FOA 1861 FINAL PROJECT BRIEFING
BIG DATA ANALYSIS OF SYNCHROPHASOR DATA

PMU-Based Data Analytics Using Digital Twin and Phasor Analytics Software

Presenter: Philip Hart [1]

Weizhong Yan[1], Tianyi Wang[1], Vijay Kumar[1], Pengyuan Wang[1], Lijun He[1], Kareem Aggour[1], Arun Subramanian[1], Xian Guo[1], Gang Zheng[2], Maddipour Farrokhifard[2]

[1] GE Research [2] GE Digital
philip.hart@ge.com
July 28, 2021

GE Research

NETL NATIONAL ENERGY TECHNOLOGY LABORATORY

Acknowledgement

Acknowledgment: "This material is based upon work supported by the Department of Energy under Award Number DE-OE0000915."

Outline

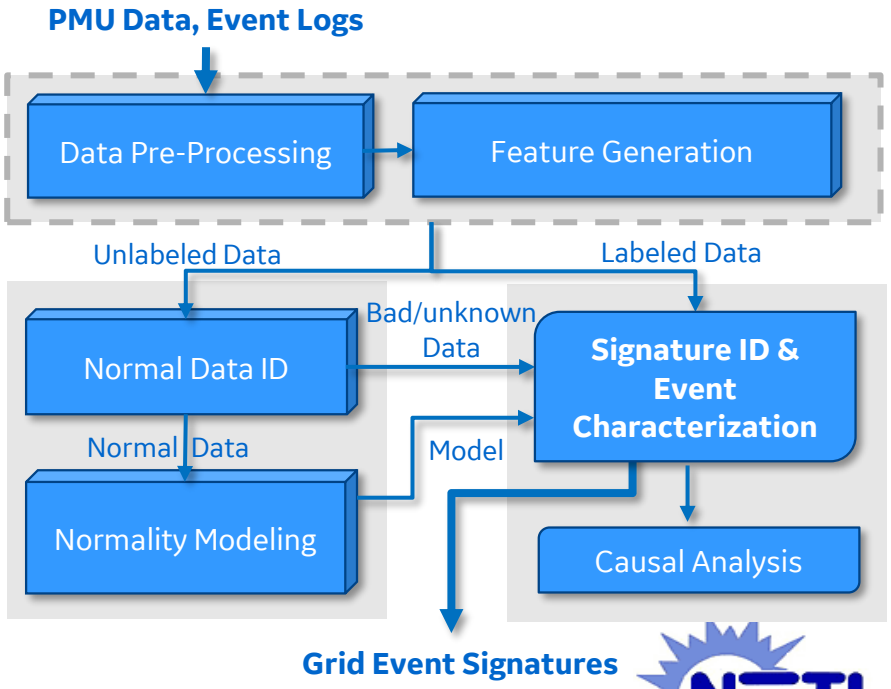
- Project Overview
  - Project Objectives
  - Technical Strategy, Significance & Impact
- Summary of Experimental Results
  - Experimental Results for Project Specific Tasks
  - Assessment of Datasets – Results, Findings, Challenges
  - Validation Results
- Summary of Technical Accomplishments
  - Specific objectives reached
- Value of Work
- Readiness for Commercialization
- Readiness for ML & BD Analytics
- Lessons Learned and Next Steps

Project Overview / Summary

- Team: GE Research & GE Digital
- Objectives:
  1. Leverage existing GE platforms, and off-the-shelf big data & ML tech
  2. Extract insights pertaining to data quality
  3. Identify and validate key grid event signatures
  4. Characterize key grid events
  5. Identify grid events not included in the event log
  6. Investigate causal / correlated factors
  7. Assess the suitability of commercialization of big data / ML technology for the power grid
- Significance & Impact:
  - ***Enhancement of grid reliability, security, and efficiency***

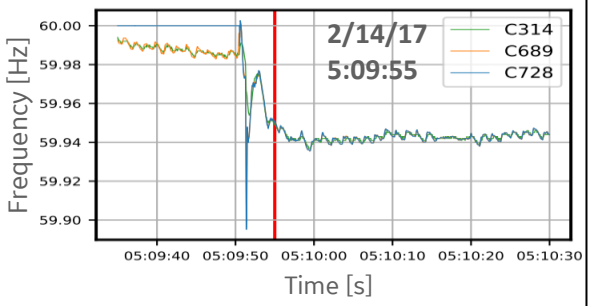
Software & Hardware Platforms:

Strategy for Event Signature I.D.:

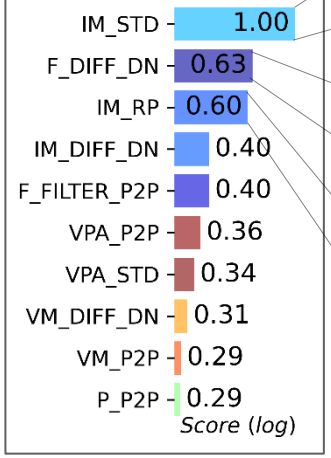


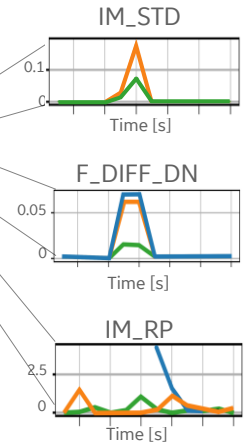
Results I—Extracted Signatures for Key Grid Events

Generator Events (E) [1]



EVENT SIGNATURE





Cause: Any

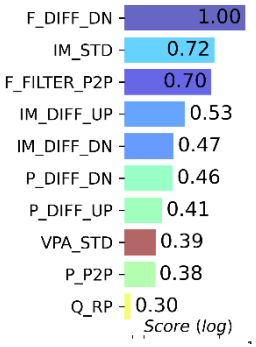
# of Events: 131 (67)

[1] E = Eastern Interconnect, W = Western

Freq. Events (W) [1]

Cause: Any

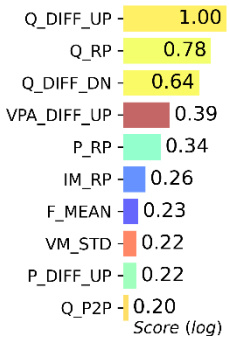
No. of Events: 465 (172)



Oscillations (W)

Cause: Any

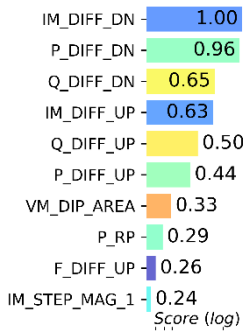
No. of Events: 100 (31)



X-Fmr Events (W)

Cause: Trip

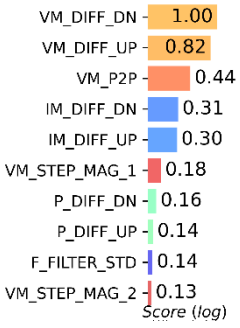
No. of Events: 451 (34)



Line Events (E)

Cause: Equipment

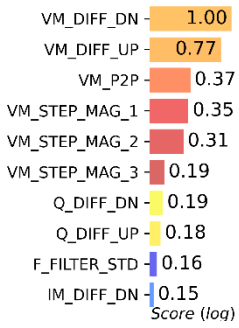
No. of Events: 96 (34)



Line Events (E)

Cause: Fault

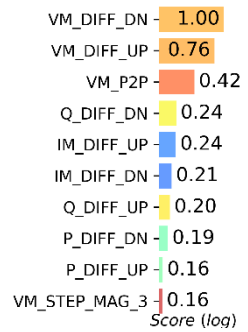
No. of Events: 93 (50)



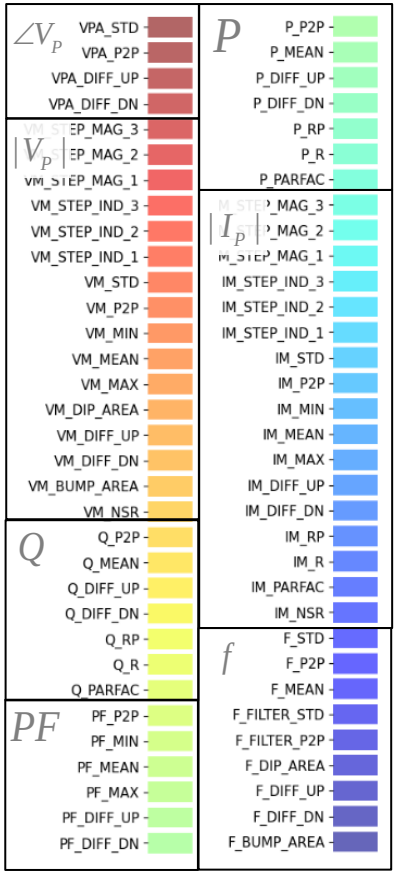
Line Events (E)

Cause: Lightning

No. of Events: 263 (110)



Features (66):

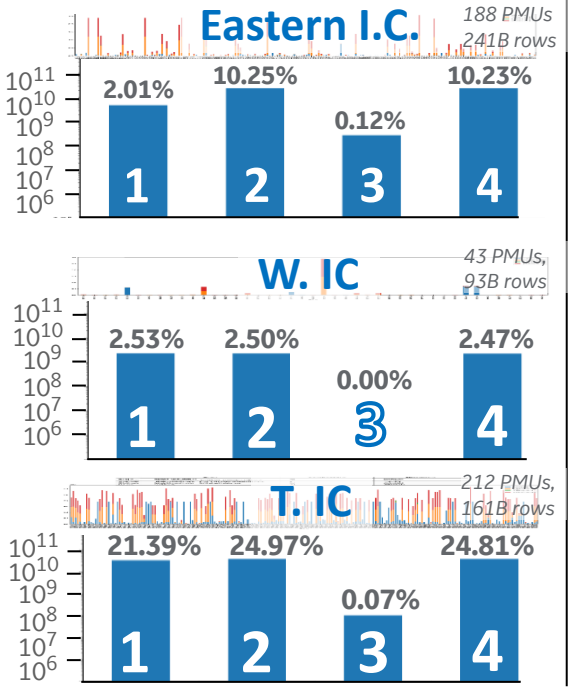


Distinct signatures for key events: shows promise for detection, classification

Results II—Dataset Quality, Event Signature Validation, Event Characterization

DATASET QUALITY:

1	-> Row has 'status' != 0
2	-> Row has 'unreasonable' value
3	-> Row has non-numerical value
4	-> Row has missing value



EVENT SIGNATURE VALIDATION:

- Event signatures used as basis for anomaly detection, binary classifier ensemble
- Line, generator event classifiers built

Eastern I.C. Training Dataset Results:

Line events classified	78/79 (**99%**)
Generator events classified	17/21 (**81%**)

Eastern I.C. Test Dataset Results:

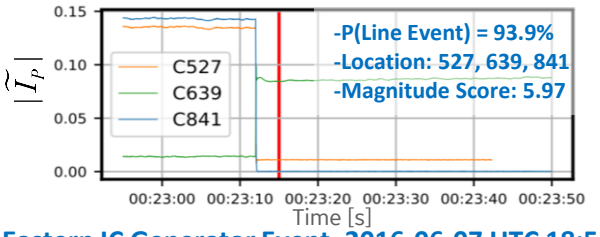
No. of rolling windows evaluated	>3,000,000
Anomalies (Unfiltered)	31,729
Anomalies (Filtered)	9,917
Events in Test event log	597
Detection: true pos. detections	**264 (44%)**
Detection: unlabeled anomalies	**9675 (98%)**
Line events	384
Line events detected	217 (57%)
Generator events	41
Generator events detected	21 (51%)
Line events classified	**196 (90%)**
Generator events classified	**16 (76%)**

DISCOVERY OF NEW EVENTS:

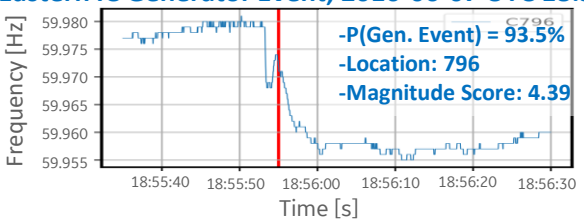
- **9,675 anomalies** discovered in Eastern interconnect Test Dataset (2016 to 2017) that were not listed in the event log

Example Events Not Found in Event Log:

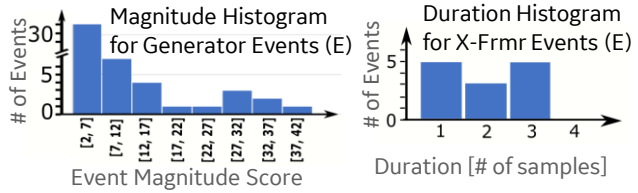
Eastern IC Line Event, 2016-02-23 UTC 00:23:15



Eastern IC Generator Event, 2016-06-07 UTC 18:55:55



EVENT CHARACTERIZATION:



High variation in data quality between ICs; classifier accuracy: 76% to 90% in Test Dataset (81% to 99% using only labeled events in Training Dataset, split 70/30).

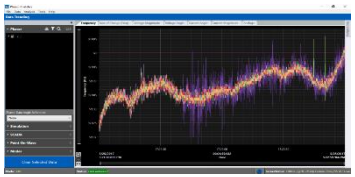
Technical Accomplishments I

1 Leveraged existing GE platforms and off-the-shelf big data & ML tools

- HDFS, Apache Hive, Spark with custom Python-based APIs
- >0.639 Trillion rows of data ingested
- ML and data analytics from GE Digital Twin ecosystem
- GE WAMS software, Python for time-series visualization

Big Data Toolkit:

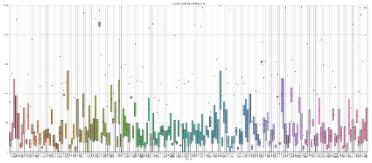
GE WAMS Software:



2 Comprehensive data quality analysis, global statistical analysis for all three I.C.s

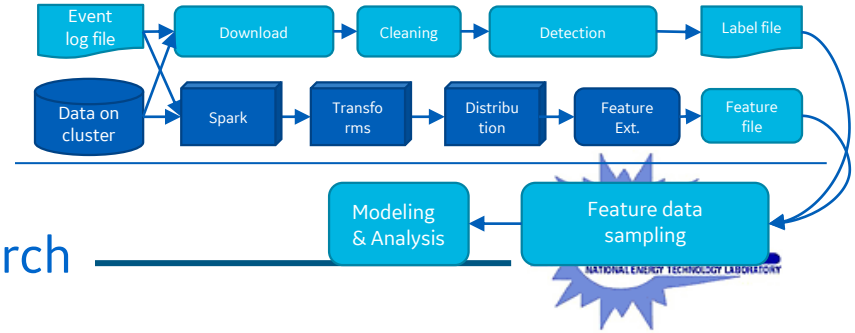
- 4 data quality queries, quartiles, min/max/average for vp_m, ip_m, f, and df across all three interconnects
- Histogram count of 'status' variable values
- Several difficult-to-detect data quality issues identified

Ex.: 6-point statistical analysis of ip_m for all PMUs in East. IC dataset:



3 Identified key grid event signatures for Eastern and Western Interconnects

- Developed requisite infrastructure/pipelines:
  - 66 time-series feature functions (7 feature 'batches')
  - Parallel feature generation pipeline with fast Pandas <-> Spark conversion. *Ex.*: feature batch for complete Western IC: 89MM feature values (23.5 GB) in 35 minutes versus 2.5 days
  - Developed normality modeling pipeline
  - Developed signature identification pipeline, processing thousands of labeled events
- Refined labeled event times for complete Eastern and Western interconnect datasets
- Over 15 signatures identified for key grid events; developed binary classifier ensemble, including decision fusion; applied to Test Dataset to validate two signatures (line event and generator event)
- Preliminary signature robustness studies (e.g. Eastern vs. Western IC) for frequency, oscillation events

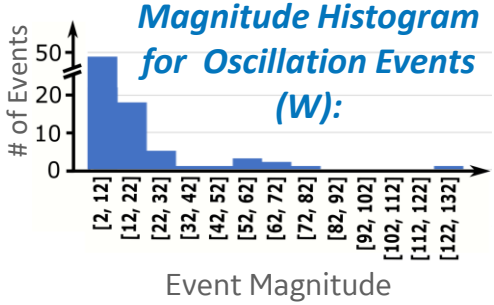


Technical Accomplishments II

4 Characterization of key grid events

- Used event signatures to characterize the ***(1) magnitude, (2) location, and (3) duration*** of 1,000s of labeled & newly-identified grid events in East or West I.C.
- Ex.: Eastern I.C.:
  - Generator
  - Line (equipment)
  - Line (fault)
  - Line (lightning)
  - Line (not lightning)
  - Oscillation

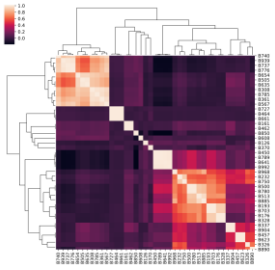
Magnitude Histogram for Oscillation Events (W):



5 Correlation & causal analysis

- Investigated correlations & causal relations between transformer failure and SNR-based features
- Investigated seasonal trends in modal characteristics (frequency, damping), relationship to active power flow (100 oscillation events)

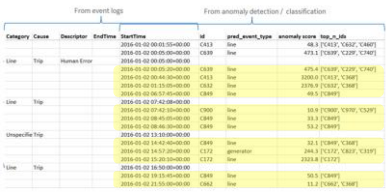
Sample Cluster Map of PMU Correlations:



6 Identified new events not included in the event log

New Events:

- Used binary classifier ensemble to label and characterize unlabeled anomalies (line, generator)

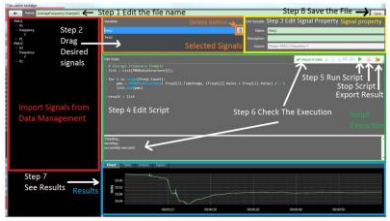


- 10s of thousands of unlabeled anomalies identified throughout entirety of Eastern IC Training and Test Datasets. Three million windows evaluated in Eastern Interconnect Test Dataset.

7 Completed commercialization report

SW Development:

- 23-page report; effort led by GE Digital
- Reviews existing software capabilities, software development considerations



- Proposes path to commercialization for event signatures, features, and big data technology developed on FOA 1861 project

Value of Work I

Q1: What additional benefit can this work bring to utilities?

1. Real-time grid event *detection*, *classification*, and *characterization* from streaming PMU data using event signatures derived from rigorous, state-of-the-art big data analytics & machine learning
   - Immediate end-use applications: situational awareness, online model calibration, WAMPAC
   - Possible future application: equipment health monitoring
2. Commercialization of a customized big data platform and machine learning tools for grid event signature identification & other applications
   - Massively parallel feature generation
   - Universal 'feature function' database and recommended hyperparameters (e.g., window length, stride)
   - Event signature database
   - Use of anonymized PMU datasets & event logs (geographically paired)

Value of Work II

Q2: What can utilities gain by sharing data with each other?

- Rarity of 'high-quality', labeled events makes it valuable to share data
- May allow for validation of event labeling

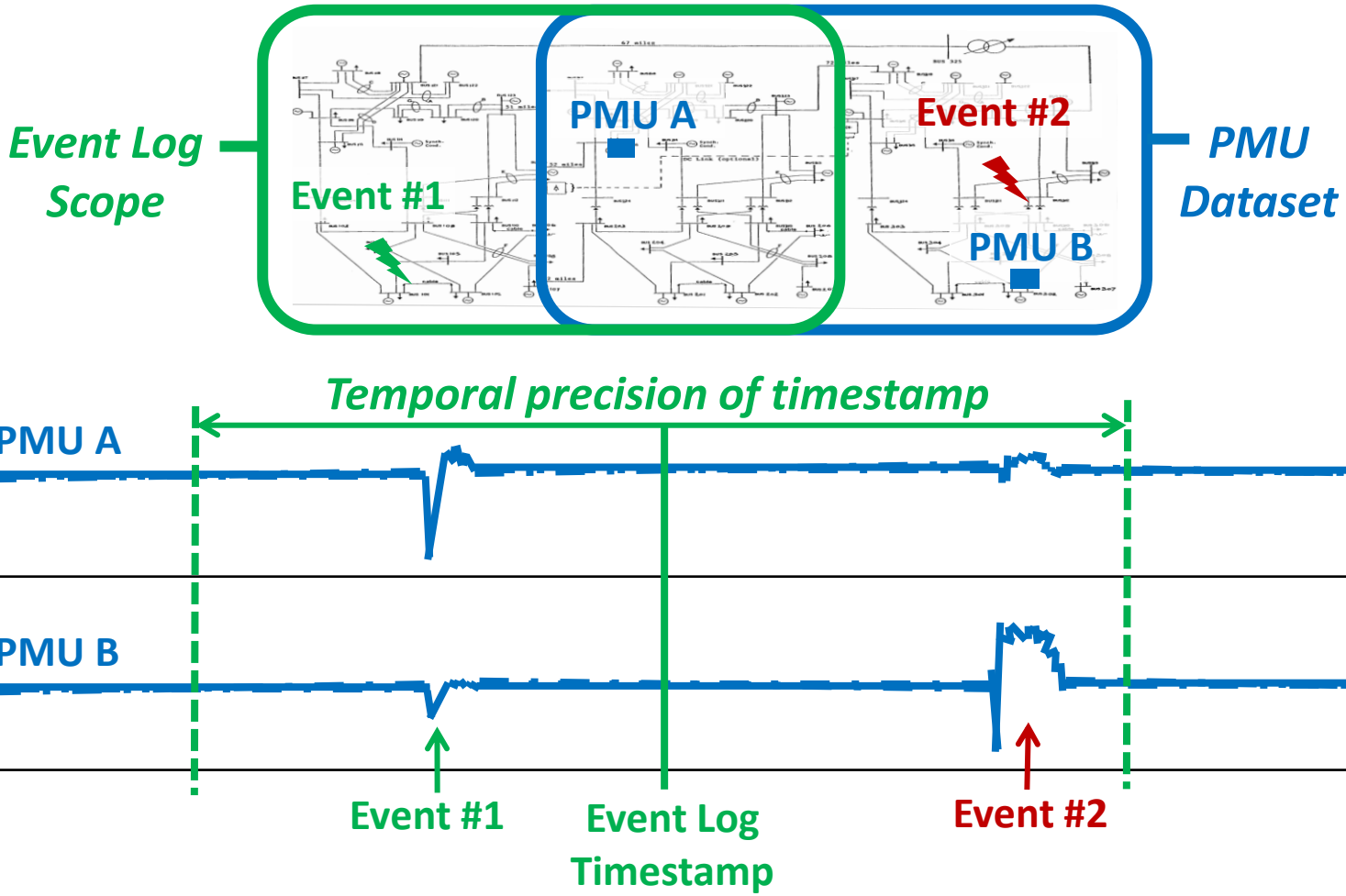
Q3: Is it worthwhile for utilities to share anonymized data?

- Despite limitations of anonymized data, it permits some progress on signature ID, event detection & classification
- Permits evaluation of data quality, feature extraction techniques, etc.

Q4: What are the limitations of working with anonymized data?

- Anonymization (no spatial information in label) compounds the issues introduced by (i) temporal imprecision of label and (ii) large number of unlabeled events

Value of Work III: Limitations Associated with Anonymized Data—Example



Anonymization + temp. imprecision + unlabeled events -> difficulty in determining which anomaly (Event #1 or #2) should be associated with the event log entry

Readiness for Commercialization I

Q5: What would be the next steps for making the results from the projects available to use by utilities?

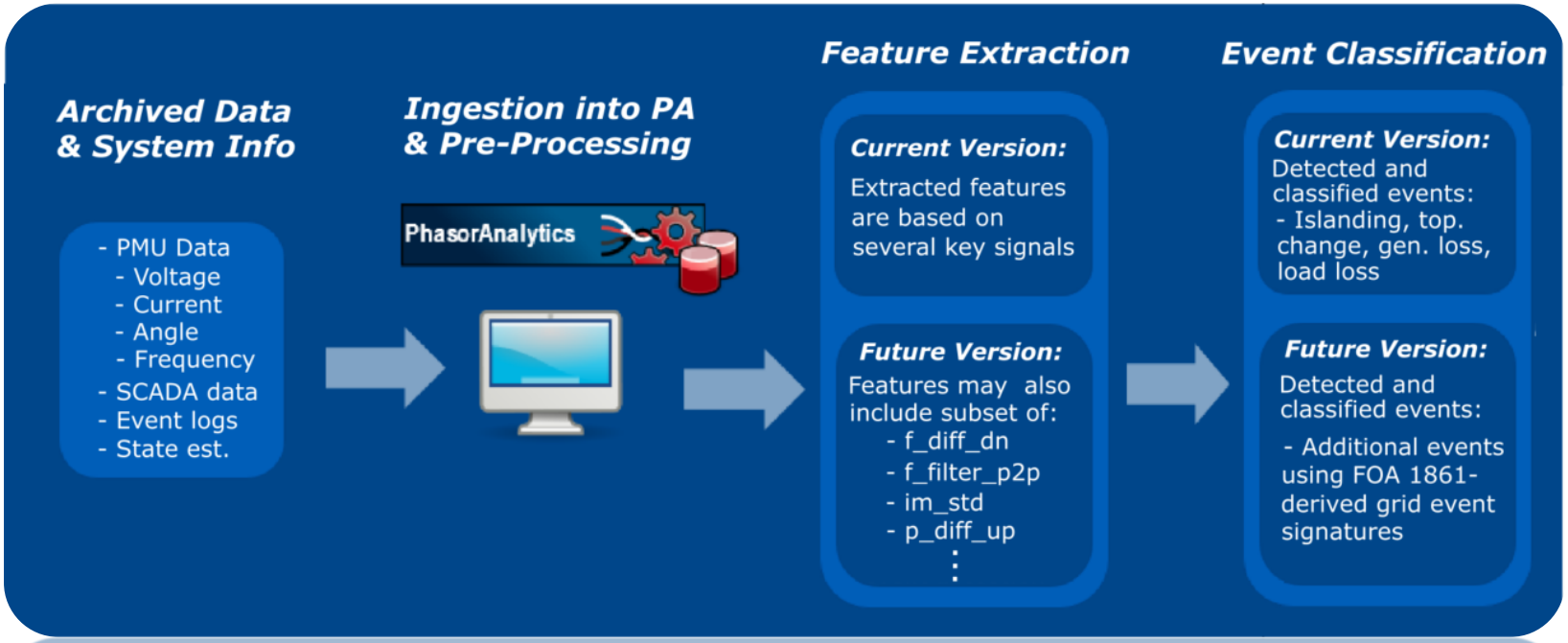
- Some major components developed on this project can be commercialized and universally applied to either anonymized or deanonymized datasets

Component	Technology Readiness Level (TRL)
Feature generation pipeline	TRL 5
Anomaly detection pipeline	TRL 5
Big data platform (APIs, performance optimizations)	TRL 5
Full process (data preprocessing to sign. ID)	TRL 3

- However, full end-to-end process is still at lower TRL:
  - Proof of concept complete. Presently adapted to meet specific data quality challenges and constraints of this project. If customer provided additional information (less anonymization), there would be strong motivation to adapt / adjust our technical strategy to achieve optimal performance (e.g. modify specific steps in normal data identification / normality modeling)

Readiness for Commercialization II

Q6: How do you anticipate transitioning your research to tools that are available to utilities? In the near-term?



Near term opportunity to integrate developed tools into GE WAMS software

Being Ready for ML & BD Analytics I

Q7: Do off-the-shelf machine learning models achieve good performance for PMU data analytics?

- Not without significant customization. We had to customize data cleaning, relabeling (spatial, temporal localization), normal data identification procedures in order to achieve satisfactory performance

Q8: What are key challenges for AI/ML in the context of power system data?

Data Challenge	Description/comments	Possible impact on successful application of ML to PMU data (1 to 10 with 10 being most detrimental)	Mitigation strategy employed here	Extent mitigated (1-10, with 10 being fully mitigated)	Suggestions to Dataset Providers, Aggregators & Assemblers	Suggestions for Future R&D Work & Priority (1-10 w/ 10 = highest priority)
1. Basic dataset quality issues	1. Status value !=0 2. Missing data 3. Non-numerical 4. Unreasonable values	Impact: 2. In the worst case scenario (when imputation is not possible), reduces usable dataset and usable events.	Basic data cleansing	Extent mitigated: 9	(1) Regular inspection of PMU data to verify functionality, adding flag if mis-operation suspected; (2): If service or maintenance is performed on PDC, archival system, or PMU and/or its related sensors, record day and time of maintenance in event log.	Priority: 4; Can revisit and advanced data imputation
2. 'Advanced' dataset quality issues	Difficult-to-detect data quality issues. Examples: -1000x step-changes in values -Unexplained slow fluctuations	Impact: 7. Pollution of 'normal' data used to develop normality model.	Visual inspection, additional filtering rules	Extent mitigated: 7-8 (Eastern); 5-6 (Western)		Priority: 8; Additional unsupervised learning for normal data ID; historical big data visualization.

Key challenge: difficult-to-detect data quality issues

Being Ready for ML & BD Analytics II

Q8: What are key challenges for AI/ML in the context of power system data?

Data Challenge	Description / Comments	Possible impact on successful application of ML to PMU data (1 to 10 with 10 being most detrimental)	Mitigation strategy employed here	Extent mitigated (1 to 10, with 10 being fully mitigated)	Suggestions to Dataset Providers, Aggregators & Assemblers	Suggestions for Future R&D Work & Priority (1-10 w/ 10 = highest priority)
3b. Anonymization of dataset: lack of spatial information in event label	Dataset and event log spans an entire interconnect. Spatial information (physical location of event) is scrubbed from event log, and location of PMUs is hidden, due to security concerns.	Impact: 8. Significant impact on signature identification. Impact may be significantly lowered if other challenges (temporal imprecision and large number of unlabeled events) are addressed.	Used anomaly detection to locate the PMUs most relevant to labeled event.	Estimate for extent mitigated: 5—degree of success remains uncertain.	Almost any supplemental spatial information may help. For example: (1) Identify the PMU geographically (electrically) closest to the labeled event; and/or (2) Increased granularity: break ICs into several regions, associating sections of the anonymized raw data with sections of the anonymized event log; and/or (3) Discard event log label if event is of sufficient electrical distance from any PMU	Priority: 9--We recommend continued application of technical strategy given new spatial information (if any of the suggestions (1)-(3) are found acceptable to data providers), or given additional temporal precision information, and/or additional event labels. If none of the aforementioned information can be easily granted, new research may be necessary to better quantify the impact of dataset anonymization and explore potential mitigation techniques.
3b. Anonymization of dataset: network topology information concealed	No information is provided regarding network structure or characteristics, due to security concerns.	Impact: 4. Not expected to be strictly necessary for event signature ID, but may be more important for other objectives (e.g. equipment health monitoring)				

Key challenge: anonymization of dataset

Being Ready for ML & BD Analytics III

Q8: What are key challenges for AI/ML in the context of power system data?

Data Challenge	Description/comments	Possible impact on successful application of ML to PMU data (1 to 10 with 10 being most detrimental)	Mitigation strategy employed here	Extent mitigated (1-10, with 10 being fully mitigated)	Suggestions to Dataset Providers, Aggregators & Assemblers	Suggestions for Future R&D Work & Priority (1-10 w/ 10 = highest priority)
4. Limited temporal resolution of label	Timestamps in the event log may be associated with the timestamp of a SCADA event alarm in control room, resolution of ~1 min or better; may be associated with relay records.	Impact: 8. Would be lower, but problems is compounded by the large number of unlabeled events. Event is ignored if anomaly is not sufficiently close to timestamp: contributes to reduced number of 'high quality' events.	Used anomaly detection to refine the timestamp of the event	Extent mitigated: 7	Provide additional information confirming the temporal precision of timestamps (seconds vs minutes vs hours)	Priority: 6— We recommend continued application of technical strategy with possible adjustments to labeled event window, given additional temporal precision information.
5. Large number of unlabeled events	Eastern IC test dataset: 10s of thousands of events with severity equal to or greater than severity of labeled events (severity determined by mag. of features in event signature)	Impact: 9. -Compounds the difficulties introduced by challenges 3 & 4. -Pollutes the normality model by interfering with normal data identification	Use anomaly detection only within a limited window around labeled events.	Extent mitigated: 7	Dataset providers may be able to help label the 'normal data'; i.e., they can perhaps identify regions of the dataset estimated to be event-free, to the best of their knowledge	Priority: 8—Adapt technical strategy to include additional unsupervised learning for normal data identification.

Key challenges: uncertain temporal precision in labels and large # of unlabeled events

Lessons Learned and Next Steps I

Q9: How important is it to synergistically combine machine learning models with power systems domain knowledge?

- Very important. Domain knowledge can impact every step of the signature identification strategy. Deep learning can address modeling and, to some extent, feature engineering; domain expertise is still critical to address data quality, pre-processing, etc.

Q10: What recommendations do you have for the FOA 1861 dataset moving forward?

FOA 1861 dataset is an excellent resource—work should continue with this dataset, but additional information from data providers may unlock more value:

1. Explore possibilities for adding limited spatial information to event log entries, without sacrificing anonymity (see Key Challenges table).
2. Explore possibilities for providing more comprehensive information regarding temporal precision of labels in event log. Some general info was provided during Datasets Webinar; but ideally, we would have a temporal precision estimate for each entry in the event log.

Lessons Learned and Next Steps II

Q11: How to improve the interpretability of data-driven models for power system event detection and classification?

- Choose a methodology that lends itself to interpretability, such as feature ranking approach employed here. Alternative approach: explainable AI

Q12: What would be next steps for research of related technologies?

To summarize the table of challenges and suggested high-priority research:

1. Investigate additional unsupervised learning techniques to better-filter hard-to-detect data quality issues and unlabeled events prior to identifying normal data
2. Big data visualization tools for large volumes of historical PMU data to further streamline normal data identification (volume is issue, not speed)
3. Continued application of technical strategy for sig. ID, event detection / classification, equipment health monitoring, prognostics, causal analysis, etc.

Publications, Presentations

- V. S. Kumar, T. Wang, K. S. Aggour, P. Wang, P. J. Hart and W. Yan, "Big Data Analysis of Massive PMU Datasets: A Data Platform Perspective," *2021 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)*, 2021, pp. 1-5.
- Currently drafting two additional publications:
  1. "Grid Event Signature Identification Through Large, Real-World PMU Datasets" (manuscript to be submitted in consideration for publication in IEEE Transactions)
  2. "Challenges with Large-Scale, Real-World PMU Datasets and Mitigation Strategies" (manuscript to be submitted to upcoming IEEE conference)
- Seminars / panel presentations (slides can be made available upon request):
  1. "Big Data Analysis of Synchrophasor Datasets" seminar at Clarkson University ECE Department, April 9th, 2021.
  2. "Experiences in applications of AI and ML to analysis of synchrophasor data," at *IEEE International Conference on Smart Grid Synchronized Measurements and Analytics*, May 25-27th, 2021.
  3. "Big Data Analytics Panel Session", at *NASPI Working Session* (no slides available)
  4. "Big Data Analysis of Synchrophasor Data: Experience from the U.S." at *2021 PES GM*, July 27th

Thank You