Application of Big Data Analytics and Machine Learning to Large-Scale Synchrophasor Datasets: Evaluation of Dataset ‘Machine Learning-Readiness’

PHILIP HART, LIJUN HE, TIANYI WANG, VIJAY S. KUMAR, KAREEM AGGOUR (Member, IEEE), ARUN SUBRAMANIAN, AND WEIZHONG YAN (Senior Member, IEEE)

GE Research, Niskayuna, NY 12309 USA
CORRESPONDING AUTHOR: P. HART (philjhart@gmail.com)

This work was supported by the Department of Energy under Award DE-OE0000915.

ABSTRACT
This manuscript presents a data quality analysis and holistic ‘machine learning-readiness’ evaluation of a representative set of large-scale, real-world phasor measurement unit (PMU) datasets provided under the United States Department of Energy-funded FOA 1861 research program. A major focus of this study is to understand the present-day suitability of large-scale, real-world synchrophasor datasets for application of commercially-available, off-the-shelf big data and supervised or semi-supervised machine learning (ML) tools and catalogue any major obstacles to their application. To this end, dataset quality is methodically examined through an interconnect-wide quantifications of basic bad data occurrences, a summary of several harder-to-detect data quality issues that can jeopardize successful application of machine learning, and an evaluation of the adequacy of event log labeling for supervised training of models used for online event classification. A global ‘six-point’ statistical analyses of several key dataset variables is demonstrated as a means by which to identify additional hard-to-detect data quality issues, also providing an example successful application of big data technology to extract insights regarding reasonable operational bounds of the US power system. Obstacles for application of commercial ML technologies are summarized, with a particular focus on supervised and semi-supervised ML. Lessons-learned are provided regarding challenges associated with present-day event labeling practices, large spatial scope of the dataset, and dataset anonymization. Finally, insight into efficacy of employed mitigation strategies are discussed, and recommendations for future work are made.

INDEX TERMS
Phasor measurement units, synchrophasor datasets, statistical analysis, data quality analysis, label quality, nonymized datasets, supervised machine learning, big data, wide area monitoring system.

I. INTRODUCTION
The power system is accelerating towards a future that includes increased amounts of intermittent renewable generation, fast-acting inverter-based resources, stochastic and increasingly bidirectional power flows, growing loads due to electrified transportation, and heightened threat of natural disaster or cyberattack. These trends are driving a need for improved situational awareness, equipment health monitoring, and rapid, automated control action in response to contingencies. To this end, data-driven, machine learning (ML)-based technologies, leveraging online measurements and historical datasets derived from sensor networks such as synchrophasor-based wide-area monitoring systems (WAMS), are destined to play an ever-greater role in the reliability and resiliency of the future grid. However, despite the proliferation of phasor measurement units (PMUs) within the power system over the past several decades, and robust body of academic literature regarding proposed ML applications in power systems (often validated using synthetic datasets), actual industry application of ML
and big data technologies to real-world PMU datasets has been relatively slow [1], [2]. Access to real datasets has been cited as a major barrier to power systems adoption of ML and big data. There are cyber-physical security concerns that often discourage public release of power systems datasets [3]. Even if access to datasets is provided, the power systems data is known to have low value density as events of interest are relatively rare, and even large-scale datasets may not provide sufficient samples to allow for adequate model training. Anonymization of datasets can potentially help reduce security concerns and permit the pooling of datasets, increasing the number of relevant training samples. However, the impediments to supervised or semi-supervised machine learning posed by PMU dataset anonymization have not yet been sufficiently addressed in the literature. While the quantity of labeled power systems events of interest is frequently cited as a bottleneck, other aspects of label quality (such as temporal precision) have not been sufficiently studied, particularly in the context of anonymized datasets.

Besides the aforementioned issue of low ‘value density’, other barriers to real-world application of ML and big data for PMU datasets may include the remaining four of the ‘five V’s’ [4]: volume (petabytes of data), variety (e.g., multiple modalities of PMU data file format or varieties of compression tools used by data historians and archival tools), velocity (30 to 60 Hz sampling), and veracity (data quality issues). Another power systems-specific bottleneck suggested is the poor robustness of applied ML solutions due to lack of power systems domain expertise in the training process. However, there do not appear to be many case study reports in the literature that offer specific details as to how some of the barriers, such as the variety of file format modalities, typically manifest in the context of PMU data. Authors in [3] suggest that the labor involved in dataset assembly and preparation is a major barrier to power system big data analytics, pointing to survey results that claim data scientists routinely attribute up to 80% of their labor to these tasks.

As claimed in [5], sluggish industrial adoption of commercial off-the-shelf ML and big data tools for PMU datasets may be partially attributable to the ‘veracity’ issue, i.e. the prevalence of significant data quality issues within these datasets. Synchrophasor data quality issues have been extensively studied, e.g. in [6], [7], and [8]. In [7], listed synchrophasor data quality challenges include incompleteness (PMU/PDC damage, communication error), measurement inaccuracy (e.g., due to calibration issues), availability (device failure, network latency, possibly due to routing traffic). While [8] provides a useful and comprehensive taxonomy of data quality issues that may affect synchrophasor datasets, the relative prevalence of these individual data quality issues within example real-world datasets is not provided. Likely due to limited access to data, few authors appear to provide methodical investigations into the data quality issues that manifest within real-world, large-scale PMU datasets. Fewer authors still have reported on the minimum level of data quality required to achieve acceptable performance of ML-based technologies for various applications within the WAMS domain, such as event detection or equipment health monitoring. Authors in [8] recommends the benchmarking of applications with datasets containing data quality, but a comprehensive ‘impact benchmarking’ does not appear to be available yet in the literature for ML applications related to PMU datasets.

Reports of ML technological innovations that include real-world PMU dataset validation, such as in [9], [10], and [11], necessarily have to address data quality issues within the respective datasets they have used for validation. However, due to the limited size and spatial scope of the dataset, data quality issues addressed in [9], [10], and [11] are not necessarily a representative survey of challenges observed across the broader industry.

In comparison, this work offers a more detailed quantification of the occurrence of data quality issues in several large-scale sets of synchrophasor data, which have been drawn from a multitude of data providers spanning three major interconnects. This work aims to transcend the limitations of a typical case study by providing a perspective on data quality challenges within a representative collection of PMU datasets from a geographically diverse group of anonymous data providers located across the continental United States. Using the big data platform described in [12], a successful application of a big data-driven statistical analysis is described which allows for more convenient discovery of additional hard-to-detect data quality issues that may present additional barriers to machine learning. This work also extends previous PMU data quality investigations found in the literature, such as in [13], by also evaluating the ‘machine learning-readiness’ of these representative datasets: i.e., identifying and cataloguing any major characteristics of the datasets and associated event logs that may work together to hinder convenient application of off-the-shelf ML tools. In particular, lessons-learned are provided regarding the impact of dataset anonymization on model training. Finally, a ranking of the relative impact of each of the identified challenges is provided, along with an evaluation of the efficacy of the techniques chosen to mitigate the identified challenges, and provides suggestions to improve the ML-readiness of datasets to be collected or assembled in the future.

In Section II, a high-level overview of the datasets is presented and a summary is provided regarding the major pre-processing steps that took place during the initial assembly of the datasets. Section III provides a quantitative analysis of bad data discovered within each of the three interconnect-wide datasets that could be mitigated during preliminary screening efforts, where-as Section IV provides a summary of harder-to-detect data quality challenges that eluded this preliminary screening and were first identified during the application of an industry-validated semi-supervised machine learning strategy [14]. Section V documents example six-point statistical analysis results from a successful application of the big data analytics platform described in [12].
Major obstacles for application of an industry-validated, semi-supervised machine learning strategy for event signature identification are summarized in Section VI, with special consideration paid to the challenges associated with present-day event labeling practices, large spatial scope of the dataset, and dataset anonymization. Finally, mitigation strategies, lessons learned, and recommendations for future work are discussed in Section VI, and conclusions are provided in Section VII.

II. DATASET DESCRIPTION AND PREPROCESSING STEPS PERFORMED BY THE DOE

The United States Department of Energy (DOE) Funding Opportunity Announcement (FOA) 1861 [1] program enlisted the support of data providers nationwide to assemble and distribute three large-scale, anonymized synchrophasor datasets—referred to herein as the ‘FOA 1861 datasets’—collected from 443 PMUs distributed across the three major North American interconnects: the Texas Interconnect (Interconnect A), the Western Interconnect (Interconnect B) and Eastern Interconnect (Interconnect C). Each interconnect-wide dataset consists of PMU data provided by multiple active transmission operators (TOs), and was assembled into a single dataset and preprocessed by the Department of Energy (DOE) staff prior to distributing the dataset to the FOA 1861 project teams.

Prior to distribution of the dataset to the project teams, the following assembly and pre-processing steps [15] were performed on the PMU data by staff at the DOE’s Pacific Northwest National Laboratory (PNNL), not necessarily in the following order:

a) Transformation of the file format of the data into comma separated value (CSV) files. Data providers had used a variety of formats in storing the PMU datasets, such as CSV files, proprietary binary archival files, and SQL server database files. Azure Databricks was used to transform the raw files into a common format prior to transformation into a Parquet file format used for dataset distribution.

b) In the case that the TOs provided multiple PMUs per data file (some files contained up to 1000 columns), splitting of the file was necessary, as well as reordering of the data fields, removal of extra fields.

c) Conversion of bad/missing string or numerical values to an empty string value. Data that was found by the assemblers to be missing was also represented as an empty string value. If PMUs were found to be producing more than 50% bad data, the entire dataset for that PMU was generally discarded.

d) Conversion of voltage and current units to consistent units of V and A (in some cases, samples were provided in units of kV and kA).

e) Conversion of dataset and event log timestamps to a consistent time zone, while also eliminating temporal offsets introduced by daylight savings time.

f) Modification of the provided event logs to ensure consistency of event label format.

g) Anonymization of the dataset in order to preserve the anonymity of the data providers: i.e., modification of dataset PMU labels to remove information regarding the absolute or relative spatial location of the PMUs within the interconnect.

The dataset comprises measurements gathered from these PMUs over 2 full years (2016–2017). The data is mostly 30 samples/s, but also contains 60 sample/s data. Each measurement is a timestamped record of up to 19 data channels (fields), including positive sequence and three-phase magnitude and angle for voltage and current, frequency (‘f’), rate-of-change-of-frequency (‘df’), and status variable (‘status’); in the following sections, ip_m and ip_a represent positive sequence current magnitude and phase angle, respectively. Most of the dataset (80%) contained only positive-sequence data: only 20% includes the three-phase quantities. The dataset was broken in 8 week intervals, with 6 weeks dedicated to the training dataset and 2 weeks dedicated to the test dataset. Table 1 shows the size of the Training Dataset component, which is the subject of analysis of this report. Doubly-compressed, the dataset requires 18.5 TB of storage space; if fully uncompressed, the dataset is estimated to grow to a significant fraction of a petabyte.

In addition to PMU data, data providers from each interconnect also provided event logs spanning the same time period as the dataset and containing thousands of event entries. The interconnect-wide event logs were processed by the DOE to ensure (a) consistency of terminology and (b) preservation of anonymity of the data provider and consolidated into a single aggregated event log for each interconnect. Network topology was not provided.

III. BAD DATA ANALYSIS

This section describes the results obtained from a quantitative data quality analysis of the FOA 1861 datasets following their distribution by DOE / PNNL. Note that this analysis was conducted following the aforementioned post-processing steps (a) – (g). Due to the discarding of some data discussed in step (c), this analysis is likely underestimate the full extent of bad data found within the files originally provided by the TOs. It is possible, however unlikely, that additional bad data was inadvertently introduced to the dataset during these previous data handling operations (a) – (g).

TABLE 1. Size of FOA 1861 [1] training dataset.

| Interconnect | # PMUs | # Records | Compressed data size (Terabytes) |
|--------------|--------|-----------|---------------------------------|
| IC_A         | 212    | 160,809,031,796 | 2.9                             |
| IC_B         | 43     | 93,353,826,102  | 4.7                             |
| IC_C         | 188    | 241,437,700,843 | 11.0                            |
| Total        | 443    | 495,600,558,741 | 18.5 TB                         |
TABLE 2. Bad data definitions BD1 – BD2.4.

| Type of Bad Data | Definition                                                                 |
|------------------|----------------------------------------------------------------------------|
| BD 1             | For each PMU, rows with: status $\neq 0$                                  |
| BD 2.1           | For each PMU, rows with: $(status = 0)$ AND (at least 1 unreasonabLe value for $vp_m, vp_a, ip_m, OR f$) |
| BD 2.2           | For each PMU, rows with: $(status = 0)$ AND (at least 1 non-numerical non-empty value for $vp_m, vp_a, ip_m, f$ OR $df$) |
| BD 2.3           | For each PMU, rows with: $(status = 0)$ AND (at least 1 empty string for $vp_m, vp_a, ip_m, f$ OR $df$) |

A. SUBDEFINITIONS OF BAD DATA

Table 2 describes several categories of bad data (BD) that were observed within the dataset distributed by DOE/PNNL, including empty string values; not-a-number (‘NaN’) or some other non-numeric, non-empty strings; ‘unreasonable’ values; and ‘non-zero status’ values.

Regarding the ‘non-zero status’ values, the synchrophasor standard IEEE C37.118.2 [16] provides an overview of the recommended use of the 16-bit ‘status’ variable. Non-zero values for certain bits within the status variable can serve as flags for the following: PMU errors; loss-of-synchronism with a UTC traceable time source; general data quality issues; PMU set to test mode; data modification in a post-processing stage; PMU configuration change (or impending change); PMU data sorted by arrival instead of by timestamp; or status of a trigger mode.

While bit 14 of the status variable is recommended to serve as a single ‘catch-all’ flag for self-detected data quality issue, other bits can also be associated with data quality issues. Any deviation from an all-zero condition was treated by this work as an aberration that, at the very least, deserves special consideration and/or dedicated pre-processing. Therefore, in order to simplify the presentation of data quality analysis in this work, any data entry with even one single bit of the status variable set high is treated as a ‘non-zero status value’ and a bad data entry of type BD 1. In this work the term ‘status = 0’ is used to denote the condition in which all 16 bits within the status variable are zero. For data entries that meet this ‘status = 0’ condition, the associated measurements may be considered ‘good data’, assuming that the associated measurements do not qualify as ‘unreasonable’, per the additional bad data subdefinitions defined in this section.

While a threshold for what is a realistic value is sometimes difficult to determine, extreme values that are highly unlikely to actually occur within the dataset can be considered bad data. In Table 2, ‘==’ means ‘equal to’, ‘!==’ means ‘not equal to’, and ‘unreasonable value’ for each selected field has been subjectively defined in the context of this work as:

1. Unreasonable $vp_m$: ($vp_m < -1 \times 10^8$) OR ($vp_m > 1 \times 10^8$)
2. Unreasonable $vp_a$: ($vp_a < -180$) OR ($vp_a > 180$)
3. Unreasonable $ip_m$: ($ip_m < -1 \times 10^7$) OR ($ip_m > 1 \times 10^7$)
4. Unreasonable $ip_a$: ($ip_a < -180$) OR ($ip_a > 180$)
5. Unreasonable $f$: ($f < 0$) OR ($f > 100$)

It should be noted any particular instance of bad data, including those contained in the four categories discussed here, may actually be an artifact of an event of interest, such as a failure or misoperation of power systems equipment, failure or misoperation of a communication network or archival equipment, or a cyber attack. No special effort is made here to distinguish these more interesting events from mundane bad data occurrences or to determine root cause of each type of bad data.

B. RESULTS OF BAD DATA ANALYSIS

Using the big data platform described in [12], dataset-wide queries were launched to quantify the instances of bad data subtypes BD 1, BD 2.1 BD 2.2 and BD 2.3, as defined in Table 2.

Among the $1.61 \times 10^{11}$ rows of data in Interconnect A, the first section of Table 3 shows that 21.39% of the rows contain BD 1, which means more than 1/5 of total PMU rows have non-zero status (i.e., the PMU has self-detected issues or specialized settings). Regarding the zero-status (i.e., no self-detected PMU issues or specialized settings) PMU rows, while BD 2.2 issues (rows with non-numerical non-empty values, as defined previously) are negligible (0.07%/(1-21.39%) = 0.09%), roughly 1/3 (24.97%/(1-21.39%) = 31.8%) of zero-status PMU rows present missing value issues. Therefore, bad data accounts for a significant fraction of the data in Interconnect A. Table 3 indicates how at least one instance of BD 1 occurs in all 212 PMUs in the Interconnect A dataset. For this interconnect, Figure 1 shows how the percentage of rows affected by BD1 in each PMU...
can reach as high as 100%. For 25 of the 212 PMUs – A183, A196, A227, A283, A288, A291, A332, A348, A359, A387, A417, A451, A469, A579, A598, A608, A612, A687, A697, A717, A730, A858, A876, A883, A971 – almost 100% of the rows present zero status, and there is very limited amount of data eligible for event analysis at all. BD 2.1 and BD 2.3 issues occur to 193 out of 212 (91%) PMUs in Interconnect A. Row counts for BD 2.1 and BD 2.3 are very similar; in general, the percentage of rows affected in each PMU vary from 0-85%, and a few PMUs that contain a significant BD 2.1 or 2.3 issues are A189, A268, A412, A441, A639, A842, and A937. BD 2.2 issue occurs to 8 PMUs (<5%) in Interconnect A, and for those 8 PMU affected, the percentage of rows with BD 2.2 are insignificant.

As shown in Table 3, among the $9.3 \times 10^{10}$ rows of data in the Interconnect B dataset, only 2.53% of PMU rows have non-zero status values; and with regards to data rows with zero status (97.47%), only a negligible amount of data presents as BD 2.1, 2.2, or BD 2.3 issues. Although overall BD issues are insignificant for Interconnect B as detailed previously, there are 15 PMUs (35% of total PMUs) that present BD 1 issues, and all 45 PMUs present BD 2.1 and BD 2.3 issues to certain extent. In particular, a few PMUs draw our attention. For example, in B635, roughly 100% of the rows present BD 2.1 and BD 2.3 issues, and therefore, there is negligible amount of data can be used for further analysis. In another instance, the BD 1 issue is significant for B209, B850, and B864, more than 40% of rows need be properly filtered out prior to application of machine learning or data analytics algorithms.

Table 3 also indicates that for Interconnect C, a majority (97.99%) of the $2.4 \times 10^{11}$ PMU rows have zero-value status, among which BD 2.1 and BD 2.3 issues are dominating and the total percentage of affected rows falls between (10.5% and 21.0 %). As shown in the last four rows of Table 3, it is observed that of all 188 PMUs in Interconnect C, 177 (94%) present a BD 1 issue, more than 50% presents BD 2.1 or BD 2.3 issues, and 10% present a BD 2.2 issue. Figure 3 shows that a few PMUs merit special attention: for instance, certain PMUs, such as C123, C141, C232, C294, C320, C484, and C733, have very significant quantities of BD 2.1 or BD 2.3, with almost 100% rows affected, and for those PMUs there is very limited data left for meaningful application of machine learning or data analytics. In addition, about 50% of data rows in C620 have a BD 1 issue (non-zero status).

Comparing the bad data statistics for all three datasets, it can be concluded that the overall quantity of bad data ranges widely among three FOA 1861 interconnects, and is the most severe for the Interconnect A dataset, followed by Interconnect C dataset, and then Interconnect B dataset. Comparing different bad data categories across all three interconnects, it is noticed that BD 1 (non-zero status issue), BD 2.1 (unreasonable numerical value issue), and BD 2.3 (missing value issues) are the major bad data issues reported in the training
dataset, while BD 2.2 (non-numerical non-empty value issue) is negligible in the given training dataset. The results of this section provide some insight into the types of bad data issues observed through the United States PMU network and underscore the strong need for bad data filtering prior to application data analytics and machine learning algorithms.

It should be noted that the BD 1 definition introduced in Subsection II.A may be unfairly broad, and that there may be instances in which the measurements could still be considered to be ‘good data’ even if one or more flags in the status variable are set high. It is therefore worth considering a narrower definition of BD 1 defined only by bits 13, 14 and 15 of the status variable. As stated in [16], bit 13 indicates a loss-of-synchronism with the time reference, bit 14 indicates a PMU error or test mode, and bit 15 indicates a PMU error with no information about the integrity of the measurement—these may be considered flags associated with the ‘most egregious’ self-reported data quality conditions. If BD 1 were to be redefined to denote status variable entries in which only these bit 13, 14, and/or 15 is set to 1 (with all other ‘less egregious’ flags ignored), the BD 1 percentages would be only 1.0%, 2.1%, and 0.19% for Interconnects A, B, and C, respectively, instead of the 21.39%, 2.53%, and 2.0% reported in Table 3: an order of magnitude reduction in BD 1 for Interconnects A and C, but only a small reduction for Interconnect B.

Note that any redefinition of BD 1 may also modify the BD 2.1 to 2.3 counts, since BD 2.1 to 2.3 were defined to be mutually exclusive from BD 1. However, the BD 1 bad data counts calculated using the original, broader version of the BD 1 definition (i.e., the definition in Section II.A and in Table 2, which is used to generate the results in the figures of this section) only account for 21.39% of the measurements from Interconnect A, and far less for Interconnects B and C. Therefore, if the narrower definition of BD 1 were used, and BD 2.1 to 2.3 were subsequently applied to the expanded number of potentially ‘good’ measurements that were no longer pre-emptively excluded as BD 1, it is not anticipated that the quantities of BD 2.1 to 2.3 bad data entry counts would change very much.

Due to the unique circumstances surrounding the assembly and provision of the synchrophasor datasets, the FOA 1861 project teams were only provided a limited understanding of the origin of the raw synchrophasor datasets prior to their distribution by the DOE. Consequently, FOA 1861 project teams were not provided detailed information regarding the practices of the data providers that may provide valuable insight into the data quality variation among the three interconnects. This variation may be attributed to factors such as PMU hardware procurement and commissioning practices, equipment management or maintenance practices, PMU data archival practices, and/or regulatory factors. While there were opportunities to discuss some of these details with several of the individual data providers during the FOA 1861 program, the identity and geographic location of the data providers was deliberately obfuscated in order to protect their privacy, and it was not possible to associate information obtained from any of the data providers with any specific interconnect. Consequently, at this time, it is not easy to speculate on the root cause of the variation of data quality between the three interconnects.

IV. ADDITIONAL ‘HARD-TO-DETECT’ BAD DATA
In addition to relatively-straightforward data quality challenges described in the previous section which could largely be mitigated through automated data cleansing and imputation processes (BD 1 and BD 2.1-2.3), there were also examples of ‘hard-to-detect’ bad data. These more confounding manifestations of bad data, if left unaddressed, have strong potential to jeopardize the downstream applications of ML.

For example, a number of PMUs contains data significantly and consistently lower than any reasonable voltage levels; they are also consistently lower than the PMU’s own median $vp_m$ levels over the 2-year period. A number of such PMUs with suspicious $vp_m$ values were identified that lasted more than 7 days. A snapshot of this phenomenon is provided in Figure 4, where multiple PMUs are exhibiting seemingly random drops in voltage magnitude, which occur suddenly and persist over multiple days or weeks before returning to a value closer to the median (y-axis information has been removed from this plot to protect the identity of the data-providers). Cursory analysis of the event log generally reveals no labeled event in the vicinity of the anomalous step changes. In several cases, the voltage drops by a factor of almost exactly 1000×, possibly suggesting that the voltage signal unit of measurement is being ‘toggled’ between kilovolts and volts, or that the unit conversion employed during dataset assembly, as discussed in Section I, was not completely successful for all data rows. Occasional data handling errors can be considered conceivable or even likely, considering the dataset contains approximately 0.5 trillion data records. As voltage drops such as these are similar to those
associated with real-world power system phenomena such as a fault or equipment failure event, manual intervention is often required to categorize this anomalous data, resulting in a time-consuming mitigation effort.

In the case that a distribution-based method is employed for feature data, it is hard to find a common threshold for all features. This is because the features (as opposed to raw data) tend to have very extreme tails that actually represent abnormal conditions, and thus should not be considered as bad data. With a poorly-selected threshold, the abnormal data will be dropped together with bad data. During normality modeling, anomaly detection and feature ranking, a rule-based approach can be employed to detect bad or “uncommon” samples on feature data, which could come from undetected bad raw data, or from the feature calculation function itself.

Additional examples of hard-to-detect dataset quality issues included event log entries in which the impact of daylight savings time was not fully taken into account during dataset assembly, causing one-hour temporal displacement of the event log timestamp. It should be also noted that manual intervention was required to address other challenging cases of bad data: for example, PMU C260 is completed excluded from normality modeling and downstream analysis due to persistent, wide-range, slow fluctuation in its \( v_p_m \) signal. Finally, there were additional examples of suspicious frequency or voltage magnitude behavior, highlighted by a global statistical analysis in the following section.

V. DESCRIPTIVE STATISTICAL ANALYSIS
This section presents descriptive statistical analyses on \( ip_m \) and \( f \), and \( v_p_m \) fields for each PMU in the training dataset using the big data analytics platform described in [4]. One intent of this analysis is to demonstrate a successful application of the big data analytics to extract meaningful insight into system operation after addressing the basic data quality issues defined in Section II. However, this analysis is also an appropriate ‘next step’ after basic data cleansing, as it can provide insight into remaining hard-to-detect data quality issues and/or equipment mis-operation such as those described in Section IV. Such ‘hard-to-detect’ bad data may be evidenced by statistical anomalies that merit further inspection. Descriptive statistics calculated over the two-year timespan include minimum value, maximum value, 1st quartile (Q1), 2nd quartile or median (Q2), and 3rd quartile (Q3), and average value of the data for each selected field (\( f \), \( ip_m \), and \( v_p_m \)) for each PMU. Therefore, the descriptive statistical analysis is referred to here as a 6-pt statistical analysis. Note that this 6-point statistical analysis is run on a refined dataset that excludes the basic subtypes of bad data defined in Section II.

Results from the 6-pt statistical analyses for frequencies measured by each PMU are shown in Figures 5 through 8. Each box plot shows quartile Q1, Q2, and Q3 values of the frequency field of the selected PMU. The min or max frequency value of the selected PMU will be represented by black whiskers if it falls within lower and upper limit of the data range (Q1 - 1.5IQR = lower limit, and Q3 + 1.5IQR = upper limit, where IQR denotes the interquartile range); otherwise, it will be treated as an outlier and represented as a black diamond in the box plot. The average frequency value for the selected PMU is marked by a red ‘x’. For a symmetric distribution (e.g. normal distribution), median and average values are overlapping. If there is any positive or negative skew in the distribution, average value will deviate from median values.

For Interconnect A, it is observed that the distribution of frequency for most PMUs is closely centered around the nominal frequency value 60Hz with IQR < 0.02Hz, except in the case of a few anomalies indicative of bad data. For example, both Q1 and median for A288 are lower than the 60Hz nominal value. Also, most frequency values for A106 and A834 are around 10Hz; most frequency values for A412 above 90Hz; the majority of frequency values of A639 are widely spread between 60 90Hz; and the majority of frequency values for A417 are around 55Hz.

![FIGURE 5. (a) Six-point statistical analysis of frequency measured by each PMU in Interconnect A training dataset across a two-year timespan. (b) Six-point statistical analysis of frequency measured by each PMU in Interconnect A training dataset across a two-year timespan (magnified).]
For Interconnect B, the distribution of frequency for all 45 PMUs are closely centered around nominal 60Hz with IQR < 0.2Hz. For Interconnect C, similarly, the distribution of frequency for all 45 PMUs are closely centered around
nominal 60Hz with IQR < 0.2Hz. Results from the 6-pt statistical analyses for PMU current magnitudes ($i_{p_m}$) are shown in Figures 8 through 10. The greater spread of current magnitude for each PMU can be observed, which is dependent upon fluctuations in system loading across two years of PMU data. In addition, Q3 (75%) current values for most PMUs are less than 2000A, except A288, who current are among 6000A to 13000A. For Interconnect B and Interconnect C, similarly, Q3 (75%) of current magnitudes are less than 2000A and widely distributed compared to voltage and frequency magnitude quartiles.

While global 6-point statistical analyses on voltage magnitude were also conducted, detailed results are not presented here in order to preserve the anonymity of data providers. While the voltage distribution of most PMUs in are closely centered around their nominal voltage ratings, there are a few anomalies, such as A940, A782, A417, A972, where IQR of each PMU’s $v_{p_m}$ value is significant, close to or even more than the absolute voltage ratings themselves. In addition, for A876, all of the voltage 6-pt statistics are negative and all fixed at $-1,310,680$V. Beyond the mentioned PMUs above, there are also a few others where their average values in Interconnect A deviate from the corresponding medium values significantly and might indicate data pattern anomalies, such as A631, A176, A353, A655, A301, A810, A888. In the case of Interconnect B, while the voltage distribution of most PMUs are closely centered around their nominal voltage ratings, a few PMU contain abnormal and unusual wide spread of $v_{p_m}$ distribution, such as B370, B727, B661, and B464. For Interconnect C, while the voltage distribution of most PMUs in are closely centered around their nominal voltage ratings, PMU C260 contain abnormal and wide spread of $v_{p_m}$ distribution. There are also a few others where their average values in interconnect C deviate from the corresponding medium values significantly, such as C764, C117, C851, C816, and C838.

Regarding the choice of statistical analysis employed in this section, a decision was made to use quartiles instead of...
### TABLE 4. Key FOA 1861 [1] dataset quality challenges encountered in example application of semi-supervised machine learning Algorithm, employed mitigation approaches, and suggestions to dataset providers.

| Dataset Challenge | Description/comments | Observed impact on successful application of ML to PMU data (1 to 10 with 10 being most detrimental) | Mitigation strategy employed and extent mitigated (1-10, with 10 being fully mitigated) | Suggestions to Dataset Providers, Aggregators and/or Assemblers |
|-------------------|----------------------|---------------------------------------------------------------|------------------------------------------------------------------|--------------------------------------------------------------------------------|
| 1. Basic dataset quality issues | 1. Status value \( \neq 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; impact can be higher if there are very few labeled events. | Basic data cleansing. Extent mitigated: 9 | Regular inspection of PMU data to verify functionality, adding flag if mis-operation suspected. |
| 2. Hard-to-detect dataset quality issues | Difficult-to-detect data quality issues that can masquerade as events, including 1000x step-changes in magnitude and slow fluctuations | Impact: 7. Pollution of ‘normal’ data used to develop normality model. | Visual inspection, additional filtering rules. Extent mitigated: 7-8 (Interconnect C); 5-6 (Interconnect B). | Regular inspection of PMU data to verify functionality, adding flag if mis-operation suspected; 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. |
| 3a. 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 spatial 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. |
| 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). | 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) |
| 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 impact 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. | Use anomaly detection only within a limited window around labeled events. Extent mitigated: 7 | Employ mechanisms or strategies to label the unlabeled events. Alternatively, dataset providers may also 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. |
| 5. Large number of unlabeled events | Tens of thousands of unlabeled events with severity equal to or greater than that of the several thousand labeled events (severity determined by magnitude of features contained 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 | |

Other groupings (such as tertiles or deciles) in order to achieve as much insight as possible without excessive computational burden. Tertiles (grouping the data into thirds) also offers nearly as much insight and would be computationally easier to compute than quartiles, but quartiles allow for more insight, and have the added benefit of also providing the median, which is a frequently-used and intuitive statistical measure. Furthermore, quartiles are compatible with the well-known box-and-whisker plot method, which is familiar to many and allows for an expedient and intuitive understanding of the statistical results. Deciles (grouping data into tenths) also provides the median, and provides greater resolution, but would demand a far greater computational time.

### VI. EVENT LABELING CONSIDERATIONS AND HOLISTIC SUMMARY OF DATASET READINESS FOR ML

An industry-validated semi-supervised machine learning strategy for event signature identification was applied to the FOA 1861 Interconnect B and C datasets, as described in [14] and [17]. Table 4 reviews key dataset-related challenges in the context of this example ML application, an estimation of the impact of each challenge on the successful application of the semi-supervised machine learning strategy, the tested mitigation approach, and estimation of the extent to which each challenge was successfully mitigated, and suggestions to dataset providers.
Many of the rows within Table 4 relate to the event log information, including the quantity of labels, as well as their temporal and spatial precision. While the event logs contained thousands of labels carefully curated by the DOE during dataset assembly (a majority of which were line events), there were tens of thousands of additional unlabeled events discovered in the dataset of equal or greater magnitude than the labeled events. Additionally, some labeled events were relatively rare despite the large size of the datasets: generator events were limited in number to several hundred across the three interconnects, and only about half of these were be detectable in the dataset.

As described in Table 4, a combination of dataset characteristics, including dataset anonymization, temporal imprecision, and a large number of unlabeled events can work together to compound the difficulties of applying supervised or semi-supervised machine learning to synchrophasor datasets. An illustrative example is shown in Figure 11: hypothetical waveforms are shown derived from a fictitious dataset that includes an event log with only one label (Event #1) and time series data from only two PMUs (PMU A and PMU B). As shown in the figure, there is an imperfect geographical pairing between the event log and the PMU dataset: the event log associated with the region where PMU A is located has been provided, but the event log associated with the region where PMU B is located has not been provided. However, since the dataset is anonymized—i.e., information regarding the location of the events and location of the PMUs have been removed from the dataset—this imperfect association of the event log and the PMU data is not known to the dataset recipient.

When the positive sequence voltage magnitude waveforms associated with each of the PMUs are examined, two anomalies can be observed: Event #1 and Event #2. The single event in the event log is meant to be associated with only one of the anomalies, Event #1. The second anomaly in the data, Event #2, is associated with an event outside of the scope of the provided event log, but this fact is not known to the recipient of the anonymized dataset. Due to the finite temporal precision of the labels in the event log, if Events #1 and #2 occur in close-enough succession, it may be difficult to determine which anomaly (Event #1 or #2) should be associated with the event log entry. It should be noted that even in the case in which there is a perfect geographical pairing between the event log and the dataset, unlabeled events are a very likely possibility, and the same challenge can occur. The problem becomes aggravated as the size of the anonymized dataset increases and potentially includes data and event logs from multiple regions and data providers. If the dataset spans an entire interconnect, chances are increased that unlabeled events occur in synchronicity with labeled events. This phenomenon can be detrimental for the signature identification process, and result in inaccurate event classification and characterization.

If the dataset was not fully anonymized, it may be possible to infer an association between Event #1 and the event log label based on knowledge of relative electrical proximity between the event and the PMU. However, since this knowledge is not available, it becomes much more challenging to determine which anomaly is related to the event log label. Improved temporal precision of the event label would be very helpful, and even a better understanding of the event log label temporal precision would be helpful. While some estimates were provided by the TOs regarding temporal precision, there is likely to be variation in temporal precision based on mechanism of entry of event log data, and a temporal precision estimate for every label would be helpful. As mentioned in Table 4, almost any additional spatial information that the data providers could provide would helpful, such as identification of the PMU geographically or electrically closest to the labeled event; and/or increased granularity of the dataset (i.e., break ICs into several regions, associating sections of the anonymized raw data with sections of the anonymized event log); and/or discarding of an event log label if the event is of sufficient electrical distance from any PMU.

### VII. CONCLUSION

A comprehensive data quality analysis and qualitative ‘machine learning-readiness’ evaluation of several large-scale, real-world phasor measurement unit (PMU) datasets have been provided. Significant variation has been found in the overall quantity of bad data across the three datasets, and in the distribution between major categories of bad data. Several examples of ‘hard-to-detect’ bad data observed within the datasets have been described in detail. Using a customized, state-of-the-art big data environment [4], evidence of several additional hard-to-detect data quality challenges have been uncovered in a 6-point statistical analysis of the positive sequence voltage, current and frequency for all 443 PMUs across all three interconnects and two years of operation.

Remaining obstacles for convenient application of off-the-shelf (supervised and semi-supervised) machine learning tools have been summarized, and new insights are provided regarding challenges associated with present-day event labeling practices, large spatial scope of the dataset, and dataset anonymization. Mitigation strategies, including
relabeling, have enabled the successful application of an industry-validated, feature-generation-intensive semi-supervised machine learning strategy to the dataset, allowing for grid event signature identification, classification of unlabeled events, as well as production of meaningful data-driven insights into event severity, spatial breadth, and duration of labeled and classified events [14]. However, there is more value ready to be unlocked from the FOA 1861 datasets—e.g., for equipment health monitoring applications and causal analysis—if the challenges in Table 4 are more satisfactorily addressed by data providers or additional research and development.

Successful application of supervised and semi-supervised machine learning strategies depends upon a large number of labeled events. To the extent possible, additional labels should be applied to the many observable, unlabeled events within the dataset. Additionally, high-quality PMU data should be sought for rarer event types, such as generator events. To supplement the limited number of these rarer, labeled events, a shared PMU data and event log database could be created by a federation of like-minded data providers to pool this valuable data, such as in [18]. Since a fully-de-anonymized, shared database can represent a security concern, new techniques for ‘semi-anonymization’ could be explored that provide sufficient spatial information to permit unrestrained application of supervised and semi-supervised machine learning (circumventing challenge #3 in Table 4, and thereby also mitigating the impact of challenges #4 and #5) while continuing to preserve the anonymity of the data providers and concealing sensitive information such as network topology. The authors believe that there are solutions, such as the ‘increased spatial granularity’ approach described in Table 4, that address the important security concerns of data providers while unlocking substantial additional value from the dataset.

ACKNOWLEDGMENT

Disclaimer: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring of the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

REFERENCES

[1] Funding Opportunity Announcement (FOA) Number: DE-FOA-0001861, Big Data Analysis of Synchrophasor Data, Amendment 000001, Dept. Energy (DOE) Office Electr., Washington, DC, USA, Oct. 2018.
[2] Y. Guo, Z. Yang, S. Feng, and J. Hu, “Complex power system status monitoring and evaluation using big data platform and machine learning algorithms: A review and a case study,” Complexity, vol. 2018, pp. 1–21, Sep. 2018.
[3] H. Akhavan-Hejazi and H. Mohsenian-Rad, “Power systems big data analytics: An assessment of paradigm shift barriers and prospects,” Energy Rep., vol. 4, pp. 91–100, Nov. 2018.
[4] L. Cheng and T. Yu, “A new generation of AI: A review and perspective on machine learning technologies applied to smart energy and electric power systems,” Int. J. Energy Res., vol. 43, no. 6, pp. 1928–1973, May 2019.
[5] M. Wang and J. H. Chow, “Data quality and privacy enhancement,” in Advanced Data Analytics for Power Systems, A. Tajeri, S. M. Pelaza, and H. V. Poor, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2021, pp. 261–278.
[6] C. Huang et al., “Data quality issues for synchrophasor applications Part I: A review,” J. Modern Power Syst. Clean Energy, vol. 4, no. 3, pp. 342–352, Jul. 2016.
[7] A. Sundararajan, T. Khan, A. Moghadasi, and A. I. Sarwat, “Survey on synchrophasor data quality and cybersecurity challenges, and evaluation of their interdependencies,” J. Modern Power Syst. Clean Energy, vol. 7, no. 3, pp. 449–467, May 2019.
[8] L. Miller et al., “Synchrophasor data quality attributes and a methodology for examining data quality impacts upon synchrophasor applications,” NASPI PMU Appl. Requirements Task Force, Richland, WA, USA, Tech. Rep. NASPI-2017-TR-002/PNNL-26313, 2017.
[9] D. Nguyen, R.arella, S. A. Wallace, X. Zhao, and X. Liang, “Smart grid line event classification using supervised learning over PMU data streams,” in Proc. 6th Int. Green Sustain. Comput. Conf. (IGSC), Las Vegas, NV, USA, Dec. 2015, pp. 1–8.
[10] M. Cui, J. Wang, J. Tan, A. R. Florita, and Y. Zhang, “A novel event detection method using PMU data with high precision,” IEEE Trans. Power Syst., vol. 34, no. 1, pp. 454–466, Jan. 2019.
[11] D.-I. Kim, T. Y. Chua, S.-H. Yoon, G. Lee, and Y.-J. Shin, “Wavelet-based event detection method using PMU data,” IEEE Trans. Smart Grid, vol. 8, no. 3, pp. 1154–1162, May 2017.
[12] 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,” in Proc. IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT), Washington, DC, USA, Feb. 2021, pp. 1–5.
[13] B. P. Leao, D. Fraadkin, Y. Wang, and S. Suresh, “Big data processing for power grid event detection,” in Proc. IEEE Int. Conf. Big Data (Big Data), GA, USA, Dec. 2020, pp. 4128–4136.
[14] T. Wang et al., “Phasor-measurement-based data analytics using digital twin and PhasorAnalytics software: Training and test datasets report,” Gen. Electr. Res., New York, NY, USA, Tech. Rep. DOE-GE-0000915-1, Aug. 2021, doi: 10.2172/1828164.
[15] J. S. Banning et al., “FOA 1861 data curation overview,” PNNL, Richland, WA, USA, Tech. Rep. PNNL-32303, 2021, doi: 10.2172/1860940.
[16] IEEE Standard for Synchrophasor Data Transfer for Power Systems, Standard C37.118.2–2011, IEEE Power and Energy Society, Dec. 2011.
[17] T. Wang et al., “Grid event signature identification through large, real-world PMU datasets,” Gen. Electr. Res., Niskayuna, NY, USA, Tech. Rep. 2021092068790, 2022.
[18] E. Litvinov et al., “A cloud-hosted synchrophasor data sharing platform,” in Power System Grid Operation Using Synchrophasor Technology, S. Nuthalapati, Ed. Cham, Switzerland: Springer, 2018, pp. 477–498.