Deep Learning Methods for Unsupervised Time-Series Anomaly Detection

Prof. Narges Armanfard

Thi Kieu Khanh Ho (PhD Student)

Hadi Hojjati (PhD Student)

Department of Electrical and Computer Engineering, McGill University

Mila - Quebec AI Institute, Montreal, QC, Canada

OUTLINE

I. What is Time Series Anomaly Detection (TSAD)?

II. Challenges in TSAD

III. Some TSAD Applications

IV. Deep Learning and TSAD: A Tale of Triumph and Turmoil

V. Deep Learning-based Methods

VI. Conclusion and Discussions

What is Time-Series (TS)?

- A time-series is a collection of observations made over a period of time, with each observation associated with a specific timestamp or time interval.

- There are two types of time series data: (1) univariate, and (2) multivariate time series
  - Univariate time series consists of a single variable recorded at successive time stamps/intervals.
  - Multivariate time series involves multiple variables recorded at each time step/interval.

- Mathematically,

A univariate TS: $\mathbf{x} = (x_1, x_2, \dots, x_N), \quad \mathbf{x} \in \mathbf{R}^{1xN}$

A multivariate TS: $\boldsymbol{X} = \left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(K)}\right), \mathbf{x}^{(i)} = \left(x_1^{(i)}, x_2^{(i)}, \dots, x_N^{(i)}\right), \boldsymbol{X} \in \mathbf{R}^{KxN}$

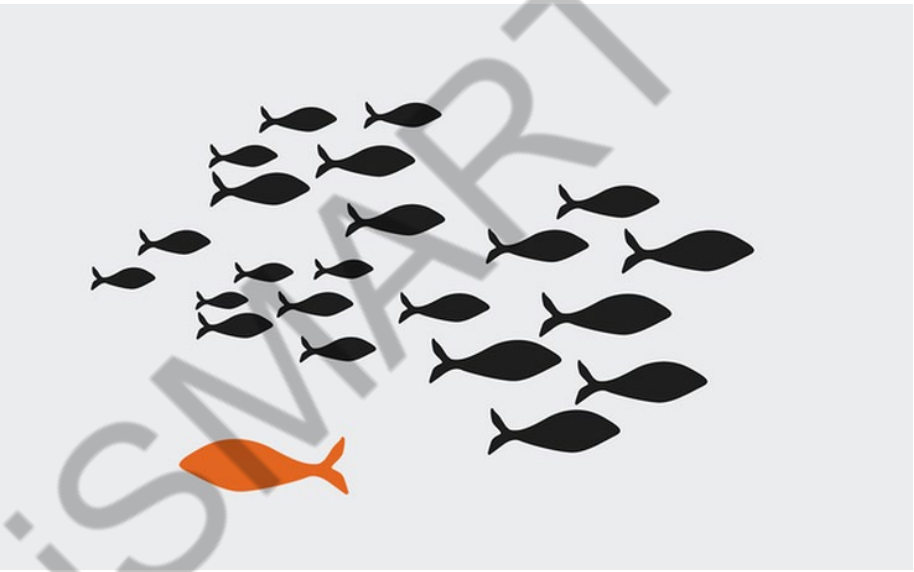
N is the number of time stamps/intervals

K is the number of variables

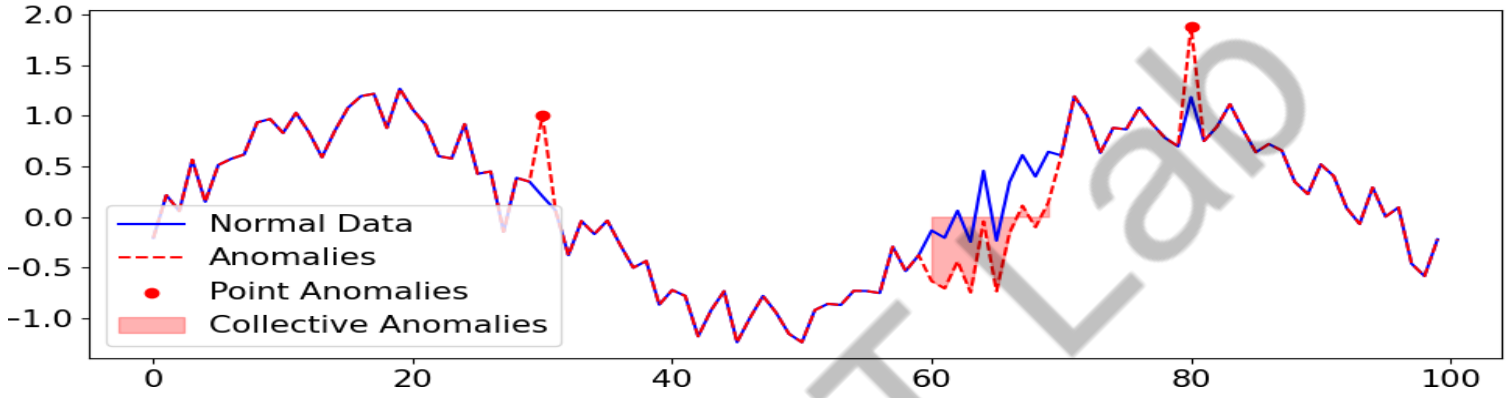
What is Anomaly Detection (AD)?

- Anomaly detection is the process of identifying patterns or instances that significantly deviate from the expected or normal behavior within a dataset.

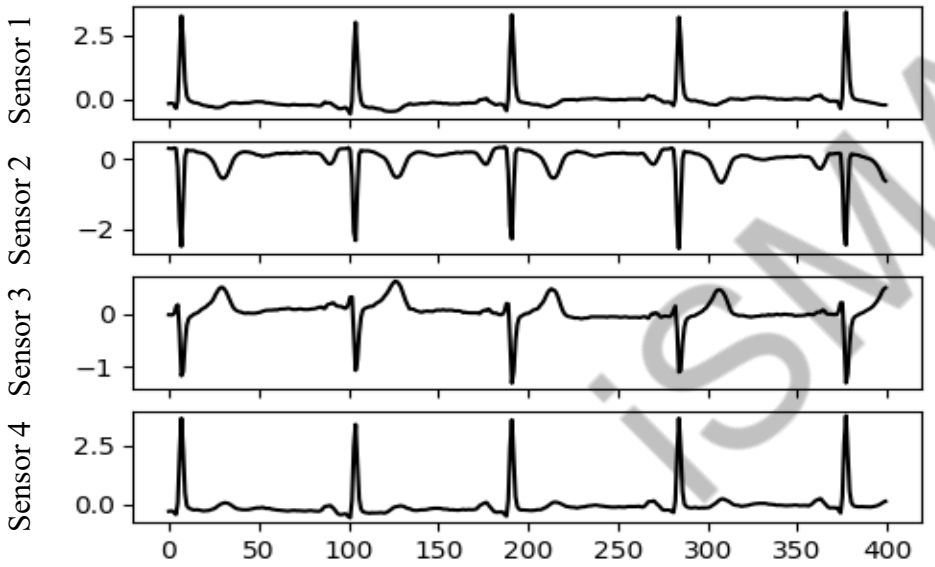
- Time-Series Anomaly Detection (**TSAD**) is the process of detecting unusual patterns that do not conform to the expected behavior within time-series data.



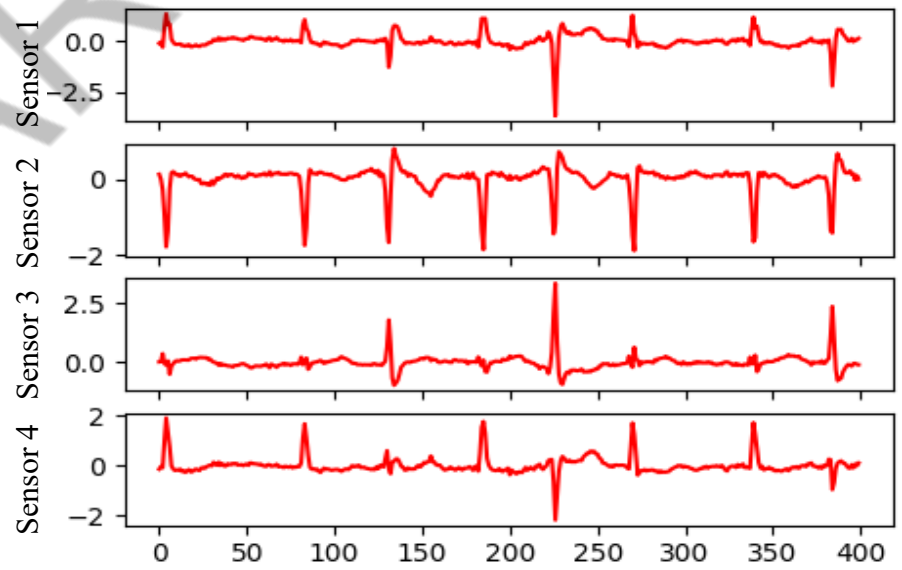
Visualization of TSAD



Univariate TS



Normal Time Interval



Abnormal Time Interval

Multivariate TS

Applications of TSAD

TRANSPORTATION

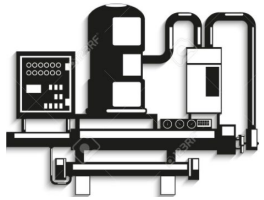
Detect anomalies in vehicles to identify mechanical failures or accidents.

HEALTHCARE

Detect anomalies in physiological signals indicating health issues

TELECOMMUNICATION

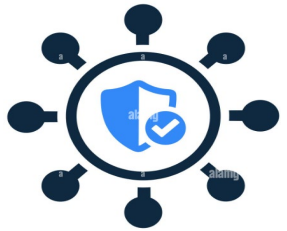
EQUIPMENTS



E-COMMERCE

FINANCE

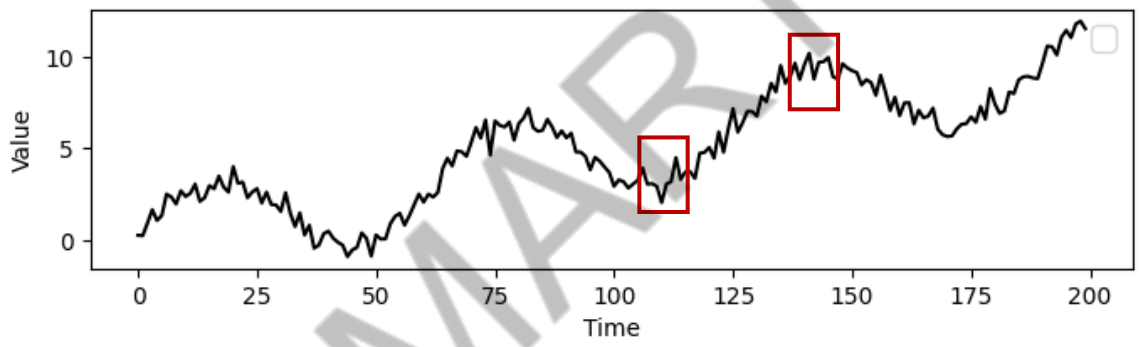
NETWORK SECURITY



SMART CITIES

TSAD Challenges

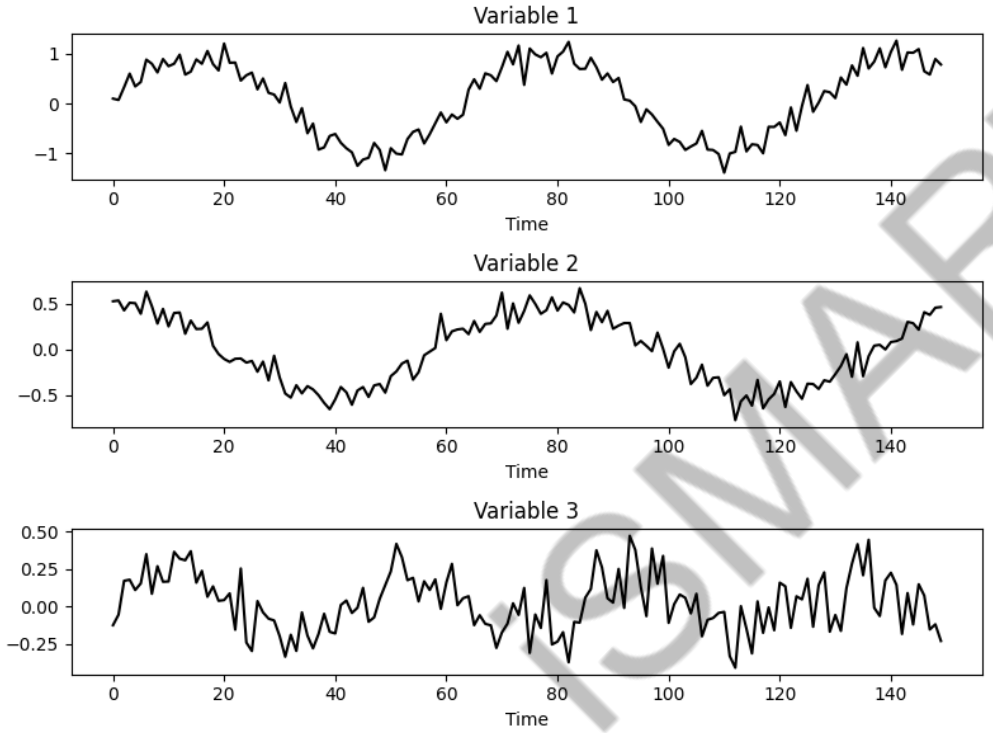
❖ **Intra-variable Dependencies** (aka temporal dependencies) where there exist correlations between previous, current and future values.

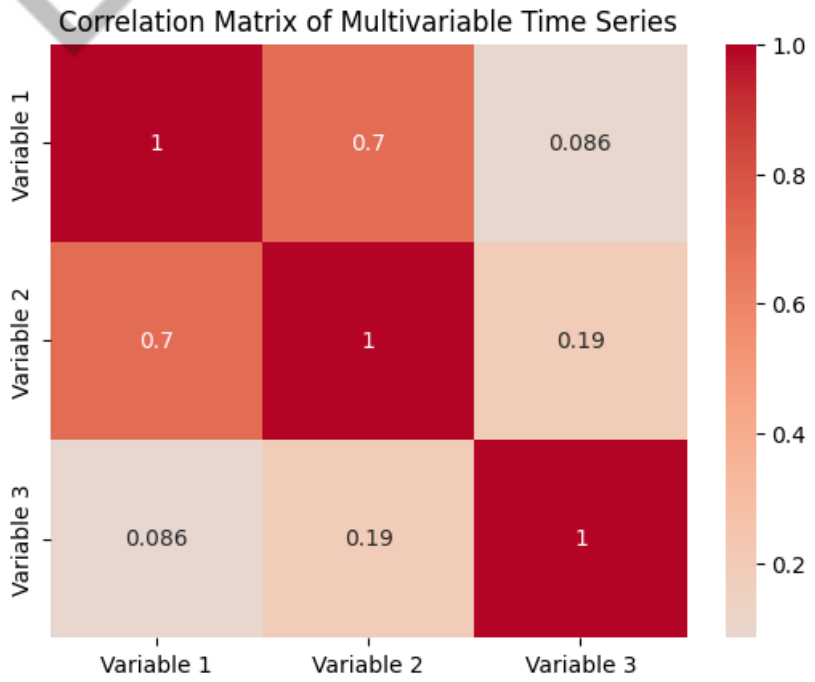


TKK Ho, A Karami, N Armanfard, Graph-based Time-Series Anomaly Detection: A Survey and Outlook.

TSAD Challenges

- **Inter-variable Dependencies** (aka spatial dependencies) where there exist relationships between multiple variables.

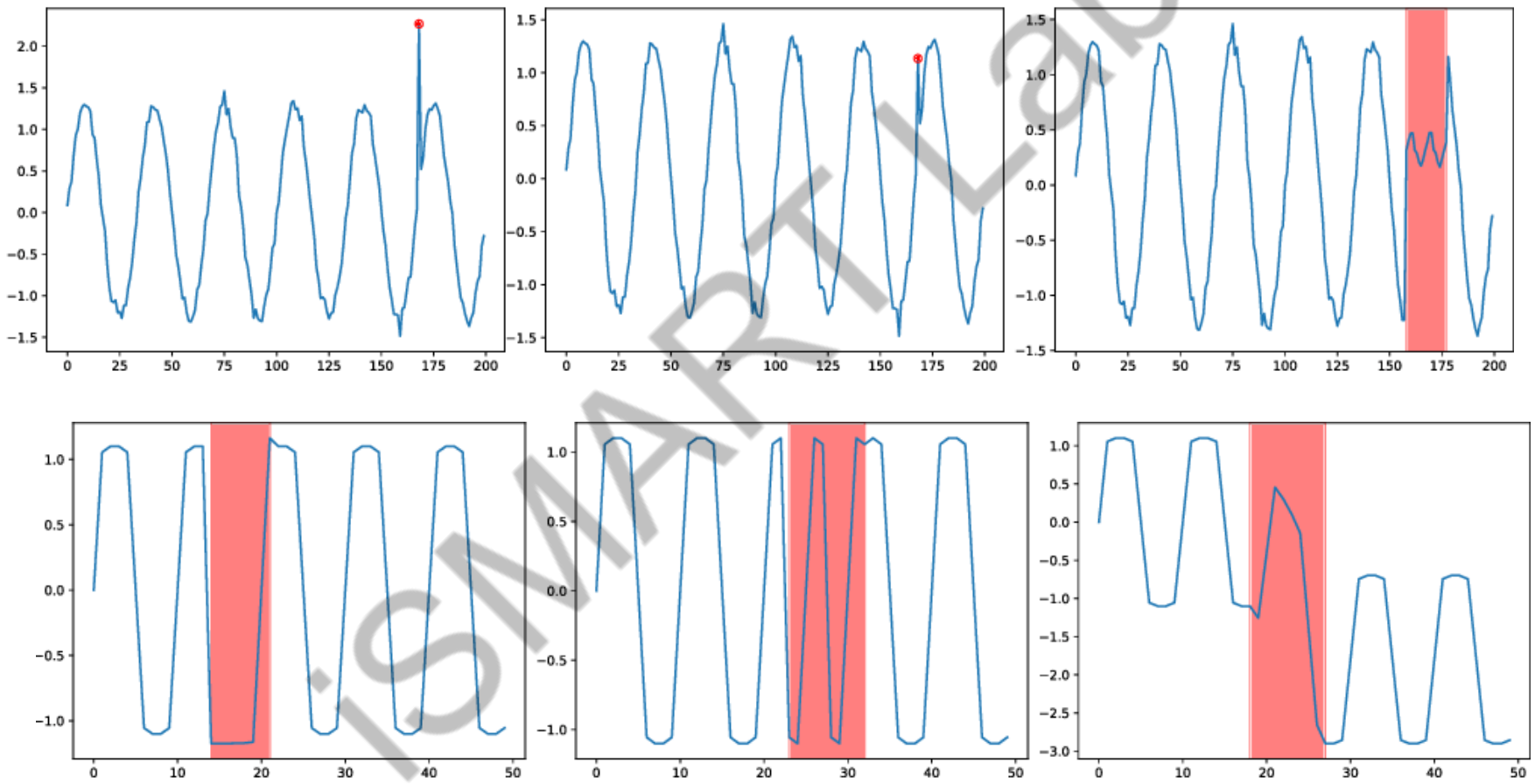




TKK Ho, A Karami, N Armanfard, Graph-based Time-Series Anomaly Detection: A Survey and Outlook.

TSAD Challenges

Types of TS Anomalies



Lai et al., Revisiting Time Series Outlier Detection: Definitions and Benchmarks, NeurIPS 2021

TSAD Challenges

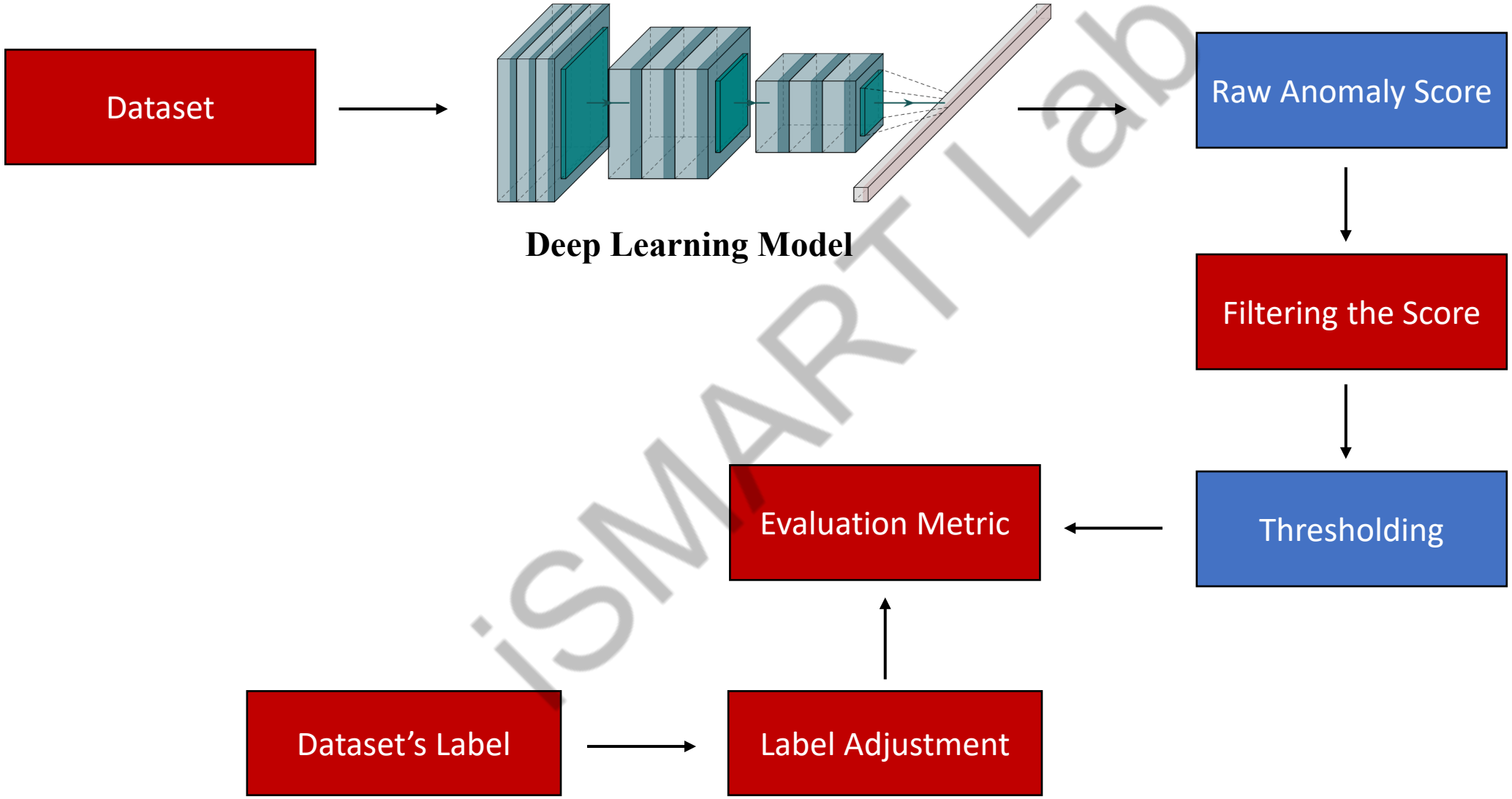
- **Noise**: Noise in TS can arise from various sources such as sensor errors or measurement inaccuracies. This noise can obscure real anomalies.
- **Concept Drift:** TS can change over time due to shifting patterns or evolving environments.
- **Handling missing values:** Missing values can disrupt intra-, inter-variable dependencies.
- **Scalability:** Analyzing large volumes of time-series data can strain computational resources, especially when applying complex algorithms or models.
- **Interpretable Models:** The complexity of the models can make it challenging to interpret why a certain prediction was made, which is essential for many applications.
- **Real-Time Detection:** In applications where real-time detection is crucial, models need to process data quickly and efficiently to identify anomalies without significant delays.
- And many more…

McGill Mila

iSMART Lab

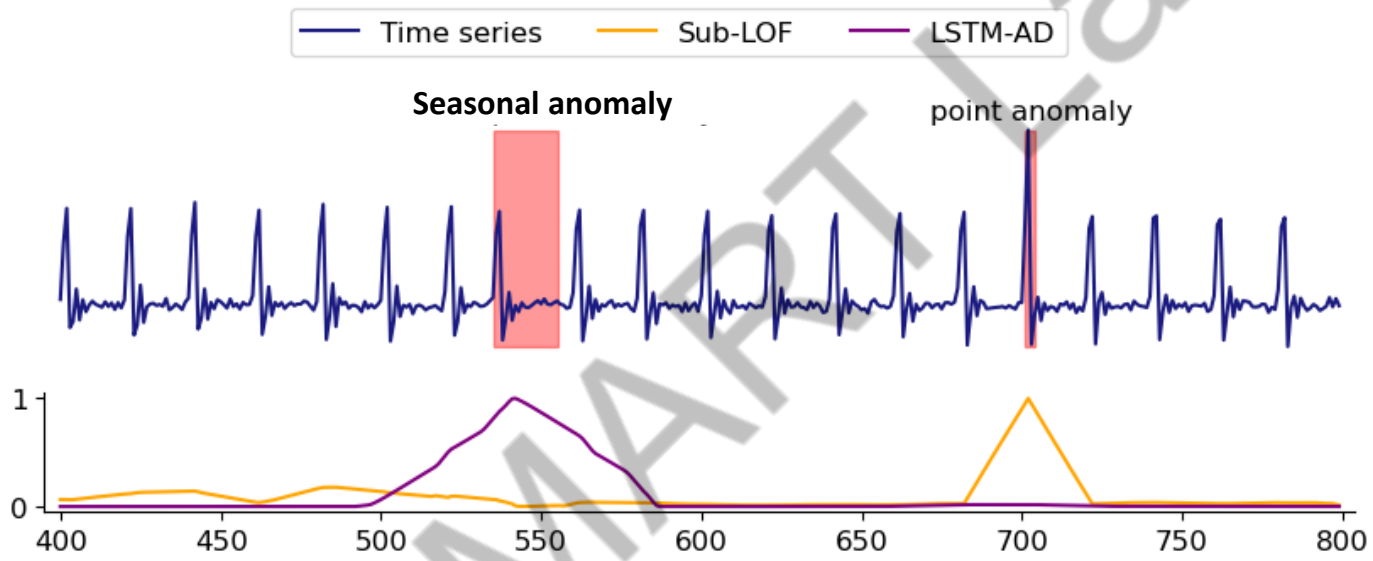
Deep Learning and TSAD: A Tale of Triumph and Turmoil

Overview of TSAD with Deep Learning



Researchers' Playground: Dataset's Anomalies

- **The type of anomaly does matter in algorithm performance!**



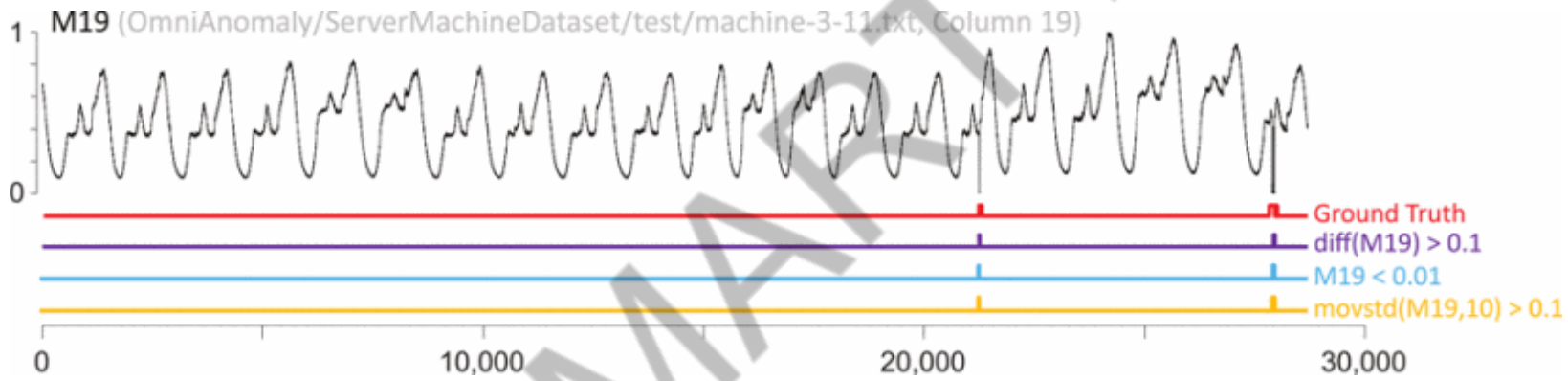
- Sub-LOF: Simple algorithm/ detects the point anomaly very well
- LSTM-AD: Complex algorithm/ detects the seasonal anomaly

→ A wide range of anomalies should be used for evaluation!

Schmidl et al., Anomaly Detection in Time Series: A Comprehensive Evaluation, VLDB 2022

Researchers' Playground: Easy Dataset

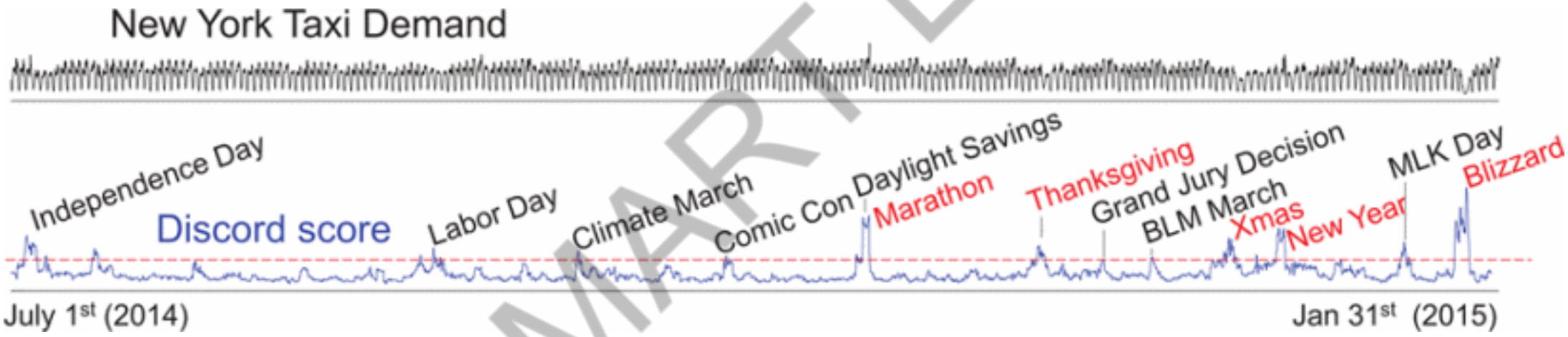
- **Too easy to solve** $\Longrightarrow$ A one line code can solve it!



Wu and Keogh, Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress, ICDE 2022

Researchers' Playground: Dataset's Label

- **Subjective Labelling**

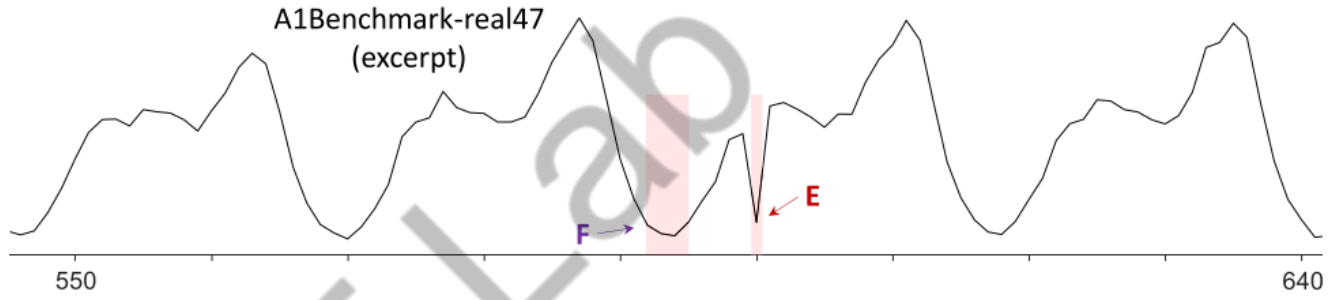


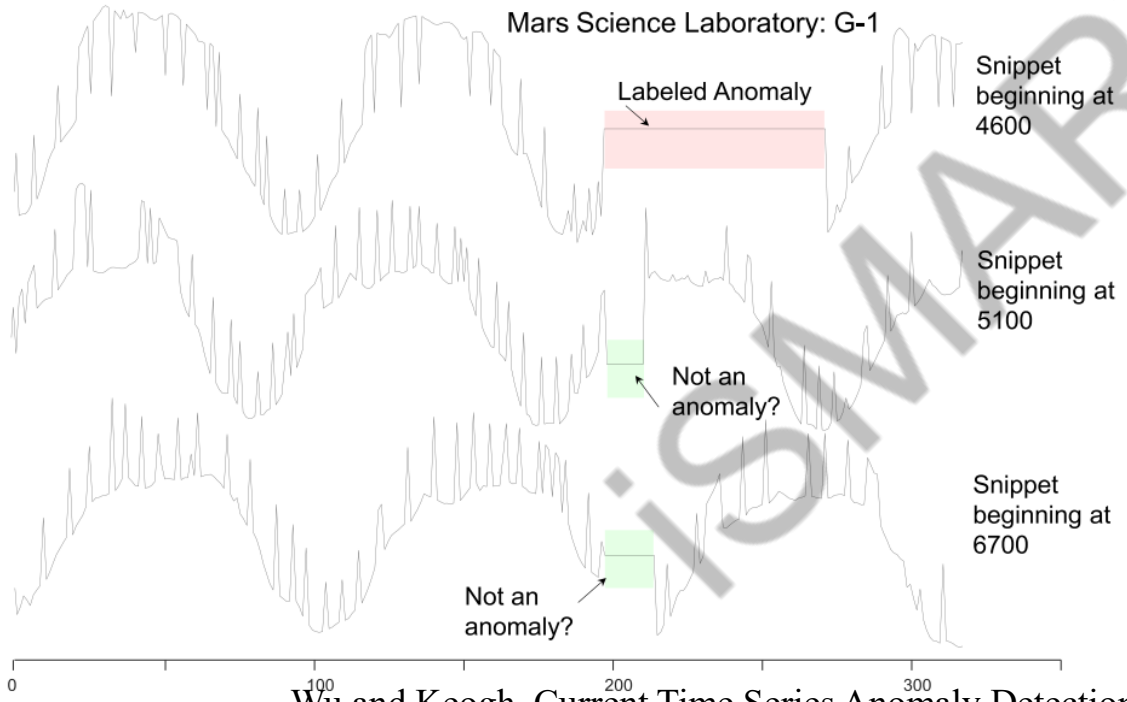
- **In many instances, it is impossible to define the true label for each data point!**

Wu and Keogh, Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress, ICDE 2022

Researchers' Playground: Dataset's Label

- **Subjective Labelling**





Wu and Keogh, Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress, TKDE 2022

Researchers' Playground: Bias in Dataset

- **Run to Failure Bias**

  Most anomalies appear at the end of the time series!

  Many real-world systems are run-to-failure ➔ there is no data to the right of the last anomaly!

A naïve algorithm that simply labels the last point as an anomaly has an excellent chance of being correct!!!

Wu and Keogh, Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress, ICDE 2022

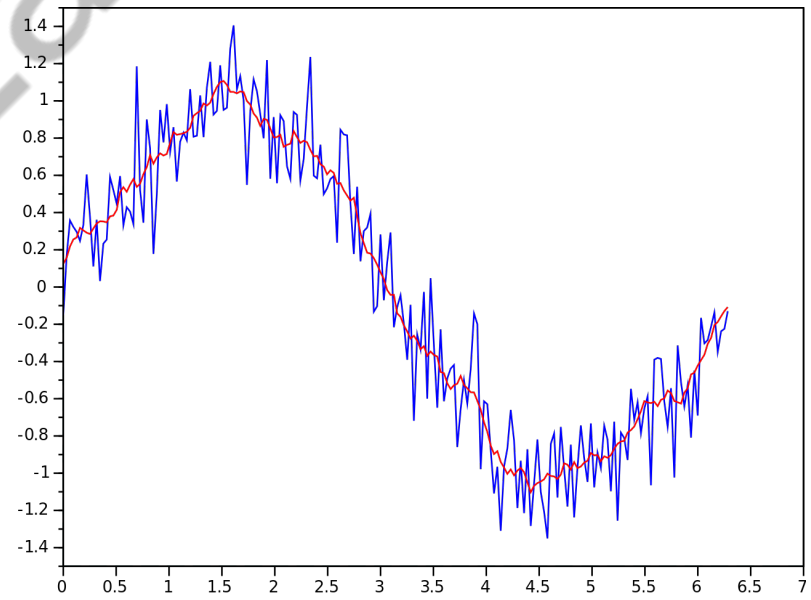
Researchers' Playground: Dataset and its label

Inappropriate Benchmark Datasets

Many published comparisons of anomaly detection algorithms may be unreliable!

Researchers' Playground: Filtering the Anomaly Score

- Raw output of NN is noisy and does not consider the temporal dependencies
- We can use a kernel, aka scoring function, to filter and smoothen the anomaly score
- Many researchers incorporate this step as one module of their algorithm



Researchers' Playground: Filtering the Anomaly Score

- **Gauss-S**

$$\mathbf{A}_t^i = -\log\left(1 - \Phi\left(\frac{\mathbf{Er}_t^i - \hat{\mu}^i}{\hat{\sigma}^i}\right)\right); \quad \mathbf{a}_t = \sum_{i=1}^{m} \mathbf{A}_t^i.$$

- **Gauss-D**

$$\hat{\mu}_t^i = \frac{1}{W}\sum_{j=0}^{W-1} \mathbf{Er}_{t-j}^i; \quad \left(\hat{\sigma}_t^i\right)^2 = \frac{1}{W-1}\sum_{j=0}^{W-1}\left(\mathbf{Er}_{t-j}^i - \hat{\mu}_t^i\right)^2$$

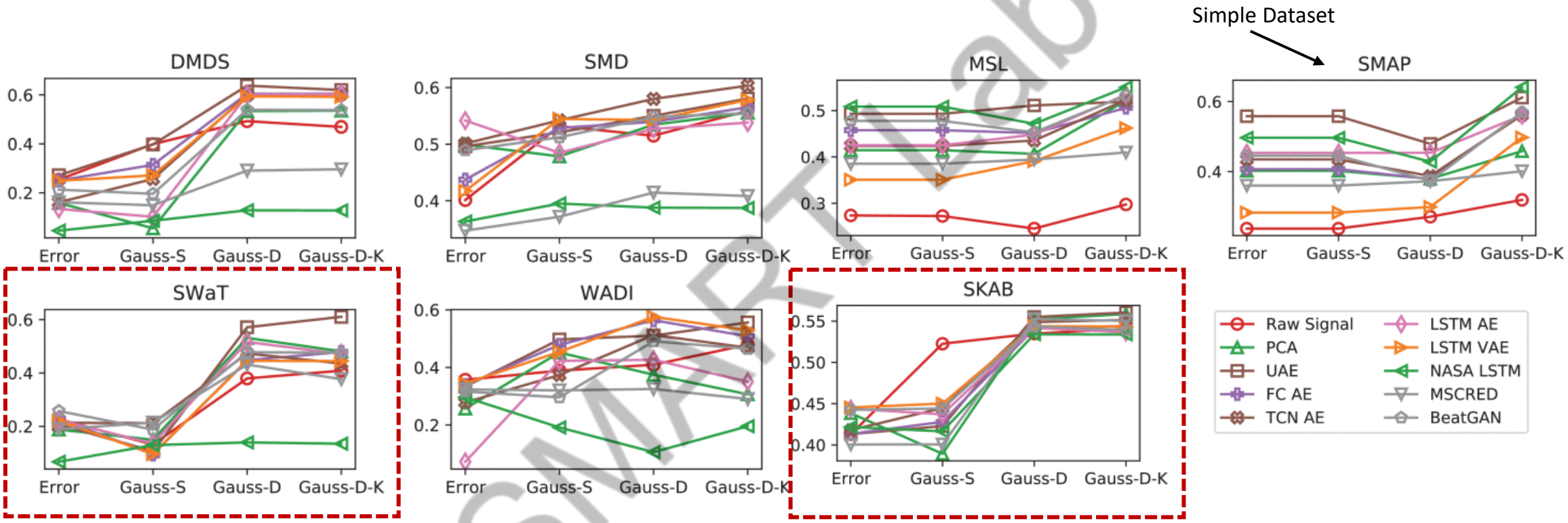
- **Gauss-D-K**

$$G\left(u;\sigma_k\right) = e^{-\frac{1}{2}\left(\frac{u}{\sigma_k}\right)^2}; \quad \mathbf{A}_t^i = G * \mathbf{A}_{t,\text{Gauss-}D}^i$$

Kim et al., Towards a Rigorous Evaluation of Time-Series Anomaly Detection, AAAI 2022

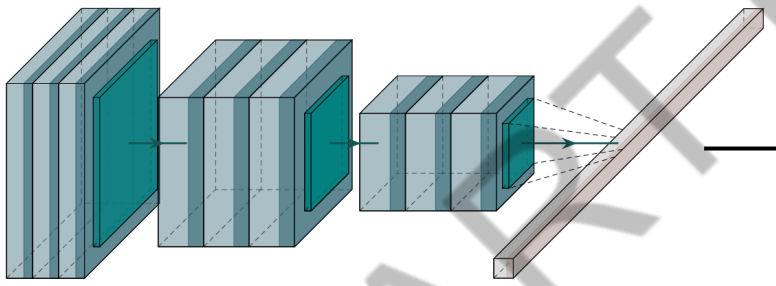
Researchers' Playground: Filtering the Anomaly Score

- Common TSAD Benchmark datasets: SWAT/WADI/MSL/SMAP/SKAP/SMD/DMDS



Garg et al., An Evaluation of Anomaly Detection and Diagnosis in Multivariate Time Series, TNNLS 2022

Labeling is easy in CV but not in TS!



Normal (Dog): 99.7%
Abnormal (Cat): 0.3%

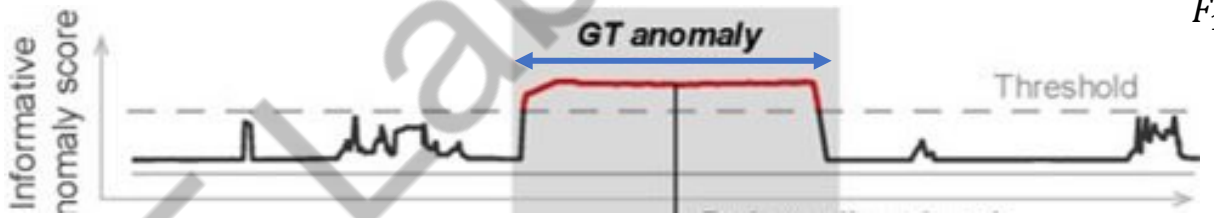
Good Classifier

Researchers' Playground: Label Adjustment/evaluation

- The anomaly score is defined for each time-stamp
- F1 can be used to evaluate the performance

$$P = \frac{TP}{TP + FP} \qquad R = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2P.R}{P + R}$$



$F_1 \approx 1$

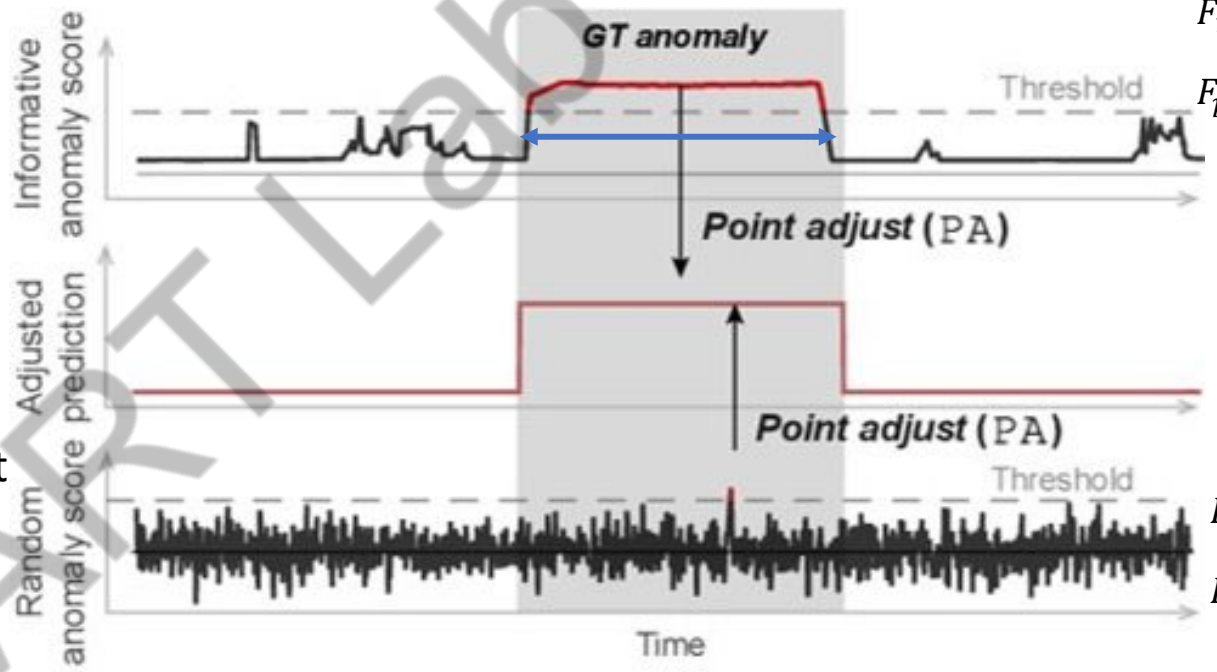
Kim et al., Towards a Rigorous Evaluation of Time-Series Anomaly Detection, AAAI 2022

Researchers' Playground: Label Adjustment/evaluation

- The anomaly score is defined for each time-stamp
- F1 can be used to evaluate the performance

$$F_1 = \frac{2P.R}{P+R}$$

Fpa Issue:



$F_1 \approx 1$

$F_{pa} = 1$

$F_1 = 0.01$

$F_{pa} = 1$

- PA protocol (peculiar!):

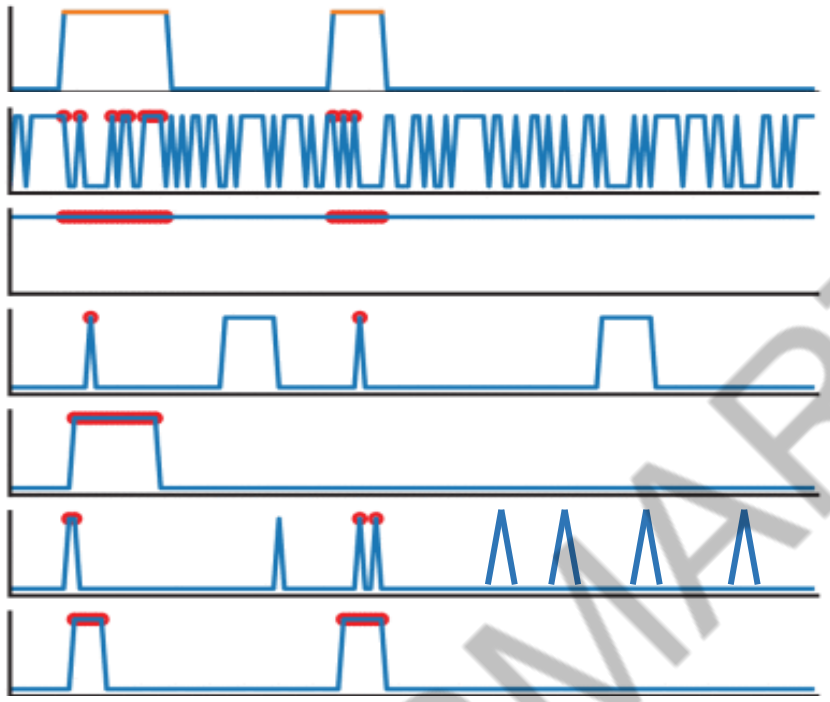
  - A single alert within an anomaly period is sufficient to take action for system recovery!
  - Re-label: assign labels at the segment level!
  - Measure F1 score based on the relabeled data!

- **Point Adjustment:** Even if the model detects one anomaly in the segment, label the rest of the segment as an anomaly! ⟶ F_{pa}

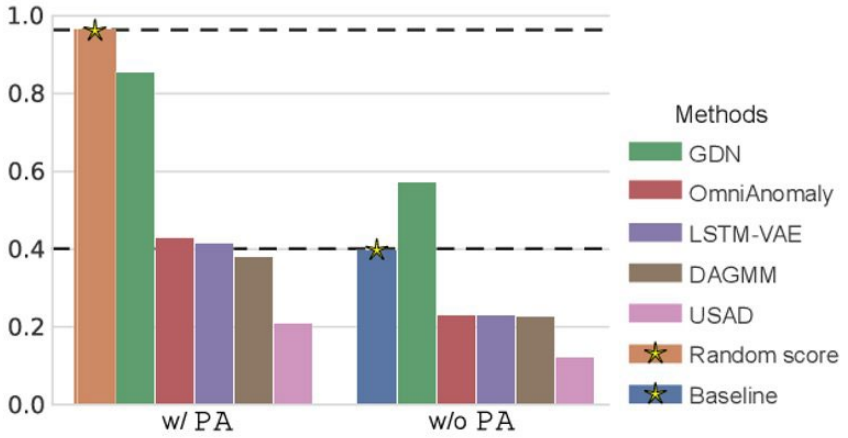
Kim et al., Towards a Rigorous Evaluation of Time-Series Anomaly Detection, AAAI 2022

Researchers' Playground: Label Adjustment/evaluation

Ground truth



F1	Fpa
0.21	0.44
0.33	0.32
0.08	0.74
0.70	0.81
0.23	**0.91**
0.64	1.00



Garg et al., An Evaluation of Anomaly Detection and Diagnosis in Multivariate Time Series, TNNLS 2022

Is Deep Learning the best solution for TSAD?

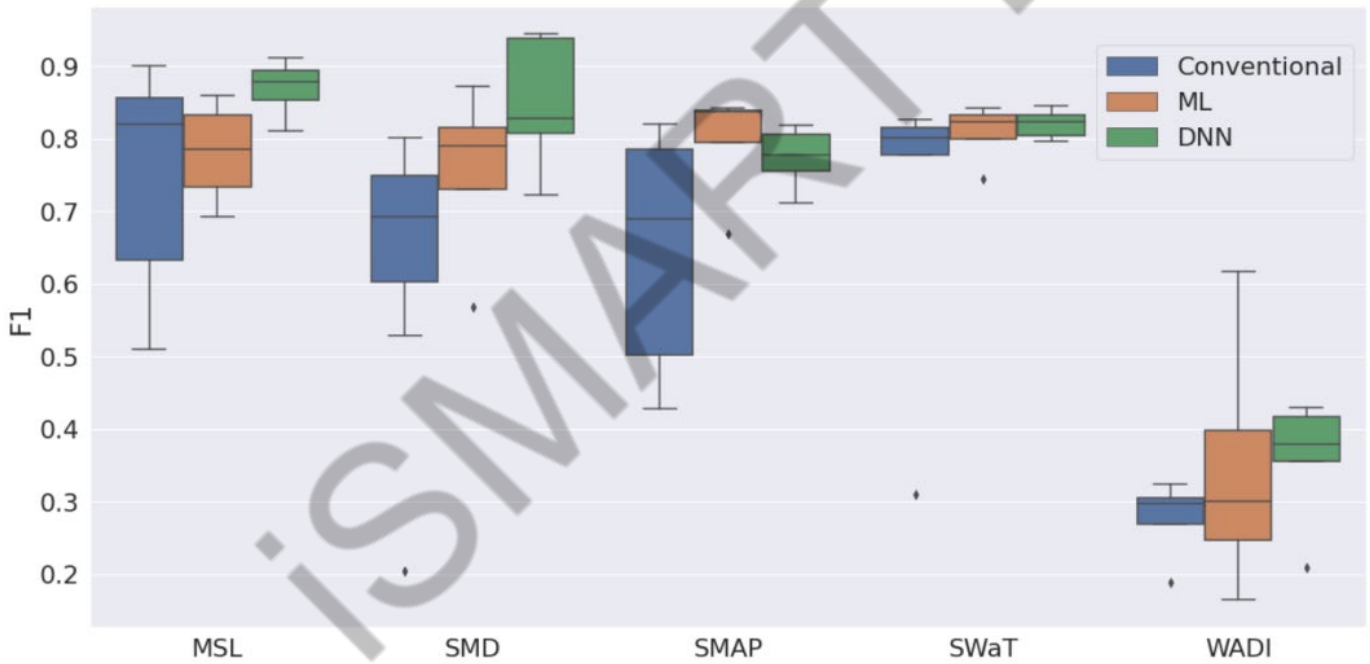
Challenges of Deep Learning in TSAD

- Data Requirements: A large amount of data for training is required. Can overfit, especially when the dataset is small.
- Computational Resources: Deep learning models are complex, with dozens of parameters.
- Fine-Tuning and Hyperparameter Tuning: DL requires extensive experiments to fine-tune the pre-trained models and fine-tune hyperparameters.
- Interpretability: DL models are complex, making them hard to interpret.
- And many more…

Is Deep Learning the best solution for TSAD?

- Compare 16 conventional, non-deep (ML) and, deep neural network approaches on 5 real-world open simple datasets (SWaT, WADI, SMD, SMAP, MSL).

→ No evidence of methods outperforms the others!



Audibert et al. *"Do Deep Neural Networks Contribute to Multivariate Time Series Anomaly Detection?"*, Pattern Recognition, 2022

The way forward!

- A unique opportunity for researchers!
- Deep Learning is still an attractive option for TSAD
- Purely numerical results are not the goal, we need more interpretability *wrt* data, application, algorithm
- Learning from mistakes:
  - We need new benchmark datasets!
  - Comparing the performance with simple, trivial, and even random baselines is of high value.
  - Filtering function should be uniform across all baselines.
  - Evaluation protocol should be the same in all baselines.
  - No free lunch! Each application requires its own method design.

McGill Mila

iSMART Lab

Deep Learning for TSAD: Ideas and Methods

Why Deep Learning (DL) for TSAD?

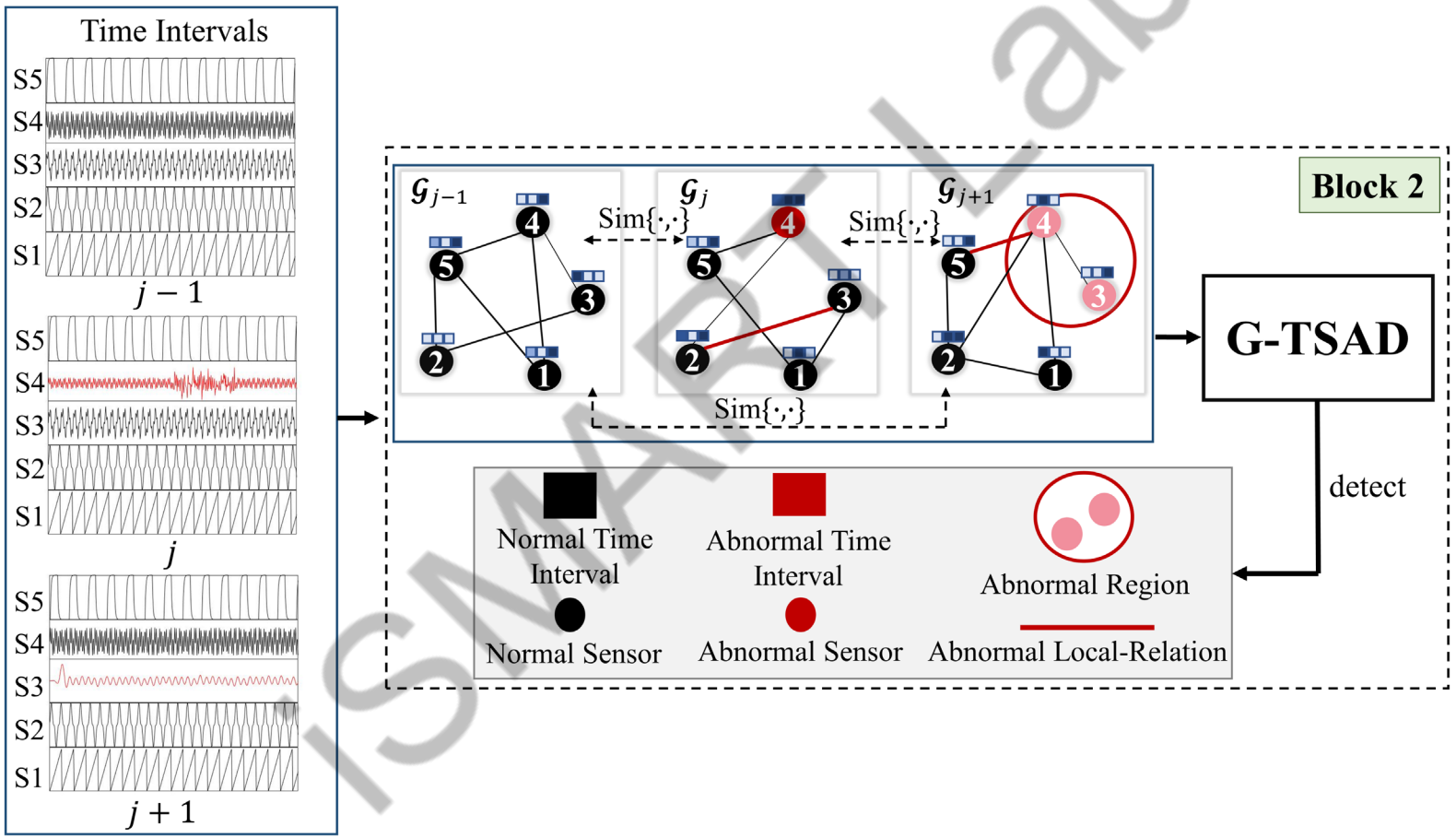
- Labeled anomalies are not required in the training phase.
- Capturing Complex Patterns: DL excels at capturing intra- and inter-variable dependencies.
- Feature Learning: DL automatically learns relevant features from raw data.
- Handling High-Dimensional Data: DL is effective for high-dimensional time series data without manual feature engineering.

Why Deep Learning (DL) for TSAD?

- Robustness to Noise: DL can learn robust representations that are less sensitive to noise.
- Transfer Learning: Pre-trained deep models can be fine-tuned for time series anomaly detection.
- Adaptation to New Data: DL can adapt to new data distributions and changing patterns.
- Anomaly Types and Diversity: DL can handle diverse types of anomalies, from subtle to complex.
- And many more…

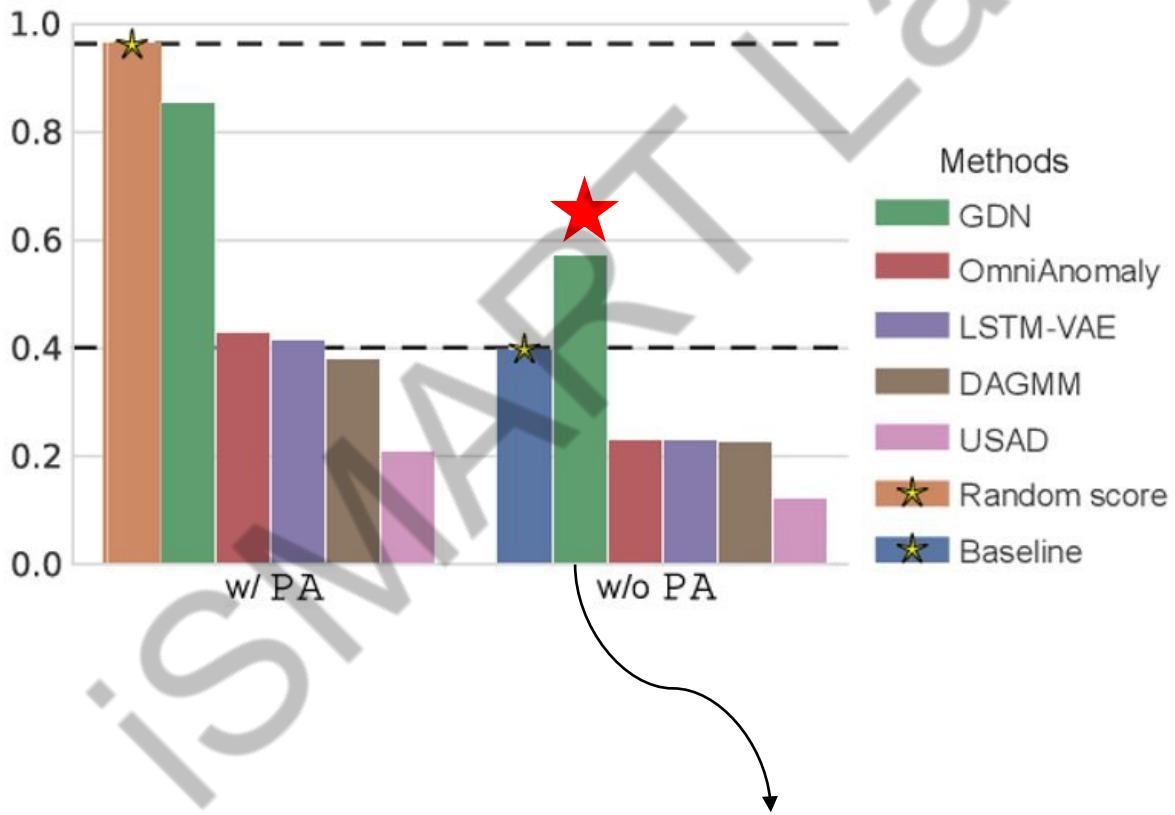
Graph-Based TSAD

Why Graphs?



TKK Ho, A Karami, N Armanfard, Graph-based Time-Series Anomaly Detection: A Survey and Outlook.

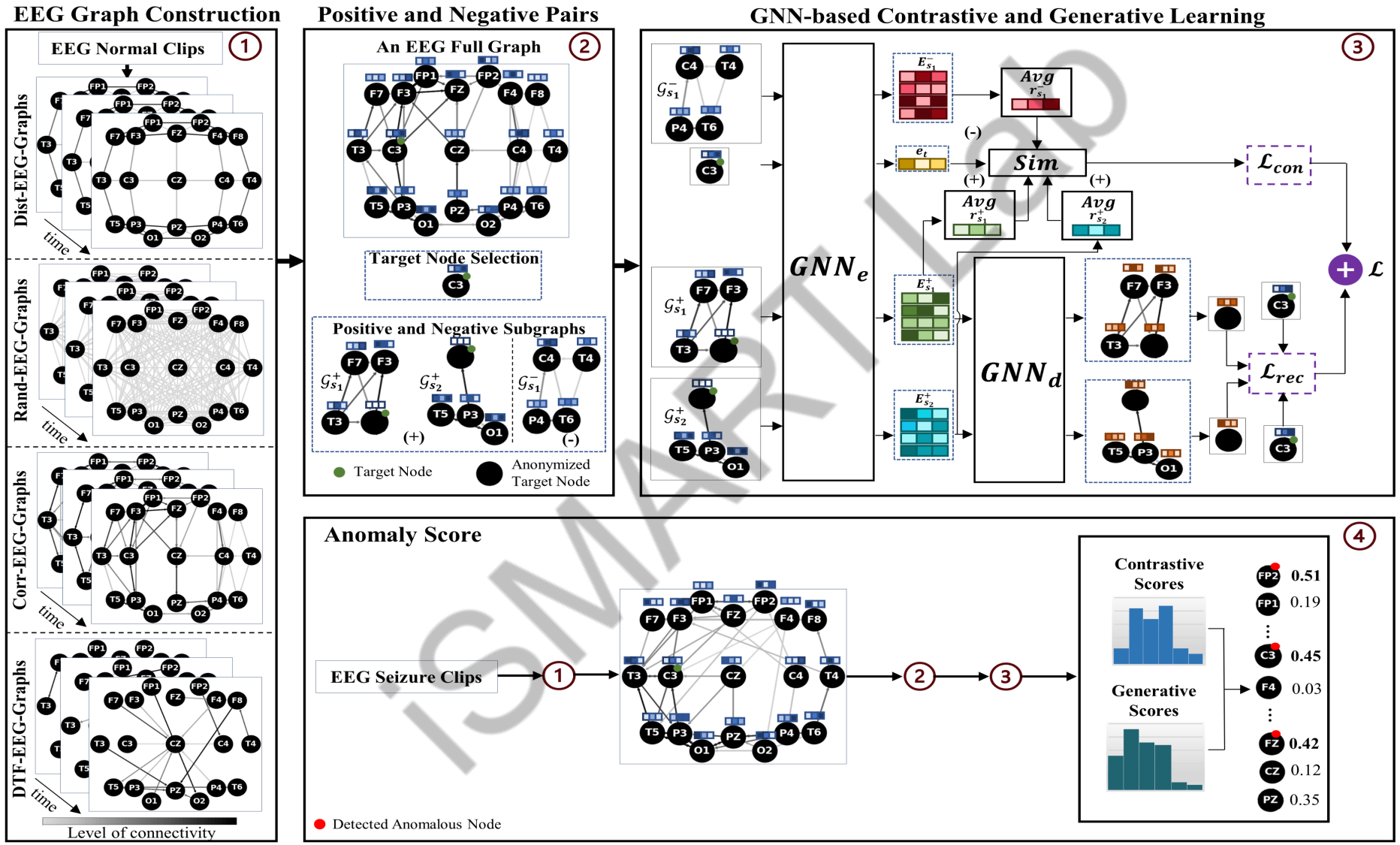
Why Graph-Based Methods?



Deng and Hoi, Graph Neural Network-Based Anomaly Detection in Multivariate Time Series, AAAI 2021

Kim et al., Towards a Rigorous Evaluation of Time-Series Anomaly Detection, AAAI 2022

TSAD in EEG Data with Graphs



TKK Ho, N Armanfard, Self-supervised Learning for Anomalous Channel Detection in EEG Graphs: Application to Seizure Analysis, AAAI-23 (Oral)

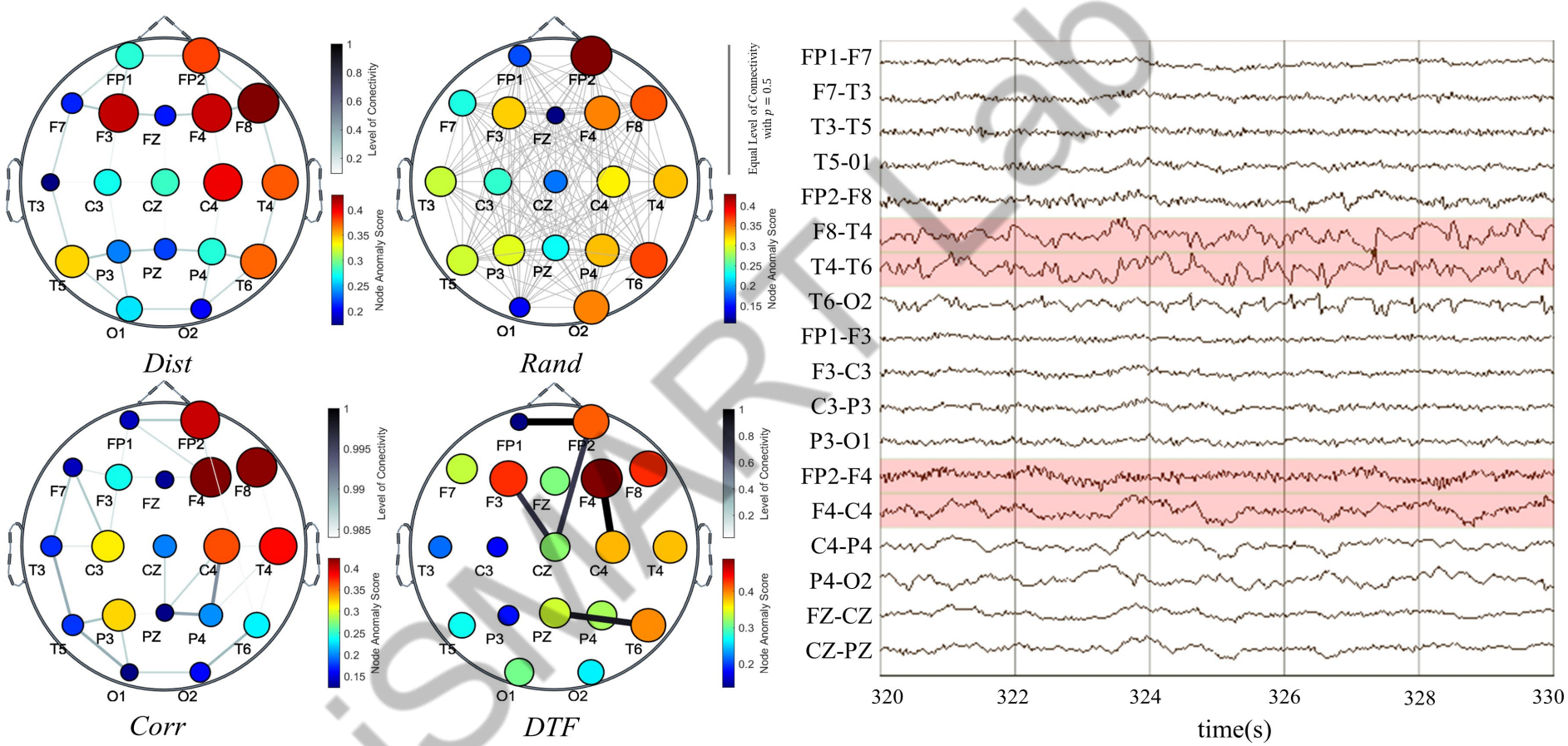
TSAD in EEG Data with Graphs

Approach	Method	F1	Sen	Spec
Supervised	EEGNet	0.474	0.299	0.902
	EEG-TL	0.420	NA	NA
	Dense-CNN	0.404	0.451	0.869
	LSTM	0.365	0.463	0.814
	CNN-LSTM	0.330	0.363	0.857
	Dist-DCRNN	0.341	0.326	0.932
	Corr-DCRNN	0.448	0.457	0.900
Self-Supervised (ours)	EEG_d-CGS	0.487	0.481	0.932
	EEG_r-CGS	0.496	0.465	0.942
	EEG_c-CGS	$0.521^{\dagger}$	$0.497^{\dagger}$	0.942
	EEG_f-CGS	0.516	0.474	$0.952^{\dagger}$
	EEG_t-CGS	**0.534**	**0.501**	**0.974**

Seizure clips detection. The best and the second-best metrics are denoted in bold and †, respectively.

TKK Ho, N Armanfard, Self-supervised Learning for Anomalous Channel Detection in EEG Graphs: Application to Seizure Analysis, AAAI-23 (Oral)

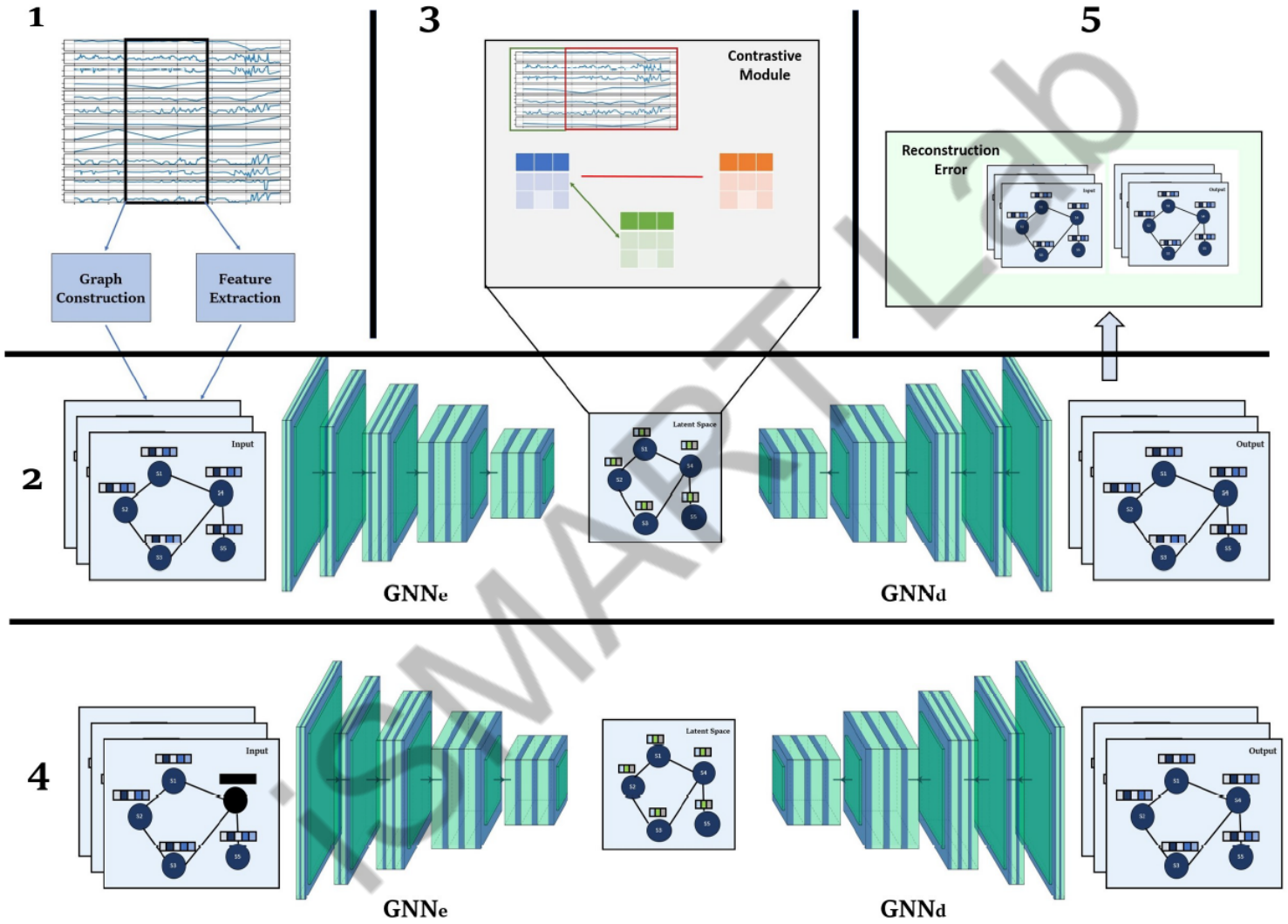
TSAD in EEG Data with Graphs



Visualization of abnormal focal-seizure region detection. Ground-truth channels are shaded in pink.

TKK Ho, N Armanfard, Self-supervised Learning for Anomalous Channel Detection in EEG Graphs: Application to Seizure Analysis, AAAI-23 (Oral)

Graph-Based Methods for Vehicle Data



H. Hojjati, M. Sadeghi, N Armanfard, Multivariate Time-Series Anomaly Detection with Temporal Self-Supervision and Graphs: Application to Vehicle Failure Prediction, ECML PKDD 2023

Graph-Based Methods for Vehicle Data

Algorithm	Metric	V_1	V_2	V_3	V_4	V_5	*Average*
OCSVM	Prec	0.56	0.48	0.49	0.63	0.46	0.52
	Recall	0.81	0.74	0.71	0.83	0.77	0.77
	F1	0.66	0.58	0.58	0.71	0.57	0.62
AE	Prec	0.58	0.51	0.53	0.67	0.58	0.57
	Recall	0.88	0.83	0.85	0.91	0.81	0.85
	F1	0.69	0.83*	0.65	0.71	0.676	0.68
LSTM-AE	Prec	0.63	0.59	0.63	0.67	0.62	0.62
	Recall	0.93	0.88	0.86	**0.97**	0.91	0.91
	F1	0.75	0.70	0.72	0.79	0.73	0.74
TCN-AE	Prec	0.61	0.52	0.57	0.63	0.59	0.58
	Recall	**0.96**	**0.93**	0.91	0.95	0.93*	0.93
	F1	0.74	0.66	0.70	0.75	0.72	0.71
USAD	Prec	0.68	0.57	0.64*	0.70	0.59	0.63
	Recall	0.92	0.90	0.94*	0.95	0.88	0.91
	F1	0.78	0.69	0.76*	0.80	0.706	0.75
GDN	Prec	0.72*	0.63*	0.61	0.72*	0.68*	0.67*
	Recall	0.94*	0.92*	0.92	0.98	0.93*	0.93*
	F1	0.81	0.74	0.73	0.83*	0.78*	0.78*
mVSG-VFP	Prec	**0.85**	**0.74**	**0.71**	**0.84**	**0.78**	**0.78**
	Recall	0.94*	0.92*	**0.95**	0.96*	**0.96**	**0.94**
	F1	**0.89**	**0.82**	**0.81**	**0.89**	**0.86**	**0.85**

H. Hojjati, M. Sadeghi, N Armanfard, Multivariate Time-Series Anomaly Detection with Temporal Self-Supervision and Graphs: Application to Vehicle Failure Prediction, ECML PKDD 2023

Conclusion

- AI and Deep Learning have achieved unprecedented success in NLP and CV
- Applying DL to time-series requires careful considerations
- There are a flurry of novel ideas but their real performance is still overshadowed
- We hope that this tutorial can spark new ideas in the literature

Thank you for your attention!

https://ismart.ece.mcgill.ca/