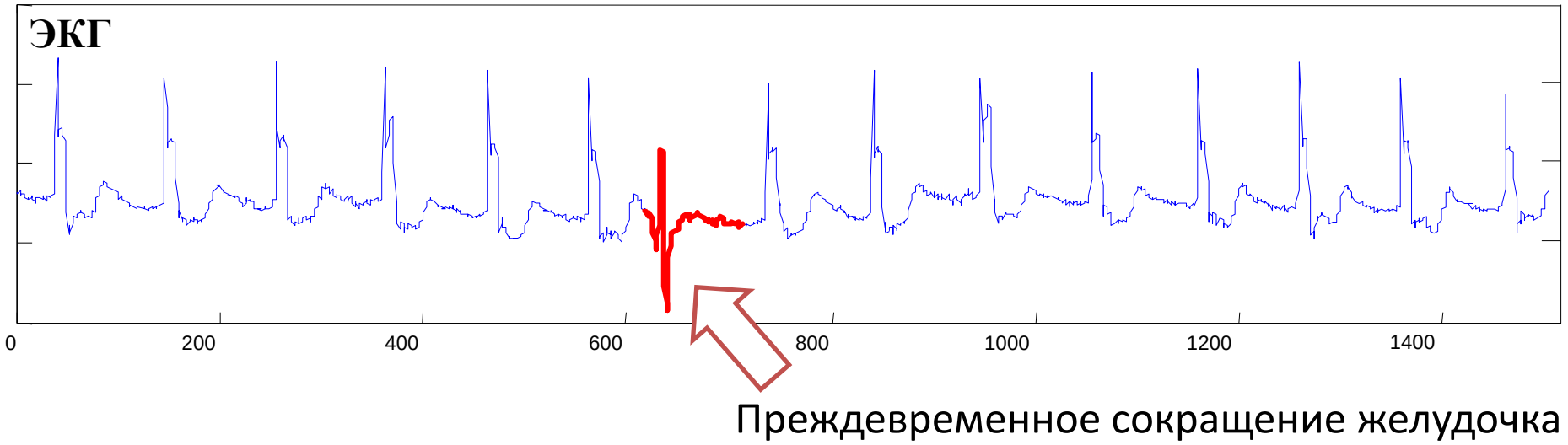
Международная научная конференция
Параллельные вычислительные технологии (ПаВТ'2020)
Пермь, 31 марта – 2 апреля 2020 г.

Поиск аномалий в сверхбольших временных рядах на высокопроизводительном кластере с многоядерными ускорителями

М.Л. Цымблер, А.В. Гренц, Я.А. Краева, С. Кумар
Южно-Уральский государственный университет (Челябинск)

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 20-07-00140) и Министерства образования и науки РФ (гос. задание FENU-2020-0022).

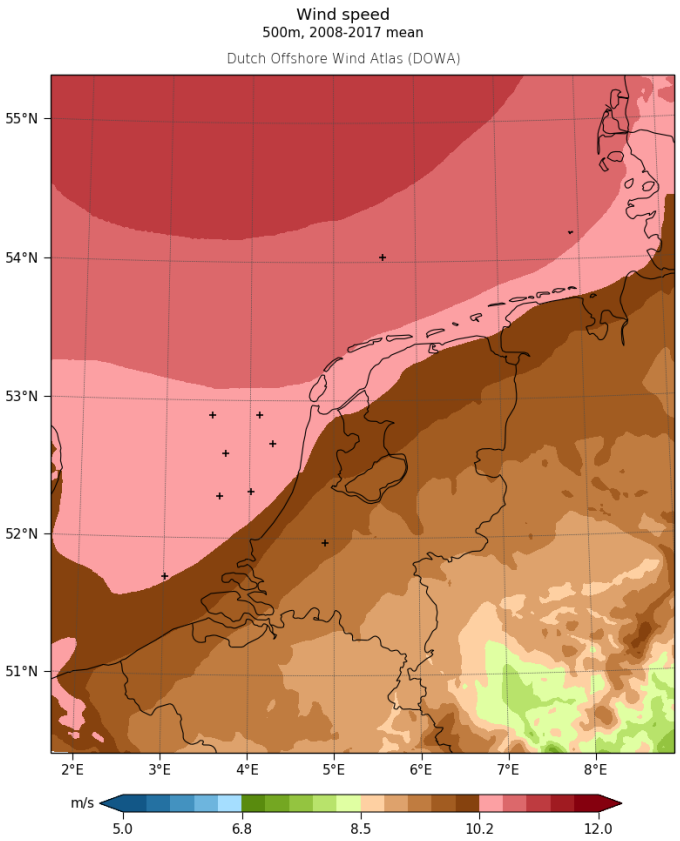
Аномалии во временных рядах



Преждевременное сокращение желудочка

- Поиск аномалий в данных мониторинга ЭКГ **за одни сутки** позволяет выявить зарождение инфаркта
- Длина ряда: **20 млн.**
- Размер данных: **200 Мб**

Аномалии в сверхбольших временных рядах



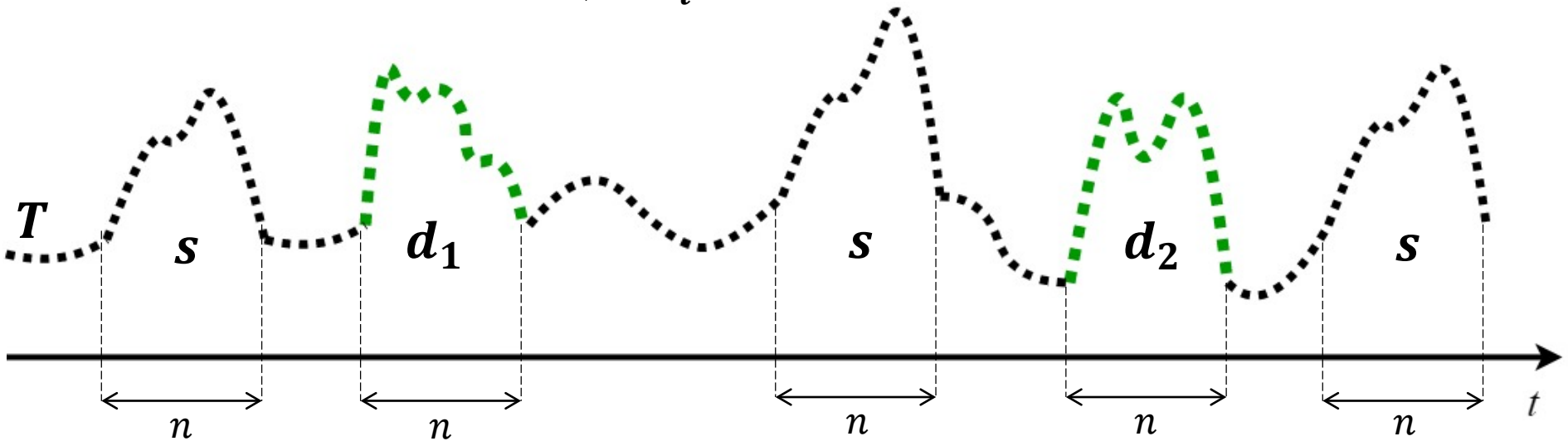
- Поиск аномалий в данных атласа морских ветров **за 10 лет**[1)] позволяет найти локации для установки прибрежных ветряных электростанций, дающие наибольшую производительность
- Длина ряда: **15 млрд.**
- Размер данных: **2.8 Тб**

[1)] De Valk C., Wijnant I.L. Uncertainty analysis of climatological parameters of the Dutch Offshore Wind Atlas (DOWA). 2019. Technical report TR-379. Royal Netherlands Meteorological Institute

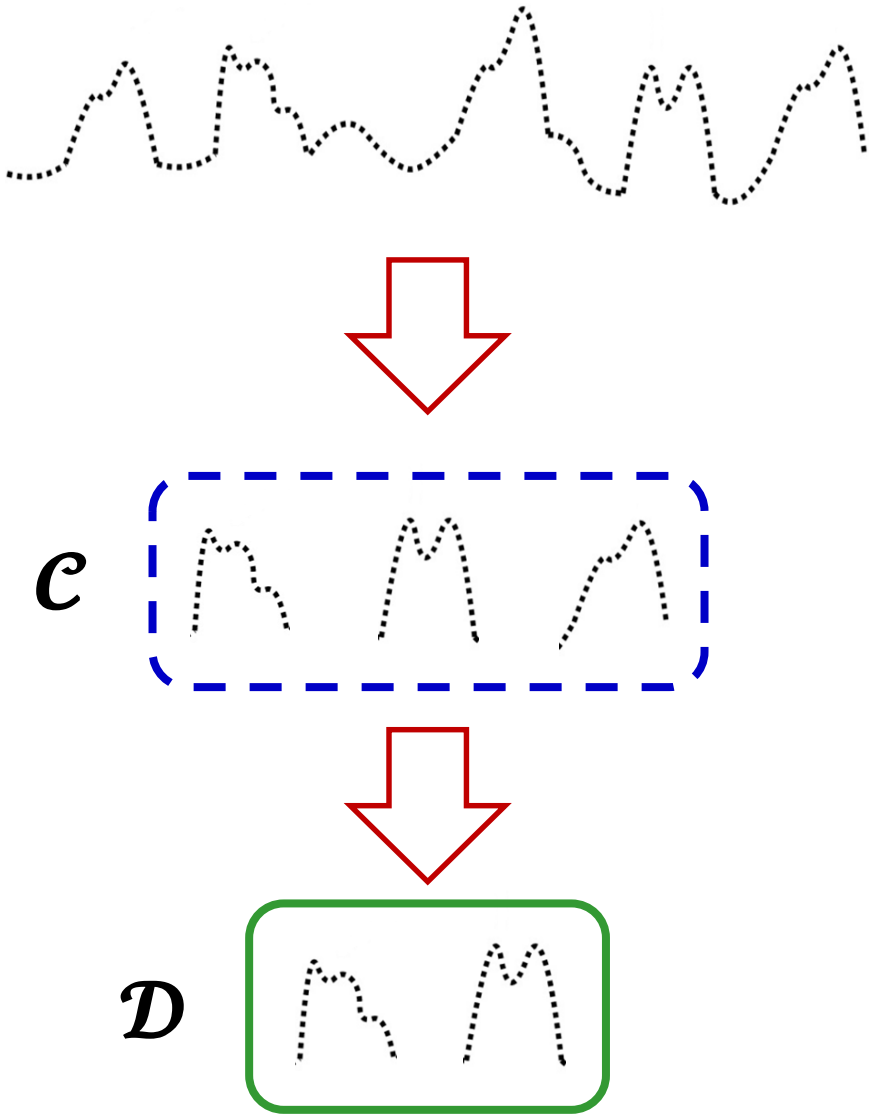
Постановка задачи

- **Диссонанс** – подпоследовательность ряда, расстояние от которой до наиболее похожей на нее подпоследовательности не ниже заданного порога
- Дано: ряд $\boldsymbol{T}$, длина диссонанса n, порог r
- Найти: $\boldsymbol{\mathcal{D}} = \{\boldsymbol{d_1}, \boldsymbol{d_2}, \dots\}$

$$\boldsymbol{d_i} \in \boldsymbol{\mathcal{D}} \Leftrightarrow \min_{\boldsymbol{s} \in \boldsymbol{T},\, \boldsymbol{s} \cap \boldsymbol{d_i} = \emptyset} \mathrm{EuclidDist}(\boldsymbol{d_i}, \boldsymbol{s}) \geq r$$



Последовательный поиск диссонансов



1. **Отбор**

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

2. **Очистка**

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

Отбор кандидатов

Сканировать ряд T:

текущая подпоследовательность s

Кандидат := TRUE

для всех $c_i \in \mathcal{C}$

если $\mathrm{ED}(s, c_i) < r$ **and** $s \cap c_i = \emptyset$ то

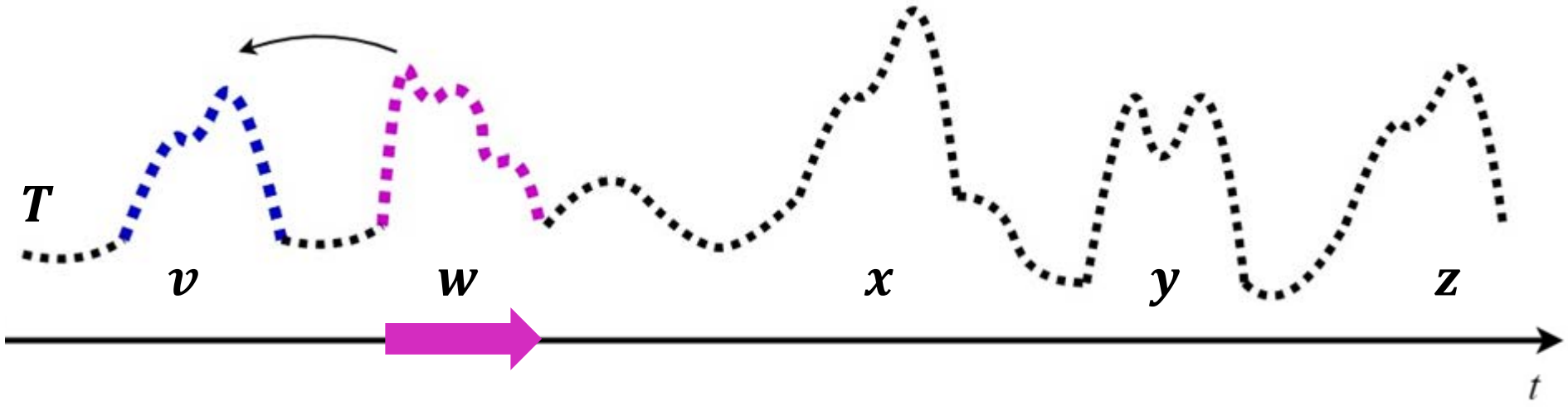
$\mathcal{C} := \mathcal{C} \setminus c_i$; Кандидат := FALSE

если Кандидат = TRUE то $\mathcal{C} := \mathcal{C} \cup s$

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

$\mathrm{ED}(w, v) \geq r$

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



Отбор кандидатов

Сканировать ряд T:

текущая подпоследовательность s

Кандидат := TRUE

для всех $c_i \in \mathcal{C}$

если $\mathrm{ED}(s, c_i) < r$ **and** $s \cap c_i = \emptyset$ то

$\mathcal{C} := \mathcal{C} \setminus c_i$; Кандидат := FALSE

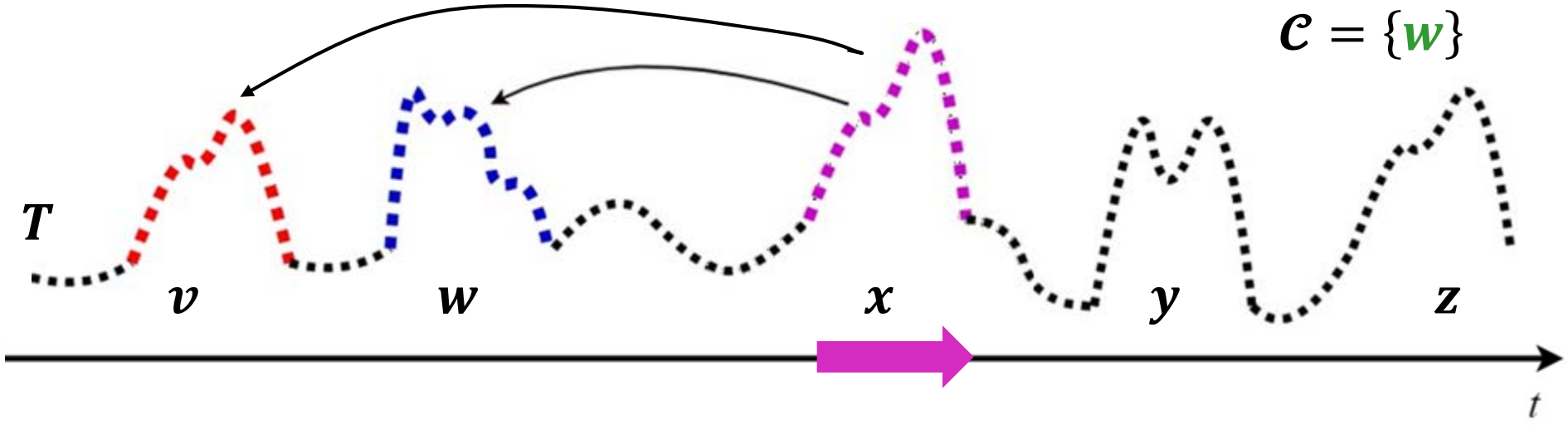
если Кандидат = TRUE то $\mathcal{C} := \mathcal{C} \cup s$

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

$\mathrm{ED}(x, v) < r$

$\mathrm{ED}(x, w) \geq r$

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



Отбор кандидатов

Сканировать ряд T:

текущая подпоследовательность s

Кандидат := TRUE

для всех $c_i \in \mathcal{C}$

если $\mathrm{ED}(s, c_i) < r$ **and** $s \cap c_i = \emptyset$ то

$\mathcal{C} := \mathcal{C} \setminus c_i$; Кандидат := FALSE

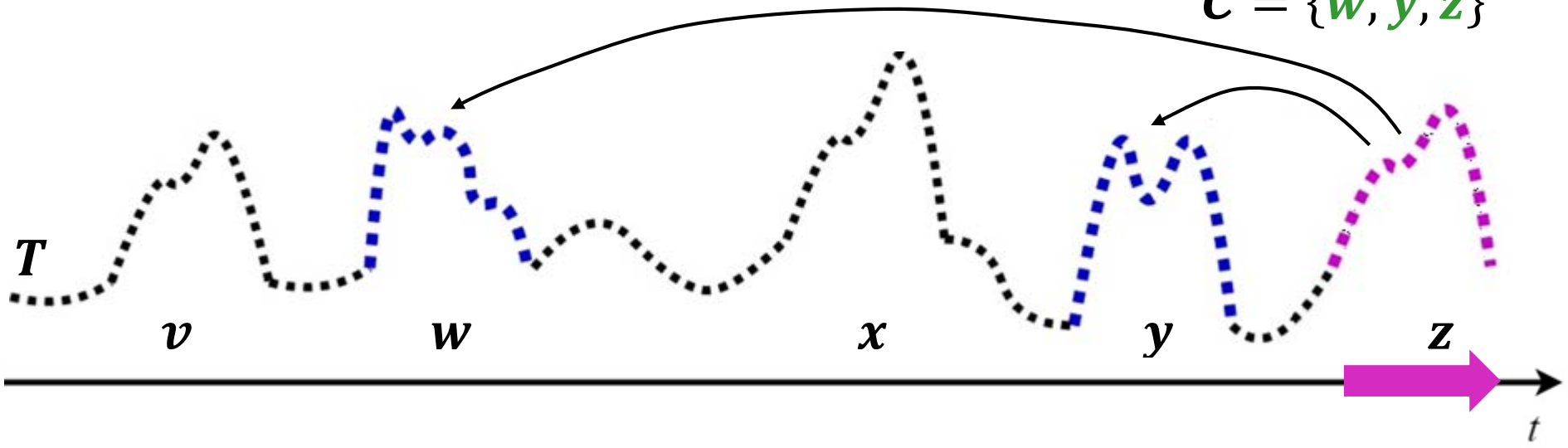
если Кандидат = TRUE то $\mathcal{C} := \mathcal{C} \cup s$

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

$\mathrm{ED}(z, w) \geq r$

$\mathrm{ED}(z, y) \geq r$

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



Очистка кандидатов

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

Сканировать ряд T:

текущая подпоследовательность s

для всех $d_i \in \mathcal{D}$

если $\mathrm{ED}(s, d_i) < r$ **and** $s \cap d_i = \emptyset$ то

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

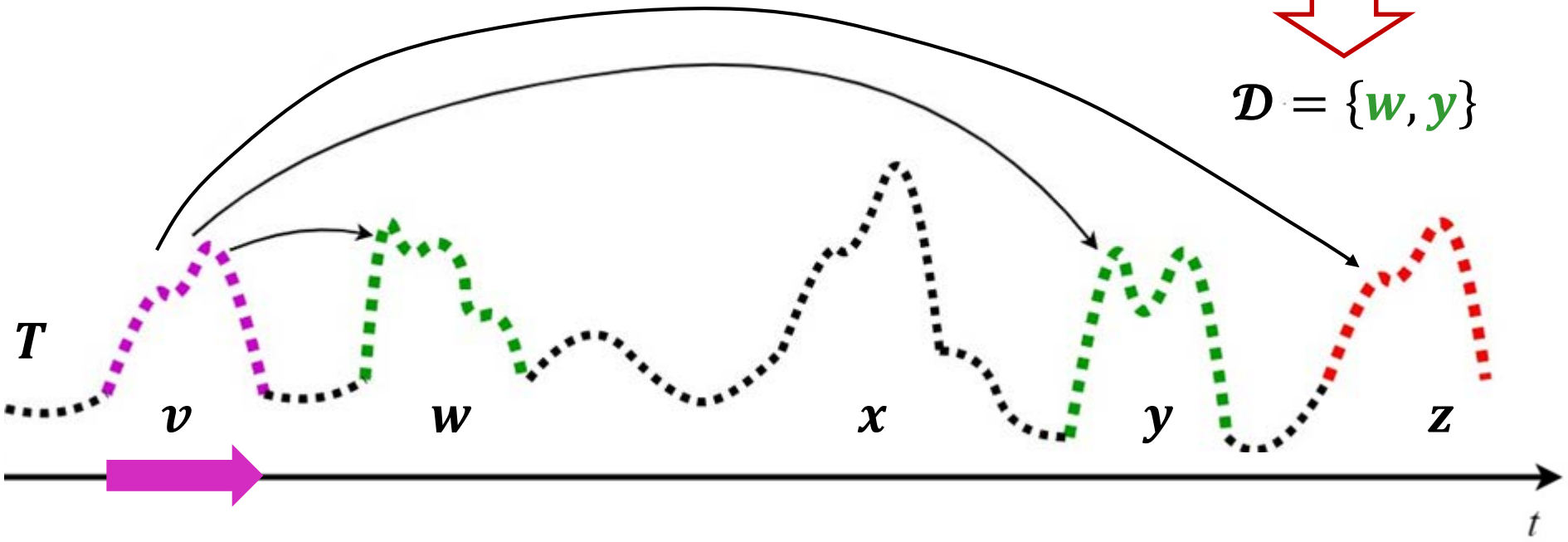
$\mathcal{D} = \{w, y, z\}$

$\mathrm{ED}(v, w) \geq r$

$\mathrm{ED}(v, y) \geq r$

$\mathrm{ED}(v, z) < r$

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



Очистка кандидатов

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

Сканировать ряд T:

текущая подпоследовательность s

для всех $d_i \in \mathcal{D}$

если $\mathrm{ED}(s, d_i) < r$ **and** $s \cap d_i = \emptyset$ то

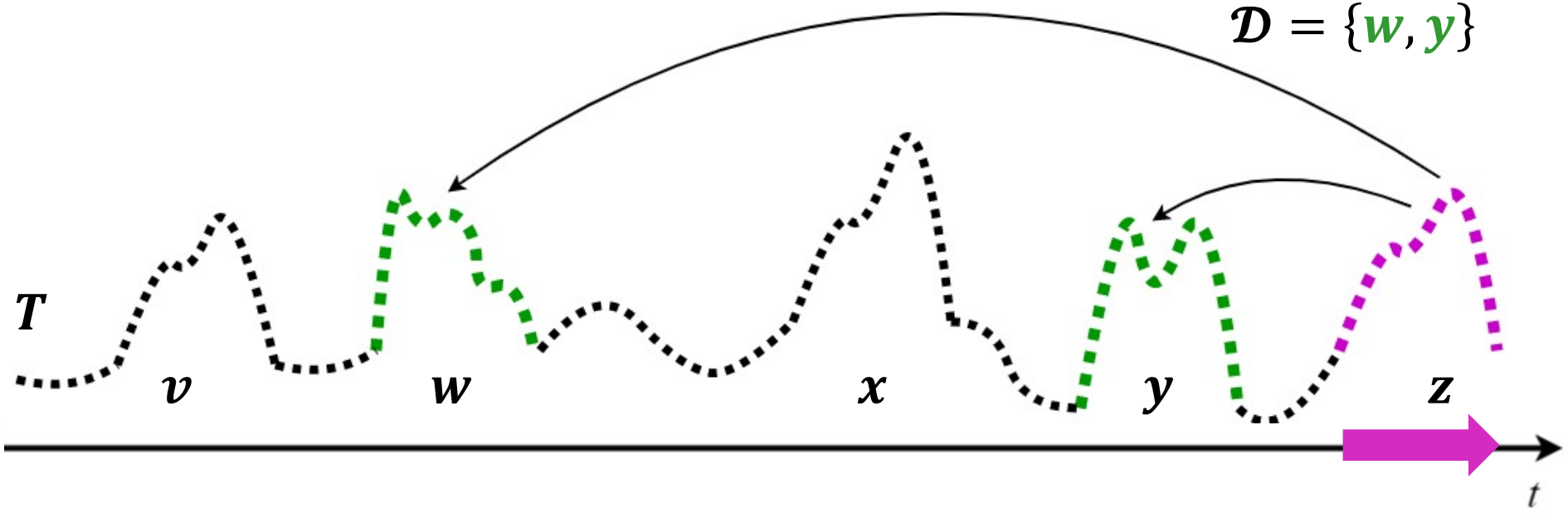
$\mathcal{D} := \mathcal{D} \setminus d_i$

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

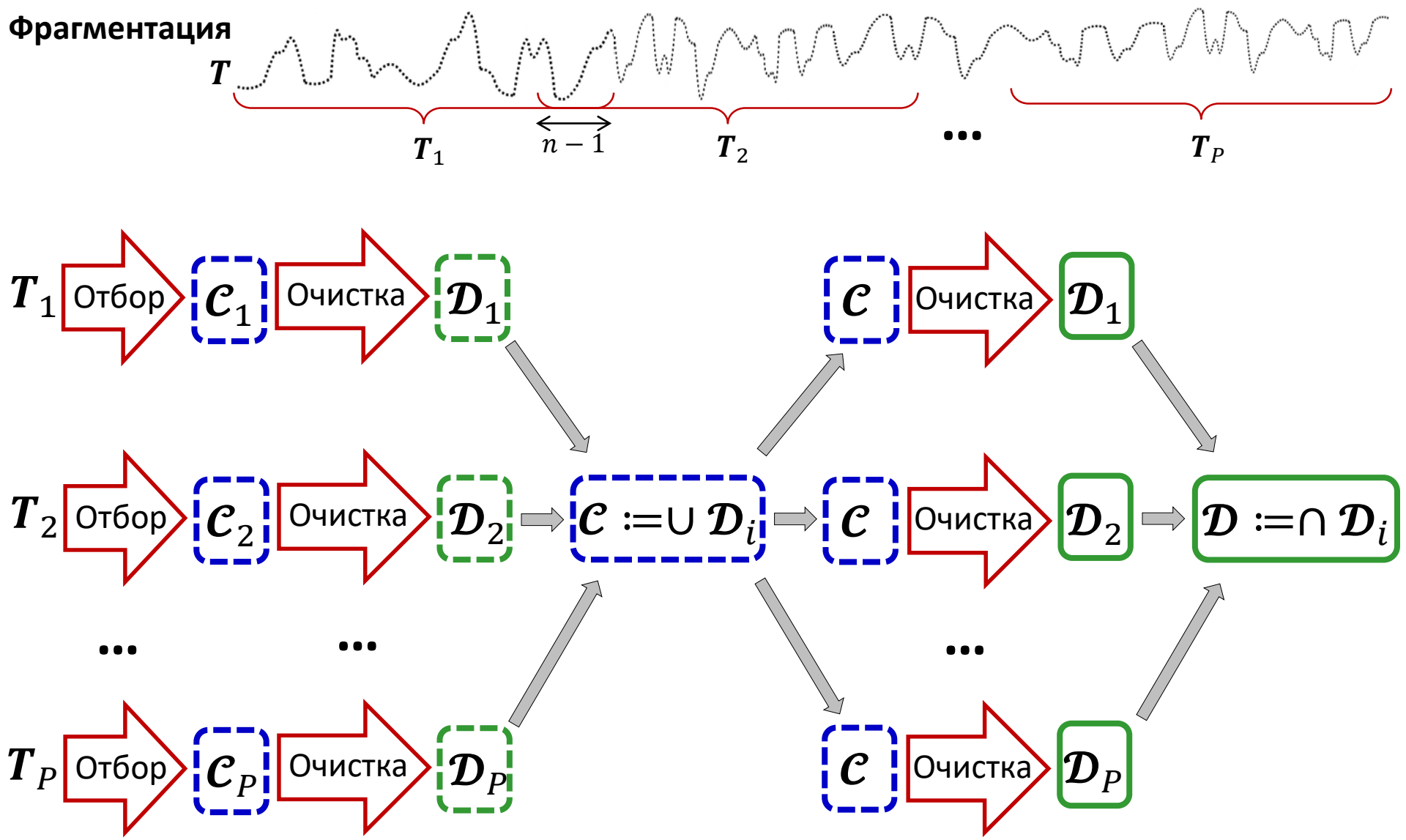
$\mathrm{ED}(z, w) \geq r$

$\mathrm{ED}(z, y) \geq r$

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



Распределенный поиск диссонансов (кластер)

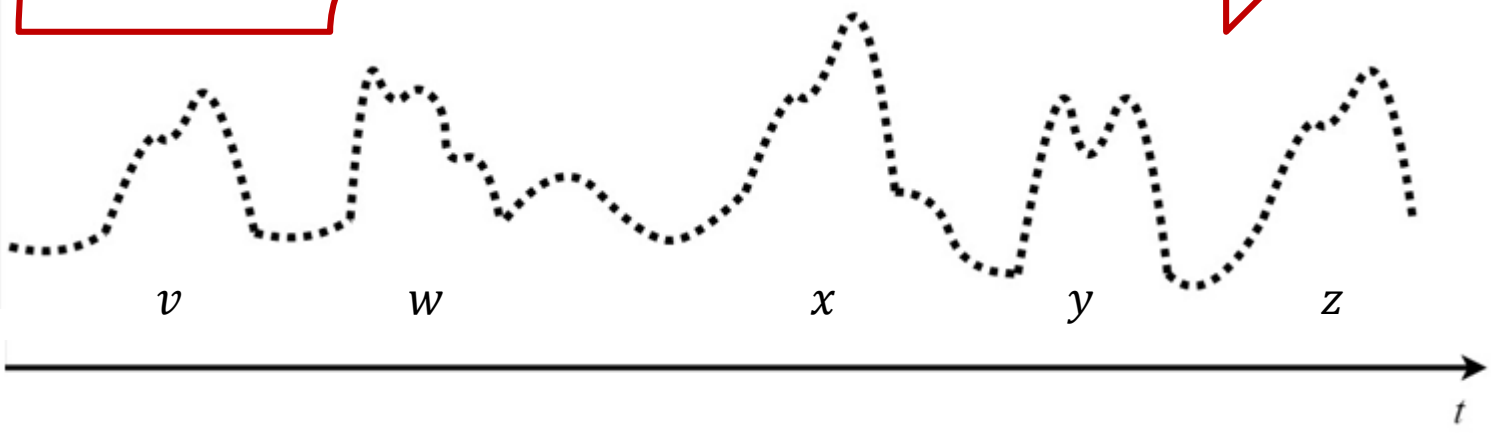


Параллельный поиск диссонансов (ускоритель)

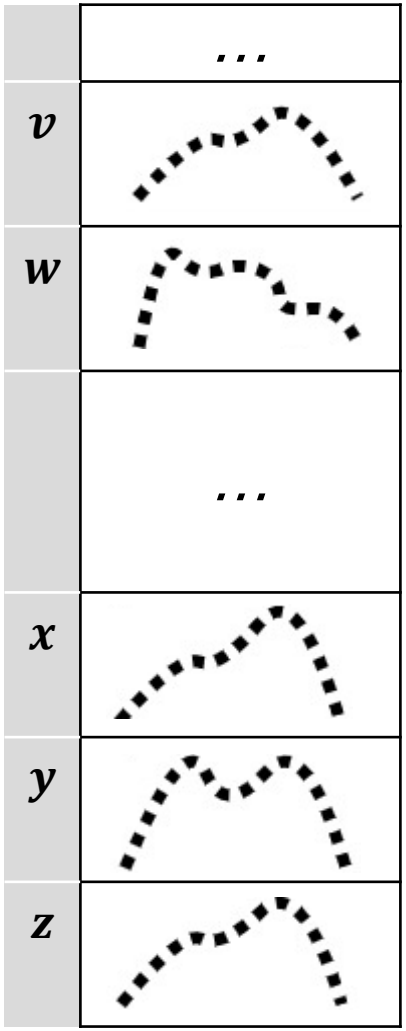
Нормализация

$$\hat{S} = (\hat{s}_1, \dots, \hat{s}_n), \qquad \hat{s}_i = \frac{s_i - \mu}{\sigma};$$

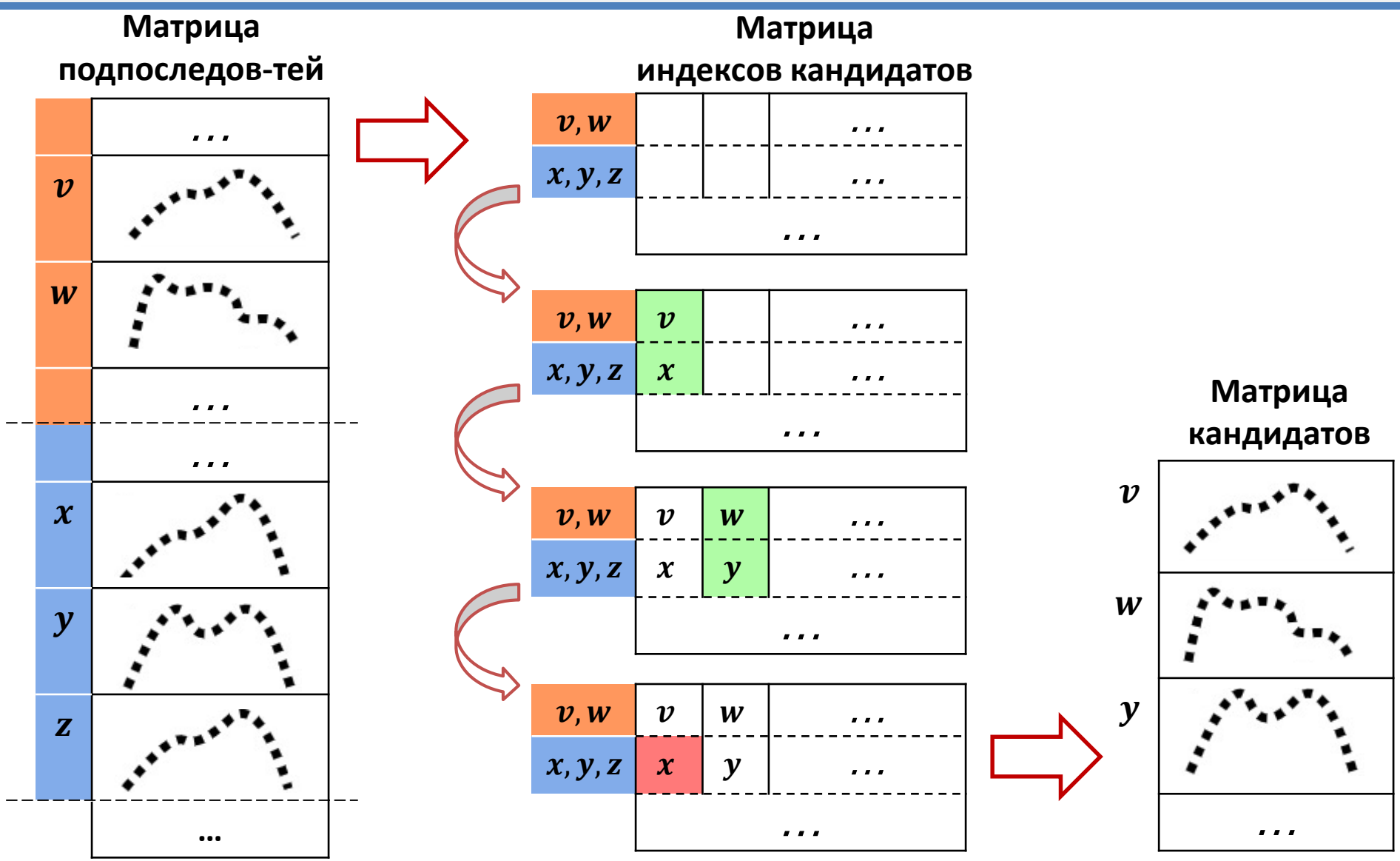
$$\mu = \frac{1}{n}\sum_{i=1}^{n} s_i, \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} t_i^2 - \mu^2$$



Матрица подпоследов-тей

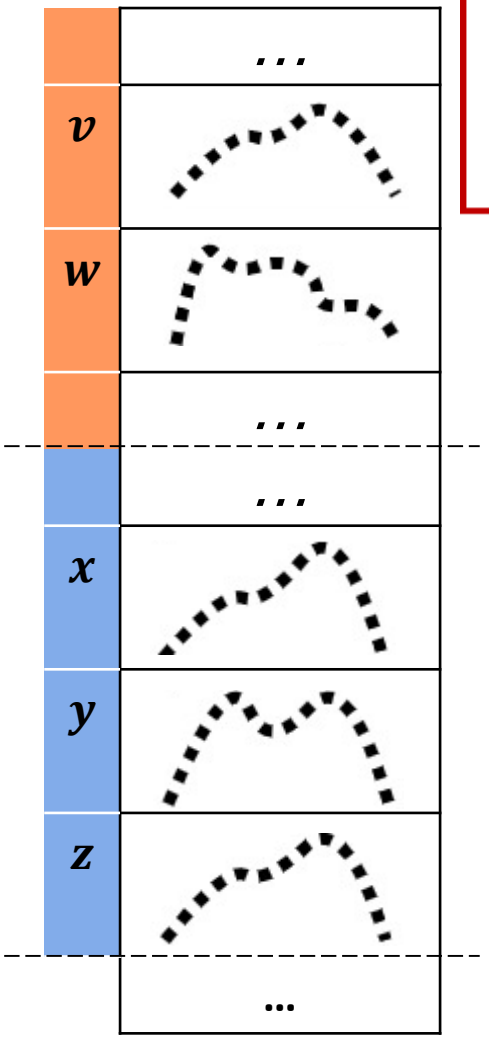


Отбор кандидатов на ускорителе



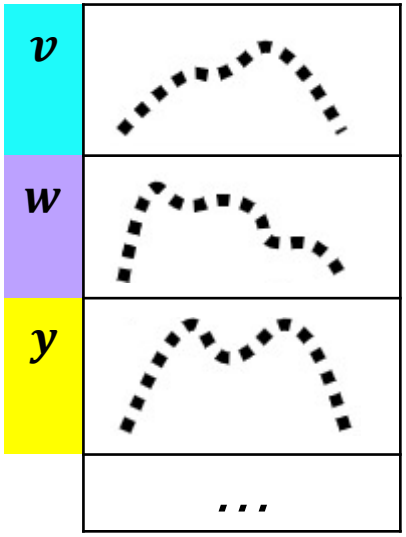
Очистка кандидатов на ускорителе

Матрица подпоследов-тей



$$B(i,j) := \bigwedge_{\substack{s \in SEG_i, \\ c_j \in \mathcal{C}}} \begin{pmatrix} s \cap c_j \neq \emptyset \\ \textbf{or} \\ \mathrm{ED}^2(s, c_j) \geq r^2 \end{pmatrix}$$

Матрица кандидатов



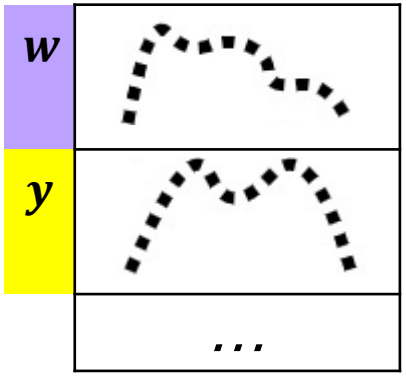
Битовая карта

	v	w	y	...
v, w	1	1	1	...
x, y, z	0	1	1	...
	...			

$\bigwedge_i B(\cdot, j)$

v	w	y	...
0	1	1	...

Матрица диссонансов

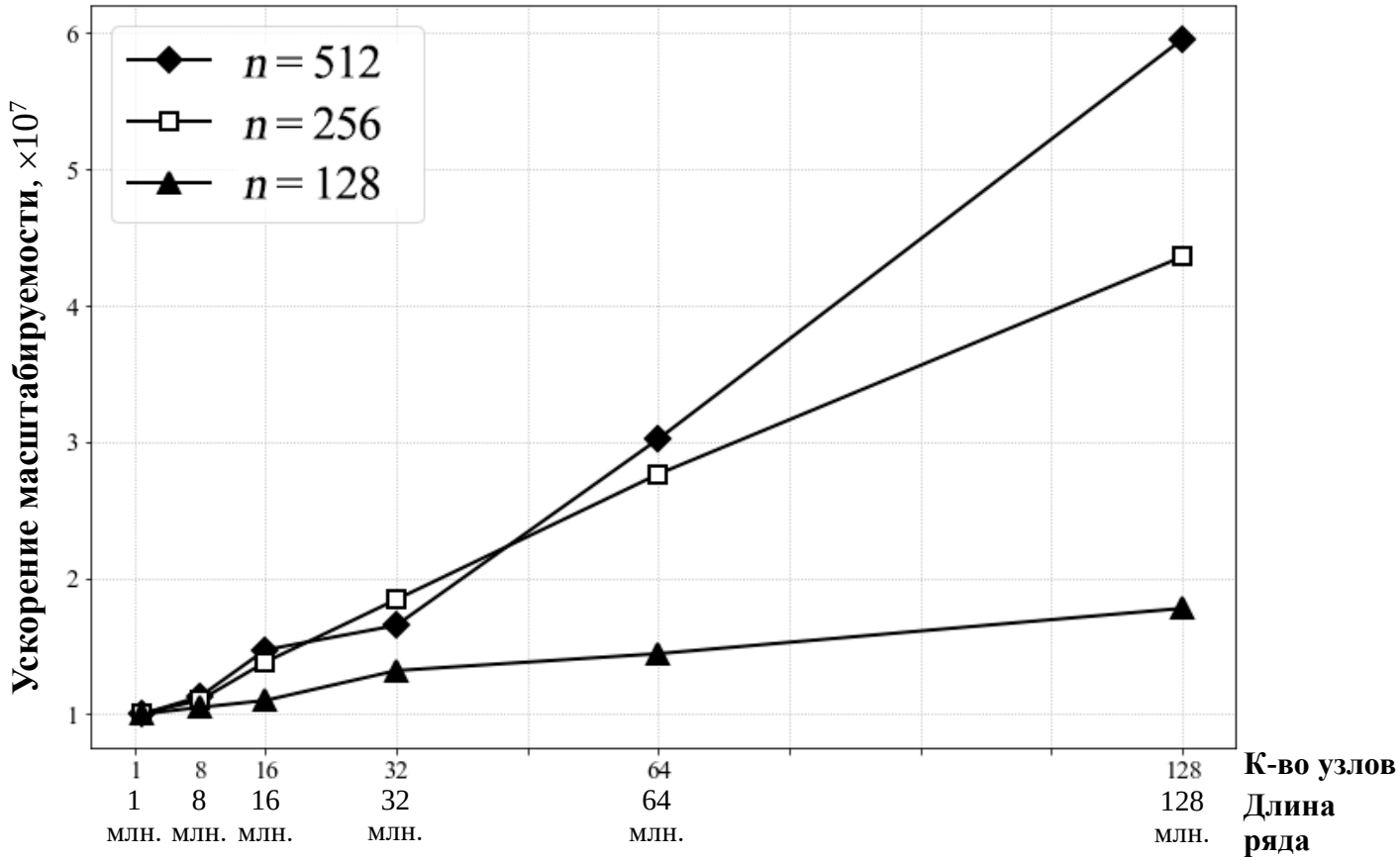


Эксперименты

- Аппаратная платформа
  - Кластер: «Торнадо ЮУрГУ», 128 узлов
  - Узел: ускоритель Intel Xeon Phi SE10X, 60 ядер @1.1 GHz, 1.076 TFLOPS
- Данные
  - ЭКГ[1)], 128 млн.
- Цели
  - Масштабируемость алгоритма

[1)] Goldberger A.L., *et al.* Physiobank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation. 101(23), e215-e220 (2000).

Ускорение масштабируемости



$$S_{scaled} = \frac{P \cdot n \cdot |\mathcal{C}|}{t_P}$$

P – количество узлов кластера
$|\mathcal{C}|$ – количество кандидатов в диссонансы
n – длина диссонанса
t_P – время работы на P узлах

Заключение

- Разработан новый параллельный алгоритм поиска диссонансов временного ряда для кластера с многоядерными ускорителями
- Проведены эксперименты, показавшие высокую масштабируемость алгоритма
- Будущие исследования
  - Поиск аномалий во временных рядах Industry 4.0
  - Разработка версии алгоритма для кластера с GPU

Спасибо за внимание!

Вопросы?