Scientific seminar

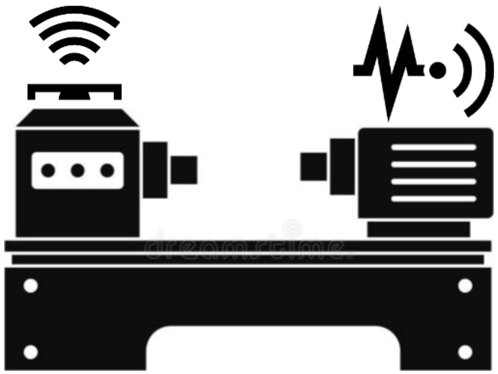
SUSU Big Data and Machine Learning Lab

Method of imputation missing values in multivariate streaming time series

Alexey Yurtin, Mikhail Zymbler

Chelyabinsk—2022

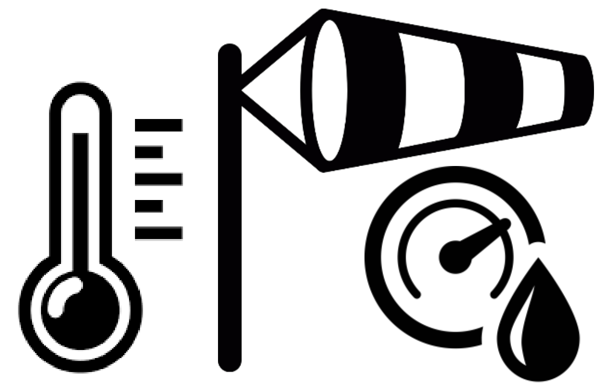
Processing of streaming time series



Predictive maintenance,
smart manufacturing

Internet
of Things

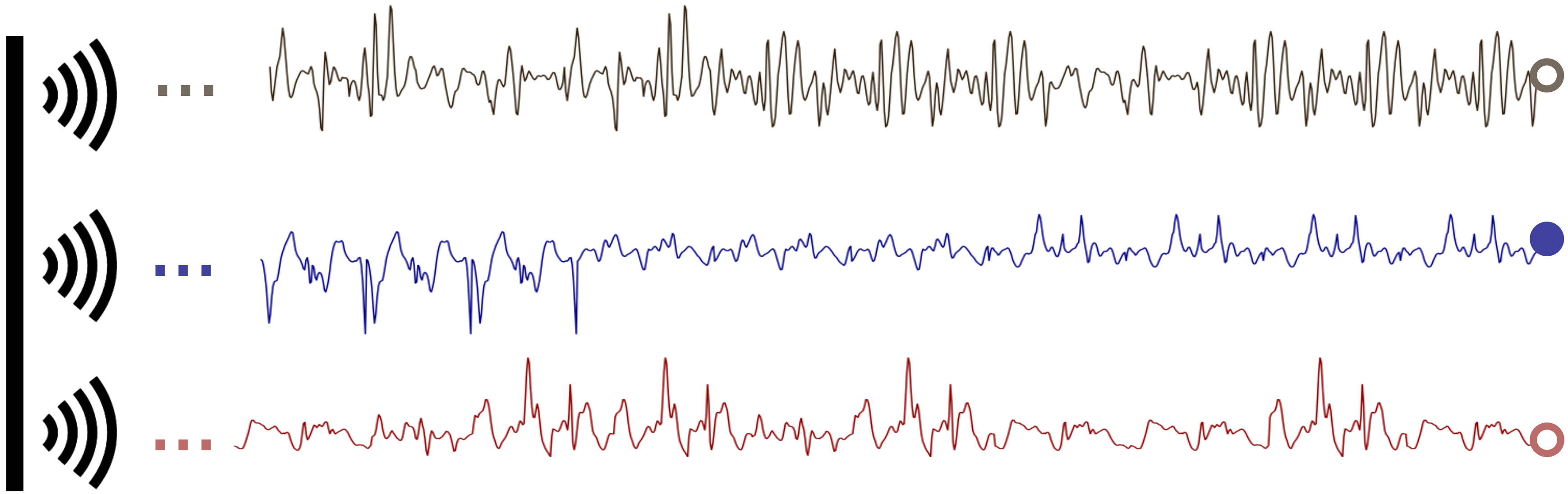
Personal
healthcare



Weather forecasting,
climate modelling

Elements of a time series
arrive one after another
in several dimensions in an online mode

Imputation of missing values

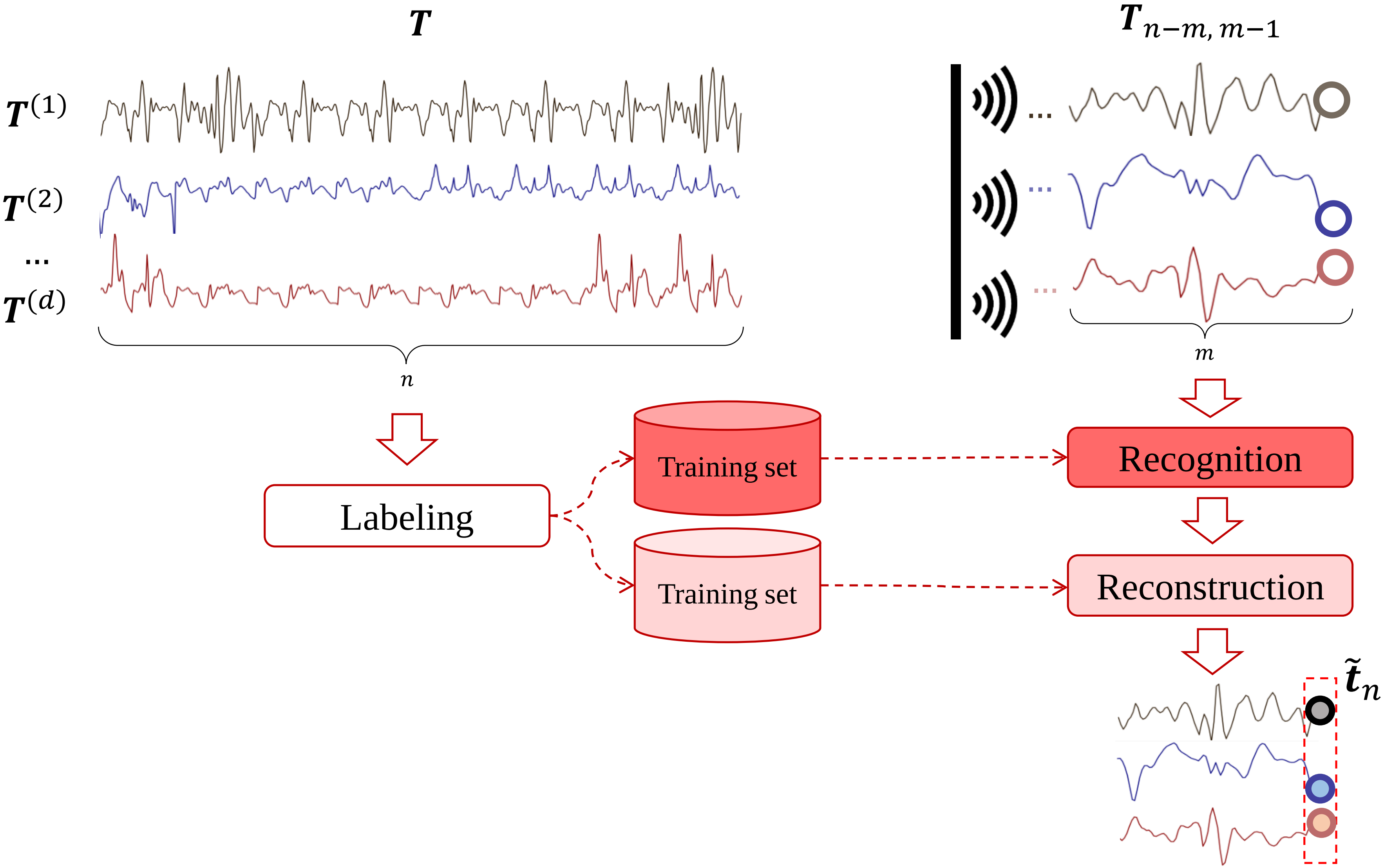


How to plausibly and quickly synthesize missing values?

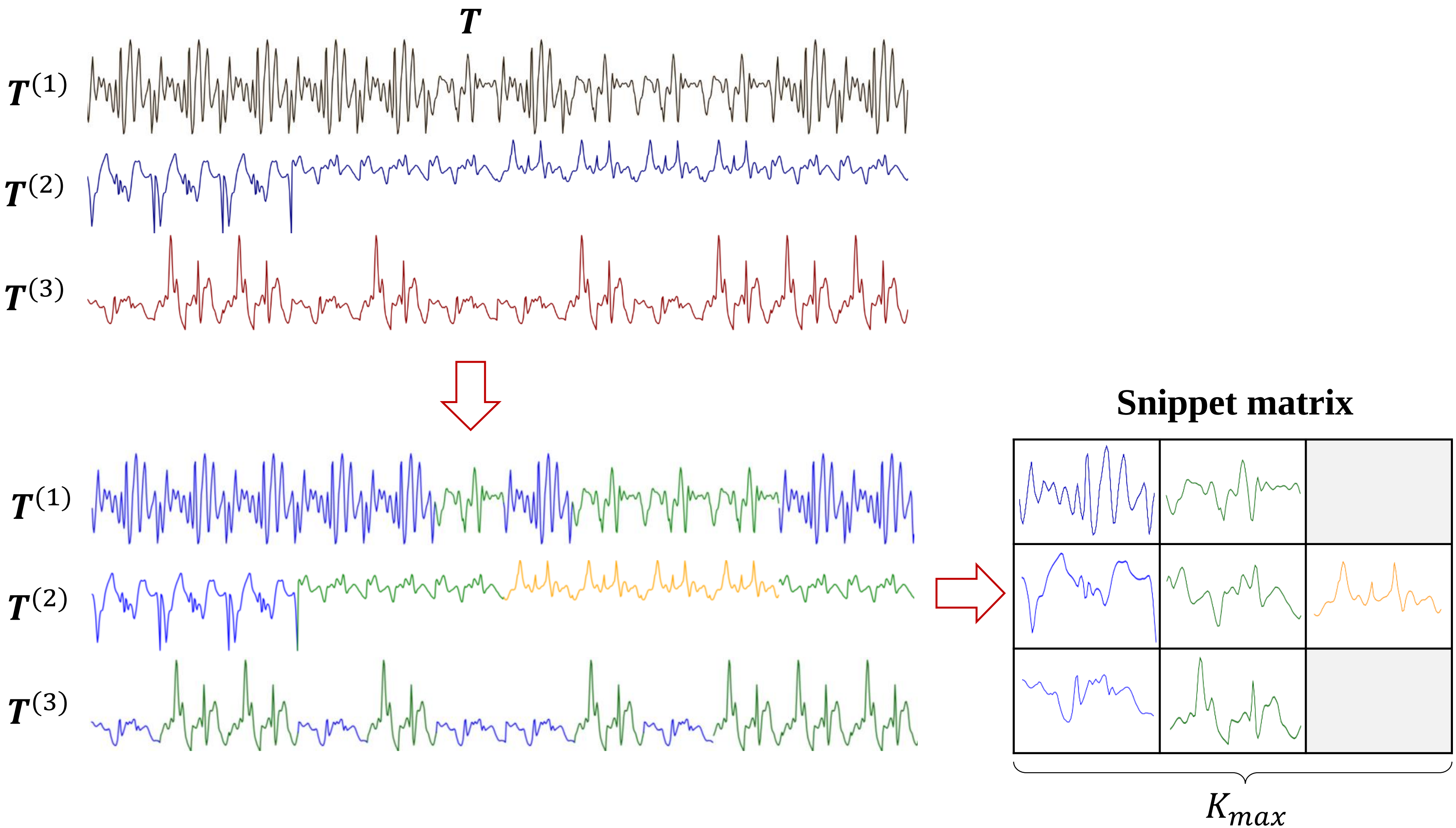
SANNI: Snippet and ANN-based Imputation

1. **Labeling**
   – Automatically label a representative fragment of the time series undergo imputation through the **snippet concept**
2. **Recognition**
   – Recognize a snippet in a subsequence before the missing value through the **convolutional neural networks**
3. **Reconstruction**
   – Generate a missing value based on the recognized snippet and **recurrent neural networks**

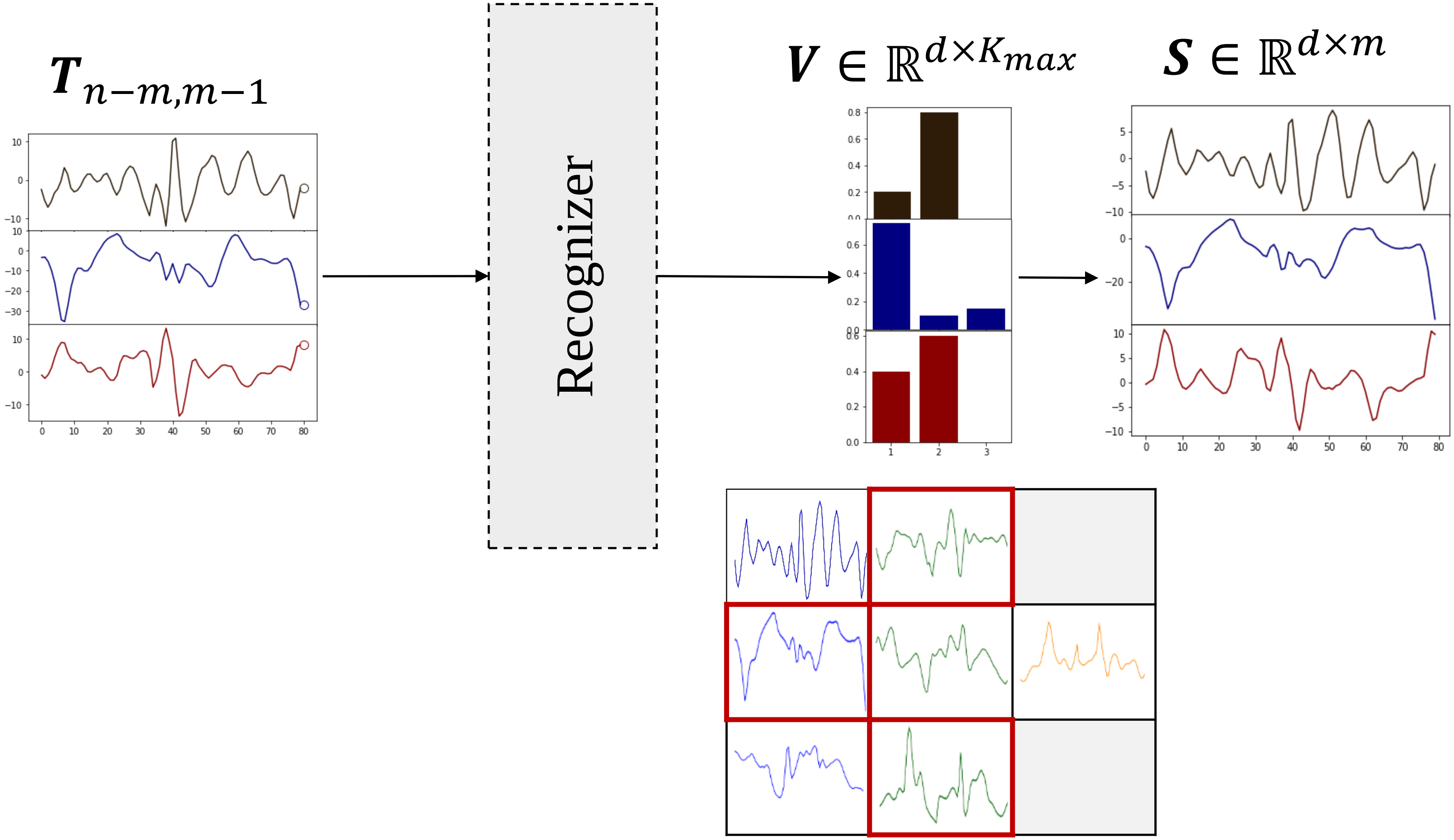
SANNI: Architecture



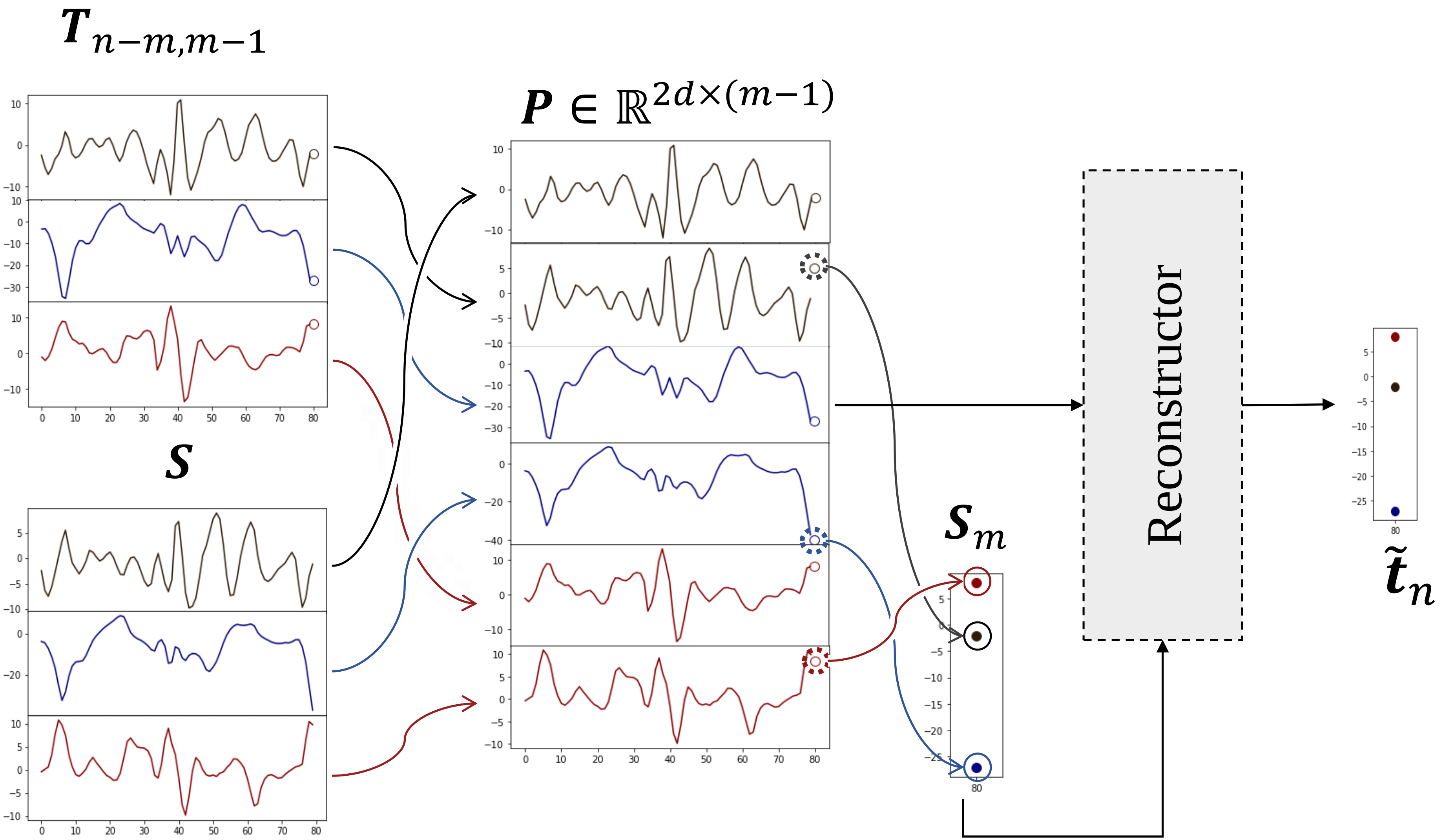
SANNI: Labeling



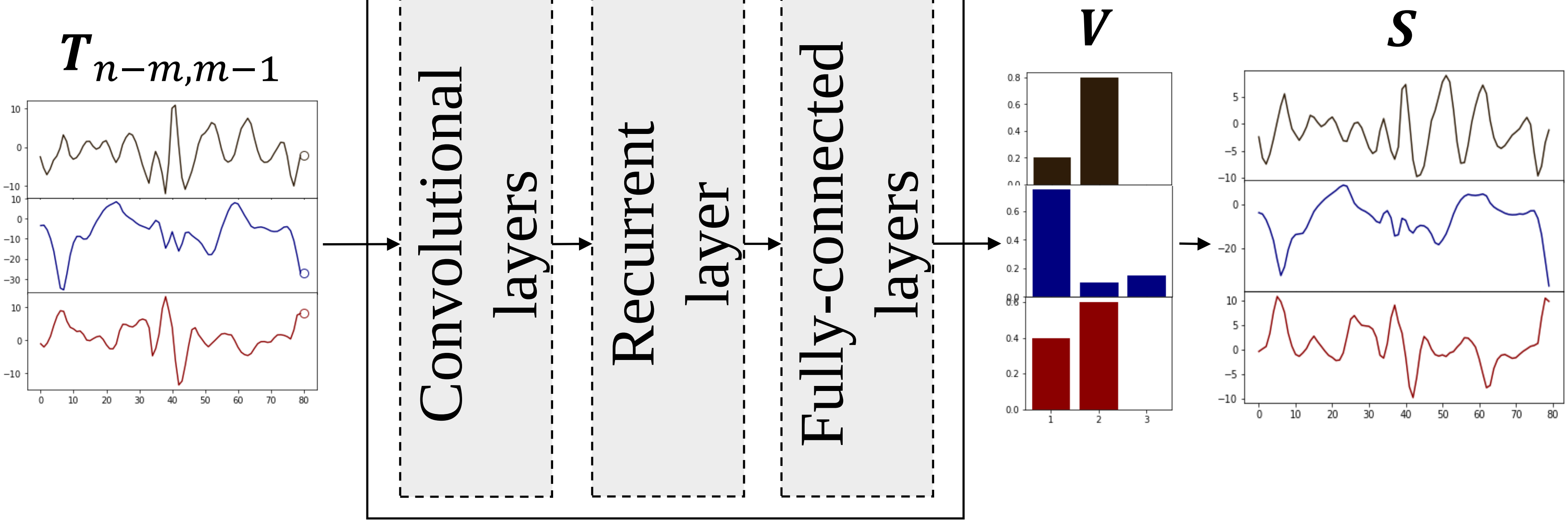
SANNI: Recognition



SANNI: Reconstruction

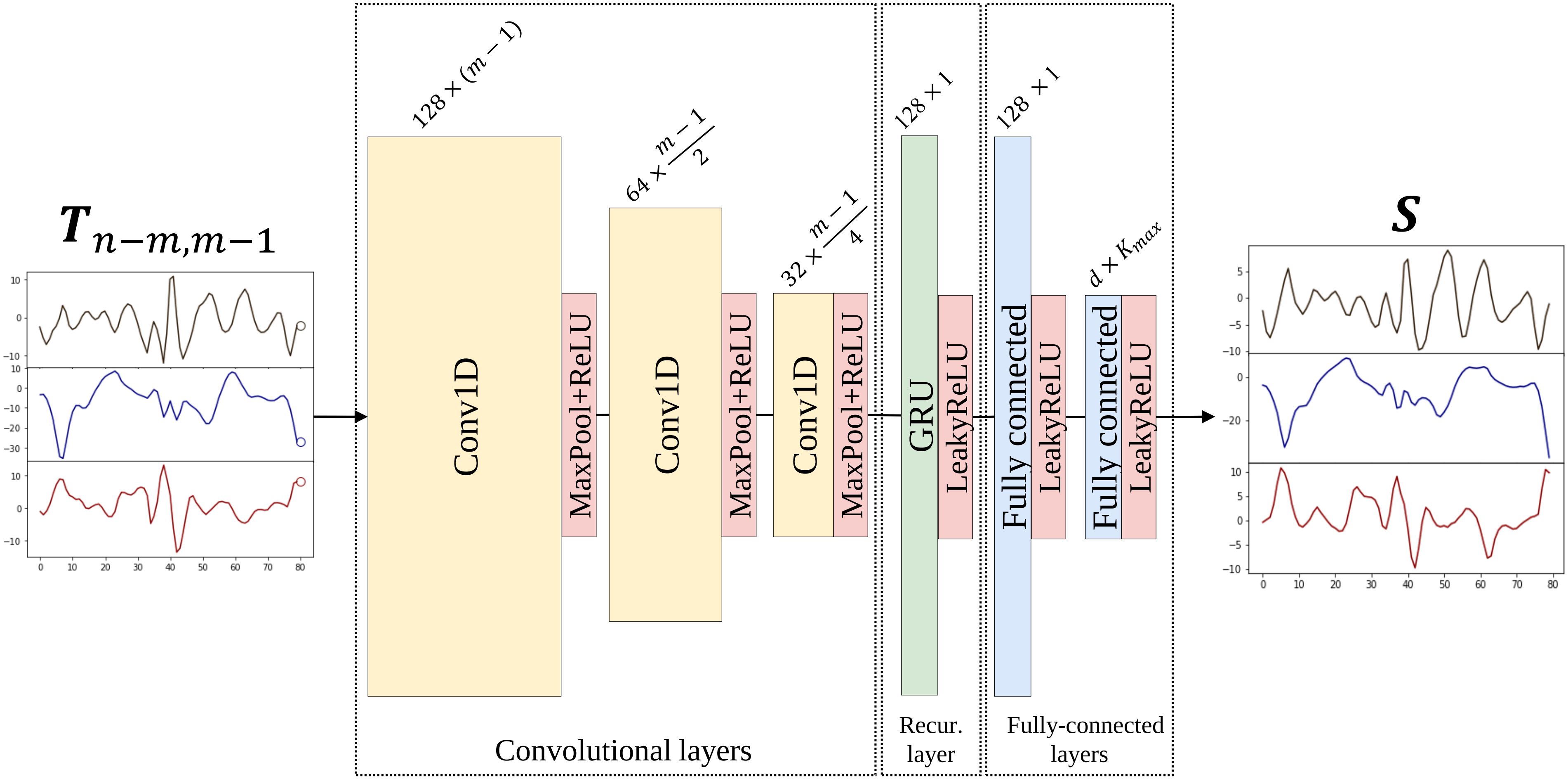


Recognizer

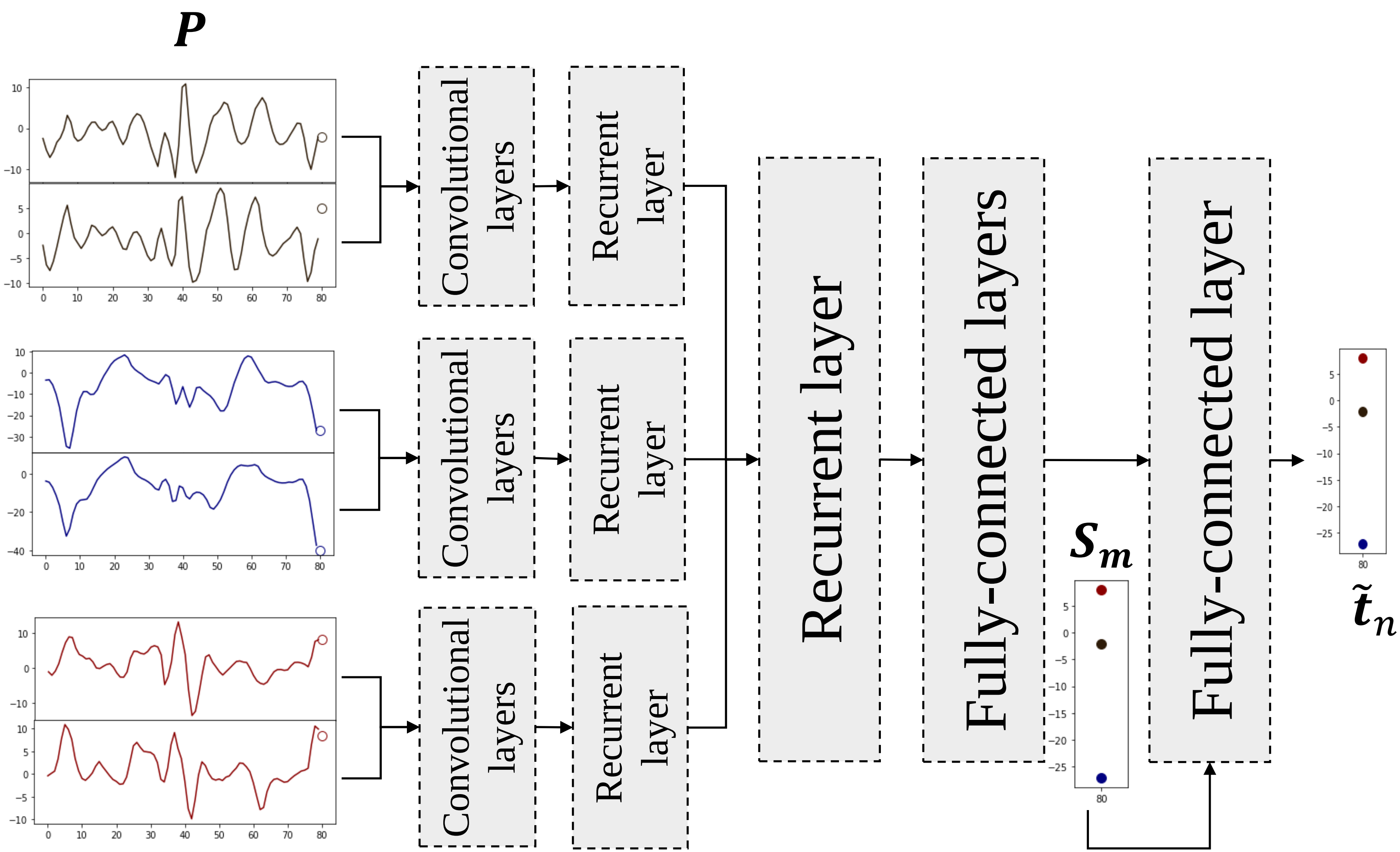


$T_{n-m,m-1}$
Convolutional layers
Recurrent layer
Fully-connected layers
V
S

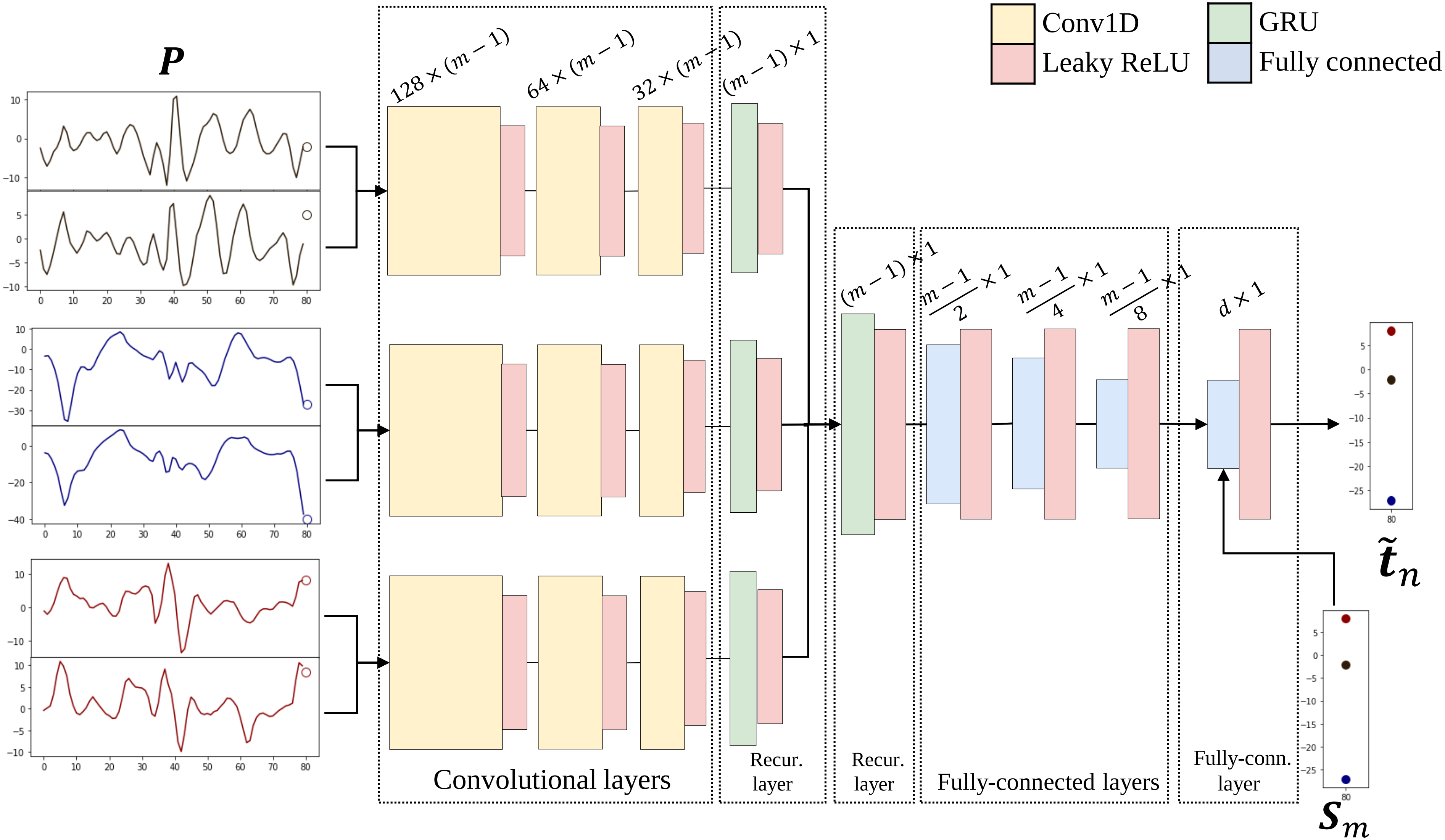
Recognizer


$T_{n-m,m-1}$
$128 \times (m-1)$
Conv1D
MaxPool+ReLU
$64 \times \frac{m-1}{2}$
Conv1D
MaxPool+ReLU
$32 \times \frac{m-1}{4}$
Conv1D
MaxPool+ReLU
128×1
GRU
LeakyReLU
128×1
Fully connected
LeakyReLU
$d \times K_{max}$
Fully connected
LeakyReLU
S
Convolutional layers
Recur. layer
Fully-connected layers

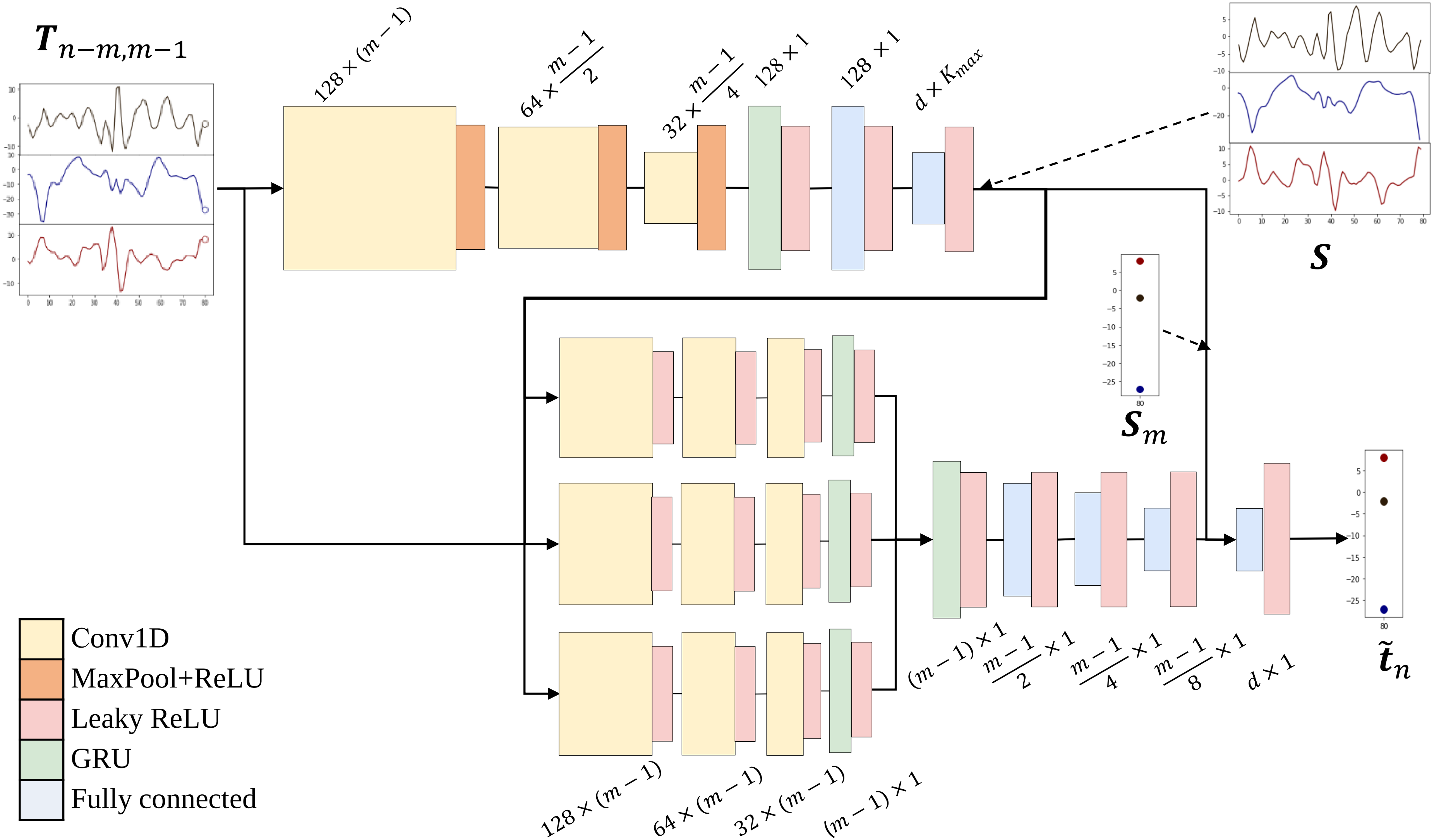
Reconstructor



Reconstructor

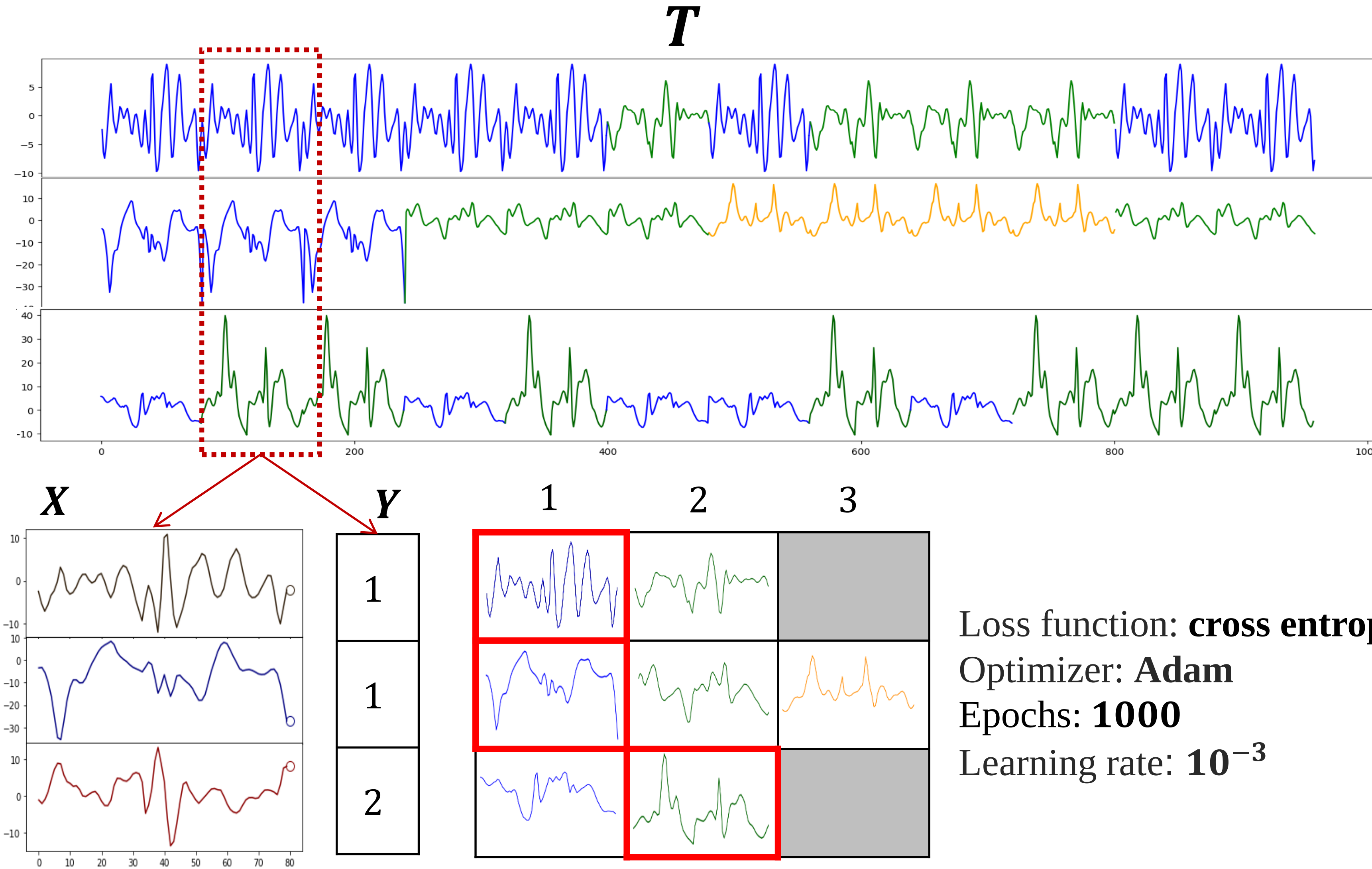


SANNI



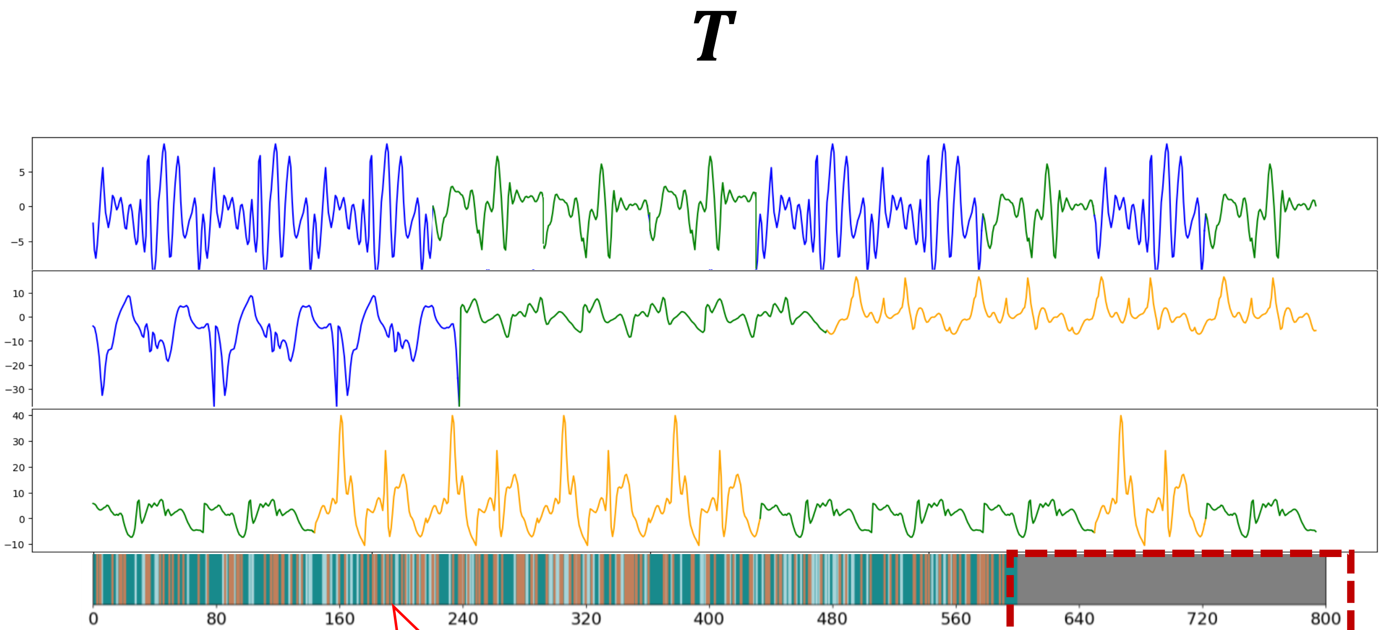
$T_{n-m,m-1}$
$128\times(m-1)$
$64\times\frac{m-1}{2}$
$32\times\frac{m-1}{4}$
128×1
128×1
$d\times K_{max}$
S
S_m
$\tilde{t}_n$
$(m-1)\times1$
$\frac{m-1}{2}\times1$
$\frac{m-1}{4}\times1$
$\frac{m-1}{8}\times1$
$d\times1$
$128\times(m-1)$
$64\times(m-1)$
$32\times(m-1)$
$(m-1)\times1$
Conv1D
MaxPool+ReLU
Leaky ReLU
GRU
Fully connected

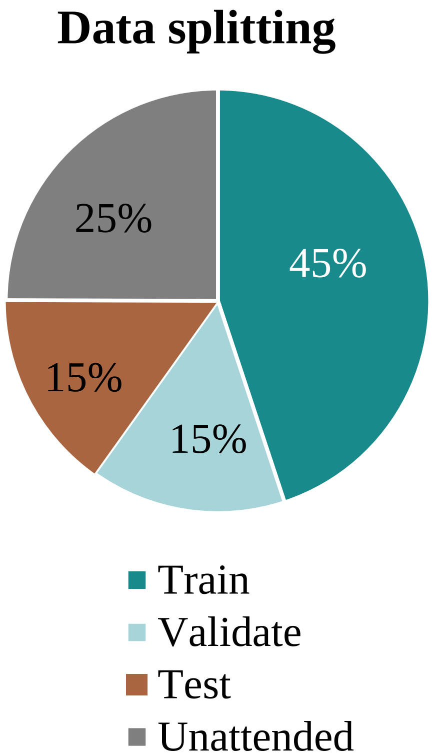
Learning of Recognizer



Loss function: **cross entropy**
Optimizer: **Adam**
Epochs: **1000**
Learning rate: $\mathbf{10^{-3}}$

Learning of Recognizer

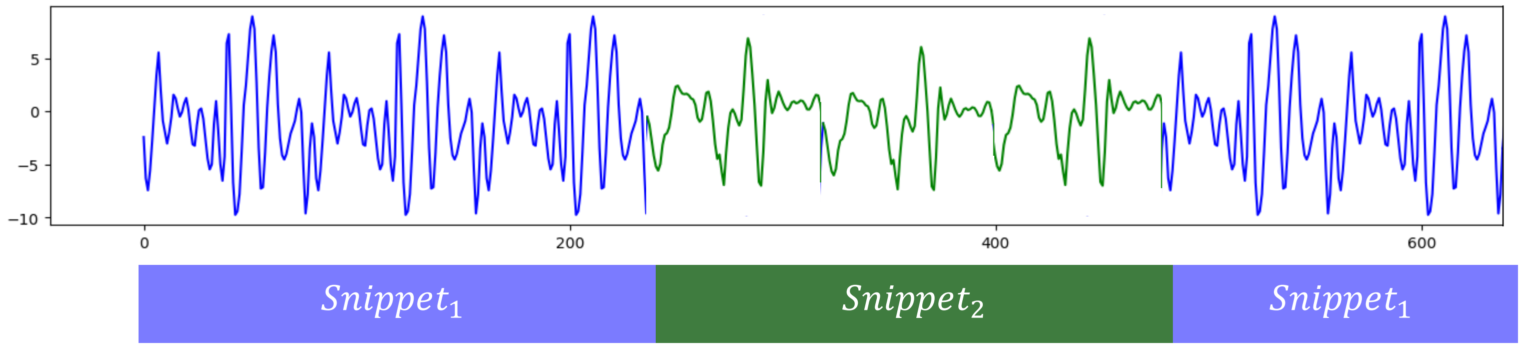


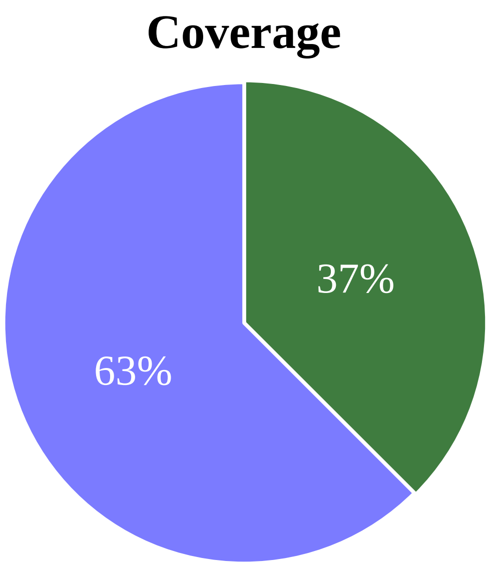


Shuffle to evenly distribute snippets across the training set

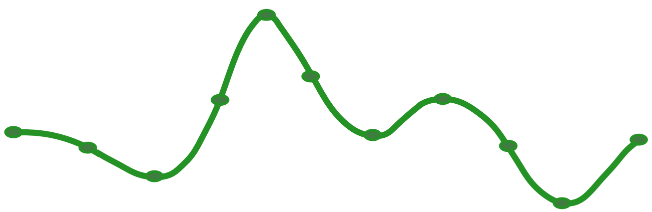
Leave for the SANNI testing

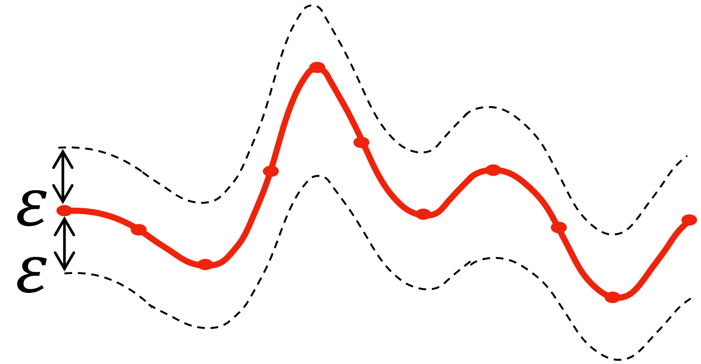
Snippets with small coverage





Augmentation of snippets with small coverage





$$\varepsilon = \mathrm{ED}(snippet, neighbor)$$

$snippet.NN$ is a set of all the nearest neighbors of the $snippet$

$$\forall neighbor \in snippet.NN\ \exists c \in R^m\colon \sum_{k=1}^{m} c_k = \varepsilon,$$

$$synthetic1 = neighbor + c, \qquad synthetic2 = neighbor - c$$

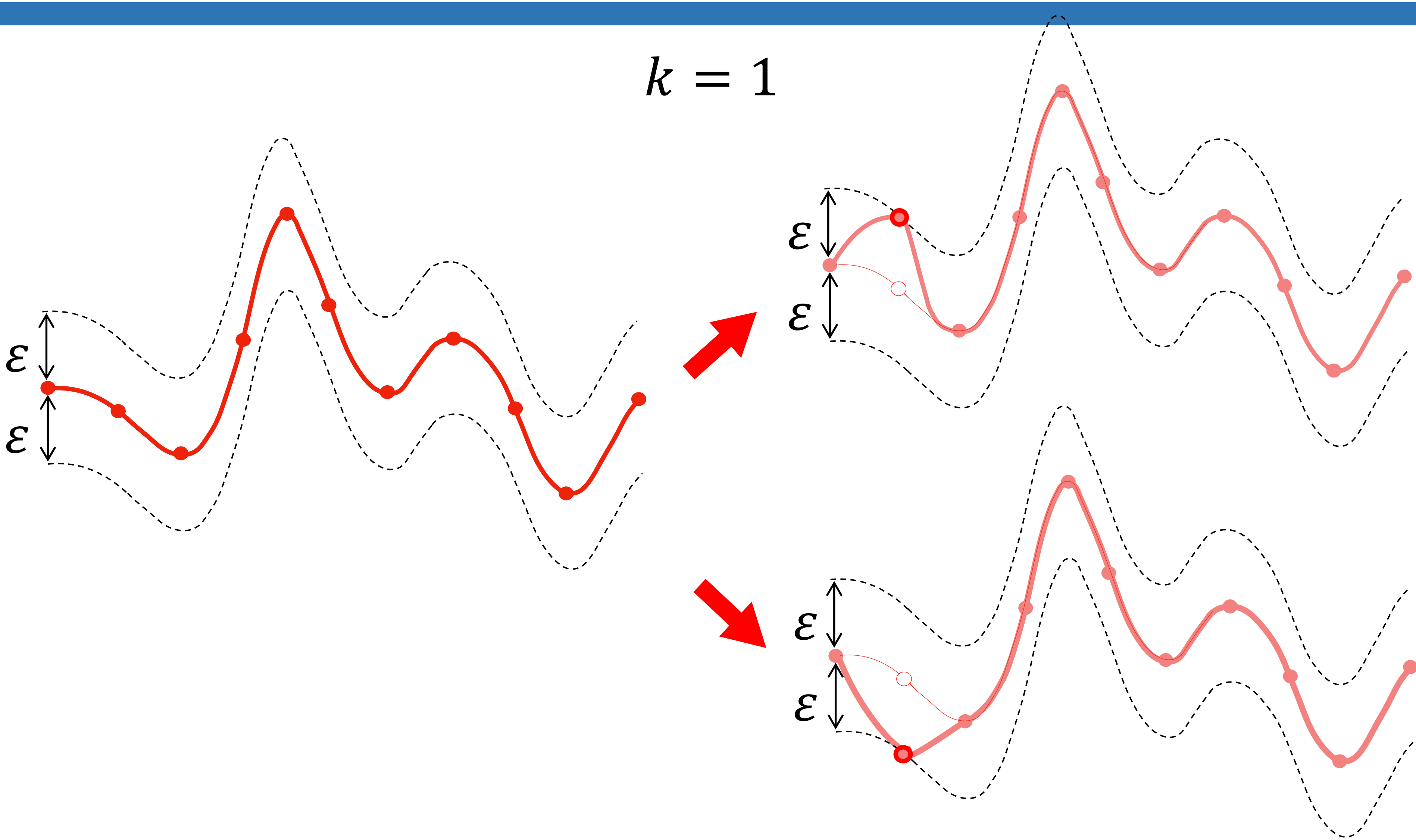
$$\mathrm{ED}(snippet, synthetic1) < \varepsilon, \qquad \mathrm{ED}(snippet, synthetic2) < \varepsilon$$

Number of synthetic neighbors is $C_m^{m+k-1} = \dfrac{(m+k-1)!}{m!\,(k-1)!}$ *

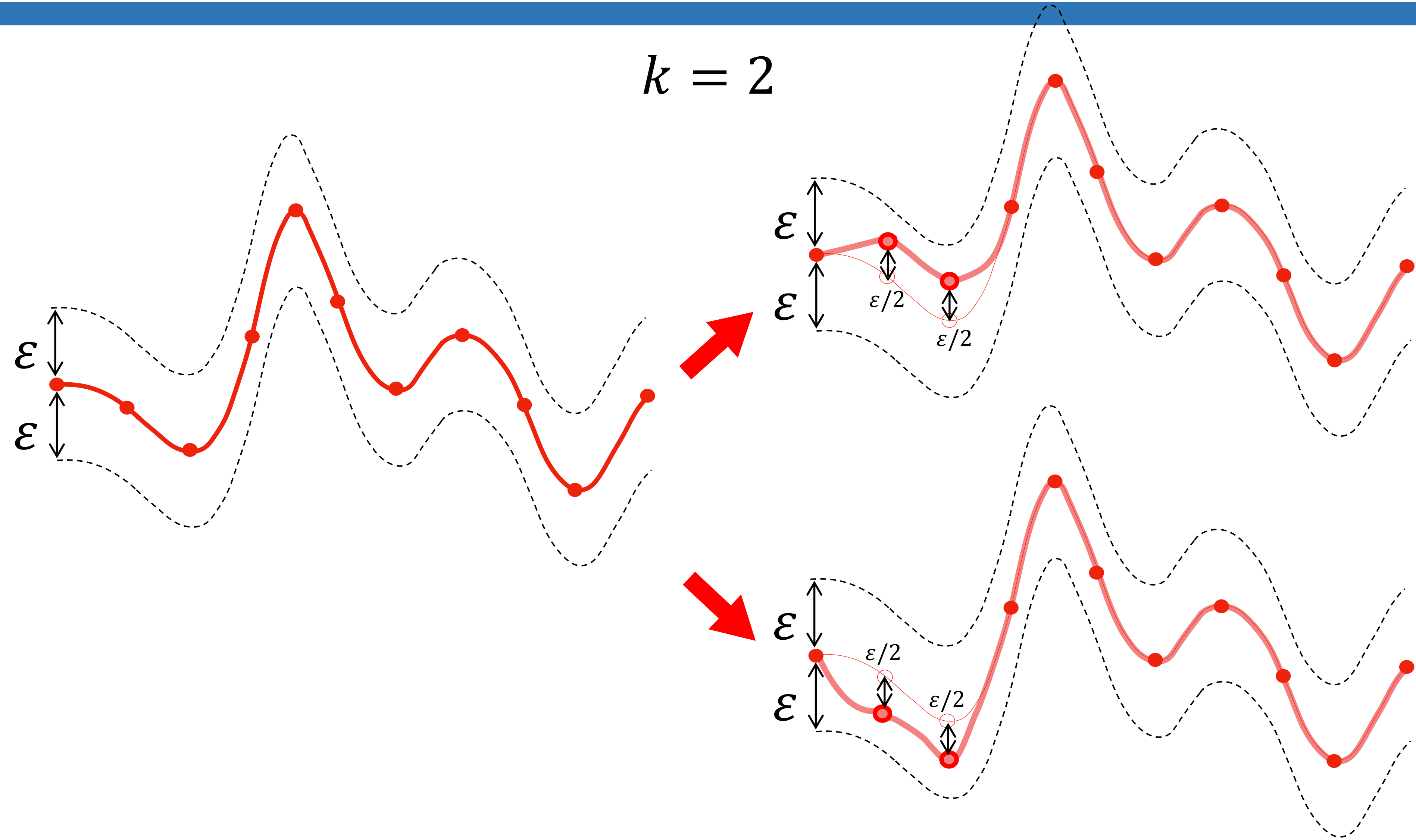
* Reingold E. *et al*. Combinatorial Algorithms: Theory and Practice. Prentice Hall, 1977. 433 p.

Synthetic neighbors of a small-coverage snippet

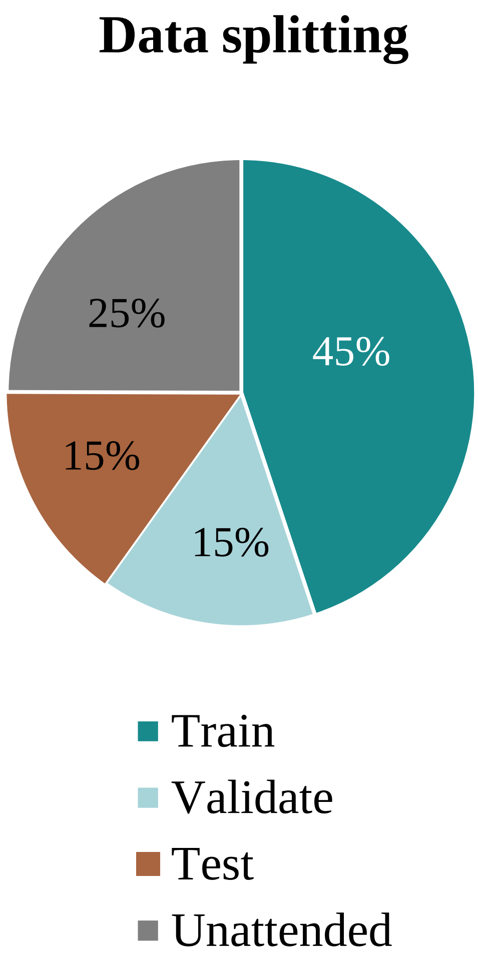
$k = 1$

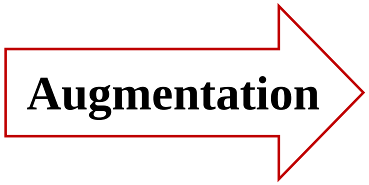


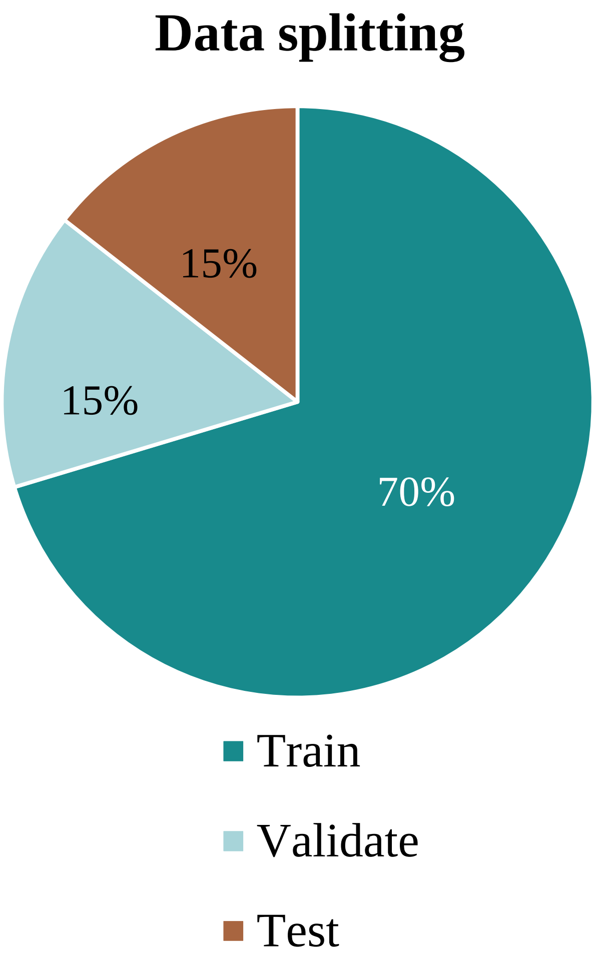
Synthetic neighbors of a small-coverage snippet



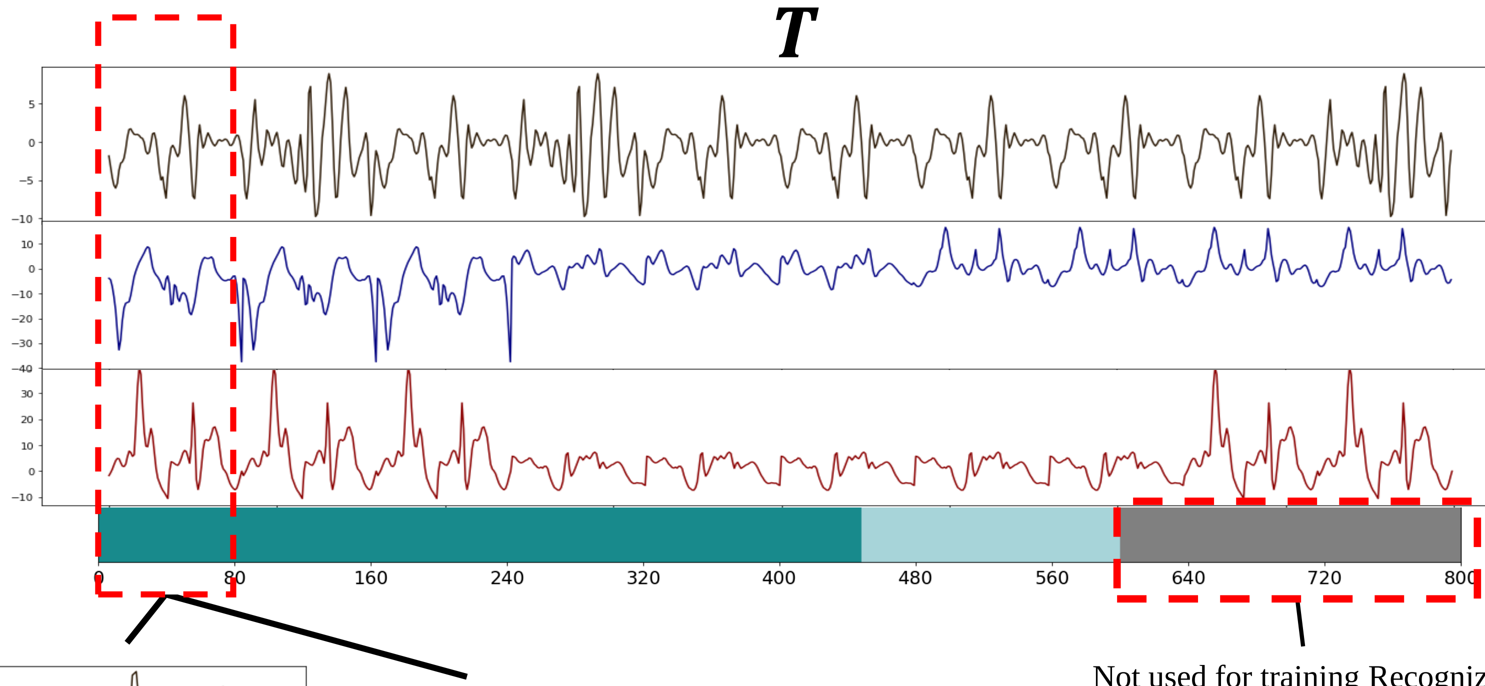
Learning Recognizer with augmentation data

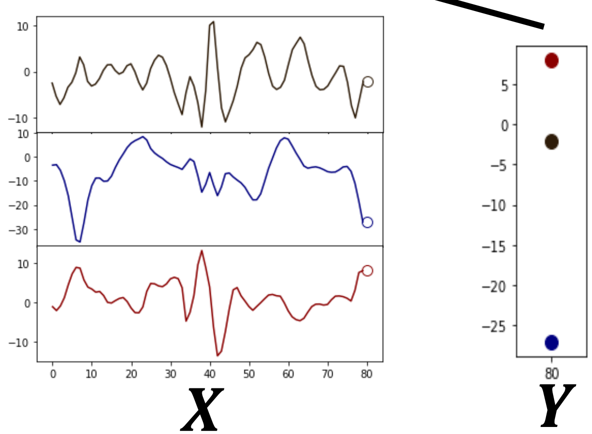




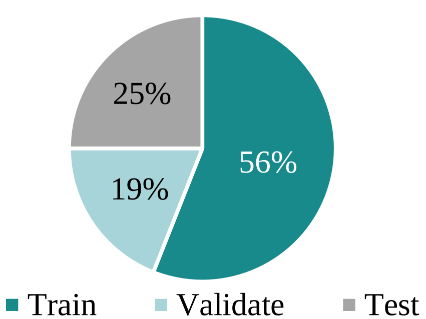


Learning of Reconstructor





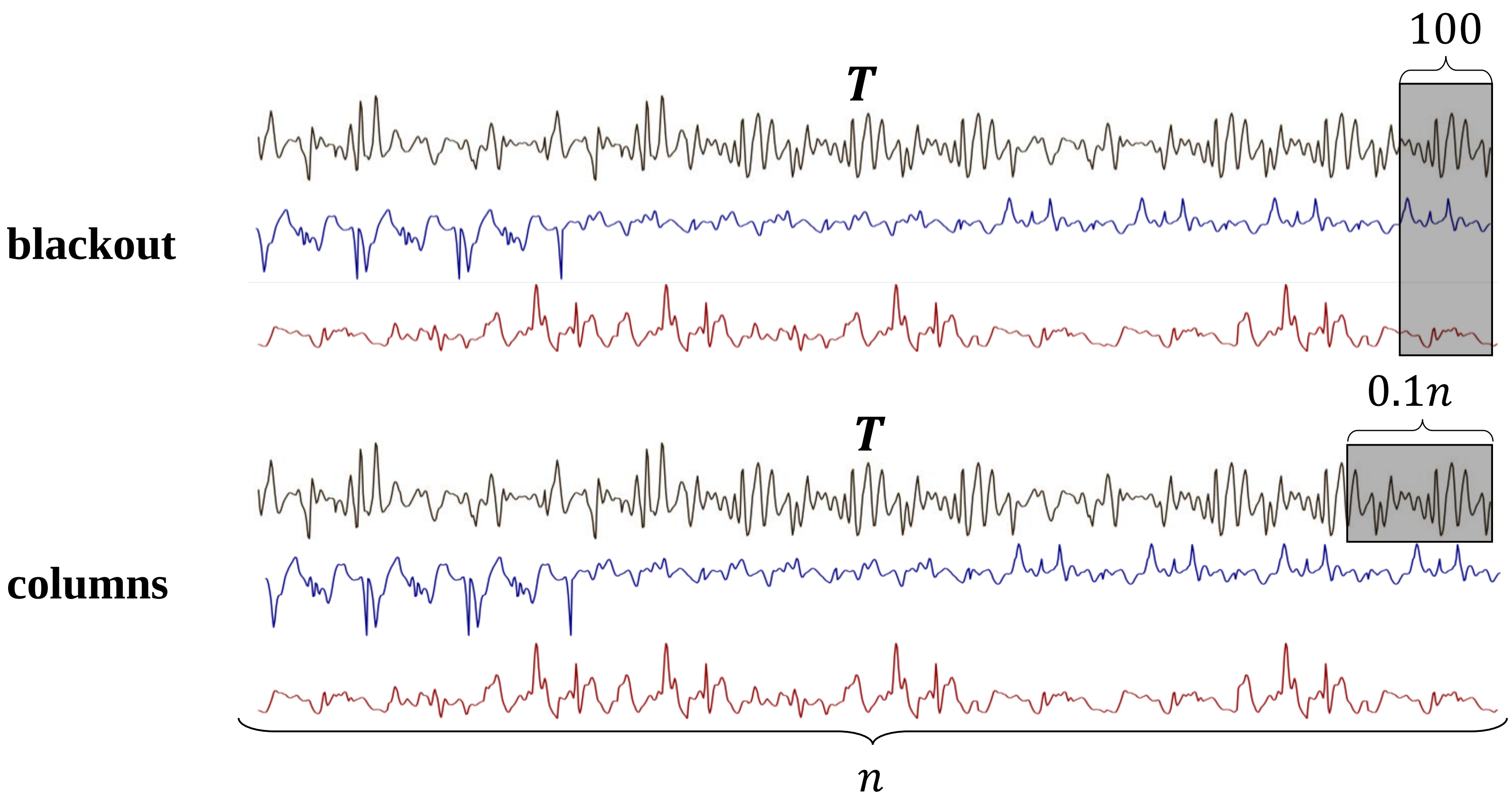
Data splitting



Loss function: **Mean Square Error**
Optimizer: **Adam**
Epochs: **1500**
Learning rate: $\mathbf{10^{-3}}$

Experiments: Setup

- GPU: NVIDIA Tesla V100 SXM2 (5120 cores @1.312 GHz, memory 32 Gb).
- Scenario under ORBITS[1] framework:



[1] Khayati M., *et al*. ORBITS: Online Recovery of Missing Values in Multiple Time Series Streams. Proc. VLDB Endow. 2020. Vol. 14, no. 3. P. 294-306.

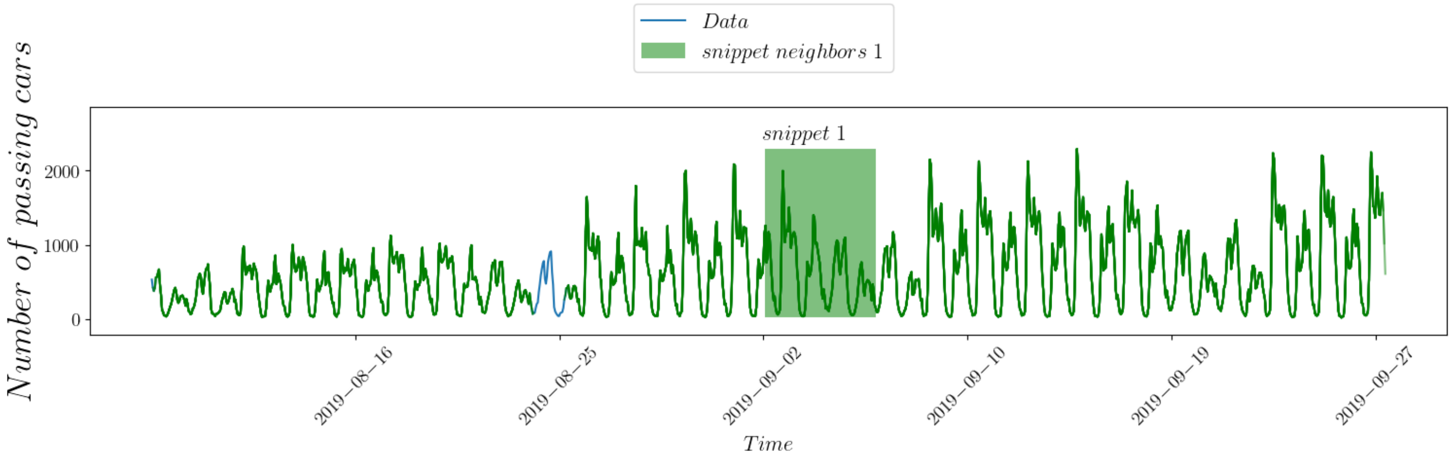
Experiments: Datasets

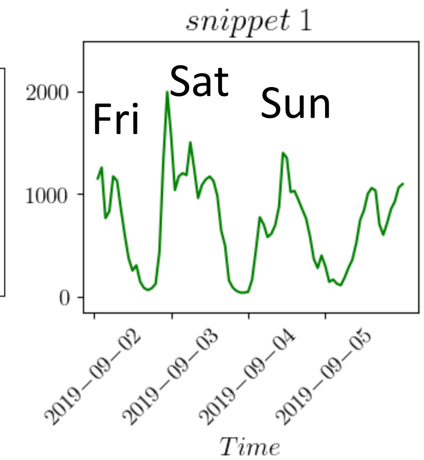
With activities
No activities

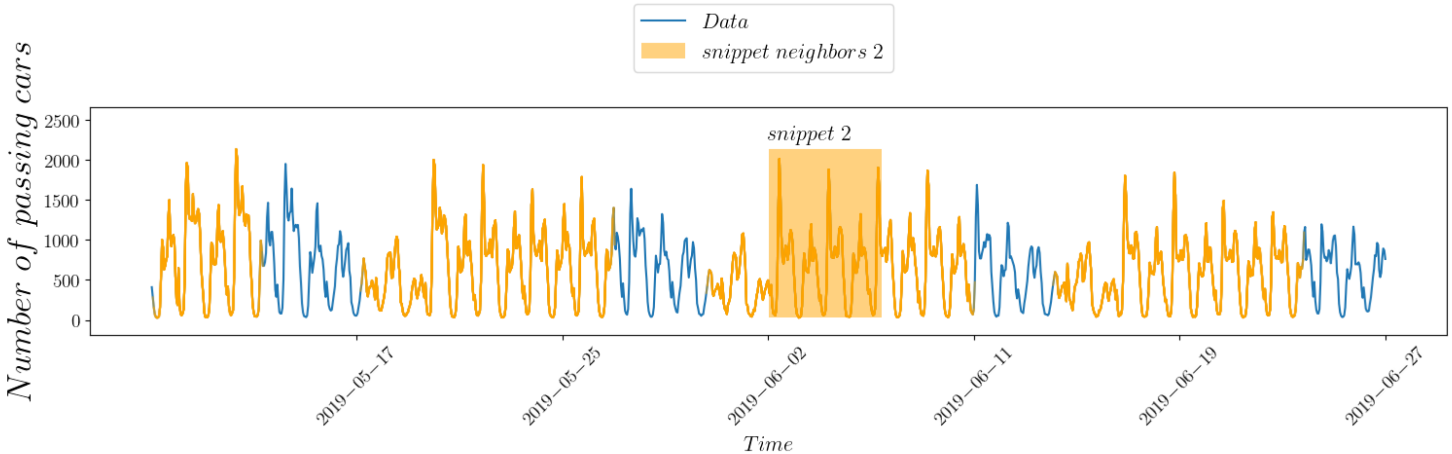
Dataset	Length n, $\times 10^3$	Dim. d	Snippet length m	Domain
Airq[1]	1	10	100	Air pollution in a city
BAFU[1]	50	12	48	Water discharge in Swiss rivers
Electricity[1]	5	20	120	Power demand of several consumers
Temp[1]	5	50	50	Daytime air temperature in different China regions
Gas[1]	3	100	100	Gas concentration measurements of a smart gas delivery platform at San Diego, US
Motion[2]	10	20	120	Smartphone accelerometer data
MADRID[2]	29	10	110	Road traffic (AVR statistics) in Madrid
WalkRun[3]	220	19	100	Accelerometer, gyroscope, and magnetometer measurements during alternation of running and walking
WalkStop[3]	140	19	250	Accelerometer, gyroscope, and magnetometer measurements during alternation of walking and still standing

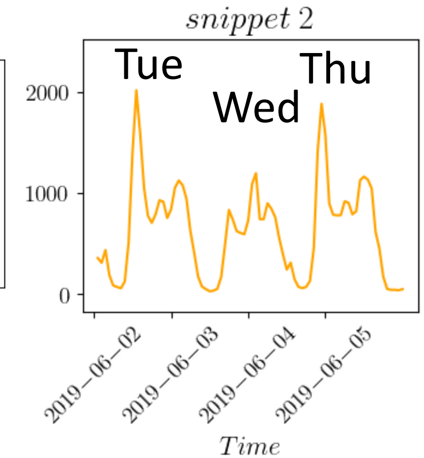
[1] Khayati M., *et al.* ORBITS: Online Recovery of Missing Values in Multiple Time Series Streams. Proc. VLDB Endow. 2020.
[2] Zymbler M.L., Poluyanov A.N., Kraeva Ya.A. Parallel Algorithm for Real-time Sensor Data Recovery for a Many-core Processor. Bulletin of the South Ural State University. Series: Computational Mathematics and Software Engineering. 2022. Vol. 11, no. 3. P. 68–89. (in Russian) DOI: 10.14529/cmse220305
[3] Our generated dataset

MADRID: dataset with activities



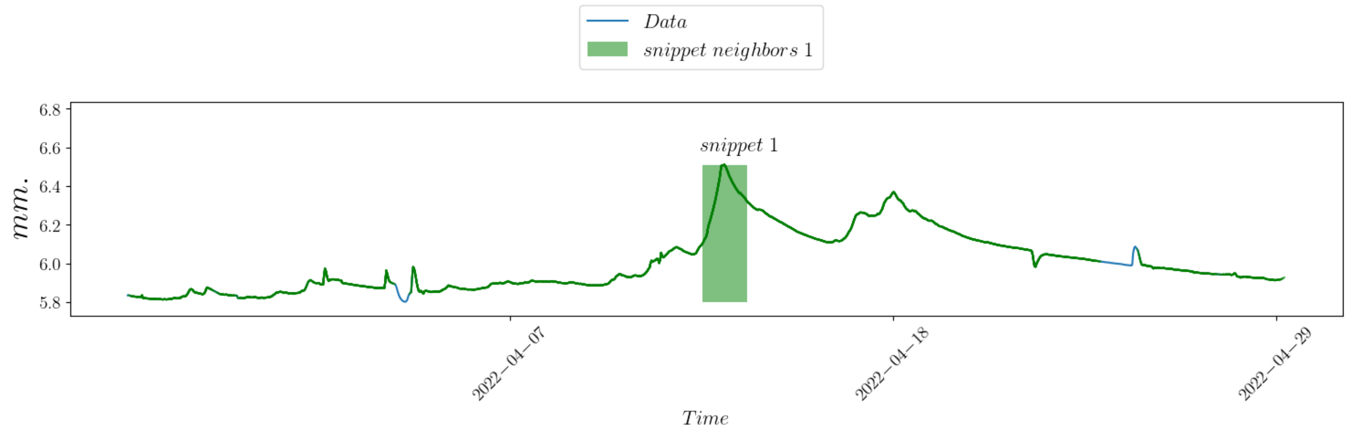


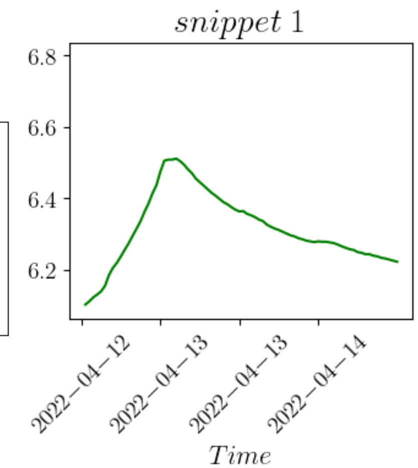


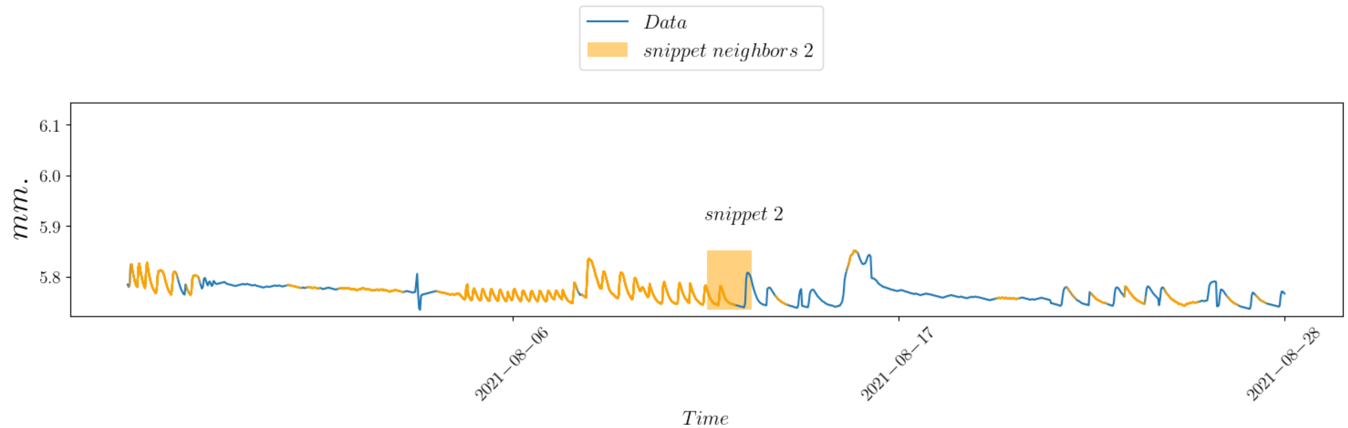


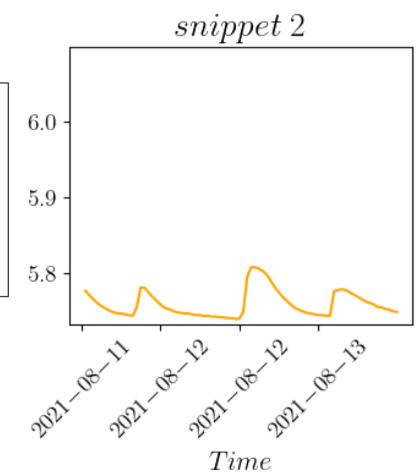
Snippet length is 110 points (4 days)

BAFU: dataset no activities









Snippet length is 48 points (3 days)

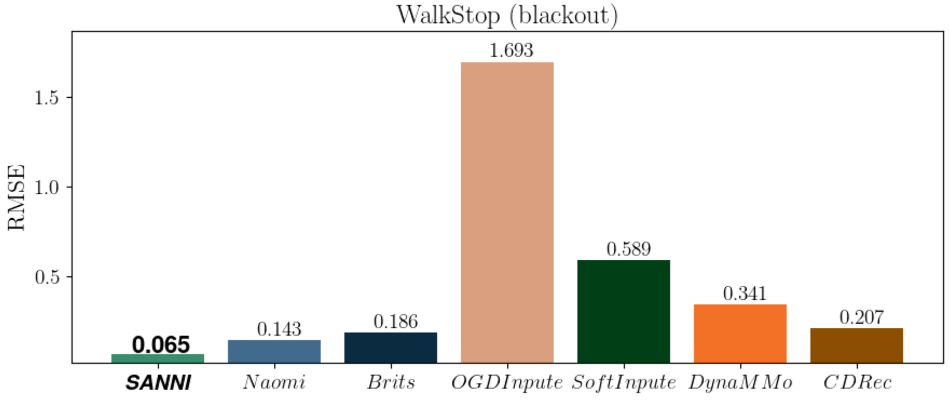
Experiments: Rivals

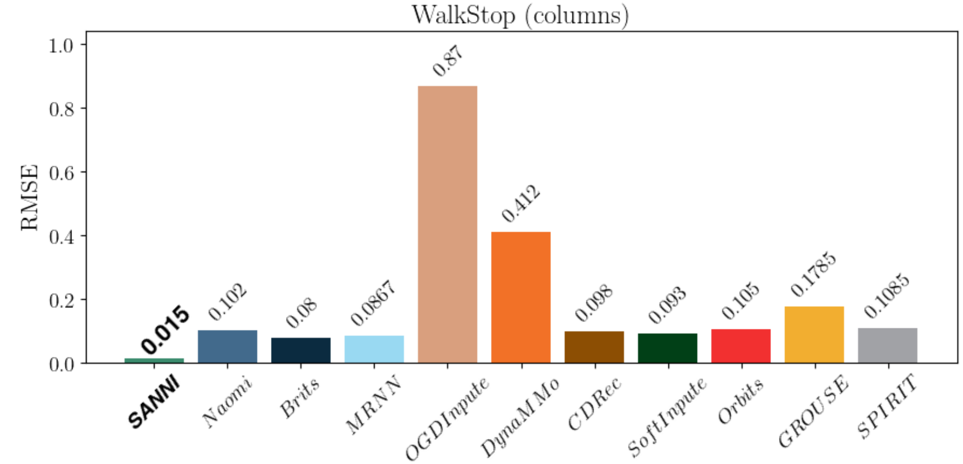
- Analytical approaches
  - CD-REC[1], OGDImpute[2], DynaMMo[3], SoftImpute[4], TKCM[5], GROUSE[6], SPIRIT[7], ORBITS[8]
- ANN-based approaches
  - BRITS[9], NAOMI[10], MRNN[11]

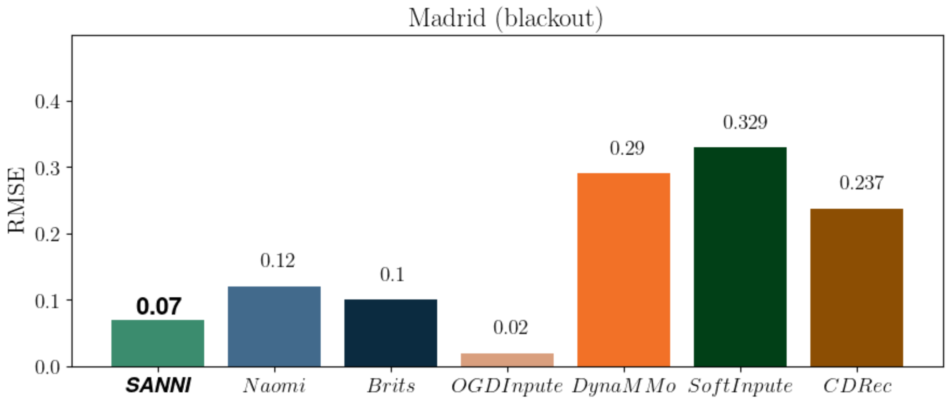
[1] Khayati M., et al. Scalable recovery of missing blocks in time series with high and low cross-correlations. Proc. Computer Science 2019.
[2] Anava O., et al. Online Time Series Prediction with Missing Data. Proc. ICML 2015.
[3] Lei L., et al.DynaMMo: mining and summarization of coevolving sequences with missing values. KDD ’09. 2009.
[4] Mazumder R., et al. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. Journal of Machine Learning Research. 2010.
[5] Wellenzohn K., et al. Continuous Imputation of Missing Values in Streams of Pattern-Determining Time Series. EDBT 2017. P. 330-341.
[6] Zhang D., et al. Global Convergence of a Grassmannian Gradient Descent Algorithm for Subspace Estimation. Electrical Engineering and Computer Science 2017.
[7] Papadimitriou S., et al. Streaming Pattern Discovery in Multiple Time-Series. VLDB 2005.
[8] Khayati M., et al. ORBITS: Online Recovery of Missing Values in Multiple Time Series Streams. Proc. VLDB Endow. 2020
[9] Wei C., et al. ORBITS: Online Recovery of Missing Values in Multiple Time Series Streams. 2018
[10] Liu Y., et al. NAOMI: Non-Autoregressive Multiresolution Sequence Imputation. 2019
[11] Yoon J., et al. Estimating Missing Data in Temporal Data Streams Using Multi-Directional Recurrent Neural Networks. IEEE. 2019

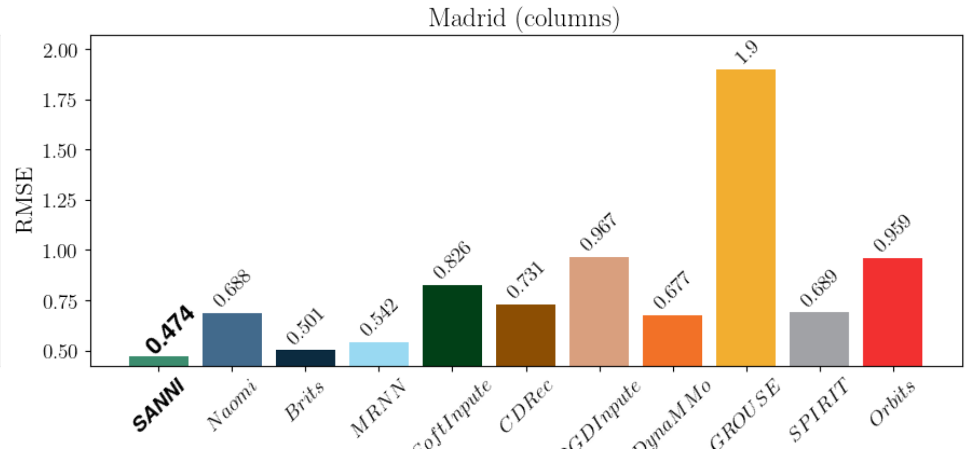
Experiments: Accuracy (activities)

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

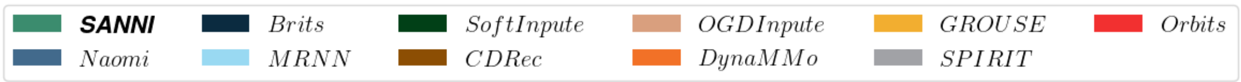






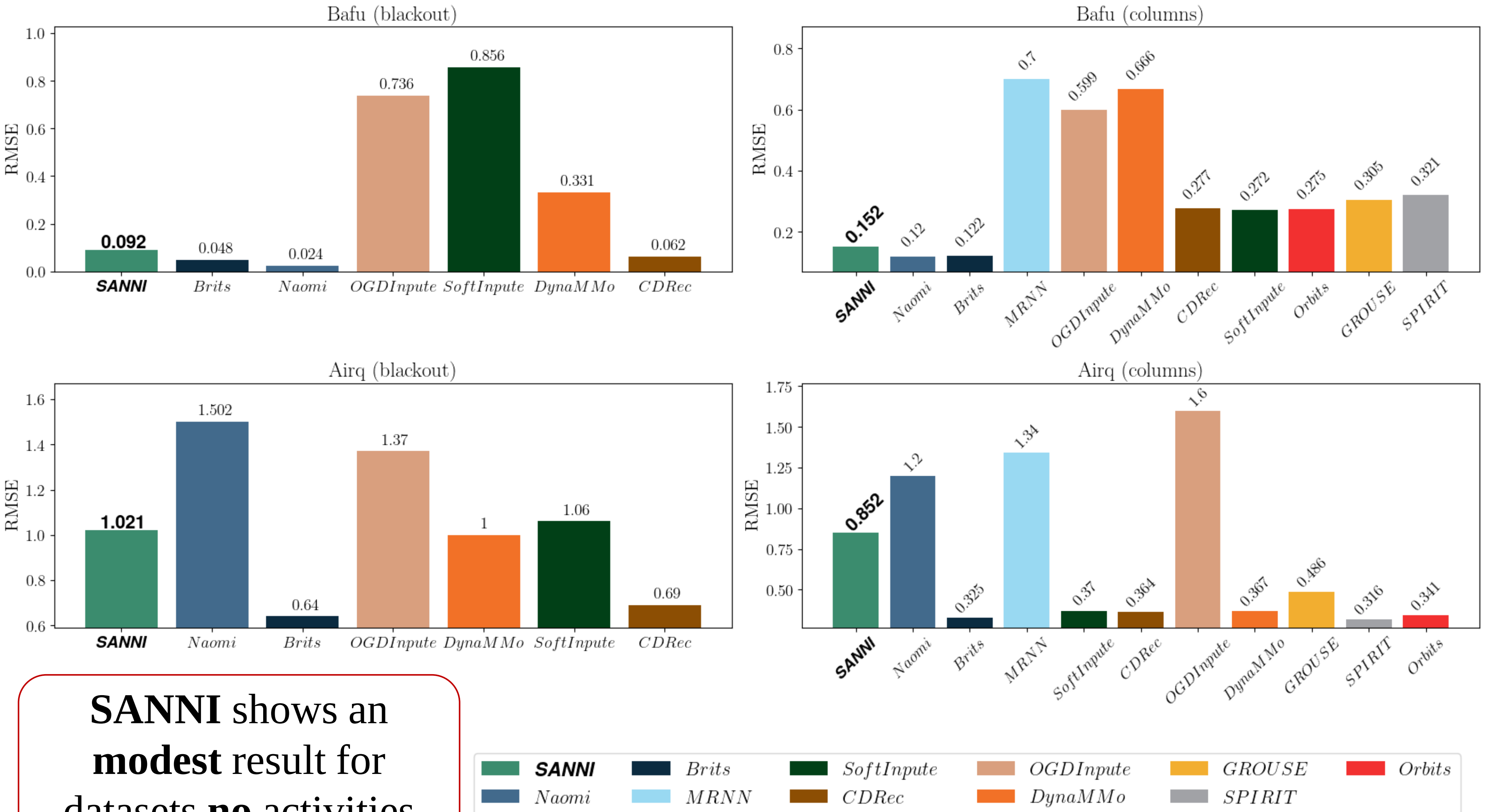


SANNI is **ahead** for datasets **with** activities



Experiments: Accuracy (no activities)

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



SANNI shows an **modest** result for datasets **no** activities

SANNI: *pro et contra*

- Pros
  - ahead of all analogs w.r.t. accuracy for the case when time series reflect activities
  - domain independence
- Cons
  - modest accuracy for the case when time series do not reflect activities

Conclusions and further research

- SANNI is novel snippet and ANN based method for imputation missing values in multivariate streaming time series that is ahead of many analogs w.r.t. accuracy for the case when measurements reflect activities
- Further research:
  - Extensive experimental evaluation of SANNI
  - Augmentation of small coverage snippet through GAN

Thank you for paying attention! Questions?

Alexey Yurtin, iurtinaa@susu.ru