

Международная научная конференция
Параллельные вычислительные технологии 2024
Челябинск, 02-04 апреля 2024

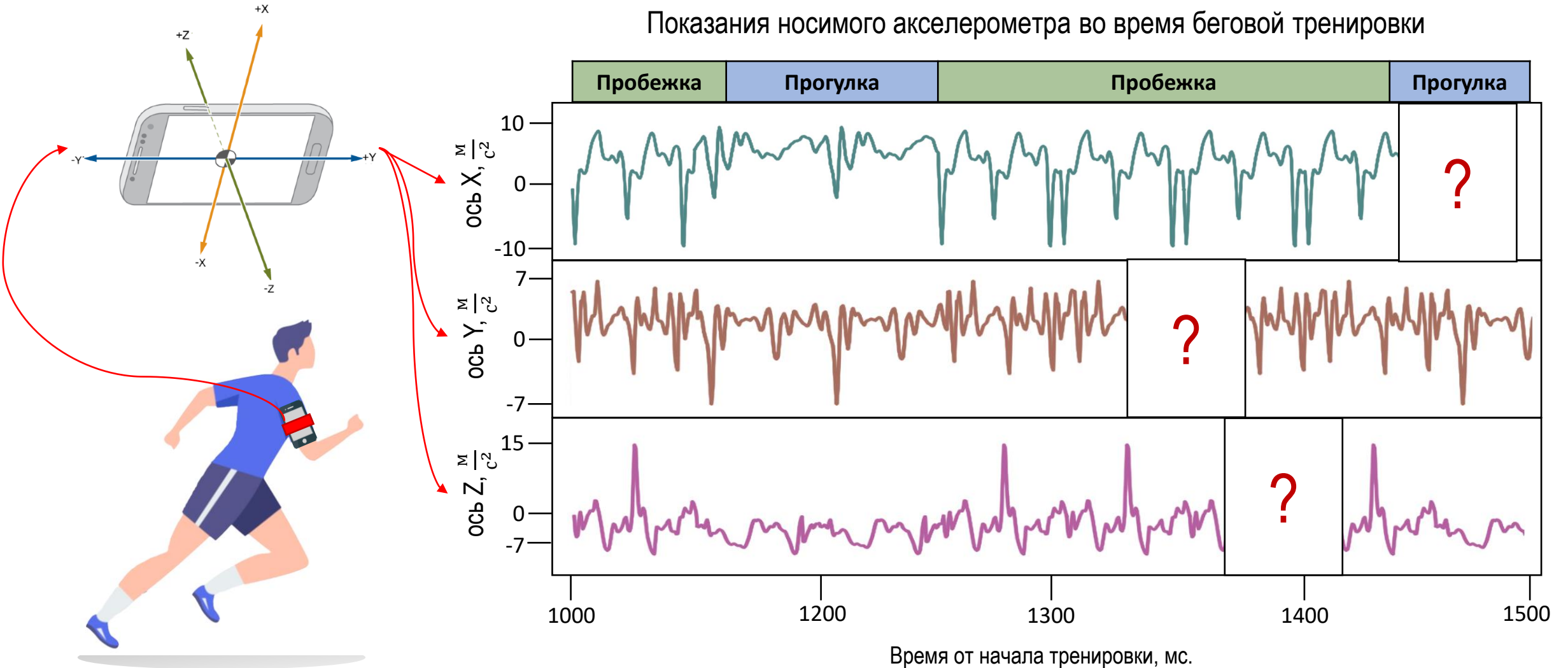
Восстановление пропусков многомерного временного ряда на основе анализа поведенческих шаблонов

А.А. Юртин, М.Л. Цымблер

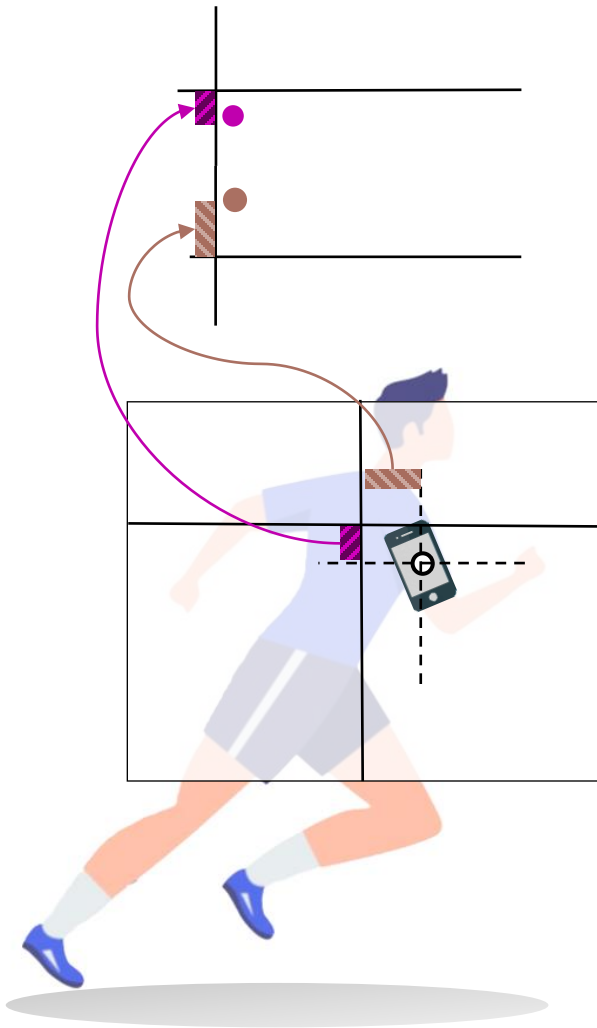
Южно-Уральский государственный университет, Челябинск

Восстановление сенсорных данных носимых устройств

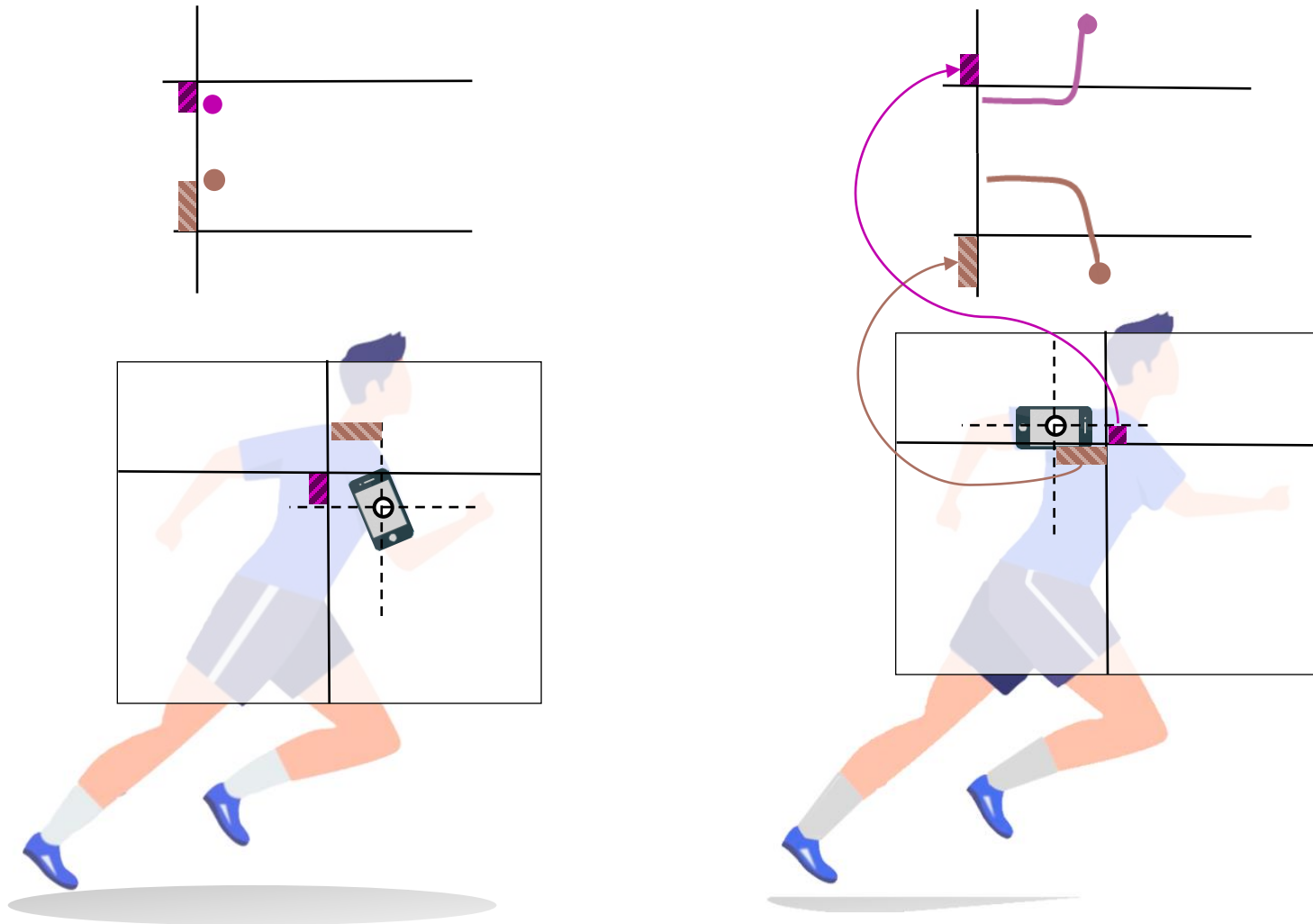
Показания носимого акселерометра во время беговой тренировки



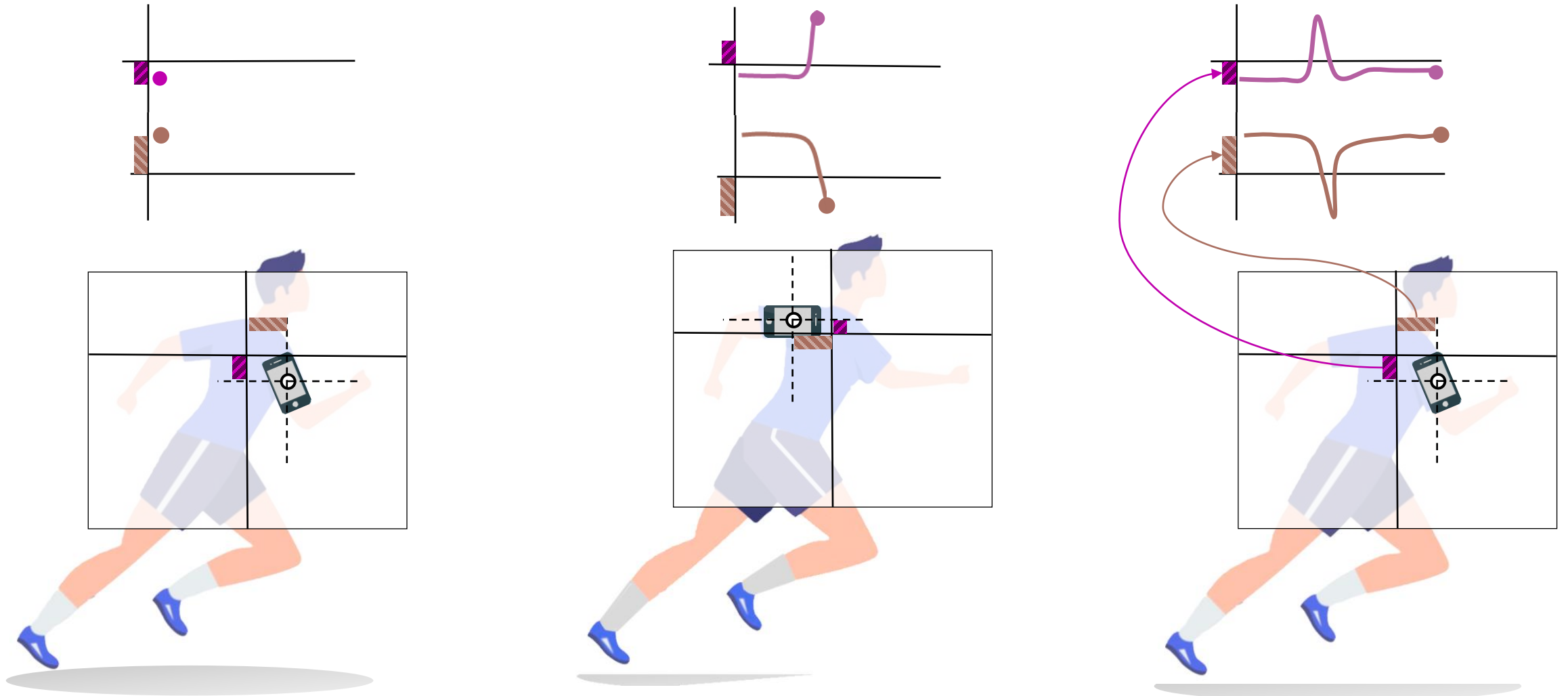
Поведенческие шаблоны



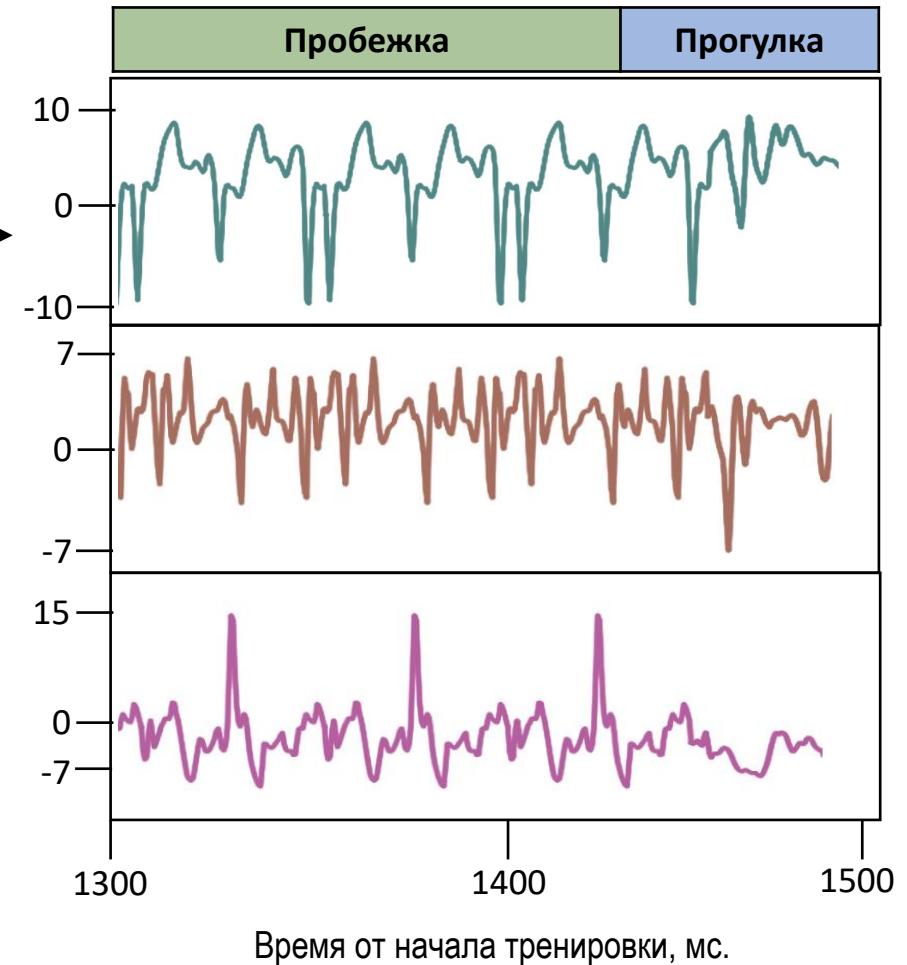
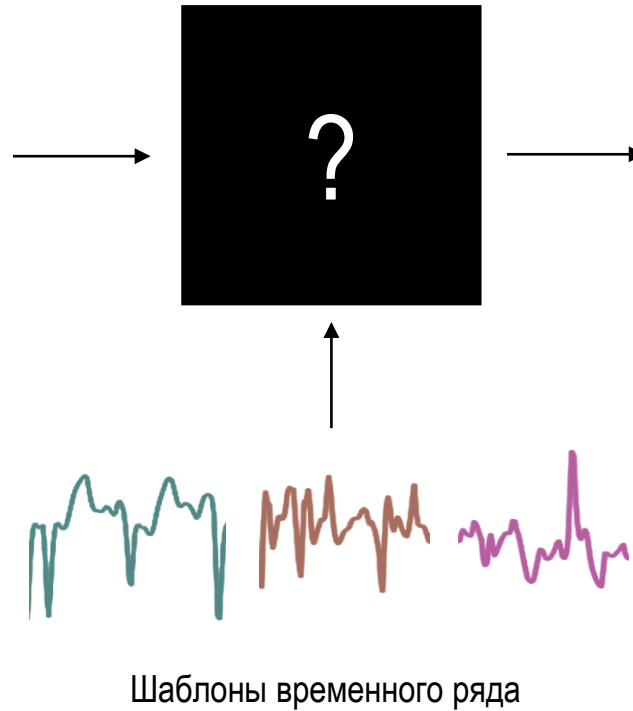
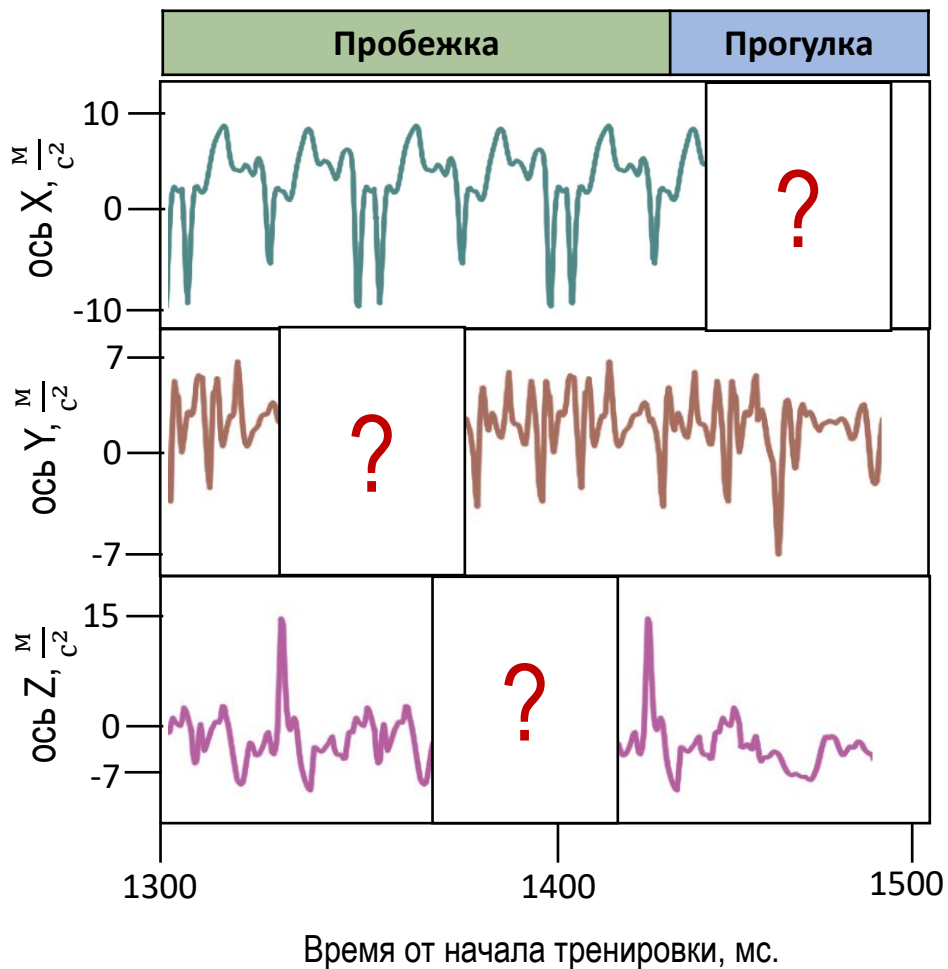
Поведенческие шаблоны



Поведенческие шаблоны

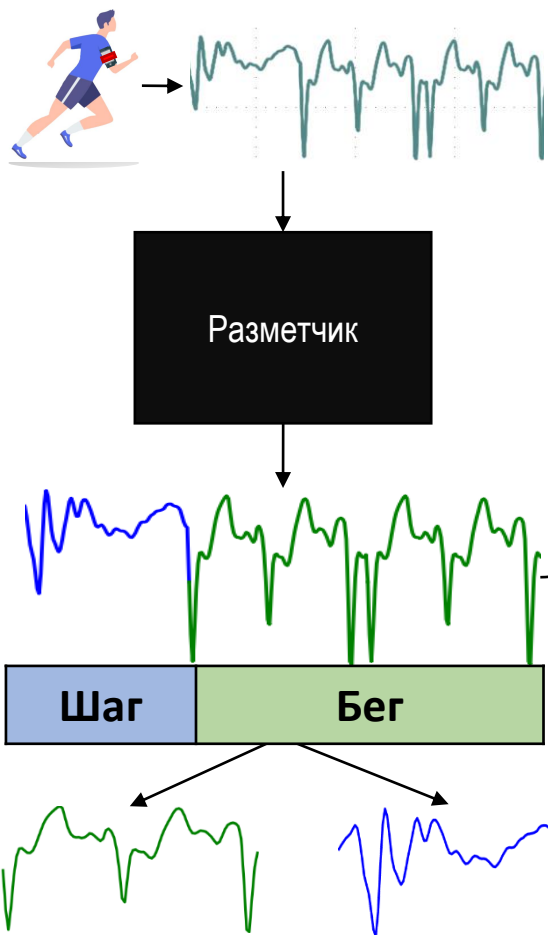


Восстановление сенсорных данных носимых устройств

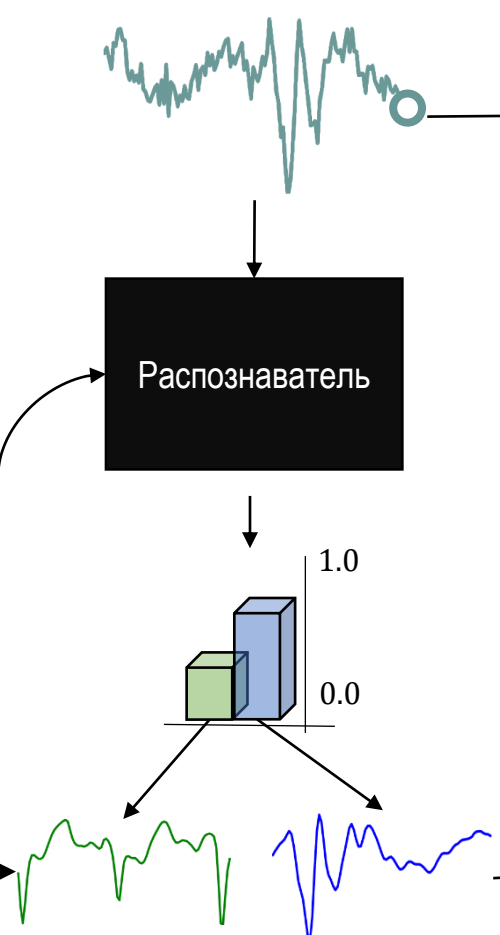


Архитектура решения

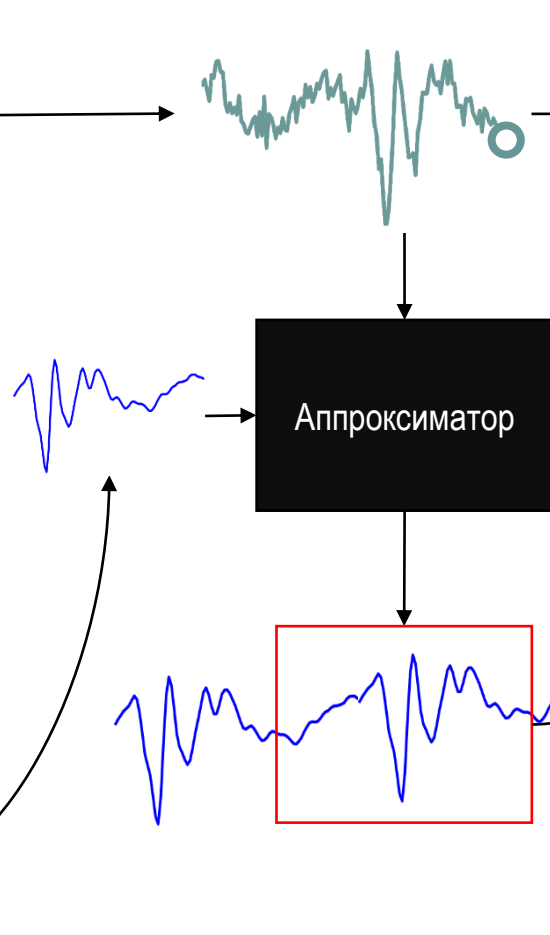
Шаг 0: Предобработка данных
Разметка исходных данных



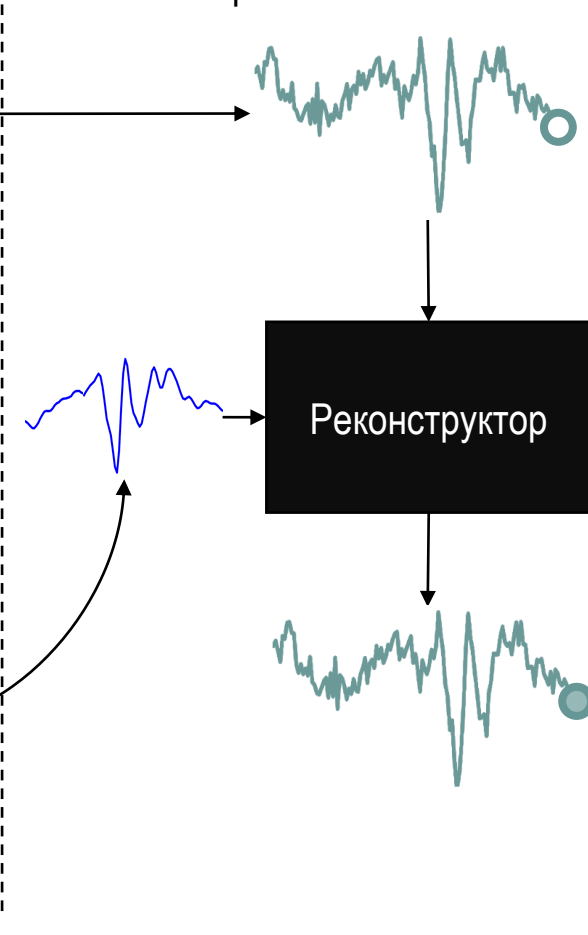
Шаг 1: Подготовка данных
Прогноз текущей активности



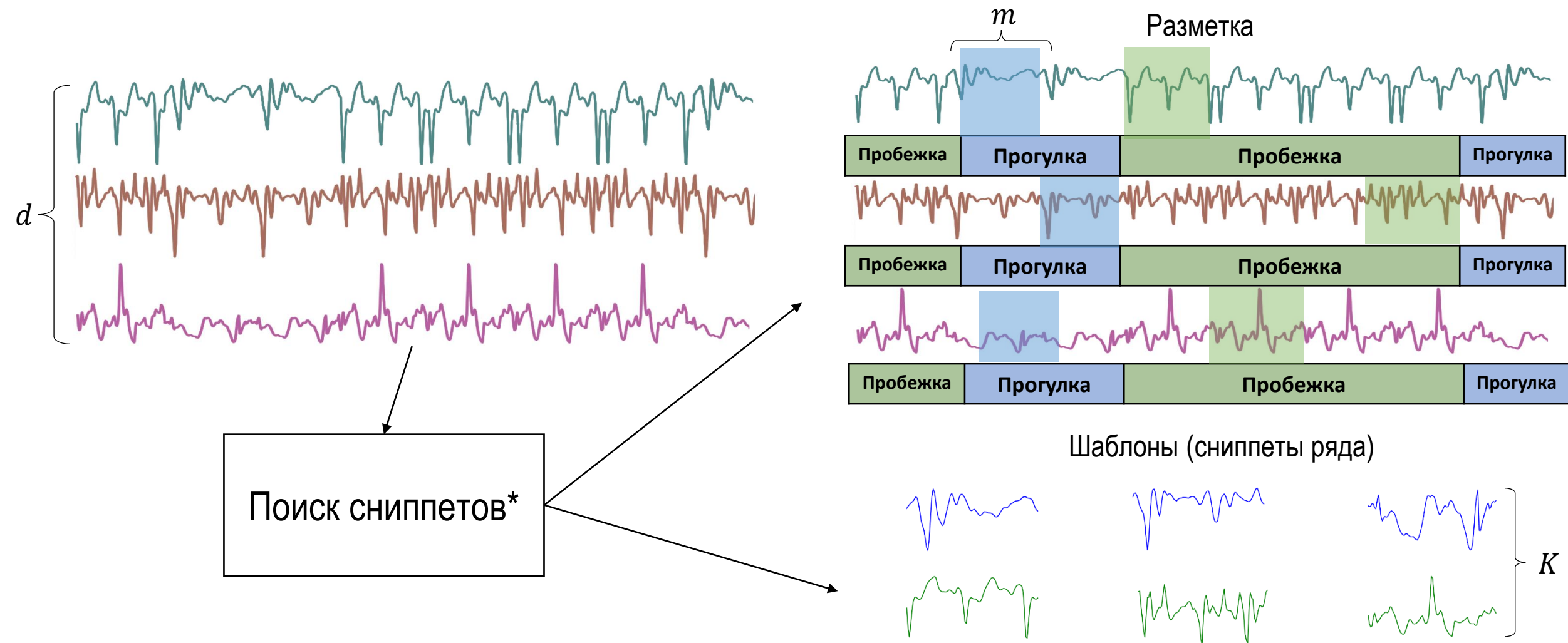
Шаг 2: Аппроксиматор
Подготовка первого приближения поведения



Шаг 3: Восстановление
Правильное использование первого приближения



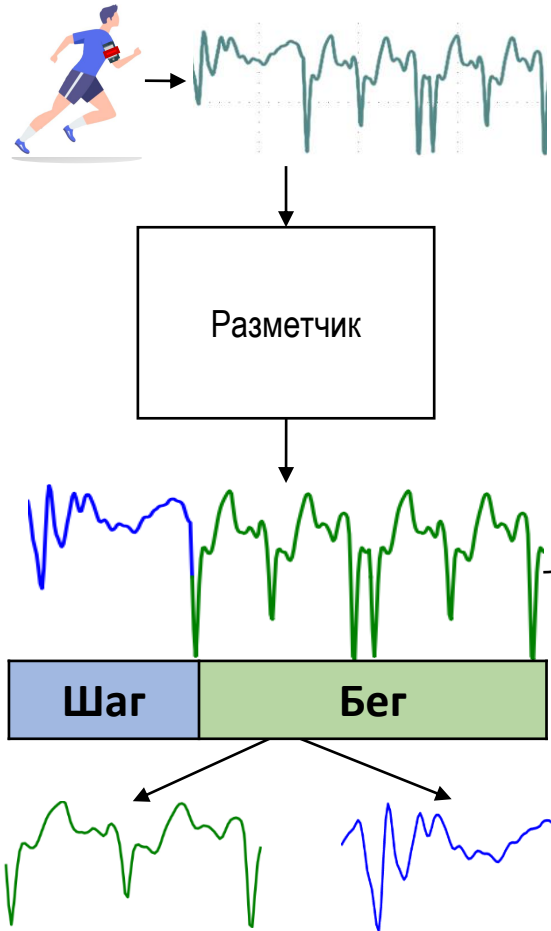
Поиск поведенческих шаблонов (сниппетов)



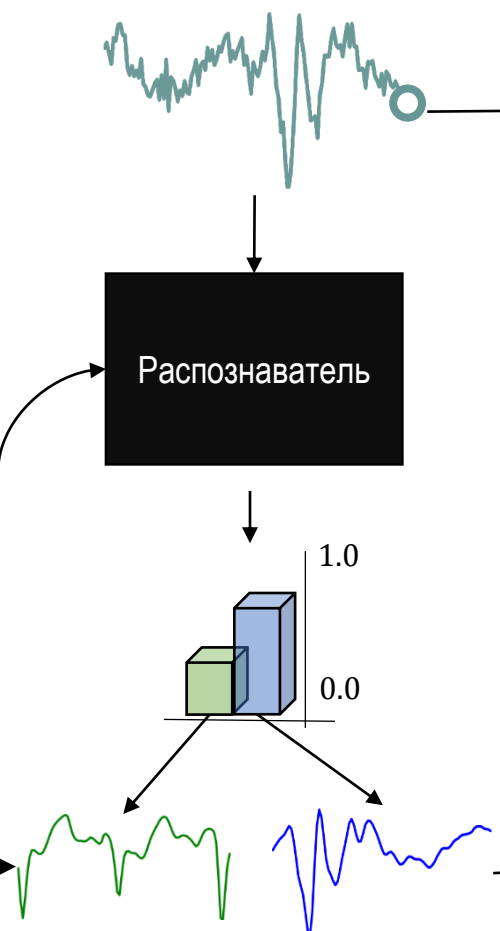
*Zymbler M., Goglachev A. Fast Summarization of Long Time Series with Graphics Processor // Mathematics. 2022. Vol. 10, No. 10. Article 1781. DOI: 10.3390/math10101781

Архитектура решения

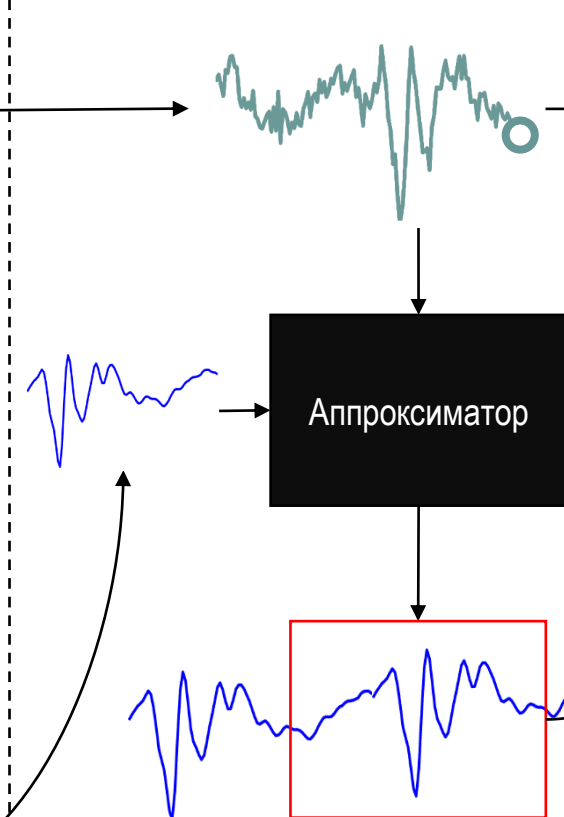
Шаг 0: Предобработка данных
Разметка исходных данных



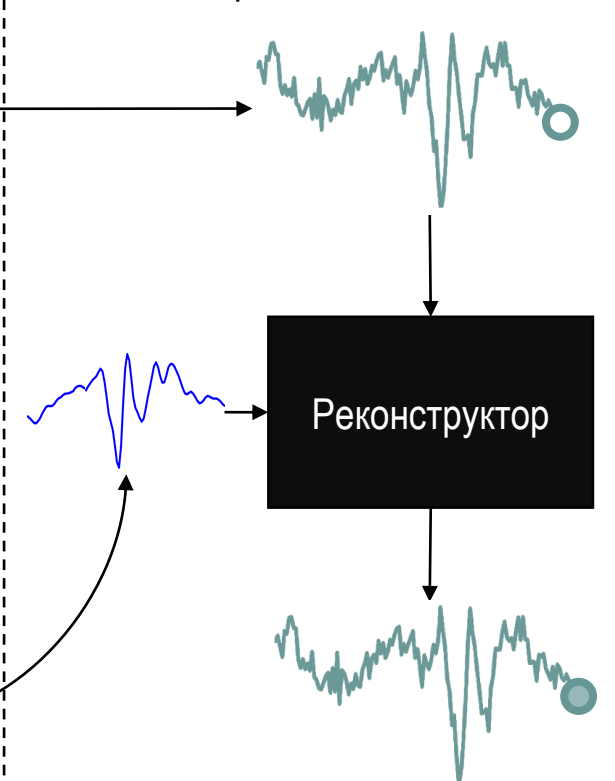
Шаг 1: Подготовка данных
Прогноз текущей активности



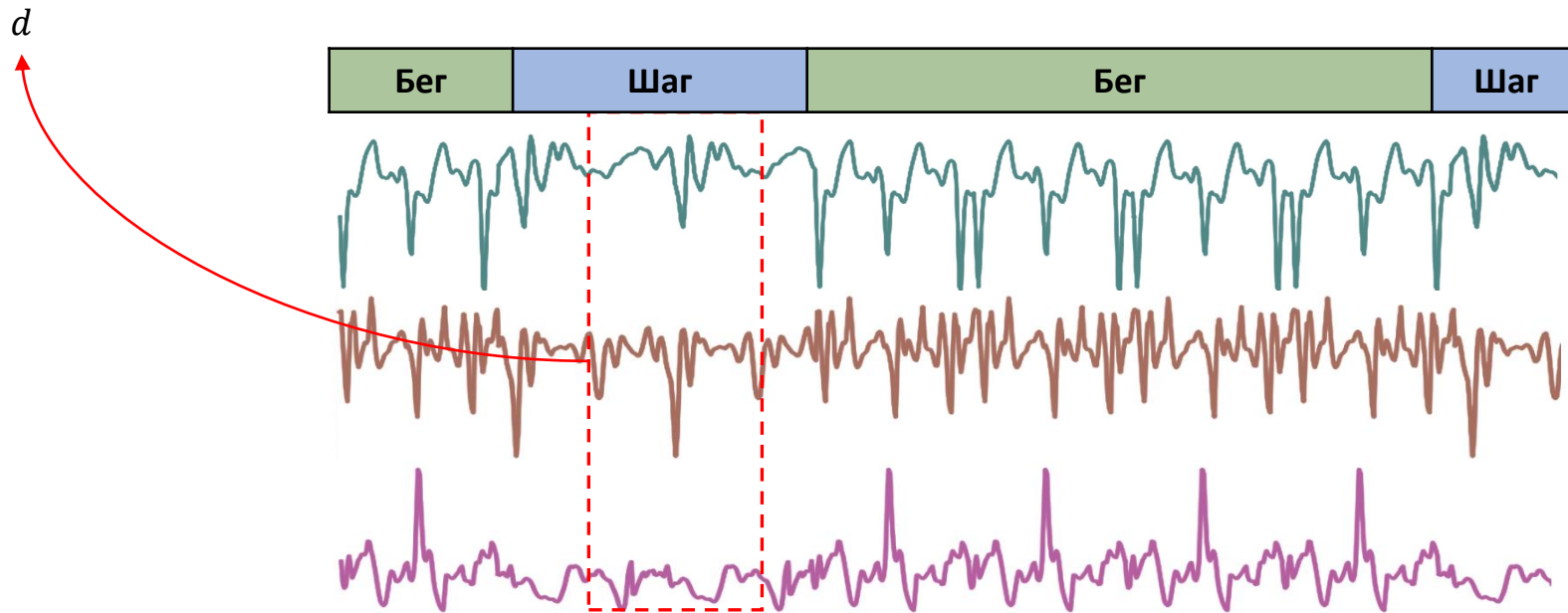
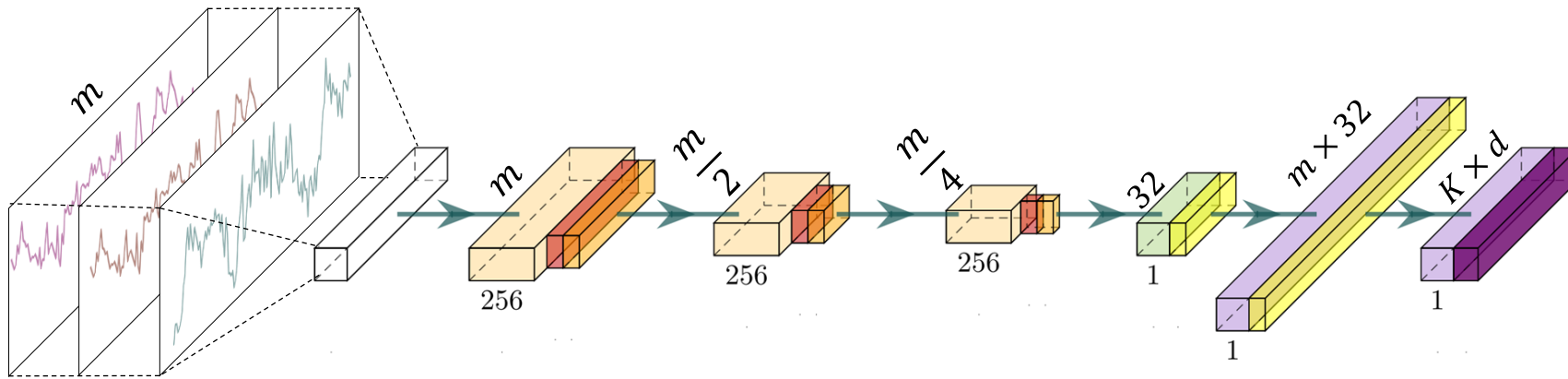
Шаг 2: Аппроксиматор
Подготовка первого приближения поведения



Шаг 3: Восстановление
Правильное использование первого приближения

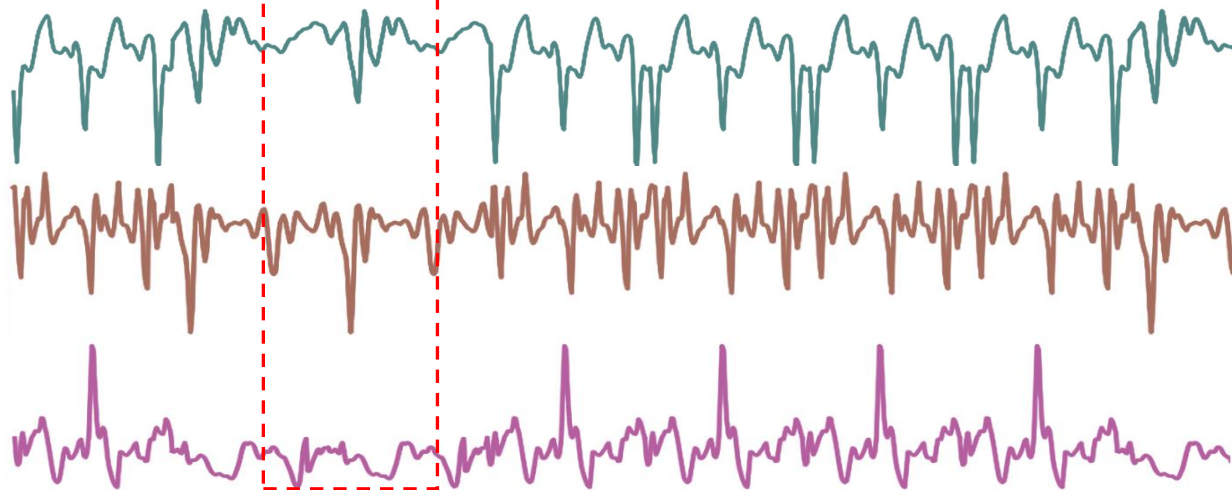
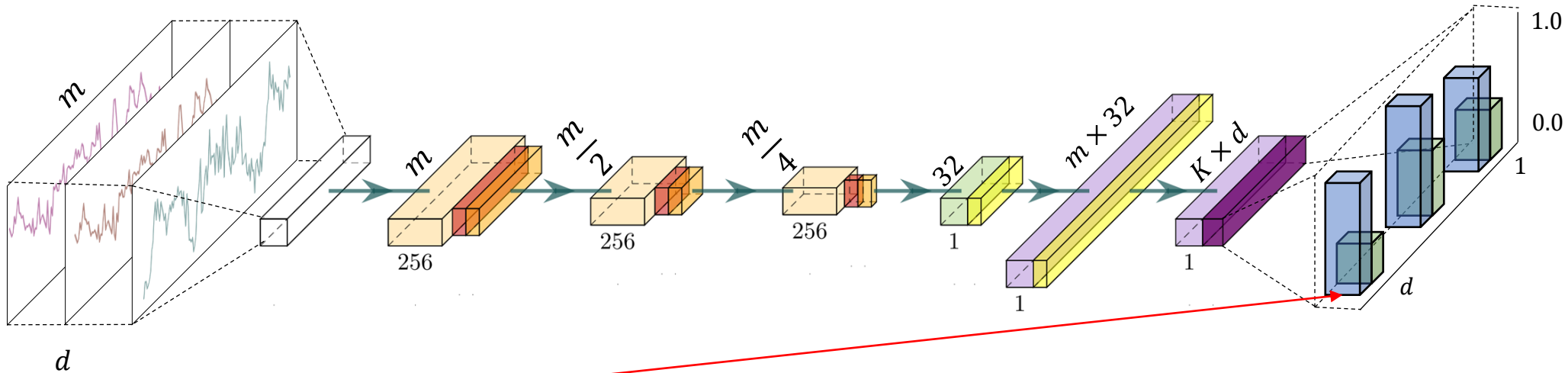


Распознаватель



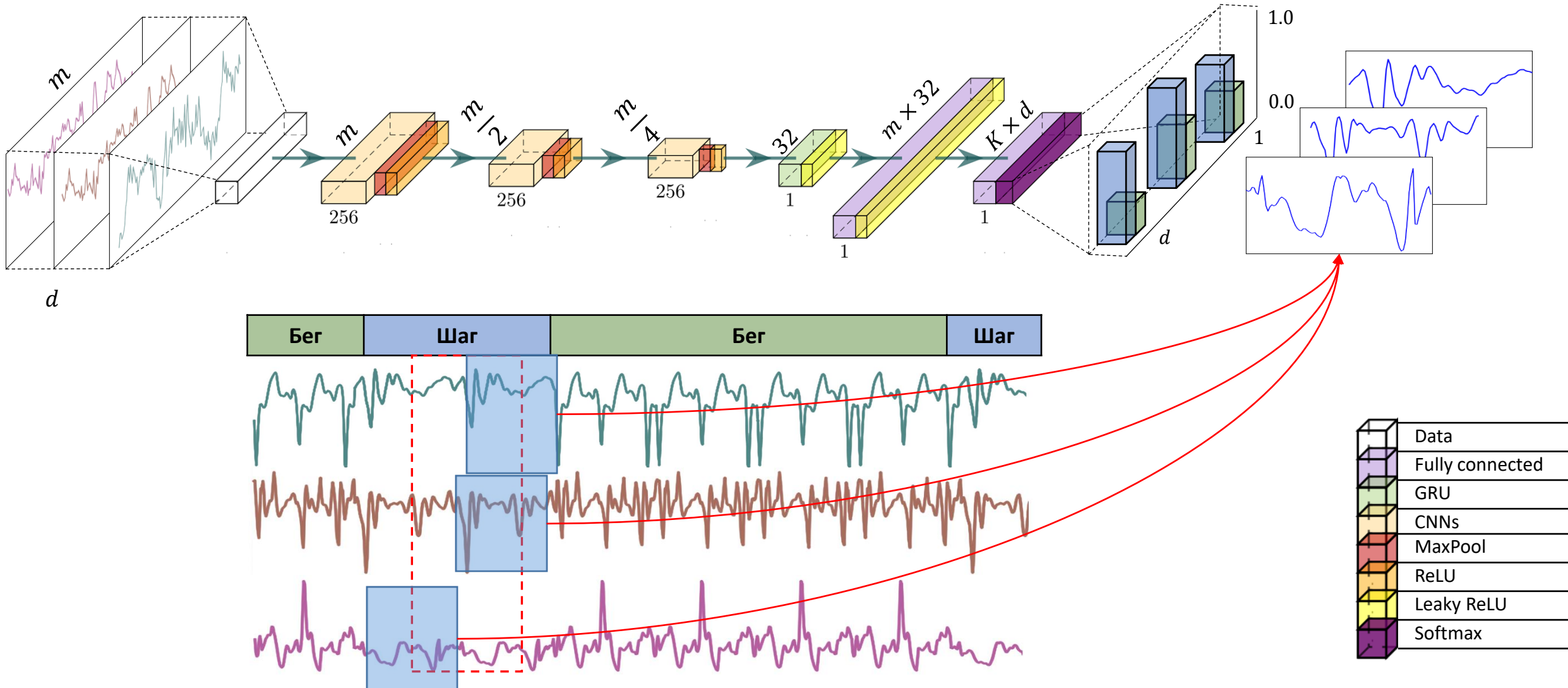
	Data
	Fully connected
	GRU
	CNNs
	MaxPool
	ReLU
	Leaky ReLU
	Softmax

Распознаватель



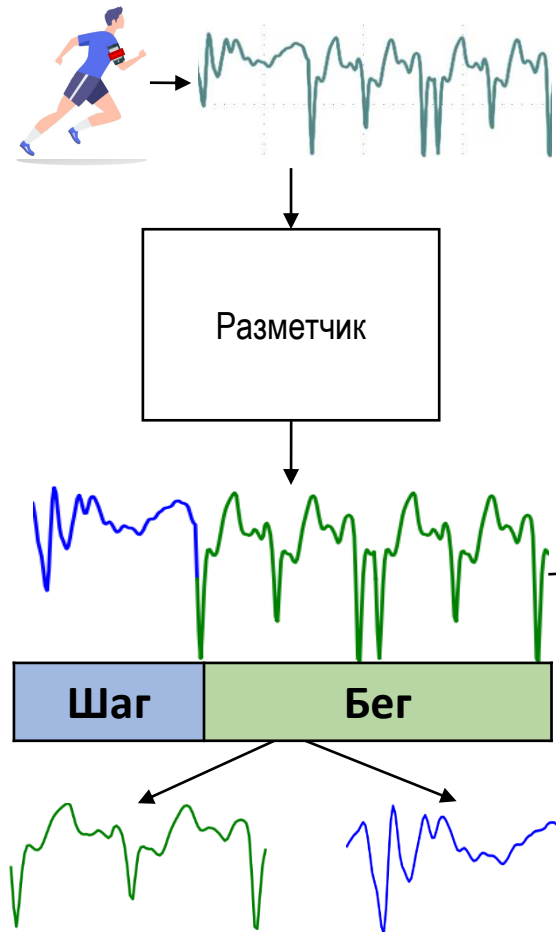
	Data
	Fully connected
	GRU
	CNNs
	MaxPool
	ReLU
	Leaky ReLU
	Softmax

Распознаватель

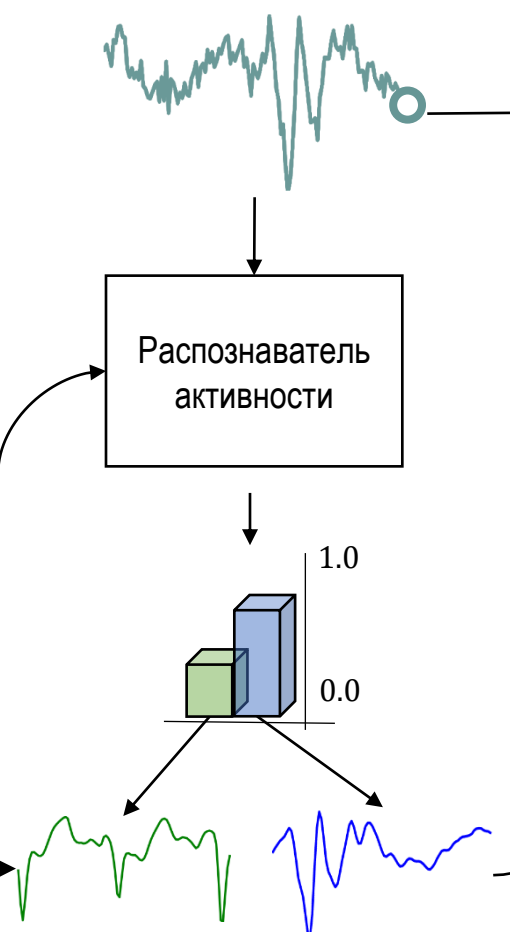


Архитектура предлагаемого решения

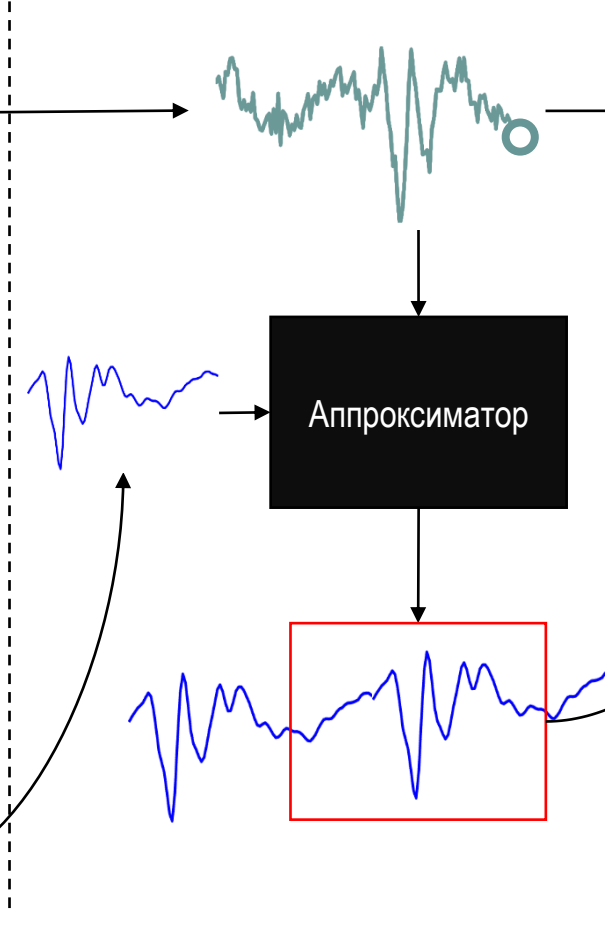
Шаг 0: Предобработка данных
Разметка исходных данных



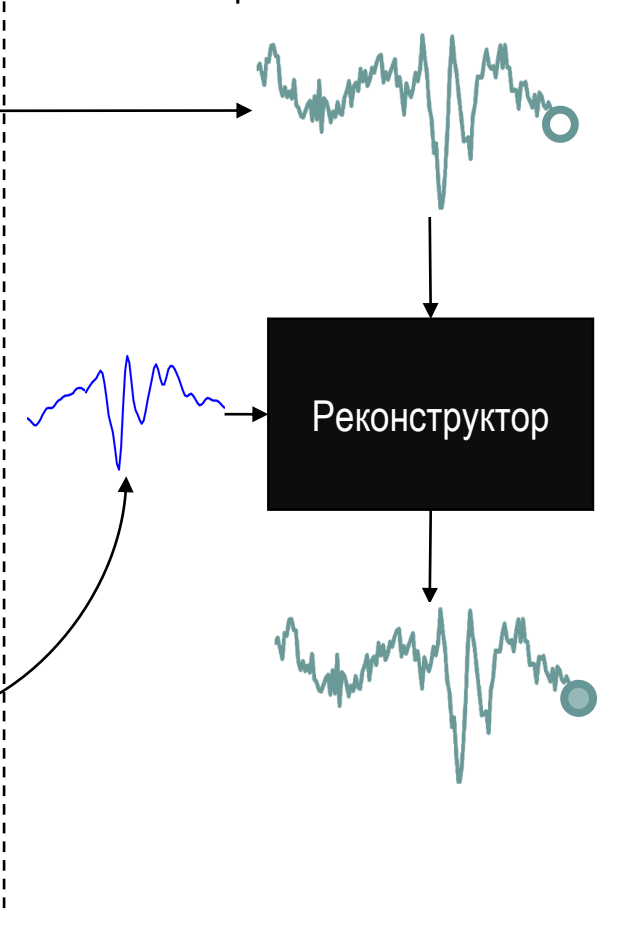
Шаг 1: Подготовка данных
Прогноз текущей активности



Шаг 2: Аппроксиматор
Подготовка первого приближения поведения

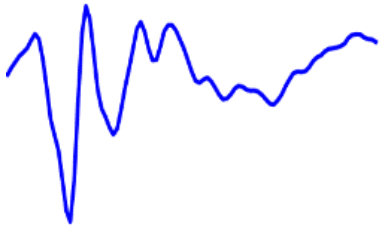
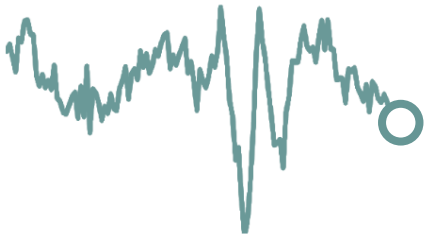


Шаг 3: Восстановление
Правильное использование первого приближения



Аппроксиматор

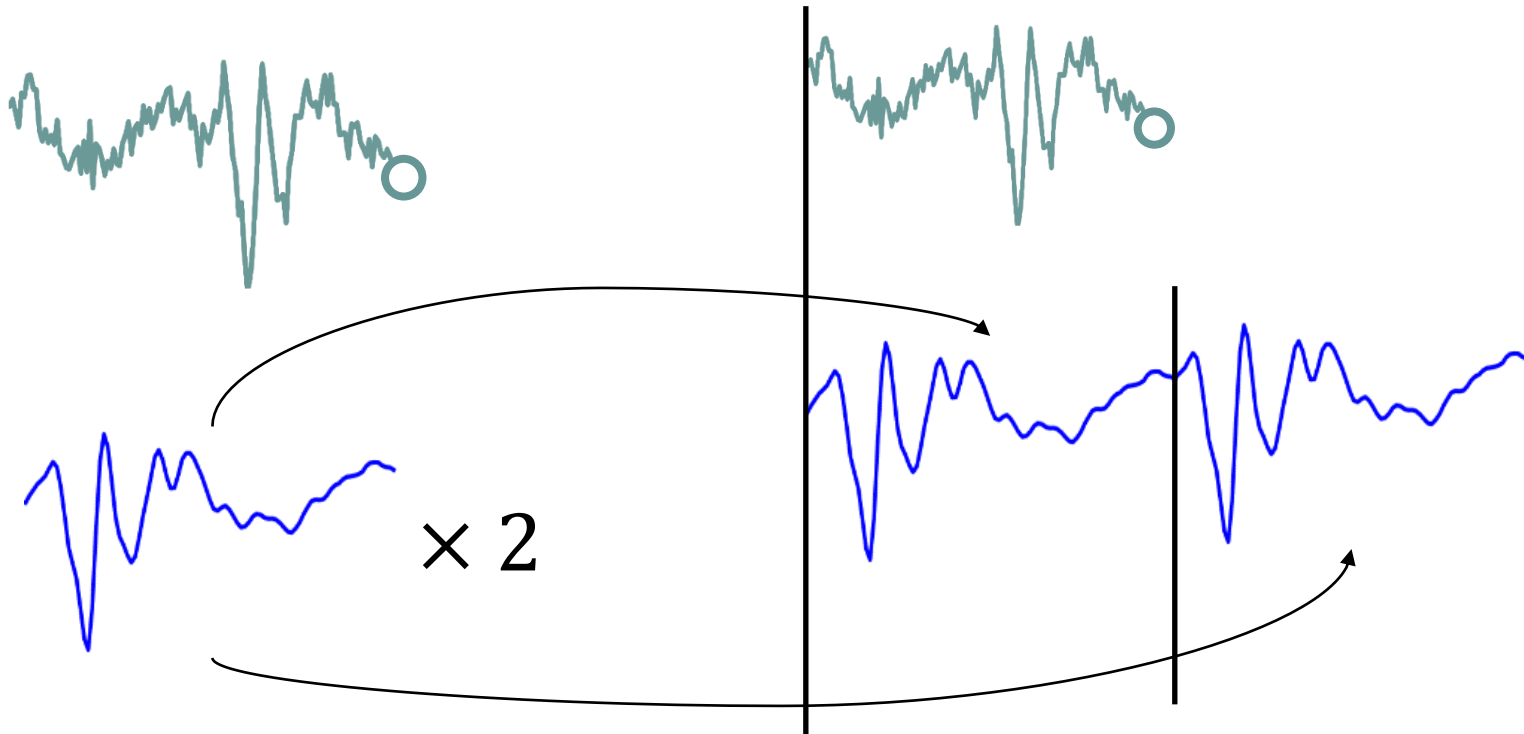
Шаг 1: Подготовка данных



Аппроксиматор

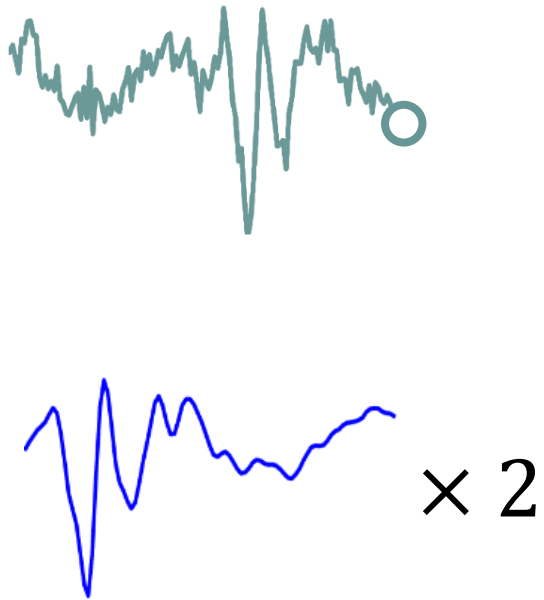
Шаг 1: Подготовка данных

Шаг 2: Поиск подходящих фрагментов

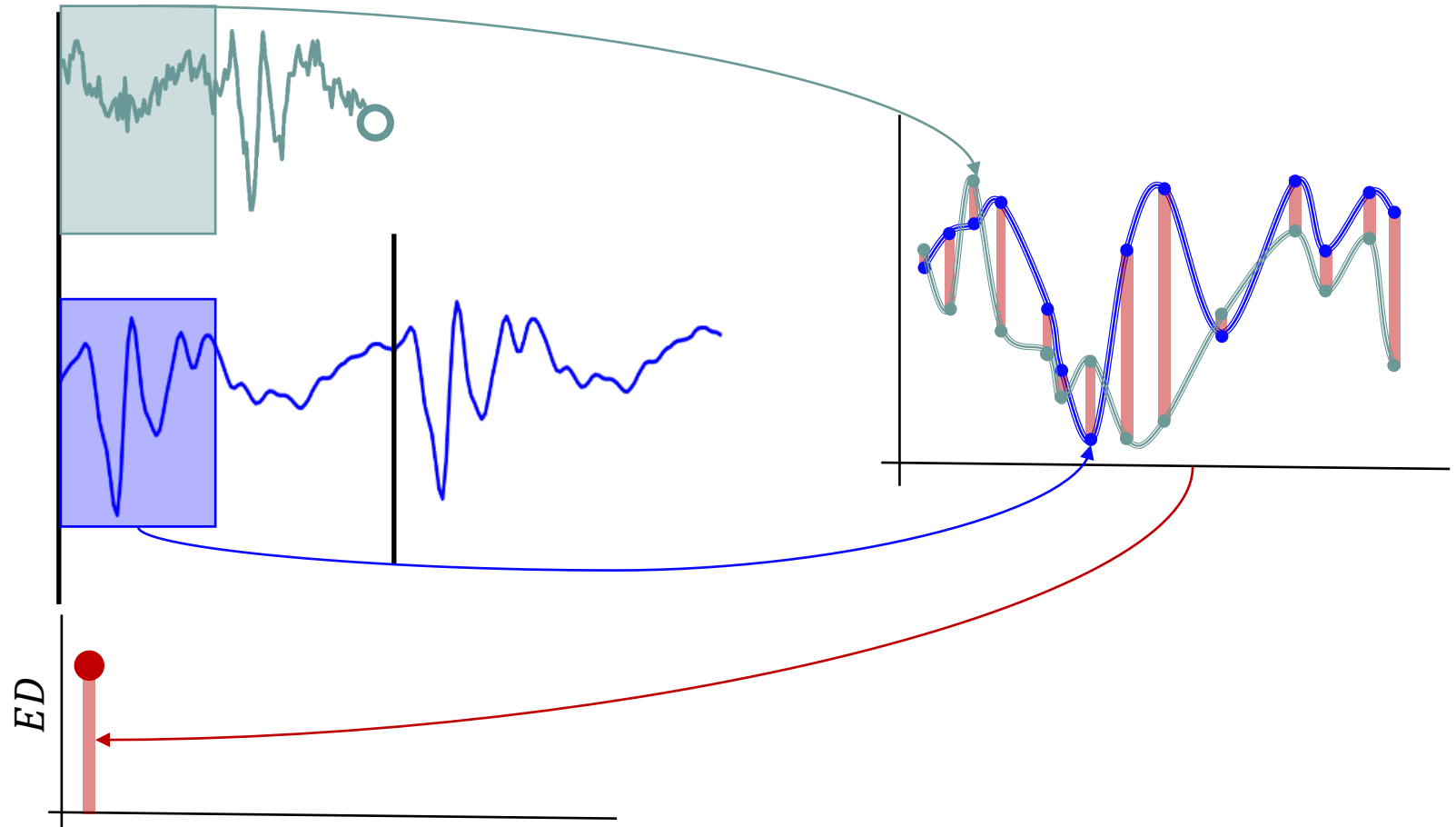


Аппроксиматор

Шаг 1: Подготовка данных

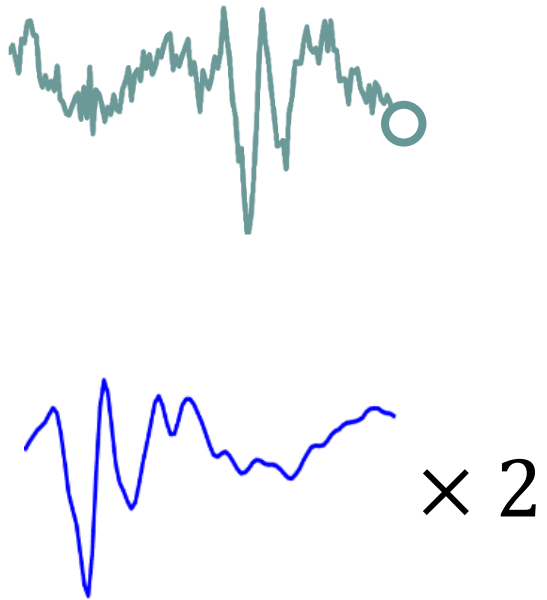


Шаг 2: Поиск подходящих фрагментов

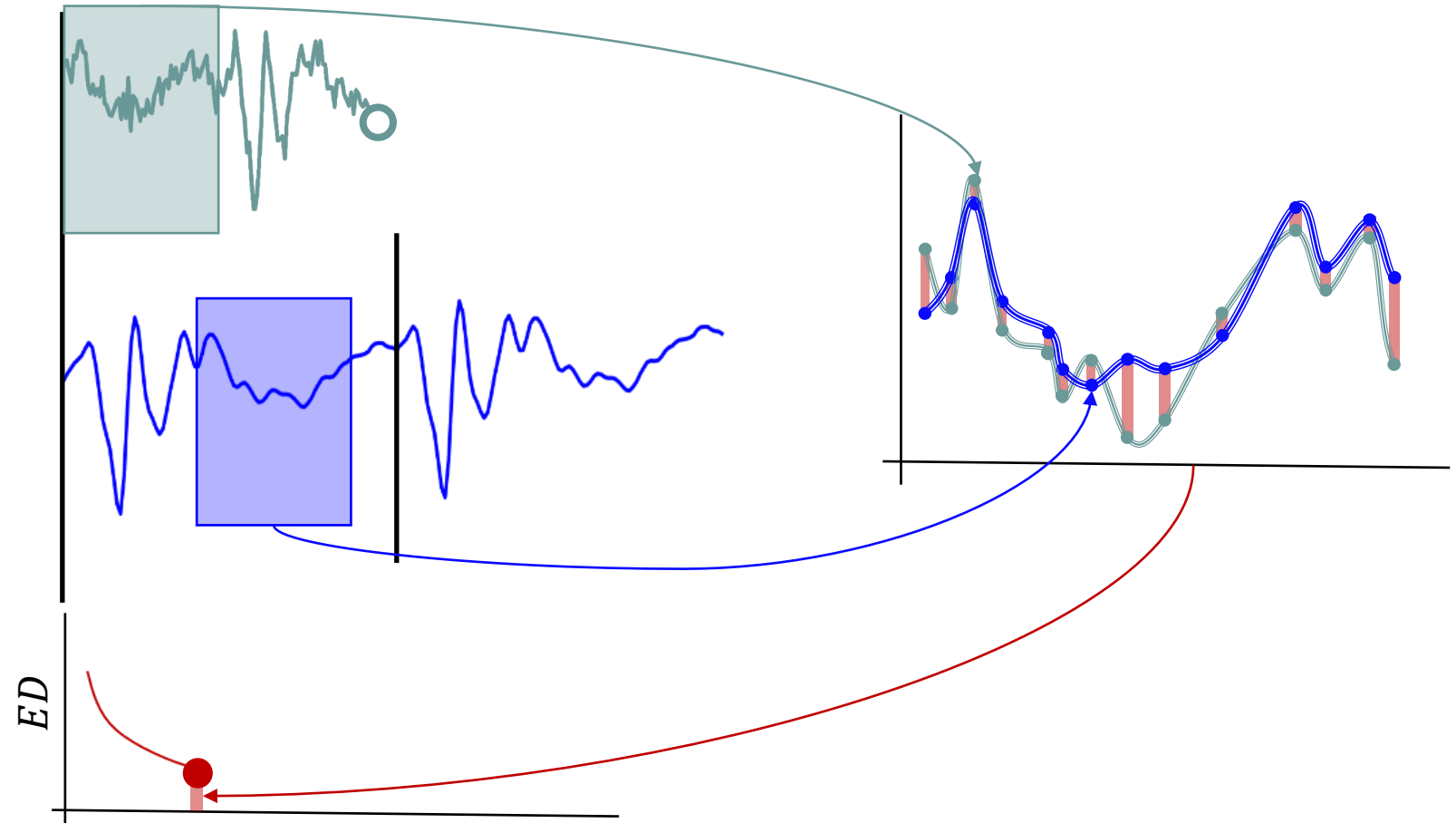


Аппроксиматор

Шаг 1: Подготовка данных

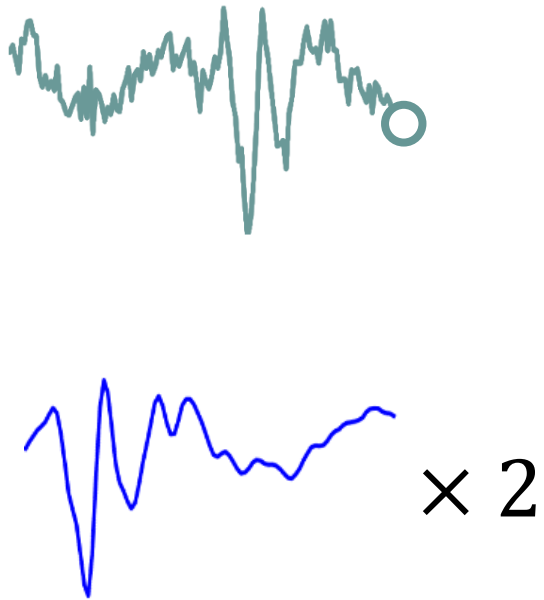


Шаг 2: Поиск подходящих фрагментов

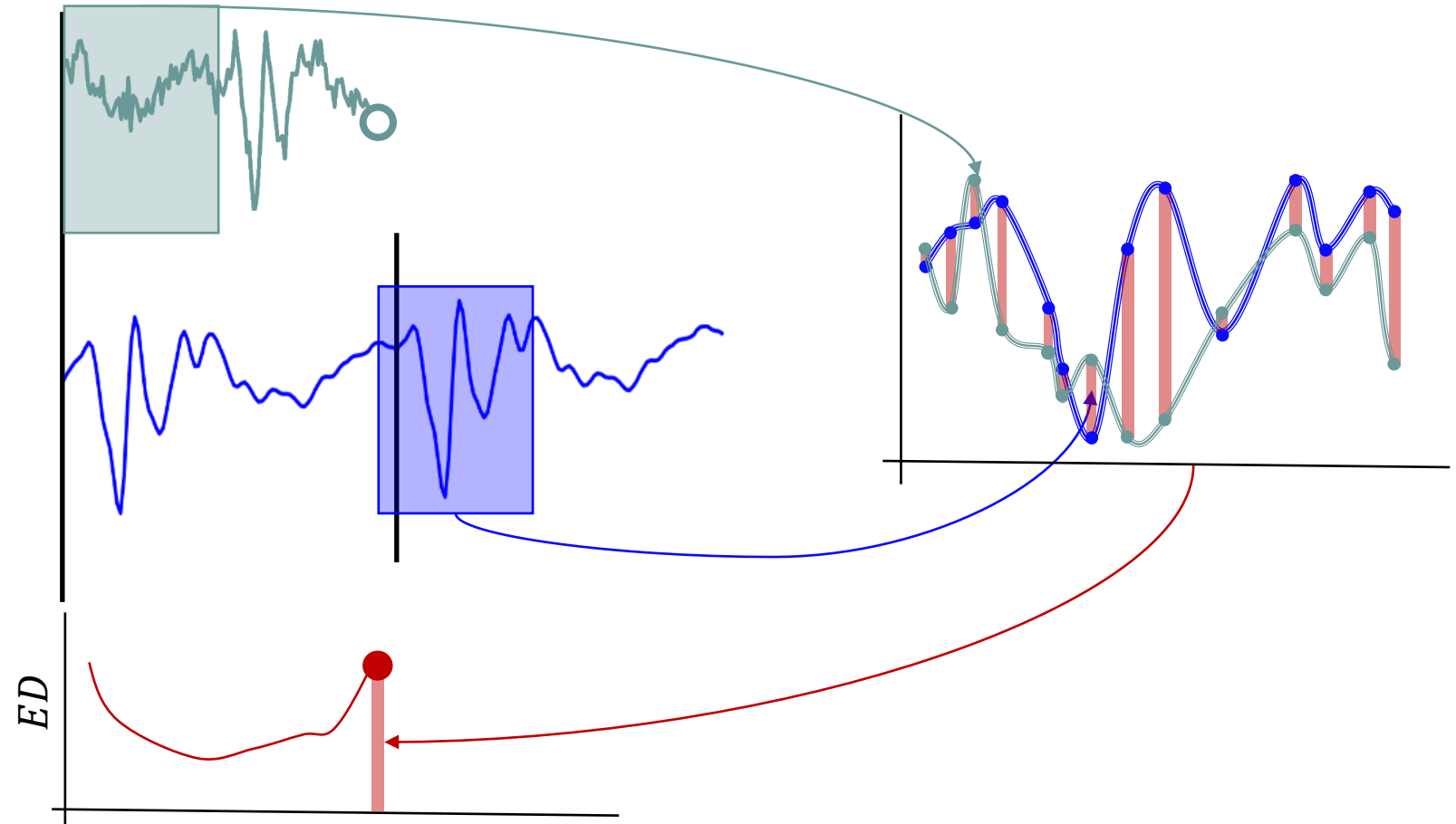


Аппроксиматор

Шаг 1: Подготовка данных

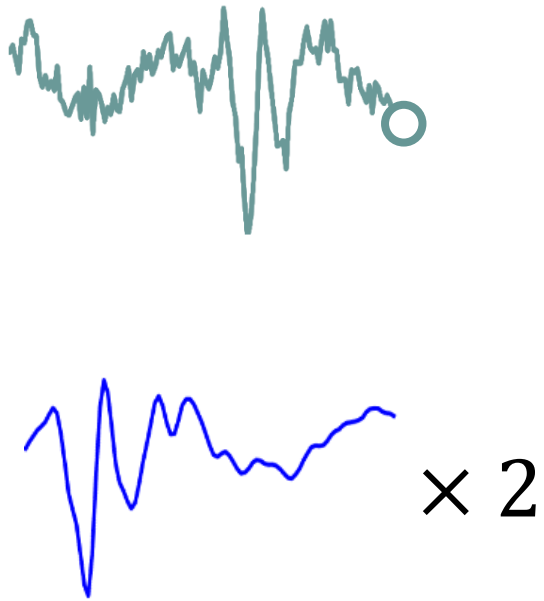


Шаг 2: Поиск подходящих фрагментов

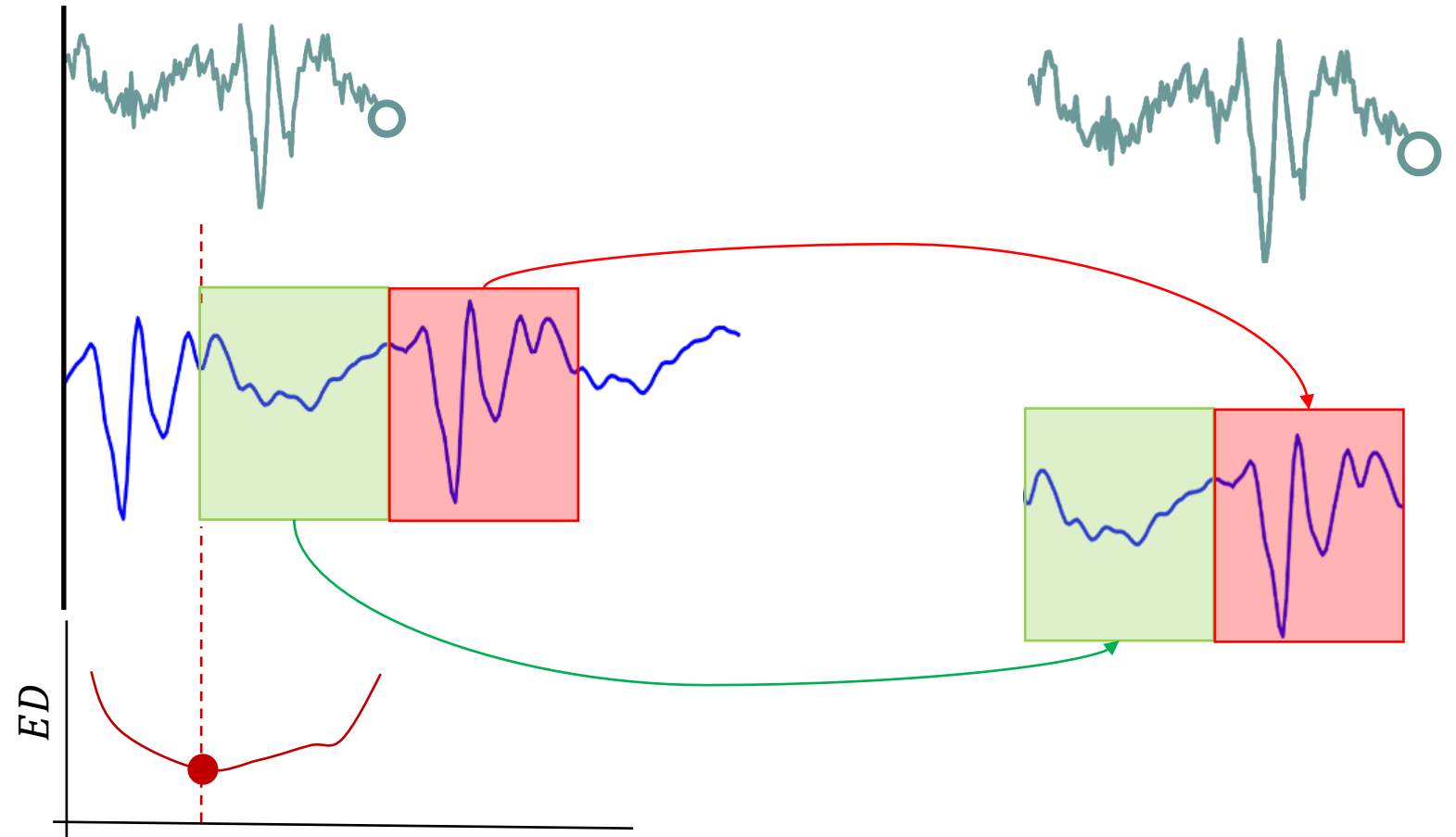


Аппроксиматор

Шаг 1: Подготовка данных

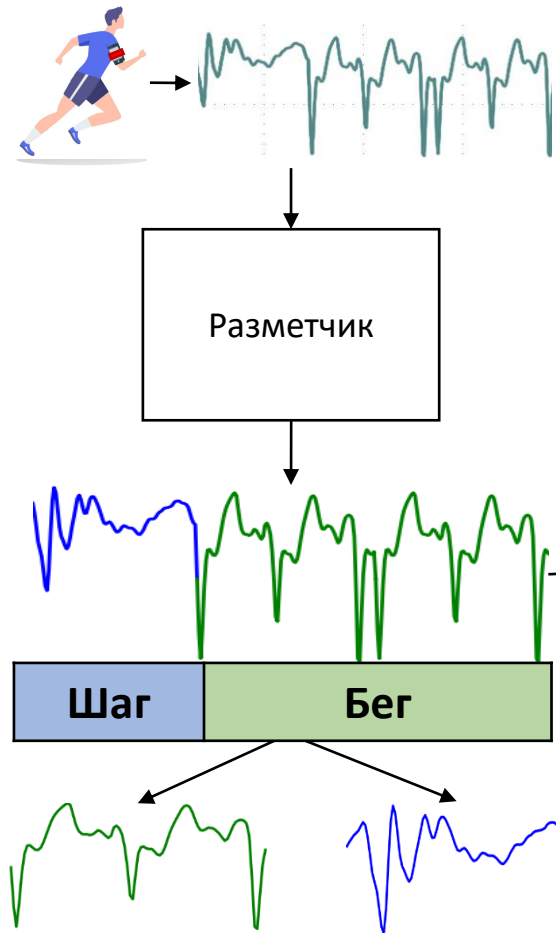


Шаг 2: Поиск подходящих фрагментов

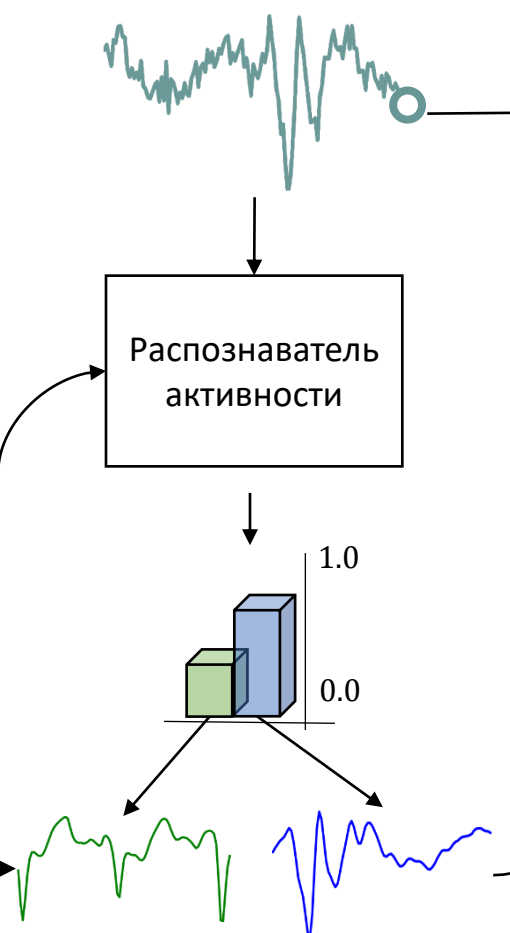


Архитектура метода

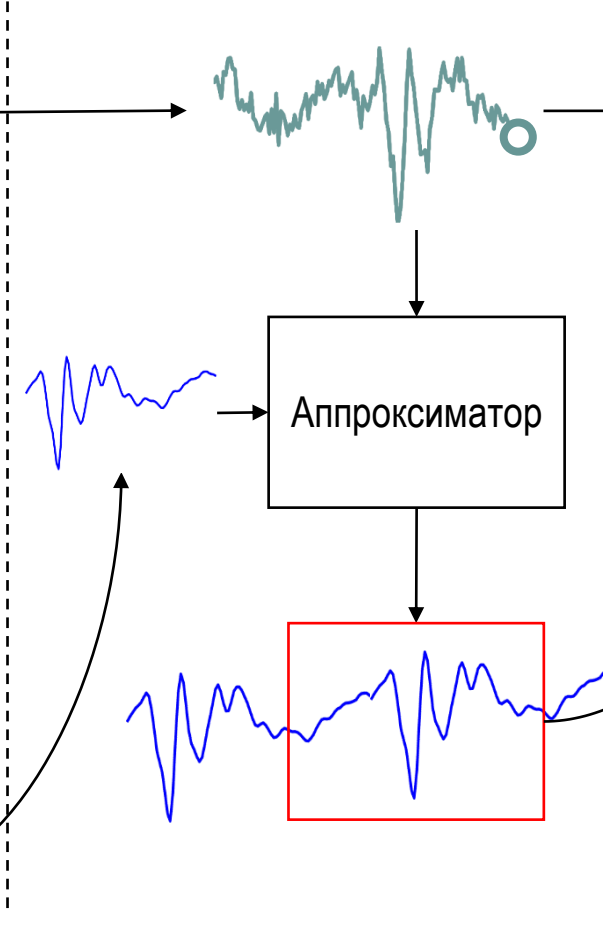
Шаг 0: Предобработка данных
Разметка исходных данных



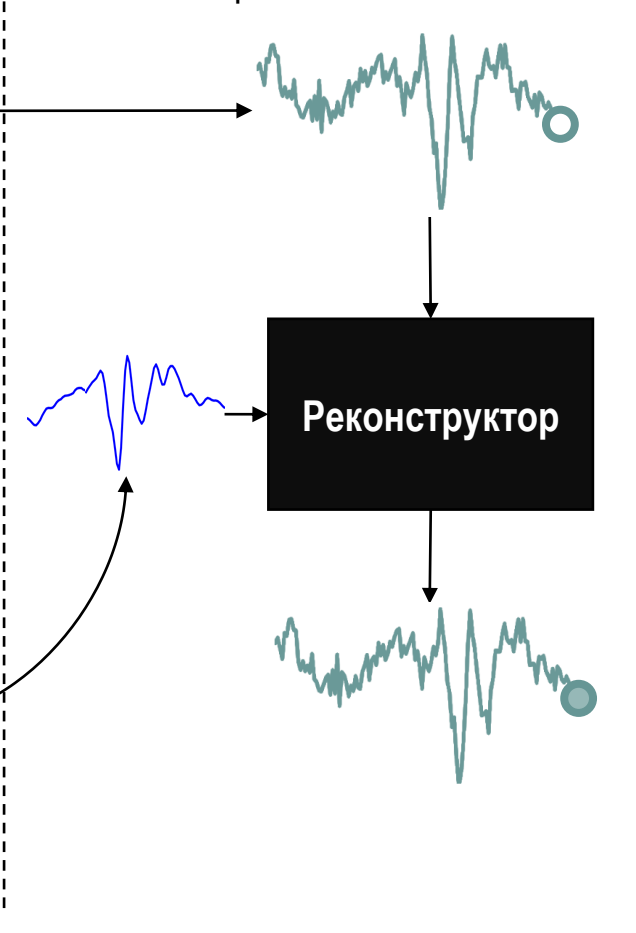
Шаг 1: Подготовка данных
Прогноз текущей активности



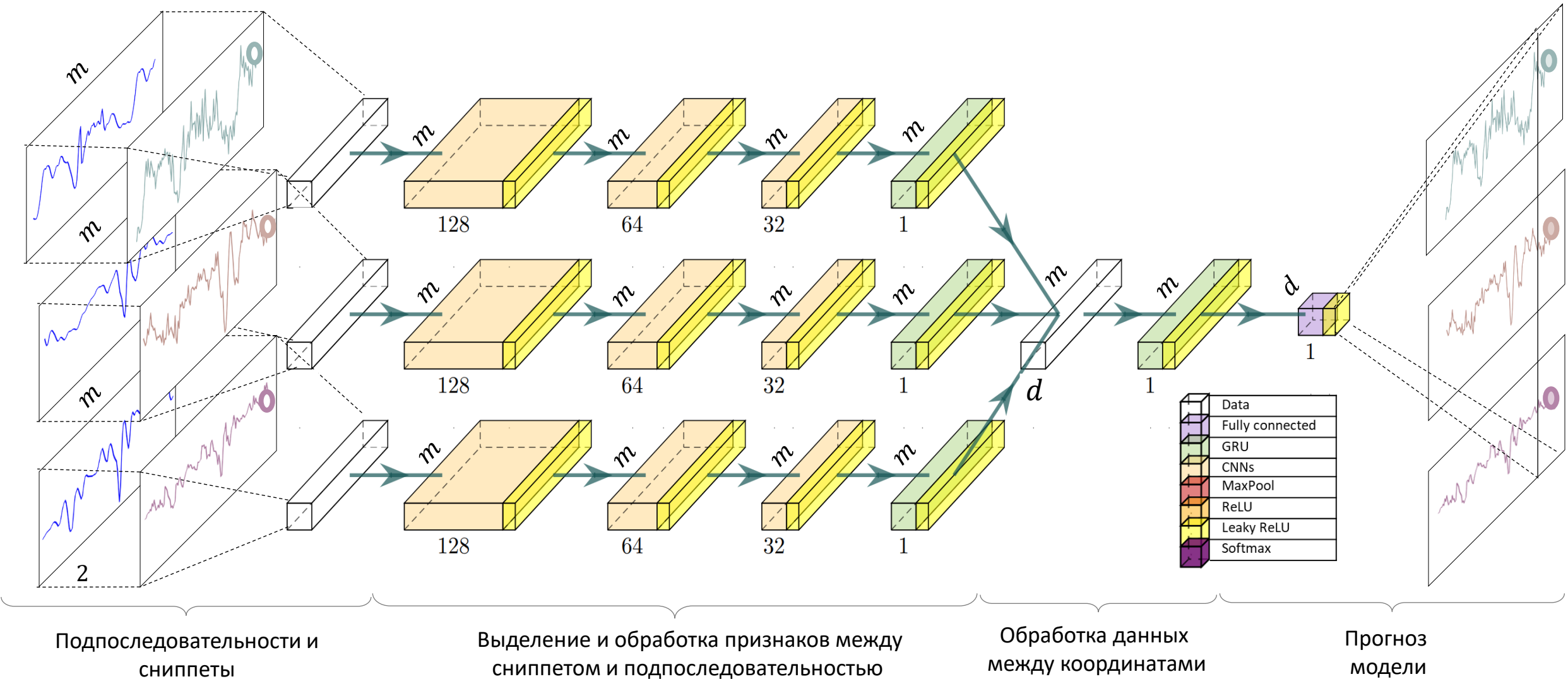
Шаг 2: Аппроксиматор
Подготовка первого приближения поведения



Шаг 3: Восстановление
Правильное использование первого приближения

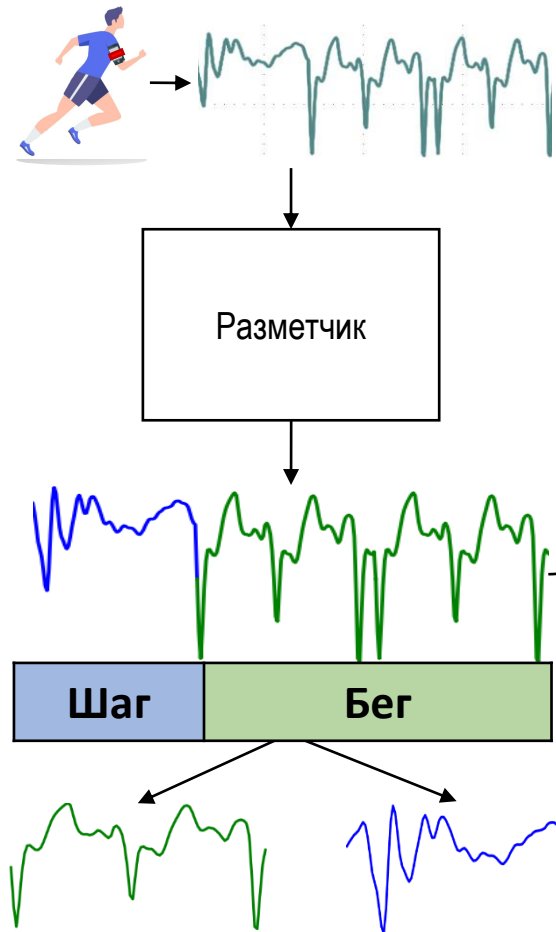


Реконструктор

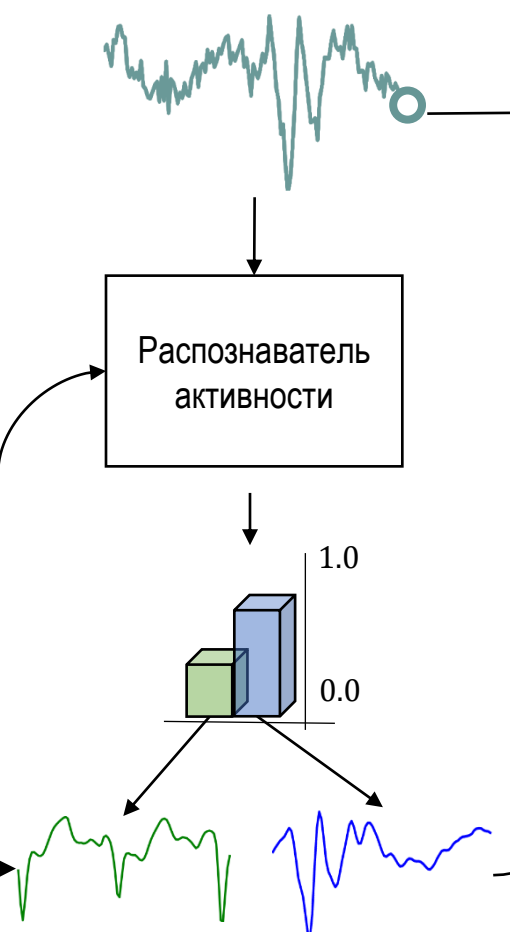


Архитектура предлагаемого решения

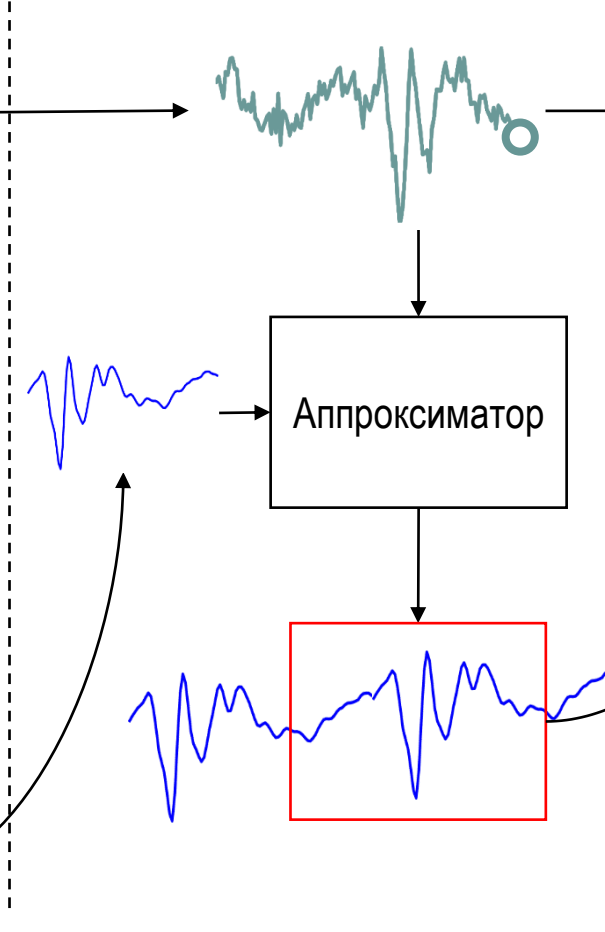
Шаг 0: Предобработка данных
Разметка исходных данных



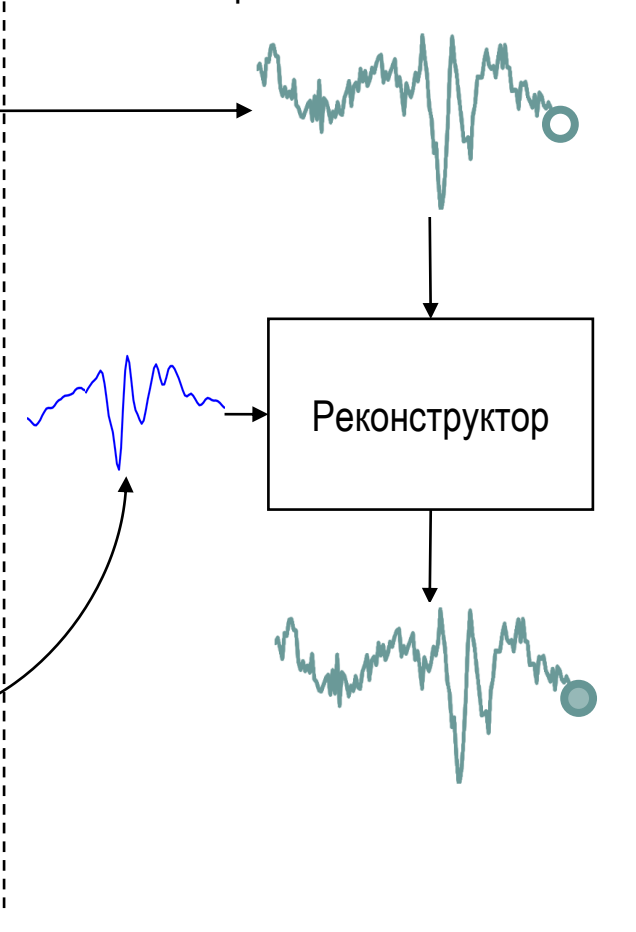
Шаг 1: Подготовка данных
Прогноз текущей активности



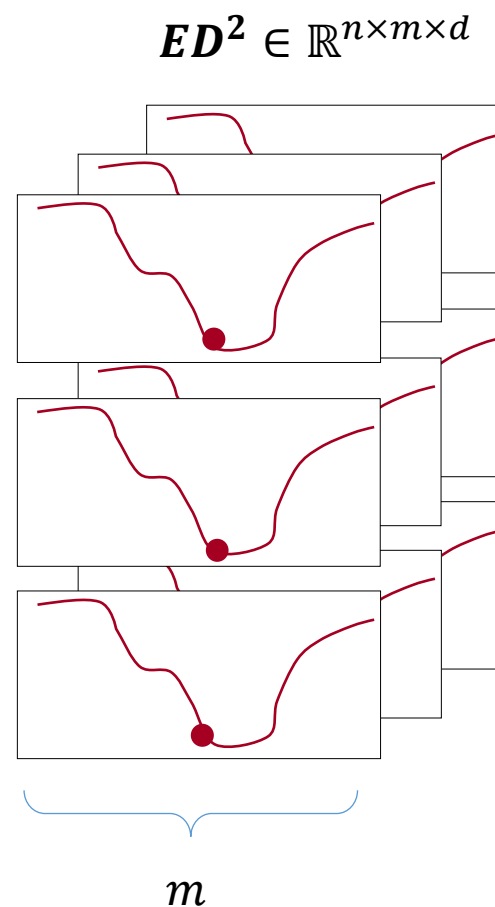
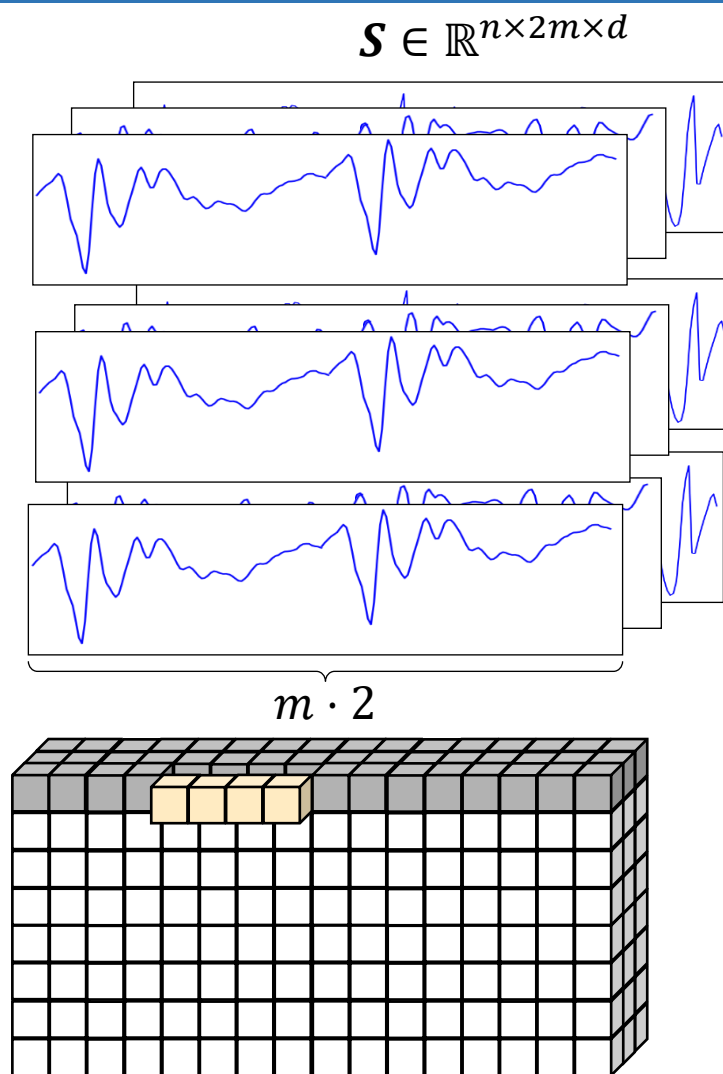
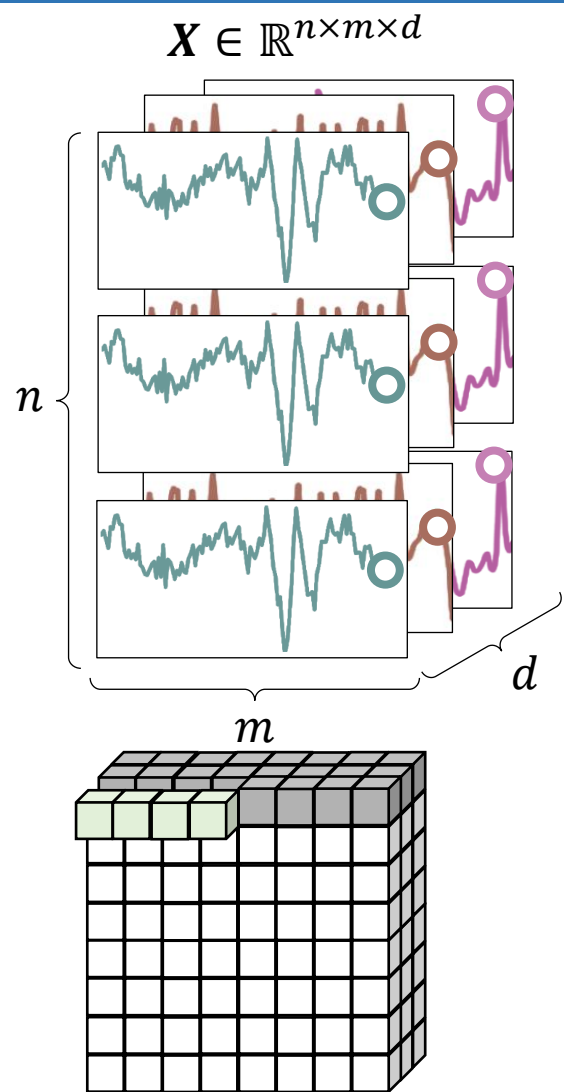
Шаг 2: Аппроксиматор
Подготовка первого приближения поведения



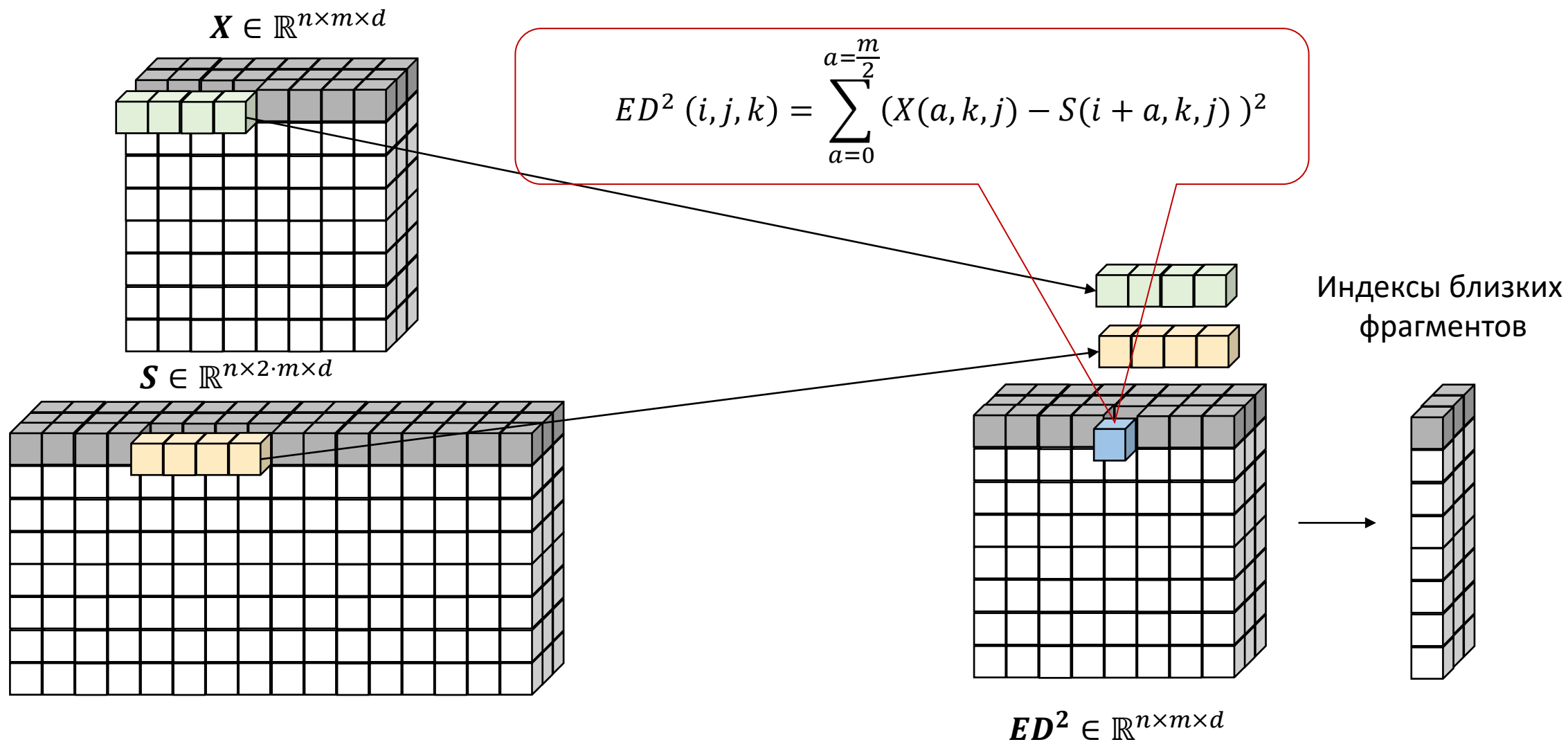
Шаг 3: Восстановление
Правильное использование первого приближения



Параллельный поиск подходящего фрагмента

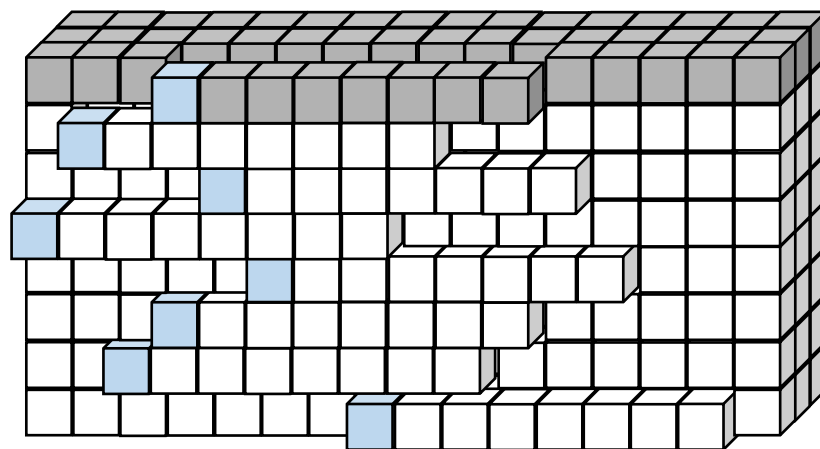
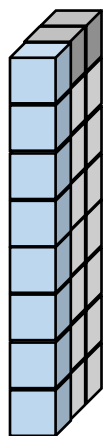


Параллельный поиск подходящего фрагмента

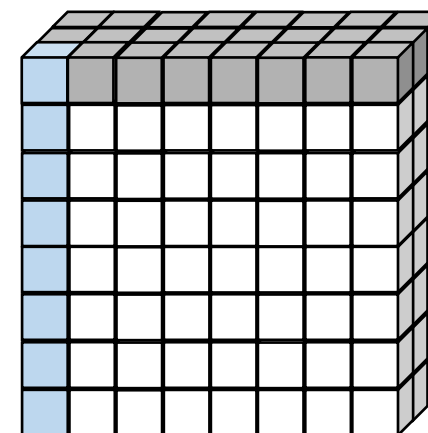


Параллельный поиск подходящего фрагмента

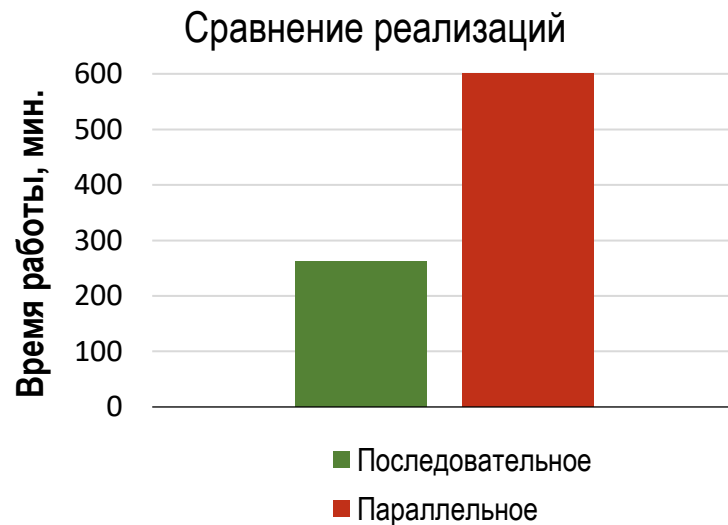
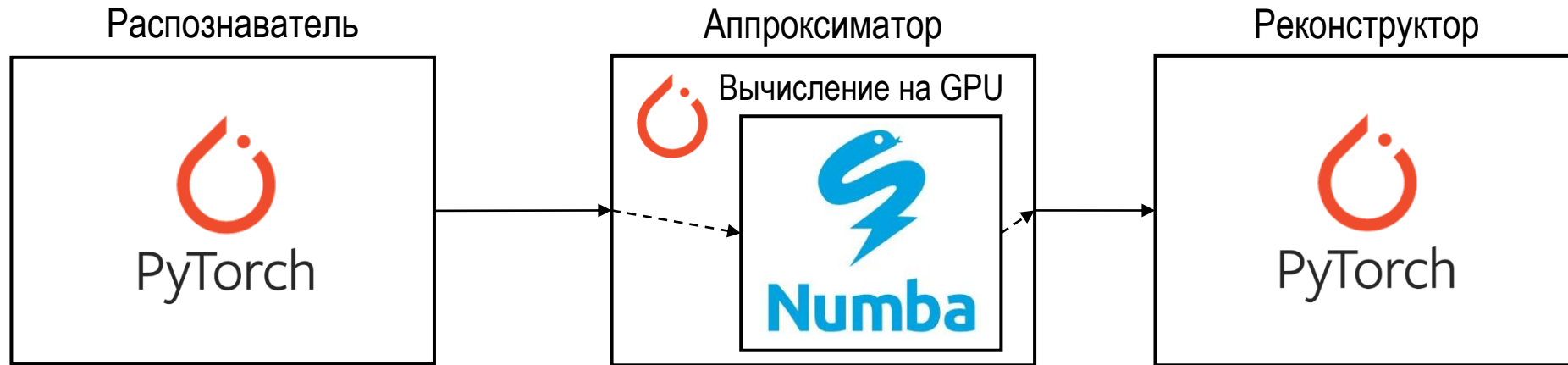
Индексы близких
фрагментов



Сконструированное поведение



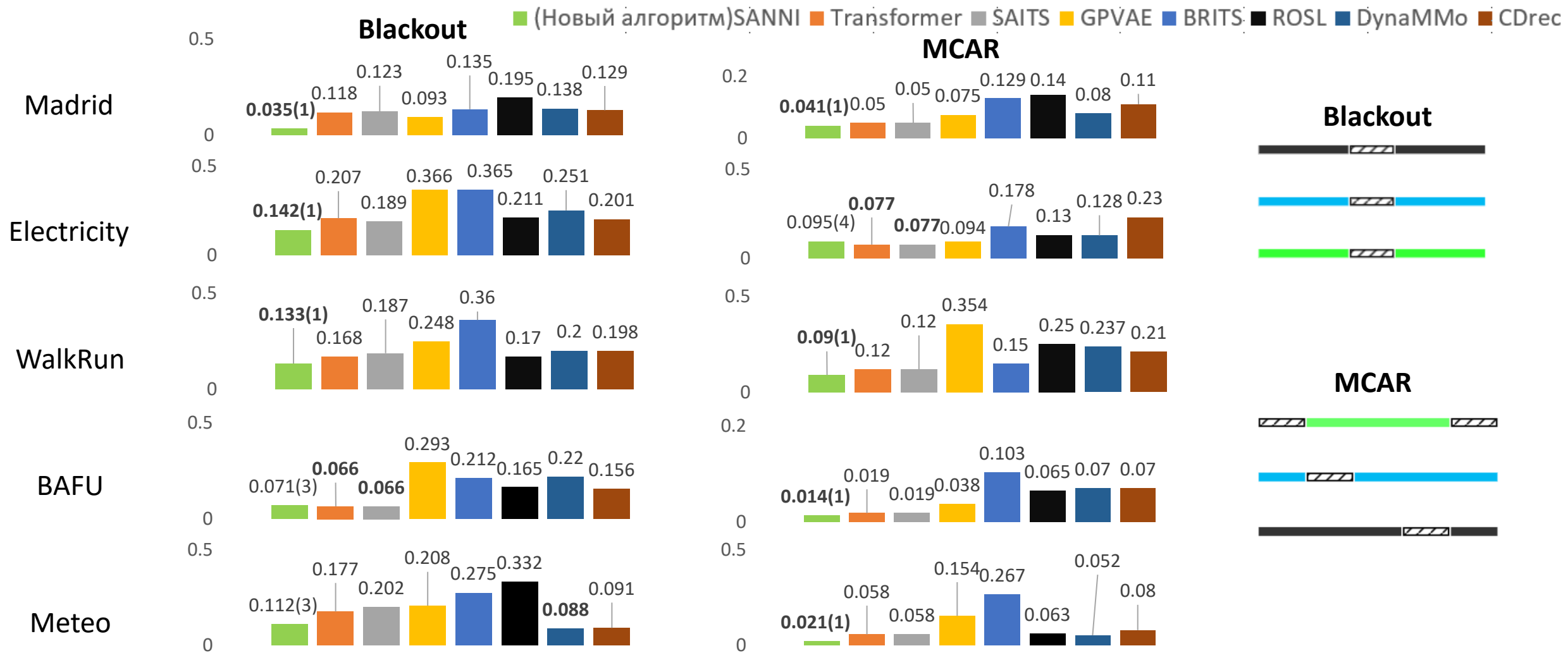
Сравнение параллельной и последовательной части



Параметры обучения

Параметры	Значения
Длина подпоследовательности, t	200
Количество активностей, K	2
Количество координат ряда, d	10
Длина ряда	25 000
Batch size, n	64

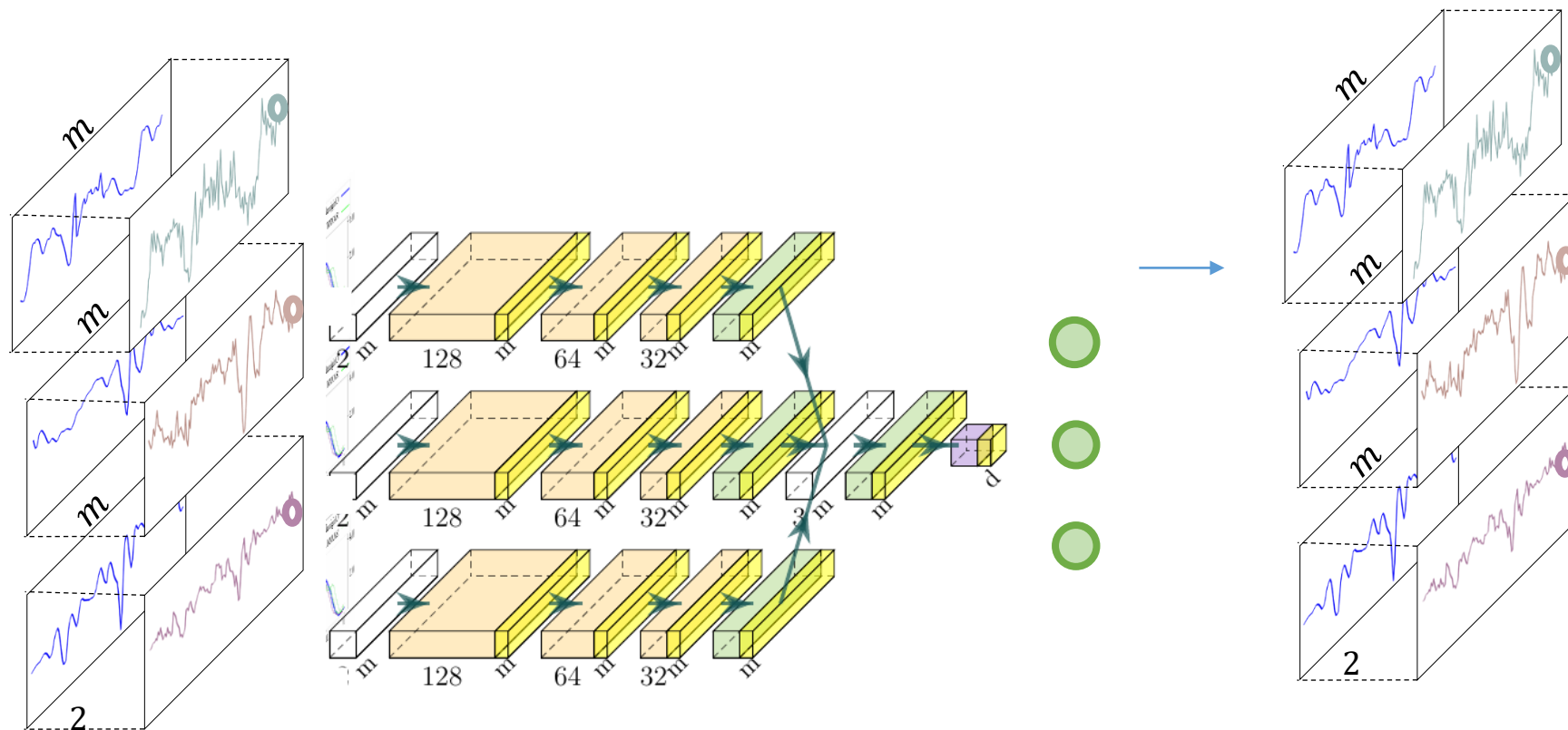
Оценка точности, RMSE



Заключение

- Предложен и реализован новый метод восстановления многомерных временных рядов, основанный на совместном использовании нейросетевых моделей и концепции поведенческих шаблонов (сниппетов)
- Будущие исследования: Исследование применимости концепции поведенческих шаблонов (сниппетов) для иных моделей и методов восстановления временных рядов
- Спасибо за внимание! Вопросы?
Алексей Юртин, yurtinalexei@yandex.ru

Абстрактный метод?

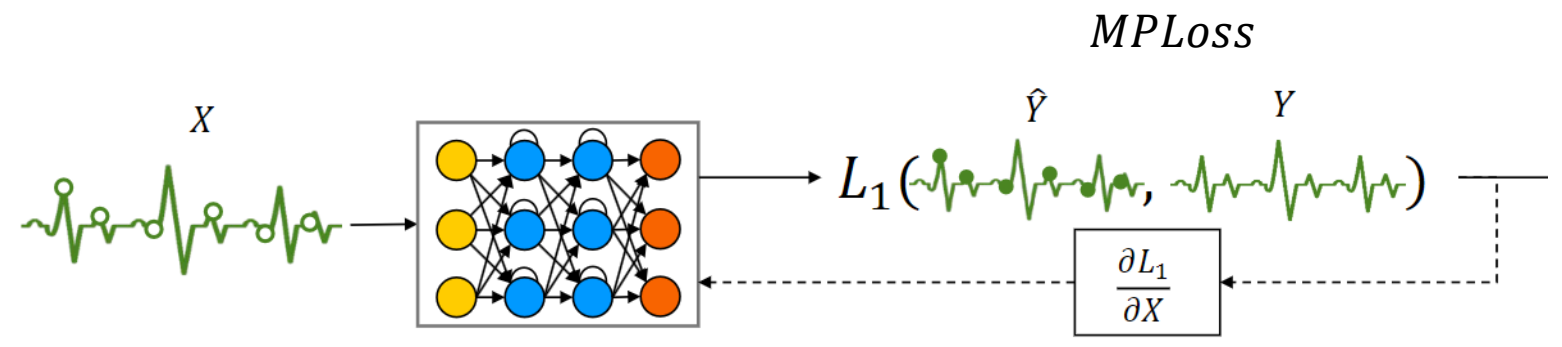
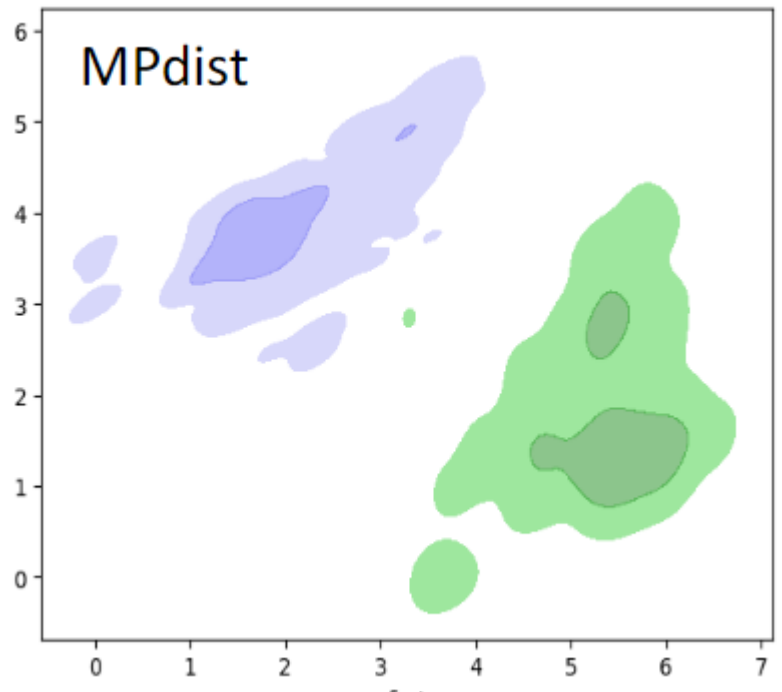
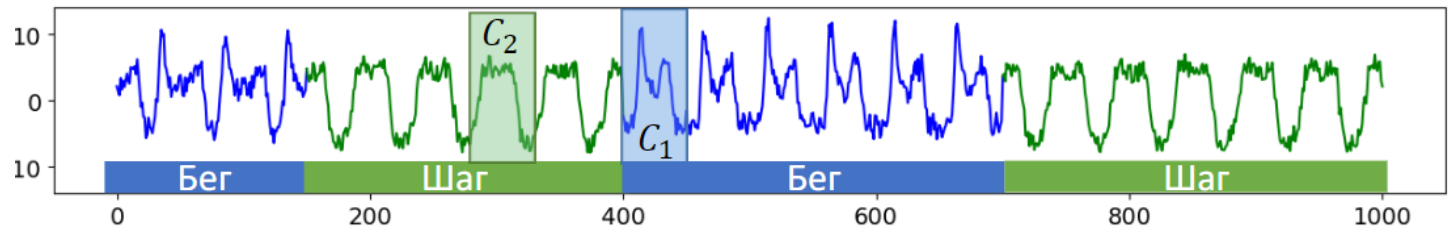


Любая модель/метод
Восстановления
временных рядов

Обязательно ли использовать данную модель?

Конкуrent	Средняя точность	Средняя точность с шаблонами	Прирост точности
SAITS	0,25	0,12	100%

Продолжение исследование (MPloss)



Experiments

- Hardware platform
 - SUSU neural computer 1 x Nvidia Tesla V100
- Dataset
 - 8 datasets covering various subject areas
- Analogues
 - 7 SOTA Methods for imputing missing values in time series data
- Scenario:
 - 2 Scenario of missing value generation (Blackout and MCAR)
- Evaluation method:
 - Recovery accuracy: root mean square error (RMSE)
- SAETI parameters:
 - Subsequence length: 200
 - Number of activities
 - Learning rate: 0.01
 - Optimizer: Adam
 - Number of epochs: 1000
 - Batch size: 64

Datasets

No	Dataset	Length, $n \times 10^3$	Dimension, d	Subject domain
1.	Bafu ¹	50	10	Water discharge in Swiss rivers
2.	BTC ²	2.5	12	Trading characteristics of cryptocurrencies
3.	Climate ³	5	10	Weather in North America
4.	Electricity ⁴	5	9	Power demand of a household in France
5.	Madrid ⁵	25	10	Automatic vehicle registration in Madrid
6.	MAREL ⁶	50	10	Characteristics of seawater in the English Channel
7.	Meteo ⁷	10	20	Weather in Switzerland
8.	Saaleaue ⁸	23	14	Weather in Germany

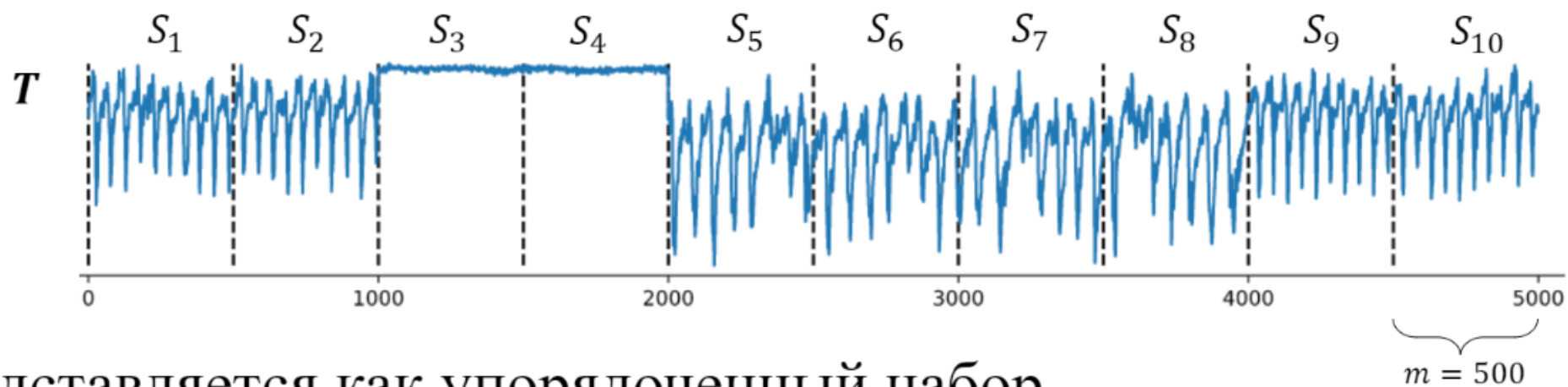
1. Bundesamt für Umwelt – Swiss Federal Office for the Environment <https://www.hydrodaten.admin.ch/>. Accessed: 2023-09-03
2. Historical Crypto Data. <https://www.cryptodatadownload.com/data/>. Accessed: 2023-09-03
3. Lozano, A.C., et al. Spatial-temporal causal modeling for climate change attribution, in: IV, J.F.E., Fogelman Soulié, F., Flach, P.A., Zaki, M.J. (Eds.), Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, June 28 - July 1, 2009, ACM. pp. 587–596. DOI: [10.1145/1557019.1557086](https://doi.org/10.1145/1557019.1557086)
4. Hebrail, G., Berard, A., 2012. Individual household electric power consumption. UCI Machine Learning Repository. DOI: [10.24432/C58K54](https://doi.org/10.24432/C58K54). accessed: 2023-09-03
5. Lañna, I., et al., 2018. On the imputation of missing data for road traffic forecasting: New insights and novel techniques. Transportation Research Part C: Emerging Technologies 90, 18–33. URL: DOI: [10.1016/j.trc.2018.02.021](https://doi.org/10.1016/j.trc.2018.02.021).
6. Lefebvre, A., 2015. MAREL Carnot data and metadata from Coriolis Data Centre. SEANOE. URL: DOI: [10.17882/39754](https://doi.org/10.17882/39754). accessed: 2023-09-03.
7. MeteoSwiss. (Year of dataset). [Title of dataset]. Federal Office of Meteorology and Climatology. <https://www.meteoswiss.admin.ch/> Accessed: 2023-09-03
8. Weather Station Saaleaue, Max Planck Institute for Biogeochemistry, Germany. https://www.bgc-jena.mpg.de/wetter/weather_data.html. Accessed: 2023-09-03

Конкуренты

№	Model	Basic principle
Analytical		
1.	CDrec ¹	Centroid Decomposition
2.	DynaMMo ²	Калман и EM
3.	ROSL ³	PCA
Neural network-based		
4.	BRITS ⁴	Bidirectional recurrent neural network
5.	GPVAE ⁵	Combination of Autoencoders and Gaussian Process
6.	SAITS ⁶	Transformer
7.	Transformer ⁶	Transformer

1. Khayati M., Cudré-Mauroux P., Böhlen M. H. Scalable recovery of missing blocks in timeseries with high and low cross-correlations // Knowledge and Information Systems. – 2020. – Т. 62. – С. 2257-2280
2. Li L. et al. Dynammo: Mining and summarization of coevolving sequences with missing values // Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. – 2009. – С. 507-516
3. Shu X., Porikli F., Ahuja N. Robust orthonormal subspace learning: Efficient recovery of corrupted low-rank matrices // Proceedings of the IEEE conference on computer vision and pattern recognition. – 2014. – С. 3874-3881
4. Cao W., Wang D., Li J., et al. BRITS: Bidirectional recurrent imputation for time series // Proc. 32nd Conf. on Neural Inf. Proc. Systems (NeurIPS 2018), Montréal, Canada, 2018. https://proceedings.neurips.cc/paper_files/paper/2018/file/734e6bfcd358e25ac1db0a4241b95651-Paper.pdf. Cited June 18, 2023.
5. Fortuin V. et al. Gp-vae: Deep probabilistic time series imputation // International conference on artificial intelligence and statistics. – PMLR, 2020. – С. 1651-1661.
6. Du W., et al. SAITS: self-attention-based imputation for time series // Expert Syst. Appl. 2023. 219, Article Number 119619. DOI: [10.1016/j.eswa.2023.119619](https://doi.org/10.1016/j.eswa.2023.119619)

Сниппеты



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

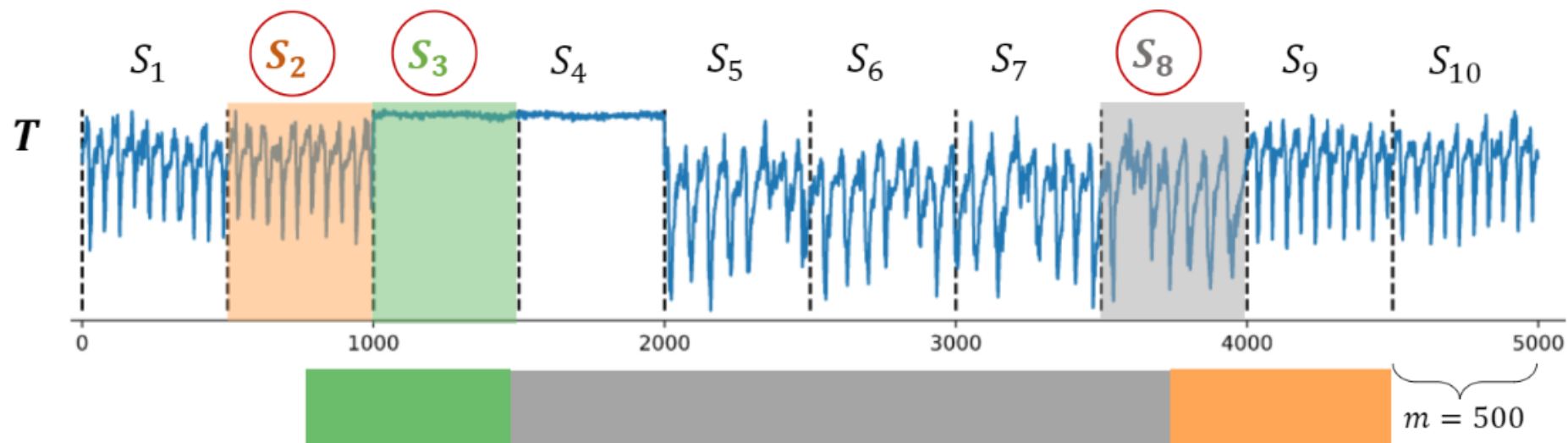
$$S = \{S_i\}_{i=1}^{n/m}, S_i = T_{m(i-1)+1, m}$$

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

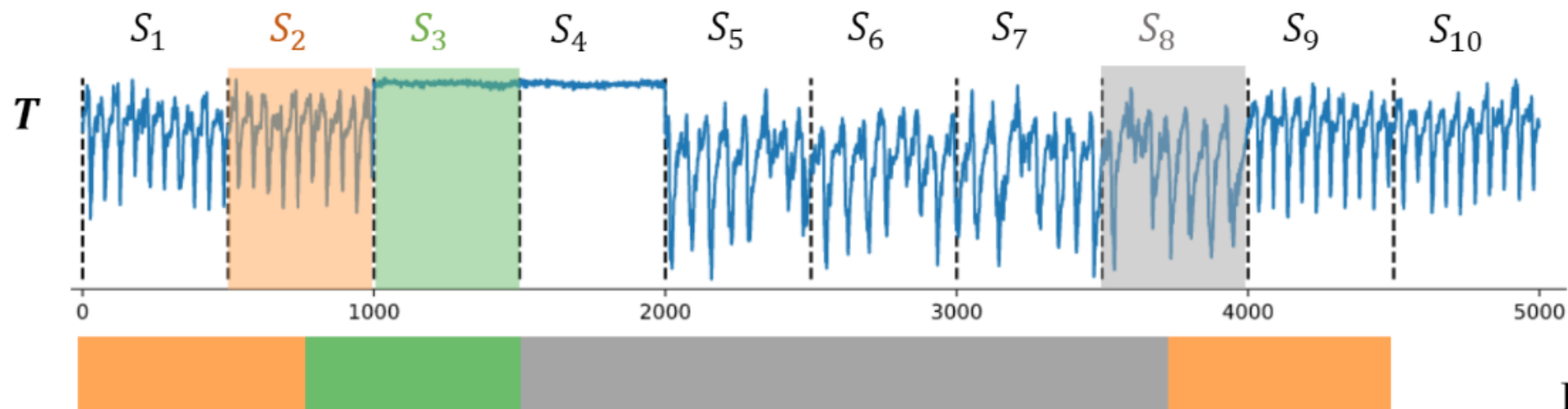


1. Ряд представляется как упорядоченный набор непересекающихся сегментов заданной длины
2. Для каждого сегмента найдем его ближайших соседей в смысле MPdist

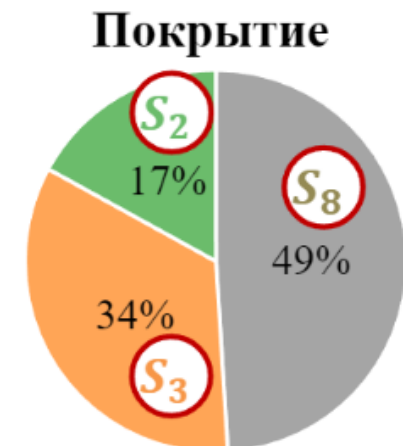
Imani S. *et al.* Introducing time series snippets: a new primitive for summarizing long time series. *Data Min. Knowl. Discov.* 2020. 34(6). 1713–1743. DOI: [10.1007/s10618-020-00702-y](https://doi.org/10.1007/s10618-020-00702-y)

Сниппеты

$K = 3$



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)