HUAWEI System Software Workshop 2022
19 October 2022, Moscow, Russia

Efficient time series analytics

through DBMSs, ANNs, and parallel algorithms

明智是了解事件的人
Wise is the person who understands events.
Chinese proverb

Mikhail Zymbler, Yana Kraeva, Andrey Goglachev, Alexey Yurtin

{mzym, kraevaya, goglachevai, iurtinaa}@susu.ru

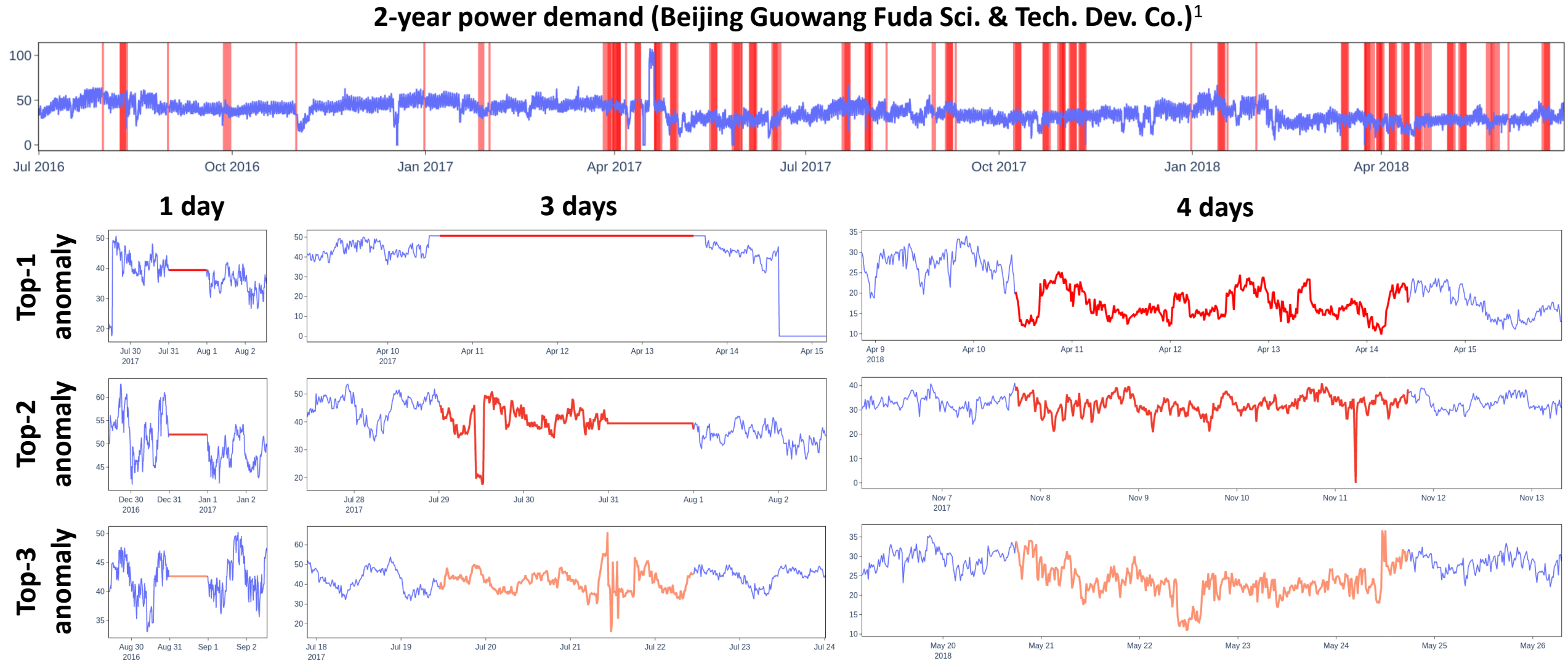
Big Data and Machine Learning Lab, South Ural State University, Chelyabinsk, Russia

This work was financially supported by the Russian Foundation for Basic Research (grant No. 17-07-00463),
and by the Ministry of Science and Higher Education of the Russian Federation (government order FENU-2020-0022)

Outline

- **Introduction**
  - Insights we can understand from time series
  - Offline and online time series analytics
- Offline time series analytics
  - Embedding time series analytics into PostgreSQL
- Online time series analytics
  - Employing ANNs together with parallel algorithms
  - Parallel time series labeling
  - Online imputation of missing values and anomaly detection
  - Parallel time series anomaly discovery
- Conclusions

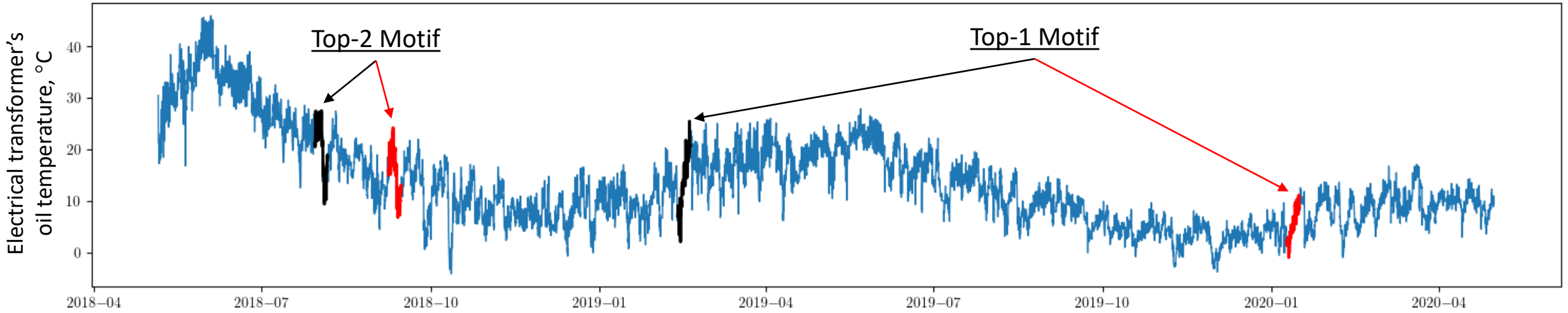
Insights from time series: Anomalies



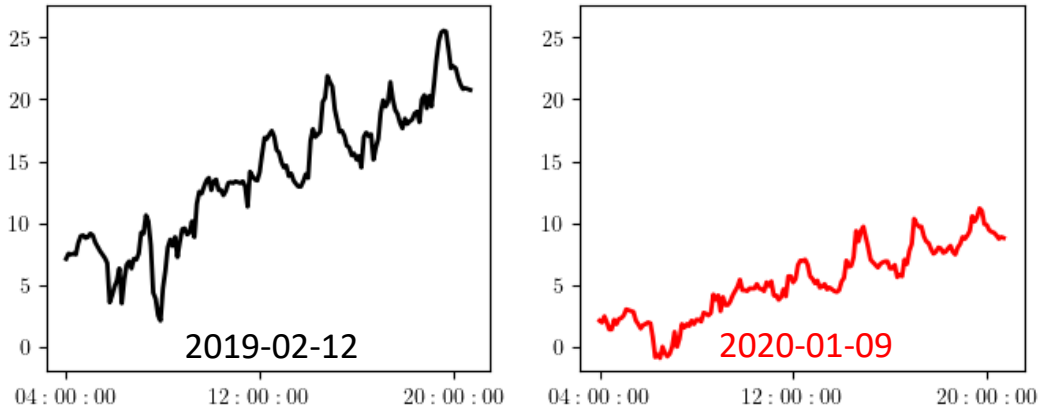
[1] Zhou H. *et al.* Informer: beyond efficient transformer for long sequence time-series forecasting. AAAI 2021: 11106-11115. DOI: 10.1609/aaai.v35i12.17325.

Insights from time series: Motifs

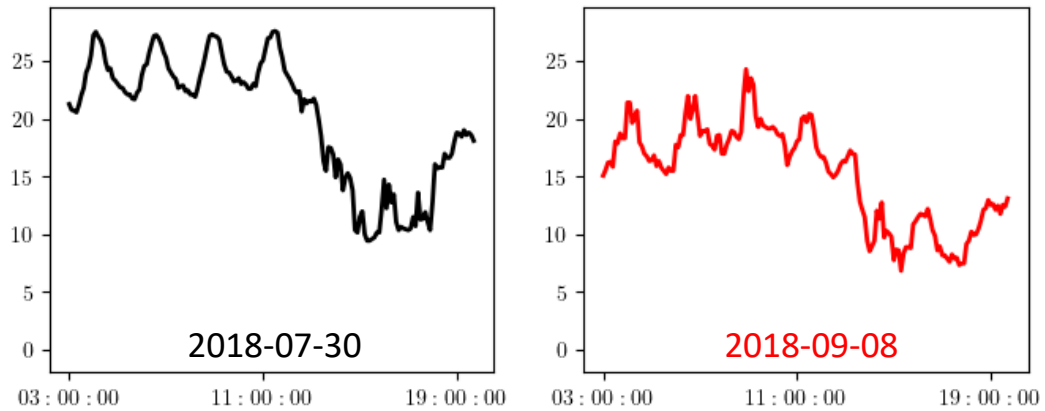
2-year power demand (Beijing Guowang Fuda Sci. & Tech. Dev. Co.)[1]



Top-1 Motif



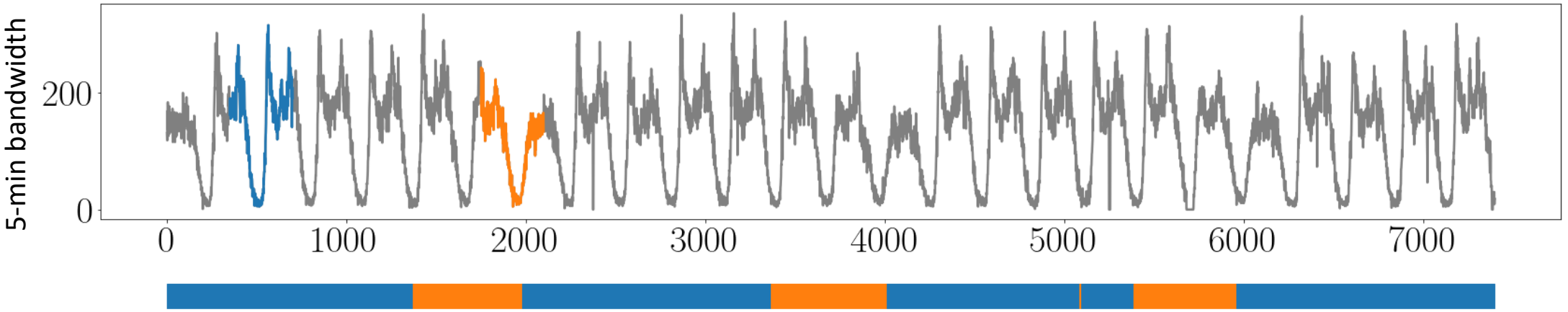
Top-2 Motif



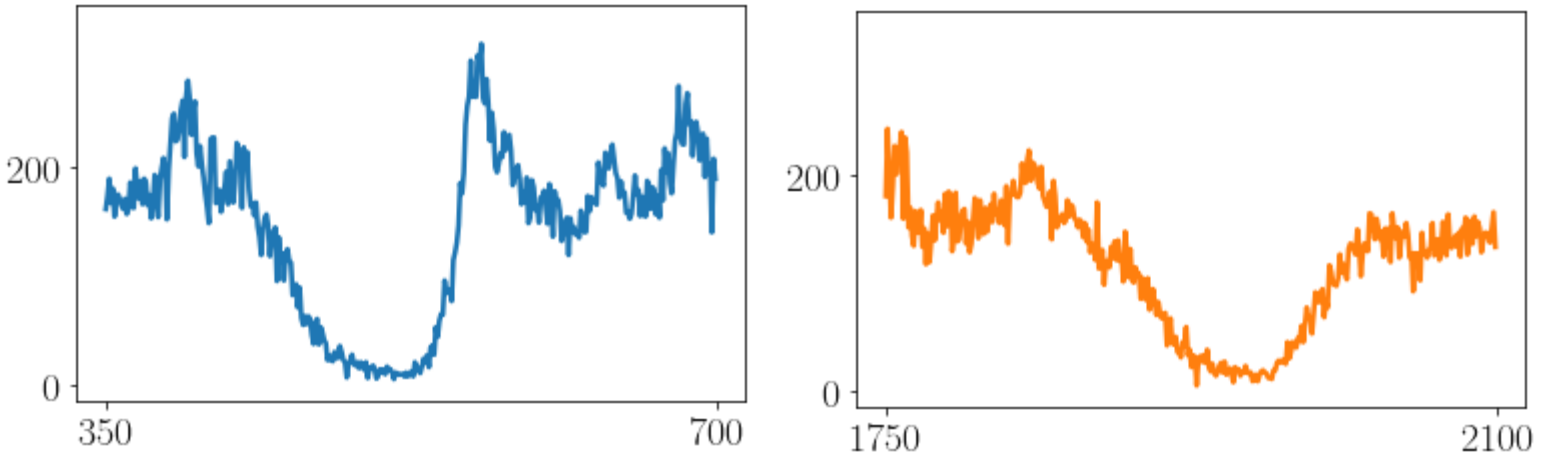
[1] Zhou H. *et al.* Informer: beyond efficient transformer for long sequence time-series forecasting. AAAI 2021: 11106-11115. DOI: 10.1609/aaai.v35i12.17325.

Insights from time series: Patterns

One-month urban traffic in Munich (gathered by the Huawei Research Center) [1]

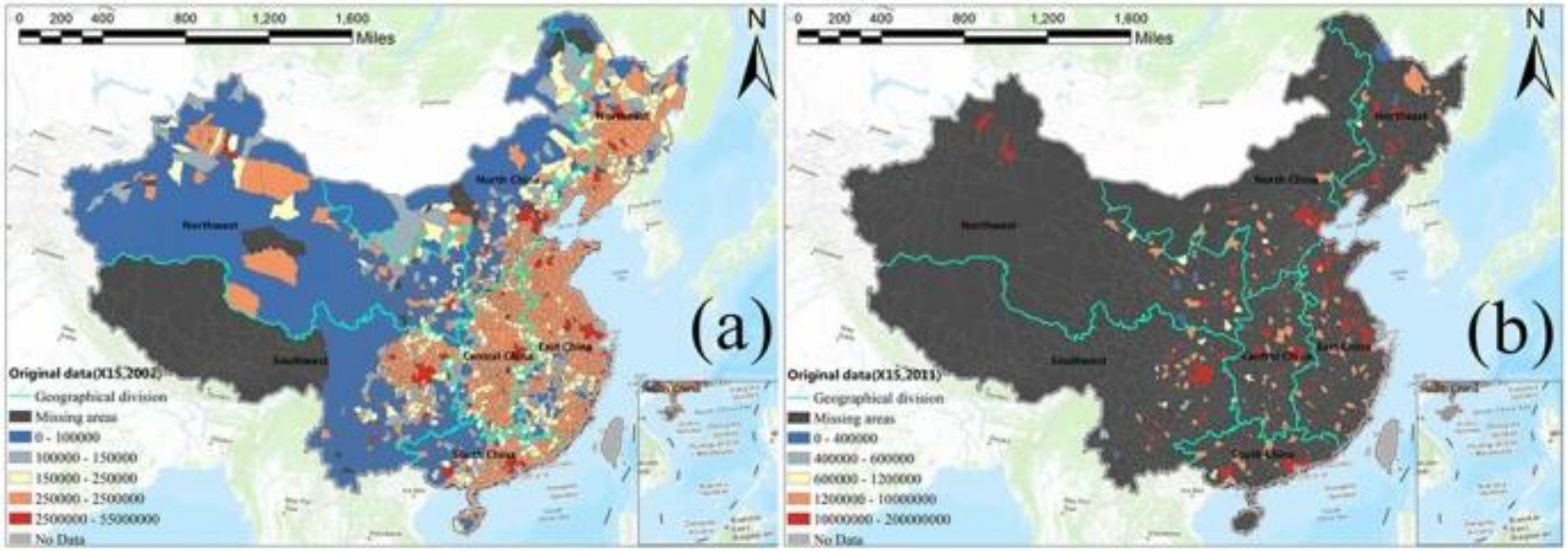


Patterns found



[1] Public (anonymized) road traffic prediction datasets from Huawei Munich Research Center. URL: https://zenodo.org/record/3653880#.Y0zZi3ZBxPa

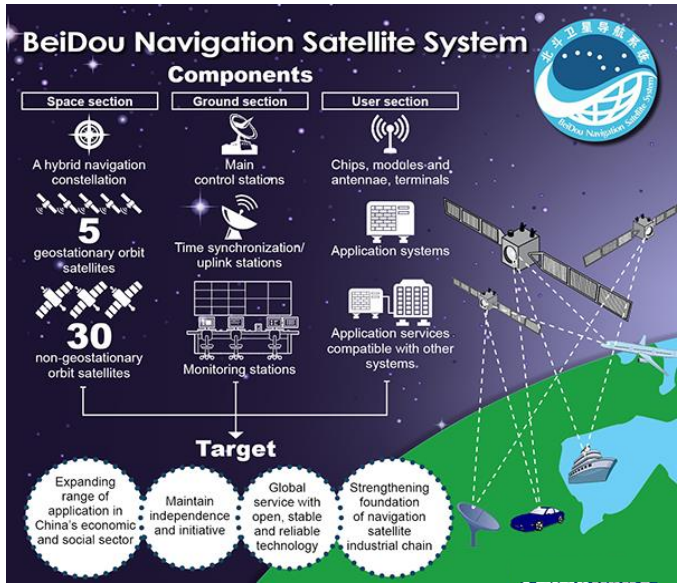
Not an insight, but a headache to get rid of: missing values



China counties with missing official statistical data (one attribute)[1]

a) 2002: less than 15% data are missing

b) 2011: more than 85% data are missing

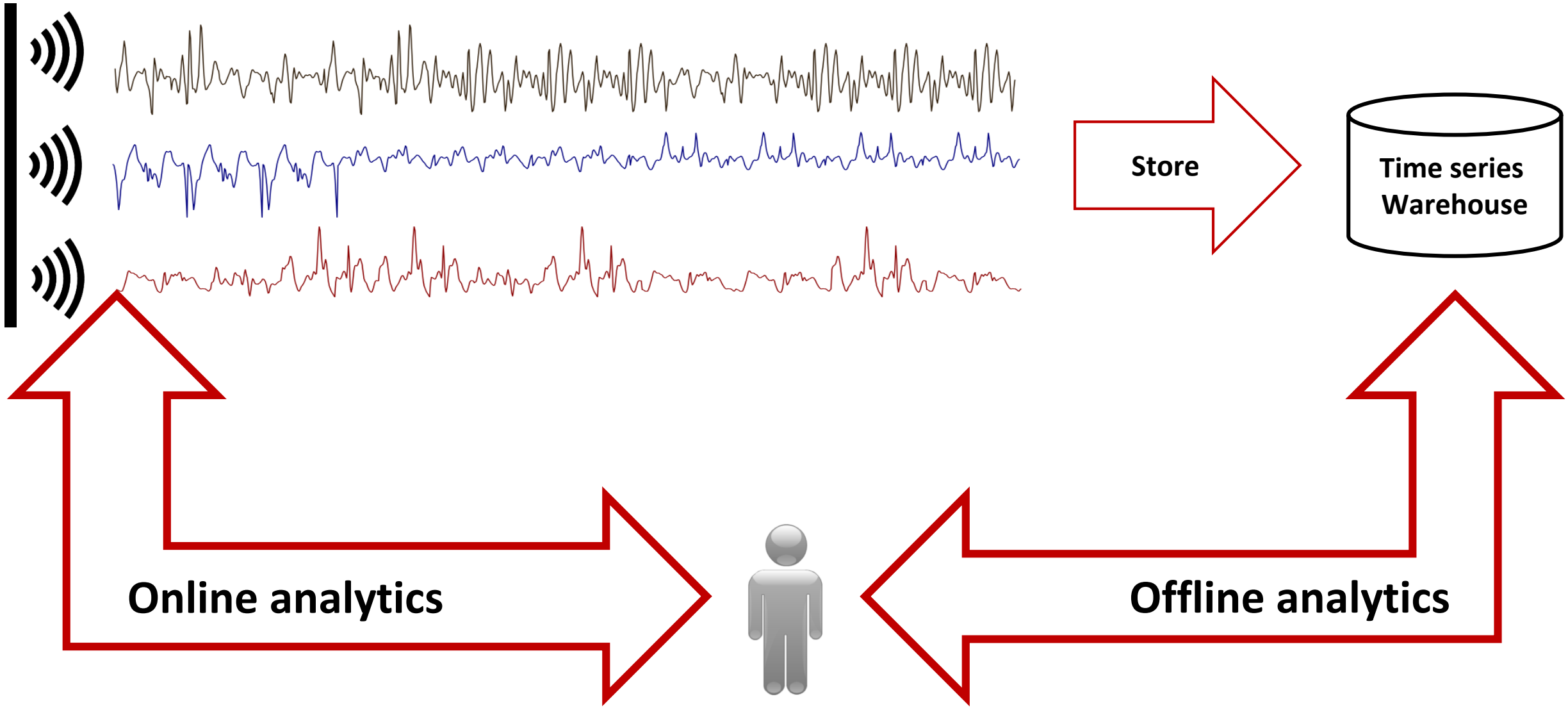


BeiDou satellite transmission link suffers the packet loss[2]

[1] Song C. *et al.* Estimating missing values in China's official socioeconomic statistics using progressive spatiotemporal Bayesian hierarchical modeling. Sci. Rep. 2018. Vol. 8, article 10055. DOI: 10.1038/s41598-018-28322-z

[2] Liu S. *et al.* A novel BeiDou satellite transmission framework with missing package imputation applied to smart ships. IEEE Sensors Journal. 2022. Vol. 22, no. 13. P. 13162-13176. DOI: 10.1109/JSEN.2022.3177167.

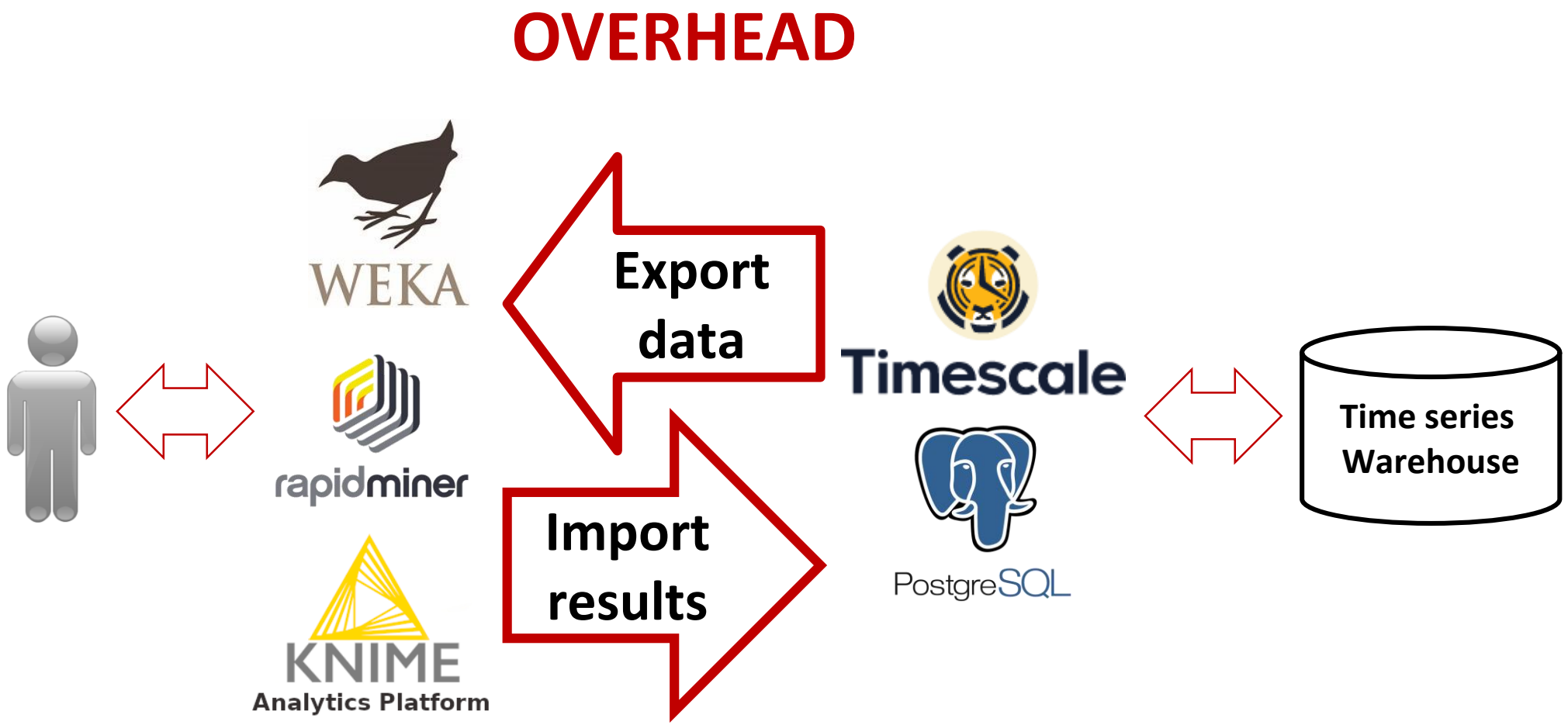
Time series analytics: online *vs.* offline



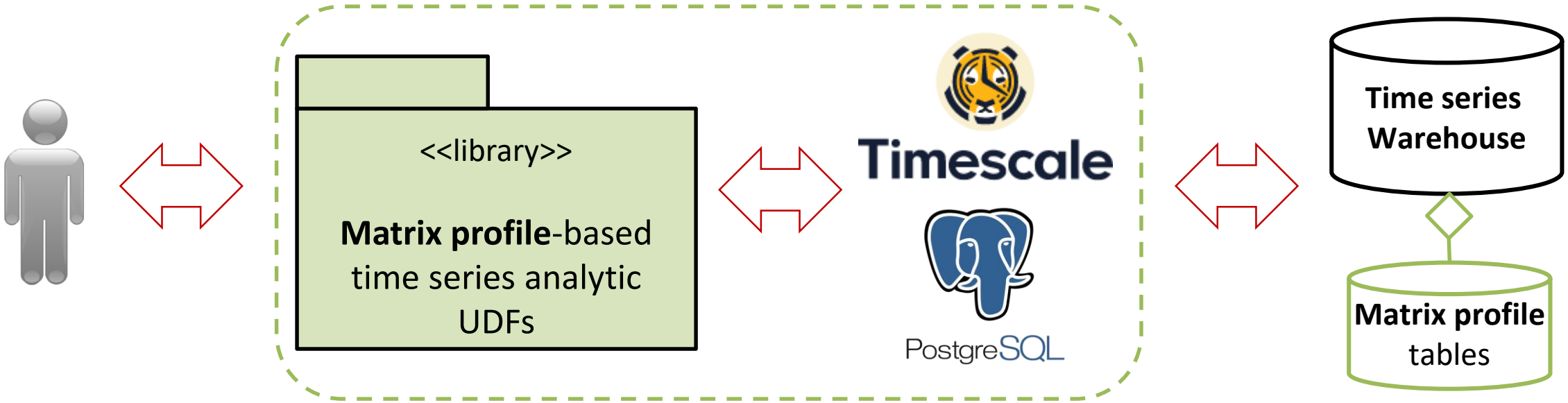
Outline

- Introduction
- **Offline time series analytics**
  - Embedding time series analytics into PostgreSQL
- Online time series analytics
  - Employing ANNs together with parallel algorithms
  - Parallel time series labeling
  - Online imputation of missing values and anomaly detection
  - Parallel time series anomaly discovery
- Conclusions

What's wrong with **offline** time series analytics?

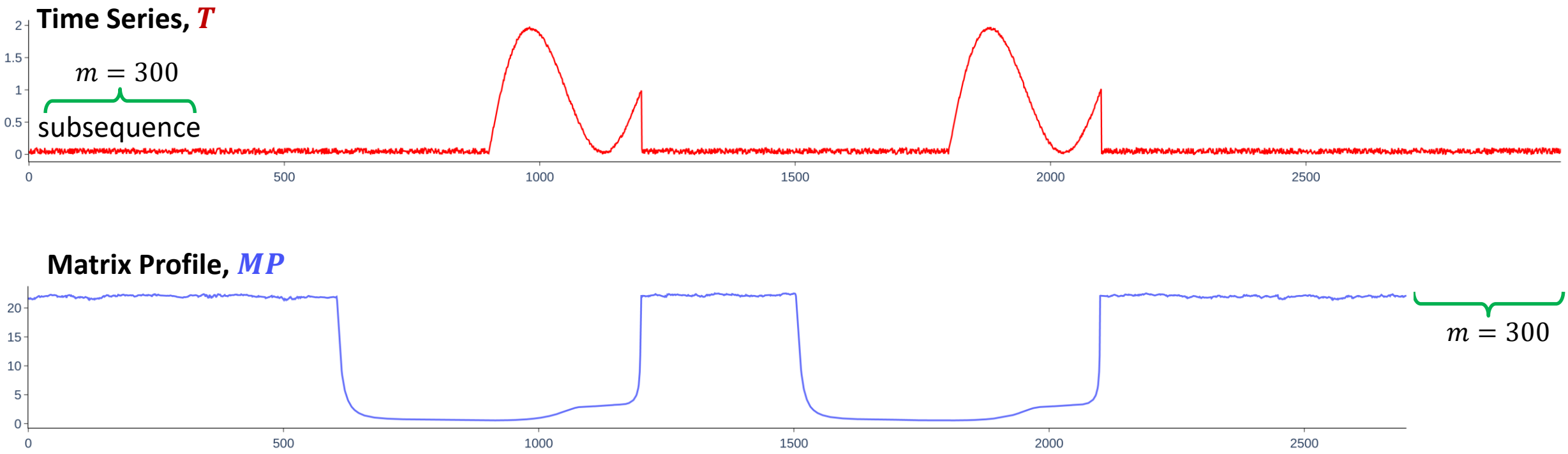


Offline time series analytics **inside** a DBMS



骑驴找马
Looking for a horse, ride a donkey.
Chinese proverb

Matrix profile[1,2] of a time series



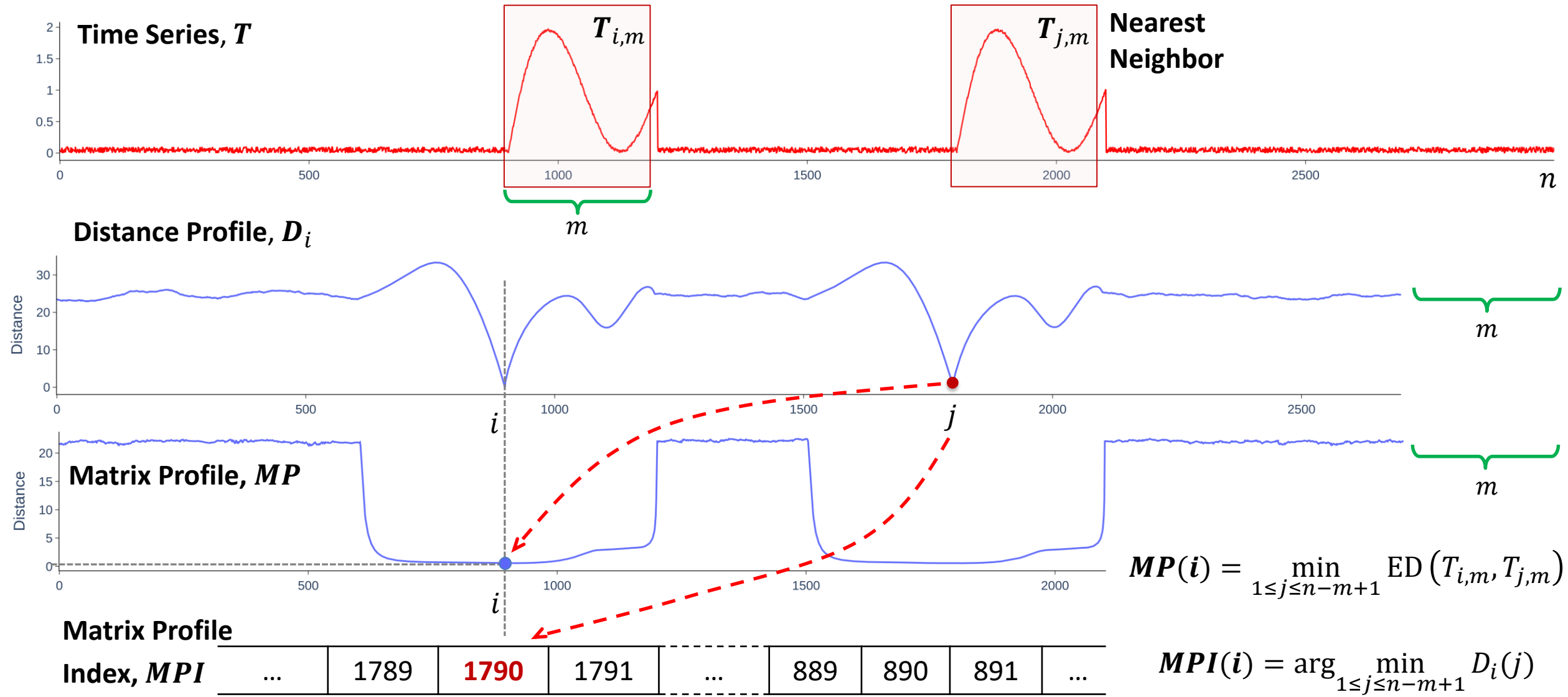
$$MP(i) = \min_{1 \le j \le n-m+1} \mathbf{ED}\left(T_{i,m}, T_{j,m}\right)$$

Matrix profile's i-th element is the **Euclidean distance** between i-th subsequence and its nearest non-overlapping neighbor

[1] Yeh C.M. *et al.* Time series joins, motifs, discords and shapelets: A unifying view that exploits the matrix prole. Data Min. Knowl. Discov. 32(1), 83-123 (2018). DOI: 10.1007/s10618-017-0519-9

[2] Zimmerman Z. *et al.* Matrix Profile XIV: Scaling time series motif discovery with GPUs to break a quintillion pairwise comparisons a day and beyond. SoCC 2019. DOI: 10.1145/3357223.3362721

Matrix profile of a time series

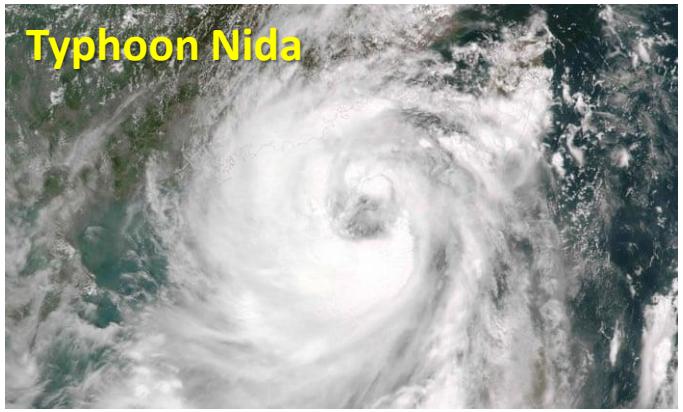


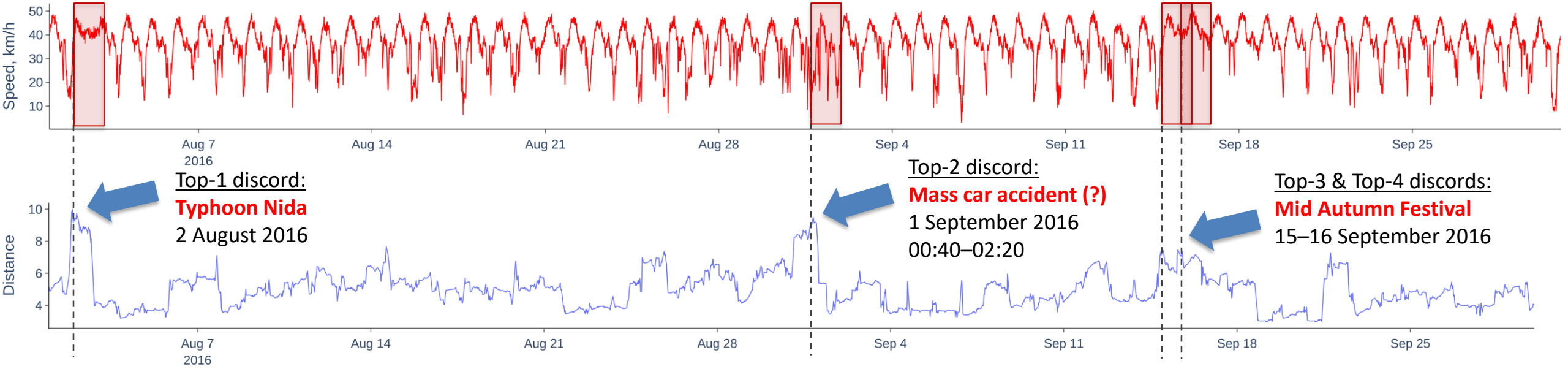
$$\boldsymbol{MP}(\boldsymbol{i}) = \min_{1 \le j \le n-m+1} \mathrm{ED}\left(T_{i,m}, T_{j,m}\right)$$

$$\boldsymbol{MPI}(\boldsymbol{i}) = \arg\min_{1 \le j \le n-m+1} D_i(j)$$

Insights from the Matrix profile: Discords

Urban traffic speed in Guangzhou[1]

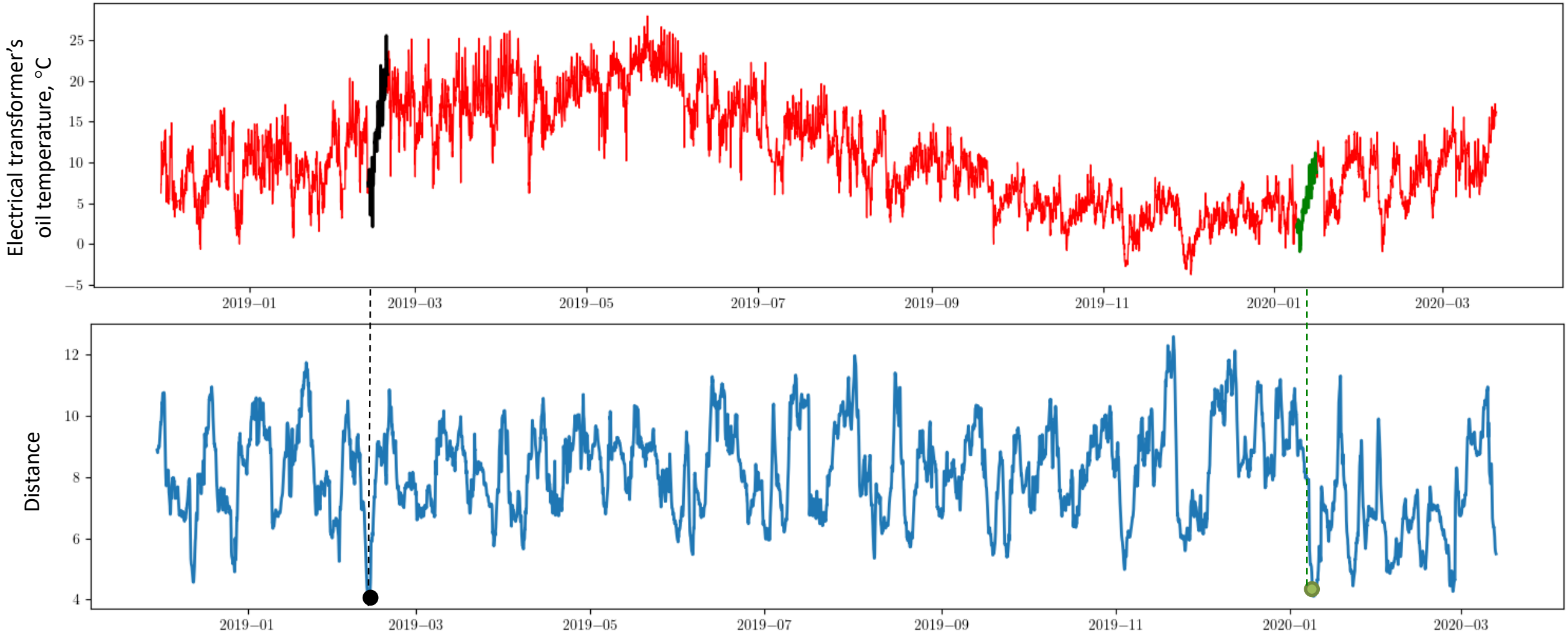




[1] Chen X, Chen Y, He Z. Urban traffic speed dataset of Guangzhou, China. 2018. DOI: 10.5281/zenodo.1205229.

Insights from the Matrix profile: Motifs

2-year power demand (Beijing Guowang Fuda Sci. & Tech. Dev. Co.)[1]



[1] Zhou H. *et al.* Informer: beyond efficient transformer for long sequence time-series forecasting. AAAI 2021: 11106-11115. DOI: 10.1609/aaai.v35i12.17325.

Matrix profile-oriented warehouse

TIME SERIES DATA

Time Series Directory	
ID	BIGSERIAL
name	TEXT
len	BIGINT

ts_name (a time series)	
num	BIGSERIAL
stamp	TIMESTAMP
val	REAL

MATRIX PROFILE DATA

Matrix Profile Directory	
ID	BIGSERIAL
tsID*	BIGINT
subseqLen	INT

mp_name_subseqLen (a matrix profile)	
num	BIGSERIAL
nnDist	REAL
nnIdx	BIGINT

TIME SERIES ANALYTICS API

discoverDiscords (**name** TEXT, **subseqLen** INT)

nnDist	REAL
idx	BIGINT
discord	REAL[subseqLen]

discover … (**name** TEXT, **subseqLen** INT)

…	…

…

discoverMotifs (**name** TEXT, **subseqLen** INT)

nnDist	REAL
idxLeft	BIGINT
motifLeft	REAL[subseqLen]
idxRight	BIGINT
motifRight	REAL[subseqLen]

matrixProfile (**name** TEXT, **subseqLen** INT)

num	BIGSERIAL
nnDist	REAL
nnIdx	BIGINT

Time series representation

TIME SERIES DATA

Time Series Directory	
ID	BIGSERIAL
name	TEXT
len	BIGINT

ts_name (a time series)	
num	BIGSERIAL
stamp	TIMESTAMP
val1	REAL

Time Series Directory

ID	name	len
1	DailyExchangeRate	5000
2	MinutelyPulse	9000

ts_DailyExchangeRate

num	stamp	val
1	10.02.2008	23.35
2	11.02.2008	24.05
…	…	…
5000	19.10.2022	63.63

ts_MinutelyPulse

num	stamp	val
1	10.10.2022 00:01	135
2	10.10.2022 00:02	105
…	…	…
9000	16.10.2022 06:00	139

Matrix profile representation

MATRIX PROFILE DATA

Matrix Profile Directory	
ID	BIGSERIAL
tsID*	BIGINT
subseqLen	INT

mp_name_subseqLen (Matrix Profile)	
num	BIGSERIAL
nnDist	REAL
nnIdx	BIGINT

mp_DailyExchangeRate_7

num	nnDist	nnIdx
1	0.02	2071
2	0.05	2078
…	…	…
4994	12.22	184

mp_DailyExchangeRate_30

num	nnDist	nnIdx
1	346.32	3546
2	358.15	4278
…	…	…
4971	0.10	1286

Time Series Directory

ID	name	len
1	DailyExchangeRate	5000
2	MinutelyPulse	9000

Matrix Profile Directory

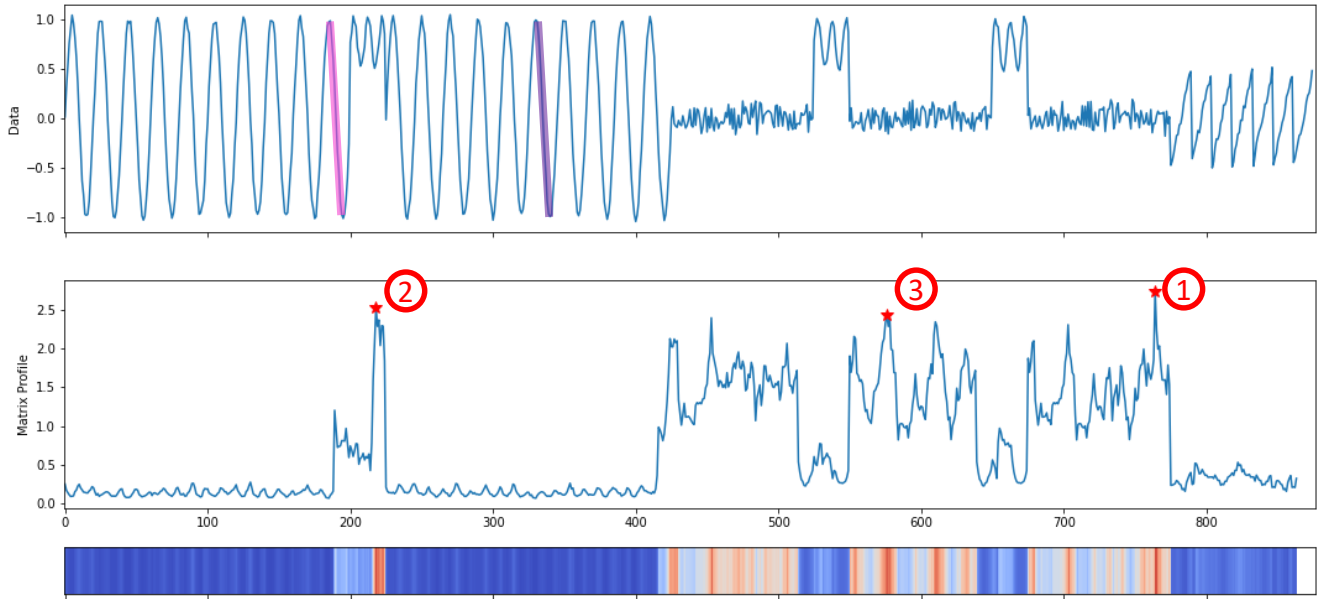
ID	tsID	subseqLen	
1	1	7	weekly
2	1	30	monthly
3	2	60	hourly

mp_MinutelyPulse_60

num	nnDist	nnIdx
1	23.04	145
2	39.78	7658
…	…	…
8941	11.65	7856

Time series analytics API

SELECT matrixProfile (sensorData, 12);



ts_sensorData

mp_sensorData_12

SELECT discoverDiscords (sensorData, 12, 3);

NNDIST	IDX	DISCORD
2.75 ①	764	[-0.05, 0.02, 0.06, 0.04, 0.04, 0.04, 0.05, -0.18, 0.06, -0.05, 0.01, -0.47]
2.53 ②	218	[0.55, 0.64, 0.74, 0.87, 1.03, 0.98, 0.85, -0.01, 0.35, 0.61, 0.81, 0.96]
2.43 ③	576	[0.05, -0.16, 0.01, -0.08, -0.09, -0.02, 0.05, -0.03, 0.01, -0.01, -0.06, -0.08]

SELECT discoverMotifs (sensorData, 12, 1);

NNDIST	IDXLEFT	MOTIFLEFT	IDXRIGHT	MOTIFRIGHT
0.069	185	[0.97,0.98,0.79,0.59,0.31,-0.01, -0.34,-0.54,-0.85,-0.97,-1.01,-0.96]	330	[0.99,0.97,0.82,0.60,0.32,-0.02, -0.32,-0.56,-0.84,-0.97,-0.99,-0.98]

Outline

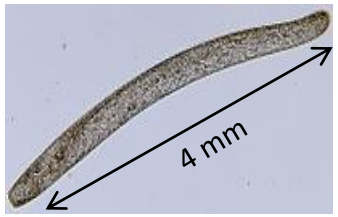
- Introduction
- Offline time series analytics
- **Online time series analytics**
  - Employing ANNs together with parallel algorithms
  - Parallel time series labeling
  - Online anomaly detection and imputation of missing values
  - Parallel time series anomaly discovery
- Conclusions

What's wrong with **online** time series analytics?

- ☺ No data labeling and model learning overheads
- ☹ Fast, but not enough to fit online analytics
- ☹ Accurate, but not like ANNs

Method	Labeling ⇨ Learning	⇨ Tuning	⇨ Mining
Parallel algorithms	子曰、攻乎异端、斯害也已。 Learning the wrong views is harmful Confucius 	 Turtle **15-35 km/h**, running/swimming	 Cheetah **100 km/h**

- ☺ Fast to fit online analytics
- ☺ High accuracy
- ☹ Data labeling and model learning overheads

Method	Labeling	⇨ Learning	⇨ Tuning	⇨ Mining
Artificial Neural Networks	 Chén Dìng (racewalker) **15.2 km/h**	 Turtle **15-35 km/h**, running/swimming	 Turtle **15-35 km/h**, running/swimming	 Spirostomum ambiguum **724 km/h**

Online time series analytics by parallel algorithms **plus** ANNs

情义无价
Friendship has no price.
Chinese proverb

Parallel algorithms

Labeling

[Cheetah](#)
100 km/h

Training set

Artificial Neural Networks

Learning ⇨ **Tuning** ⇨ **Mining**

[Turtle](#)
15-35 km/h,
running/swimming

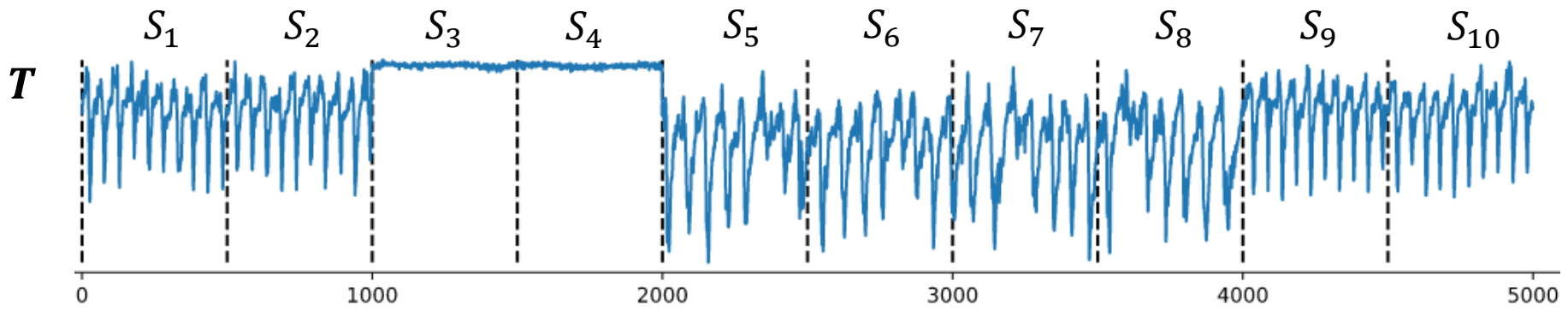
[Turtle](#)
15-35 km/h,
running/swimming

[Spirostomum ambiguum](#)
724 km/h

Outline

- Introduction
- Offline time series analytics
- **Online time series analytics**
  - Employing ANNs together with parallel algorithms
  - Parallel time series labeling
  - Online anomaly detection and imputation of missing values
  - Parallel time series anomaly discovery
- Conclusions

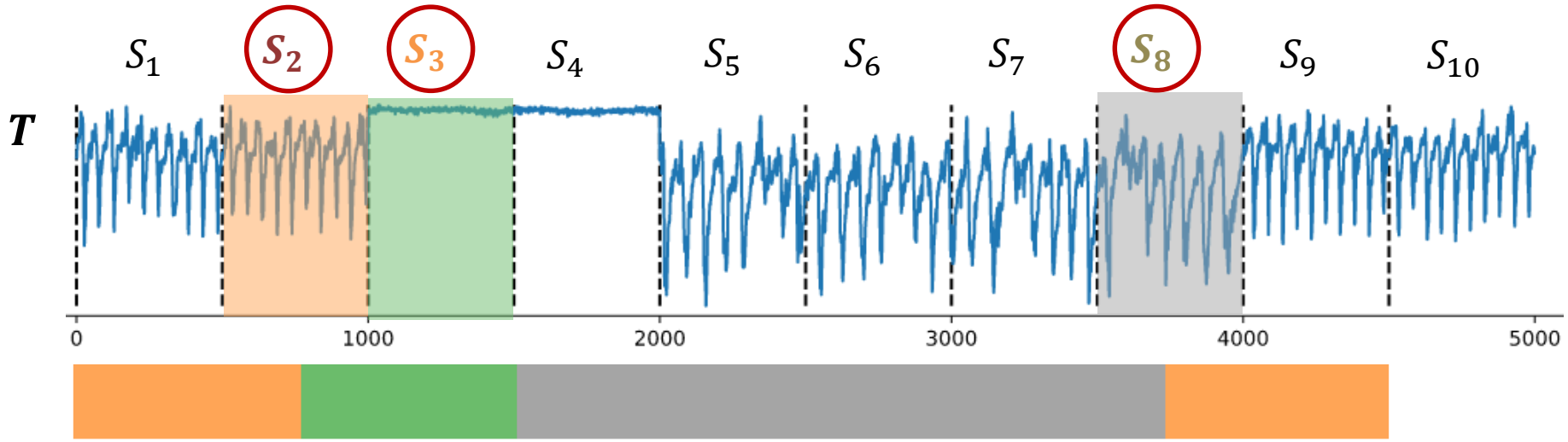
Time series labeling through snippets[1]



1. Let us represent a time series as a set of non-overlapped segments of a given length

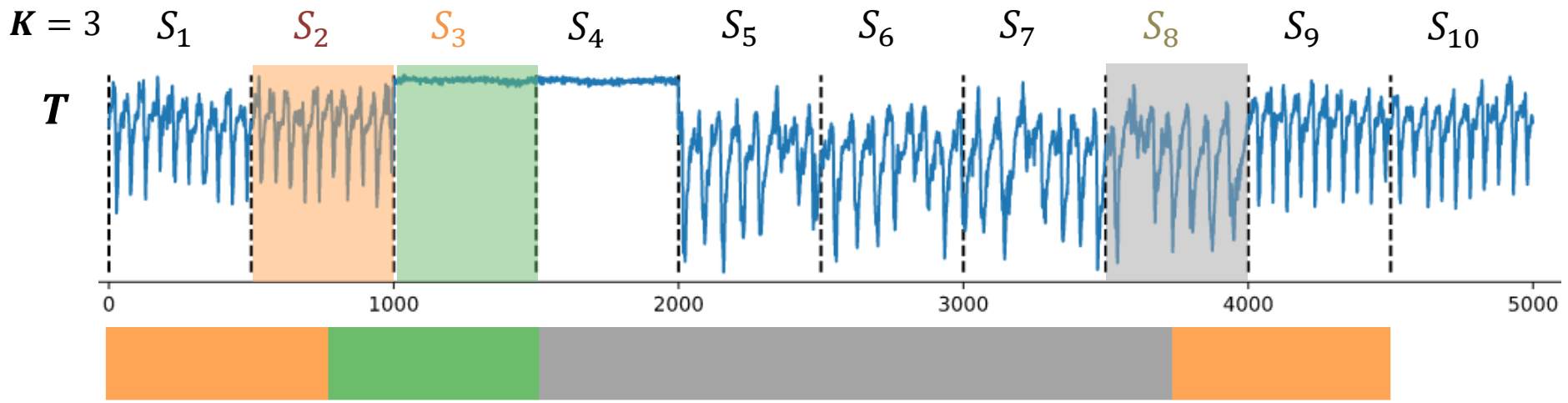
[1] Imani S. *et al.* Introducing time series snippets: a new primitive for summarizing long time series. Data Min. Knowl. Discov. 2020. Vol. 34, no. 6. P. 1713-1743. DOI: 10.1007/s10618-020-00702-y

Time series labeling through snippets

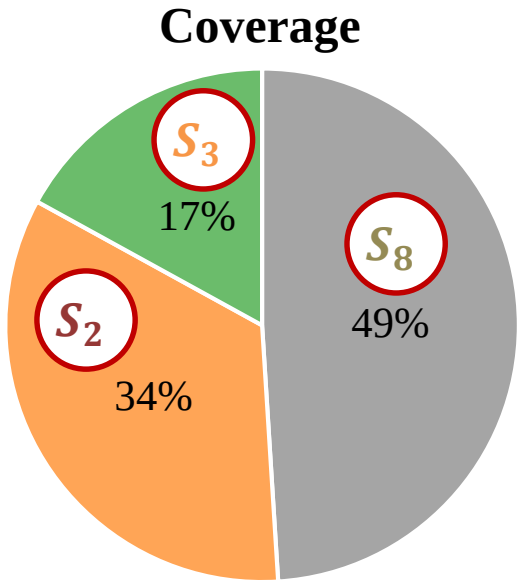


1. Let us represent a time series as a set of non-overlapped segments of a given length
2. For each segment, let us find its **nearest neighbors**

Time series labeling through snippets



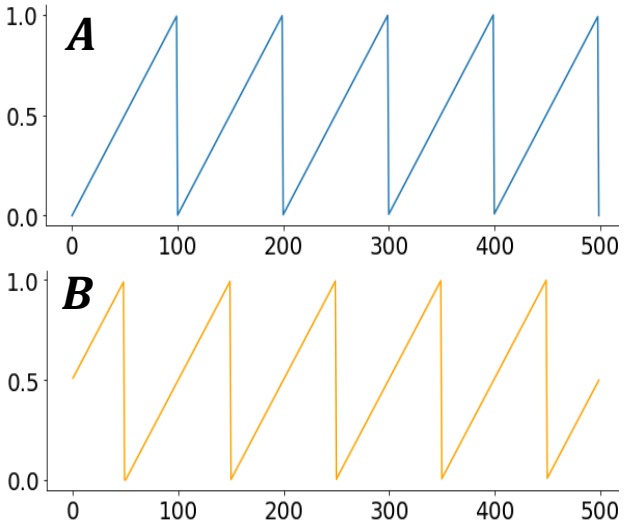
1. Let us represent a time series as a set of non-overlapped segments of a given length
2. For each segment, let us find its **nearest neighbors**
3. For each segment, let us compute its coverage and take top-K segments



Time series similarity through MPdist[1]

- Two m-**length time series** are more similar w.r.t. **MPdist**, the more ℓ-**length** $(3 \leq \ell \leq m)$ **subsequences** close to each other w.r.t. the **normalized Euclidean distance**, are in them
- MPdist is a distance measure (not a metric), i.e. it holds the identity and symmetry axioms but not the triangle inequality
- MPdist is phase-invariant

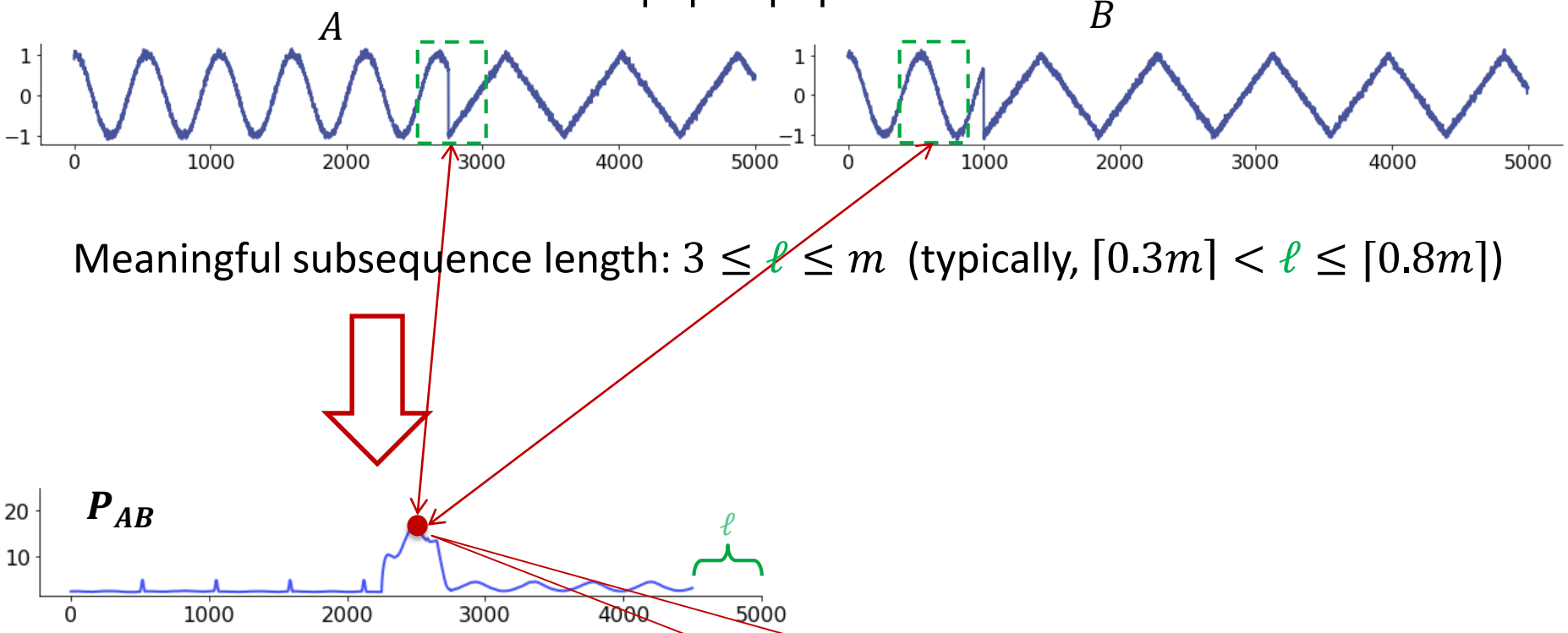
$\mathrm{ED}(A, B)$	11.2
MPdist(A, B)	**0**



1 Gharghabi S. *et al.* An ultra-fast time series distance measure to allow data mining in more complex real-world deployments. Data Min. Knowl. Discov. 2020. Vol. 34. P. 1104–1135. DOI: 10.1007/s10618-020-00695-8

MPdist: Matrix profile AB

$$|A| = |B| = m$$

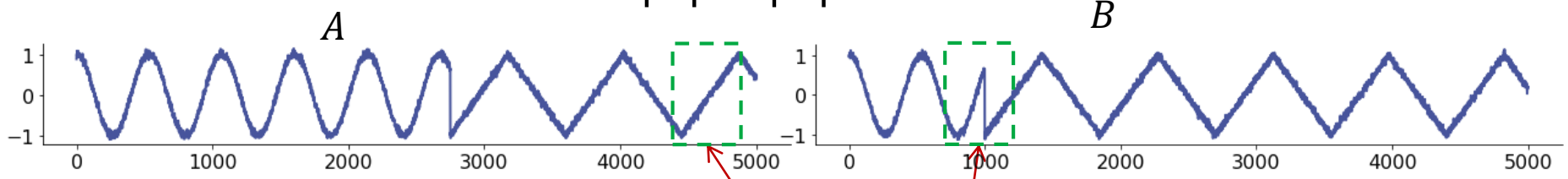


Meaningful subsequence length: $3 \leq \ell \leq m$ (typically, $\lceil 0.3m \rceil < \ell \leq \lceil 0.8m \rceil$)

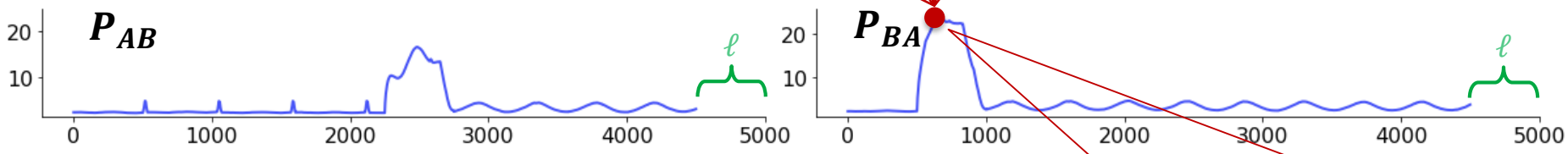
Normalized Euclidean distance between i-th ℓ-length subsequence in A and its nearest ℓ-length subsequence in B

MPdist: Matrix profile BA

$$|A| = |B| = m$$

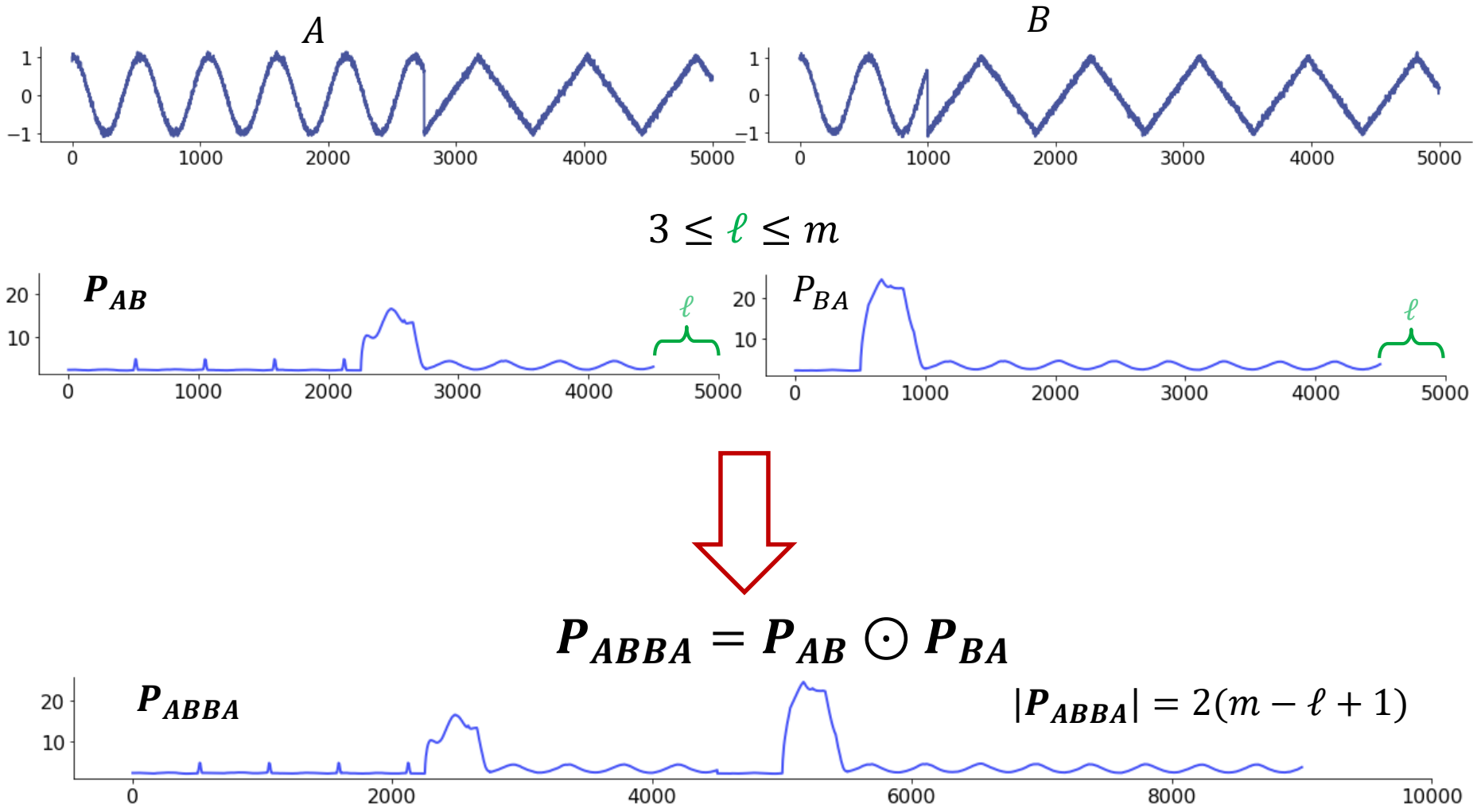


Meaningful subsequence length: $3 \leq \ell \leq m$ (typically, $\lceil 0.3m \rceil < \ell \leq \lceil 0.8m \rceil$)

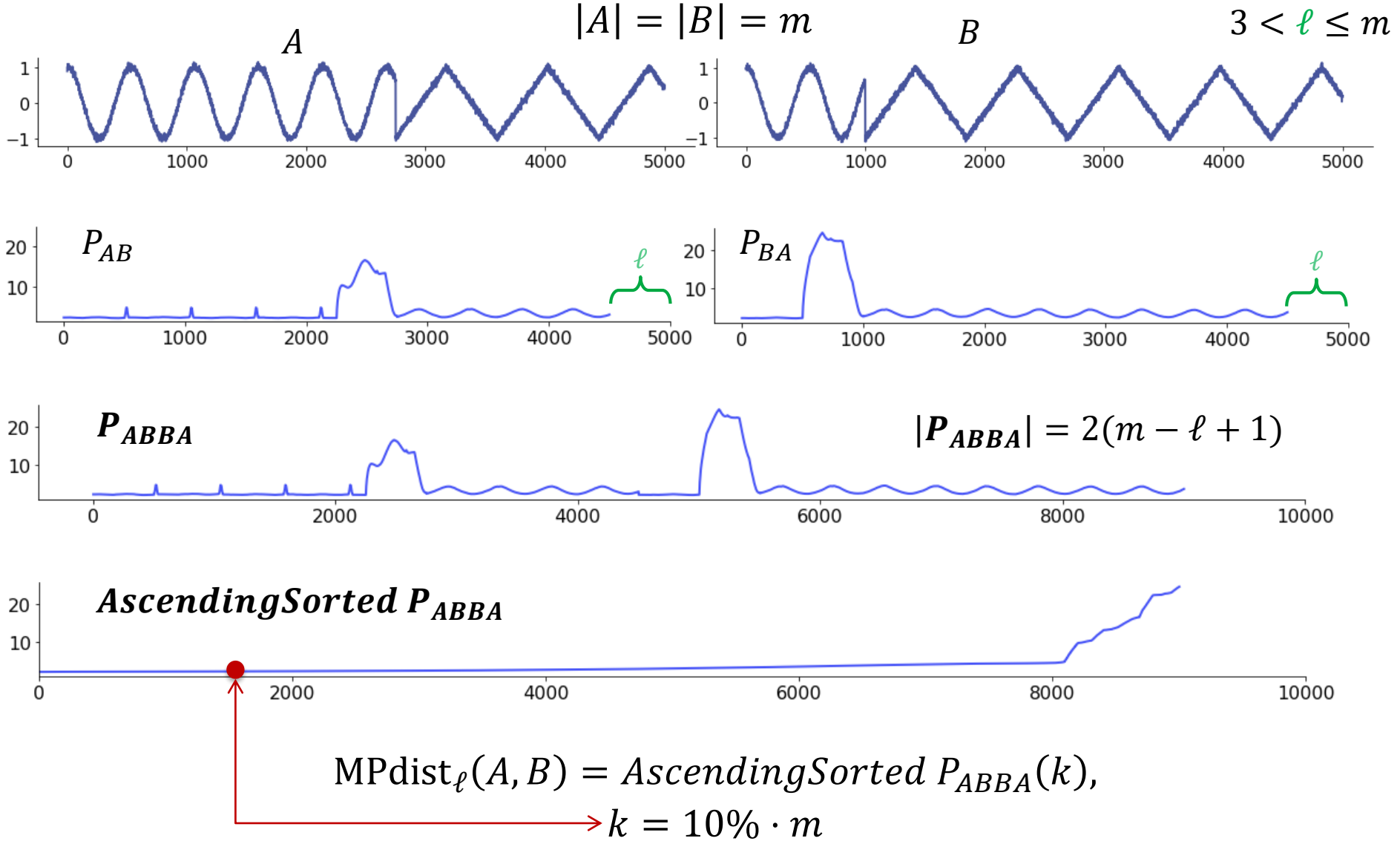


Normalized Euclidean distance between i-th ℓ-length subsequence in A and its nearest ℓ-length subsequence in B

MPdist: Matrix profile ABBA



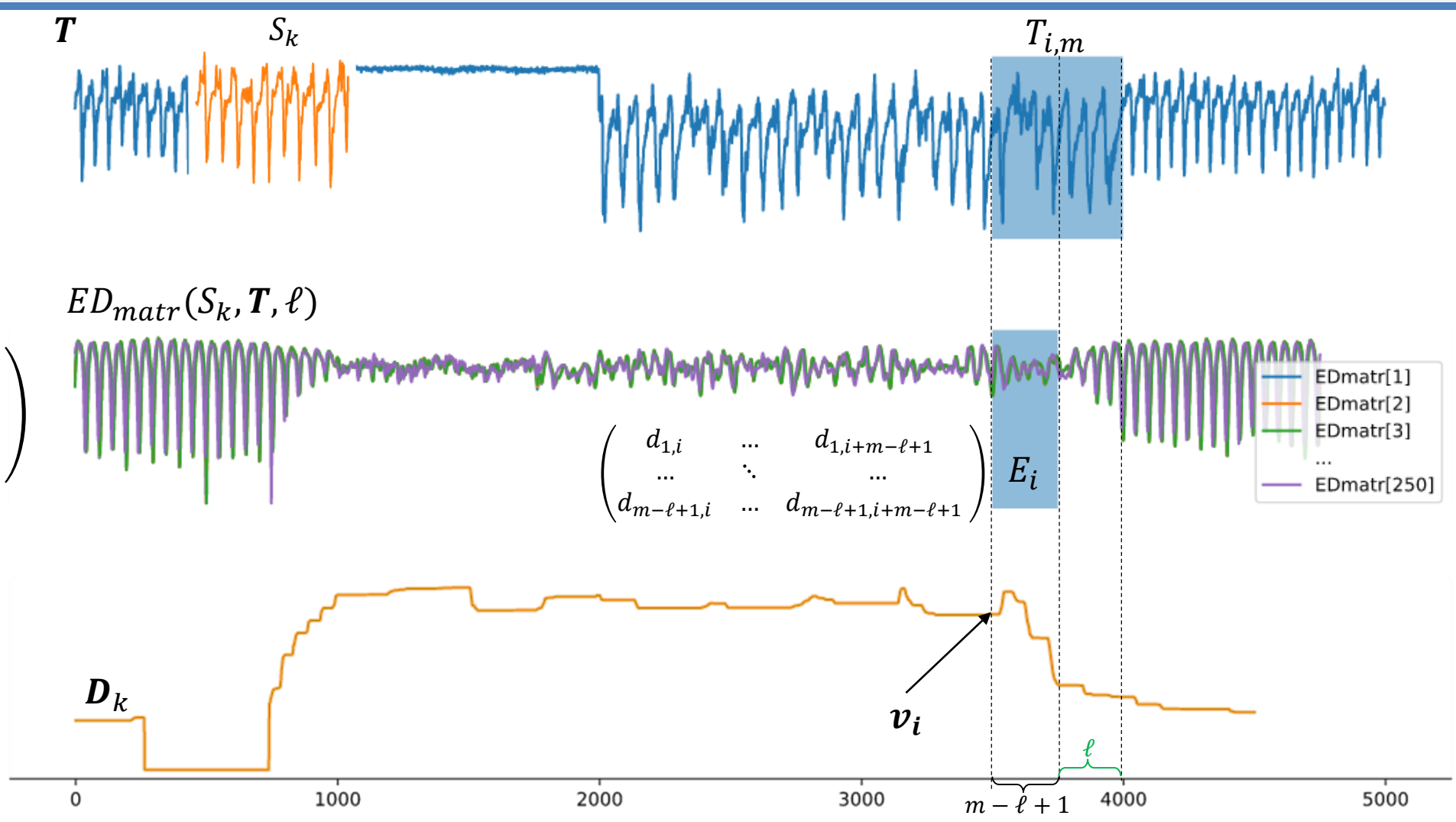
MPdist: Eventual calculation



Parallel labeling: MPdist profile

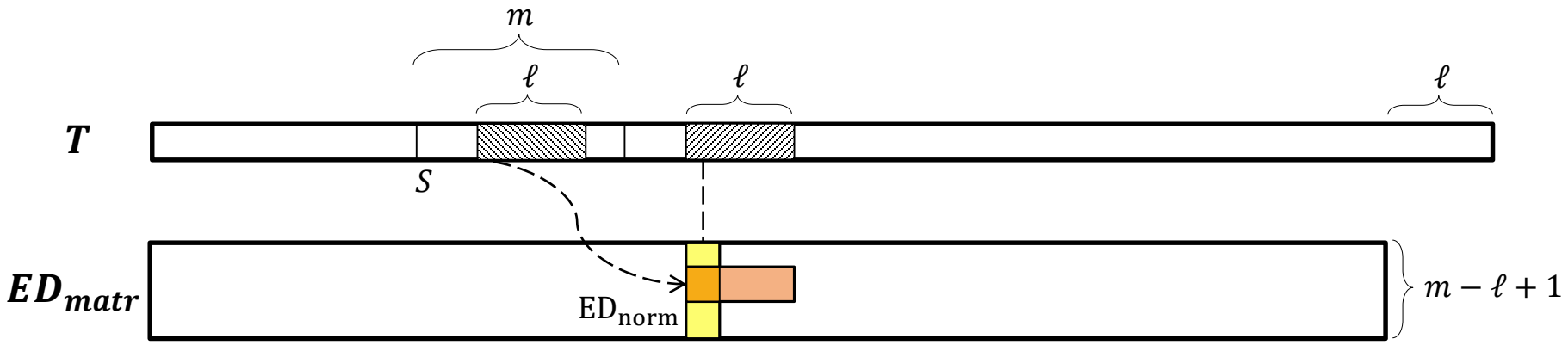
$$\begin{pmatrix} d_{1,1} & \cdots & d_{1,n-\ell+1} \\ \cdots & \ddots & \cdots \\ d_{m-\ell+1,1} & \cdots & d_{m-\ell+1,n-\ell+1} \end{pmatrix}$$

$$d_{i,j} = \mathrm{ED}_{\mathrm{norm}}(S_{i,\ell}, T_{j,\ell})$$



$$D_k(S_k, T) = [v_1, v_2, \ldots, v_{n-m+1}], \qquad v_i = \mathrm{MPdist}_\ell(S_k, T_{i,m})$$

Parallel snippet discovery: ED_{matr}



$$\mathrm{ED}_{\mathrm{norm}}(T_{i,m}, T_{j,m}) = \sqrt{2m(1 - P_{i,j})}$$

$$P_{i,j} = \overline{QT}_{i,j} \cdot \frac{1}{\|T_{i,m} - \mu_i\|} \cdot \frac{1}{\|T_{j,m} - \mu_j\|},$$

$$T_{i,m} - \mu_i = (t_i - \mu_i, \ldots, t_{i+m-1} - \mu_i),$$

$$\mu_i = \frac{1}{m} \sum_{j=i}^{i+m} t_j,$$

$$dg_0 = 0; dg_i = (t_{i+m-1} - \mu_i) + (t_{i-1} - \mu_{i-1}),$$

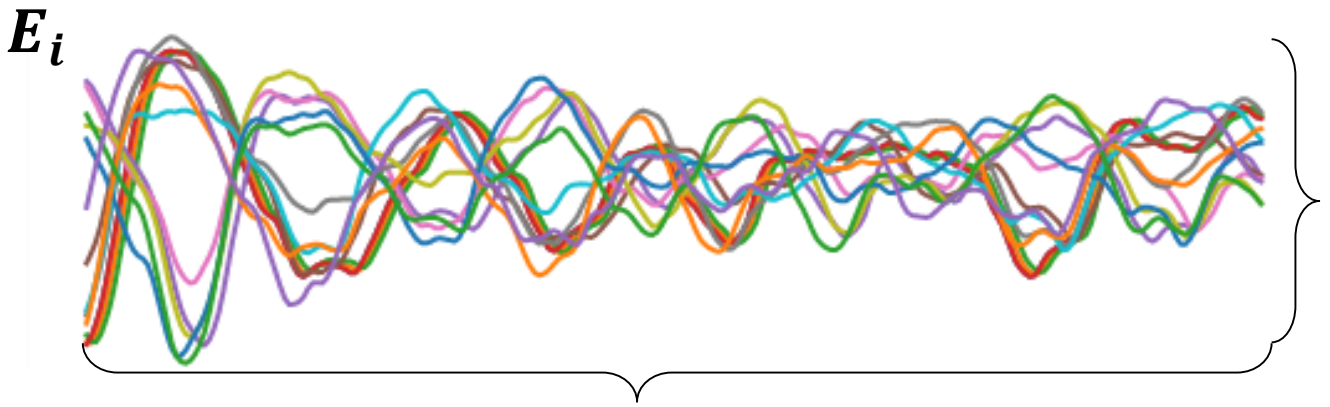
$$df_0 = 0; df_i = \frac{t_{i+m-1} - t_{i-1}}{2},$$

$$\overline{QT}_{i,j} = \overline{QT}_{i-1,j-1} + df_i \cdot dg_j + df_j \cdot dg_i,$$

[1]

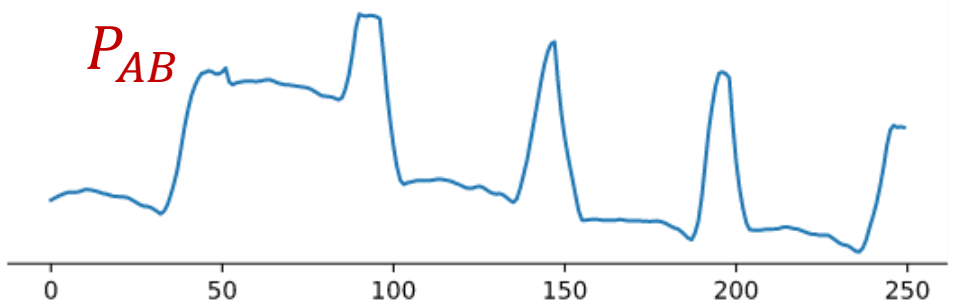
[1] Zimmerman Z. *et al.* Matrix Profile XIV: Scaling time series motif discovery with GPUs to break a quintillion pairwise comparisons a day and beyond. SoCC 2019. P. 74–86. DOI: 10.1145/3357223.3362721

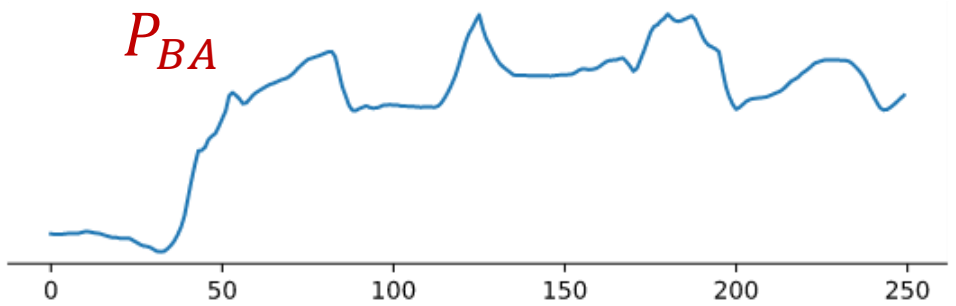
Parallel labeling: MPdist profile



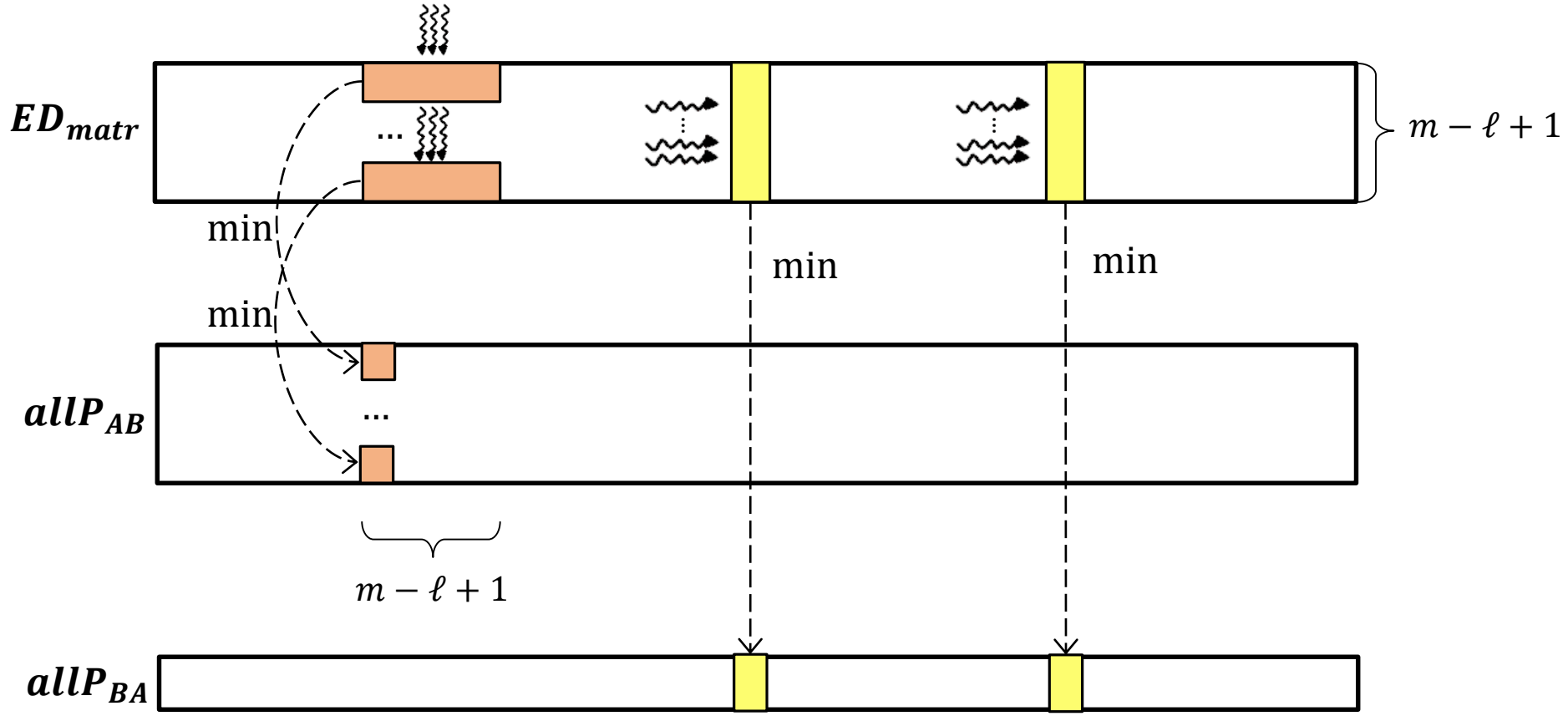
Column-wise minima results in P_{AB}

Row-wise minima results in P_{BA}



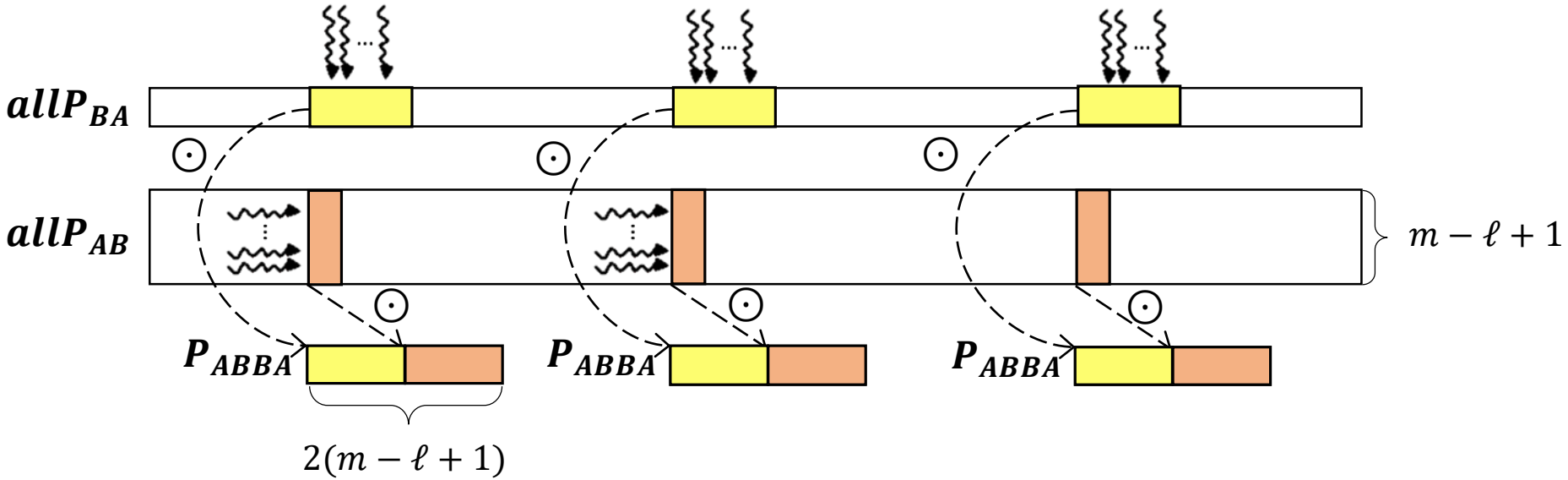


Parallel labeling: $allP_{AB}$ and $allP_{BA}$

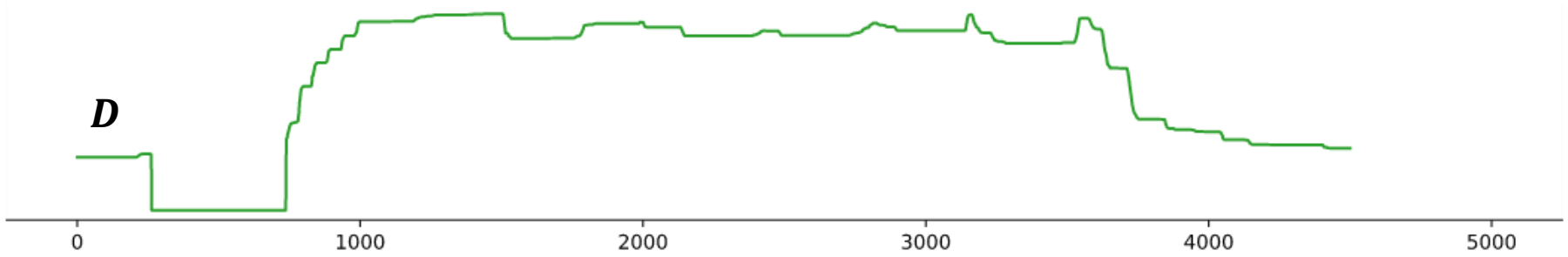


$$allP_{AB}(i,j) = \min_{j \le c \le j+m-\ell+1} ED_{matr}(i,c) \qquad allP_{BA}(j) = \min_{1 \le i \le m-\ell+1} ED_{matr}(i,j)$$

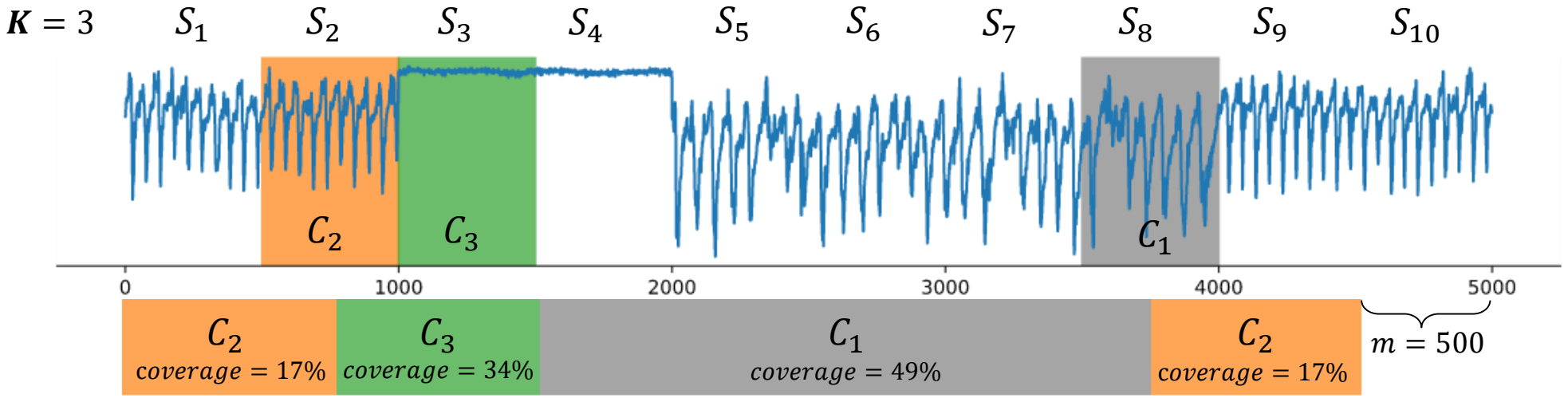
Parallel labeling: P_{ABBA}



$$\text{MPdist}_{\ell}(A, B) = AscendingSorted\ P_{ABBA}(k),$$
$$\text{where } k = 10\% \cdot m$$



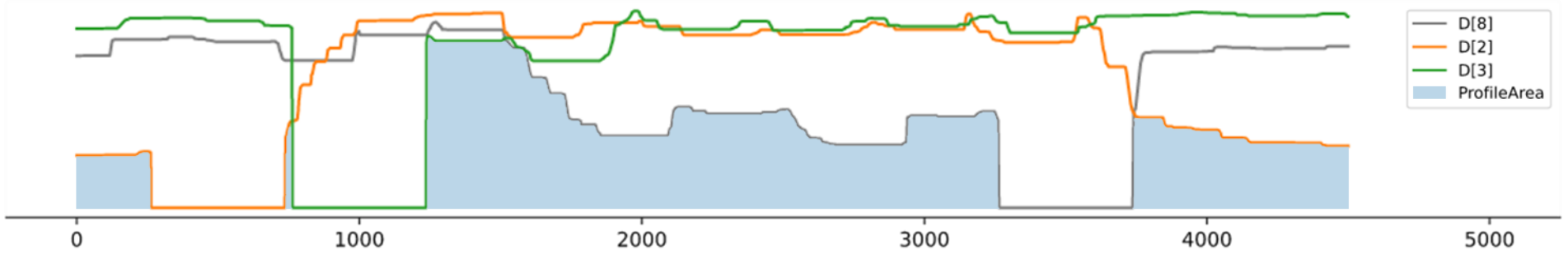
Parallel labeling: Results



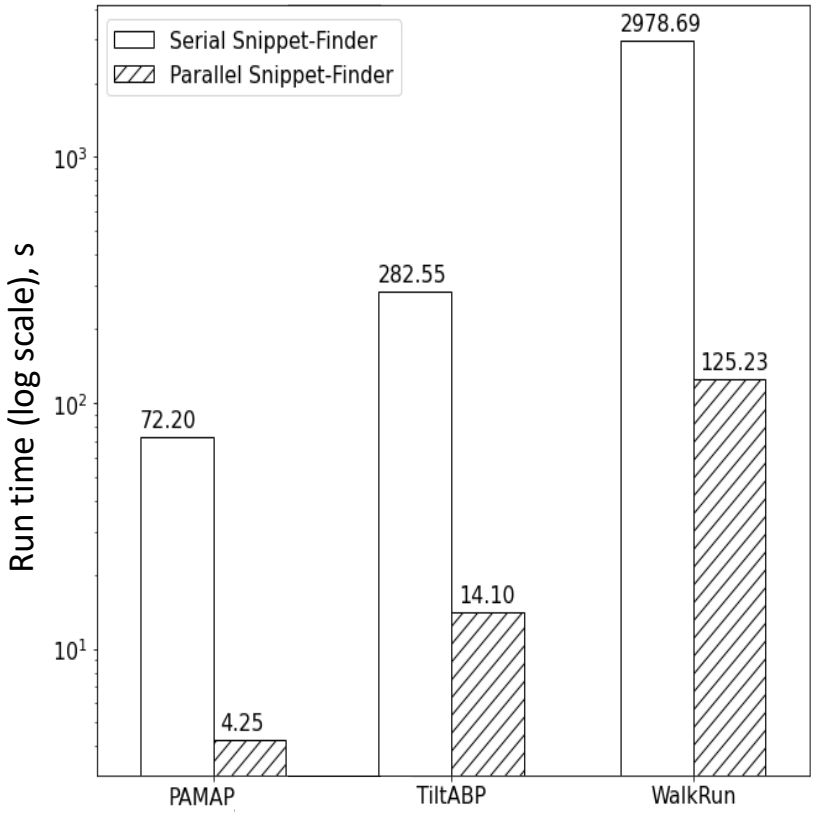
$$i = 1: \quad C_1 = \arg\min_{1 \le j \le n/m} ProfileArea(\{D_j\})$$

$$i = 2: \quad C_2 = \arg\min_{1 \le j \le n/m} ProfileArea(\{D_{C_1}, D_j\})$$

$$3 \le i \le K: \quad C_i = \arg\min_{1 \le j \le n/m} ProfileArea(\{D_{C_1}, \ldots, D_{C_{i-1}}, D_j\})$$



Parallel labeling: Experiments on performance



Parallel labeling algorithm is **20 times ahead** of serial analog

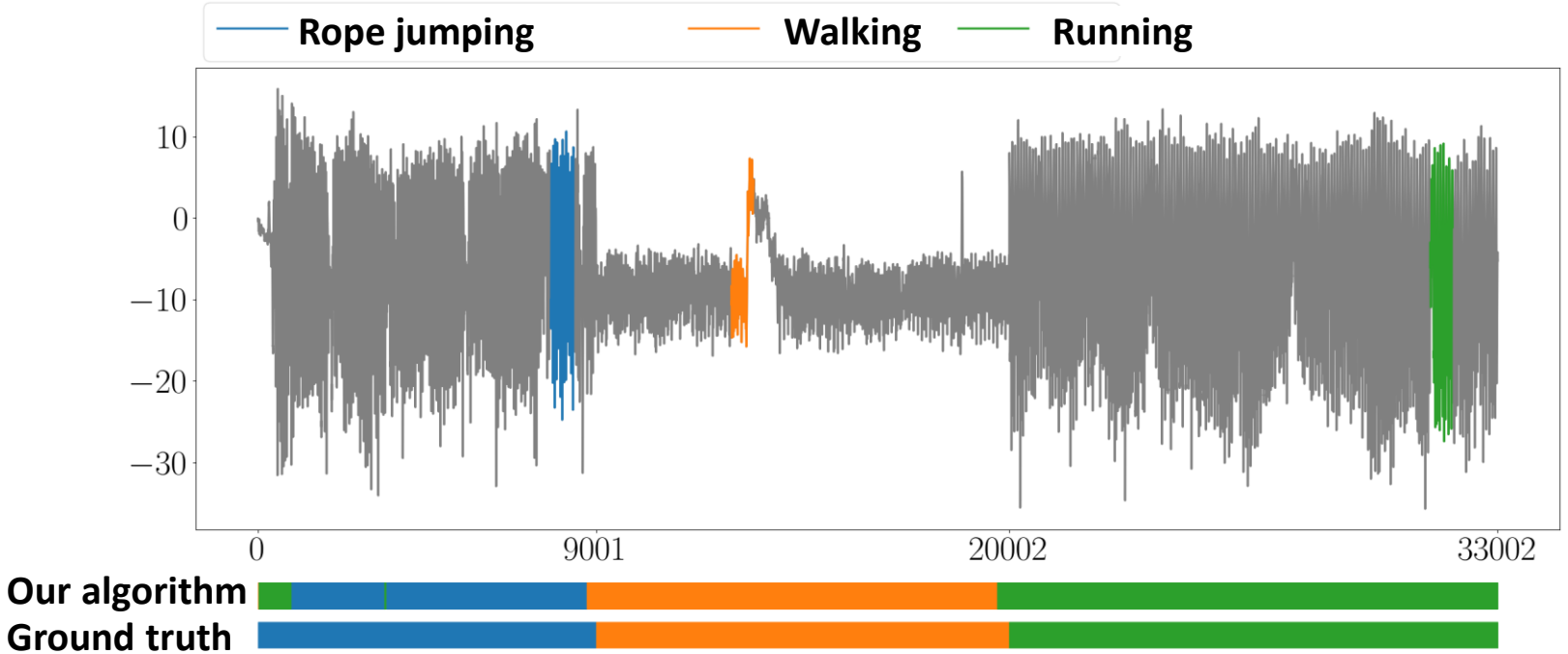
Hardware: NVIDIA Tesla V100 SXM2 (5120 cores @1.3 GHz)

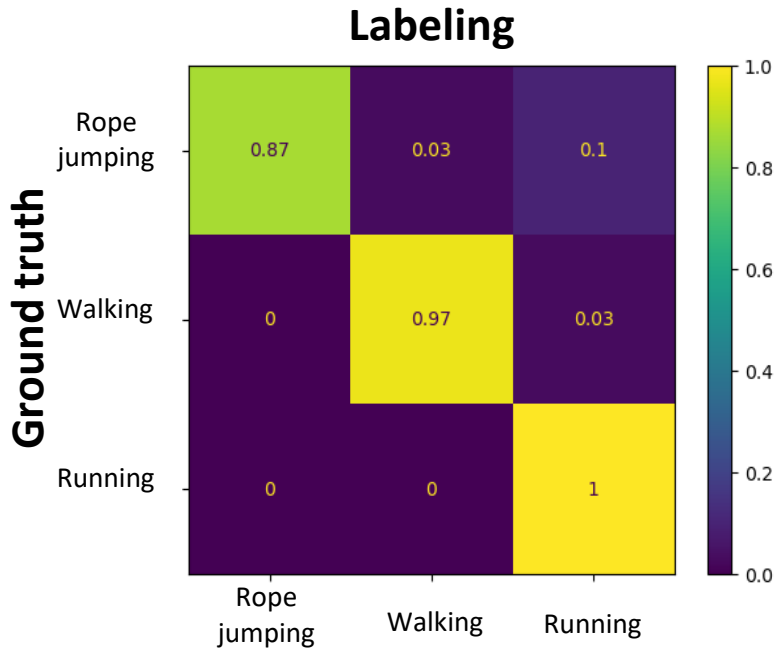
Time series	Length	Snippet length	Domain
TiltABP[1]	40 000	630	Human blood pressure during rapid tilts
PAMAP[2]	20 002	600	Wearable accelerometer readings during various types of human physical activity
WalkRun[2]	100 000	240	

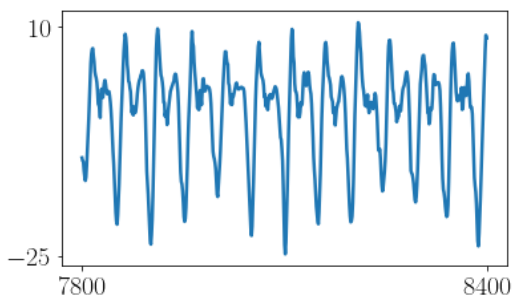
[1] Imani S. *et al.* Introducing time series snippets: a new primitive for summarizing long time series. Data Min. Knowl. Discov. 2020. Vol. 34, no. 6. P. 1713-1743. DOI: 10.1007/s10618-020-00702-y

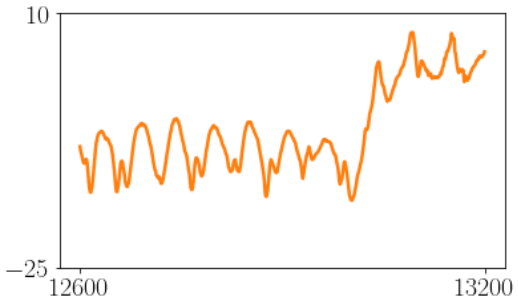
[2] Reiss A., Stricker D. Introducing a new benchmarked dataset for activity monitoring. ISWC 2012. P. 108–109. DOI: 10.1109/ISWC.2012.13

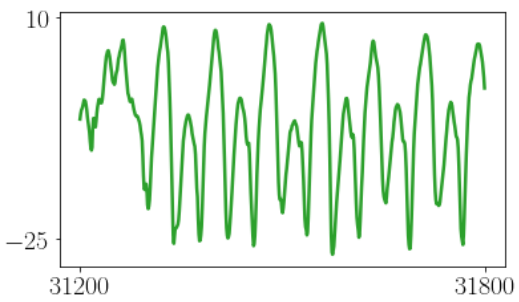
Parallel labeling: Experiments on accuracy











Parallel labeling algorithm demonstrates good accuracy

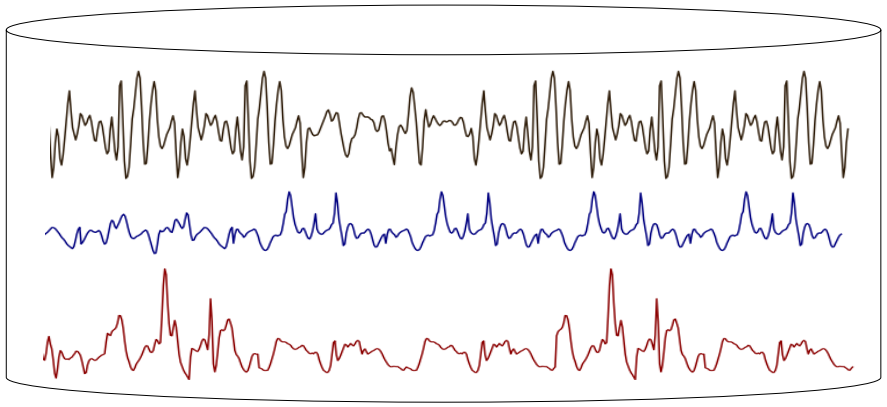
* Reiss A., Stricker D. Introducing a new benchmarked dataset for activity monitoring. ISWC 2012. P. 108–109. DOI: 10.1109/ISWC.2012.13

Outline

- Introduction
- Offline time series analytics
- **Online time series analytics**
  - Employing ANNs together with parallel algorithms
  - Parallel time series labeling
  - Online imputation of missing values and anomaly detection
  - Parallel time series anomaly discovery
- Conclusions

Online time series imputation

Representative fragment of time series



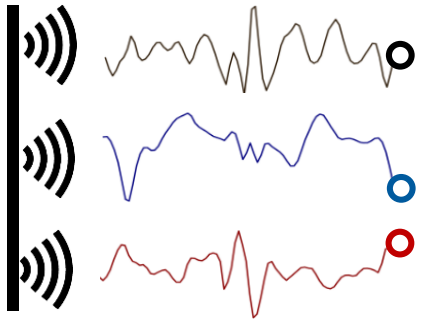
Parallel Labeling

Training sets

recognize

reconstruct

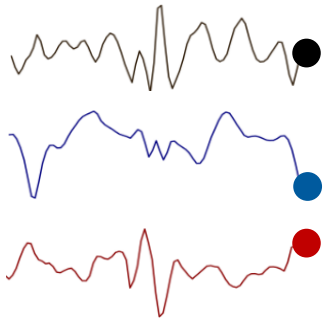
Subsequence with **missing** values



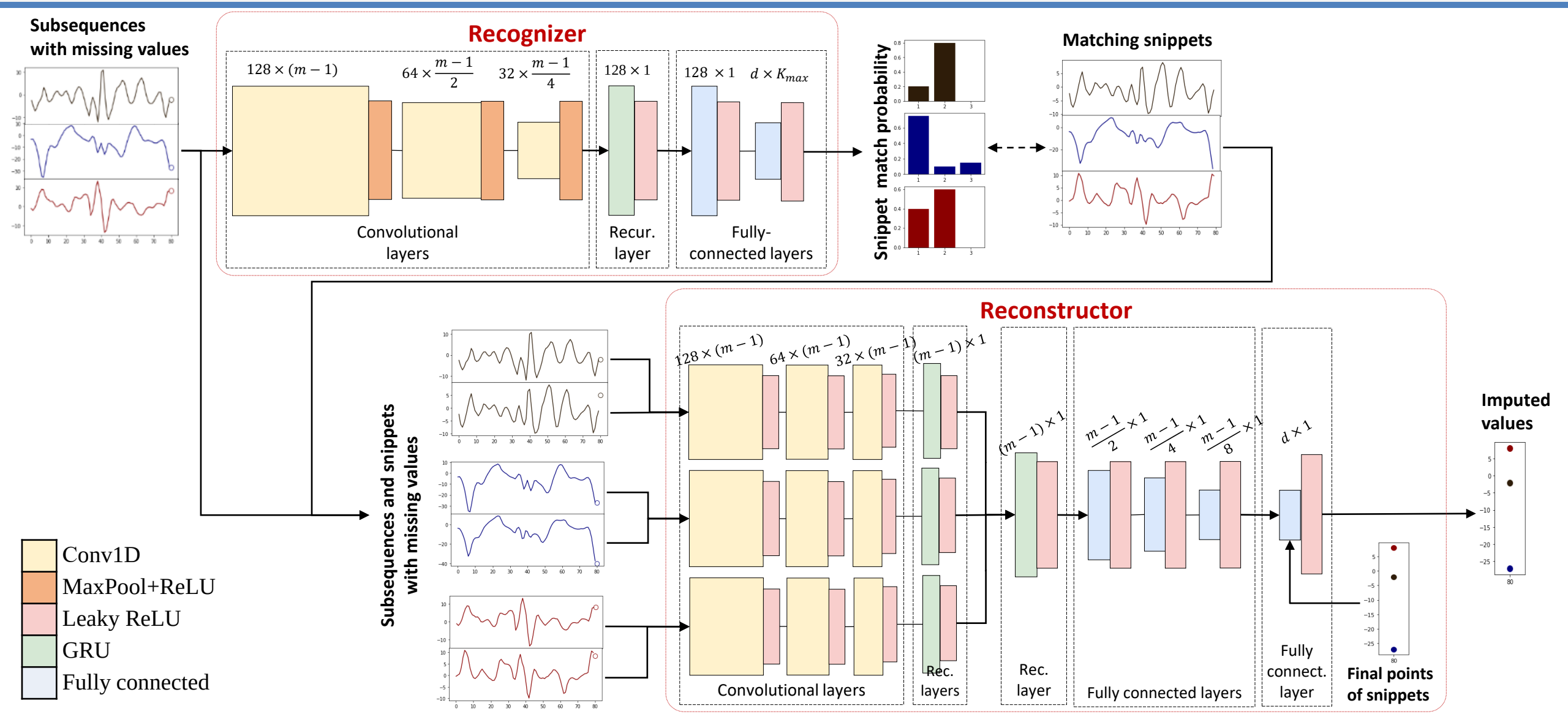
Recognizer ANNs

Reconstructor ANNs

Subsequence with **imputed** values



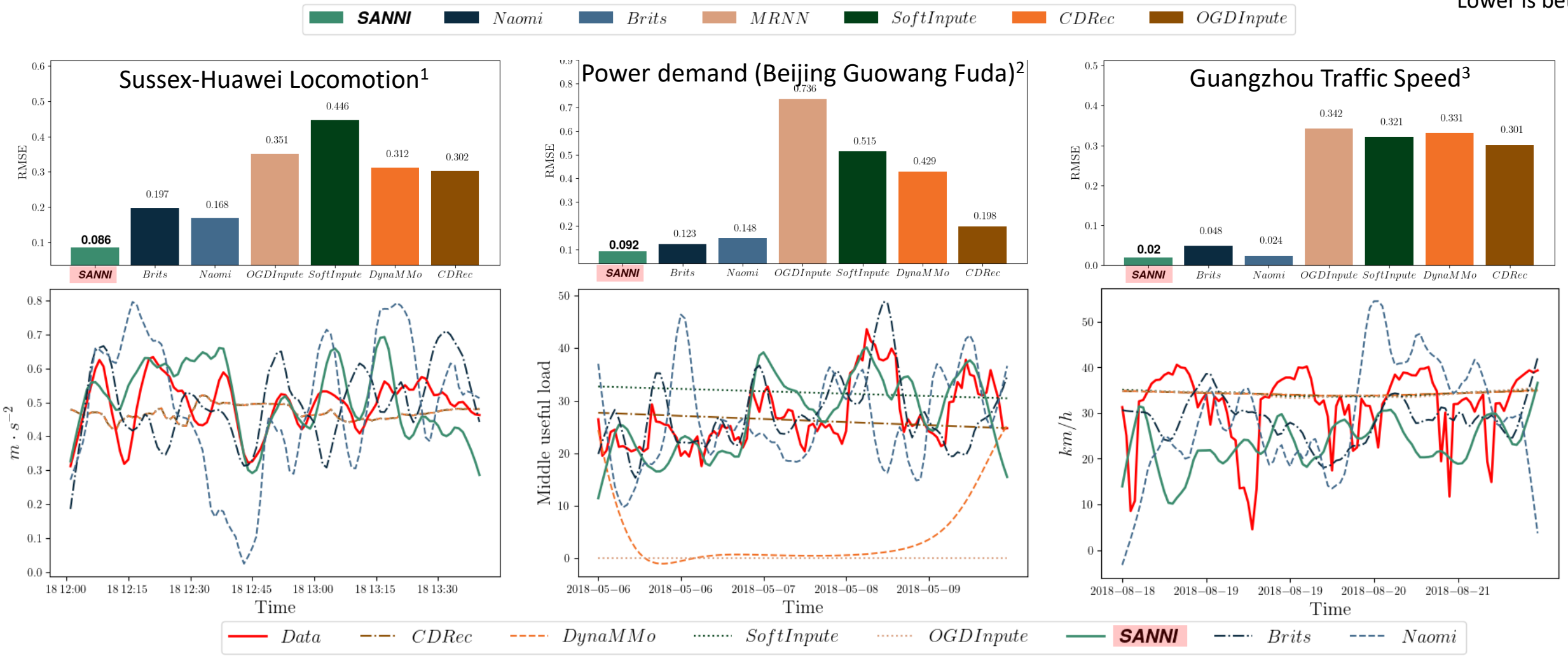
SANNI: Snippet & ANN-based Imputation



Online imputation: Experiments on accuracy

$$\mathrm{RMSE} = \sqrt{\frac{1}{h}\sum_{i=1}^{h}(t_i - \tilde{t}_i)^2}$$

Lower is better



[1] Rogger D. *et al.* The University of Sussex-Huawei Locomotion (SHL) dataset. URL: http://www.shl-dataset.org/

[2] Zhou H. *et al.* Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. AAAI 2021. P. 11106-1115. DOI: 10.1609/aaai.v35i12.17325

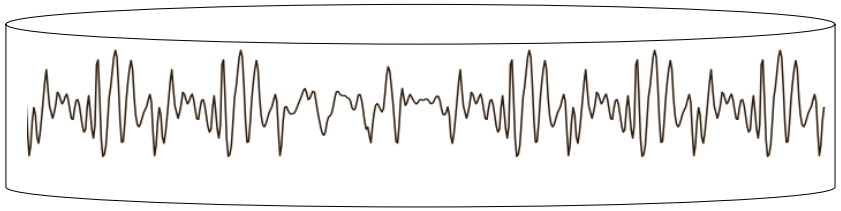
[3] Chen X. *et al.* A Bayesian tensor decomposition approach for spatiotemporal traffic data imputation. 2018 Transportation Research Part C: Emerging Technologies. DOI: 10.1016/J.TRC.2018.11.003

Outline

- Introduction
- Offline time series analytics
- **Online time series analytics**
  - Employing ANNs together with parallel algorithms
  - Parallel time series labeling
  - Online imputation of missing values and anomaly detection
  - Parallel time series anomaly discovery
- Conclusions

Online time series anomaly detection

Representative fragment of time series



Parallel Labeling

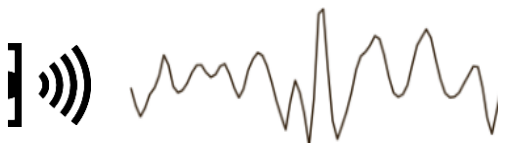
Discovers snippets

Parallel Discord Discovery

Removes anomalies

Training set

Potentially anomalous subsequence



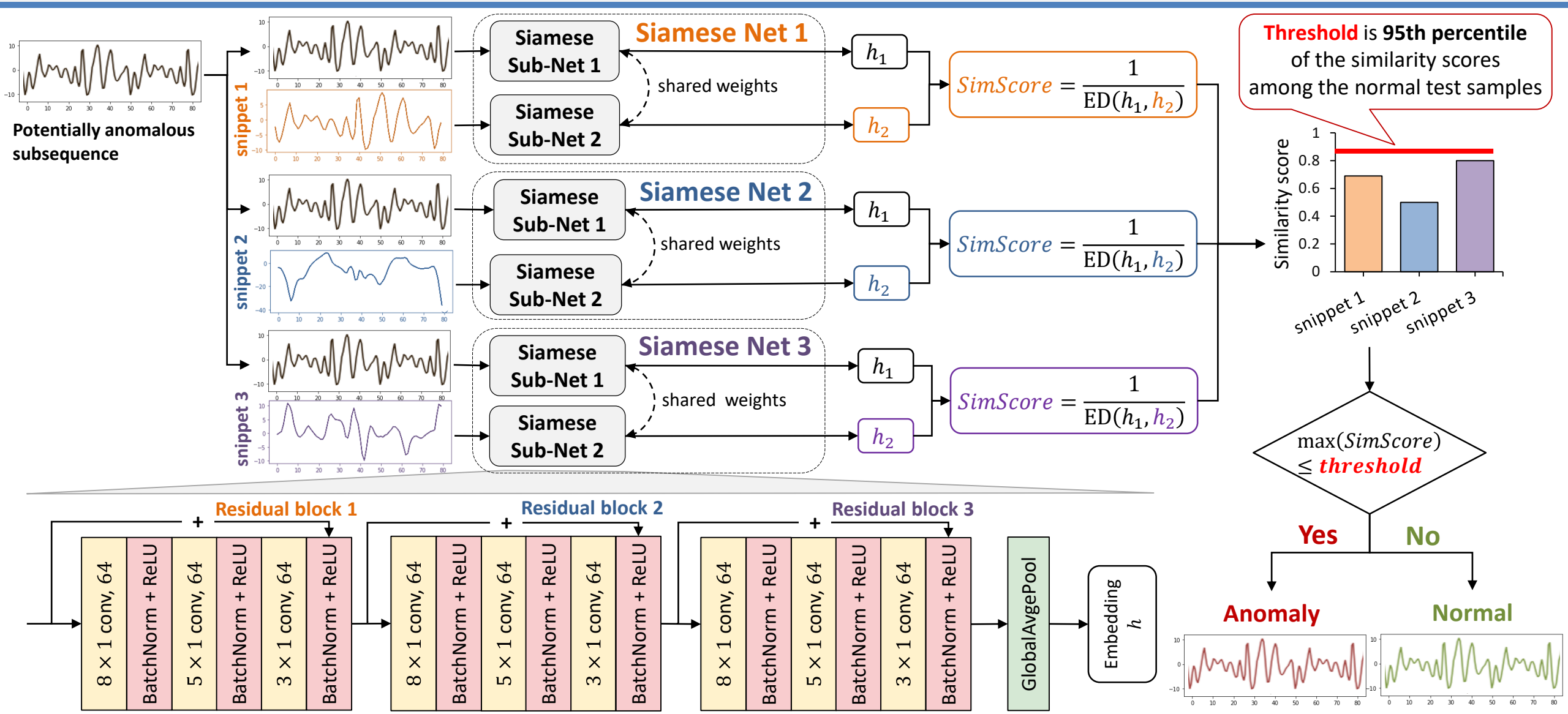
Anomaly detector ANN

Norm

OR

Anomaly

DiSSiD: Discord, Snippet, and Siamese Net-based Detector of anomalies

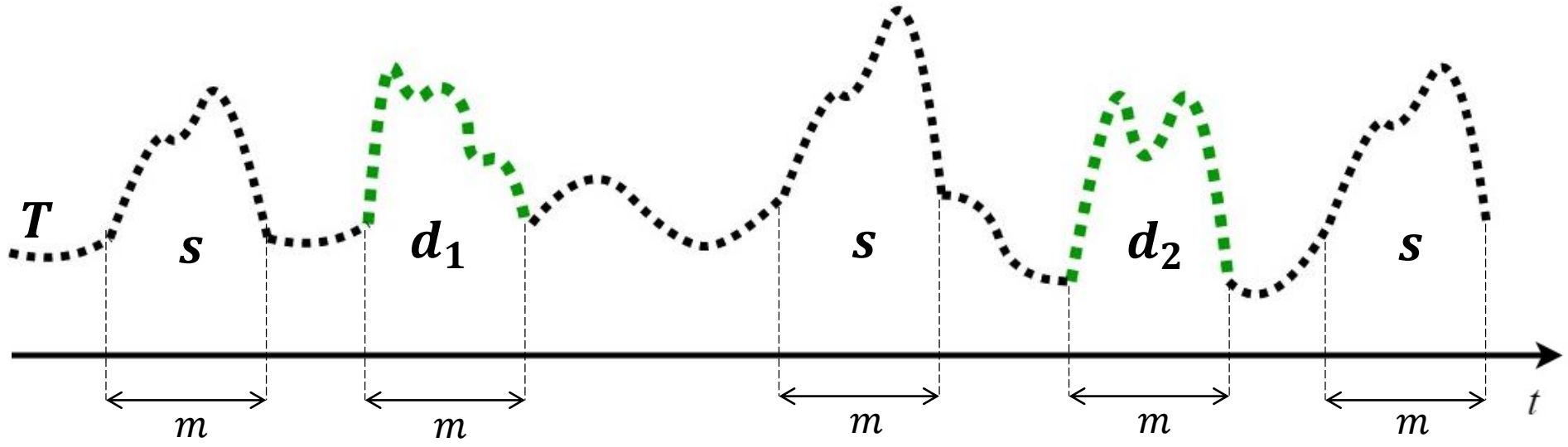


Outline

- Introduction
- Offline time series analytics
- **Online time series analytics**
  - Employing ANNs together with parallel algorithms
  - Parallel time series labeling
  - Online imputation of missing values and anomaly detection
  - Parallel time series anomaly discovery
- Conclusions

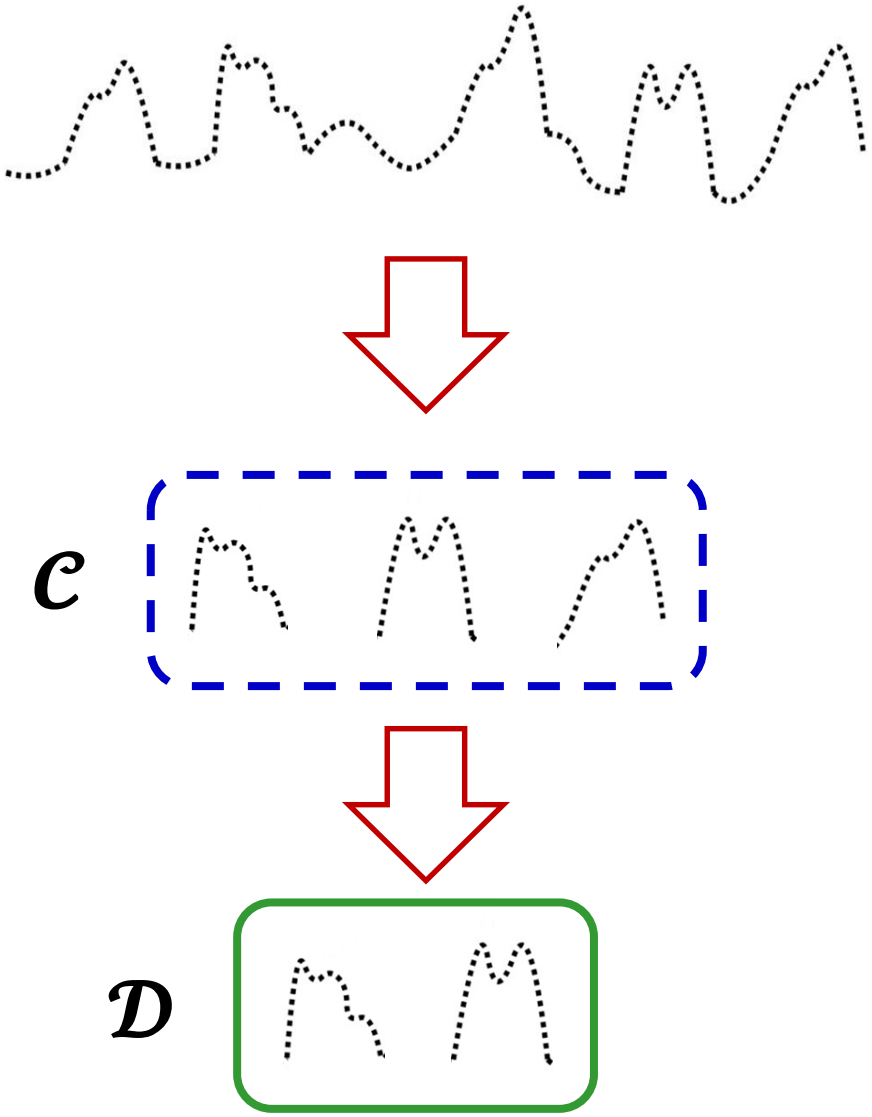
How to formalize anomalies? Discords[1]

- Discord is a subsequence whose nearest neighbor is at least at a given threshold far away
- We are given: $\boldsymbol{T}$, discord length m, threshold r
- We are to find: $\boldsymbol{\mathcal{D}} = \{\boldsymbol{d_1}, \boldsymbol{d_2}, \ldots\}, \quad \boldsymbol{d_i} \in \boldsymbol{\mathcal{D}} \Leftrightarrow \min_{\boldsymbol{s} \in \boldsymbol{T},\, \boldsymbol{s} \cap \boldsymbol{d_i} = \emptyset} \mathrm{ED}(\boldsymbol{d_i}, \boldsymbol{s}) \geq r$



[1] Yankov D. *et al.* Disk aware discord discovery: finding unusual time series in terabyte sized datasets. Knowl. Inf. Syst. 17(2): 241–262. 2008. DOI: 10.1007/s10115-008-0131-9

Discord discovery



1. Selection

Through one full scan of the time series, create a **set of candidates** to discords

2. Refinement

Through one full scan of the time series, **prune false positives** from the set above

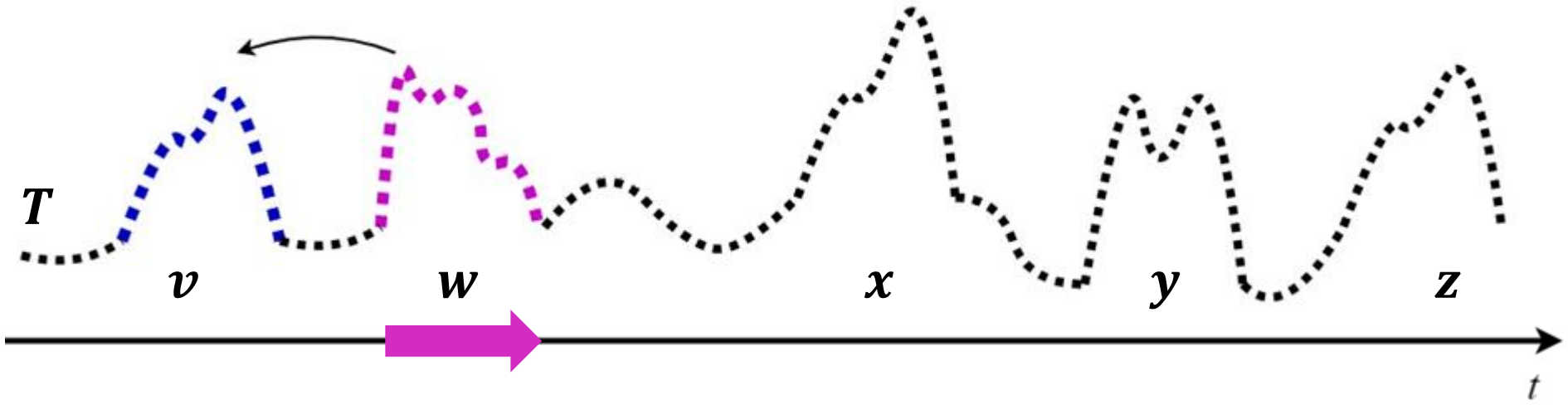
Discord discovery: Selection

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

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

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

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



Discord discovery: Selection

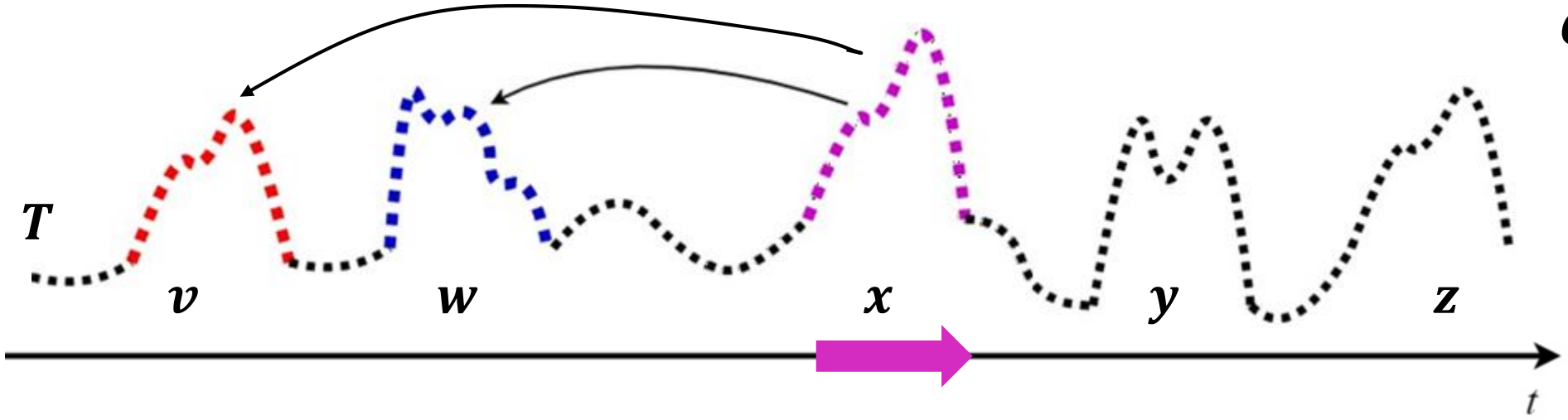
while not end of T
  get current subsequence s
  $isCandidate := \text{TRUE}$
  for each $c_i \in \mathcal{C}$ **and** $s \cap c_i = \emptyset$
    if $\text{ED}(s, c_i) < r$ **then**
      $\mathcal{C} := \mathcal{C} \setminus c_i$; $isCandidate := \text{FALSE}$
  if $isCandidate = \text{TRUE}$ **then**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

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

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

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

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



Discord discovery: Selection

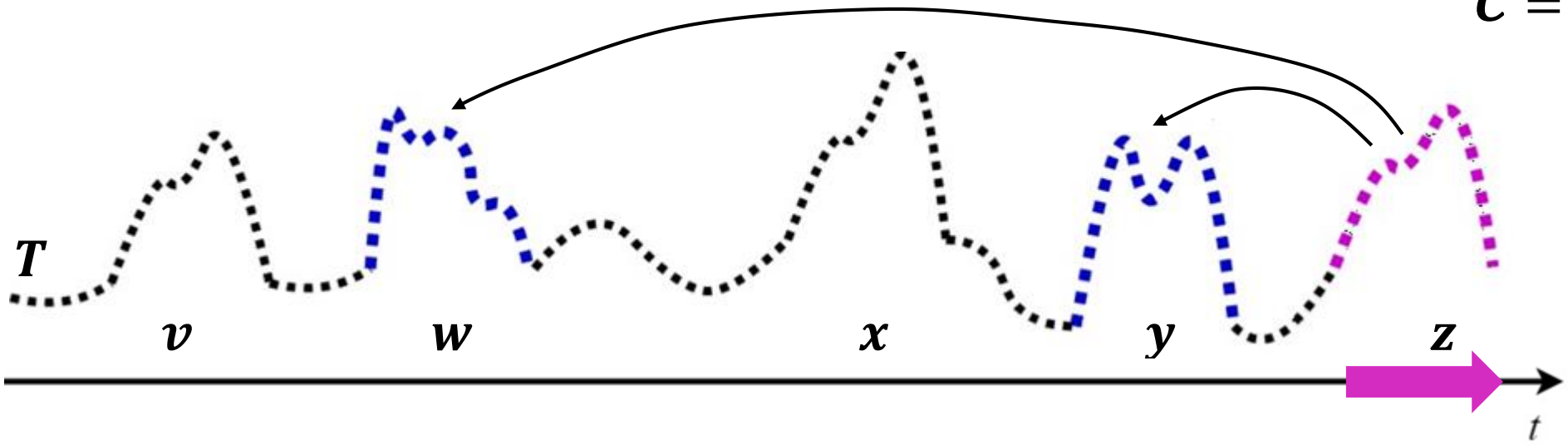
while not end of T
  get current subsequence s
  $isCandidate := \text{TRUE}$
  for each $c_i \in \mathcal{C}$ **and** $s \cap c_i = \emptyset$
    if $\text{ED}(s, c_i) < r$ **then**
      $\mathcal{C} := \mathcal{C} \setminus c_i$; $isCandidate := \text{FALSE}$
  if $isCandidate = \text{TRUE}$ **then**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

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

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

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

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



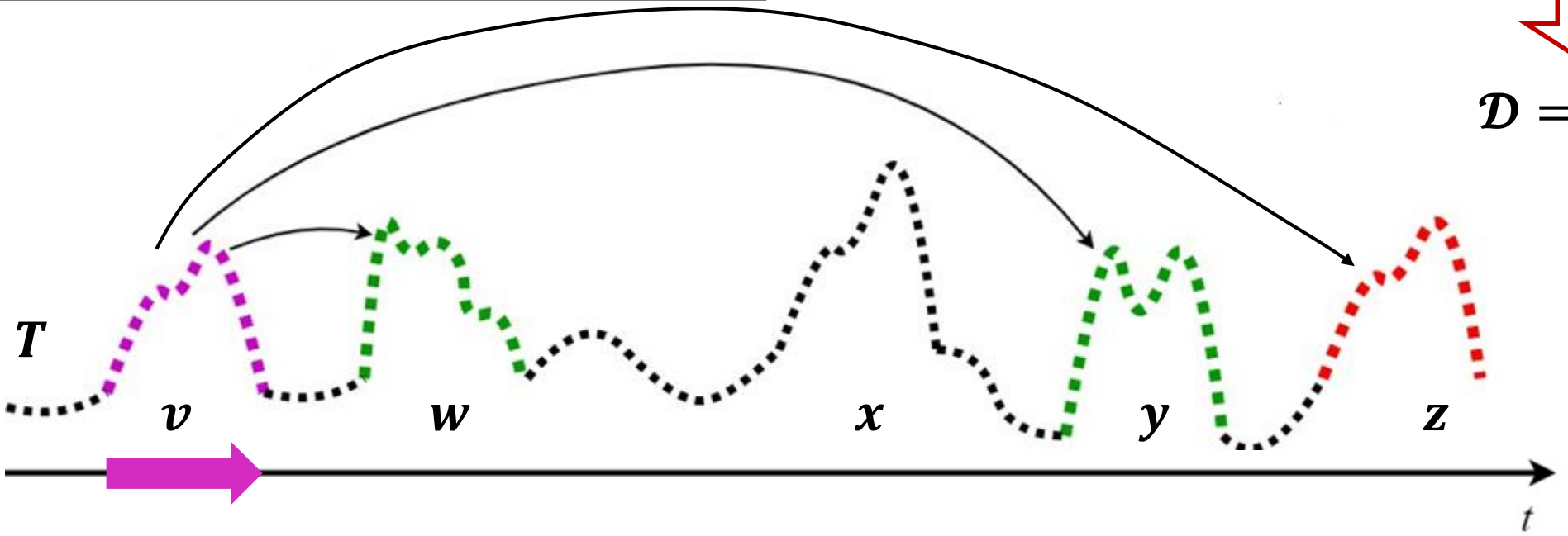
Discord discovery: Refinement

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

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

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

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



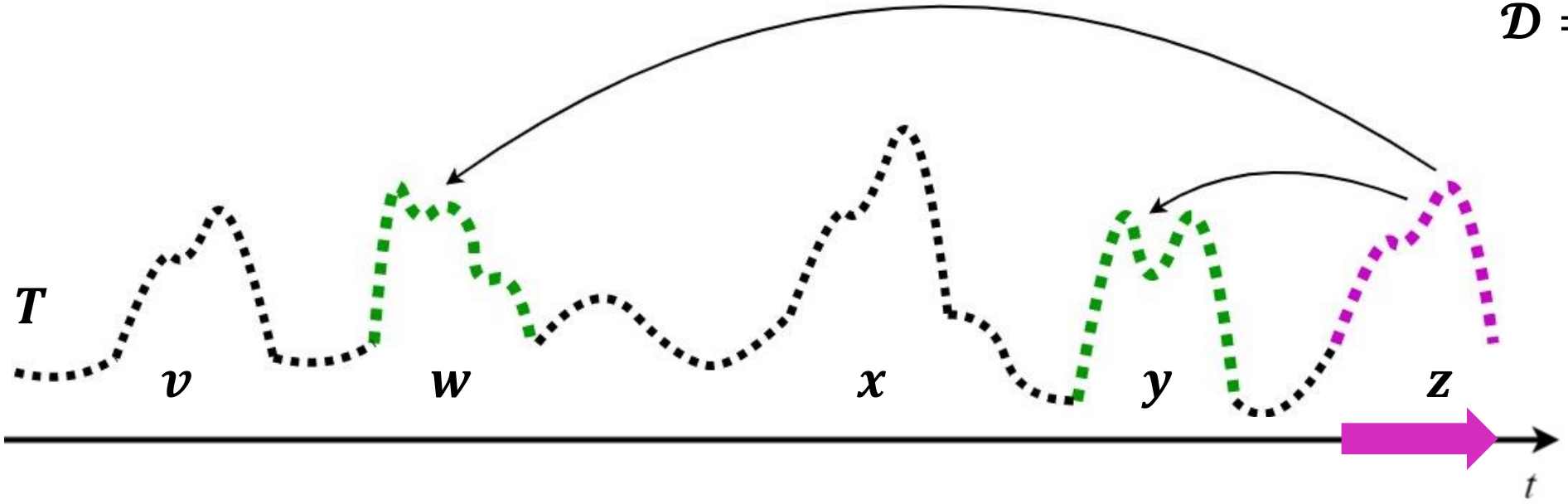
Discord discovery: Refinement

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

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

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

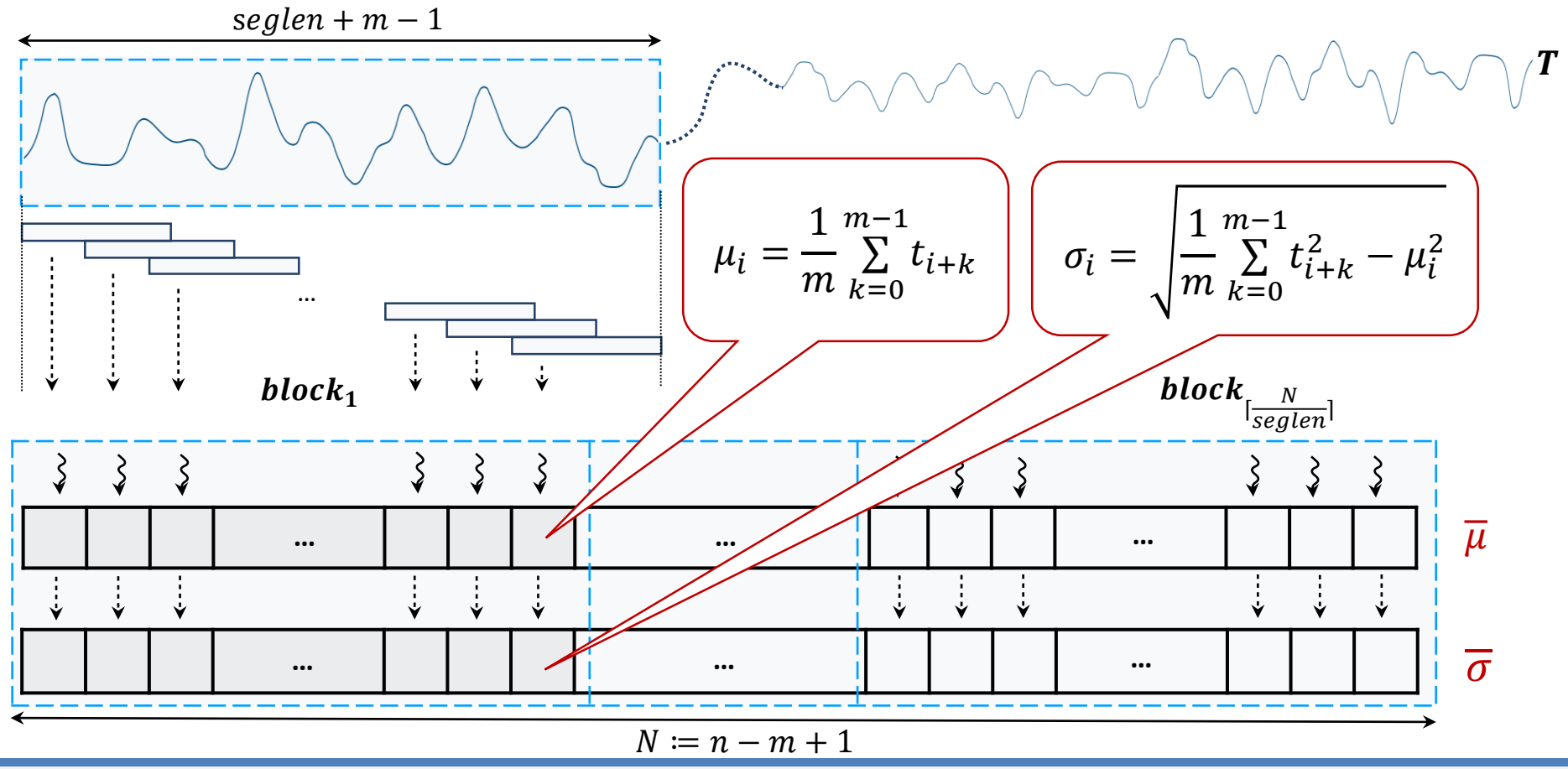
$\mathcal{D} = \{w, y\}$



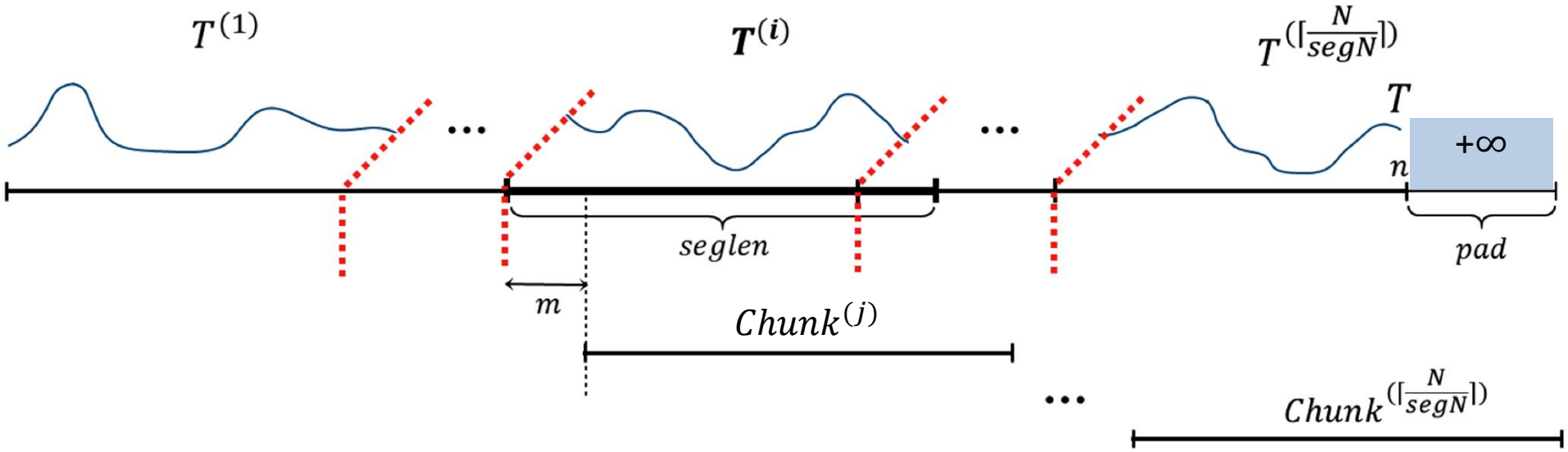
Parallel discord discovery: Preprocessing

For highest possible performance, let us use **quadratic ED**

$$\mathrm{ED}^2_{\mathrm{norm}}(T_{i,m}, T_{j,m}) = 2m\left(1 - \frac{< T_{i,m}, T_{j,m} > - m\mu_i\mu_j}{m\sigma_i\sigma_j}\right)$$

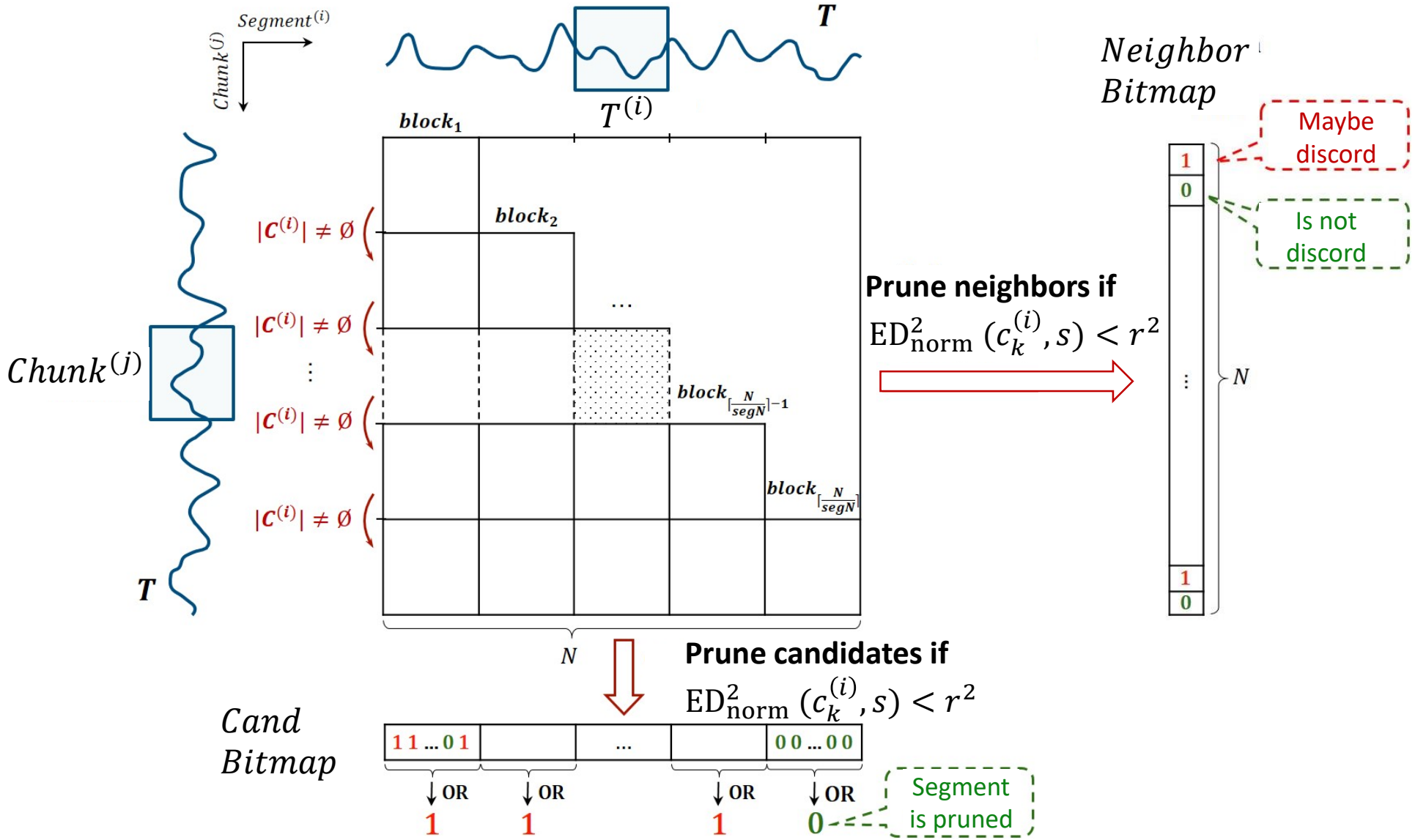


Parallel discord discovery: Segmentation



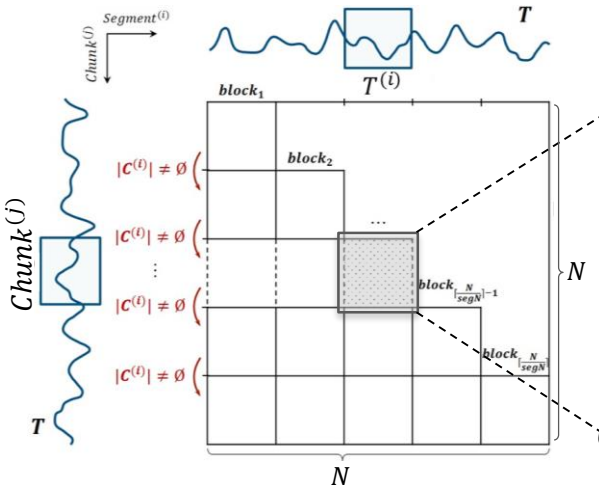
Parameter	Semantic
$T^{(i)}$	Segment to select (prune) candidates
$seglen = segN + m - 1$	Segment length
$segN = k \cdot warpsize$	# candidates in a segment
$warpsize = 32$	# threads in a group within a thread block
$Chunk^{(j)}$	Interval to test its subsequences against a segment candidates
pad	# dummy elements

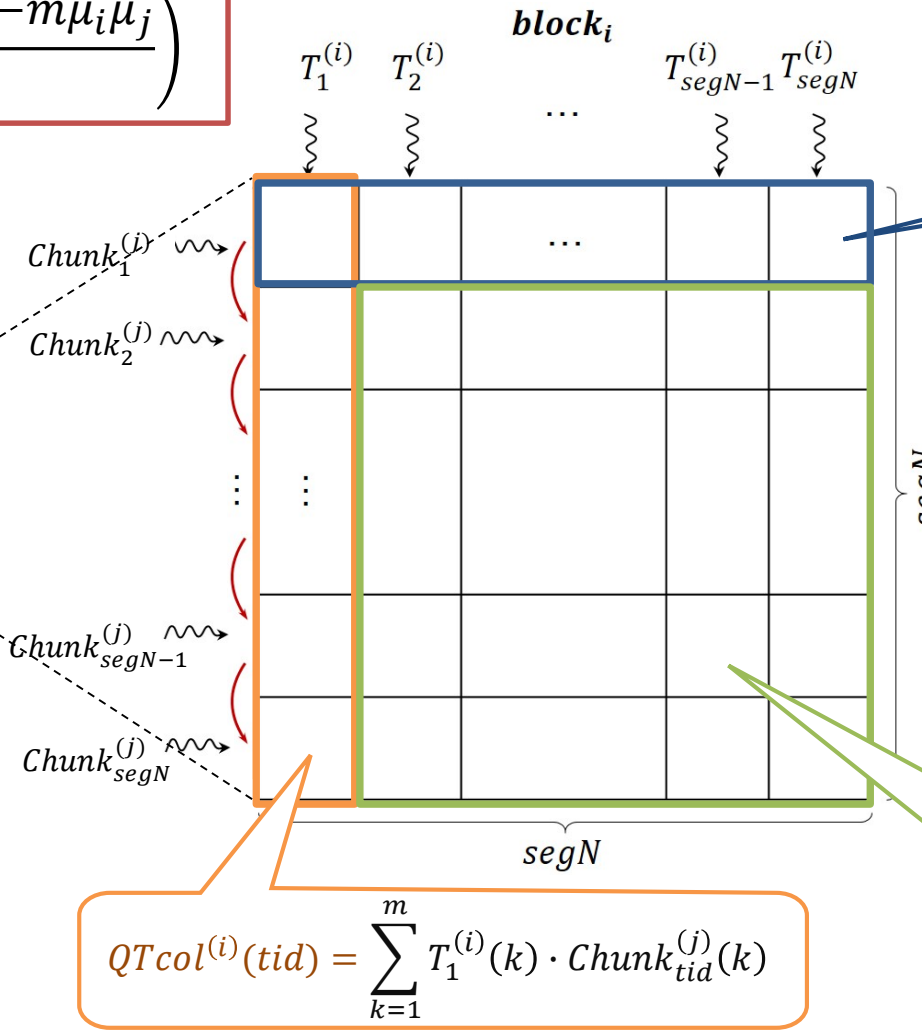
Parallel discord discovery: Selection (blocks)



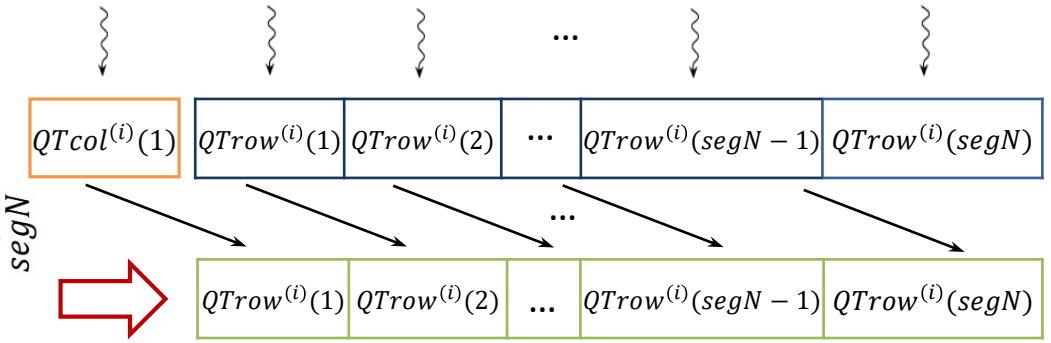
Parallel discord discovery: Selection (threads)

$$d_{i,j} = \mathrm{ED}^2_{\mathrm{norm}}\left(T_{i,m}, T_{j,m}\right) = 2m\left(1 - \frac{< T_{i,m}, T_{j,m} > - m\mu_i\mu_j}{m\sigma_i\sigma_j}\right)$$





$$QTrow^{(i)}(tid) = \sum_{k=1}^{m} T^{(i)}_{tid}(k) \cdot Chunk^{(j)}_1(k)$$

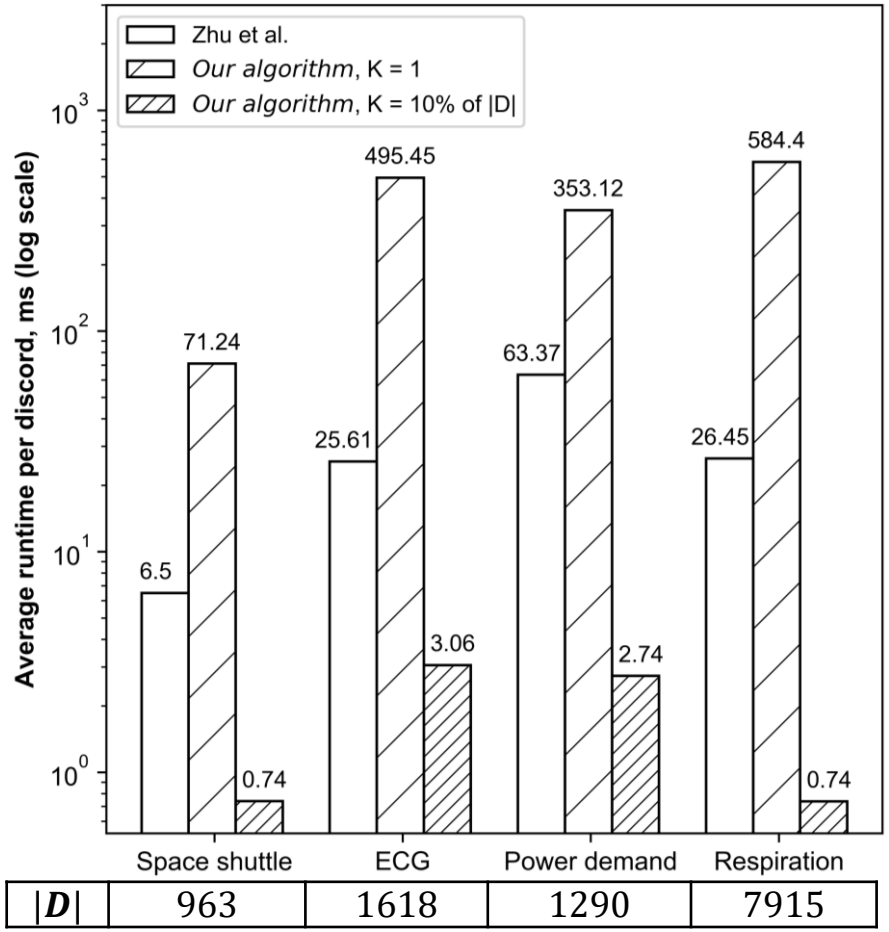


Complexity is $\boldsymbol{O(1)}$ instead of $\boldsymbol{O(m)}$!

$$QTrow^{(i)}(tid) = QTrow^{(i)}(tid-1) - T^{(i)}_{tid-1}(1) \cdot Chunk^{(j)}_{tid-1}(1) + T^{(i)}_{tid}(m) \cdot Chunk^{(j)}_{tid}(m)$$

$$QTcol^{(i)}(tid) = \sum_{k=1}^{m} T^{(i)}_1(k) \cdot Chunk^{(j)}_{tid}(k)$$

Parallel discords discovery: Experiments on performance



$\|D\|$	963	1618	1290	7915

Parallel algorithm
is up to **30 times ahead** of rival
w.r.t. average running time to discover one discord

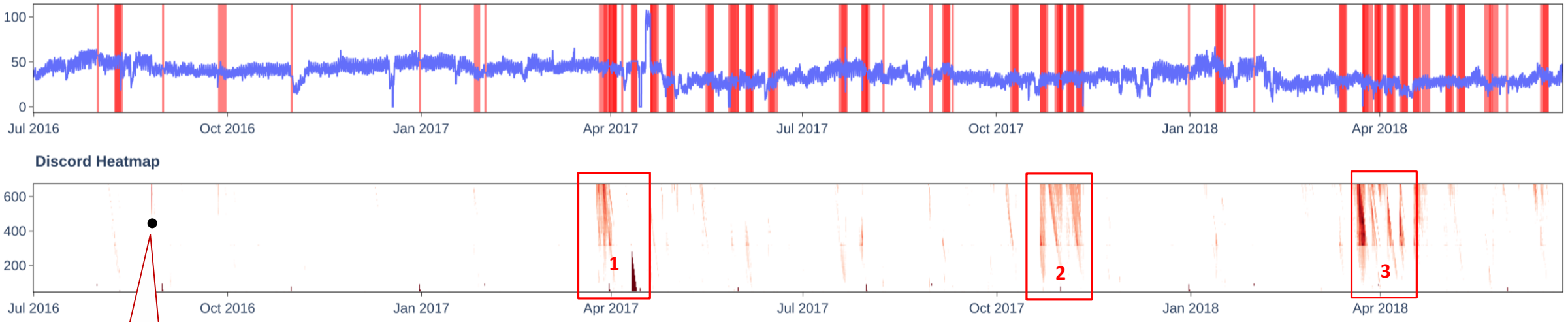
Hardware: NVIDIA Tesla P100 (3584 cores @1.19 GHz)

Time series	Length	Discord	Domain
Space shuttle	5 000	150	Measurements of a sensor on the NASA spacecraft[1]
ECG	45 000	200	ECG of an adult patient[2]
Power demand	33 220	750	Annual energy consumption of an office[3]
Respiration	24 125	250	Human breathing by chest expansion[4]

[1] Ferrell B., *et al.* NASA shuttle valve data 2005. URL: http://www.cs.fit.edu/~pkc/nasa/data/
[2] Goldberger A., *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation 101(23): 215–220. DOI: 10.1161/01.CIR.101.23.e215
[3] van Wijk J.J., *et al.* Cluster and calendar based visualization of time series data. INFOVIS'99: 4–9. DOI: 10.1109/INFVIS.1999.801851
[4] Keogh E., *et al.* HOT SAX: Finding the most unusual time series subsequence: Algorithms and applications. ICDM 2004: 440–449. URL: http://www.cs.ucr.edu/~eamonn/discords/
[5] Zhu B. *et al.* A GPU Acceleration framework for motif and discord based pattern mining. IEEE Transactions on Parallel and Distributed Systems 32(8): 1987–2004. 2021. DOI: 10.1109/TPDS.2021.3055765

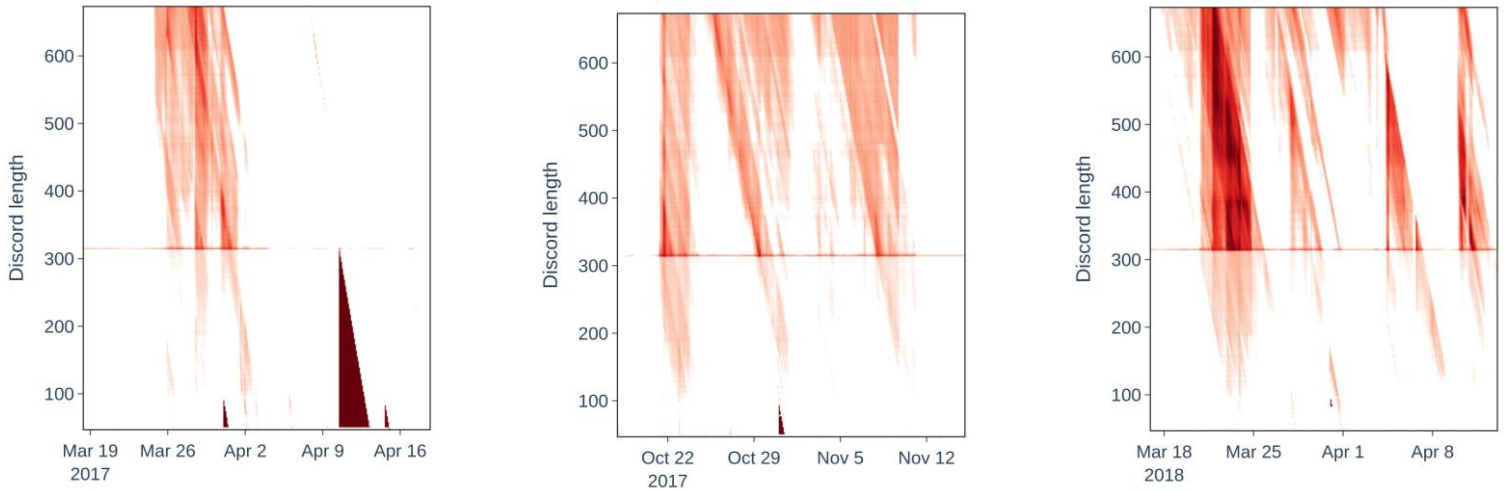
Parallel discords discovery: Experiments on accuracy

2-year power demand (Beijing Guowang Fuda Sci. & Tech. Dev. Co.)[1]



Minimax normalization

$$heatmap_{i,m} = \frac{T_{i,m}.nndist - \min\limits_{d \in D_i} d.nndist}{\max\limits_{d \in D_i} d.nndist - \min\limits_{d \in D_i} d.nndist}, T_{i,m} \in D_i$$



[1] Zhou H. *et al.* Informer: beyond efficient transformer for long sequence time-series forecasting. AAAI 2021: 11106-11115. DOI: 10.1609/aaai.v35i12.17325.

Conclusions: We are ready for further collaboration

- Applying time series analytics to Smart DBMS
  - Anomaly detection to monitor database activities
  - Hardware lifecycle prediction
  - Workload prediction based on resource usage patterns
- Embedding time series analytics into DBMS
  - In-DBMS matrix profile support
  - Table data imputation: on the fly and/or in background
- Online time series analytics of mobile users
  - Activity recognition
  - Anomaly detection
  - Data imputation

Big Data and Machine Learning Laboratory

Mikhail Zymbler
Dr.Sci.
Assist. Prof.

Yana Kraeva
MSc,
2yr PhD student

Andrey Goglachev
MSc,
1yr PhD student

Alexey Yurtin
MSc,
1yr PhD student