

# The Elephant in the Room:

Towards A Reliable Time-series Anomaly  
Detection Benchmark

---

*Qinghua Liu*

*John Paparrizos*

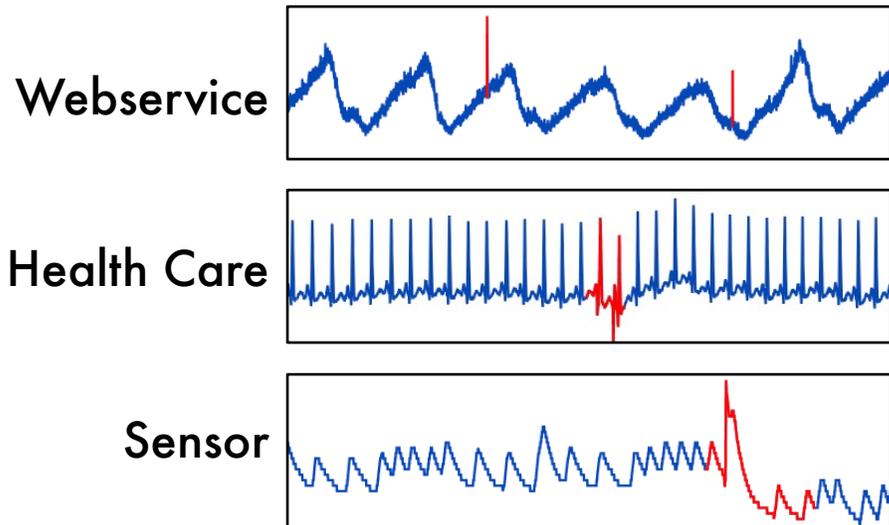


THE OHIO STATE  
UNIVERSITY



# Background and Overview

Rising Demand for Time-Series Anomaly Detection

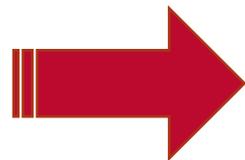


...

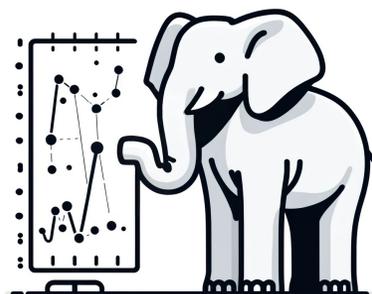
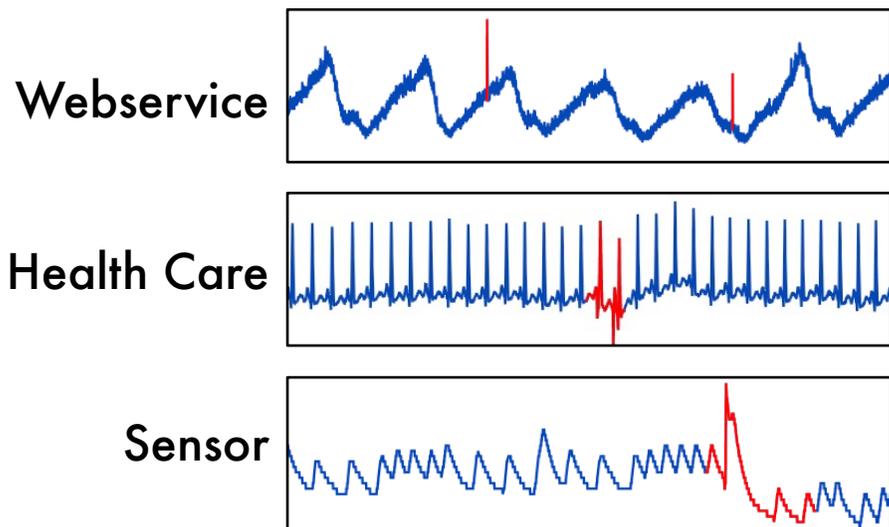
[1] Qinghua Liu, Paul Boniol, Themis Palpanas, and John Paparrizos.  
Time-Series Anomaly Detection: Overview and New Trends. VLDB 2024.

# Background and Overview

Rising Demand for Time-Series Anomaly Detection



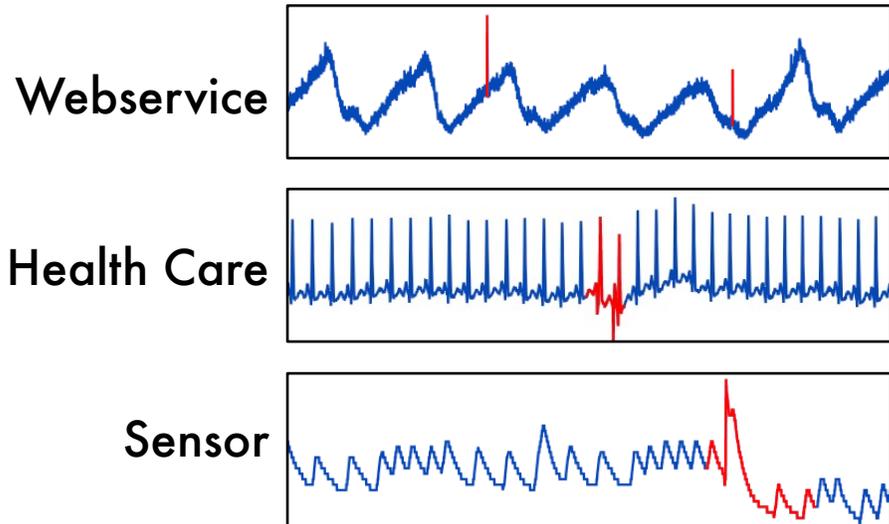
The Elephant in the Room



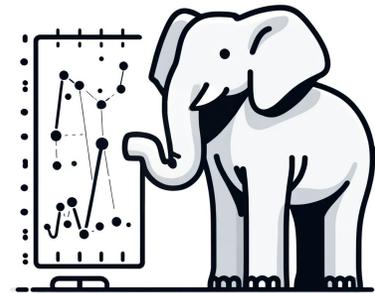
[1] Qinghua Liu, Paul Boniol, Themis Palpanas, and John Paparrizos.  
Time-Series Anomaly Detection: Overview and New Trends. VLDB 2024.

# Background and Overview

Rising Demand for Time-Series Anomaly Detection



The Elephant in the Room



Flaws in Dataset

Problematic Evaluation Measures

Inconsistent Benchmarking Practice

[1] Qinghua Liu, Paul Boniol, Themis Palpanas, and John Paparrizos.  
Time-Series Anomaly Detection: Overview and New Trends. VLDB 2024.



# **DATASET INTEGRITY**

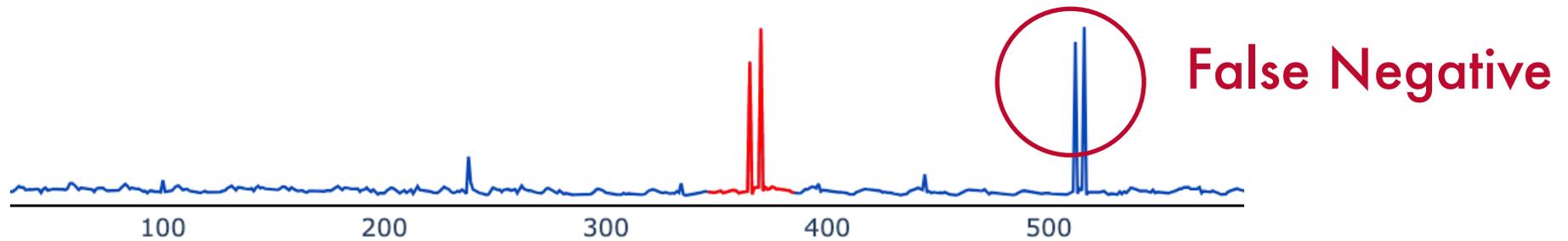


# Common Flaws in Dataset

Mislabeling

Bias

Feasibility

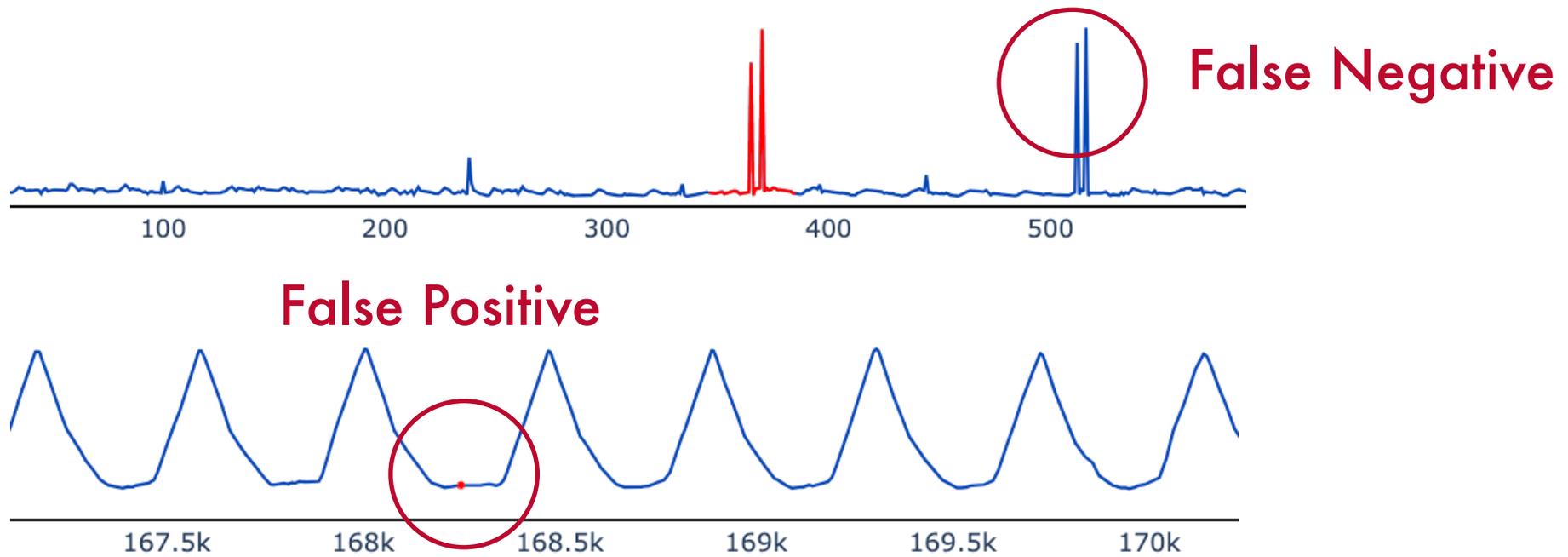


# Common Flaws in Dataset

Mislabeling

Bias

Feasibility

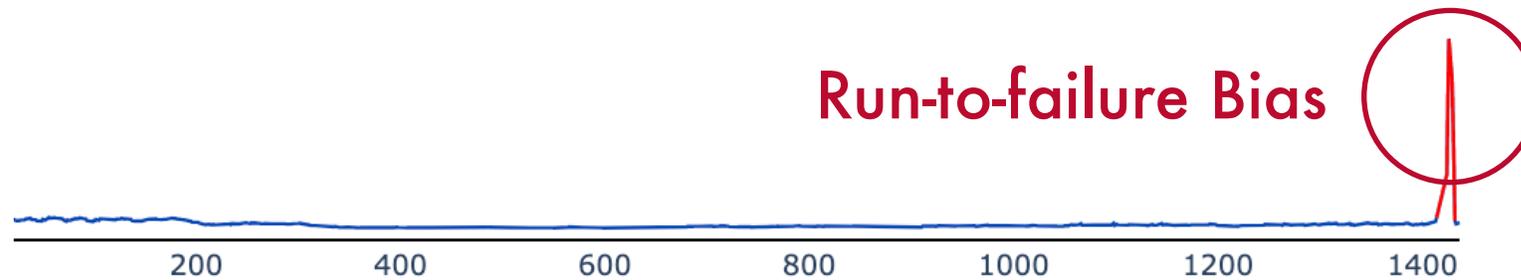


# Common Flaws in Dataset

Mislabeling

Bias

Feasibility



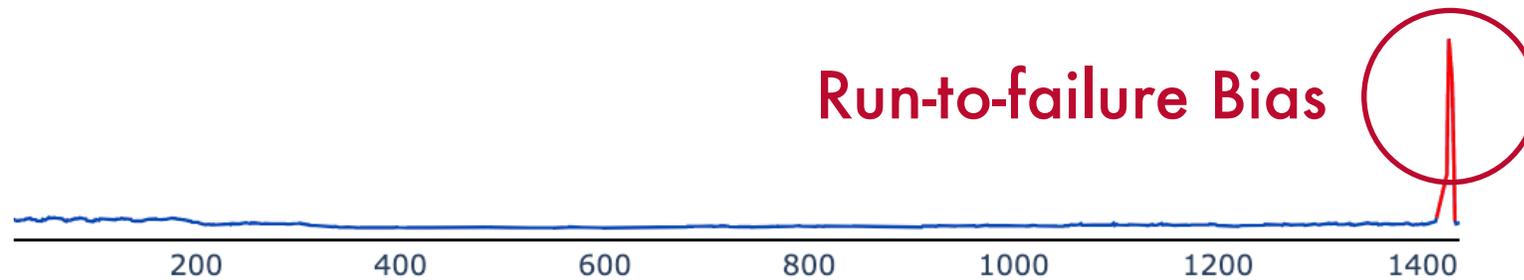
# Common Flaws in Dataset

Mislabeling

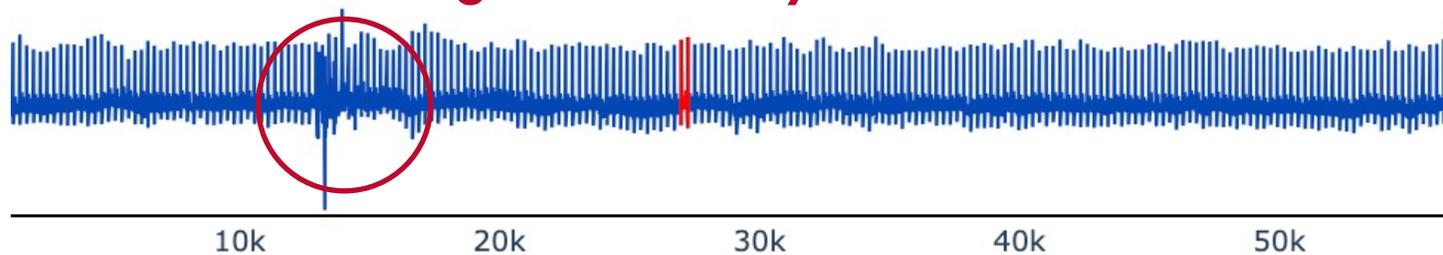
Bias

Feasibility

Run-to-failure Bias



Single Anomaly Bias



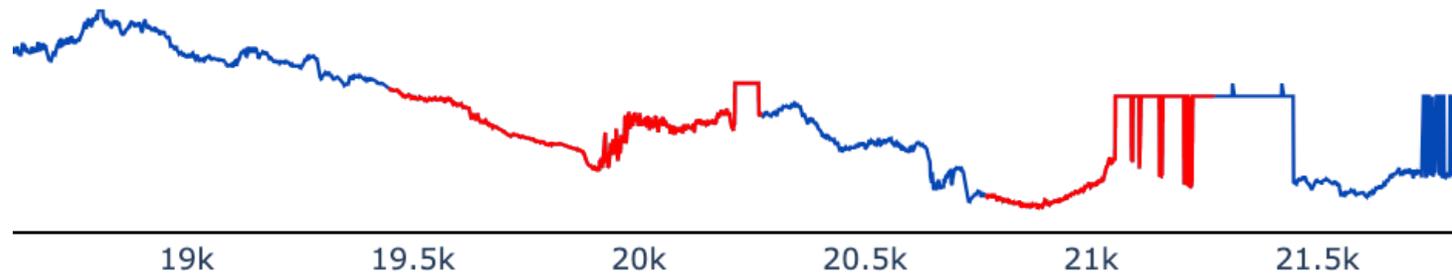
# Common Flaws in Dataset

Mislabeling

Bias

Feasibility

Lack of In-context Data



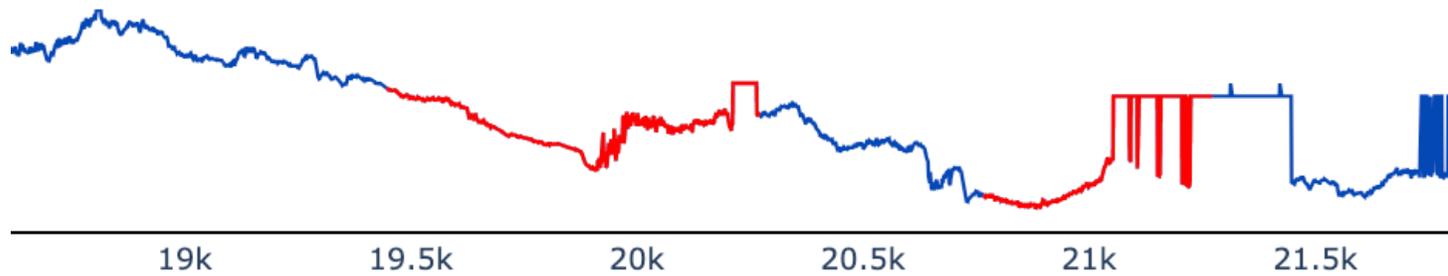
# Common Flaws in Dataset

Mislabeling

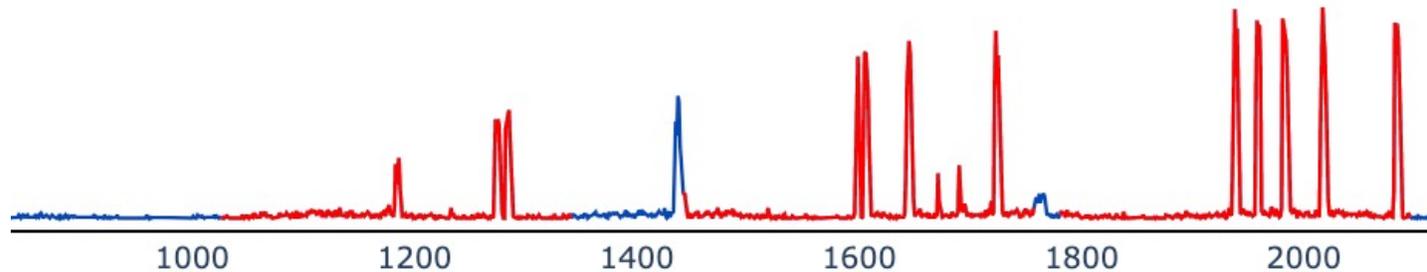
Bias

Feasibility

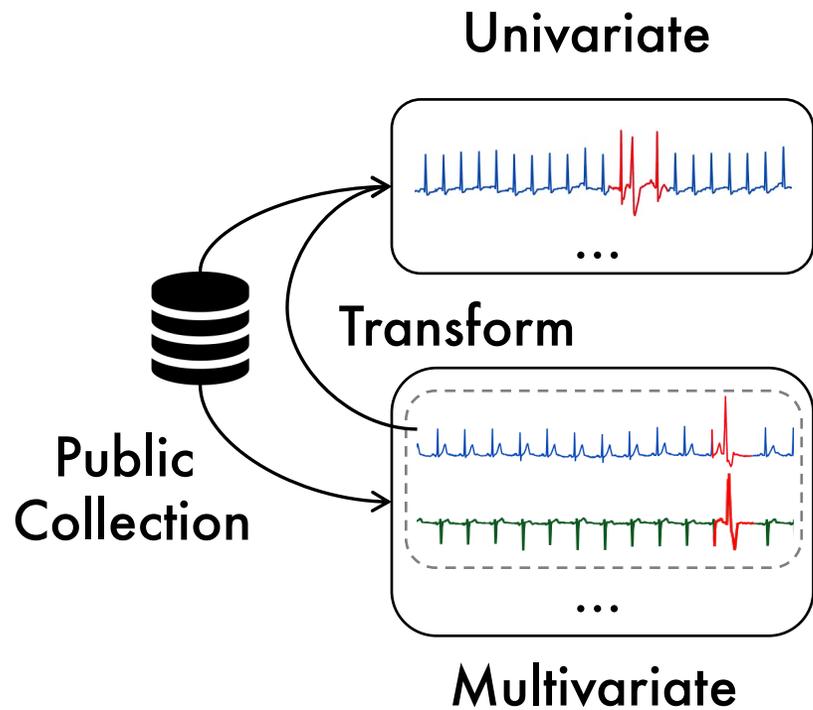
Lack of In-context Data



Unrealistic Anomaly Ratio

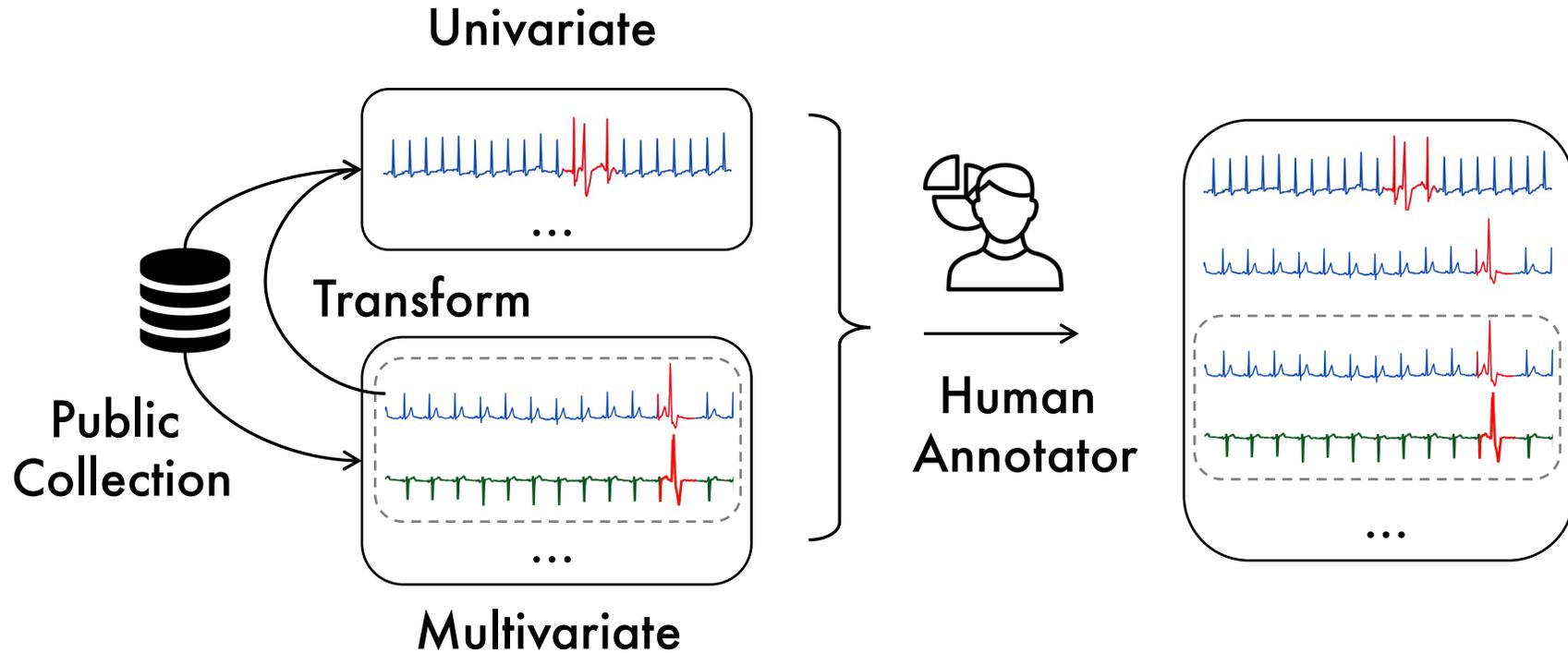


# Dataset Construction Pipeline



Step 1:  
Dataset Collection

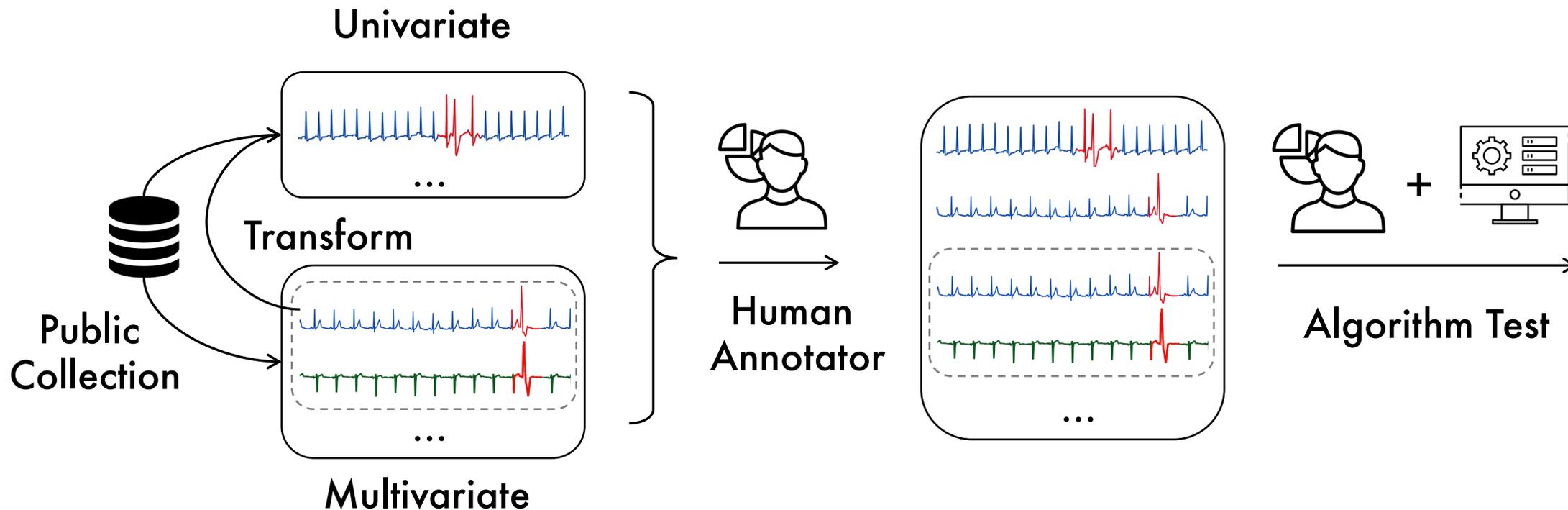
# Dataset Construction Pipeline



Step 1:  
Dataset Collection

Step 2:  
Flaws Identification

# Dataset Construction Pipeline



Step 1:  
Dataset Collection

Step 2:  
Flaws Identification

Step 3:  
Label Quality Assessment

# TSB-AD Dataset Overview

- **TSB-AD** is all you need !
- **Largest** heterogenous and curated time-series anomaly detection dataset

Category	Split	# TS	Avg Length	Avg Anomaly Length	Avg # Anomalies	Anomaly Ratio
TSB-AD-U	All	870	38814.1	179.5	39.7	2.4%
	Eval	350	51886.7	321.3	46.6	4.5%
	Tuning	48	47143.3	185.9	82.6	3.5%
TSB-AD-M	All	200	107760.4	582.6	71.1	5.1%
	Eval	180	108826.7	591.2	67.7	5.0%
	Tuning	20	98164.1	504.7	101.1	5.7%

# TSB-AD Dataset Overview

- **Double** the size of the previous largest collection
- **Four times** the number of existing curated datasets

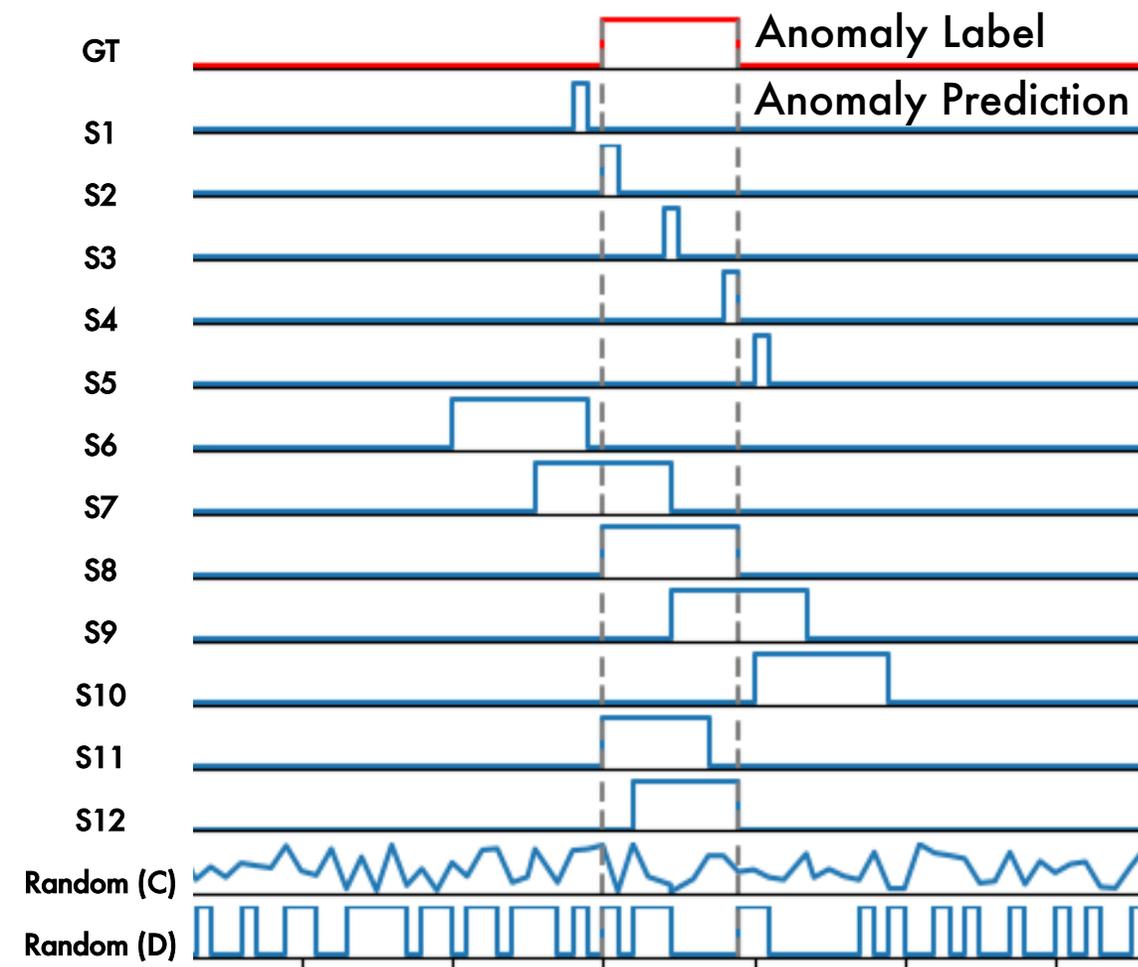
Category	Split	# TS	Avg Length	Avg Anomaly Length	Avg # Anomalies	Anomaly Ratio
TSB-AD-U	All	870	38814.1	179.5	39.7	2.4%
	Eval	350	51886.7	321.3	46.6	4.5%
	Tuning	48	47143.3	185.9	82.6	3.5%
TSB-AD-M	All	200	107760.4	582.6	71.1	5.1%
	Eval	180	108826.7	591.2	67.7	5.0%
	Tuning	20	98164.1	504.7	101.1	5.7%



# MEASURE RELIABILITY

---

# Flaws in Evaluation Measures

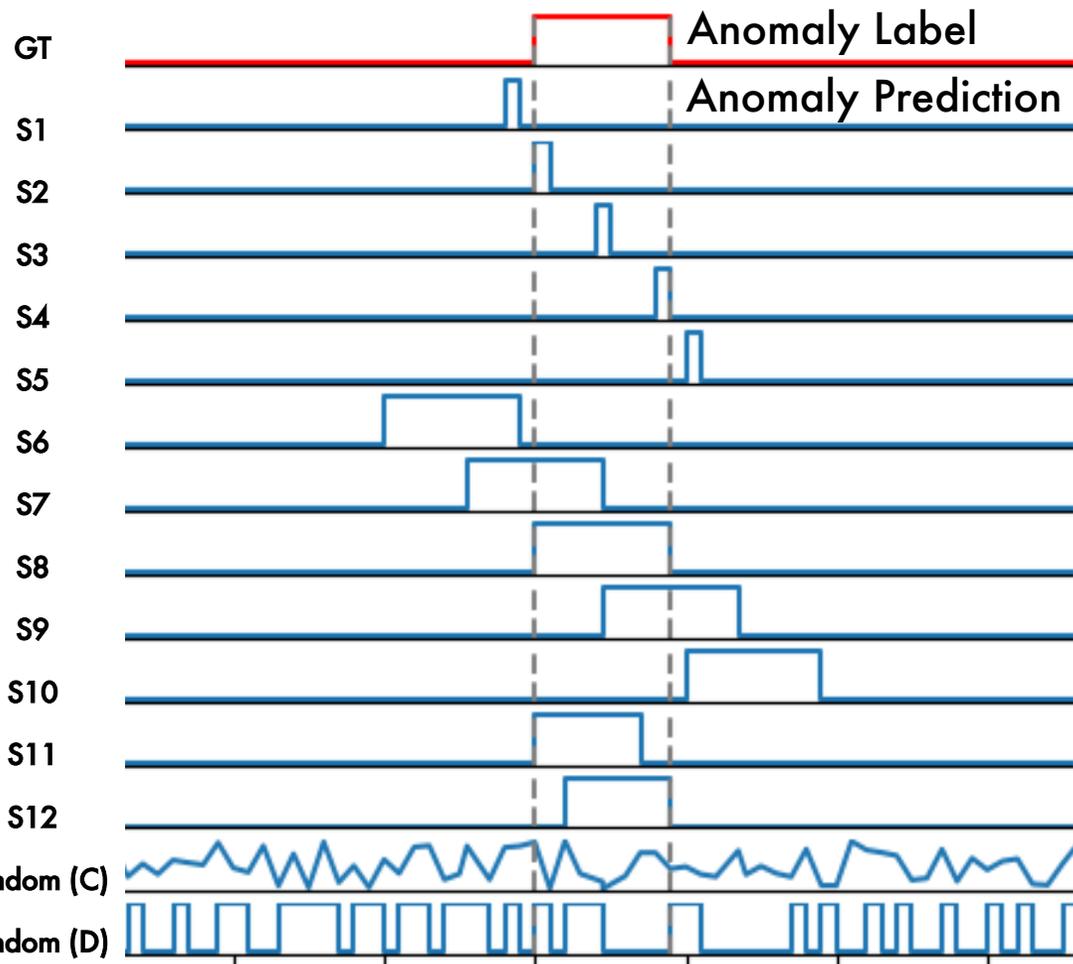


# Flaws in Evaluation Measures

Bias

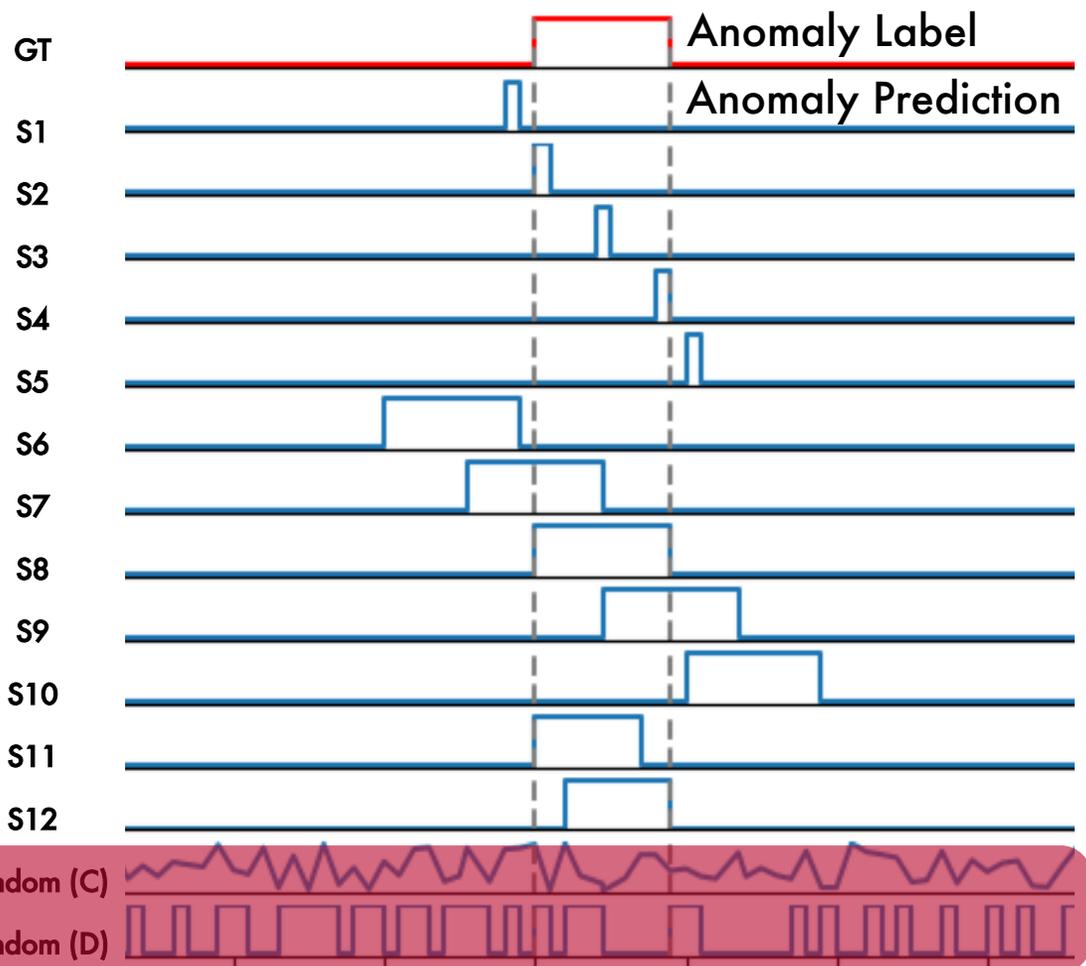
Indiscrimination

Lack of Adaptability



Scenario	Threshold-independent					Threshold-dependent				
	AUC-PR	AUC-ROC	VUS-PR	VUS-ROC	PATE	Standard-F1	PA-F1	Event-based-F1	R-based-F1	Affiliation-F1
S1	0.04	0.50	0.13	0.54	0.03	0.00	0.00	0.00	0.00	0.95
S2	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S3	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.99
S4	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S5	0.04	0.50	0.13	0.54	0.18	0.00	0.00	0.00	0.00	0.95
S6	0.04	0.48	0.18	0.62	0.03	0.00	0.00	0.00	0.00	0.93
S7	0.27	0.74	0.61	0.85	0.68	0.50	0.80	0.67	0.55	0.98
S8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S9	0.27	0.74	0.61	0.85	0.60	0.50	0.80	0.67	0.55	0.98
S10	0.04	0.48	0.18	0.62	0.14	0.00	0.00	0.00	0.00	0.93
S11	0.81	0.90	0.81	0.90	0.96	0.89	1.00	1.00	0.91	1.00
S12	0.81	0.90	0.81	0.90	0.91	0.89	1.00	1.00	0.91	1.00
Random (C)	0.06	0.56	0.09	0.72	0.06	0.12	0.73	0.21	0.16	0.70
Random (D)	0.04	0.51	0.08	0.66	0.34	0.08	0.15	0.08	0.09	0.68

# Flaws in Evaluation Measures



Bias

Indiscrimination

Lack of Adaptability

Scenario	Threshold-independent					Threshold-dependent				
	AUC-PR	AUC-ROC	VUS-PR	VUS-ROC	PATE	Standard-F1	PA-F1	Event-based-F1	R-based-F1	Affiliation-F1
S1	0.04	0.50	0.13	0.54	0.03	0.00	0.00	0.00	0.00	0.95
S2	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S3	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.99
S4	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S5	0.04	0.50	0.13	0.54	0.18	0.00	0.00	0.00	0.00	0.95
S6	0.04	0.48	0.18	0.62	0.03	0.00	0.00	0.00	0.00	0.93
S7	0.27	0.74	0.61	0.85	0.68	0.50	0.80	0.67	0.55	0.98
S8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S9	0.27	0.74	0.61	0.85	0.60	0.50	0.80	0.67	0.55	0.98
S10	0.04	0.48	0.18	0.62	0.14	0.00	0.00	0.00	0.00	0.93
S11	0.81	0.90	0.81	0.90	0.96	0.89	1.00	1.00	0.91	1.00
S12	0.81	0.90	0.81	0.90	0.91	0.89	1.00	1.00	0.91	1.00
Random (C)	0.06	0.56	0.09	0.72	0.06	0.12	0.73	0.21	0.16	0.70
Random (D)	0.04	0.51	0.08	0.66	0.34	0.08	0.15	0.08	0.09	0.68

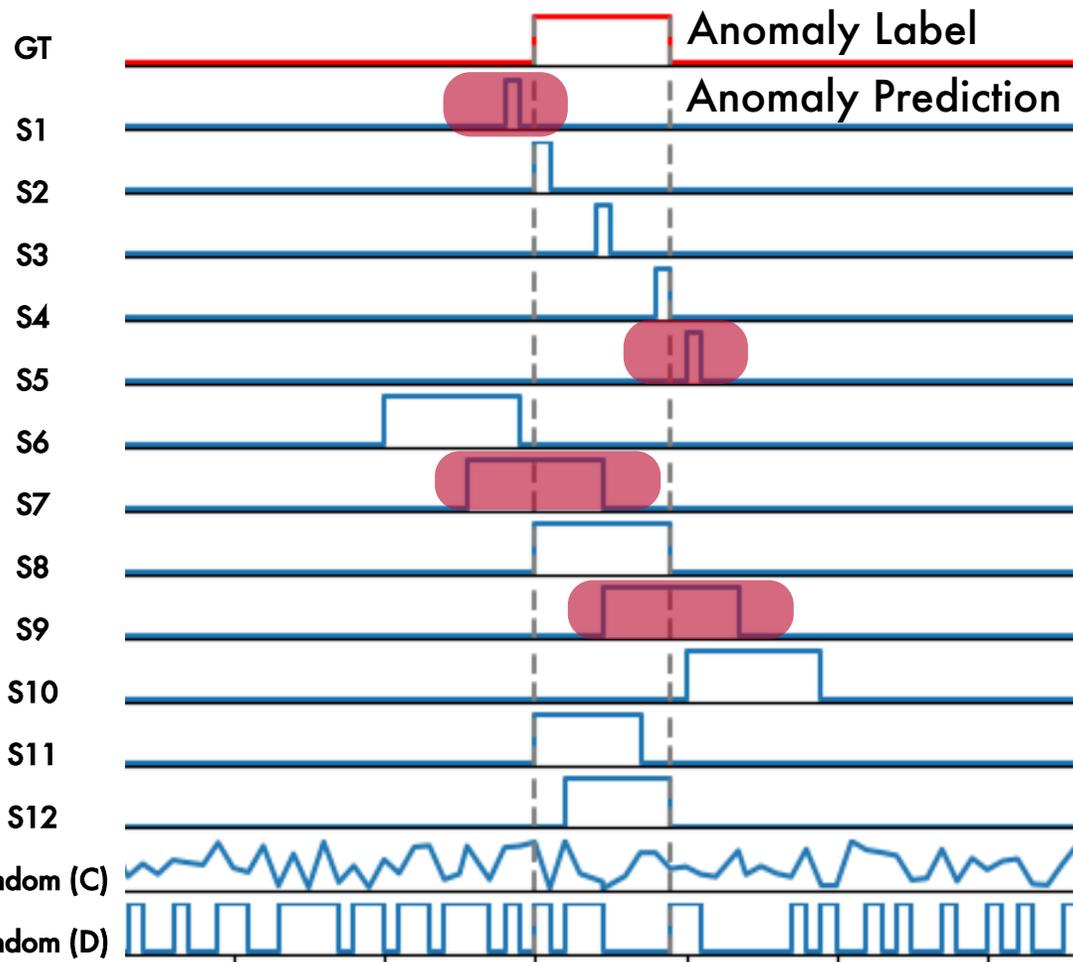
Bias towards random score

# Flaws in Evaluation Measures

Bias

Indiscrimination

Lack of Adaptability



Scenario	Threshold-independent					Threshold-dependent				
	AUC-PR	AUC-ROC	VUS-PR	VUS-ROC	PATE	Standard-F1	PA-F1	Event-based-F1	R-based-F1	Affiliation-F1
S1	0.04	0.50	0.13	0.54	0.03	0.00	0.00	0.00	0.00	0.95
S2	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S3	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.99
S4	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S5	0.04	0.50	0.13	0.54	0.18	0.00	0.00	0.00	0.00	0.95
S6	0.04	0.48	0.18	0.62	0.03	0.00	0.00	0.00	0.00	0.93
S7	0.27	0.74	0.61	0.85	0.68	0.50	0.80	0.67	0.55	0.98
S8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S9	0.27	0.74	0.61	0.85	0.60	0.50	0.80	0.67	0.55	0.98
S10	0.04	0.48	0.18	0.62	0.14	0.00	0.00	0.00	0.00	0.93
S11	0.81	0.90	0.81	0.90	0.96	0.89	1.00	1.00	0.91	1.00
S12	0.81	0.90	0.81	0.90	0.91	0.89	1.00	1.00	0.91	1.00
Random (C)	0.06	0.56	0.09	0.72	0.06	0.12	0.73	0.21	0.16	0.70
Random (D)	0.04	0.51	0.08	0.66	0.34	0.08	0.15	0.08	0.09	0.68

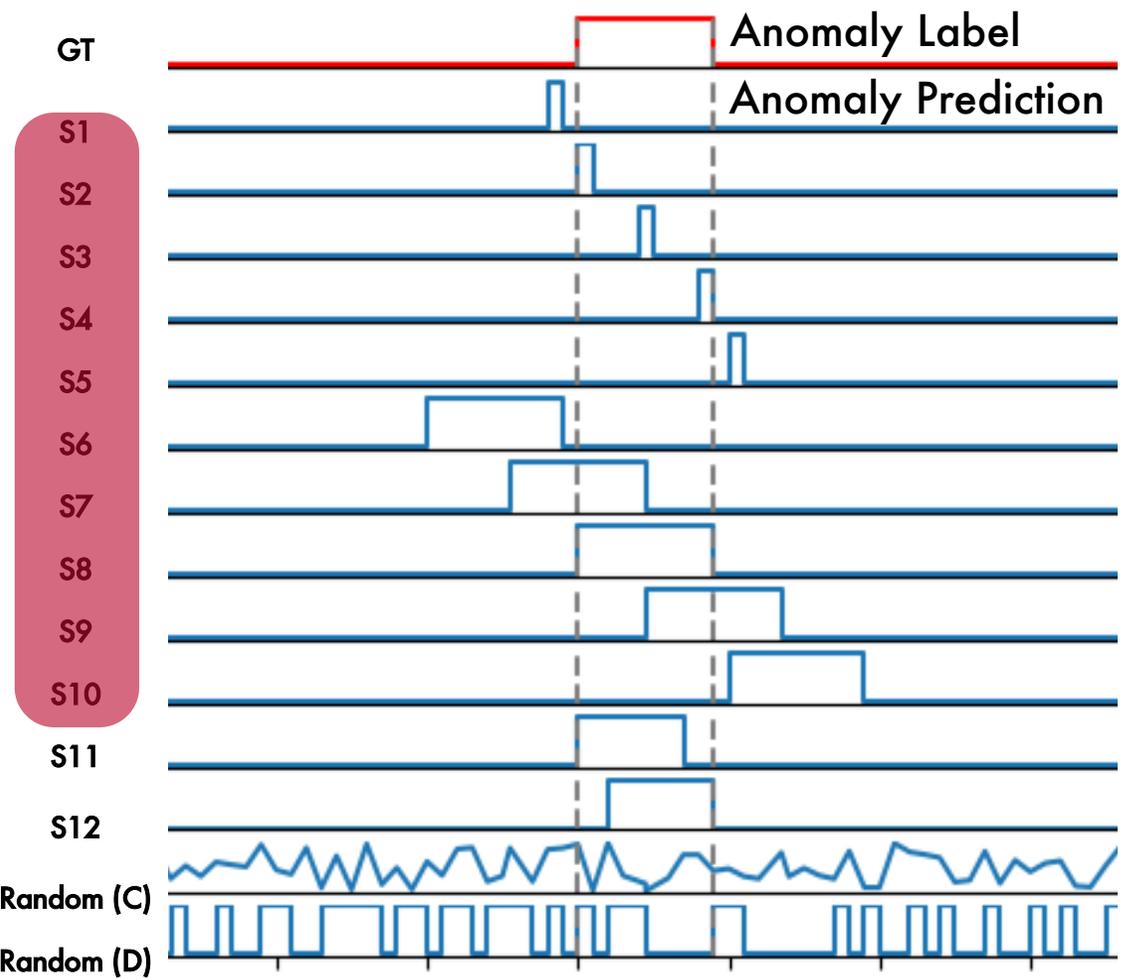
Inconsistent evaluation across different scenarios

# Flaws in Evaluation Measures

Bias

Indiscrimination

Lack of Adaptability



Scenario	Threshold-independent					Threshold-dependent				
	AUC-PR	AUC-ROC	VUS-PR	VUS-ROC	PATE	Standard-F1	PA-F1	Event-based-F1	R-based-F1	Affiliation-F1
S1	0.04	0.50	0.13	0.54	0.03	0.00	0.00	0.00	0.00	0.95
S2	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S3	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.99
S4	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S5	0.04	0.50	0.13	0.54	0.18	0.00	0.00	0.00	0.00	0.95
S6	0.04	0.48	0.18	0.62	0.03	0.00	0.00	0.00	0.00	0.93
S7	0.27	0.74	0.61	0.85	0.68	0.50	0.80	0.67	0.55	0.98
S8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S9	0.27	0.74	0.61	0.85	0.60	0.50	0.80	0.67	0.55	0.98
S10	0.04	0.48	0.18	0.62	0.14	0.00	0.00	0.00	0.00	0.93
S11	0.81	0.90	0.81	0.90	0.96	0.89	1.00	1.00	0.91	1.00
S12	0.81	0.90	0.81	0.90	0.91	0.89	1.00	1.00	0.91	1.00
Random (C)	0.06	0.56	0.09	0.72	0.06	0.12	0.73	0.21	0.16	0.70
Random (D)	0.04	0.51	0.08	0.66	0.34	0.08	0.15	0.08	0.09	0.68

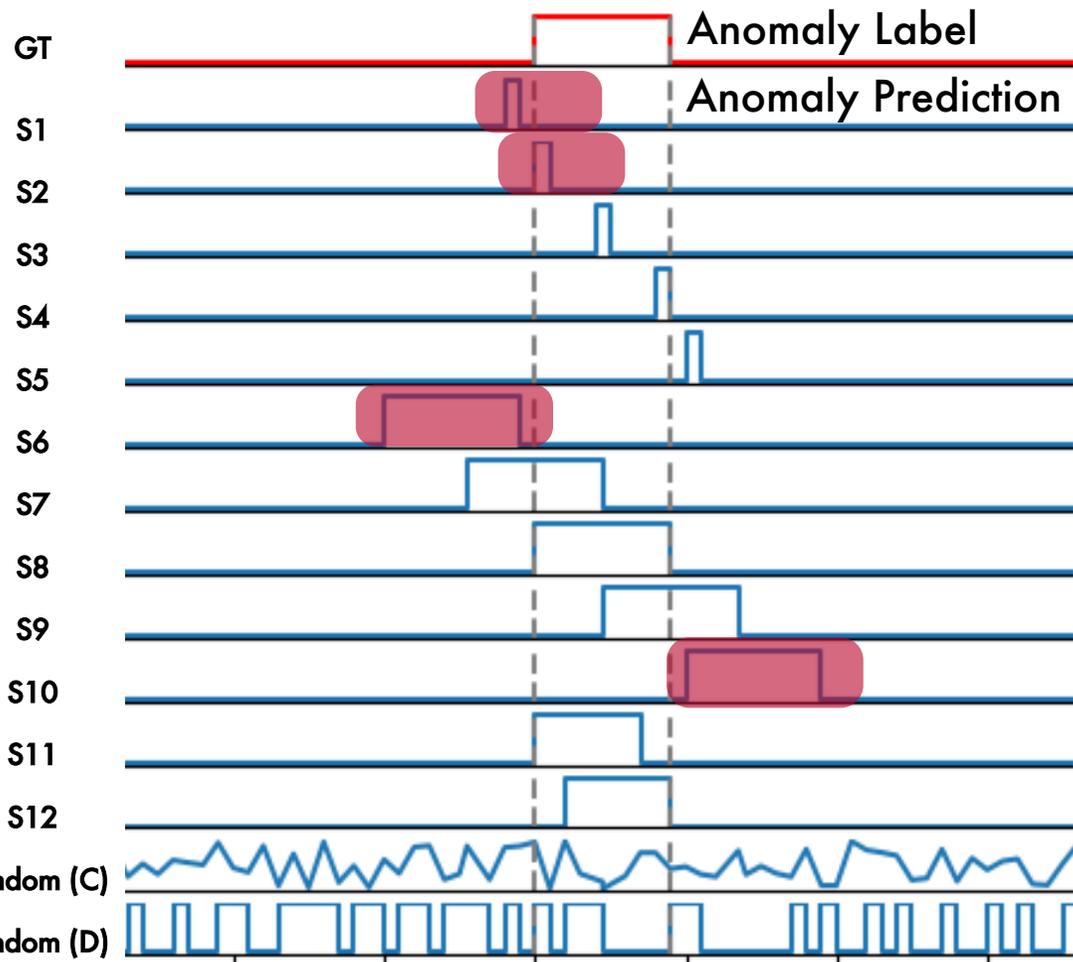
No differentiation across S1 to S10

# Flaws in Evaluation Measures

Bias

Indiscrimination

Lack of Adaptability

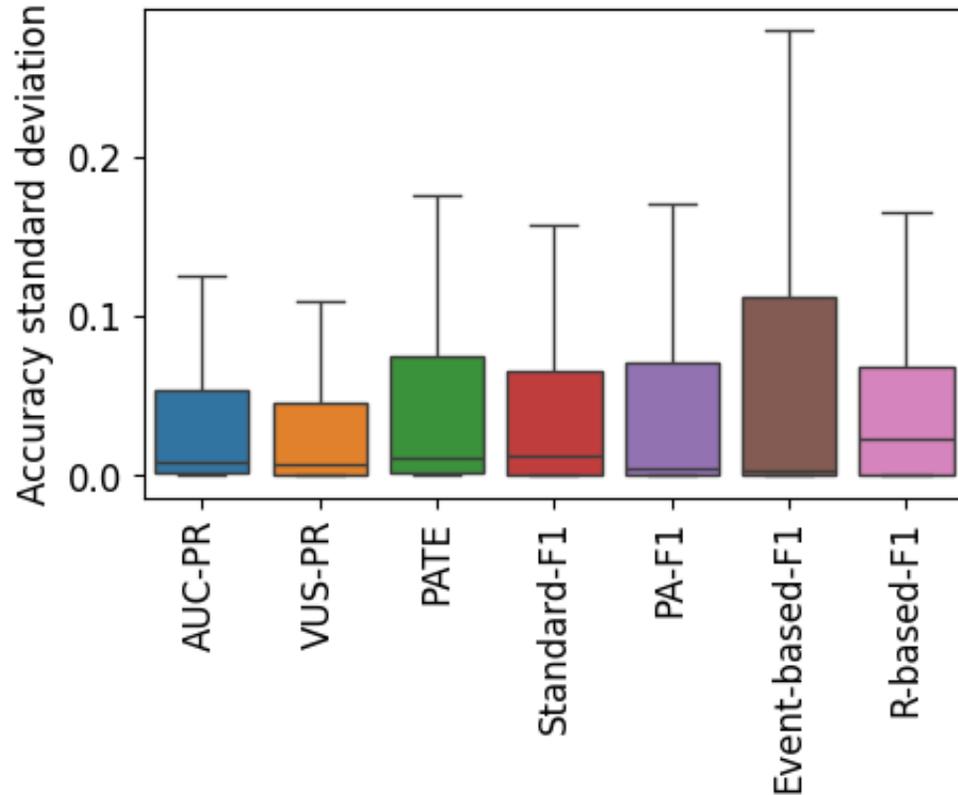


Scenario	Threshold-independent					Threshold-dependent				
	AUC-PR	AUC-ROC	VUS-PR	VUS-ROC	PATE	Standard-F1	PA-F1	Event-based-F1	R-based-F1	Affiliation-F1
S1	0.04	0.50	0.13	0.54	0.03	0.00	0.00	0.00	0.00	0.95
S2	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S3	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.99
S4	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S5	0.04	0.50	0.13	0.54	0.18	0.00	0.00	0.00	0.00	0.95
S6	0.04	0.48	0.18	0.62	0.03	0.00	0.00	0.00	0.00	0.93
S7	0.27	0.74	0.61	0.85	0.68	0.50	0.80	0.67	0.55	0.98
S8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S9	0.27	0.74	0.61	0.85	0.60	0.50	0.80	0.67	0.55	0.98
S10	0.04	0.48	0.18	0.62	0.14	0.00	0.00	0.00	0.00	0.93
S11	0.81	0.90	0.81	0.90	0.96	0.89	1.00	1.00	0.91	1.00
S12	0.81	0.90	0.81	0.90	0.91	0.89	1.00	1.00	0.91	1.00
Random (C)	0.06	0.56	0.09	0.72	0.06	0.12	0.73	0.21	0.16	0.70
Random (D)	0.04	0.51	0.08	0.66	0.34	0.08	0.15	0.08	0.09	0.68

Fail to account for time series nature

# Investigation of Evaluation Measures

## Sensitivity to Lags



Scenario	Threshold-independent					Threshold-dependent				
	AUC-PR	AUC-ROC	VUS-PR	VUS-ROC	PATE	Standard-F1	PA-F1	Event-based-F1	R-based-F1	Affiliation-F1
S1	0.04	0.50	0.13	0.54	0.03	0.00	0.00	0.00	0.00	0.95
S2	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S3	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.99
S4	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S5	0.04	0.50	0.13	0.54	0.18	0.00	0.00	0.00	0.00	0.95
S6	0.04	0.48	0.18	0.62	0.03	0.00	0.00	0.00	0.00	0.93
S7	0.27	0.74	0.61	0.85	0.68	0.50	0.80	0.67	0.55	0.98
S8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S9	0.27	0.74	0.61	0.85	0.60	0.50	0.80	0.67	0.55	0.98
S10	0.04	0.48	0.18	0.62	0.14	0.00	0.00	0.00	0.00	0.93
S11	0.81	0.90	0.81	0.90	0.96	0.89	1.00	1.00	0.91	1.00
S12	0.81	0.90	0.81	0.90	0.91	0.89	1.00	1.00	0.91	1.00
Random (C)	0.06	0.56	0.09	0.72	0.06	0.12	0.73	0.21	0.16	0.70
Random (D)	0.04	0.51	0.08	0.66	0.34	0.08	0.15	0.08	0.09	0.68

# Investigation of Evaluation Measures

VUS-PR emerges to be the most accurate and reliable evaluation measure

Scenario	Threshold-independent					Threshold-dependent				
	AUC-PR	AUC-ROC	VUS-PR	VUS-ROC	PATE	Standard-F1	PA-F1	Event-based-F1	R-based-F1	Affiliation-F1
S1	0.04	0.50	0.13	0.54	0.03	0.00	0.00	0.00	0.00	0.95
S2	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S3	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.99
S4	0.14	0.55	0.17	0.56	0.58	0.18	1.00	1.00	0.44	0.98
S5	0.04	0.50	0.13	0.54	0.18	0.00	0.00	0.00	0.00	0.95
S6	0.04	0.48	0.18	0.62	0.03	0.00	0.00	0.00	0.00	0.93
S7	0.27	0.74	0.61	0.85	0.68	0.50	0.80	0.67	0.55	0.98
S8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
S9	0.27	0.74	0.61	0.85	0.60	0.50	0.80	0.67	0.55	0.98
S10	0.04	0.48	0.18	0.62	0.14	0.00	0.00	0.00	0.00	0.93
S11	0.81	0.90	0.81	0.90	0.96	0.89	1.00	1.00	0.91	1.00
S12	0.81	0.90	0.81	0.90	0.91	0.89	1.00	1.00	0.91	1.00
Random (C)	0.06	0.56	0.09	0.72	0.06	0.12	0.73	0.21	0.16	0.70
Random (D)	0.04	0.51	0.08	0.66	0.34	0.08	0.15	0.08	0.09	0.68

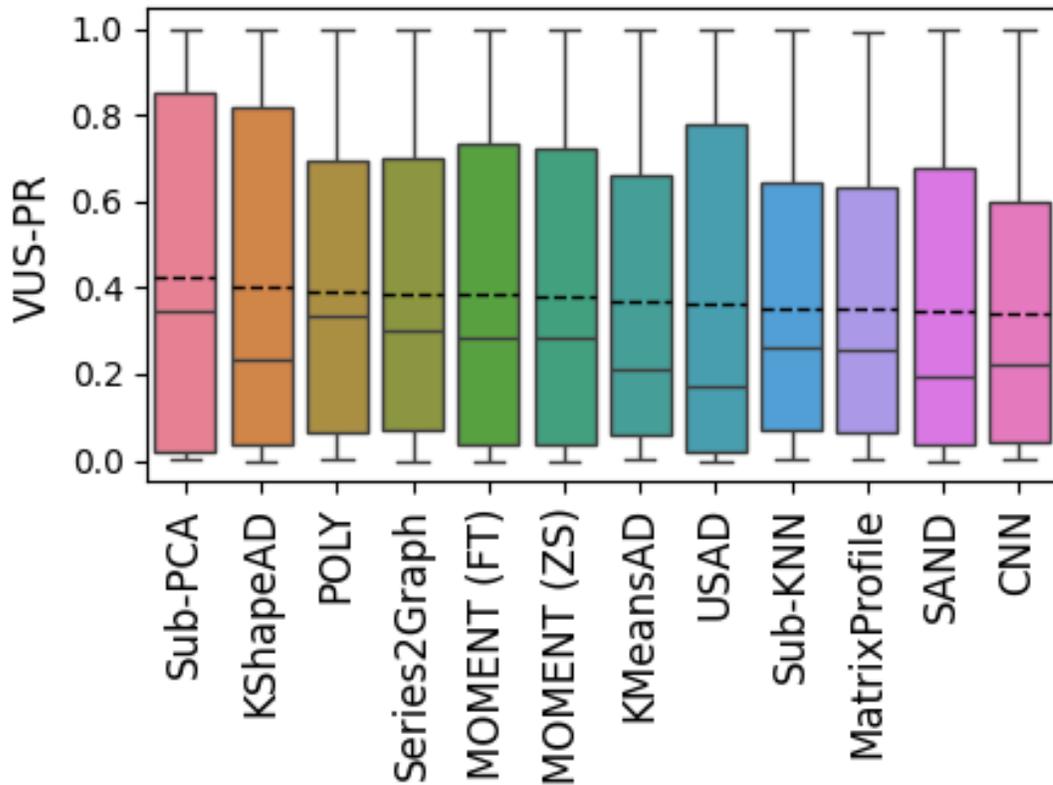


# BENCHMARKING

---

# Benchmark Accuracy Evaluation

## TSB-AD-U

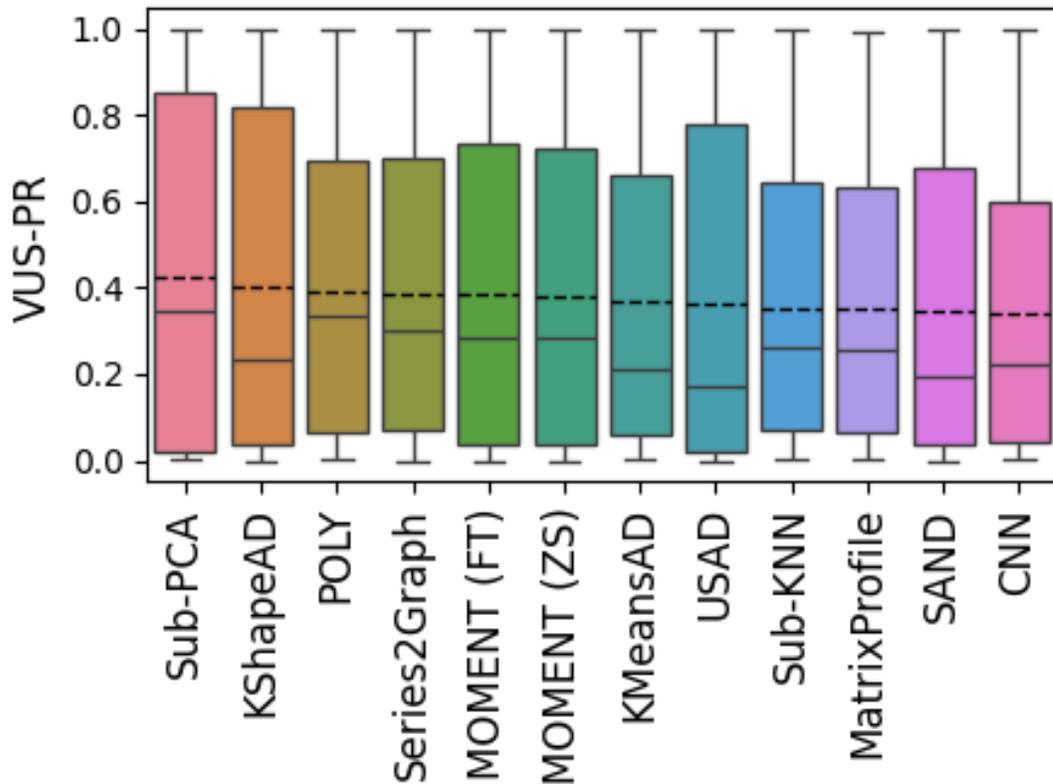


## VUS-PR Ranking

1	Sub-PCA
2	KShapeAD
3	POLY
4	Series2Graph
5	MOMENT (FT)
6	MOMENT (ZS)
7	KMeansAD
8	USAD
9	Sub-KNN
10	MatrixProfile
11	SAND
12	CNN

# Benchmark Accuracy Evaluation

## TSB-AD-U



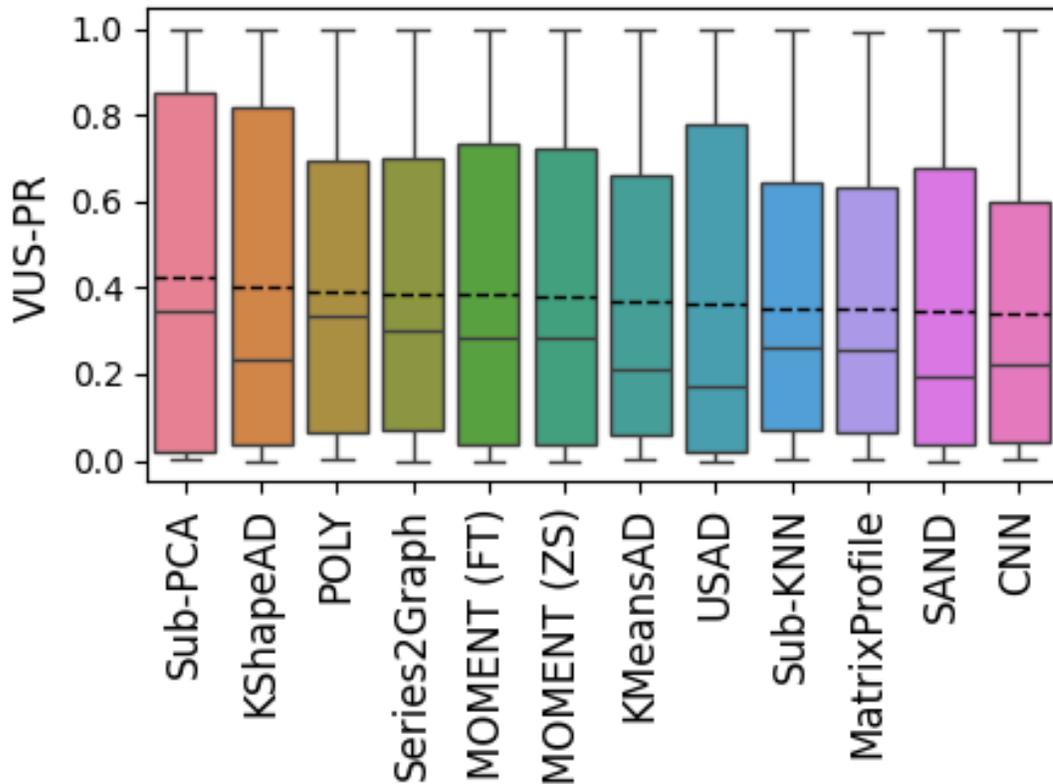
## VUS-PR Ranking

1	Sub-PCA
2	KShapeAD
3	POLY
4	Series2Graph
5	MOMENT (FT)
6	MOMENT (ZS)
7	KMeansAD
8	USAD
9	Sub-KNN
10	MatrixProfile
11	SAND
12	CNN

① Top-performing methods been overlooked for many years

# Benchmark Accuracy Evaluation

## TSB-AD-U



## VUS-PR Ranking

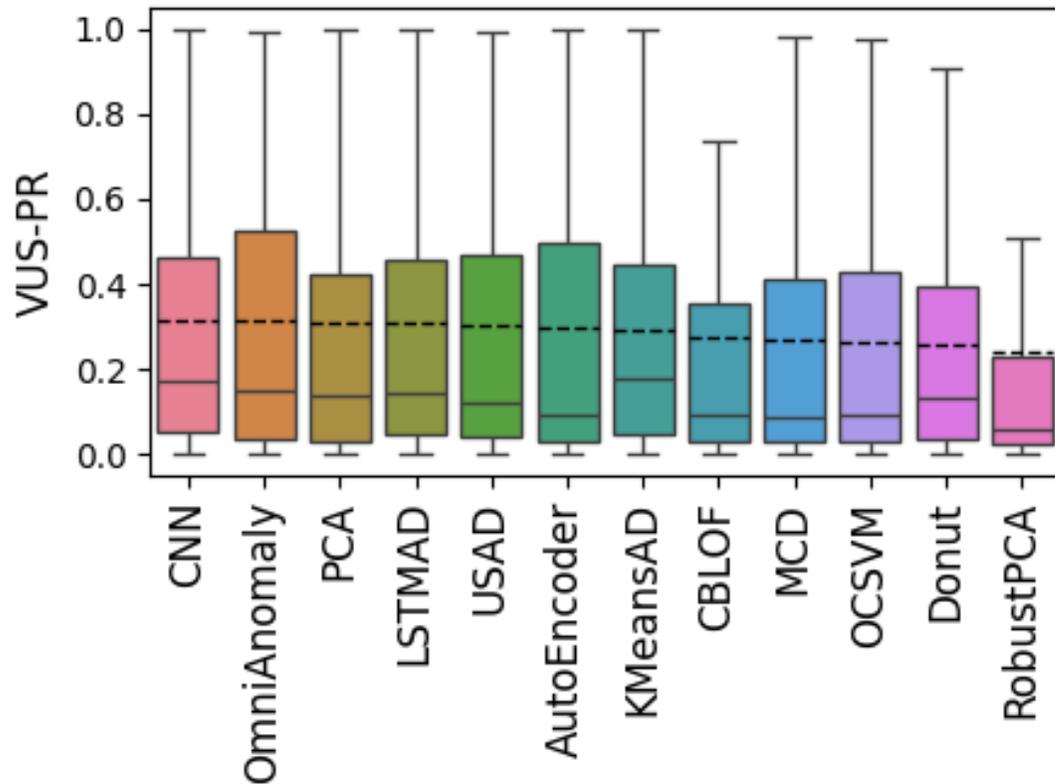
1	Sub-PCA
2	KShapeAD
3	POLY
4	Series2Graph
5	MOMENT (FT)
6	MOMENT (ZS)
7	KMeansAD
8	USAD
9	Sub-KNN
10	MatrixProfile
11	SAND
12	CNN

① Top-performing methods been overlooked for many years

② Performance of time-series foundation models shows promise

# Benchmark Accuracy Evaluation

## TSB-AD-M



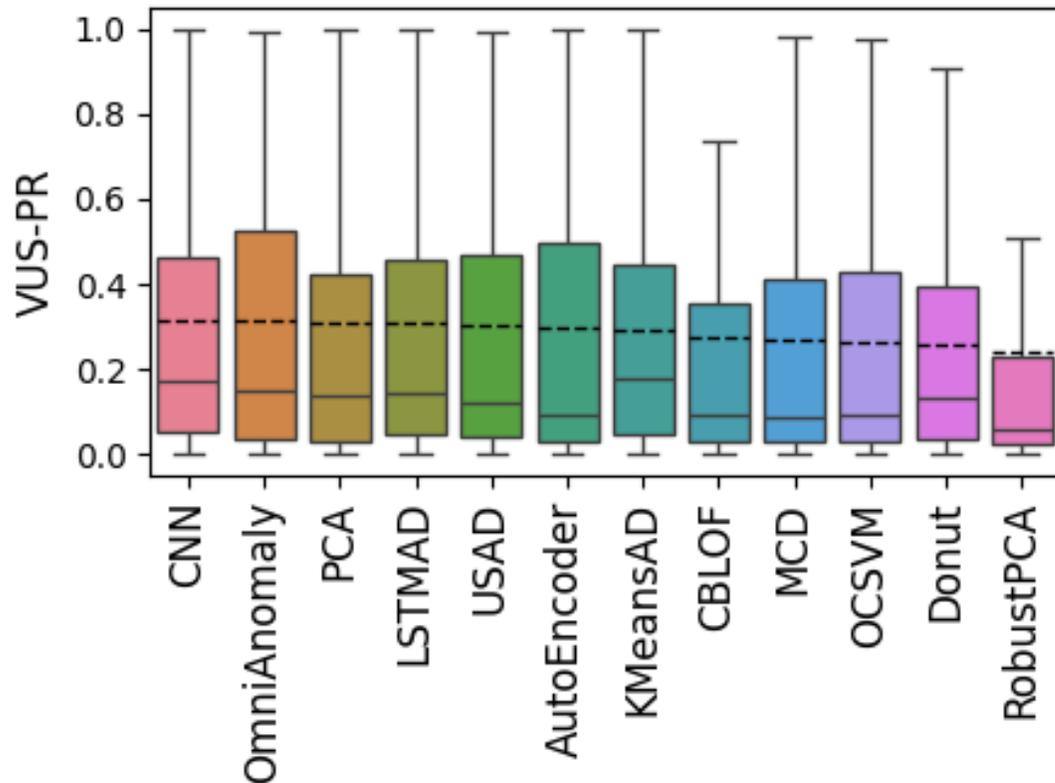
## VUS-PR Ranking

1	CNN
2	OmniAnomaly
3	PCA
4	LSTMAD
5	USAD
6	AutoEncoder
7	KMeansAD
8	CBLOF
9	MCD
10	OCSVM
11	Donut
12	RobustPCA

③ Neural-network-based methods strive in multivariate cases

# Benchmark Accuracy Evaluation

## TSB-AD-M



## VUS-PR Ranking

1	CNN
2	OmniAnomaly
3	PCA
4	LSTMAD
5	USAD
6	AutoEncoder
7	KMeansAD
8	CBLOF
9	MCD
10	OCSVM
11	Donut
12	RobustPCA

③ Neural-network-based methods strive in multivariate cases

④ Simpler architectures generally outperform more complex designs

# THANK YOU



**Benchmark**



**Leaderboard**

---

**Contact:**

**liu.11085@osu.edu**

**paparrizos.1@osu.edu**