Data Quality & Data Integrity

October 18, 2022

Kaveri Mahapatra
Power System Research Engineer

Pacific Northwest National Laboratory

U.S. DEPARTMENT OF ENERGY · BATTELLE

PNNL is operated by Battelle for the U.S. Department of Energy

Outline

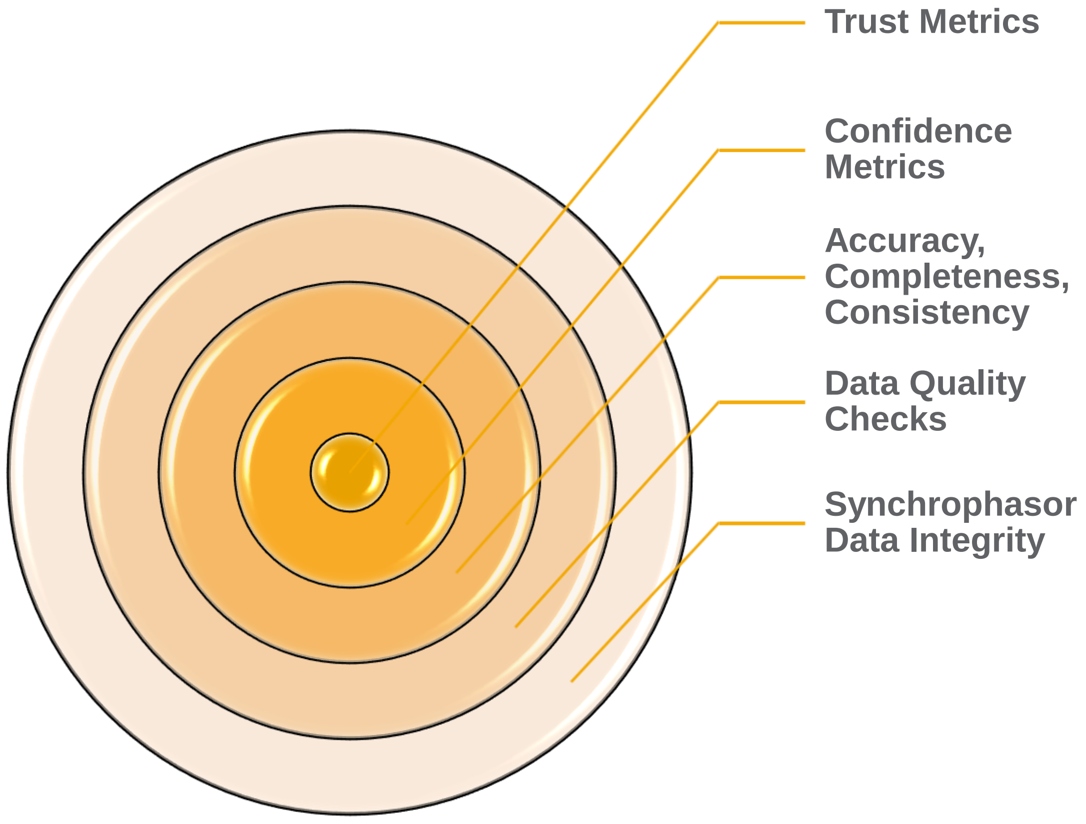
- Data Integrity and Data Quality
- Motivation for SENTIENT
- Sample Datasets
- SENTIENT Anomaly Detection Platform (ADP)
- Workflow
- Metrics
- Results
- Introduction to SENTIENT Testbed and Noise Emulation Platform
- Summary & Future work

Data Integrity and Data Quality

- Data integrity is the maintenance of, and the assurance of, data accuracy and consistency over its entire life-cycle and is a critical aspect to the design, implementation, and usage of any system that stores, processes, or retrieves data*.

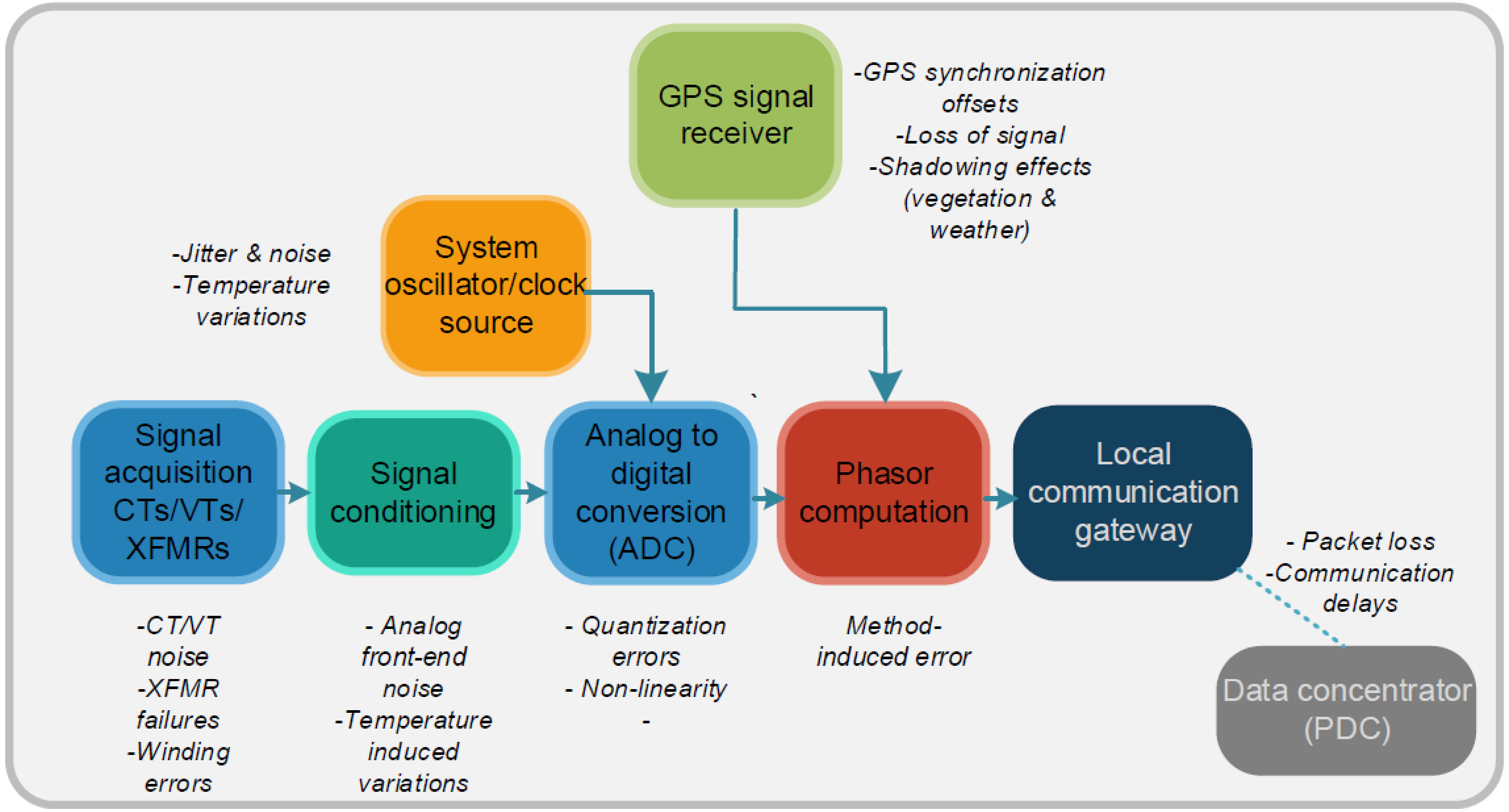
- Data quality is a measure of the condition of data based on factors such as accuracy, completeness, consistency, reliability and whether it's up to date. Measuring data quality levels helps identifying data errors that need to be resolved and assess whether the data is fit to serve its intended purpose.**

- Data from sensors is critical for advanced applications that support efficient, reliable, and resilient electric grid operations.
- Data validation is a prerequisite for data integrity.



** https://www.techtarget.com/searchdatamanagement/definition/data-quality
*https://www.talend.com/resources/what-is-data-integrity/

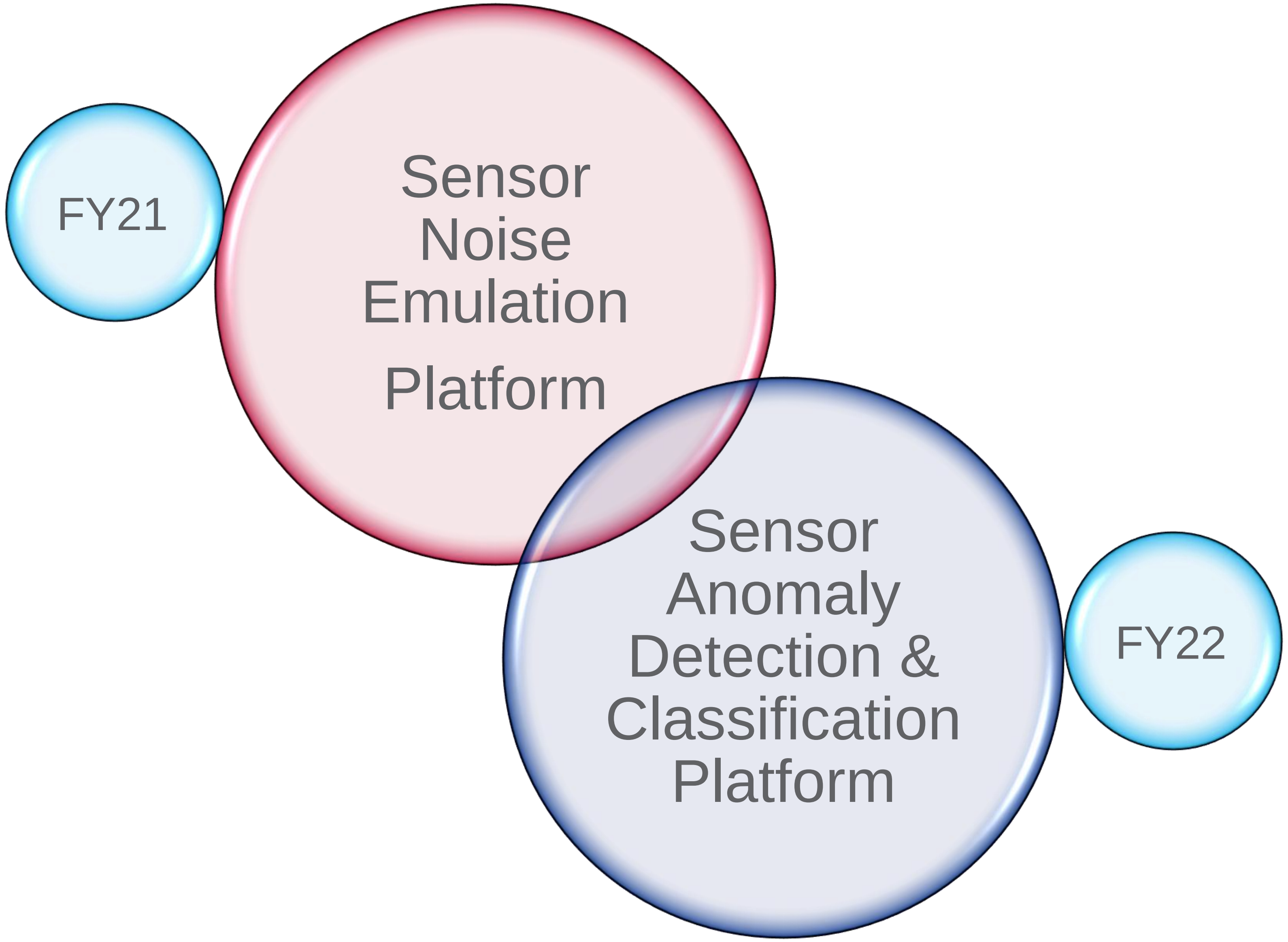
Motivation and SENTIENT objectives



Common Sources of Error in PMU devices

- Develop advanced sensing and measurement tool with data and visual analytics
- Predict sensor behavior and detect/forecast anomalies/issues
- Demonstrate SENTIENT with realistic and accurate noise and bias models.
- Perform operational comparison to discover deviations from expected behavior.
- Facilitate physics-based data-driven means to conduct operational planning and improve grid resiliency
- Use operating conditions of sensors and discover various colors of noise and bias for accurate sensor-based grid modeling.
- Ensure the detection and data analytics capabilities of the ML engine

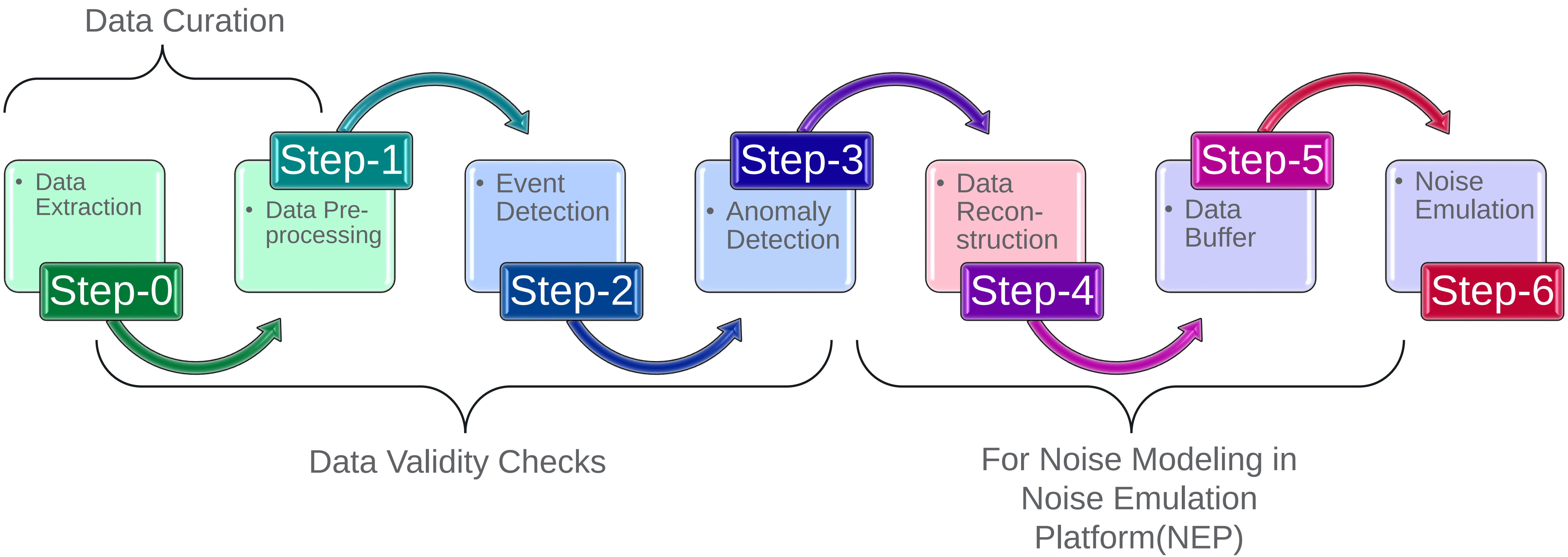
SENTIENT Project overview



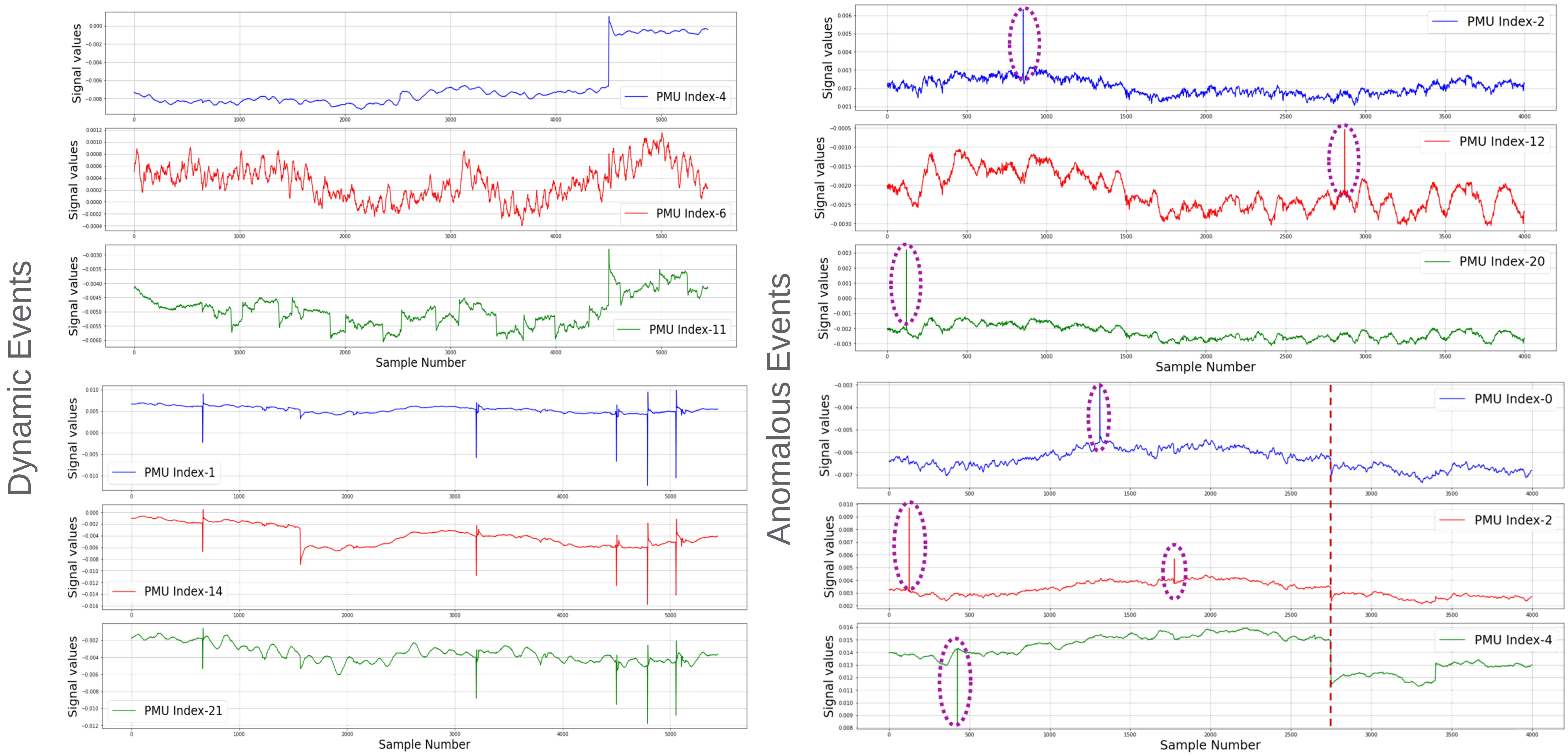
SENTIENT tasks

- Incorporate physics-based models with sensor system
- Discover abnormal deviations, degrading sensor performance, and Predict sensor failures
- Implement in a distributed application architecture

SENTIENT Anomaly Detection Platform (ADP)



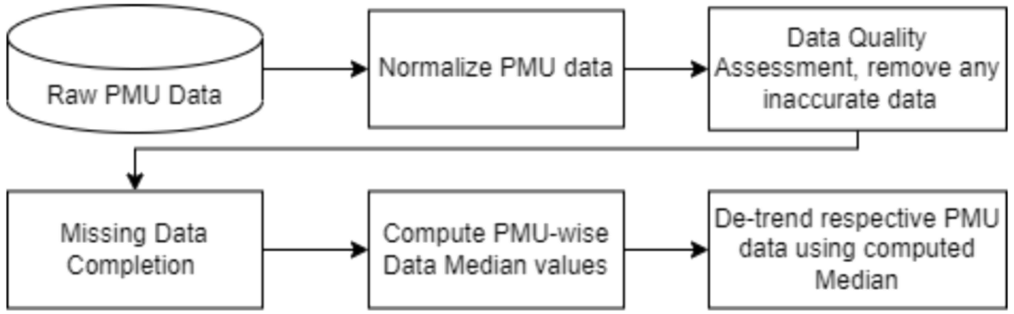
Sample Datasets used for Training the Algorithm



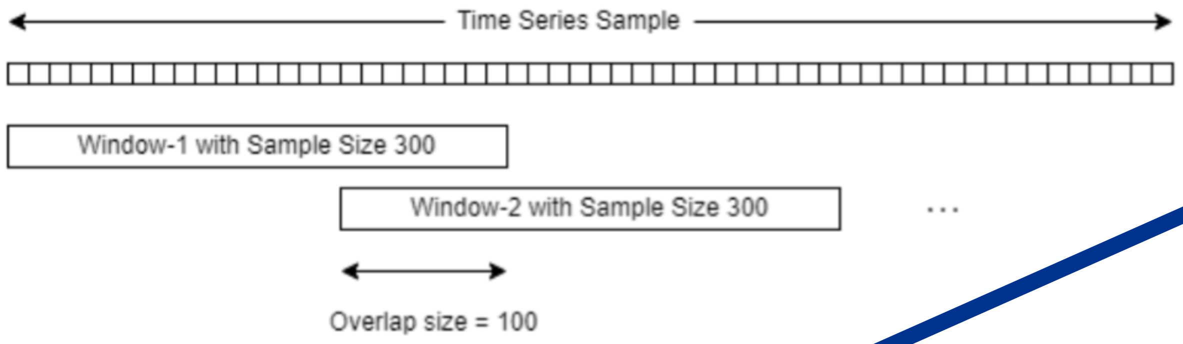
Workflow

A total of 2340 data instances (2+mins) each containing 23 PMU data channels collected over a month

Preprocessing

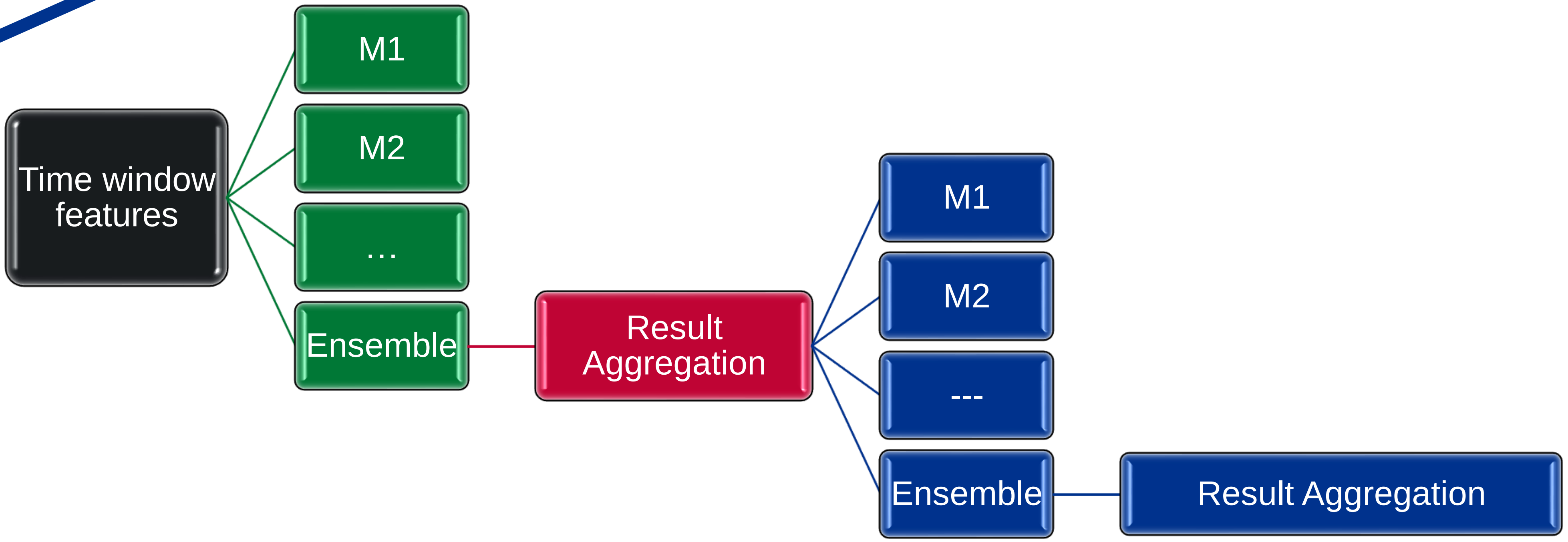


Moving window extraction



Ensemble Method

M1, M2,… are Unsupervised Binary Classifiers

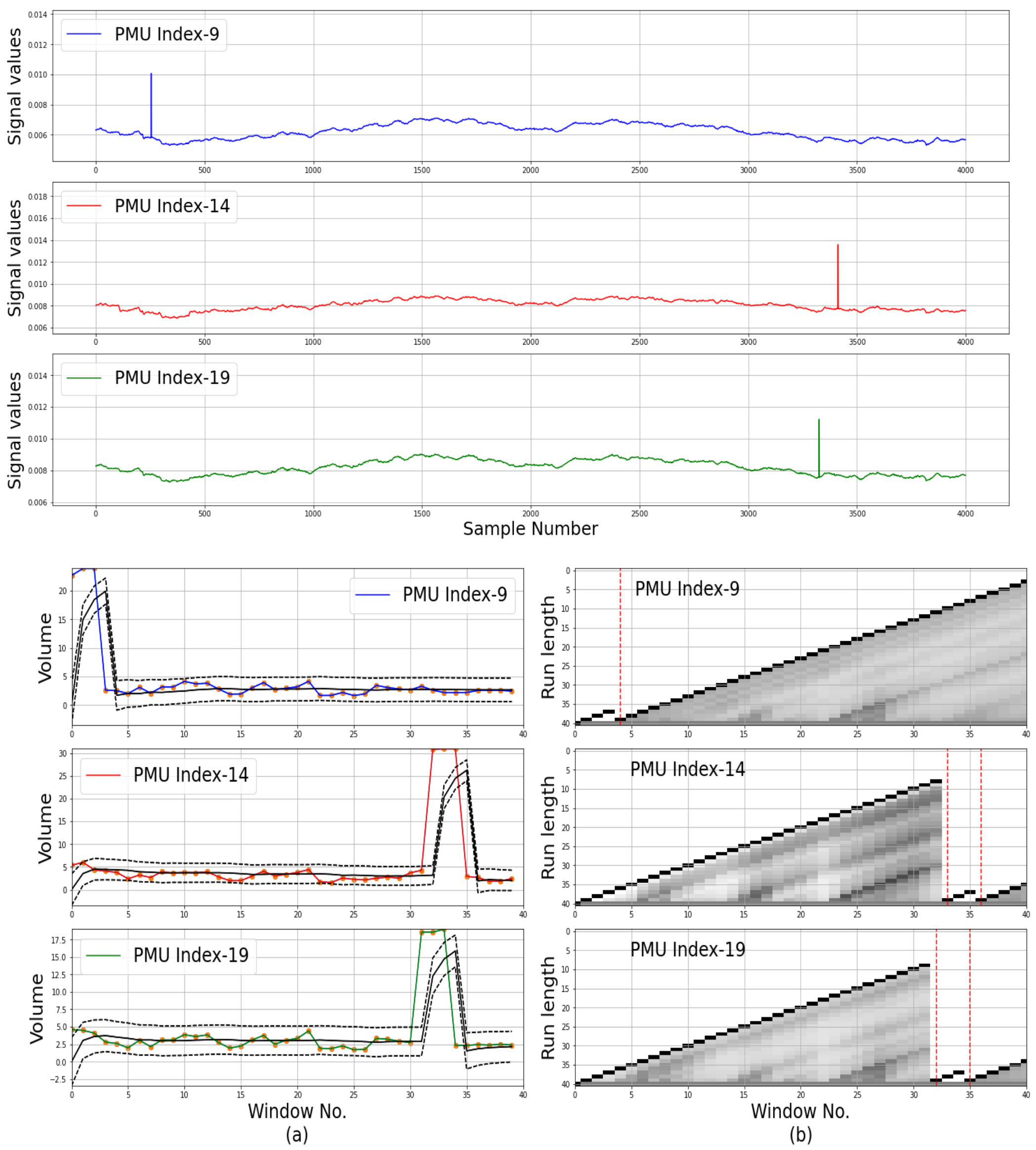


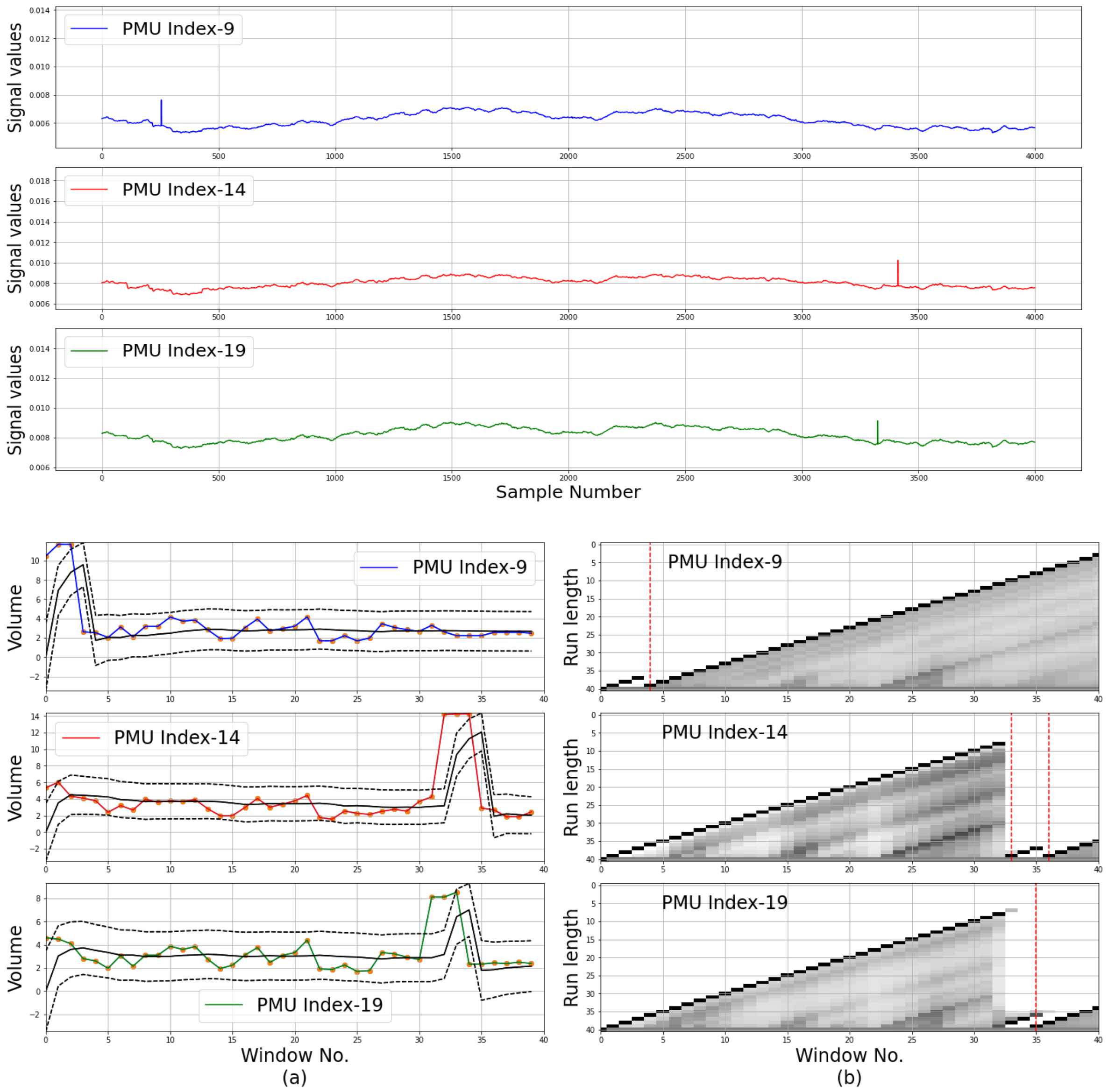
Statistical Metrics for Event/Anomaly Detection

True Label	Predictor:M1	Predictor:M2	ML based answer from Ensemble	Confidence on ML answer
0	0	1	1	Low
X	0	1	1	Low
0	0	0	0	High
1	1	0	1	Low
X	1	0	1	Low
1	1	1	1	High

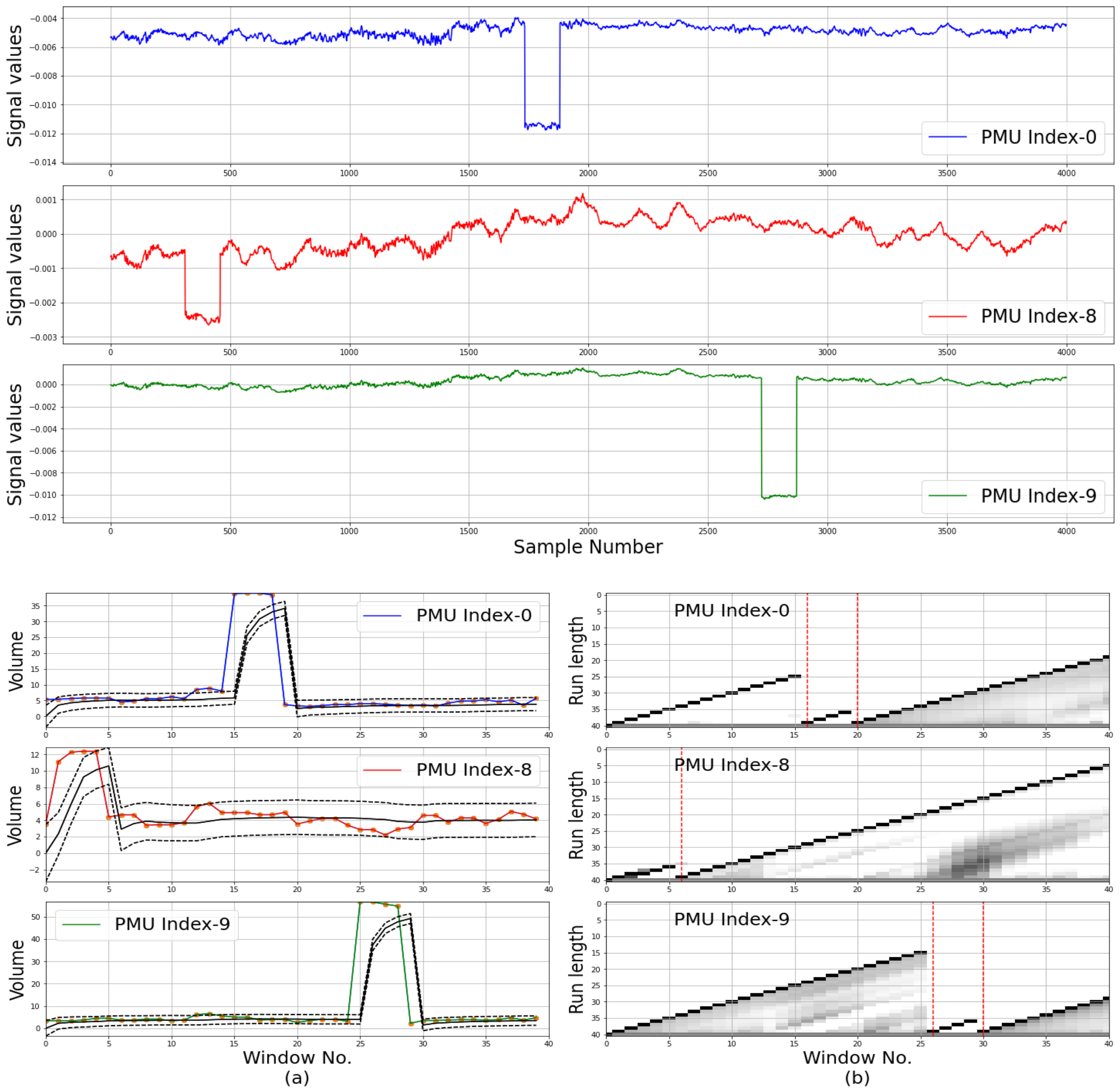
Precision, Recall, F1-Score, Confidence for no anomaly detection, Confidence for anomalies detection

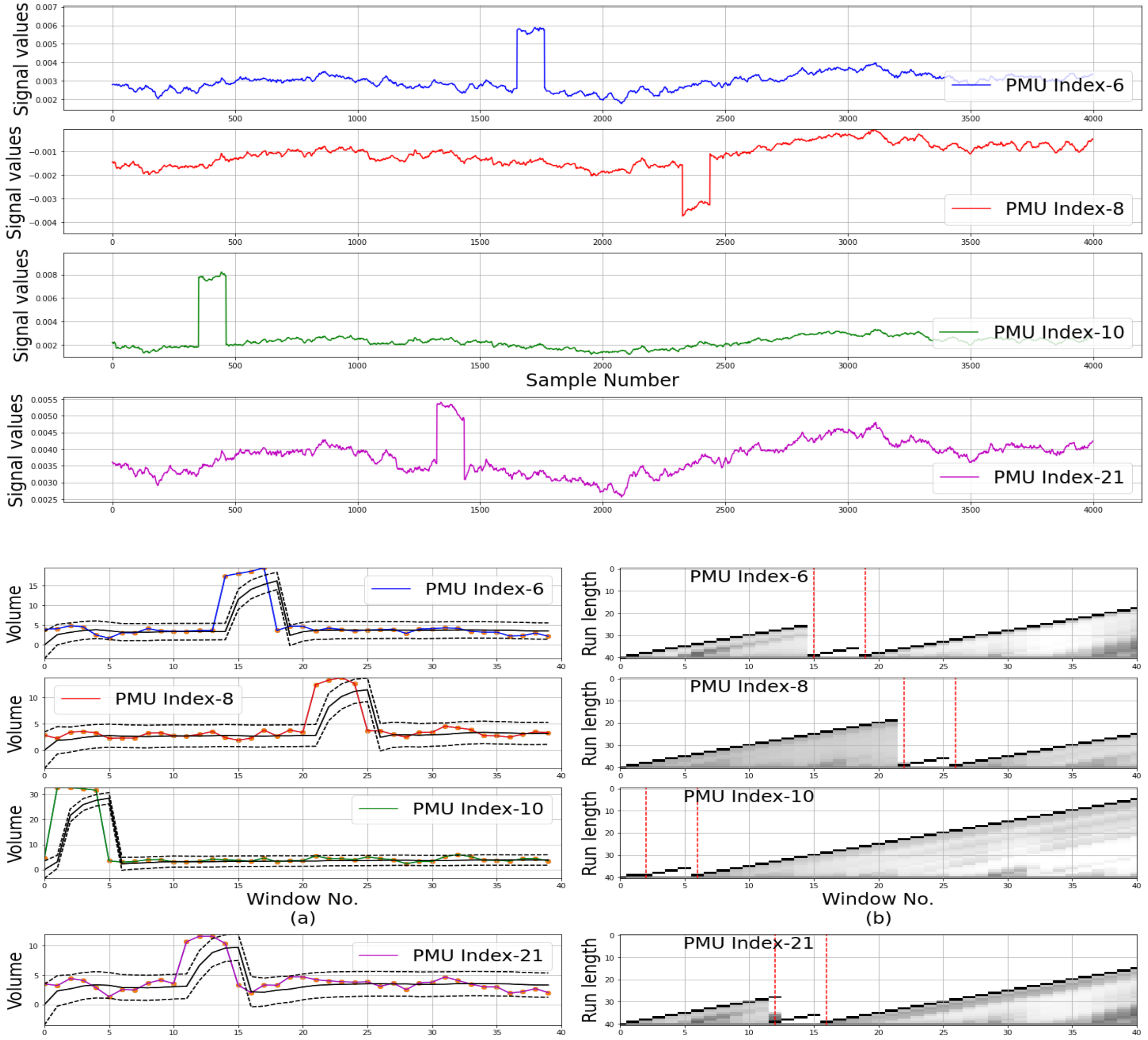
Examples


PMU Index-9
PMU Index-14
PMU Index-19
Signal values
Sample Number
Volume
Run length
Window No.
(a)
(b)


PMU Index-9
PMU Index-14
PMU Index-19
Signal values
Sample Number
Volume
Run length
Window No.
(a)
(b)

Multiple Sample Case





Synthetically introduced patterns representing data integrity problems

Results with Event Detection

Anomaly Distribution ranges	Methods	Precision	Recall	F1-score	Confidence for no anomalies	Confidence for anomalies
1 (X 0.7)	M1	93.04	76.23	83.80	67.29	93.04
1 (X 0.7)	M2	90.63	91.47	91.05	84.11	90.63
1 (X 0.7)	**Ensemble**	**88.84**	**92.13**	**90.45**	**84.54**	**88.84**
2 (X 0.3)	M1	90.90	62.01	73.73	56.25	90.90
2 (X 0.3)	M2	89.18	77.98	83.20	67.43	89.18
2 (X 0.3)	**Ensemble**	**86.80**	**79.53**	**83.00**	**67.72**	**86.80**

Results with Anomaly Detection

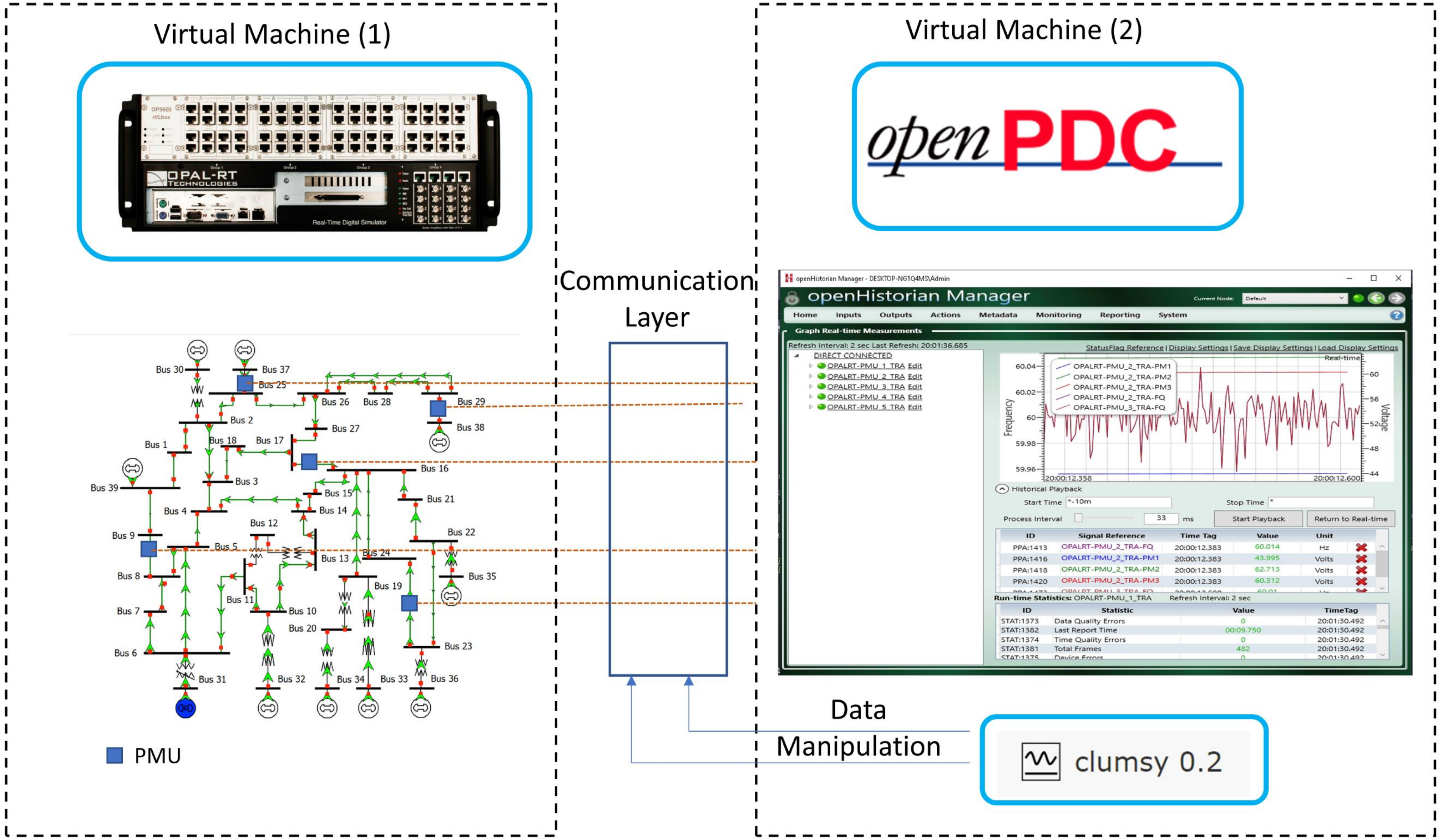
Less than 25% PMU is disturbed

Anomaly Distribution ranges	Methods	Precision	Recall	F1-score	Confidence for no anomalies	Confidence for anomalies
1 (X 0.7)	M1	90.67	64.06	75.08	66.84	90.67
1 (X 0.7)	M2	88.02	85.10	86.54	81.90	88.02
1 (X 0.7)	**Ensemble**	**85.60**	**87.85**	**86.71**	**84.10**	**85.60**
2 (X 0.3)	M1	87.08	46.96	61.02	57.85	87.08
2 (X 0.3)	M2	85.99	68.98	76.55	68.84	85.99
2 (X 0.3)	**Ensemble**	**82.68**	**71.91**	**76.92**	**69.73**	**82.68**

Motivation for SENTIENT Testbed

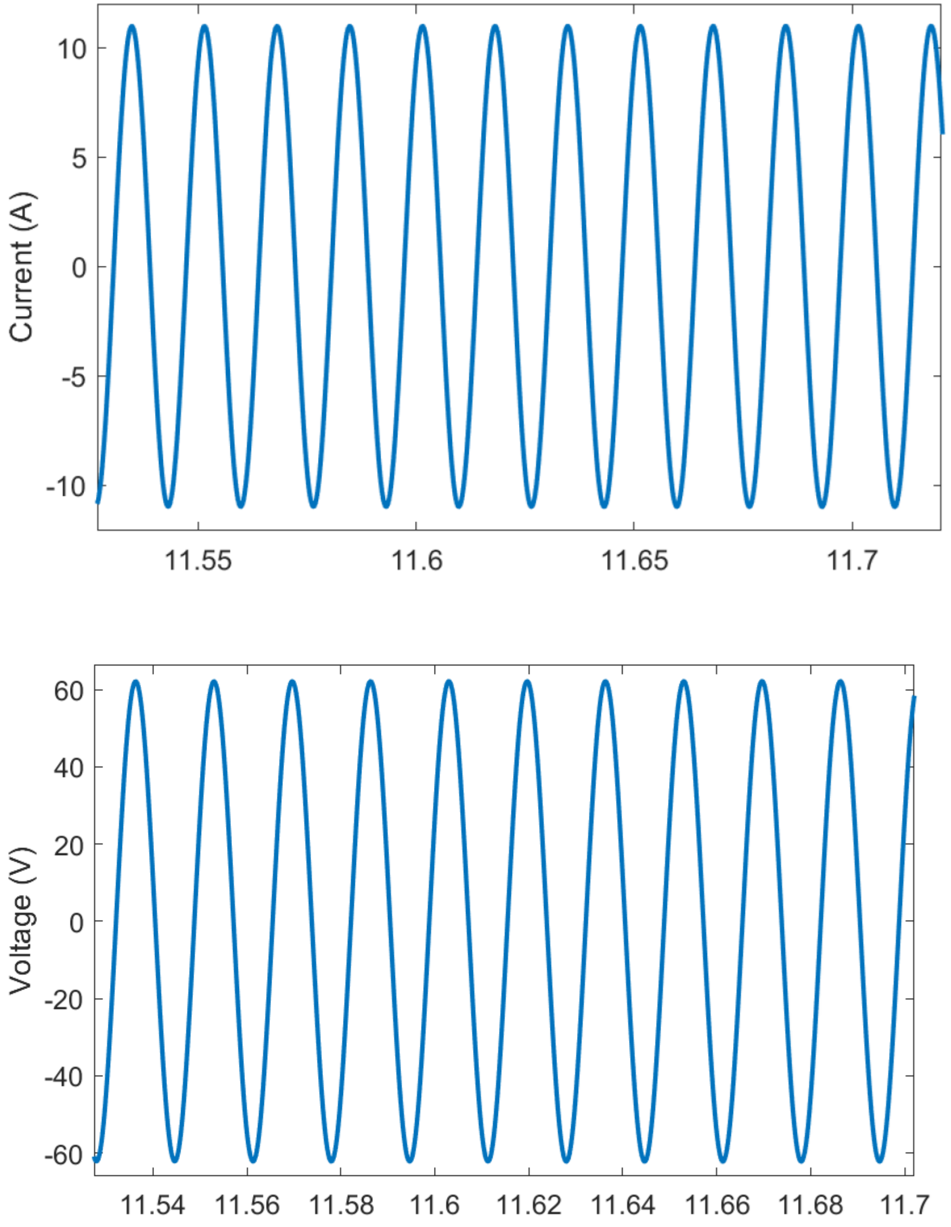
- Multiple Sensor Anomaly pattern generation and analysis
  - Instrument transformer (CT/PT) failure
  - Saturation of CVT and CT
  - Communication noise
  - Sensor noise
  - Data drops
  - Data Lag
  - Data Latency
  - Packet content manipulation

Sentient Testbed

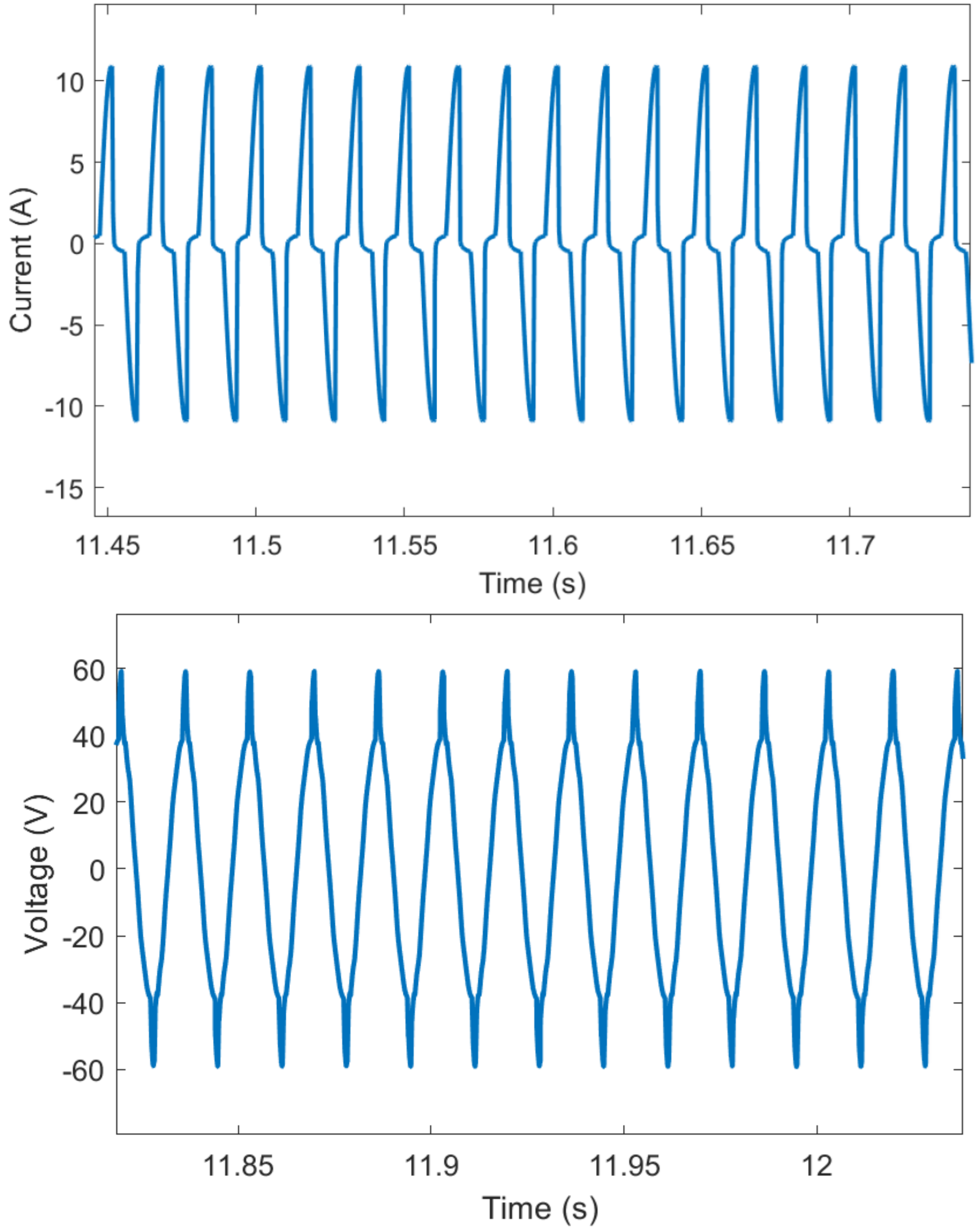


Testbed Saturation Dataset Sample POW – 20kHz

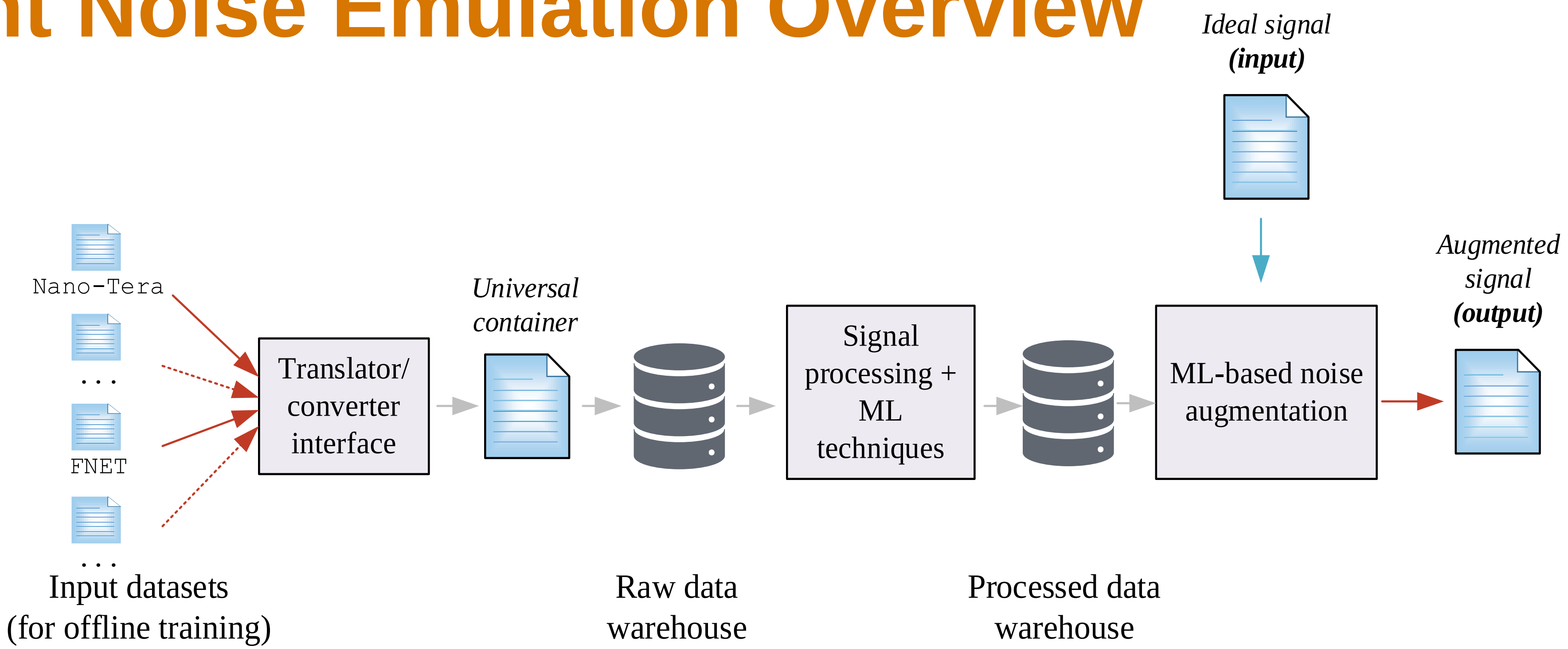
No Saturation



With Saturation



Sentient Noise Emulation Overview



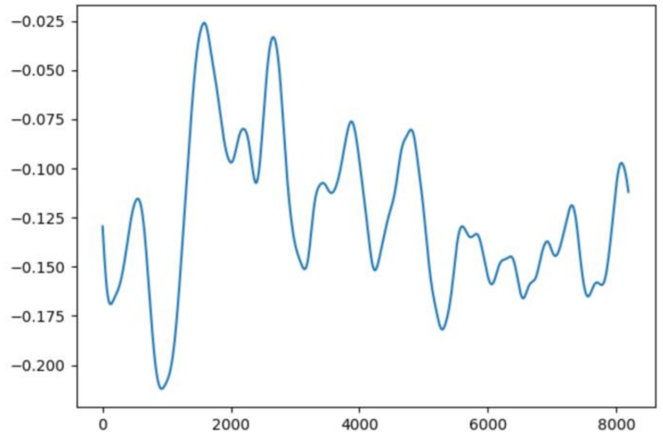
SENTENT noise emulation platform generates synthetic synchrophasor data by learning from actual data. It preforms

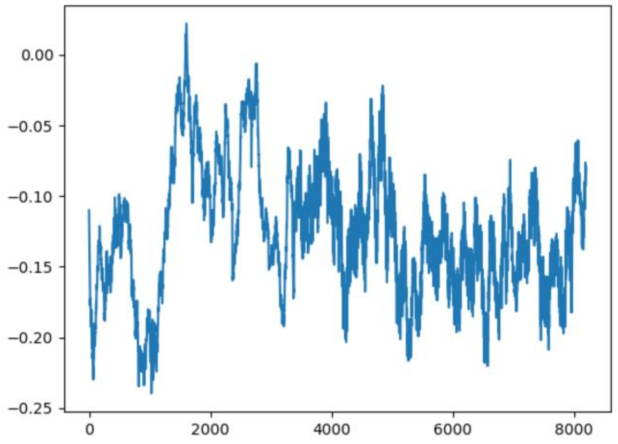
1) Preparing the signals according to similar metadata labels (event logs and data flow channel information)
2) Preprocessing to extract ambient noise components from the measurement signals (extract sensor induced information)
3) Characterizing the types of noise characteristics present in the data base and save the noise characteristics
4) Preparing ML models – GAN to emulate noise characteristics using the pre-cached noise and sensor event data

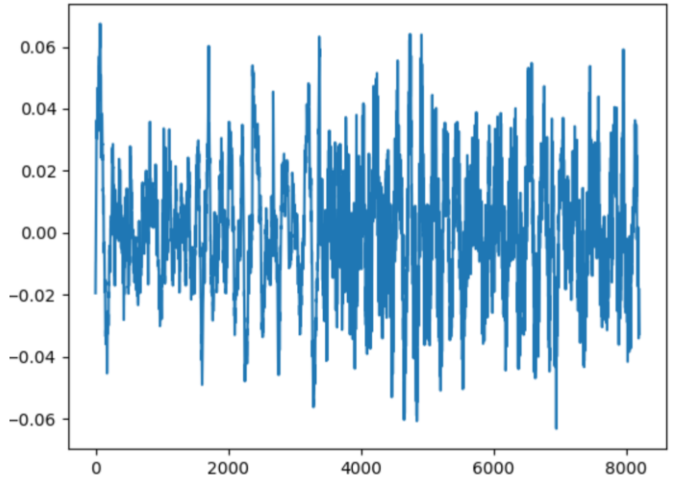
Pacific Northwest
NATIONAL LABORATORY

Noise in Synthetic PMU data

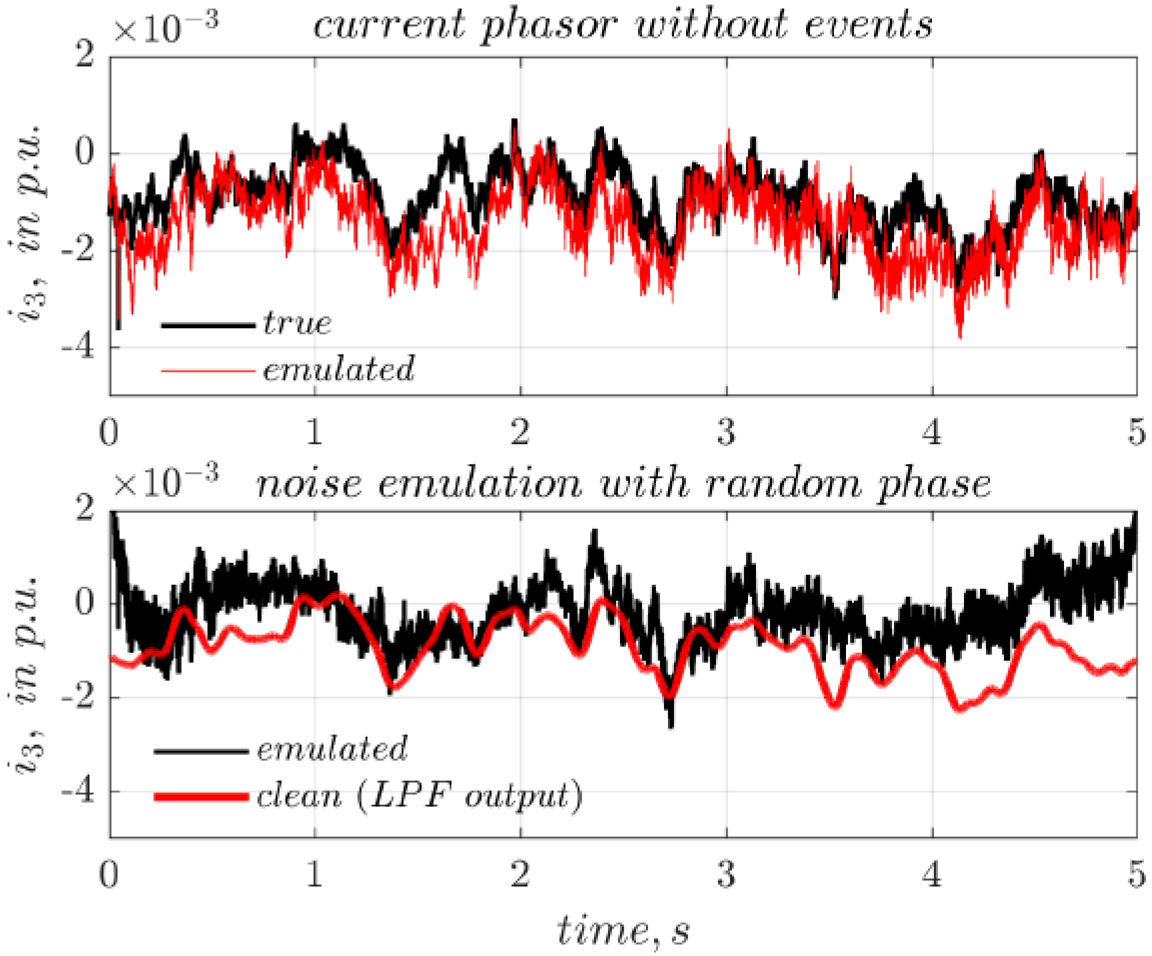
- Simulated PMU data is often ideal and does not contain noise profiles, whereas real PMU data does have noise profiles.
- These noise profiles are often not as simples as white noise. They generally have underlying critical frequencies and harmonics.

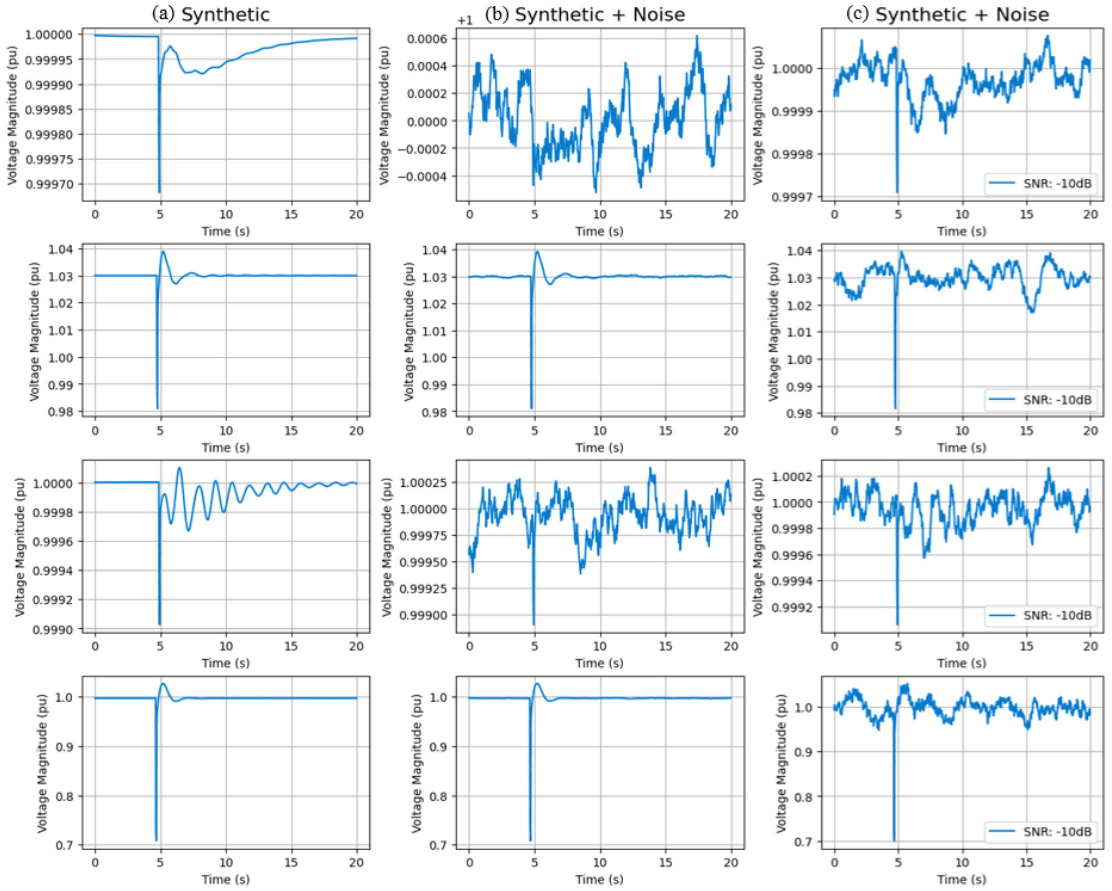






Noise Emulation Platform – Sensor Noise Modeling





Summary

- Data integrity issues for the synchrophasors
- Ensemble based detection method is proposed for identifying any anomalous behavior in the PMU dataset which combines the results of multiple unsupervised binary classifiers approaches through confidence-based aggregation
- A two-stage anomaly detection procedure is proposed.
  - Event Detection
  - Anomaly Detection/Event Characterization
- Provide confidence on proposed predictions
- Proposed SENTIENT testbed to generate these anomalies from different layers of data acquisition
- Future work involves designing trust metrics for synchrophasor event detection process

Pacific Northwest
NATIONAL LABORATORY

Thank you