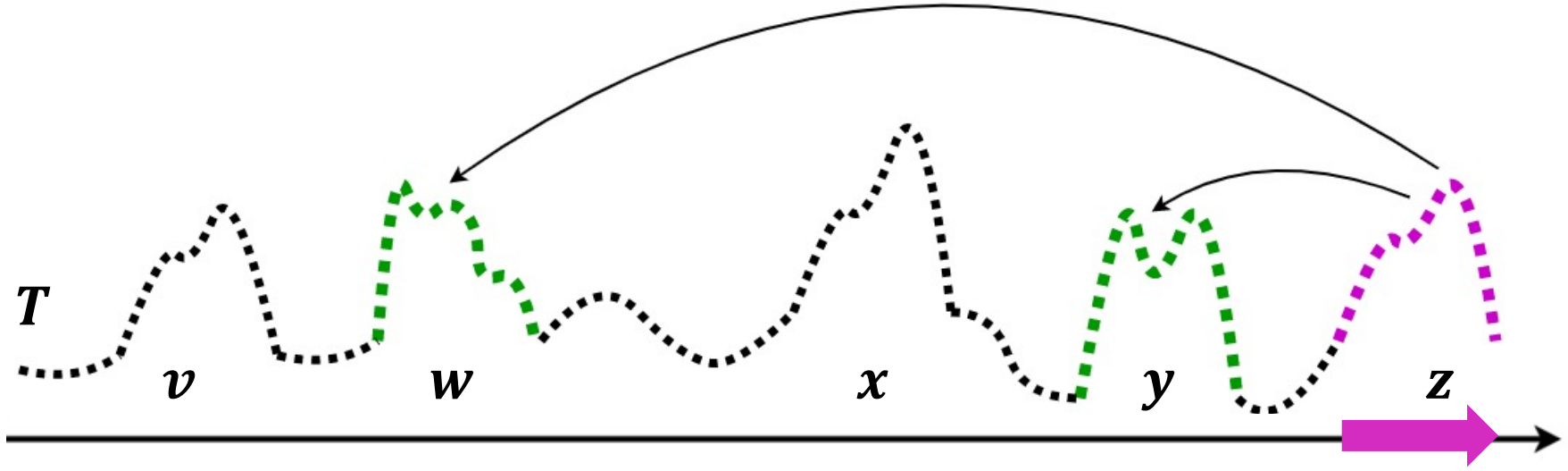
$\mathcal{D} := \mathcal{D} \setminus d_i$

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

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

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

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



Эвристический подбор параметра r

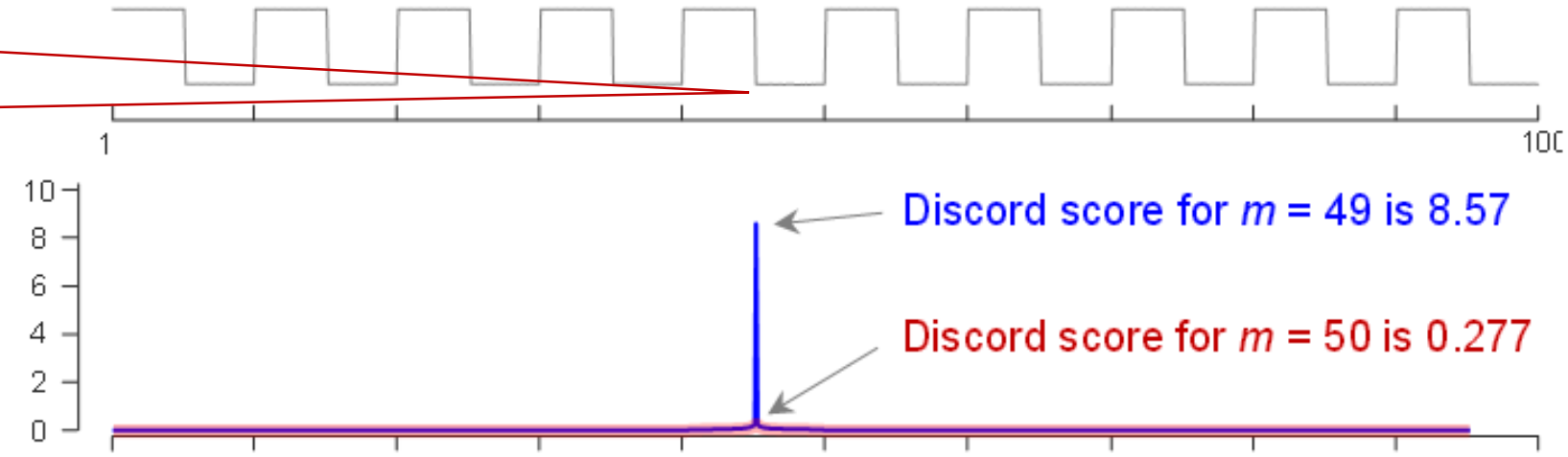
1. Выбрать случайный сегмент ряда максимальной длины, который может быть размещен в памяти
2. Найти в выбранном сегменте диссонанс с помощью алгоритма HOTSAX
3. Взять в качестве порога r расстояние от найденного диссонанса до его ближайшего соседа

Проблемы DRAG

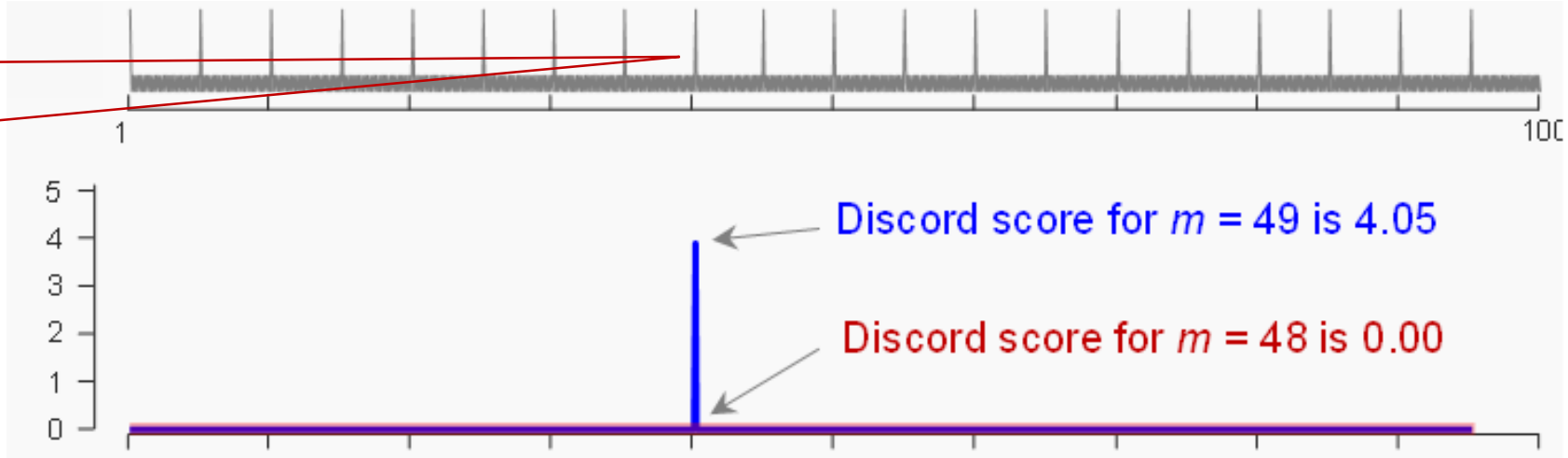
1. Ручной подбор длины диссонанса m
   - Не всегда заранее известна длина аномалии
   - Запуск DRAG для всех возможных длин вычислительно неосуществим
2. Ручной подбор порога r
   - Слишком большой порог – нет диссонансов, слишком маленький порог – много ложных диссонансов
3. Алгоритм последовательный

Чтобы найти все аномалии, нужно проверить все значения m

Мы можем корректно найти слабо различимую аномалию для $m = 49$, но не для $m + 1$

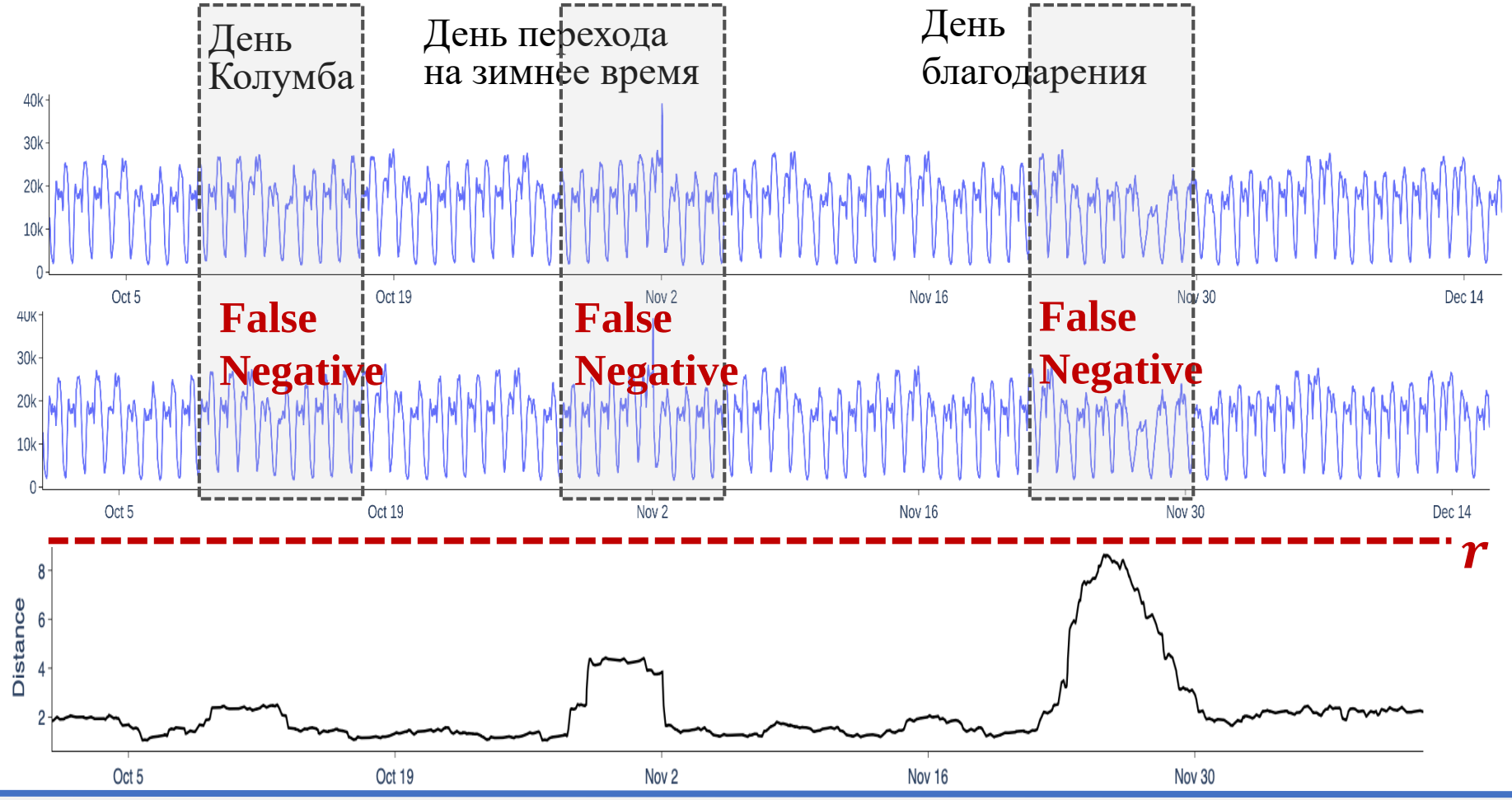


Мы можем корректно найти слабо различимую аномалию для $m = 49$, но не для $m - 1$



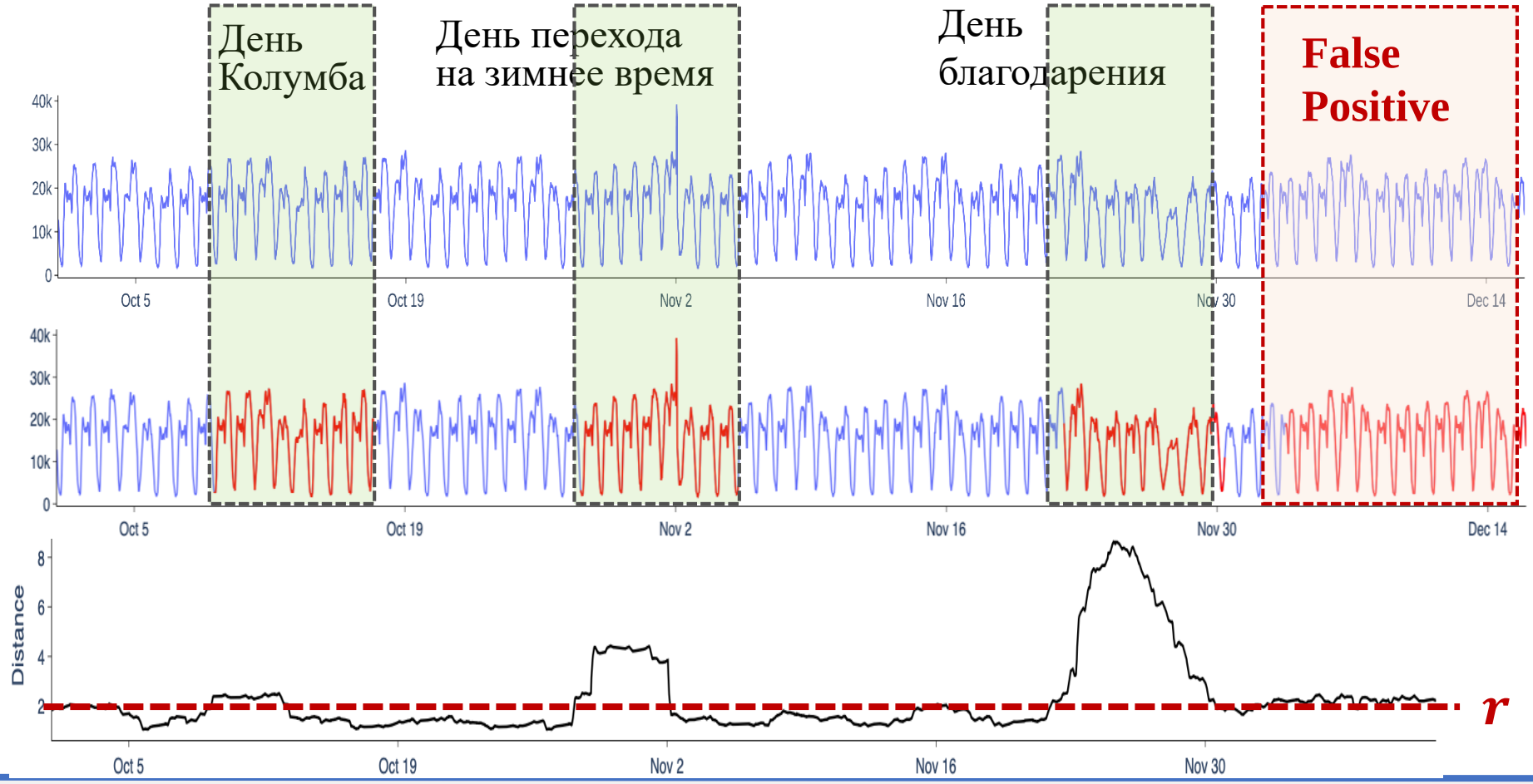
Ручной подбор порога: $r \to +\infty \Rightarrow$ нет диссонансов

Среднее число пассажиров NY такси осенью 2014 г.

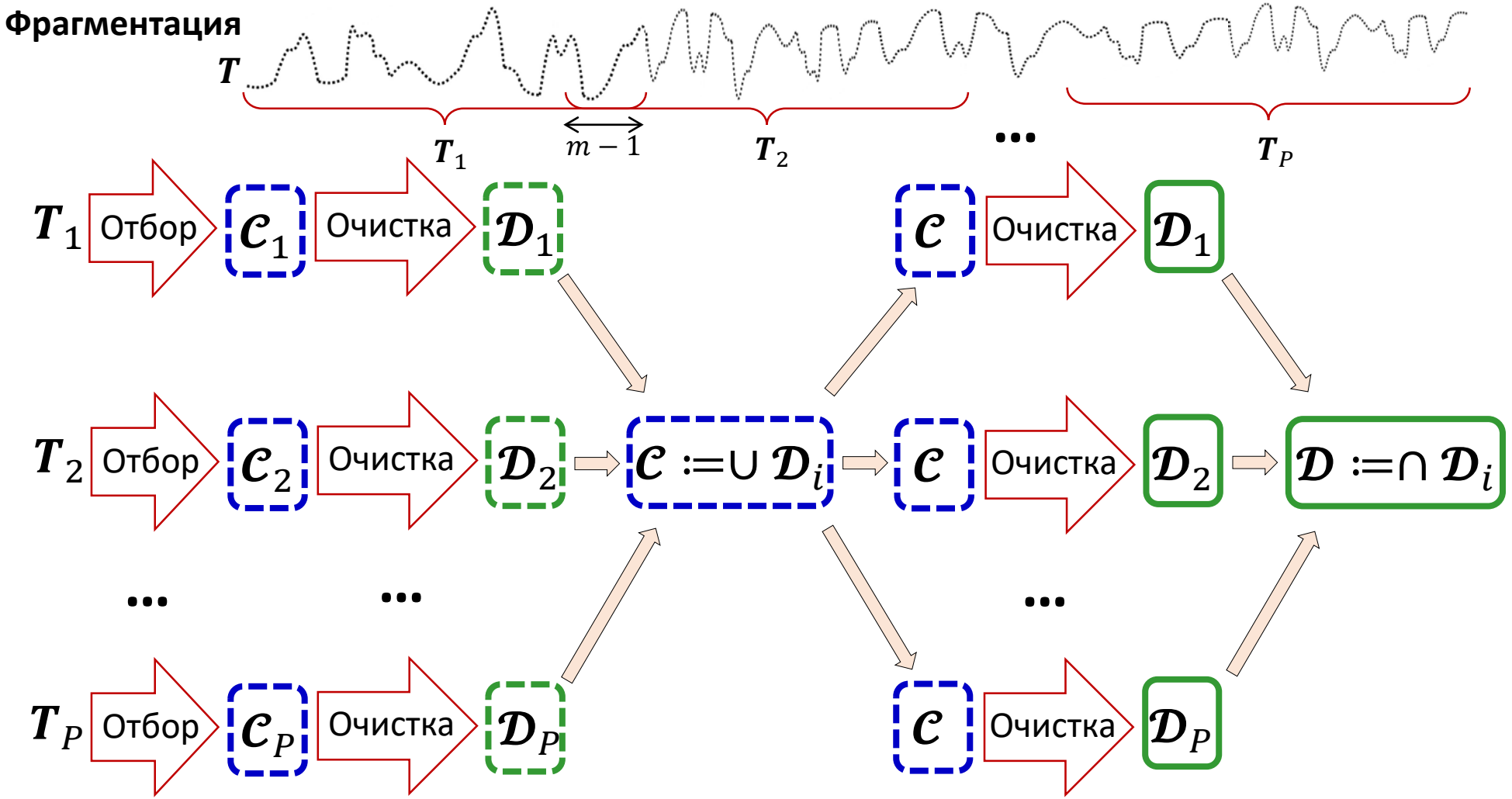


Ручной подбор порога: $r \to 0 \Rightarrow$ ложные аномалии

Среднее число пассажиров NY такси осенью 2014 г.



Распараллеливание DRAG



Содержание

- Понятия аномалии и диссонанса
- Алгоритм HOTSAX
- Алгоритм DRAG
- **Алгоритм MERLIN**

Алгоритм MERLIN

Algorithm MERLIN (IN T, $minL$, $maxL$, $topK$; OUT $\mathcal{D}$)

1: $\mathcal{D} \leftarrow \varnothing$; $r \leftarrow 2\sqrt{minL}$; $nnDist_{minL} \leftarrow -\infty$
2: **while** $nnDist_{minL} < 0$ **and** $|D_{minL}| < topK$ **do**
3: $\quad D_{minL} \leftarrow \text{DRAG}(T, minL, r)$; $\mathcal{D} \leftarrow \mathcal{D} \cup D_{minL}$; $nnDist_{minL} \leftarrow \min_{d \in D_{minL}} d.nnDist$
4: $\quad r \leftarrow 0.5 \cdot r$
5: **for** $i \leftarrow minL + 1$ **to** $minL + 4$ **do**
6: $\quad nnDist_i \leftarrow -\infty$
7: $\quad$ **while** $nnDist_i < 0$ **and** $|D_i| < topK$ **do**
8: $\quad\quad r \leftarrow 0.99 \cdot nnDist_{i-1}$
9: $\quad\quad D_i \leftarrow \text{DRAG}(T, i, r)$; $\mathcal{D} \leftarrow \mathcal{D} \cup D_i$; $nnDist_i \leftarrow \min_{d \in D_i} d.nnDist$
10: $\quad\quad r \leftarrow 0.99 \cdot r$
11: **for** $i \leftarrow minL + 5$ **to** $maxL$ **do**
12: $\quad \mu \leftarrow \text{Mean}(\{nnDist_k\}_{k=i-1}^{i-5})$; $\sigma \leftarrow \text{Std}(\{nnDist_k\}_{k=i-1}^{i-5})$; $r \leftarrow \mu - 2\sigma$
13: $\quad D_i \leftarrow \text{DRAG}(T, i, r)$; $\mathcal{D} \leftarrow \mathcal{D} \cup D_i$; $nnDist_i \leftarrow \min_{d \in D_i} d.nnDist$
14: $\quad$ **while** $nnDist_i < 0$ **and** $|D_i| < topK$ **do**
15: $\quad\quad D_i \leftarrow \text{DRAG}(T, i, r)$; $\mathcal{D} \leftarrow \mathcal{D} \cup D_i$; $nnDist_i \leftarrow \min_{d \in D_i} d.nnDist$
16: $\quad\quad r \leftarrow r - \sigma$
17: **return** $\mathcal{D}$

Шаг 1. Поиск диссонансов минимальной длины $minL$

$$r = 2\sqrt{minL}$$

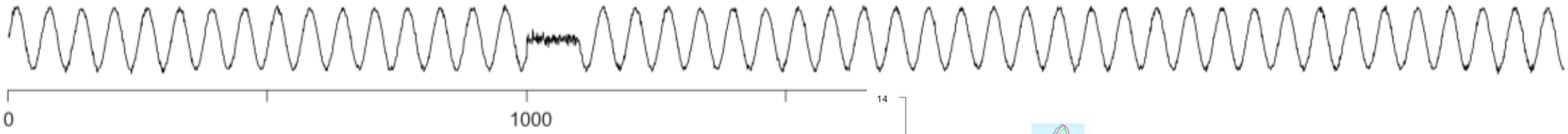
Шаг 2. Поиск диссонансов следующих четырех длин

$$r = 0.99 \cdot nndist_{m-1}$$

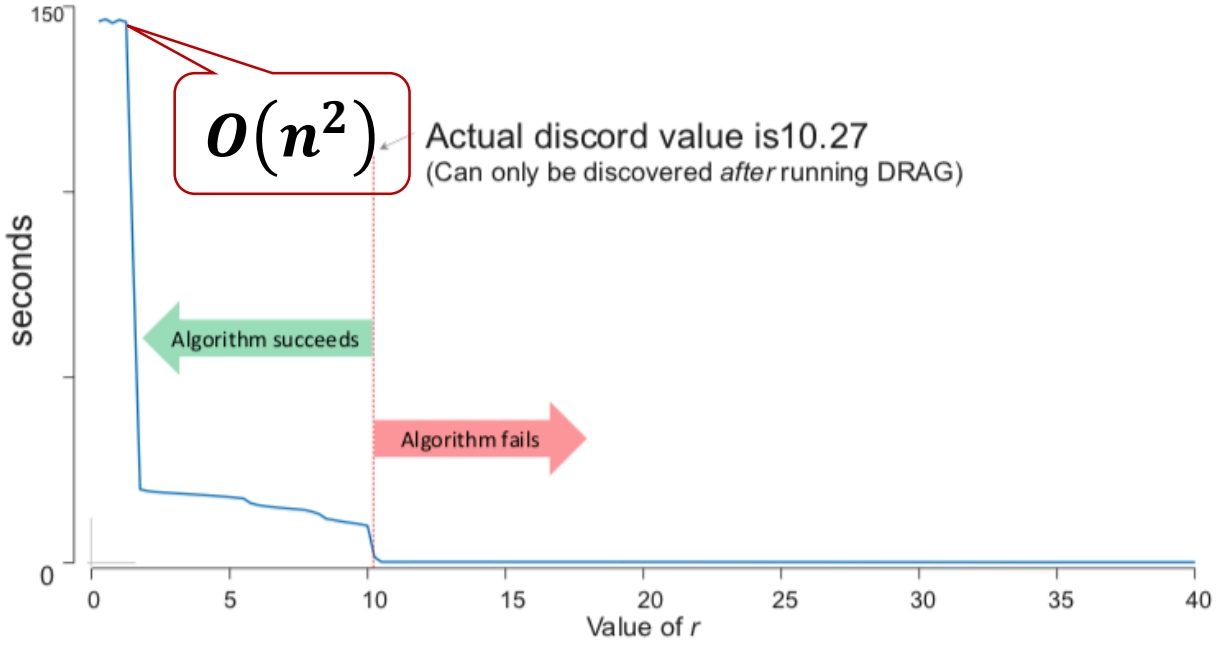
Шаг 3. Поиск диссонансов всех оставшихся длин

$$r = \mu - 2\sigma$$

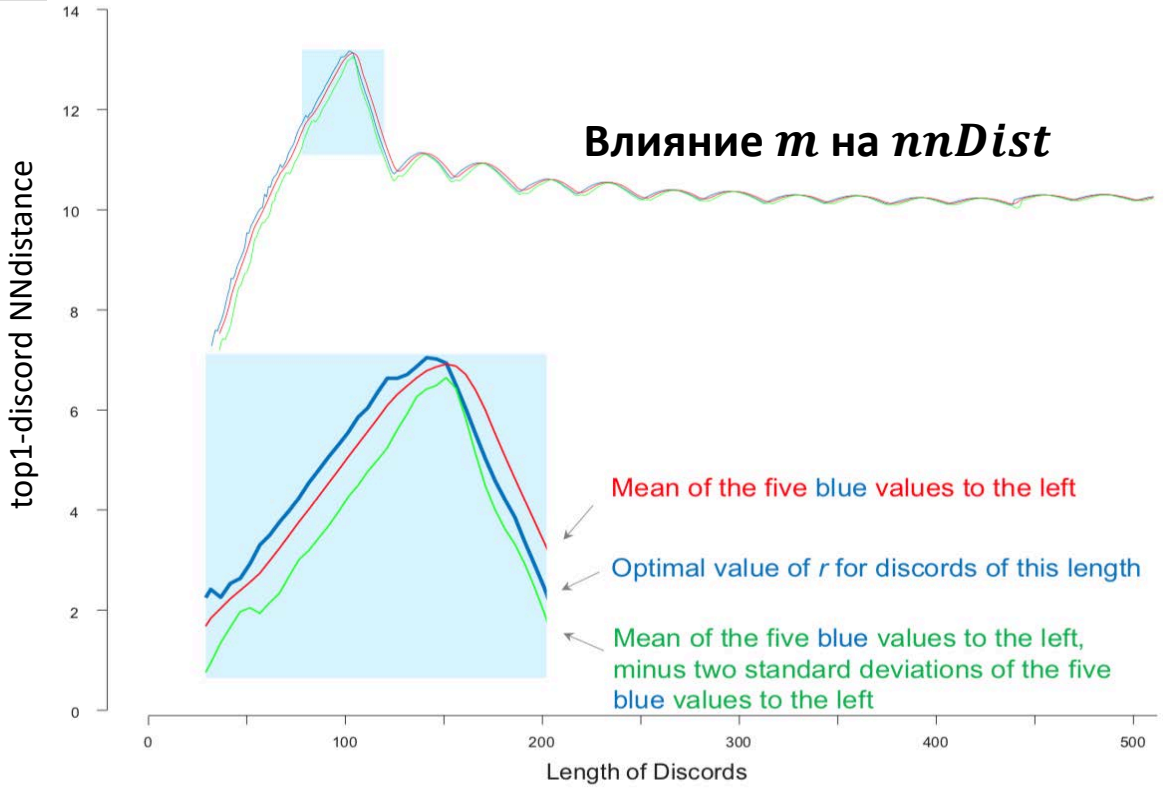
MERLIN: Подбор порога r



Влияние r на время выполнения DRAG



Влияние m на $nnDist$



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