Big Data Analytics for Future Electricity Grids

Mladen Kezunovic
Texas A&M University
College Station, TX, USA

Pierre Pinson
Technical University of Denmark
Kongens Lyngby, Denmark

Zoran Obradovic
Temple University
Philadelphia, PA, USA

Santiago Grijalva
Georgia Institute of Technology
Atlanta, GA, USA

Tao Hong
Univ. of North Carolina Charlotte
Charlotte, NC, USA

Ricardo Bessa
INESC TEC
Porto, Portugal

***Abstract* — This paper provides a survey of big data analytics applications and associated implementation issues. The emphasis is placed on applications that are novel and have demonstrated value to the industry, as illustrated using field data and practical applications. The paper reflects on the lessons learned from initial implementations, as well as ideas that are yet to be explored. The various data science trends treated in the literature are outlined, while experiences from applying them in the electricity grid setting are emphasized to pave the way for future applications. The paper ends with opportunities and challenges, as well as implementation goals and strategies for achieving impactful outcomes.**

***Index Terms*—Electricity grids, Analytics, Big data, Decision-making.**

I. INTRODUCTION

The definition of big data analytics in the electricity grid applications is involved. This is due to its broad scope and numerous data science approaches that may be considered under this umbrella (statistical analysis tools, artificial intelligence, machine learning, deep learning, etc.). The complexity also stems from the numerous approaches in different applications such asset management, operations, control, protection, market and planning decisions, among others. One underlying issue that makes it particularly difficult to delineate what constitutes big data analytics is the volume of data being considered in electricity grid applications today, which typically is at the terabyte scale, far from the petabyte scale that is considered as big data in some other domains. Finally, the concept of "data analytics" is also a bit misleading since most of the traditional applications in the electricity grid domain are based on the processing of measurement data, which may or may not be considered as data analytics in a big data context. As an example, the traditional approaches to state estimation and fault location would not necessarily qualify as big data analytics if they are only based on mathematical equations derived from physics. On the other hand, fault prediction based on data-driven models utilizing high-resolution weather and outage data may qualify for big data analytics. As a result, this survey does not attempt to provide a rigorous definition of what constitutes big data analytics for power systems, but rather intends to emphasize the reported work that uses the latest advances in data sciences as applied in the electricity grid for specific applications such as asset and outage management, and integration of renewables.

The earliest work on the use of big data in the utility industry was published in 2013 [1], but field demonstrations were reported only in the last few years [2]. The approaches surveyed in this paper are at the crossroads of novel data analytics techniques, added application benefits, and unique data sets or features used in the implementation. While the literature is saturated by papers focusing on novel data analytics approaches, the added benefit of our paper is in identifying the papers that have bridged the vast data analytics literature with a few applications that have demonstrated tangible benefits using actual utility data [3]. The data sets used for the applications are quite often traditional, and in a few instances, new data integration and management approaches were used [4]. Covering an emerging area in the electricity grid domain, the related papers are surveyed with an intent to define promising trends, and what may result in transformational ideas in the future [5]. While we recognize the contributions of other survey papers [2], [3], [5], [6], [7], [8] and recent books on the subject [9], [10], we did not attempt a rigorous approach of comparing our survey paper to such surveys to avoid any overlaps since the focus and context of our paper stems from our own deployment experiences and views that come from our own practical insights.

To facilitate the educational component of the survey, we will introduce some basic concepts of what constitutes big data in electricity grid applications, and what are some of the traditional aspects of the data properties that uniquely represent the electricity grid domain. In doing so, some well-known facts about big data are framed using relevant electricity-grid-related examples. With the same goal, we offer a classification of the data analytics fundamentals and related implementation techniques. We also share experiences from the real-life implementations of big data analytics approaches and tools for electricity grids, focusing on the various engineering issues associated with the implementation of novel applications. We then focus on specific applications reported in the literatures

and try to classify them based on their relation to the various uses in the electricity grid settings.

The main contribution of the paper is in pointing out the key implementation issues, while at the same time providing a broad overview of the trends for selected applications. For future users, the paper gives a large number of references by researchers and electricity grid professionals but does not explore the level of their scientific contributions; the intent here is to trace the references associated with practical applications. Our classification provides easier access to the applications of interest rather than a guidance to the most relevant works in the general data analytics area. Some important works may not have been reported or we may have not been able to locate it due to worldwide publication spread.

The paper is organized as follows: after the introduction in Section I, we focus on the importance and the feasibility of big data analytics in Section II, the challenges in Section III, the applications survey in Section IV, and the future opportunities and challenges in Section V. A representative list of references is provided at the end.

II. Importance and Feasibility of Big Data Analytics in Electricity Grids

A. Impact of Big Data Analytics

The changes in the electricity industry are unprecedented, including the shifts in the energy mix, more active customers, new devices and technologies, and evolving business models never seen before. These resulted in increased complexity and uncertainty, bringing to the fore new challenges as well as opportunities. At the core of addressing these challenges is the need for better decision-making in the grid operational and planning stages, including long-term investment and policy. In addition, the grid is being instrumented with capabilities for sensing and data acquisition resolution with orders of a magnitude higher than what was previously implemented. The new data with novel data analytics methods can support the electricity grid objectives of higher resilience, economic efficiency, and reduced emissions. The electricity grid data analytics emerges as the key factor that enables the technologies for better decision-making. It became a core industry capability and a strategic advantage to organizations who seek to innovate and provide higher levels of service quality and customer satisfaction. The applications of data analytics to the electricity grid are numerous and can be identified in many activities of the industry. Data analytics is, hence, a transformational step toward the future grid.

B. Novel Data Sources

Applying big data solutions in different electricity grids is focused on exploring emerging heterogeneous data sources that have distinct quality, spatial and/or temporal resolution, and information presentation. It is feasible to leverage these data by applying the following knowledge extraction approaches in different use case examples: (a) combining emerging and conventional data sources, e.g., by using data fusion theory [11]; (b) extracting and combining information from different modalities (e.g., images, texts, categorical statements) using multimodal learning [12] or a heterogeneous information network [13]; and (c) combining data from geographically distributed data sources, e.g., by using classical vector autoregressive [14] or deep learning [15] methods.

Novel data sources are emerging in different domains:

- *The grid infrastructure*: system operators are improving network observability by installing phasor measurement units (PMU) that can provide high reporting rate data (e.g., 30 measurements per second of voltage/current magnitude, phase, and frequency) and remote terminal units (RTU) in substations and smart meters at the consumer level. Sensors for remote supervision of substations are also being tested for asset condition monitoring and quality of service improvement [16].
- *Renewable power plants*: the renewable energy industry is installing and operating monitoring sensors at the wind turbine and photovoltaic panel level, which generates a large volume of data (e.g., a wind turbine can have more than 100 sensors inside the rotor, which gather more than 10,000 data points every second) that can be used for predictive maintenance (and reduce Operation and Maintenance costs); data from a grid of numerical weather predictions, geographically distributed sensors (e.g., wind turbines, pyranometers), sky cameras, and satellite images can be combined to improve power (and weather) forecasting skills in multiple time horizons [17]. In renewable generation forecasting, the scale of studies has also grown from a single site to over 100 sites [18].
- *Consumer and social media*: while at the early stages of deployment, the proliferation of the internet-of-things devices in smart homes and buildings were creating conditions for data-driven energy and non-energy services [19], whose impact depends on solving challenges such as data privacy/protection and consumer engagement. Moreover, the increased footprints of social media have enabled the power companies to better understand and engage customers than ever before [20]. Researchers have also tried to fuse Twitter data into power outage detection [21].
- *Electricity markets*: Over the last few years in Europe, the electricity market transparency has improved noticeably, and after the publication of Regulation (EU) No 5 43/2013 [22], the amount of publicly available data is increasing [23], including access to individual offers from market players (usually available with a delay of few months). The same trend is happening in the USA, with platforms such as the Form EIA-930 data collection that provides a centralized and comprehensive source for hourly operating data of the high-voltage bulk electric power grid in the lower 48 states. This open data can be used for different objectives: to improve the price forecasting skills by combining prices from different regions [24] or to assess the large-scale impact of renewable energy generation in cross-border power flow [25].

- *Environmental and ambient domains*: the weather data is of paramount importance in predicting operating conditions, including faults. The data from ground weather stations [26], satellite [27] and radar resources [28] are readily available from government databases. Specialized sensor networks, such as the national lightning detection network in the USA [29], are also sources of rather useful weather data. Several weather forecast services are also at our disposal for providing pre-calculated features of the weather data sets [30]. Additionally, data about vegetation, soil, animal migration, and other ambient conditions may be readily available from various other sources [31]. The means of utilizing high precision data by using specialized databases such as Light Detection and Ranging (LIDAR) or drone surveys are also reported in the literature [32]. Such data is not typically collected within the utility industry jurisdiction and constitute an outside data of great value to the industry. In load forecasting, the research frontier has moved from temperature collected at a single station to a variety of weather variables and multiple weather stations [33]-[35]. In solar power forecasting, sky image data are heavily used for cloud detection [36].

Researchers and practitioners nowadays focus on exploiting existing data and exploring emerging data sources and data at a larger scale to pursue improvements in electricity grid planning and operations. A large data set or a variety of data sources are not necessary to claim a research topic in big data analytics. There are many other important aspects of big data analytics, such as building algorithms that can leverage a high-performance computing environment and expanding the size of models to capture detailed features in the data [37]. Another example is to use hourly load and weather data informed long-term load forecasts, which are traditionally based on monthly data [38], [39]. A recent review article on smart meter data analytics listed 10 publicly available datasets for electricity demand [40]. Table I highlights some example data sources for general applications of power systems data analytics, with an emphasis on publicly available data to promote reproducible research. Such a list of datasets is increasing at a rapid pace under various open modeling approaches and data sharing initiatives. Such data is mostly related to electricity network measurements or properties.

In such studies, diverse data sets with quite different data properties. Fig. 1 illustrates data properties for an example of the use of one type of (synchrophasor) data. Examples of where merging diverse data sets created value may be found in the reported work on outage prediction [41]. Table II gives examples from a particular geographic region (USA). Such data may be available in many other parts of the world from local government agencies or industry services. Unlike the data in Table I, this data is characterized by not being directly related to power system measurements or properties, yet is highly correlated to the data in Table I. The importance of big data properties depicted in Fig.1 is that many such properties may be found in datasets used in an electricity grid application, which creates non-trivial data integration challenges.

TABLE I. EXAMPLES OF OPEN-SOURCE DATA SETS

Data source	Application areas
GEFCom2012 [42]	Load forecasting; wind power forecasting
GEFCom2014 [43]	Load forecasting; price forecasting; wind power forecasting; solar power forecasting
GEFCom2017 [44]	Distribution level load forecasting
Irish data [45]	Smart meter data analytics
ARPA-E GRID DATA projects [46]	Power system analysis
My Electric Avenue [47], [48]	Electric vehicles
EV Research @ Caltech [49]	Electric vehicles
ENTSO-E Transparency Platform [23], [50]	Electricity markets
European power system [51], [52]	Power system models
UK power system [53]	Power system models
Load dataset with grid data [54], [55]	Load forecasting
Sotavento wind farm, Spain [56]	Wind power forecasting
NREL wind integration toolkit [57]	Wind integration
Data sets for benchmarking solar energy forecasting methods [58]	Solar energy forecasting
Photovoltaic hourly power measurements and geographical grid (169 equally distributed points) from the Weather Research and Forecasting model [59], [60].	Solar energy forecasting

C. Important Considerations when Creating Datasets

How the data sets are created is quite important for big data analytics applications due to additional considerations such as:

- Spatiotemporal correlation and synchronism
- Scalability
- Missing data
- Bad data diversity
- Various types of uncertainties

How each of these considerations reflect on the big data applications in the electricity grid is outside the scope of this paper, but certainly is worth exploring as new data sets get added and merged.

III. BIG DATA ANALYTICS CHALLENGES

A. Data Sciences Foundations

The goal of data science is to extract value from data. The steps of the data management life cycle include data collection; preprocessing (exploration, sampling, dimensionality reduction/feature selection, feature creation, transformation, cleaning, and integration); analytical processing (modeling, which often includes multiple building blocks); interpretation; and reporting results [61]. The key skills needed in this area are often viewed as multidisciplinary at the intersection of computer science, mathematics, statistics, and the problem

domain. On the technical side, major challenges are typically related to big data, artificial intelligence, and machine learning methodologies, while the process of applied data science could also require social sciences, communications, and business skills, and it has been suggested that this intersection should include additional disciplines [62].

A holistic view of data science emphasizes that data science is "more than a combination of statistics and computer science" as "it requires training in how to weave statistical and computational techniques into a larger framework, problem by problem, and to address discipline-specific questions" [63]. The same authors point out that data science requires: (1) understanding the context of data, (2) appreciating the responsibilities involved in using private and public data; and (3) clear communication on what can and cannot be inferred from a dataset.

The core components of data science are machine learning-based methods for finding patterns in data that may provide insights into the phenomena described by the data, and predictions regarding future events of interest. In machine learning, the objective is to learn a function that maps the given input data (explanatory variables) to the observed output (response). A simplified representation of reality created for this purpose, called a model, is used to estimate the unknown response for new cases based on observed explanatory variables of interest, and this process is called inference or, more simply, prediction.

Machine learning techniques typically address applications where traditional analytics are inappropriate due to data size, high dimensionality, heterogeneity, diversity, or other challenges. Methods are developed to address various aspects of these challenges. In some methods, independence of data records is assumed (e.g., when data types are multidimensional numerical tables, tables with categorical or mixed attributes, or text). Otherwise, specialized data science methods have been developed to model implicit or explicit dependencies in data sets common in time-series, discrete sequences, or spatial, spatio-temporal, or network applications.

Machine learning objectives are often grouped into descriptive tasks and predictive tasks. Descriptive tasks aim to discover interpretable patterns that describe past data, and predictive tasks are those where the goal is to identify patterns observed in training data in order to estimate future predictions of risks and other outcomes. Descriptive tasks are usually unsupervised, meaning that only explanatory variables are considered in the analysis. Common descriptive objectives include data clustering [64], [65], association discovery [66], and detection of deviations from normal behavior, including extreme value analysis, outlier detection, and identification of emerging patterns [67]. Prediction tasks are supervised, such that they require not only explanatory variables but also the value of the dependent variable that is being predicted. Practical examples include risk assessment [68] and diagnostics [69].

There are also semi-supervised [70] and self-training methods [71], where training data includes some labeled data and much more unlabeled data.

In classification, the response variables being predicted are a class (e.g., one of several kinds of data labels), or in the case of regression, it is a continuous value. One of the commonly used approaches for classification is the induction-based decision tree. Hunt's algorithm [67], one of the earliest decision tree methods, proposed the general procedure of partitioning

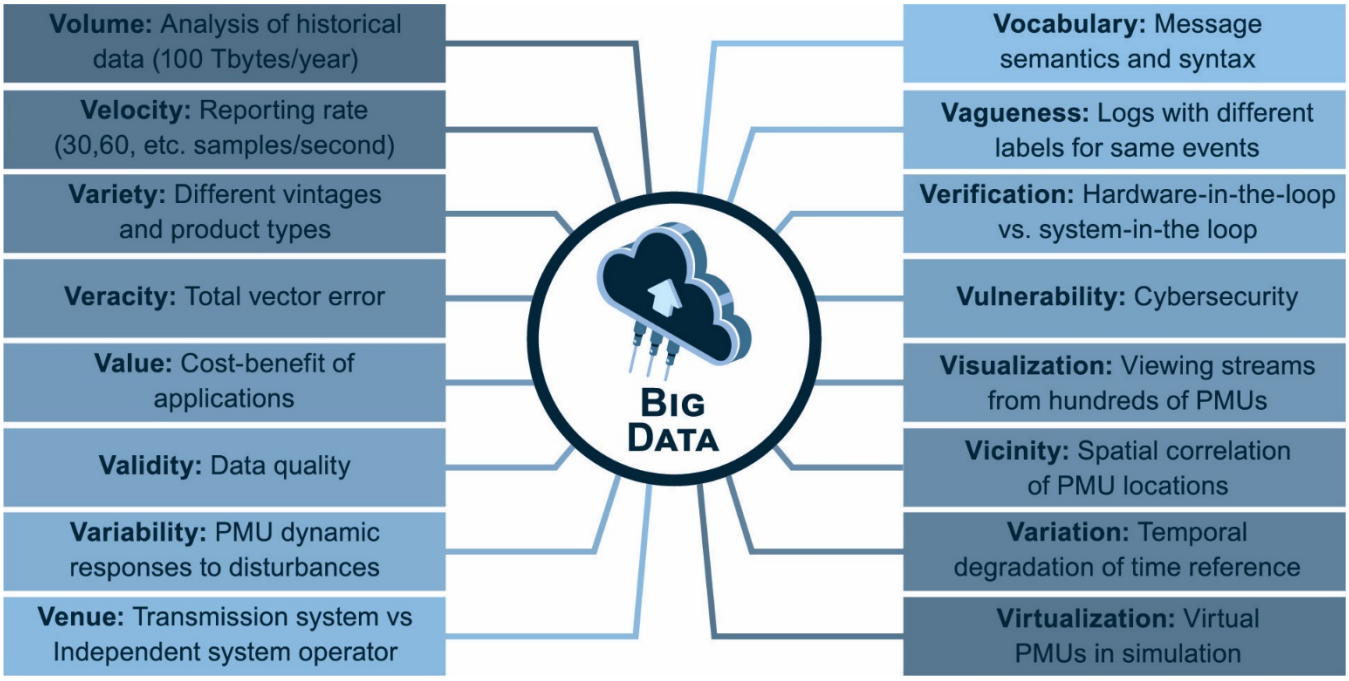

Figure 1: Big data properties

data, based on the value of a single attribute and proceeding recursively on subsets if the class is not sufficiently pure at the subsets. Many methods have been proposed to measure a data subset impurity and to determine the next split (e.g., entropy in CART [72] or Gini index in ID3 [73] and C4.5 [74]) as well as to prune the tree, thereby improving model generalization. Decision trees are easy to interpret, are quite inexpensive to construct, and are very fast at classifying unknown cases. They are also robust to noise and can handle redundant or irrelevant attributes, but they do not account for interactions between attributes. One of the limitations of decision trees is that they require pruning, as otherwise they grow too big resulting in overfit problems. This limitation is successfully addressed by Random Forests, built as an ensemble of decorrelated decision trees [75], [76]. This ensemble is inspired by the Bagging method, an aggregation method based on bootstrap sampling, developed for reducing variance without enlarging bias [77]. In Random Forests, this idea is further extended by limiting each node to consider only a small random subset of attributes. The resulting solution is empirically shown to be more accurate than the AdaBoost algorithm, which is an effective ensemble-based classifier that in training adjusts to the weight of an observation based on the previous classification [78].

A popular alternative technique that can handle interactions among explanatory variables is to classify a new case by computing distance to k-nearest neighbors in the training set and to predict the class based on a majority or a weighted

TABLE II. EXAMPLES OF THE DATA SETS AVAILABLE FROM GOVERNMENT AND COMMERCIAL SOURCES [79]

			VELOCITY	**VOLUME**	
	Source	**Data Type**	**Temporal Resolution**	**Spatial Resolution**	**Measurements**
V A R I E T Y	Automated Surface Observation System	Land-Based	1 min	900 stations	Air temperature, Dew point, Relative humidity, Wind direction, Speed and gust, Sea level pressure, Sky, Precipitation
	Level-2 Next Generation Weather Radar	Radar Data	5 min	160 high-resolution Doppler radar sites	Precipitation and atmospheric movement
	NOAA Satellite Database	Satellite Data	Hourly, daily, monthly	4 km	Cloud coverage, Hydrological observations (precipitation, cloud liquid water, total precipitable water, snow cover…), Pollution monitoring
	Vaisala U.S. National Lightning Detection Network	Lightning Data	Instantaneous	Median location accuracy <200m	Date and time, Latitude and longitude, Peak amplitude, Polarity; Type of event: Cloud or cloud to ground
	National Digital Forecast Database	Weather Forecast Data	3 hours	5 km	Wind speed, direction, and gust, Temperature, Relative humidity, Tornado probability, Probability of severe thunderstorms, etc.
	Texas Parks & Wildlife Department	Texas Ecological Mapping System	Static	10 m	Distribution of different tree species
	Texas Natural Resources Information System	NAIP	Year	50 cm – 1 m	High resolution imagery
	National Aeronautics and Space Administration	3D Global Vegetation Map	Static	1 km	Canopy height data
	National Cooperative Soil Survey	gSSURGO	Static	10 m	Soil type
	Utility Data	Historical Outage Data	Instantaneous	Feeder section	Location, Start and end time and date; Number of customers affected; Cause code
		Tree Trimming Data	Day	Feeder	Feeder location, Date, Trimming period, Number of customers affected, Cost of trimming
		Network GIS data	Static	Infinity (shapefile)	Poles: Location, Material/class, Height Feeders: Location; Conductor size, count, and material; Nominal voltage
		Historical Maintenance Data	Day	Tower location	Start and end date and time, Location, Type (maintenance, replacement), Cost, Number of customers affected
		Insulator asset data	Static	Infinity (shapefile)	Surge Impedances of towers and ground wires, Footing resistance, Component BIL
		In-field measurements	Instantaneous	Tower location	Leakage current magnitude, Flashover voltage, Electric field distribution, Corona discharge detection, Infrared reflection thermography, Visual inspection

majority of the identified k-neighbors whose class is known. This is a lazy learning method, since the model is not built explicitly, and inference time required to classify a new case is quite large. It also requires a comparison of each new data point to each data point in the training set. In addition, this technique is not easy to use when many attribute values are missing, since in such cases, the distance-based method of determining nearest neighbors could be unreliable. Some of these limitations could be overcome by proximity graphs, in which nodes are connected if certain geometric conditions are satisfied. In such a formulation, various efficient graph algorithms (e.g., minimum spanning trees and triangulations) could be used to identify nearest and relative neighbors more efficiently [80].

An alternative classification approach that is mathematically more rigorous is to estimate the posterior probability of the target class using Bayes' theorem. A simple but elegant and robust approach, called Naive Bayes, assumes that the attribute values are conditionally independent of each other, given the class label y. In such a case, the class-conditional probabilities of all the attributes can be factored as a product of the class-conditional probabilities of every attribute. The approach is robust to noise, missing values, and irrelevant attributes. However, conditional independence among explanatory variables is a strong assumption that is not valid in many applications. For such scenarios, a class of probabilistic graphical models called Bayesian Belief Networks were developed by modeling conditional dependencies via directed acyclic graphs. The exact inference on such graphs is NP-hard [81], and therefore, their applications are limited to smaller numbers of attributes or to special types of graph structures.

Another effective probabilistic classifier is logistic regression [82]. We have successfully used this method to predict weather-related power outage probabilities [83]. Assuming a two-class problem (the response y = 0 or 1), this approach avoids directly estimating the conditional probability of an instance, but instead estimates the ratio of posterior class probabilities $P(y=1|x)/P(y=0|x)$. A great advantage of logistic regression as compared to k-nearest neighbors is that it is applicable to high-dimensional problems, since the method does not rely on measuring similarities between data points. Another benefit of this approach is that weight parameters correspond to individual attributes and, therefore, provide fairly easy interpretability. Still, the presence of a large number of irrelevant attributes is a challenge for logistic regression, and this method is not applicable to classifying cases with missing values, which could be a serious limitation in practice.

The logistic regression model can be viewed as a case of generalized linear model. Other representationally powerful models from this category are Support Vector Machines (SVM) and Multilayer Neural Networks (MNN). In SVM, the optimization problem is formulated as finding a maximum margin separating the hyperplane for which a large region exists on each side of the decision boundary [84]. This is formulated as a constrained nonlinear programming problem expressed as a function of the coefficients of the separating hyperplane, which is solved using the Lagrange multiplier method. For non-linear classification, data is implicitly transformed into a high-dimensional space, where the problem is linearly separable. This is achieved by using the kernel trick, so as to reduce the problem effectively to a linear classification situation. Using carefully selected kernels (Gaussian, polynomial, or sigmoid) allows the approximation of arbitrary decision boundaries. The main benefits of the SVM approach are that it is robust to noise and reduces overfitting while finding the global minima of the objective function. However, the computational cost of SVM is high, and it is still challenging to use this model when descriptive variables are partially missing in observed data.

Multilayer feed-forward neural networks (FNN) are also used successfully for classification in a variety of challenging applications [85]. For example, we have successfully trained FNN to discriminate between power transformer magnetizing inrush and fault current [86]. This model has at least one layer of hidden units, each computing a nonlinear smooth and differentiable function of a weighted input sum (e.g., sigmoid function). In this model, the problem of updating the parameters when an error is observed at the output is commonly solved by error backpropagation from the output toward the previous layers. In this process, the error of a node in the hidden layer is estimated as a function of the error estimates and weights in the nodes in the previous layer, and this value is used to update the weights of this hidden node by computing an error gradient with respect to the weights in the node [87]. FNNs are able to approximate arbitrary functions, and hence are representationally more powerful than SVM. However, when designing a network, overfitting must be carefully addressed. Also, noise in data could cause training problems, as the model may converge to a local minimum, and the training process might require a long time, limiting practical applications. Another problem with classical FNN is that learning deep networks is very difficult, due to the compounding effect of saturating the sigmoid activation function when backpropagating small errors, which results in very slow convergence. Huge progress in addressing this issue, called the vanishing gradient problem, has been made in recent years. This, together with progress in GPU-based distributed computational infrastructure and availability of very large datasets, has allowed the development of effective deep neural networks, which significantly outperformed all the alternatives in many challenging applications, including computer vision, natural language processing, speech and audio recognition, and healthcare informatics [88]. Many deep learning architectures were proposed to handle various data properties. Some of the established solutions commonly applied to a wide variety of datasets include convolutional neural networks for grid-based data (e.g., imaging) [89] and recurrent neural networks for sequences and temporal data [90].

In power systems, data is often observed over space and time, and therefore, more advanced graph-based structural regression methods are used to exploit structural dependencies. For example, we have used structured learning in Gaussian Conditional Random Fields to assess the risk of insulation breakdown for a given exposure and associated weather threats in a power network [91]. The latest research in deep learning

[88], [92], [93] suggests that a broad range of applications, including structured regression, could benefit from learning latent representations for input data. In our study [94], learning representations for power system substations, based on their spatial proximity, was greatly beneficial for predicting power outages and estimating outage probabilities. In such an approach, nodes of a graph are embedded in a lower-dimensional space, where standard machine learning methods could be more easily applied. The embedding algorithms aim at conserving graph structure and simplifying the learning models by moving away from graph representations. An advantage of using such methodologies is that they can potentially uncover more complex spatial dependencies that include some long-range interactions in addition to influences of the local neighborhood.

The node embedding process represents the original graph in a new feature space, which best-describes the spatial relationships of the nodes in the original graph. This characteristic of the node embedding aims to capture the essential relationships of the original graph structure while simplifying representation to a lower-dimensional list of feature vectors.

There are several algorithms to obtain such an embedding; Two commonly used algorithms are DeepWalk [95] and Node2Vec [96]. Both algorithms rely on community information obtained by random walks, which were used to learn latent space representations. In addition, DeepWalk is able to perform local exploration efficiently and can accommodate small changes in graph structure without global recomputation. Node2Vec is an algorithmic framework that generalizes the DeepWalk process to provide a flexible notion of a node's neighborhood, which allows learning richer representations by effectively exploring diverse neighborhoods. This solution was successfully employed to develop a novel approach to solar radiation forecasting, based on spatial and temporal embeddings using the Node2Vec model for graph data [97]. This approach simplifies the learning models by moving away from complex graphs. The model was developed for forecasts ranging from 3 to 12 hours ahead. The model predicted solar irradiance with very high accuracy in the summer, when there are more clear sky days. During the winter months, the accuracy had a slight drop, but was still high and remained robust even when observational data was missing both spatially and temporally.

B. Engineering Aspects

While big data analytics relies on strong data science foundations, there are also several important aspects for those methods to be used in practice. Interacting with practitioners and those in the industry that try to rip tangible benefits from using data science, one often hears that the actual data science part may only consist 10% of the work, while 90% relates to setting up the workflow, data management (and storage) as well as computing aspects. Therefore, it is of utmost importance here to observe some of the engineering-related aspects of big data analytics. They have been defined as the main challenges for the success of big data analytics [98], [99].

At the core of the concept of big data analytics is the underlying idea that the data to be handled is "big". For an attempt at properly defining big data and its essential features, the reader is referred to [100]. Typical examples relate to the collection of PMU data, as well as high-resolution data at the asset level, (e.g., from wind turbines, Photovoltaic (PV) inverters, smart meters, etc.). The collection rate of these data is at the second to minute time scales, and for many geographical locations simultaneously, while also consisting of many different types of variables. In general, such data in the electricity grid includes point measurements, images, and possibly text. Some of these aspects of big data for power systems, from challenges to applications, were recently covered by [9].

In most of the scientific literature describing electricity applications and beyond, it is assumed that data is available and is of good quality. However cumbersome it may be, ensuring communication of data, ensuring data integrity, as well as assuring data quality, are necessary steps before designing and deploying a data-driven solution [101]. In contrast, those aspects related to data availability and quality have been considered for quite a while by the computer science community, for which a number of methodologies were proposed [102]. Data cleansing and modifications of datasets are then often involved, though one should be aware that these actions may actually affect the original information in the datasets, for instance in terms of its statistical properties [103]. For applications in the electricity markets related to renewable energy, a classical problem for instance is filling gaps in time-series, i.e., there may be periods where data is simply not available, due to failures in logging, storing, or transmitting the data. To fill these gaps, various methods have been proposed, taking advantage of data surrounding that period, data with spatiotemporal dependencies (especially relevant for weather-driven renewable energy generation), data availability at different aggregation levels, as well as physical relationships among variables of interest (e.g., in an optimal power flow problem). For a recent example related to electric load data, see discussion in [104].

Besides these data-related aspects, data-driven approaches used on large datasets require substantial computing power to solve the simulation and optimization problems involved. To centralize the data, these problems can be solved through High-Performance Computing (HPC), which is becoming increasingly common for the electricity grid and electricity market applications [105]. Notable examples include optimal power flow [106] and transmission-switching problems [107]. Solving those large-scale problems with HPC will involve some form of decomposition techniques, which have also become popular in power system operations, markets, and planning problems [108]. However, there may also be a number of applications for which this is neither possible, nor is it desired to centralize the data. In that case, similar approaches may be used to solve these problems in a decentralized manner, though most likely at the cost of increased communication burden due to the inherently iterative nature of distributed optimization approaches. Notable

examples include optimal power flow problems [109] and distributed learning for renewable energy forecasting [110]. Most likely in the future, relevant setups will not be fully centralized or fully decentralized and will be relying on cloud-based, fog and edge computing [111].

Big data analytics is to be placed in the bigger picture of problem-solving. Indeed, in practice, it is only an additional tool to support operations and decision-making. Therefore, before investing in specific big data setups and analytical tools, the problem and related problem-solving approach should be well-defined. For example, if forecasts are there to support decision-making, the type of forecasts (e.g., deterministic or probabilistic) and the forecast products (resolution, normalization, etc.) should be decided upon based on the decision problem at hand. Also, going from data to analytics, there is often the need for an additional layer of extracting the right type and level of information from the raw data. This may be done based on filtering and smoothing, feature engineering, etc. A typical example would be that of event detection based on data streams, to then be used as input to some other analytical problems.

Additional requirements may bring another level of complexity to big data analytics. A crucial requirement linked to the data itself (and related data streams) is how to handle cybersecurity and privacy in the electricity grids. Today, cybersecurity represents a crucial component of future distributed power systems, on which big data analytics may be performed [112]. Consequently, setups for big data analytics, as well as the tools employed, need to be robust to be able to withstand the removal of important data or falsification of data. Also, data privacy is of increasing concern because if the data being collected is shared, one could infer information about specific assets or consumers, which was never meant to be known. Privacy concerns are especially valid now that smart meters are being widely deployed [113], which is potentially allowing one to gain knowledge about consumer-habits for targeted marketing and criminal activities. Another key requirement relates to the need to have interpretable models and outcomes. This has recently triggered a new body of work related to interpretable machine learning and physics-informed machine learning.

C. Decision Making Framework

The use of big data analytics inevitably leads to enhanced decision making. Therefore, in the proposed data analytics, special attention is given to the visualization of the results.

As an example, one approach used in the prediction of outages is to develop risk maps such as the ones shown in Fig. 2. They represent the weather hazard, vulnerability, and the risk calculated as the product of the two

A *risk assessment* framework is formulated by the United Nations Disaster Relief Office (UNDRO), which was explored recently by the United Nations Office for Disaster Risk Reduction (UNISDR)) in their related report [114], was later adopted by the Federal Emergency Management Agency (FEMA) in the USA, since the main focus is on climate-related impacts to infrastructure, society, and environment. This introduces the State of Risk (SoR) as:

$$\text{Risk} = \text{Hazard} \times \text{Vulnerability} \times \text{Consequences} = P(T) \times P(C|T) \times u(C),$$

where P(T) is the Hazard or probability of a given threat intensity (T); u(C) is the Loss (social, economic, or environmental) associated to the level of Consequences (C), associated to the threat intensity (T); and P(C|T) is the Vulnerability or probability of experiencing a consequence

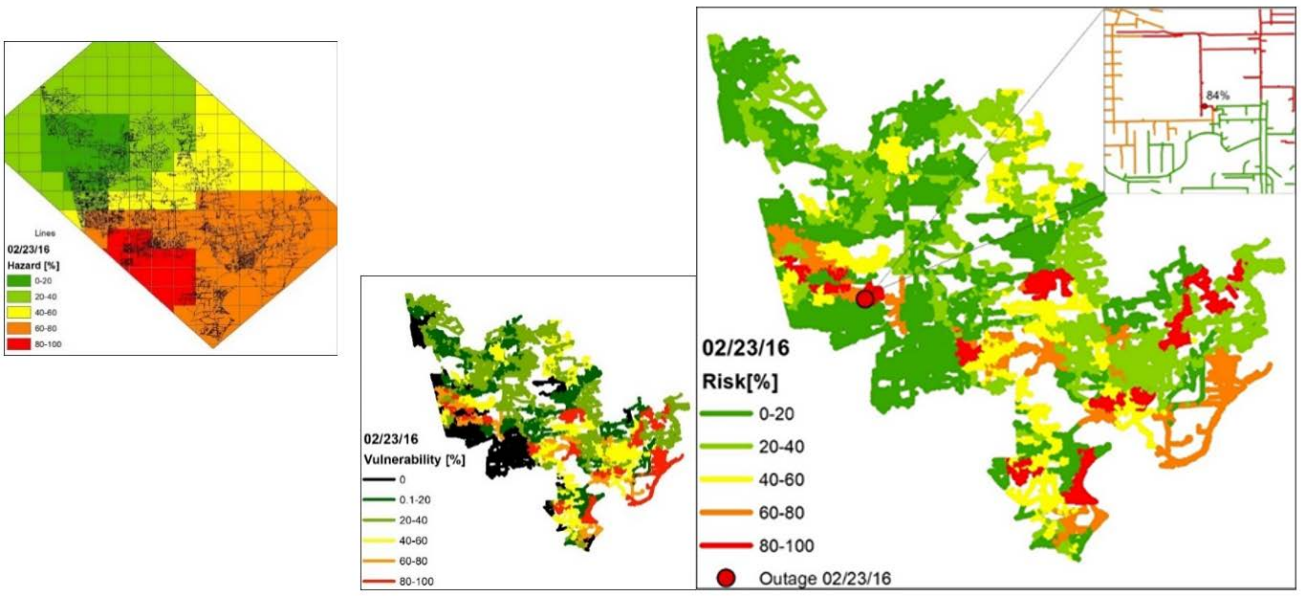

Figure 2: Maps of hazard and vulnerability (left, respectively) resulting in the risk map (right)

level (C) given a threat intensity (T). The risk units are, therefore, expressed in the units of the losses.

In this context, the decision making is related to an optimization process where objective function and constraints aimed at mitigation actions are defined. A broader framework for such decision making for the aforementioned examples of the outage prediction is shown in Fig. 3.

IV. Applications

A. *Asset and Outage Management*

Competitive electricity markets, privatization, and regulatory or technical requirements mandate power utilities to optimize their operation, outage, and asset management practices and develop the requisite decision plans techno-economically. In today's utility practice, asset and outage management are handled by different groups and are viewed as long term vs. short term planning efforts, respectively. We have kept the discussion of such seemingly unrelated issues together to emphasize their close correlation in terms of the use of data, since outage management may utilize the same data as asset management and vice versa since the underlaying status of assets drives both applications:

1) Asset management classification

Asset management in electric power systems can be broadly classified into four main categories based on the possible time scales, i.e., real-time, short-term, mid-term, and long-term [115].

Real-time asset management mainly covers the key power system resiliency principles and deals with the unexpected outages of power system equipment and grid disruptions. By enhancing situational awareness, the electricity grid operators can effectively monitor and control the system. *Short-term* asset management strives to maximize the rate of return associated with asset investments. The value mainly depends on the uncertain market prices through various market realizations. Market risk assessment is a key consideration, and the revenue/profit distributions are gained through a profitability analysis. Optimized maintenance scheduling falls within the *mid-term* asset management. It guides the maintenance plans toward satisfactorily meeting the system-wide desired targets.

The efforts are focused on optimizing the allocation of limited financial resources where and when needed for an optimal outage management without sacrificing the system reliability. Extensive deployment of smart sensors and monitoring technologies is to be used for health and reliability assessment of system equipment over time and to optimize the maintenance plans accordingly [116]. The investment in power system expansion planning, as well as the wide deployment of distributed generations, fall within the scope of *long-term* asset management where the self-interested players, investors, and competitors are invited to participate in future economic plans.

2) Weather impacts on outages

Weather impacts on outages in power systems can be classified into direct and indirect [117]:

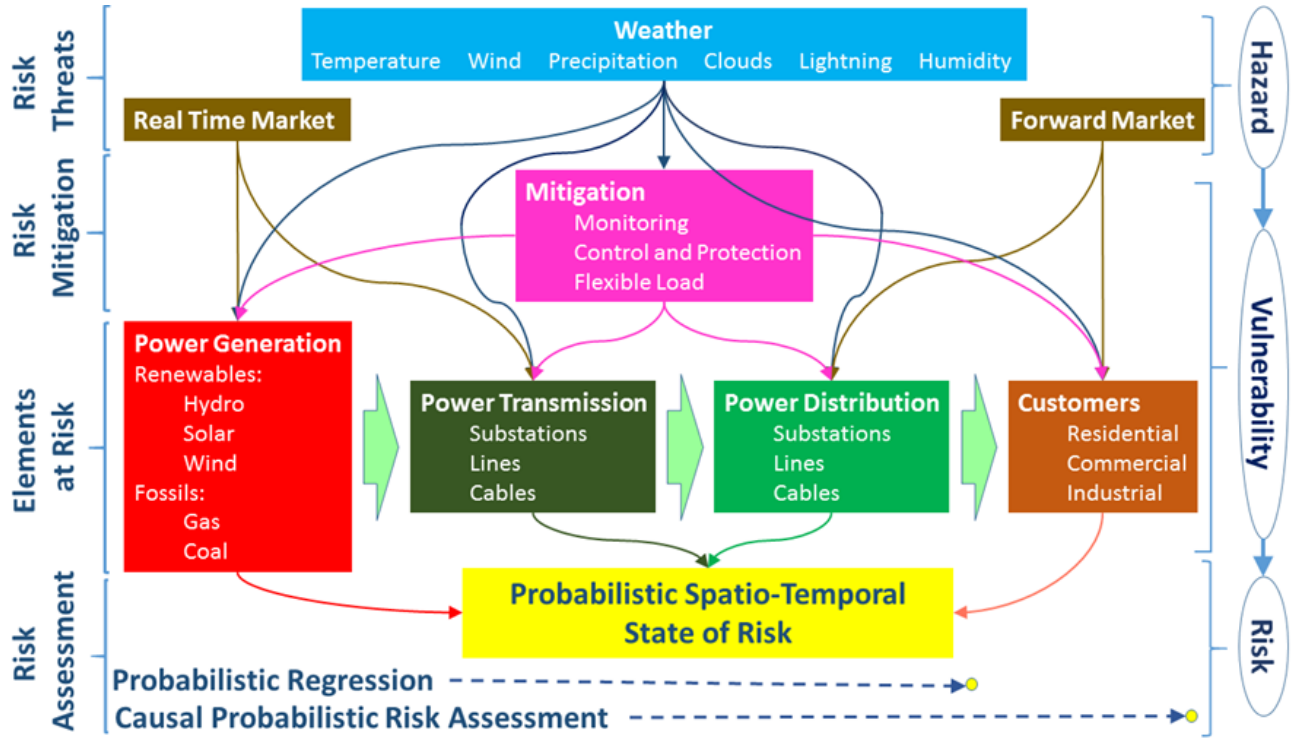

Figure 3: Decision making framework for predictive risk assessment

- Direct impact to utility assets: This type of impact includes all the situations where severe weather conditions directly caused the component to fail. Examples are: lightning strikes to the utility assets, wind impact making trees or tree branches come in contact with lines, etc. These types of outages are marked as weather caused outages.
- Indirect impact to utility assets: This type of impact accrues when weather creates a situation in the network that indirectly causes the component to fail. The examples are: hot weather conditions increasing the demand thus causing the overload of the lines resulting in line sags, increasing the risk of faults due to tree contact, exposure of assets to long term weather impacts causing component deterioration, etc. These types of outages are marked as equipment failure.

3) Outage management background

The ability to track multiple-weather threats synchronously as they develop and to assess associated multiple-consequence impacts to utility industry assets, infrastructure, and the lifelines they support is critical in the utility sector. Electricity grids are spread across wide regions with generation typically located in remote areas. Major consumption in metropolitan areas means that the transmission grid has to bring the power from remote generation sites to the consumption centers, and distribution systems must provide the utility lifelines to the individual customers. To accomplish this, the grid goes through different operating states (Fig. 4) [118]. The corresponding electricity market states are shown in Table III [119]. By combining asset and outage management, one deals with the impacts most effectively [94].

4) Transmission line outage prediction

The knowledge from historical data can be utilized to issue predictions of weather-related transmission outages 1-3 hours ahead. Spatial embeddings are added to the input data set [94] to capture the spatial interdependencies between nodes and events. Consider the example with historical outage data

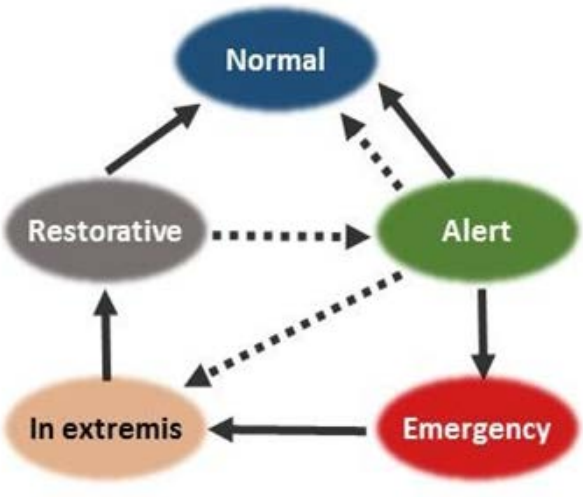

Figure 4: The Electricity Grid States

TABLE III. THE ELECTRICITY MARKET STATES.

Type	Configuration	Market Parameters
Normal	All MPs* complete	Within limits
Emergency	All MPs complete	One or more parameters violate the limits
Restorative	Structure incomplete	Within limits

*MPs (Market Participants) include generator companies, transmission owners, load-serving entities, and other non-asset owners such as energy traders.

collected from Bonneville Power Administration (BPA) [120]. The Automated Surface Observing Systems (ASOS) program [26] data was used to collect the historical weather measurements for the following parameters: Wind Direction [degrees], Wind Speed [knots], Wind Gust [knots], Temperature [F], Dew Point [F], Relative Humidity [%], Pressure [mb], Precipitation/Hour [inch], and Present Weather Codes. The National Digital Forecast Database (NDFD) [121] was used to extract the historical weather forecast data that is used for the testing of the real-time outage probability mapping system utilized in the, insulation coordination study [41].

The optimal placement of line surge arresters is aimed at minimizing the overall risk of lightning-related outages and disturbances while staying within the required budgetary limits [41]. The network and its surrounding impacts are modeled using a multi-modal weighted graph that uses data coming from various sources. The developed risk model considers the accumulated impact of past lightning disturbances to produce a more accurate estimate of insulator strength and predicts insulator performances for the future lightning-caused overvoltage, using Gaussian Conditional Random Fields (GCRF) [122]. The predictive data-driven method for vegetation management in distribution is introduced in [123]. A model for spatio-temporal correlation of a variety of data is developed, which enables real-time analysis of the vegetation impact on the distribution feeders based on predictive risk maps. A prediction algorithm is based on the GCRF regression predictor. The optimization algorithm is used to find the most cost-effective dynamic tree trimming schedule that minimizes the overall risk of the network for each quarter.

5) Transformer health assessment

The traditional approaches for transformer health assessment were developed by using domain knowledge of the physical and chemical processes occurring inside the oil tanks of the transformer, later validated with empirical studies. Some examples are Duval's triangle [124], IEC gas ratios [125], or

the Key Gas analysis [126]. The increasing data (e.g., periodic dissolved gas and oil analysis, sensors that collect real-time data, etc.) collected by electrical utilities motivated the development of data analytics methods based on supervised learning to classify transformer condition and type of fault. Some examples are multi-layered artificial neural networks [127], support vector machine [128], and Bayesian networks [129]. The failures of step-down transformers (22.9KV-220V) used in the distribution sectors in South Korea are studied in [130].

The application of supervised learning algorithms faces the following challenges: (i) the majority of the algorithms offer low interpretability to decision-makers, which is a fundamental requirement in this problem; (ii) lack of available failure data with high quality; and (iii) labeled data about transformer condition in most cases is not available or is defined by a human (i.e., subjective classification of the condition). The use of unsupervised learning is an alternative and appealing solution, but the literature remains limited. The first approach of unsupervised learning was the Health Index described in reference [131] that summarizes the overall health of the asset by combining the results of operational observations, field inspections, and site and laboratory testing into a single index. However, the main limitations of this index are: (i) empirical definition of the weights for each criterion, and (ii) the lack of information about the type of fault. Other alternatives are clustering [132] and semi-supervised learning with Low Dimensional Scaling [133]. Moreover, research in deep learning can contribute to data augmentation [134] of transformer data and deliver new techniques for unsupervised learning like Siamese networks [135]. It is important to emphasize that we should expect a lower performance from unsupervised techniques when compared to supervised learning, but a larger application potential in real-world datasets.

6) Predictions of catastrophic infrastructure damage causing outages

Big data analytics may be used to predict catastrophic asset failures due to inclement weather events such as hurricanes, cyclones, tornados and tsunamis [136], [137]. Such studies are mostly related to the prediction of the infrastructure damage, including the number of toppled poles, destroyed substations, and other damages that require full reconstruction of the electricity grid infrastructure. An implicit assessment is also associated with the outages since a reconstruction is needed first to restore the service.

B. Smart Meter Data Analytics

This application domain has a variety of use cases:

- Applications of a single smart meter analytics
- Applications of groups of smart meters
- Smart meters connected to grid models

Today's penetration of smart meters in the US alone exceeds 50%, and in some European countries over 50%, providing important opportunities for data analytics to enhance customer management and grid operations and planning. Smart meters provide readings of energy, power, and voltage at temporal granularities typically of one hour or 15 minutes. Electric utilities, energy services providers, customers, and researchers have identified numerous use cases for smart meter analytics such as forecasting, customer load profiling and classification, load estimation, and enhanced grid modeling [40]. When combined with other data sources and utility systems, smart meter analytics can further expand its benefits to utility operations enterprise-wide. A summary of smart meter application is illustrated in Fig. 5.

Below, we provide a summary of the salient applications of smart meter data analytics with corresponding references. Some of the applications will be expanded in the following sections under separate titles.

1) Load Forecasting

The power industry has been using load forecasting for grid operation and planning and for customer management. Smart meter data caught the attention of researchers in the past decade for both point and probabilistic load forecasting [138], [139]. A more comprehensive treatment of this topic is in the subsection.

2) Customer Load Profiling

Load profiling refers to the classification of the historical readings of customer demand into groups based on their behavior. Clustering techniques such as k-means, hierarchical clustering, and self-organizing maps have been utilized for load profiling [140]-[142]. Time-varying models combined with clustering have been utilized to develop complex power load modeling using Advance Metering Infrastructure (AMI) data [143].

3) DER Analytics

A significant amount of distributed energy resources (DER) are being connected to the grid, including solar PV, energy storage, and electric vehicles. These resources within the grid create new challenges for utility providers including voltage variability [144], thermal limit violations, reverse flows, and impacts on the expected life of the infrastructure such as transformers and voltage regulators. It is crucial for utilities to have accurate information related to Distributed Energy Resources (DERs) at the distribution circuit and behind-the-meter (BTM). Researchers have demonstrated the possibility of detecting rooftop PV [145] and electric vehicles charging at the consumer end [146] using non-intrusive analytics on smart meter data. Deep neural networks have been utilized to detect size, tilth, and azimuth parameters of solar PV installations based on AMI data [147].

4) Grid Applications

Smart meter data can be utilized in conjunction with distribution feeder data to obtain refined models for distribution planning, or to obtain insight into specific modeling problems. For instance, smart meter data has been utilized for anomaly detection [148] such as drastic changes in demand or voltage. Smart meter outage data has been utilized for analytics and optimal outage restoration [149] and feeder reconfiguration design. Smart meter analytics has also been

used for transformer connection correction [149], phase identification [150], topology identification [151], and parameter estimation [152].

C. Load Forecasting

In a utility analytics survey conducted in 2017 among 136 utilities from 24 countries, 52% of the respondents considered energy forecasting as a high-priority application, the highest percentage among all applications in the survey [153]. Forecast improvement can lead to better operational and planning decisions and thus to monetary savings or system reliability enhancement. Depending upon the factors such as the size of the company and the magnitude of error reduction, forecast error reduction may result in annual savings up to millions of dollars [154]. In the big data era, the growth of data, advancement of computing technologies, and breakthroughs in advanced analytics further stimulate the improvement of energy forecasting techniques and methodologies. Many of these recent developments were recognized as winning entries in the Global Energy Forecasting Competitions (GEFCom) [42]-[44].

Utilities have been practicing load forecasting for over a century [155]. Following the growth of the electric grid footprints and the power industry, long-term load forecasting, or spatial load forecasting, in particular, has been considered as a crucial component to power systems planning in the late 20th century to early 2000s [156]-[158], when the load data were collected from distribution transformers and in low resolution, such as monthly or annual. On the other hand, the increasingly sophisticated operational needs started to require accurate short-term load forecasts [159]. Artificial intelligence techniques, such as artificial neural networks, took the majority of the literature in the 1990s [160]. Although many models were developed for Short-term Load Flow (STLF), most were of little practical value. A notable success developed from academia in the 1990s was a neural network model, which was later commercialized and is still being used in the industry today [161]. Another recent academic research discovery that has been commercialized and deployed worldwide during the 2010s is a regression-based modeling framework [39], [162].

As distribution automation and smart grid technologies made high spatiotemporal resolution data available to load forecasters, the research in load forecasting flourished too. For instance, retail electricity providers started to use hourly data for long-term load forecasting [163]. Some market operators and utilities also relied on hourly or sub-hourly data to forecast load at a high voltage level [38], [39], [164]. These high-resolution load data allow forecasters to build models with hundreds of parameters that can capture many salient features in the load [37]. They also enabled the load forecasters to improve aggregate forecasts by leveraging meter level load information [138].

While exploiting the use of high-resolution load data, researchers and practitioners also invested some efforts on

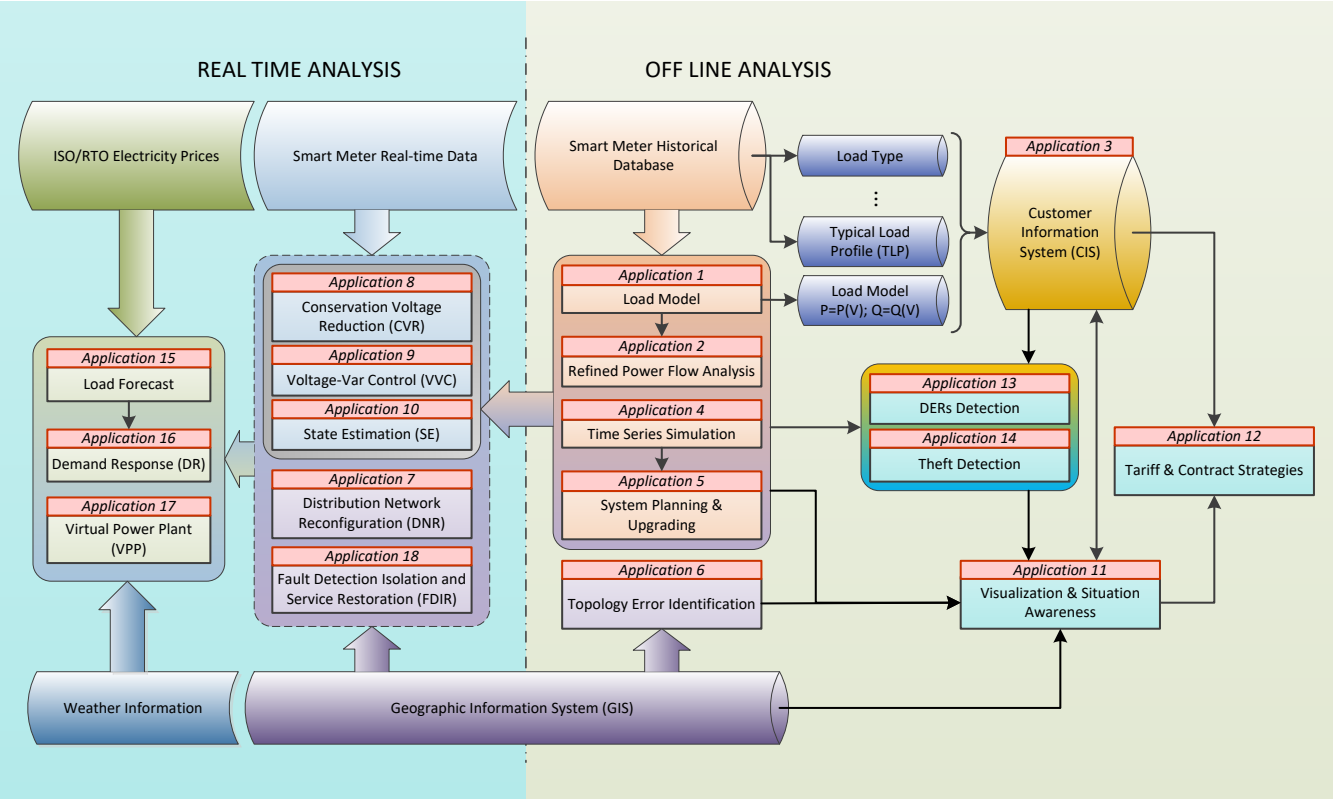

Figure 5. Illustration of Smart Meter Data Analytics Applications

innovative ways of leveraging weather data. Traditionally, only a small number of weather stations are being used to forecast load in a specific region. In GEFCom2012, temperature series from 11 weather stations were released to the research community [42]. A weather station selection methodology was then proposed and shown to add value to state-of-the-art load forecasts. The method was then used by a winning team in GEFCom2014 [165]. Recently, it was further improved by other researchers [166], [167]. Since weather data can be used for many load forecasting models, these weather station selection methods are beneficial to both long- and short-term load forecasting.

While the main research trend of load forecasting research is to leverage big data, another trend is to embed comprehensive information about the future into load forecasts. Probabilistic load forecasting, which provides forecasts in quantile, probability interval, or density function, is certainly a hot topic in the past decade [168]. These probabilistic load forecasts can be generated via simulating residuals [169], [170], generating input scenarios [171], or combining point forecasts [172]. Quantile regression has also been adopted to produce probabilistic load forecasts [139] [172], [173]. Some methodological aspects of forecasting have also been studied specifically for probabilistic load forecasting, such as feature selection [174], [175] and forecast combination [176].

In addition to improving load forecasts at high voltage levels, researchers and practitioners have also devoted many efforts to load forecasting at medium and low voltage levels. Some of them are at a delivery point level [42], [44], while others are at individual meters [40], [139], [177]. Another popular topic is to forecast building level load, ranging from academic buildings [178], [179] to residential buildings [180], [181]. While the weather still plays a major role in many building level load forecasting models, its effects on industrial load are rather minimal. Load forecasting for factories and industrial plants is another emerging topic in this category [182]-[184], and includes reactive power load forecasting.

D. Renewable Energy Analytics and Forecasting

The deployment of renewable energy generation capacities has continued at a sustained pace over the past decade or two. Owing to the variability of power generation from renewable energy sources, and also due to the limited predictability, the emphasis has been placed on developing approaches to integrate those renewables. Besides aspects related to grid code, and some of the analytics and novel energy management approaches covered in other sections, a large part of the efforts has been on how to optimally use the wealth of data available to improve knowledge about renewable power generation, most often with a view on forecasting. One of the first papers proposing dynamic models for predicting wind speed and corresponding power generation is [185]. Even though focusing on simple models and a fictitious setup, that manuscript laid the groundwork for a wealth of subsequent developments. Obviously, at the time, big data aspects were not discussed, and the dimensionality of the models involved was small.

Today for wind farms, especially offshore, it is standard to collect data at the turbine level at a one-second resolution. Similarly, for solar power plants, data can be collected at the inverter level, and with a similar second-level resolution. Those data at a very fine level are to be leveraged to improve analytics and forecasts at the wind farm (or solar power plant) level [186]. In addition, since renewable energy generation capacities start to be numerous and geographically dispersed in a dense manner, one may also accommodate all the data collected at the site levels to improve forecasts [14], [18], [187].

To this should be added a wealth of other data sources of relevance, mainly related to meteorological observations and forecasts, which describe complex processes and yield very large data volumes. On the side of meteorological observations:

(i) sky imagers have shown great potential for high-resolution modeling and forecasting of solar power generation since they are tracking moving clouds and their impact on solar panels [188]. They may give an image of the sky above the solar power plants every 30 seconds;

(ii) weather radars have similarly demonstrated their interest in appraising and modeling dynamic regimes for application to high-resolution wind power prediction [189]. Depending upon technology, radar images may be available between every minute and every 10 minutes, with an image radius between 60 and 250 kms;

(iii) LIDARs are increasingly seen as highly relevant for wind measurements upwind of wind turbines and integration in forecasting methodologies [190], or more generally as new potential observations of wind profiles to be used in weather and renewable energy prediction [191]. LIDARs provide wind measurements for the cone they scan (vertically or horizontally, depending on the way they are set-up) every few seconds.

One could additionally mention satellite images, of potential interest for wind, solar, and wave energy. Their lower frequency of update makes them less relevant for the time being though. The information from these various types of devices is referred to as remotely sensed information.

A first and complex challenge when handling remotely sensed information is dimension reduction. This may be performed: (i) based on statistical and signal processing techniques, e.g., Independent Component Analysis–ICA, for motion fields in weather radar images; (ii) by extracting physical features like clouds in sky images [188] or precipitation systems and their characteristics in radar images [192]; or finally (iii) through functional models like wind profiles for LiDAR vertical measurements.

Besides meteorological observations, new high-dimensional input data may also take the form of weather forecasts. First of all, relevant information in weather forecasts

may not only be for the closest point to a site of interest but may be provided over the whole area of that site of interest [59]. Secondly, to express forecasts in a probabilistic manner, ensemble weather forecasts consist of a set of alternative and equally likely trajectories (typically, between 10 and 100) for relevant weather variables. They can be interpreted as sample realizations from multivariate probabilistic forecasts to feed in renewable energy forecasting approaches. These are available over large areas. e.g., all of Europe, providing information on the multivariate space-time dependencies in renewable power generation [193], [194]. Finally, very high resolution (spatial resolution in the order of tens to hundreds of meters and temporal resolution in the order of seconds to minutes) weather forecasts are bringing new opportunities for renewable energy forecasting, as for the recent example in [186] and applications at offshore wind farms.

Today, the availability of such quantities of data calls for fundamental changes in renewable energy analytics and forecasting, both in terms of the methods involved but also in terms of the business models. We expect to see many innovative works appearing in the future proposing approaches based on stochastic differential equations, deep learning, distributed and federated learning as well as data markets.

E. Energy Optimization and Efficiency

Smart grid technologies offer a large potential to boost energy efficiency in different sectors. However, presently, energy efficiency actions are mainly confined to the implementation of ISO 50001 certification as follows: (i) install additional equipment (meters, sensors, etc.) to measure energy consumption; (ii) install new hardware and replace equipment; and (iii) visualize the data in a seamless user interface, find anomalous patterns; and identify energy-intensive processes. This standardized practice provides monitoring and awareness of energy consumption to human decision-makers, but it does not enable prescriptive analysis and autonomous process control.

Different works in the literature explore model-driven energy optimization techniques, such as integer programming for peak load reduction in steel-plants [195] and mixed-integer linear programming for thermal domestic appliances [196]. Model-driven approaches have the disadvantage of requiring a (mathematical) model of the physical process and constraints, which might be complex to obtain in some cases. They do not allow continuous improvement of control policies.

The advent of internet-of-things technology offers technical conditions for data-driven modeling in energy optimization. Some examples are the use of batch reinforcement learning (fitted Q-iteration) for controlling a cluster of domestic electric water heaters for demand response services [197]; fitted Q-iteration combined with auto-encoders for energy optimization in electric water heaters [198]; deep reinforcement learning for predictive energy optimization of wastewater pumping stations [199]; and tree-based modeling of building energy consumption for optimal heating system scheduling [200]. This approach does not require full modeling of the process equations since its understanding is made in real-time through data. However, data availability and the time (e.g., number of interactions with the physical system) required to "train" data-driven optimization methods remain as practical challenges for industry adoption.

Besides data-driven energy optimization, additional data-driven energy services can be offered by energy service companies (ESCOs), aggregators, and retailers to their customers aimed at maximizing end-user awareness to energy efficiency actions and extracting business value from data.

An important contribution to increase end-user awareness is non-intrusive load monitoring from smart meter data, which can be applied to detect and estimate residential PV installations [145], heat pump consumption [201], and individual load profiles from feeder load curves [202]. Customer segmentation with load profiling can be applied by aggregators and retailers to identify target customers, design tailor-made dynamic or real-time tariffs [203], model dynamic behavior of controllable loads/appliances [204], or inform consumers if they are facing an abnormal change of their load profile. The standard techniques are batch time series clustering, but online clustering is a fundamental requirement due to the dynamic nature in the consumption behavior [205].

Smart meter data combined with exogenous variables (e.g., outdoor temperature) can support retailers and aggregators to estimate the demand response potential of their customers. In [206], a stochastic knapsack problem is formulated for customer selection in DR programs using consumption data, as well as to estimate the probability of achieving a load reduction target. Causality inference between Distributed Resource (DR) tariffs and consumption is used in [207] to estimate consumers' elasticity and to identify if dynamic tariffs influence the consumers' usual consumption diagram; a similar goal is attained with a correlation-based approach [208]. Moreover, online learning can be used to dynamically adjust price signal to obtain a desirable usage behavior, e.g., by formulating an online convex optimization problem [209] or via a parametric utility model [210].

The data collected from energy efficiency audits and certification is very valuable for different stakeholders and services [211]. For instance, the Department of Energy (DOE) Buildings Performance Database can be used for different goals [212]: energy efficiency score benchmarking of different building types and geographical location; estimate energy savings potential associated to specific retrofit actions; and portfolio-level impact assessment of energy technologies. In fact, data-driven techniques can be used for an *a priori* assessment of retrofit measures in terms of energy savings and "de-risk" investments in energy technologies. For instance, multi-linear regression can produce a probabilistic estimation of the return-on-investment associated, with different retrofit measures, considering the building's characteristics and systems [213].

F. Synchrophasor Data Analytics and Event Analysis

To appreciate the importance of Phasor Measurement Unit (PMU) data, one has to go back to the reports on the causes for

the major blackout in the Northeast USA on August 13, 2003 [214]. The U.S.-Canada Power System Outage Task Force has concluded that one of the main reasons for the occurrence of the blackout was the lack of situational awareness. This, in turn, was attributed to the limited operator view of the power system events and associated dynamics enabled at the time by the by Energy Management System (EMS) through the field measurements provided by the Supervisory Control and Data Acquisition (SCADA) systems and substation recording devices such as digital fault recorders, digital relays, and sequence of events recorders. The key deficiency was the measurement reporting rate (SCADA Remote Terminal Units), lack of differentiated time stamping (SCADA database), and inability to timely update the power system model to reflect cascading switching events (state estimator). In addition, the fault recording devices were not adequately time-synchronized and time-stamped. To remedy such shortcomings, the PMUs and synchrophasor-based Wide Area Monitoring, Protection and Control (WAMPAC) were identified as an adequate measurement infrastructure. With a boost in funding from the American Recovery and Reinvestment Act of 2009, many PMUs were installed across the USA reaching more than 2,500 units today. An even larger number of PMUs were installed in China, and plans for a large number of installations are underway in India.

One of the main strengths of PMUs is that they provide time-synchronized measurements of real-time voltage and current phasors at much higher reporting rates compared to SCADA Remote Terminal Units (RTU). Insight into how this is done can be obtained from Fig. 6 where the PMU generic architecture is shown [215]. The key characteristics of PMU are the one pulse per second (1PPS) and the time-code provided by the Global Positioning Satellite (GPS) clock receiver. The accurate clock signal at the rate of 1PPS feeds the sample and hold (S/H) circuits at all PMUs, enabling synchronous sampling of input waveforms across not only the inputs of one PMU but inputs of all PMUs installed in a given power system. The time-code representing absolute time adds an ability to time-stamp all reference times. Such calculated phasor values are reported at a rate per second of 30, 60, 120, or even higher, giving two key advantages: a) providing streaming measurements of high-fidelity data, and b) assuring

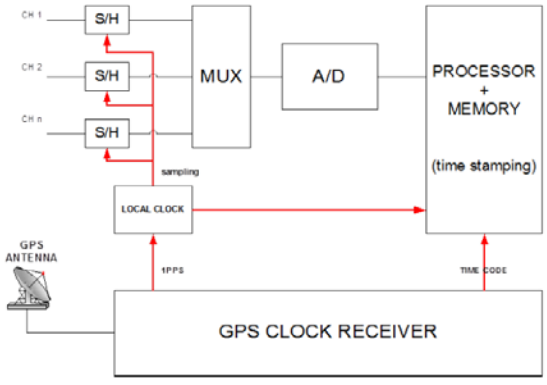

Figure 6: Generic architecture of a PMU [215]

high accuracy calculation of the phase difference between any location in the system and the reference phasor.

This provides a unique opportunity for post-event analysis, especially in the case of complex events (for example, cascading faults caused by equipment failures leading to blackouts). In addition, capturing and analyzing oscillations in a power system is important since it announces a possibility of system collapse [216]. The PMUs play an important role in disturbance recording [217] with the recent development of disturbance identification and classification technologies [217]-[220]. In recent years, power system capabilities have been extended to include renewable resources, increased energy demand, electric vehicle integration, etc. All of these technologies impose novel challenges to the system operation. PMU data provide a valuable source that could help meet the challenges and increase the resilience of the composite grid.

Thus, the data collected from PMUs plays a vital role in applications such as system monitoring, control, protection, state estimation, stability assessment, and fault detection [221]. However, the post-event analysis is still performed manually in many cases with limited or no capability of prediction. The proposed techniques embedded in the mentioned software solutions are aimed at automated analysis capable of not only classifying past events but also predicting future contingencies using the records of streaming data.

The majority of predictive methods for PMU data analysis in literature are focused on enabling more meaningful situational awareness than what is covered by EMS SCADA, in turn assuring dynamic stability of the system [222]-[229]. The prediction methods can extend the operators' capability to differentiate types of events, ranging from normal operation to operation in extremes by capturing PMU waveform features and automatically identifying events during the classification process. This enables the applications of big data, AI, and machine learning to help in predicting multiple types of alert and emergency events in power systems, in addition to tracking normal operation and dynamic stability extremes. The initial insight can be used as the guide for future research in this area and potentially open the door for a more thorough exploration of prediction algorithms based on PMU for other applications, such as asset management and monitoring, or outage management and prediction.

One challenge that all entities involved with the development of online and offline applications are facing is the proliferation of an extremely large amount of data coming from PMUs. To improve the systems reliability, security, and efficiency, automated tools will need to be developed that are capable of both analyzing past events and informing decisions in real-time. This means that the proposed approaches need to differentiate between techniques suitable for accessing and processing a large amount of historical data where the processing time is less important vs the techniques aimed at real-time processing of streaming data where the computational efficiency is crucial.

As computational resources advance, learning and extracting useful patterns from big data creates new

opportunities. Deep learning [88], [92], [93] allowed for finding more abstract patterns from big complex, and even heterogeneous data [230]. A deep model learns in a multi-layered fashion. Each time the new data or an excreted set of features is passed through a layer, an additional level of data abstraction is introduced. Thus, through many consecutive layers, deep models learn richer, high-level representations of low-level raw data such as images, sound, and text. In other words, deep models provide an end-to-end framework for automated, complex, feature extraction at a high level of abstraction. The deep models do not extract predefined representations—on the contrary, they tend to find invariant patterns by removing variation in data [231]. The ultimate hypothesis in this regard suggests that the more data there is, the more knowledge is being extracted [232], and thus the greater the generalization capability that can be achieved. The learned representations are compact, which requires less computation and, thus, makes further learning quite efficient.

The overall architecture of the PMU data automated analysis process is given in Fig. 7.

G. Grid Applications

The investment in Smart Grid technologies (e.g., smart meter, PMU, and intelligent electronic devices [IED]) for distribution and transmission grids are enhancing their monitoring and control capabilities [233]. In parallel, new market and regulatory frameworks are being tested (at pilot level) and implemented in different countries to support the integration of flexible DER.

1) Observability

Despite all the technological advances, observability of low voltage (LV) grids remains a major bottleneck to fully explore the potential from these technologies and integrate flexible DER to electricity markets and power system management. This can be divided into two main challenges: (i) topology and grid's parameters characterization; and (ii) real-time monitoring.

Information from smart meters installed in LV customers and in secondary substation feeders can be used to construct grid topology using a variety of methods, such as: probabilistic graphical model and LASSO linear regression [234], [235]; and power flow combined with mutual information [151]. In some cases, information about grid topology is available, but the electrical parameters exhibit gross errors or are unknown (i.e., parameters from cable catalogues are used), and thus, it is necessary to conduct a robust (and data-driven) estimation of grid parameters [236], [237]. This problem also requires the development of methods for planning sensor placement and minimization of monitoring costs [238]. Another source of uncertainty is the connection phase of each customer. Clustering 15-min voltage magnitude measurements from smart meters can identify groups of customers connected to the same phase and reduce the workforce cost for phase identification in the field [239]. Smart meter data can also be used to estimate the rated power of behind-the-meter PV panels [240] and detect electric vehicles charging [241].

In terms of real-time monitoring, information about voltage profiles is very important to generate alarms to human operators in LV dispatch centers. However, even with communication protocols such as Power Line Communication (PLC) PRIME or General Packet Radio Service (GRPS) either is technically unfeasible to collect every 15-min information from hundreds of customers or the communication costs are very high. Data-driven state estimation functions help to estimate the voltage magnitude in every LV node in real-time using only information from a subset of smart meters (i.e., 10-20%) with real-time communication [242]. Load and generation forecasts can be used as input in a standard state estimation algorithm and produce a probabilistic estimation of the nodal voltage [243].

Other voltage levels are also benefiting from data-driven state estimation functions for real-time monitoring. Some examples include the following: a robust data-driven state estimation was proposed in [244] to exploit historical data collected by substation meters and PMUs, by finding system similarities in a supervised learning framework with kernel ridge regression; a dynamic state estimator with forecasting capability is described in [245]; the combination of SCADA data from primary substations (normally available in central databases with a few seconds/minutes delay) and secondary substation (normally available in central databases with a delay of hours) can extrapolate quasi-real-time operating conditions of Medium Voltage (MV) grids [246], [247]; and a mosaic of local competitive auto-encoders estimates the status of grid switchers based on a set of local electrical measurements [248].

Furthermore, traditional functions, such as outage detection, can benefit from data analytics functions that integrate textual, temporal, and spatial information from social media [21] and PMU data [218].

2) Controllability and decision-support

Data-driven functions are also being integrated in traditional energy management (EMS) and distribution management systems (DMS) to complement or replace classical grid analytical functions such as power flow, network reconfiguration, optimal power flow, etc. In some cases, a machine learning model is used as a proxy for traditional functions, e.g., to estimate power flow of a Jacobian matrix with PMU data [249] and unit commitment computations [250]. A potential disadvantage of this approach is that a large dataset needs to be collected or simulated to fit the models. However, after being fitted, it can be integrated into different applications and provide fast estimations.

In other cases, machine learning is used to learn from (or imitate) historical data of human operators control actions and decision-aid [251] or to explore new (and better) solutions by using expert systems [252]. An important advantage of these two approaches, developed for Transmission System Owners (TSOs), is its high interpretability to human decision-makers and the capacity to exploit expert knowledge (i.e., past decisions and control heuristics). However, both fall into the imitation learning paradigm i.e., supervised learning applied to decisions from an expert, and do not explore (or search for)

new solutions. An alternative is reinforcement learning, which provides a trade-off between exploitation and exploration and in some cases, like the case described in [253], can be combined with imitation learning.

Some use cases for this data-driven control approach are: fast ranking of higher-order contingencies (according to their risk) to better prioritize power systems simulations [254]; and causality analysis on measurement data to implement optimal node attack strategies [255].

Finally, the integration of renewable energy systems and the wide adoption/implementation of IEC 61850 is increasing the volume of alarm data and the number of alarms that require attention and control actions in control rooms [256]. Traditional approaches are based in rule-based expert systems [257], but without the capacity to provide valuable insights into the information contained in an alarm sequence and reduce the cognitive load of human operators. In [258], an unsupervised rough classification technique is proposed to reduce the volume of substation data and messages received during emergency scenarios and improve decision-making, but do not provide suggestions of control actions. This last step of decision-making (i.e., define a sequence of manual maneuvers by the operator) is missing in most approaches that handle data from protection schemes and is fundamental to generate the actual impact from data analytics functions, such as reduce time to make the first decision and System Average Interruption Duration Index (SAIDI).

V. Future Opportunities and Challenges

The survey indicates many applications that are at different stages of practical implementation. While the number of references on the general subject of big data is already large, it is expected that the work will be progressing at an unprecedented pace, and the number of references will grow even more in the near future. However, it is not clear in some of the published surveys, what the end applications of some of the surveyed techniques are, and what are the associated benefits and business drivers. We see the benefit of future research in its focus on end-applications while answering the fundamental issues at the same time to advance the following opportunities eventually;

- Predicting events ahead of time and allowing mitigation strategies to be implemented and risk quantification with uncertainty forecasts to be calculated
- Monetizing historical data for the benefits of owners and users
- Combining physical and data models for improved root-cause analysis
- Preventing power system outages and operational constraints currently costing billions of dollars
- Making the users aware of upcoming electricity supply constraints and utilizing distributed resources more effectively to mitigate outages and other contingencies
- Spurring innovation by utilizing cross-discipline experiences from seemingly different domains but with a strong correlation of data properties and analytics requirements
- Utilizing data from many disparate ubiquitous sources such as tablets, smartphones, and other personal electronic devices making the owners and stakeholders of the data analytics enterprise the participating actors
- Replacing some of the tasks performed by operating agents and experts today by algorithms and automated processes

At the same time, many challenges need to be overcome:

- Lack of real-world data to make future studies more practically focused
- Lack of open data and benchmark models for different use cases
- The use of synthetic data may have to be carefully evaluated for yielding any meaningful outcomes
- Agreeable metrics for evaluating big data analytics' results for their validity and impact (e.g., return-on-investment of big data technologies)
- Offer model's interpretability and sufficient accuracy to decision-makers
- Availability of open-source platforms for data management and data analytics implementation
- Limited viewing capabilities for large amounts of data points in scalable data models
- Consideration of computational requirements to find an optimal trade-off between centralized and decentralized computing
- Cybersecurity and privacy of the data management in data analytics enterprises.

VI. Conclusions

This survey has strived to achieve several goals:

- It addressed both the breadth and depth of practical big data analytics application in the electricity grid.
- It made a review of the key issues in implementing big data analytics in the selected grid domains.
- It gave ample examples of recent trends in the decision-making framework and predictive analytics.
- It enumerated a number of references that may be used by researchers to further their own research.
- It gave some direction of future research and outlined challenges and opportunities.

The paper is by no means an exhaustive account of all the published works or trends. If interested in the subject matter, the readers of this survey should explore additional aspects not mentioned here.

Acknowledgment

Special thanks are extended to Mr. Milad Soleimani, a Ph.D. student in Dr. Kezunovic's lab, for tirelessly helping with paper formatting and managing references during multiple paper revisions. Additional acknowledgments are due to a number of reviewers who provided a set of

comments and suggestions that significantly contributed to improving the paper.

REFERENCES

[1] M. Kezunovic, L. Xie, and S. Grijalva, "The role of big data in improving power system operation and protection," in *2013 IREP Symposium Bulk Power System Dynamics and Control - IX Optimization, Security and Control of the Emerging Power Grid*, Rethymno, 2013.

[2] M. Ghorbanian, S. H. Dolatabadi, and P. Siano, "Big data issues in smart grids: A survey," *IEEE Systems Journal*, vol. 13, no. 4, pp. 4158-4168, Dec. 2019.

[3] Y. Zhang, T. Huang, and E. F. Bompard, "Big data analytics in smart grids: a review," *Energy Informatics*, vol. 1, no. 8, pp. 1-24, 2018.

[4] M. Kezunovic, Z. Obradovic, T. Dokic, B. Zhang, J. Stojanovic, P. Dehghanian, and P.-C. Chen, "Predicating spatiotemporal impacts of weather on power systems using big data science," in *Data Science and Big Data: An Environment of Computational Intelligence*. Heidelberg: Springer, 2017.

[5] H. Akhavan-Hejazi and H. Mohsenian-Rad, "Power systems big data analytics: An assessment of paradigm shift barriers and prospects," *Energy Reports*, vol. 4, pp. 91-100, Nov. 2018.

[6] L. Duchesne, E. Karangelos, and L. Wehenkel, "Recent Developments in Machine Learning for Energy Systems Reliability Management," *Proceedings of the IEEE*, in press.

[7] H. Jiang, K. Wang, Y. Wang, M. Gao, and Y. Zhang, "Energy big data: A survey," *IEEE Acccess*, vol. 4, pp. 3844-3861, 2016.

[8] M. Ghofrani, A. n. Steeble, C. Barrett, and I. Daneshnia, "Survey of big data role in smart grids: Definitions, applications, challenges, and solutions," *The Open Electrical & Electronic Engineering Journal*, vol. 12, pp. 86-97, 2018. [Online]. Available: www.benthamopen.com/TOEEJ/.

[9] R. Arghandeh and Y. Zhou, *Big data application in power systems*. Elsevier Science, 2017.

[10] A. F. Zobaa and T. J. Bihl, *Big data analytics in future power systems*. CRC Press, 2018.

[11] A. S. Costa, A. Albuquerque, and D. Bez, "An estimation fusion method for including phasor measurements into power system real-time modeling," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1910-1920, 2013.

[12] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. of Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012, pp. 2222-2230.

[13] Z. W. H. Sun, J. Wang, Z. Huang, N. Carrington, J. Liao, "Data-driven power outage detection by social sensors," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2516 – 2524, Sept. 2016.

[14] L. Cavalcante, R. J. Bessa, M. Reis, and J. Browell, "LASSO vector autoregression structures for very short-term wind power forecasting," *Wind Energy*, vol. 20, no. 4, pp. 657–675, 2017.

[15] Q. Zhu, J. Chen, D. Shi, L. Zhu, X. Bai, X. Duan, and Y. Liu, "Learning temporal and spatial correlations jointly: A unified framework for wind speed prediction," *IEEE Transactions on Sustainable Energy*, 2019.

[16] A. Leitão, P. Carreira, J. Alves, F. Gomes, M. Cordeiro, and F. Cardoso, "Using smart sensors in the remote condition monitoring of secondary distribution substations," in *Proc. of the 23rd International Conference on Electricity Distribution (CIRED 2015)*, Lyon, France, June 2015.

[17] C. Sweeney, R. J. Bessa, J. Browell, and P. Pinson, "The future of forecasting for renewable energy," *Wiley Interdisciplinary Reviews: Energy and Environment*, vol. 9, no. 2, art. no. e365, 2020.

[18] J. W. Messner and P. Pinson, "Online adaptive lasso estimation in vector autoregressive models for high dimensional wind power forecasting," *International Journal of Forecasting*, vol. 35, no. 4, pp. 1485-1498, 2019.

[19] E. Ahmed, I. Yaqoob, A. Gani, M. Imran, and M. Guizani, "Internet-of-things-based smart environments: state of the art, taxonomy, and open research challenges," *IEEE Wireless Communications*, vol. 23, no. 5, pp. 10-16, October 2016.

[20] A. Moreno-Munoz, F. J. Bellido-Outeirino, P. Siano, and M. A. Gomez-Nieto, "Mobile social media for smart grids customer engagement: Emerging trends and challenges," *Renewable and Sustainable Energy Reviews*, vol. 53, pp. 1611-1616, 2016.

[21] H. Sun, Z. Wang, J. Wang, Z. Huang, N. Carrington, and J. Liao, "Data-driven power outage detection by social sensors," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2516 – 2524, Sept. 2016.

[22] *COMMISSION REGULATION (EU) No 543/2013 of 14 June 2013*, O. J. o. t. E. Union, 15.6.2013.

[23] L. Hirth, J. Mühlenpfordt, and M. Bulkeley, "The ENTSO-E transparency platform – A review of Europe's most ambitious electricity data platform," *Applied Energy*, vol. 225, pp. 1054–1067, Sep. 2018.

[24] F. Ziel and R. Weron, "Day-ahead electricity price forecasting with high-dimensional structures: Univariate vs. multivariate modeling frameworks," *Energy Economics*, vol. 70, pp. 396-420, Feb. 2018.

[25] M. Zugno, P. Pinson, and H. Madsen, "Impact of wind power generation on European cross-border power flows," *IEEE Transactions on Power Systems*, vol. 28, no. 4, pp. 3566-3575, Nov. 2013.

[26] Automated surface observing system (ASOS). National Centers for Environmental Information ,National Oceanic and Atmospheric Administration. [Online] Available: https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/automated-surface-observing-system-asos

[27] National Oceanic and Atmospheric Administration. (2017). Satellite Data. [Online] Available: https://www.ncdc.noaa.gov/data-access/satellite-data

[28] National Oceanic and Atmospheric Administration. Radar data in the NOAA big data project. [Online] Available: https://www.ncdc.noaa.gov/data-access/radar-data/noaa-big-data-project

[29] Vaisala, "National lightning detection network – Technical specification," 2017. [Online]. Available: http://www.vaisala.com/en/products/thunderstormandlightningdetectionsystems/Pages/NLDN.aspx.

[30] National Weather Service. (2017). National digital forecast database (NDFD) Tkdegrib and GRIB2 DataDownload and ImgGen tool tutorial. National Oceanic and Atmospheric Administration (NOAA). [Online] Available: https://www.weather.gov/media/mdl/ndfd/ndfd_tutorial.pdf

[31] TNRIS maps & data. Texas Natural Resources Information System (TNRIS). [Online] Available: https://data.tnris.org/

[32] TNRIS lidar data. Texas Natural Resources Information System (TNRIS). [Online] Available: https://data.tnris.org/

[33] T. Hong, P. Wang, and L. White, "Weather station selection for electric load forecasting," *International Journal of Forecasting*, vol. 31, no. 2, pp. 286-295, April-June 2015.

[34] J. Xie, Y. Chen, T. Hong, and T. D. Laing, "Relative humidity for load forecasting models," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 191-198, 2016.

[35] J. Xie and T. Hong, "Wind speed for load forecasting models," *Sustainability*, vol. 9, no. 5, p. 795, 2017.

[36] J. Kleissl, "Solar energy forecasting and resource assessment," *Academic Press*, 2013.

[37] P. Wang, B. Liu, and T. Hong, "Electric load forecasting with recency effect: a big data approach," *International Journal of Forecasting*, vol. 23, no. 3, pp. 585-597, July-September 2016.

[38] R. J. Hyndman and S. Fan, "Density forecasting for long-term peak electricity demand," *IEEE Transactions on Power Systems*, vol. 25, no. 2, pp. 1142-1153, 2009.

[39] T. Hong, J. Wilson, and J. Xie, "Long term probabilistic load forecasting and normalization with hourly information," *IEEE Transactions on Smart Grid*, vol. 5, no. 1, pp. 456-462, 2014.

[40] Y. Wang, Q. Chen, T. Hong, and C. Kang, "Review of smart meter data analytics: Applications, methodologies, and challenges," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3125-3148, May 2019.

[41] M. Kezunovic and T. Dokic, "Big data framework for predictive risk assessment of weather impacts on electric power systems," in *Grid of the Future, CIGRE US National Committee*, Atlanta, November, 2019.

[42] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," *International Journal of Forecasting*, vol. 30, no. 2, pp. 357-363, April-June 2014.

[43] T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond," *International Journal of Forecasting*, vol. 32, no. 3, pp. 896-913, July-September 2016.

[44] T. Hong, J. Xie, and J. Black, "Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting," *International Journal of Forecasting*, vol. 35, no. 4, pp. 1389-1399, October-December, 2019.

[45] Commission for Energy Regulation (CER). (2012). CER smart metering project - Electricity customer behaviour trial, 2009-2010 [dataset]. 1st Edition. Irish Social Science Data Archive, . [Online] Available: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/

[46] ARPA-E. (2016). Grid data. [Online] Available: https://arpa-e.energy.gov/?q=arpa-e-programs/grid-data

[47] My Electric Avenue. (2016). [Online] Available: http://myelectricavenue.info/

[48] J. D. Cross and R. Hartshorn, "My electric avenue: Integrating electric vehicles into the electrical networks," in *6th Hybrid and Electric Vehicles Conference (HEVC 2016)*, London, UK, 2-3 Nov. 2016.

[49] Z. J. Lee, T. Li, and S. H. Low, "ACN-Data: Analysis and applications of an open EV charging dataset," in *Proc. of the Tenth International Conference on Future Energy Systems, e-Energy '19*, Phoenix, Arizona, U.S.A., June 2019. [Online]. Available: https://ev.caltech.edu/dataset.

[50] ENTSO-E transparency platform. European Network of Transmission System Operators for Electricity. [Online] Available: https://transparency.entsoe.eu/

[51] T. V. Jensen and P. Pinson, "RE-Europe, a large-scale dataset for modeling a highly renewable European electricity system," *Scientific Data*, vol. 4, no. 170175, 2017.

[52] T. V. Jensen, H. d. Sevin, M. Greiner, and P. Pinson. (2015). The RE-Europe data set. [Online] Available: https://zenodo.org/record/35177#.Xlp626j7RaR

[53] G.B. national grid status. Gridwatch. [Online] Available: http://www.gridwatch.templar.co.uk/

[54] Z. Wang. (2019). Iowa Distribution Test Systems. Iowa State University. [Online] Available: http://wzy.ece.iastate.edu/Testsystem.html

[55] F. Bu, Y. Yuan, Z. Wang, K. Dehghanpour, and A. Kimber, "A Time-series distribution test system based on real utility data," in *2019 North American Power Symposium (NAPS)*, Wichita, KS, USA, 2019, pp. 1-6.

[56] Sotavento wind farm historical data. Sotavento. [Online] Available: http://www.sotaventogalicia.com/en/technical-area/real-time-data/historical/

[57] C. Draxl, A. Clifton, B.-M. Hodge, and J. McCaa, "The wind integration national dataset (WIND) toolkit," *Applied Energy*, vol. 151, no. 1, pp. 355-366, August 2015.

[58] H. T. C. Pedro, D. P. Larson, and C. F. M. Coimbra, "A comprehensive dataset for the accelerated development and benchmarking of solar forecasting methods," *Journal of Renewable and Sustainable Energy*, vol. 11, no. 036102, 2019.

[59] J. R. Andrade and R. J. Bessa, "Improving renewable energy forecasting with a grid of numerical weather predictions," *IEEE Transactions on Sustainable Energy*, vol. 8, no. 4, pp. 1571–1580, 2017.

[60] J. R. Andrade and R. J. Bessa. (14.04.2020). Solar power forecasting: measurements and numerical weather predictions. INESC TEC research data repository.

[61] C. Aggarwal, *Data mining, the textbook*. Springer, 2018.

[62] D. Taylor, "Battle of the data science Venn diagrams," *KDnuggets*, vol. 16, no. 33, 2016. [Online]. Available: https://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html.

[63] D. Blei and P. Smyth, "Science and data science," in *Proceedings of the National Academy of Sciences of the USA (PNAS)*, Aug. 15, 2017, vol. 114, no. 33, pp. 8689-8692.

[64] N. Albarakati and Z. Obradovic, "Multi-domain and multi-view networks model for clustering hospital admissions from the emergency department," *International Journal of Data Science and Analytics*, vol. 7, no. 4, pp. 385-403, 2019.

[65] D. Gligorijevic, J. Stojanovic, and Z. Obradovic, "Disease types discovery from a large database of inpatient records: A Sepsis study," *Methods*, vol. 111, pp. 45-55, 2016.

[66] D. Gligorijevic, J. Stojanovic, N. Djuric, V. Radosavljevic, M. Grbovic, R. J. Kulathinal, and Z. Obradovic, "Large-scale discovery of disease-disease and disease-gene associations," *Scientific Reports, Nature Publishing Group*, vol. 6, no. 32404, Aug 31, 2016.

[67] P. N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction to data mining*, 2nd edition ed. Pearson, 2019.

[68] I. Stojkovic and Z. Obradovic, "Sparse learning of the disease severity score for high dimensional data," *Complexity*, Article ID 7120691, 2017.

[69] M. Ghalwash, V. Radosavljevic, and Z. Obradovic, "Extraction of interpretable multivariate patterns for early diagnostic," in *Proc. 2013 IEEE International Conference on Data Mining (ICDM'13)*, Dallas, TX, Dec. 2013, pp. 201-210.

[70] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Machine Learning*, vol. 56 (Special Issue on Clustering), pp. 209–239, 2004.

[71] I. Triguero, S. Garcia, and F. Herrera, "Self-labeled techniques for semi-supervised learning: taxonomy," *software and empirical study, Knowledge and Information Systems*, vol. 42, no. 2, pp. 245–284, Nov. 2013.

[72] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees. Monterey*. CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[73] J. R. Quinlan, "Induction of decision trees, Machine Learning," vol. 1, pp. 81–106, 1986.

[74] J. R. Quinlan, *C4.5: Programs for machine learning* (Morgan Kaufmann Publishers). 1993.

[75] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[76] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[77] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[78] Y. Freund and R. E. Schapire, "A shot introduction to boosting," *Journal of Japaneese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771-780, Sept. 1999.

[79] T. Dokic, "Predictive risk assessment for optimal asset management in power systems," Ph.D. Dissertation, Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX, USA, May 2019.

[80] G. T. Toussaint, "Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining," *Int. J. Comput. Geometry Appl.*, vol. 15, no. 2, pp. 101–150, 2005.

[81] G. F. Cooper, "The computational complexity of probabilistic inference using Bayesian belief networks," *Artificial Intelligence*, vol. 42, no. 2-3, pp. 393–405, 1990.

[82] S. H. Walker and D. B. Duncan, "Estimation of the probability of an event as a function of several independent variables," *Biometrika*, vol. 54, no. 1/2, pp. 167–178, 1967.

[83] M. Kezunovic, Z. Obradovic, T. Dokic, and S. Roychoudhury, "Systematic framework for integration of weather data into prediction models for the electric grid outage and asset management applications," in *Proc. 51th IEEE Hawaii International Conference on System Science (HICSS)*, Big Island, Hawaii, January 2018, pp. 2737-2746.

[84] C. Cortes and V. N. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.

[85] G. P. Zhang, "Neural networks for classificaiton: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 4, pp. 451-462, Nov. 2000.

[86] L. G. Perez, A. J. Flechsig, J. L. Meador, and Z. Obradovic, "Training an artificial neural network to discriminate between magnetizing inrush and internal faults," *IEEE Trans. on Power Delivery*, vol. 9, no. 1, pp. 434-441, 1994.

[87] P. Werbos, *The roots of backpropagation: From ordered derivatives to neural networks and political forecasting*. New York: John Wiley & Sons, 1994.

[88] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, 2015.

[89] D. Ciresan, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, "Flexible, high performance convolutional neural networks for image classification," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2001, vol. 2, pp. 1237–1242.

[90] F. Gers, N. Schraudolph, and J. Schmidhuber, "Learning precise timing with LSTM recurrent networks," *Journal of Machine Learning Research*, vol. 3, pp. 115–143, 2002.

[91] T. Dokic, P. Dehghanian, P.-C. Chen, M. Kezunovic, Z. Medina-Cetina, J. Stojanovic, and Z. Obradovic, "Risk assessment of a transmission line insulation breakdown due to lightning and severe weather," in *Proc. 49th IEEE Hawaii International Conference on System Science (HICSS)*, Kauai, Hawaii, January 2016, pp. 2488-2497.

[92] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning (Vol. 1)*. Cambridge: MIT press, 2016.

[93] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural networks*, vol. 61, pp. 85-117, 2015.

[94] T. Dokic, M. Pavlovski, D. Gligorijevic, M. Kezunovic, and Z. Obradovic, "Spatially aware ensemble-based learning to predict weather-related outages in transmission," in *The Hawaii International Conference on System Sciences – HICSS*, Maui, Hawaii, January 2019.

[95] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 701-710.

[96] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 855-864.

[97] M. Alqudah, T. Dokic, M. Kezunovic, and Z. Obradovic, "Prediction of solar radiation based on spatial and temporal embeddings for solar generation forecast," in *Proc. 53th IEEE Hawaii International Conference on System Science (HICSS)*, Maui, Hawaii, January 2020.

[98] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big data: Issues and challenges moving forward," in *Proc. of 46th Hawaii International Conference on System Sciences*, 2013.

[99] J. Chen, Y. Chen, X. Du, C. Li, J. Lu, S. Zhao, and X. Zhou, "Big data challenge: a data management perspective," *Frontiers of Computer Science*, vol. 7, pp. 157-164, 2013.

[100] A. D. Mauro, M. Greco, and M. Grimaldi, "A formal definition of Big Data based on its essential features," *Library Review*, vol. 65, no. 3, pp. 122-135, 2016.

[101] L. Cai and Y. Zhu, "The challenges of data quality and data quality assessment in the big data era," *Data Science Journal*, vol. 14, p. 2, 2015.

[102] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," *ACM Computing Surveys*, vol. 41, no. 3, art. no. 16, 2009.

[103] T. Dasu and J. Loh, "Statistical distortion: consequences of data cleaning," in *Proc. of the VLDB Endowment*, 2012, vol. 5, no. 11.

[104] J. Mei, Y. D. Castro, Y. Goude, and G. Hébrail, "Nonnegative matrix factorization for time series recovery from a few temporal aggregates," in *Proc. of the 34th International Conference on Machine Learning (ICML'17)*, 2017, vol. 70, pp. 2382-2390.

[105] S. K. Khaitan and A. Gupta, *High performance computing in power and energy systems*. New York: Springer Verlag, 2013.

[106] A. Papavasiliou, S. S. Oren, and B. Rountree, "Applying high performance computing to transmission-constrained stochastic unit commitment for renewable energy integration," *IEEE Transactions on Power Systems*, vol. 30, no. 3, pp. 1109-1120, 2014.

[107] A. Papavasiliou, S. S. Oren, Z. Yang, P. Balasubramanian, and K. Hedman, "An application of high performance computing to transmission switching," in *Proceedings of IREP'2013*, 2013.

[108] A. J. Conejo, E. Castillo, R. Minguez, and R. Garcia-Bertrand, *Decomposition techniques in mathematical programming: engineering and science applications*. New York: Springer Verlag, 2006.

[109] A. Kargarian, J. Mohammadi, J. Guo, S. Chakrabarti, M. Barati, G. Hug, S. Kar, and R. Baldick, "Toward distributed/decentralized DC optimal power flow implementation in future electric power systems," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 2574-2594, 2018.

[110] P. Pinson, "Introducing distributed learning in wind power forecasting," in *Proc. of PMAPS 2016*, Beijing, China, 2016.

[111] W. Shi and S. Dustdar, "The promise of edge computing," *Computer*, vol. 49, no. 5, pp. 78-81, 2016.

[112] Z. Li, M. Shahidehpour, and F. Aminifar, "Cybersecurity in distributed power systems," in *Proc. of the IEEE*, 2017, vol. 105, no. 7, pp. 1367-1388.

[113] G. L. Ray and P. Pinson, "The ethical smart grid: Enabling a fruitful and long-lasting relationship between utilities and customers," *Energy Policy, in press*, 2020.

[114] "National disaster risk assessment," United Nations Office of Disaster Risk Reduction, 2017.

[115] J. Endrenyi, S. Aboresheid, R. N. Allan, G. J. Anders, S. Asgarpoor, R. Billinton, N. Chowdhury, E. N. Dialynas, M. Fipper, R. H. Fletcher, C. Grigg, J. McCalley, S. Meliopoulos, T. C. Mielnik, P. Nitu, N. Rau, N. D. Reppen, L. Salvaderi, A. Schneider, and C. Singh, "The present status of maintenance strategies and the impact of maintenance on reliability," *IEEE Transactions on Power Systems*, vol. 16, no. 4, pp. 638-646, Nov. 2001.

[116] Q. Yan, T. Dokic, and M. Kezunovic, "Predicting impact of weather caused blackouts on electricity customers based on risk assessment," in *IEEE Power and Energy Society General Meeting*, Boston, MA, July 2016.

[117] L. H. Fink and K. Carlsen, "Operating under stress and strain," *IEEE Spectrum*, vol. 15, pp. 48–53, Mar. 1978.

[118] "Internal EPRI communication (need to find an open source reference)."

[119] Y. Dong, V. Aravinthan, M. Kezunovic, and W. Jewell, "Integration of asset and outage management tasks for distribution systems," in *IEEE PES General Meeting*, Calgary, Canada, Jul 2009.

[120] Miscellaneous outage data and analysis. Bonneville Power Administration. [Online] Available: https://transmission.bpa.gov/Business/Operations/Outages/default.aspx

[121] National Weather Service. National digital forecast database (NDFD). National Oceanic and Atmospheric Administration. [Online] Available: https://www.weather.gov/mdl/ndfd_home

[122] V. Radosavljevic, K. Ristovski, and Z. Obradovic, "Gaussian conditional random fields for modeling patients' response to acute inflammation treatment," in *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, GA, 2013.

[123] T. Dokic and M. Kezunovic, "Predictive risk management for dynamic tree trimming scheduling for distribution networks," *IEEE Transactions on Smart Grid*, vol. 10, no. 5, pp. 4776-4785, September 2018.
[124] M. Duval, "Dissolved gas analysis: It can save your transformer," *IEEE Electrical Insulation Magazine*, vol. 5, pp. 22–27, 1989.
[125] R. K. Arora, "Different DGA techniques for monitoring of transformers," *International Journal of Electronics and Electrical Engineering*, vol. 1, pp. 299–303, 2013.
[126] *IEEE guide for the interpretation of gases generated in oil-immersed transformers*, IEEE Std. C57.104-1992, 1992.
[127] V. Miranda, A. R. Castro, and S. Lima, "Diagnosing faults in power transformers with autoassociative neural networks and mean shift," *IEEE Transactions on Power Delivery*, vol. 27, no. 3, pp. 1350–1357, 2012.
[128] K. Bacha, S. Souahlia, and M. Gossa, "Power transformer fault diagnosis based on dissolved gas analysis by support vector machine," *Electric Power Systems Research*, vol. 83, pp. 73–79, 2012.
[129] J. I. Aizpurua, V. M. Catterson, B. G. Stewart, S. D. McArthur, B. Lambert, B. Ampofo, G. Pereira, and J. G. Cross, "Power transformer dissolved gas analysis through bayesian networks and hypothesis testing," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 25, no. 2, pp. 494–506, 2018.
[130] E. H. Ko, T. Dokic, and M. Kezunovic, "Prediction model for the distribution transformer failure using correlation of weather data," in *CIGRE 5th International Colloquium Transformer Research and Asset Management*, Opatija, Croatia, October 2019.
[131] A. N. Jahromi, R. Piercy, S. Cress, J. R. R. Service, and W. Fan, "An approach to power transformer asset management using health index," *IEEE Electrical Insulation Magazine*, vol. 25, no. 2, pp. 20–34, Mar. 2009.
[132] M. M. Islam, G. Lee, and S. N. Hettiwatte, "Incipient fault diagnosis in power transformers by clustering and adapted KNN," in *2016 Australasian Universities power engineering conference (AUPEC)*, Brisbane, QLD, 2016, pp. 1-5.
[133] P. Mirowski and Y. LeCun, "Statistical machine learning and dissolved gas analysis: a review," *IEEE Transactions on Power Delivery*, vol. 27, no. 4, pp. 1791-1799, 2012.
[134] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.
[135] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *the 32nd International Conference on Machine Learning (ICML)*, Lille, France, 2015.
[136] S. M. Quiring, A. B. Schumacher, and S. D. Guikema, "Incorporating hurricane forecast uncertainty into a decision-support application for power outage modeling," *Bulletin of the American Meteorological Society*, vol. 95, no. 1, pp. 47-58, 2014.
[137] R. Nateghi, S. D. Guikema, and S. M. Quiring, "Comparison and validation of statistical methods for predicting power outage durations in the event of hurricanes," *Risk Analysis*, vol. 31, no. 12, pp. 1897-1906, 2011.
[138] F. L. Quilumba, W. J. Lee, H. Huang, D. Y. Wang, and R. L. Szabados, "Using smart meter data to improve the accuracy of intraday load forecasting considering customer behavior similarities," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 911-918, 2014.
[139] S. B. Taieb, R. Huser, R. J. Hyndman, and M. G. Genton, "Forecasting Uncertainty in Electricity Smart Meter Data by Boosting Additive Quantile Regression," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2448-2455, Sept. 2016.
[140] A. Albert and R. Rajagopal, "Smart meter driven segmentation: What your consumption says about you," *IEEE Transactions on power systems*, vol. 28, no. 4, pp. 4019–4030, 2013.
[141] B. Stephen, A. J. Mutanen, S. Galloway, G. Burt, and P. Järventausta, "Enhanced load profiling for residential network customers," *IEEE Transactions on Power Delivery*, vol. 29, no. 1, pp. 88-96, Feb. 2014.
[142] K. Zhou, S. Yang, and C. Shen, "A review of electric load classification in smart grid environment," *Renewable and Sustainable Energy Reviews*, vol. 24, pp. 103–110, 2013.
[143] X. Zhang, S. Grijalva, and M. J. Reno, "A time-variant load model based on smart meter data mining," in *2014 IEEE PES General Meeting | Conference & Exposition*, National Harbor, MD, 2014.
[144] D. Divan, R. Moghe, and H. Chun, "Managing distribution feeder voltage issues caused by high pv penetration," in *2016 IEEE 7th International Symposium on Power Electronics for Distributed Generation Systems (PEDG)*, 2016, pp. 1-8.
[145] X. Zhang and S. Grijalva, "A data-driven approach for detection and estimation of residential PV installations," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2477–2485, Sept. 2016.
[146] Z. Zhang, J. H. Son, Y. Li, M. Trayer, Z. Pi, D. Y. Hwang, and J. K. Moon, "Training free non-intrusive load monitoring of electric vehicle charging with low sampling rate," in *IECON 2014-40th Annual Conference of the IEEE Industrial Electronics Society*, 2014, pp. 5419–5425.
[147] K. Mason, M. Reno, L. Blakely, S. Vejdan, and S. Grijalva, "A deep learning approach for residential PV size, tilt and azimuth estimation," *Solar Energy*, 2019.
[148] R. Moghaddass and J. Wang, "A hierarchical framework for smart grid anomaly detection using large-scale smart meter data," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 5820–5830, 2017.
[149] Y. Jiang, C.-C. Liu, M. Diedesch, E. Lee, and A. K. Srivastava, "Outage management of distribution systems incorporating information from smart meters," *IEEE Transactions on Power Systems*, vol. 31, no. 5, pp. 4144–4154, 2015.
[150] L. Blakely, M. J. Reno, and W. Feng, "Spectral clustering for customer phase identification using AMI voltage timeseries," in *2019 IEEE Power and Energy Conference at Illinois (PECI)*, 2019.
[151] Y. Weng, Y. Liao, and R. Rajagopal, "Distributed energy resources topology identification via graphical modeling," *IEEE Transactions on Power Systems*, vol. 32, no. 4, pp. 2682-2694, July 2017.
[152] M. Lave, M. J. Reno, and J. Peppanen, "Distribution system parameter and topology estimation applied to resolve low-voltage circuits on three real distribution feeders," *IEEE Transactions on Sustainable Energy*, vol. 10, no. 3, pp. 1585–1592, 2019.
[153] T. Hong, D. W. Gao, T. Laing, D. Kruchten, and J. Calzada, "Training energy data scientists: universities and industry need to work together to bridge the talent gap," *IEEE Power and Energy Magazine*, vol. 16, no. 3, pp. 66-73, May-June 2018.
[154] T. Hong, "Crystal ball lessons in predictive analytics," *EnergyBiz*, pp. 35-37, Spring 2015.
[155] T. Hong, "Energy forecasting: past, present and future," *Foresight: The International Journal of Applied Forecasting*, vol. 32, pp. 43-48, Winter 2014.
[156] H. L. Willis and J. E. Northcote-Green, "Spatial electric load forecasting: a tutorial review," *Proceedings of the IEEE*, vol. 71, no. 2, pp. 232-253, Feb. 1983.
[157] H. L. Willis, *Spatial electric load forecasting*. CRC Press, 2002.
[158] T. Hong, "Long-term spatial load forecasting using human machine co-construct intelligence framework," Master thesis, North Carolina State University, 2008.
[159] G. Gross and F. D. Galiana, "Short-term load forecasting," *Proceedings of the IEEE*, vol. 75, no. 12, pp. 1558-1573, Dec. 1987.
[160] H. S. Hippert, C. E. Pedreira, and R. C. Souza, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Transactions on power systems*, vol. 16, no. 1, pp. 44-55, February 2001.
[161] A. Khotanzad, R. Afkhami-Rohani, and D. Maratukulam, "ANNSTLF-artificial neural network short-term load forecaster-generation three," *IEEE Transactions on Power Systems*, vol. 13, no. 4, pp. 1413-1422, 1998.
[162] T. Hong, "Short term electric load forecasting," PhD dissertation, North Carolina State University, 2010.
[163] J. Xie, T. Hong, and J. Stroud, "Long term retail energy forecasting with consideration of residential customer attrition," *IEEE*

Transactions on Smart Grid, vol. 6, no. 5, pp. 2245-2252, September 2015.

[164] T. Hong and M. Shahidehpour, "Load forecasting case study," *National Association of Regulatory Utility Commissioners*, pp. 1-171, 2015.

[165] J. Xie and T. Hong, "GEFCom2014 probabilistic electric load forecasting: an integrated solution with forecast combination and residual simulation," *International Journal of Forecasting*, vol. 32, no. 3, pp. 1012-1016, July-September 2016.

[166] M. Sobhani, A. Campbell, S. Sangamwar, C. Li, and T. Hong, "Combining weather stations for electric load forecasting," *Energies*, vol. 12, no. 8, p. 1510, April 2019.

[167] S. Moreno-Carbonell, E. F. Sánchez-Úbeda, and A. Muñoz, "Rethinking weather station selection for electric load forecasting using genetic algorithms," *International Journal of Forecasting*, 2019. [Online]. Available: https://doi.org/10.1016/j.ijforecast.2019.08.008.

[168] T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914-938, 2016.

[169] J. Xie, T. Hong, T. D. Laing, and C. Kang, "On normality assumption in residual simulation for probabilistic load forecasting," *IEEE Transactions on Smart Grid*, vol. 8, no. 3, pp. 1046-1053, May 2017.

[170] Y. Wang, Q. Chen, N. Zhang, and Y. Wang, "Conditional residual modeling for probabilistic load forecasting," *IEEE Transactions on Power Systems*, vol. 33, no. 6, pp. 7327-7330, 2018.

[171] J. Xie and T. Hong, "Temperature scenario generation for probabilistic load forecasting," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 1680-1687, May 2018.

[172] B. Liu, J. Nowotarski, T. Hong, and R. Weron, "Probabilistic load forecasting via quantile regression averaging on sister forecasts," *IEEE Transactions on Smart Grid*, vol. 8, no. 2, pp. 730-737, March 2017.

[173] D. Gan, Y. Wang, S. Yang, and C. Kang, "Embedding based quantile regression neural network for probabilistic load forecasting," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 2, pp. 1-11, 2018.

[174] J. Xie and T. Hong, "Variable selection methods for probabilistic load forecasting: empirical evidence from seven states of the United States," *IEEE Transactions on Smart Grid*, vol. 9, no. 6, pp. 6039-6046, November 2018.

[175] Y. Wang, D. Gan, N. Zhang, L. Xie, and C. Kang, "Feature selection for probabilistic load forecasting via sparse penalized quantile regression," *Journal of Modern Power Systems and Clean Energy*, vol. 7, no. 5, pp. 1200-1209, 2019.

[176] Y. Wang, N. Zhang, Y. Tan, T. Hong, D. Kirschen, and C. Kang, "Combining probabilistic load forecasts," *IEEE Transactions on Smart Grid*, vol. 10, no. 4, pp. 3664-3674, July 2019.

[177] W. Kong, Z. Y. Dong, Y. Jia, D. J. Hill, Y. Xu, and Y. Zhang, "Short-term residential load forecasting based on LSTM recurrent neural network," *IEEE Transactions on Smart Grid*, vol. 10, no. 1, pp. 841-851, 2017.

[178] H. Chitsaz, H. Shaker, H. Zareipour, D. Wood, and N. Amjady, "Short-term electricity load forecasting of buildings in microgrids," *Energy and Buildings*, vol. 99, pp. 50-60, 2015.

[179] J. Massana, C. Pous, L. Burgas, J. Melendez, and J. Colomer, "Short-term load forecasting in a non-residential building contrasting models and attributes," *Energy and Buildings*, vol. 92, pp. 322-330, 2015.

[180] P. Lusis, K. R. Khalilpour, L. Andrew, and A. Liebman, "Short-term residential load forecasting: Impact of calendar effects and forecast granularity," *Applied Energy*, vol. 205, pp. 654-669, 2017.

[181] P. Bacher, H. Madsen, H. A. Nielsen, and B. Perers, "Short-term heat load forecasting for single family houses," *Energy and buildings*, vol. 65, pp. 101-112, 2013.

[182] A. Bracale, G. Carpinelli, P. D. Falco, and T. Hong, "Short-term industrial reactive power forecasting," *International Journal of Electrical Power & Energy Systems*, vol. 107, pp. 177-185, May 2019.

[183] F. Alasali, S. Haben, V. Becerra, and W. Holderbaum, "Day-ahead industrial load forecasting for electric RTG cranes," *Journal of Modern Power Systems and Clean Energy*, vol. 6, no. 2, pp. 223-234, 2018.

[184] K. Berk, A. Hoffmann, and A. Müller, "Probabilistic forecasting of industrial electricity load with regime switching behavior," *International Journal of Forecasting*, vol. 34, no. 2, pp. 147-162, 2018.

[185] B. G. Brown, R. W. Katz, and A. H. Murphy, "Time series models to simulate and forecast wind speed and wind power," *Journal of Applied Meteorology*, vol. 23, pp. 1184–1195, 1984.

[186] C. Gilbert, J. W. Messner, P. Pinson, P. J. Trombe, R. Verzijlbergh, P. v. Dorp, and H. Jonker, "Statistical post-processing of turbulence-resolving weather forecasts for offshore wind power forecasting," *Wind Energy*, vol. 23, no. 4, pp. 884-897, 2020.

[187] J. Dowell and P. Pinson, "Very short-term probabilistic wind power forecasts by sparse vector autoregression," *IEEE Transactions on Smart Grid*, vol. 7, no. 2, pp. 763–770, 2016.

[188] C. W. Chow, B. Urquhart, M. Lave, A. Dominguez, and J. Kleissl, "Intra-hour forecasting with a total sky imager at the UC San Diego solar energy testbed," *Solar Energy*, vol. 85, no. 11, pp. 2881–2893, 2011.

[189] P. J. Trombe, P. Pinson, C. Vincent, T. Bøvith, N. A. Cutululis, C. Draxl, G. Giebel, A. N. Hahmann, N. E. Jensen, B. P. Jensen, N. F. Le, H. Madsen, L. B. Pedersen, and A. Sommer, "Weather radars - the new eyes of offshore wind farms," *Wind Energy*, vol. 17, no. 11, 2014.

[190] R. Frehlich, "Scanning doppler lidar for input into short-term wind power forecasts," *Journal of Atmospheric & Oceanic Technology*, vol. 30, no. 2, pp. 230–244, 2013.

[191] C. L. Archer, B. A. Colle, L. D. Monache, M. J. Dvorak, J. Lundquist, B. H. B. a. P. Beaucage, M. J. Churchfield, A. C. F. a. B. Kosovic, S. L. a. P. J. Moriarty, H. Simao, R. J. A. M. Stevens, D. Veron, and J. Zack, "Meteorology for coastal/offshore wind energy in the United States: Recommendations and research needs for the next 10 years," *Bulletin of the American Meteorological Society*, vol. 95, pp. 515-519, 2014.

[192] P. J. Trombe, P. Pinson, and H. Madsen, "Automatic classification of offshore wind regimes with weather radar observations," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 1, pp. 116–125, 2014.

[193] P. Pinson, "Adaptive calibration of (u, v)-wind ensemble forecasts," *Quarterly Journal of the Royal Meteorological Society*, vol. 138, no. 666, pp. 1273–1284, 2012.

[194] P. Pinson, "Wind energy: Forecasting challenges for its optimal management," *Statistical Science*, vol. 28, no. 4, pp. 564–585, 2013.

[195] S. Ashok, "Peak-load management in steel plants," *Applied energy*, vol. 83, no. 5, pp. 413-424, May 2006.

[196] M. Heleno, M. A. Matos, and J. A. P. Lopes, "Availability and flexibility of loads for the provision of reserve," *IEEE Transactions on Smart Grid*, vol. 6, no. 2, pp. 667-674, Mar. 2015.

[197] F. Ruelens, B. J. Claessens, S. Vandael, S. Iacovella, P. Vingerhoets, and R. Belmans, "Demand response of a heterogeneous cluster of electric water heaters using batch reinforcement learning," in *Proc. of the 18th Power Systems Computation Conference (PSCC 2014)*, Wroclaw, Poland, Aug 2014.

[198] F. Ruelens, B. J. Claessens, S. Quaiyum, B. D. Schutter, R. Babuška, and R. Belmans, "Reinforcement learning applied to an electric water heater: From theory to practice," *IEEE Transactions on Smart Grid*, vol. 9, no. 4, pp. 3792-3800, July 2018.

[199] J. Filipe, R. J. Bessa, M. Reis, R. Alves, and P. Póvoa, "Data-driven predictive energy optimization in a wastewater pumping station," *Applied Energy*, vol. 252, p. 113423, Oct. 2019.

[200] F. Smarra, A. Jain, T. Rubeis, D. Ambrosini, A. D'Innocenzo, and R. Mangharam, "Data-driven model predictive control using random forests for building energy optimization and climate control," *Applied Energy*, vol. 226, pp. 1252-1272, Sept. 2018.

[201] K. Kouzelis, Z. H. Tan, B. Bak-Jensen, J. R. Pillai, and E. Ritchie, "Estimation of residential heat pump consumption for flexibility market applications," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1852 – 1864, July 2015.

[202] A. Gerossier, T. Barbier, and R. Girard, "A novel method for decomposing electricity feeder load into elementary profiles from customer information," *Applied Energy*, vol. 203, pp. 752-760, Oct. 2017.

[203] H. Yang, J. Zhang, J. Qiu, S. Zhang, M. Lai, and Z. Dong, "A practical pricing approach to smart grid demand response based on load classification," *IEEE Transactions on Smart Grid*, vol. 9, no. 1, pp. 179-190, Jan. 2018.

[204] X. Z. Wang, J. Zhou, Z. L. Huang, X. L. Bi, Z. Q. Ge, and L. Li, "A multilevel deep learning method for Big Data analysis and emergency management of power system," *IEEE Int. Conference on Big Data Analysis*, pp. 1-5, 2016.

[205] G. L. Ray and P. Pinson, "Online adaptive clustering algorithm for energy consumption profiling," *Sustainable Energy, Grids and Networks*, vol. 17, p. 100181, Mar. 2019.

[206] J. Kwac and R. Rajagopal, "Data-driven targeting of customers for demand response," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2199 – 2207, Sept. 2016.

[207] K. Ganesan, J. Saraiva, and R. J. Bessa, "On the use of causality inference in designing tariffs to implement more effective behavioral demand response programs," *Energies*, vol. 12, no. 14, p. 2666, July 2019.

[208] G. L. Ray, E. M. Larsen, and P. Pinson, "Evaluating price-based demand response in practice – with application to the EcoGrid EU Experiment," *IEEE Transactions on Smart Grid*, vol. 9, no. 3, pp. 2304-2313, May 2018.

[209] S. Kim and G. B. Giannakis, "An online convex optimization approach to real-time energy pricing for demand response," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 2784 – 2793, Nov. 2017.

[210] R. Yu, W. Yang, and S. Rahardja, "A statistical demand-price model with its application in optimal real-time price," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1734 – 1742, Dec. 2012.

[211] N. Koseleva and G. Ropaite, "Big data in building energy efficiency: Understanding of big data and main challenges," *Procedia Engineering*, vol. 172, pp. 544-549, 2017.

[212] P. A. Mathew, L. N. Dunn, M. D. Sohn, A. Mercado, C. Custudio, and T. Walter, "Big-data for building energy performance: Lessons from assembling a very large national database of building energy use," *Applied Energy*, vol. 140, pp. 85-93, Feb. 2015.

[213] T. Walter and M. D. Sohn, "A regression-based approach to estimating retrofit savings using the building performance database," *Applied Energy*, pp. 996-1005, Oct. 2016.

[214] "Final report on the August 14, 2003 blackout in the United States and Canada: Causes and recommendations," U.S.-Canada Power System Outage Task Force, April 2004.

[215] M. Kezunovic, S. Meliopoulos, S. Venkatasubramanian, and V. Vittal, *Application of Time-Synchronized Measurements in Power System Transmission Networks*. Springer, 2014.

[216] J. A. d. l. O. Serna, "Synchrophasor estimation using Prony's method," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 8, Aug. 2013.

[217] O. P. Dahal, S. M. Brahma, and H. Cao, "Comprehensive clustering of disturbance events recorded by phasor measurement units," *IEEE Transactions on Power Delivery*, vol. 29, no. 3, pp. 1390-1397, 2014.

[218] M. Rafferty, X. Liu, D. Laverty, and S. McLoone, "Real-time multiple event detection and classification using moving window PCA," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2537 – 2548, Sept. 2016.

[219] M. Biswal, S. M. Brahma, and H. Cao, "Supervisory protection and automated event diagnosis using PMU data," *IEEE Transactions on Power Delivery*, vol. 31, no. 4, pp. 1855-1863, 2016.

[220] D. I. Kim, T. Y. Chun, S. H. Yoon, G. Lee, and Y. J. Shin, "Wavelet-based event detection method using PMU data," *IEEE Transactions on Smart grid*, vol. 8, no. 3, pp. 1154-1162, 2017.

[221] M. Dehghani, B. Shayanfard, and A. R. Khayatian, "PMU ranking based on singular value decomposition of dynamic stability matrix," *IEEE Transactions on Power Systems*, vol. 28, no. 3, Aug. 2013.

[222] V. Malbasa, C. Zheng, P. C. Chen, T. Popovic, and M. Kezunovic, "Voltage stability prediction using active machine learning," *IEEE Transactions on Smart Grid*, vol. 8, no. 6, pp. 3117-3124, 2017.

[223] C. Zheng, V. Malbasa, and M. Kezunovic, "Regression tree for stability margin prediction using synchrophasor measurements," *IEEE Transactions on Power Systems*, vol. 28, no. 2, pp. 1978-1987, May 2013.

[224] L. S. Moulin, A. P. A. d. Silva, M. A. El-Sharkawi, and R. J. Marks, "Support vector machines for transient stability analysis of large-scale power systems," *IEEE Transactions on Power Systems*, vol. 19, no. 2, pp. 818-825, May 2004.

[225] K. Sun, S. Likhate, V. Vittal, V. S. Kolluri, and S. Mandal, "An online dynamic security assessment scheme using phasor measurements and decision trees," *IEEE Transactions on Power Systems*, vol. 22, no. 4, pp. 1935-1943, Nov. 2007.

[226] I. Kamwa, S. R. Samantaray, and G. Joos, "Development of rule-based classifiers for rapid stability assessment of wide-area post-disturbance records," *IEEE Transactions on Power Systems*, vol. 24, no. 1, pp. 258-270, Feb. 2009.

[227] F. R. Gomez, A. D. Rajapakse, U. D. Annakkage, and I. T. Fernando, "Support vector machine-based algorithm for post-fault transient stability status prediction using synchronized measurements," *IEEE Transactions on Power Systems*, vol. 26, no. 3, pp. 1474-1483, Aug. 2011.

[228] L. Zhu, C. Lu, Z. Y. Dong, and C. Hong, "Imbalance learning machine-based power system short-term voltage stability assessment," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 5, pp. 2533-2543, Oct. 2017.

[229] T. Guo and J. V. Milanović, "Online identification of power system dynamic signature using PMU measurements and data mining," *IEEE Transactions on Power Systems*, vol. 31, no. 3, pp. 1760-1768, May 2016.

[230] Torch: A scientific computing framework for LuaJIT. Torch. [Online] Available: http://torch.ch/

[231] W. G. Hatcher and W. Yu, "A survey of deep learning: Platforms, applications and emerging research trends," *IEEE Access*, vol. 6, pp. 24411-24432, 2018.

[232] M. Gheisari, G. Wang, and M. Z. A. Bhuiyan, "A survey on deep learning in big data. ," in *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, Guangzhou, 2017, July, vol. 2, pp. 173-180.

[233] M. G. Simões, R. Roche, E. Kyriakides, S. Suryanarayanan, B. Blunier, K. D. McBee, P. H. Nguyen, P. F. Ribeiro, and A. Miraoui, "A comparison of smart grid technologies and progresses in Europe and the U.S.," *IEEE Transactions on Industry Applications*, vol. 48, no. 4, pp. 1154-1162, July-Aug. 2012.

[234] Y. Liao, Y. Weng, G. Liu, and R. Rajagopal, "Urban MV and LV distribution grid topology estimation via group Lasso," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 12-27, Jan. 2019.

[235] O. Ardakanian, V. Wong, R. Dobbe, S. H. Low, A. v. Meier, C. Tomlin, and Y. Yuan, "On identification of distribution grids," *IEEE Transactions on Control of Network Systems*, vol. 6, no. 3, pp. 950 – 960, Sept. 2019.

[236] J. Yu, Y. Weng, and R. Rajagopal, "PaToPa: A data-driven parameter and topology joint estimation framework in distribution grids," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4335 – 4347, Jul. 2018.

[237] D. Kodaira, J. Park, S. Y. Kim, S. Han, and S. Han, "Impedance estimation with an enhanced particle swarm optimization for low-voltage distribution networks," *Energies*, vol. 12, no. 6, p. 1167, 2019.

[238] R. A. Sevlian and R. Rajagopal, "Distribution system topology detection using consumer load and line flow measurements," *arXiv:1503.07224*, 2015.
[239] F. Olivier, A. Sutera, P. Geurts, R. Fonteneau, and D. Ernst, "Phase identification of smart meters by clustering voltage measurements," in *Proc. of the 20th Power Systems Computation Conference (PSCC 2018)*, Dublin, Ireland, June 2018.
[240] X. Zhang and S. Grijalva, "An advanced data driven model for residential electric vehicle charging demand," in *2015 IEEE Power & Energy Society General Meeting*, Denver, CO, USA, Jul. 2015.
[241] A. Shaw and B. P. Nayak, "Electric vehicle charging load filtering by power signature analysis," in *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, Pune, India, Feb. 2017.
[242] R. J. Bessa, G. Sampaio, J. Pereira, and V. Miranda, "Probabilistic low voltage state estimation using analog-search techniques," in *Proc. of the 20th Power Systems Computation Conference (PSCC 2018)*, Dublin, Ireland, June 2018.
[243] B. Hayes and M. Prodanovic, "State forecasting and operational planning for distribution network energy management systems," *IEEE Transactions on Smart Grid*, vol. 7, no. 2, pp. 1002 – 1011, March 2016.
[244] Y. Weng, R. Negi, C. Faloutsos, and M. D. Ilić, "Robust data-driven state estimation for smart grid," *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1956 – 1967, Jul. 2017.
[245] M. Brown, D. C. Filho, and J. S. d. Souza, "Forecasting-aided state estimation—Part I: Panorama," *IEEE Transactions on Power Systems*, vol. 24, no. 4, pp. 1667 – 1677, Nov. 2009.
[246] S. Huang, C. Lu, and Y. Lo, "Evaluation of AMI and SCADA data synergy for distribution feeder modeling," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1639 – 1647, July 2015.
[247] J. Massignan, J. L. Jr., M. Bessani, C. Maciel, A. Delbem, M. Camillo, and T. W. d. L. Soares, "In-field validation of a real-time monitoring tool for distribution feeders," *IEEE Transactions on Power Delivery*, vol. 33, no. 4, pp. 1798 – 1808, Aug. 2018.
[248] J. Krstulovic, V. Miranda, A. J. S. Costa, and J. Pereira, "Towards an auto-associative topology state estimator," *IEEE Transactions on Power Systems*, vol. 28, no. 3, pp. 3311 – 3318, Aug. 2013.
[249] Y. C. Chen, J. Wang, A. D. Domínguez-García, and P. W. Sauer, "Measurement-based estimation of the power flow jacobian matrix," *IEEE Transactions on Smart Grid*, vol. 7, no. 5, pp. 2507 – 2515, Sept. 2016.
[250] G. Dalal, E. Gilboa, S. Mannor, and L. Wehenkel, "Chance-constrained outage scheduling using a machine learning proxy," *IEEE Transactions on Power Systems*, vol. 34, no. 4, pp. 2528-2540, Jul. 2019.
[251] B. Donnot, I. Guyon, M. Schoenauer, P. Panciatici, and A. Marot, "Introducing machine learning for power system operation support," in *X Bulk Power Systems Dynamics and Control Symposium (IREP'2017 Symposium)*, Espinho, Portugal, 27 Aug. – 1 Sept. 2017.
[252] A. Marot, B. Donnot, S. Tazi, and P. Panciatici, "Expert system for topological remedial action discovery in smart grids," in *Proc. of the 11th Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MEDPOWER 2018)*, Dubrovnik, Croatia, Nov. 2018.
[253] T. Lan, J. Duan, B. Zhang, D. Shi, Z. Wang, R. Diao, and X. Zhang, "AI-based autonomous line flow control via topology adjustment for maximizing time-series ATCs," *arXiv:1911.04263v1*, 2019.
[254] B. Donnot, I. Guyon, M. Schoenauer, A. Marot, and P. Panciatici, "Anticipating contingencies in power grids using fast neural net screening," in *2018 International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, Brazil, 8-13 July 2018.
[255] Q. Li, S. Li, B. Xu, and Y. Liu, "Optimal node attack on causality analysis in cyber-physical systems: A data-driven approach," *IEEE Access*, vol. 7, pp. 16066 - 16077, January 2019.
[256] S. L. Hay, G. W. Ault, and K. Bell, "Control room scenarios on active distribution networks: Early results and next steps," in *Proc. of the 44th International Universities Power Engineering Conference (UPEC 2009)*, Glasgow, UK, 1-4 Sept. 2009.
[257] I. Dabbaghchi and R. J. Gwky, "An abductive expert system for interpretation of real-time data," *IEEE Transactions on Power Delivery*, vol. 8, no. 3, pp. 1061-1069, Jul. 1993.
[258] C. Hor and P. A. Crossley, "Unsupervised event extraction within substations using rough classification," *IEEE Transactions on Power Delivery*, vol. 21, no. 4, pp. 1809 – 1816, Nov. 2006.