Integrated Multi-Scale Data Analytics and Machine Learning for the Distribution Grid

Emma M. Stewart*, Philip Top
Lawrence Livermore National Laboratory, Livermore, CA, USA

Michael Chertkov, Deepjyoti Deka, Scott Backhaus, Andrew Lokhov
Los Alamos National Laboratory, Los Alamos, NM, USA

Ciaran Roberts, Val Hendrix, Sean Peisert
Lawrence Berkeley National Laboratory, Berkeley, CA, USA

Anthony Florita
National Renewable Energy Laboratory, Golden, CO, USA

Thomas J. King Jr
Oak Ridge National Laboratory Oak Ridge, TN, USA

Matthew J. Reno
Sandia National Laboratories Albuquerque, NM, USA

Abstract— **We consider the field of machine learning and where it is both useful, and not useful, for the distribution grid and buildings interface. While analytics, in general, is a growing field of interest, and often seen as the golden goose in the burgeoning distribution grid industry, its application is often limited by communications infrastructure, or lack of a focused technical application. Overall, the linkage of analytics to purposeful application in the grid space has been limited. In this paper we consider the field of machine learning as a subset of analytical techniques, and discuss its ability and limitations to enable the future distribution grid. To that end, we also consider the potential for mixing distributed and centralized analytics and the pros and cons of these approaches. There is an exponentially expanding volume of measured data being generated on the distribution grid, which, with appropriate application of analytics, may be transformed into intelligible, actionable information that can be provided to the right actors – such as grid and building operators, at the appropriate time to enhance grid or building resilience, efficiency, and operations against various metrics or goals – such as total carbon reduction or other economic benefit to customers. While some basic analysis into these data streams can provide a wealth of information, computational and human boundaries on performing the analysis are becoming significant, with more data and multi-objective concerns. Efficient applications of analysis and the machine learning field are being considered in the loop. This paper describes benefits and limits of present machine-learning applications for use on the grid and presents a series of case studies that illustrate the potential benefits of developing advanced local multi-variate analytics machine-learning-based applications.**

Keywords—Analytics, Machine Learning, Distribution Grid, DER, validation, verification, prediction, incipient failure

I. Introduction

A vision of the future distribution grid and its interface to buildings is one of cohesion, an interactive reliable environment where there are consumer benefits and motivations to leverage customer owned behind-the-meter assets to provide services to the grid, energy markets, other entities within the distribution feeder, and ultimately to the larger society as a whole. This future distribution grid may be a reliable, safe, and resilient energy transport platform that supports high penetration of Distributed Energy Resources (DER). The growth of communicative DER and connected behind-the-meter power electronic devices may introduce fluctuations and uncertainty not previously seen on the distribution grid if the resources operate independently, or are driven by independent communications and controls. However, these new data generating and communicative features may also offer a vast opportunity to increase the operational efficiency of both the grid and the buildings connected to it, but only if the data collected at all the various nodes can be easily transformed into intelligible, actionable information.

Considering the customer interface to the grid, and vice versa, is a key opportunity to which analytics can be applied to enable greater interaction and unlock the potential of these resources. For example, in the vision we describe of cohesion, the availability of the resource to provide a particular service must be understood for that asset to be utilized and rewarded. The average customer has no desire to perform power flow calculations and evaluate their available data on a minute by minute, or even daily basis, and a contribution to a new service such as power quality management may have a difficult transition into the language of the consumer. Application of analytics at this interface will allow the automation of this function, providing useful information to the consumer regarding what they are participating in, while also giving the utility and grid a clear accurate understanding of the resources availability and performance. The existing customers benefit from new markets, new customers can integrate more distributed resources, all customers will have improved reliability, and the utility can manage the distributed generation adequately. We consider analytics to either derive information, diagnostics, prediction or a prescription or instruction for optimal control. While existing descriptive and diagnostic analytics for example, simple fault location and outage analysis, do not require forward thinking analytics, but a processing of information to distill useful information, utilizing this data in a prediction or prescription often requires a method of correction which can be enabled by machine learning. We discuss this in later stages of this introductory paper.

Lawrence Livermore National Laboratory is operated by Lawrence Livermore National Security, LLC, for the U.S. Department of Energy, National Nuclear Security Administration under Contract DE-AC52-07NA27344

To further this point, the large spatial footprint of the distribution grid and the diverse locations of its assets and node points make observability of normal, stochastic, and dynamic behavior and monitoring and diagnosis of abnormal (faults) and even planned (demand response or DER dispatch) events challenging tasks for the existing descriptive analytics field. The lack of observability, controllability, and validation/verification of DER and other behind the meter assets including their availability due to consumer behavior, preference, and choice may be a barrier to developing new transaction based markets, where consumer resources interact with one and other to provide or receive these services.

The work being developed by the multi-national laboratory team, funded through the Grid Modernization Initiative [1] will evaluate these challenges to develop data driven solutions leveraging multi-scale machine learning based analytics. The work utilizes various data sets across the nodes within the end to end power system (e.g. generation to end use) to automatically produce accurate actionable information for the various parties and actors encompassing the power system. At the heart of the work, applied analytics are required to turn these raw data into actionable information. Machine learning is required to enable a predictive prescriptive, computationally efficient and accurately managed modernized grid.

Although the terms data and information are often used interchangeably, in the context of analytics, they differ: Data are measurements – for example voltage, current, phase angle, power, or even metadata such as location and sensor type – and information is the actionable result of an application of an analytics technique to the data – for example calculating the availability of a behind-the-meter resource, predicting availability over time or the mean time to failure of a component, verifying the success of a requested action, and determining the best course of action with available resources. We examine where further development is needed to address gaps in existing analytics techniques to successfully apply these techniques to the grid and present notional case studies to demonstrate the potential value of these data analytics.

This paper will answer key questions related to 1) Machine learning and its definition as specific to the grid and buildings interface 2) Existing state of the art in machine learning applied to building and grid datasets and 3) Key case studies of machine learning applied to building and grid datasets with value propositions to each

II. Case Studies and State of the Art in Machine Learning

A significant volume of analyses are already being proposed for the power grid and buildings interface. Analyses such as consumption, forecast of load, and outages at present often rely on single data sources, such as smart metering on/off status being utilized to diagnose an outage location. Within the existing analytics platforms where techniques such as machine learning are already implemented, there are numerous instances of siloed data sources and techniques, and the analytics developed are often specific to the architecture of one set of data and tools rather than being multi-variate and applying data fusion techniques. This is the area where we intend to work and move the industry forward, ensuring that full advantage is taken of the data that are available, and that they are transformed into actionable information.

Machine learning is a subfield of computer science that studies and constructs algorithms that can learn and make predictions from data. Machine learning traditionally is suitable for situations where a domain is poorly understood or random, and a system needs to adapt to changes in the environment. In general it is applied when there is no knowledge to draw upon to create algorithmic solutions. Adapting analytics to machine learning, and advancing machine learning to meet the needs of power systems, can solve the problems and achieve the goals defined in the previous section, delivering the right information at the right time to the right people.

Machine learning uses techniques from many disciplines, including statistics, probability, game theory, and neurobiology. The basic principle is that a computer algorithm learns through experience with a set of tasks. The algorithm's performance is measured by how much the results improve over time and the speed of calculation. Machine learning can assist in interpreting stochastic and random behavior and using this information to benefit the customer and grid operator. The case studies we describe in this paper will have cross-program application and can play a key role in enabling the modernized grid.

On a broad scale, machine learning can be categorized into three areas: supervised, unsupervised, and reinforcement learning. Supervised machine learning entails learning from labeled examples (features) and produces a function that predicts either a discrete (classification) or continuous (regression) outcome. Unsupervised learning tries to describe patterns from data, while reinforcement techniques entail learning with rewards.

Current, state-of-the-art machine-learning solutions often rely on "black box" approaches, of which there has been significant development, applied often where systematic knowledge is poor but applications are computationally intense, require ubiquitous data sets, and are often agnostic to the subject area; i.e., the same methods are used for disparate areas such as medical data, grid data, or social census data. The application of machine learning in the power system is relatively new. Existing and recently developed machine-learning algorithms and approaches for power-distribution data problems can be divided into two categories: 1) utilizing established black box machine-learning methods to solve distribution-grid problems agnostic of power-system physics, and 2) improving and developing new methods utilizing power-system theory and models.

III. Centralized versus Distributed Analytics

One of the benefits of machine learning is the flexibility between it being applied locally or in a distributed manner, at the grid edge or building to grid interface, or by improving the existing state of the art applied normally in a centralized big data stack fashion. We now introduce centralized and distributed analytics as a preface to the case study discussion.

Evaluation and maintenance of grid health currently depends on a centralized, deterministic approach in which data are collected and analyzed, and some control action is then taken. This practice is designed for the old electricity grid that

functioned as a one-way conduit from a centralized plant to customers. Centralized analytics use data to discover the state of a system or find controls through global optimization. For example, in the case of grid state estimation, centralized analytics entail finding system parameters that are most consistent with the data. The global optimization problem is huge and frequently impractical to solve. However, in some cases an exact, or sufficiently accurate, optimal solution can be found by breaking systems into pieces and solving local optimization problems that each correspond to a piece of the system and results and information are updated among the pieces until convergence is achieved. The latter, the distributed approach, is computationally advantageous (as a function of communication latency and processor availability) and, as with distributed analytics, the processing power and communication remain local when this "piecewise" solution methodology is applied.

By contrast to traditional centralized grid data monitoring and analysis, building component health relies on a decentralized analytic approach in which each building component is monitored and analyzed individually. DER and smart meters have changed the grid paradigm, adding isolated, disparate data sources in both the distribution grid and buildings. At the same time, some building components impact grid performance, for example, the power-quality impact of a high penetration of electric vehicle chargers, or new electronically commutated motors in air conditioners. Actionable, evolving information is essential at both the building and grid level to enable reliable, efficient grid and building operations.

Collecting and mining raw data centrally make it challenging to act in a timely fashion on the information embedded in those data, whereas distributed analysis is challenging for the overall systematic approach. Central data management also requires significant amounts of data storage and increases the frequency of data errors (inconsistency, incompleteness, redundancy). To achieve the speed and efficiency required for grid operations, we need to move toward a hybrid approach to the central and decentralized model in which the computation takes place at the data sources themselves, or at the central area depending on the action, and actions are taken locally and reported globally. In this resource we consider intelligence as needing to be embedded in the grid, while informing overall operations in a timely accurate manner, a hybrid approach.

IV. Limitations of Existing Approaches to Analytics and Machine Learning

There are areas where the existing black box approaches of machine learning will not currently be useful. For example, machine learning can be brittle when faced with new situations, i.e. values that were not observed during training. Examples of this on the existing distribution grid are the few cases where PV penetration has reached more than 100% of peak demand (excluding Hawaii and California) or where significant volumes of customers have engaged in transactions to provide a local distribution grid service such as phase balancing or reduction in transformer loading. Without training on the impact of this scenario, machine-learning techniques, as currently applied, cannot easily provide operators sufficient confidence in the performance of advanced analytics. However, physics-based models and new techniques presented could create training sets for algorithms for situations that have not yet been experienced by the building or grid analytics node and assist in forecasting performance. This is a key benefit of new machine-learning applications that can be used to improve robustness and constrain outputs to build confidence in the new grid paradigm for system operators, who are traditionally conservative actors. In general, operators manage the grid well within stakeholders' bounds of safety, reliability, and comfort – which is not normally the most economically beneficial approach for the customers and leaves a significant bandwidth of untapped potential for building and grid services.

V. Applications of Machine learning to the buildings to grid interface

We consider three case studies areas across the spectrum of the building to grid interface including 1) DR and DER Local Availability and Verification, 2) Incipient Failure Detection in Distribution, and 3) Topology and Parameter Estimation. For each of these cases we consider the present state of the art, the problem we are trying to solve and the potential for distributed machine learning to create benefit for consumers and grid operators.

A. DR and DER Local Availability and Verification

At the building to grid interface, the ability of customers to transact or exchange resources is being considered as a new operational paradigm, both individually and in clustered aggregate systems with the grid. At the heart of this structure is controllable load, DR and DER The ability of new controllable DER devices to reliably provide controllable action depends on several factors. For one, these devices' response to input signaling and frequency fluctuation needs to be quantified. For aggregated loads like buildings, efficient load modeling is necessary to understand their cumulative response. Finally, efficient, low-overhead control schemes need to be designed and deployed in a distributed manner to create ubiquitous ancillary services from the distribution grid. Under present conditions, achieving all of these goals would be difficult.

The decentralized approach of transactive energy systems encompasses both energy and non-energy transactions in the distribution grid and buildings. The characteristics of transactive systems are that they have distributed control, provide feedback (via DER), concurrently address multiple objectives (e.g., load vs. comfort), are multi-scaled (microseconds, years), leverage automation (e.g., local action of voltage events) with the human-in-the-loop (control actions), and engage in coordination (negotiate decisions for competing objectives). Transactive system analysis encompasses several dimensions across time, space, stakeholders, and decision of risk. Distributed machine-learning techniques can enable customers and operators to leverage monetary and non-monetary benefits (e.g., health, comfort and environmental quality) of actions, while also communicating the overall verification of response to the upper grid hierarchy.

Within building energy consumption monitoring is provided by a complex network of building occupancy and local equipment sensors plus smart metering. Providing useful information from the grid operator to the building, or vice versa,

must use a single system that conveys the right information. Our goal is to use existing information about individual buildings to build a class of statistical models that characterize operations of the buildings under different/varying grid conditions, and then to reconstruct/learn the operational parameters of the probability distributions for multiple buildings simultaneously. The building side will benefit from improved forecasting and value streams being enabled for participating in non-kilowatt-only services. Grid operators will benefit from improved forecasting of customer behavior and new methods for enhancing stability on a grid with millions of resources.

Within this case we can consider the P-Q consumption of multiple buildings and complex loads within the distribution grid, enabling buildings and behind-the-meter resources to act, in clusters, like conventional generators, thereby enhancing grid stability. [2] developed a machine learning driven estimation method to determine the electrical performance of clusters of behind the meter resource, which is often limited by utilization of weather versus electrical measurements. These types of techniques are the pre-cursors to a distributed environment with predictive resource capabilities (Fig.1)

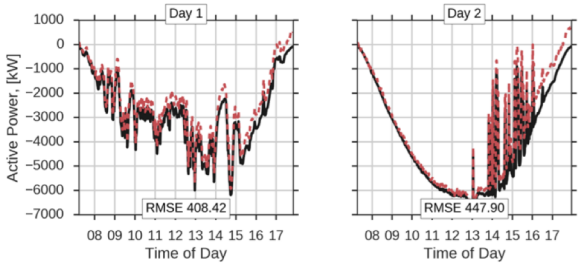

Fig. 1. Example of application of machine learning to prediction of a variable solar resource behind the meter

Building operators who are managing multiple objectives such as building comfort, system load, component failure risks, and extreme event responses need timely interpretable information about the building. [3] used a machine-learning technique called Learning-based Model Predictive Control that combined models with statistics to estimate occupancy and heating load based only on temperature measurements. To compensate for heating by occupancy control action chosen (AC on/off), [4] demonstrated a demand-response strategy synthesis that uses regression trees to partition the data space into small, manageable regions and then to partition the partitions until the data spaces can have simple models to fit them (easier for humans to interpret). Predictor variables are disturbances (weather, temperature) and controllable actions. During extreme events and peak-demand periods, utility customers could be incentivized to reduce electricity consumption. Establishing methods of reporting a customer's reduction in electricity use is critical for increasing the effectiveness of demand-response programs. Transactive control applications are generally designed to be self-organizing and localized; that is, they are decentralized and not coordinated.

Application of grid and building machine learning, could improve the accuracy and accountability for these services, which would enable an increase in the potential revenue streams and energy savings as well as enabling utility interaction with participating customers and an expansion of localized services to include ancillary services, distribution voltage regulation, and distribution balancing. The Bonneville Power Administration stated that to enable a comparable project at scale, i.e. with millions of available participating customers each with DER, the control schemes would require enhanced programmability, reliability, cost, and cost recovery for the utility. Application of advanced analytics with machine learning could address these requirements. The predicted impact of making use of these resources would be derived locally, with analytics informed by the grid and the state of the building and trained to apply to the local conditions. The current radial and manual configuration of the distribution grid state will change as the grid modernizes. The grid services being discussed here must be spatially and temporally verifiable to a particular electrical feeder, substation, and potentially phase of the distribution system. Incorporating these services requires that building and grid information be strongly linked and that services be verified, able to evolve, and repeatable.

New approaches to DER and DR controls and coordination will be essential with millions of resources available as described above. Millions of distributed smart grid assets require innovative approaches as existing theories that work on a scale of hundreds rather than millions of customers will no longer hold true. The increased volume of customer interaction would pose a significant computational and hierarchical burden using today's methods, and bounds of error for existing approaches will become untenable for large-scale control problems. Conceptual data-driven improvement of verifying and predicting DER controls could be made for example by cycling through aggregation and reinforcement learning. Machine-learning schemes built on observations collected from novel devices can help in realizing this vision. Aggregate load modeling and dynamic characterization of loads are possible using pattern-recognition schemes operating on training data. Theoretical machine learning provides the necessary mathematical foundation to develop distributed algorithms with low sample complexities that guarantee convergence of consensus and other control algorithms. These can aid in fast recovery, using load resources, after a frequency event. Further, the distributed nature of the algorithms will ensure an equitable share of the ancillary services among different distribution resources, and more reserve accountability will be enabled

B. Incipient Failure Detection in Distribution

Detection and identification of incipient failures within the electrical grid infrastructure can be considered in two realms a) direct sensing detection of failure and b) analysis of available local data for signs of failure. Proactive detection schemes can enable condition based maintenance and preventative responses that prevent potentially disruptive, costly, and potentially even catastrophic outages and failures before they occur. Large power transformers are commonly measured directly due to the high impact an outage would cause and the relative cost of that outage in comparison to a direct sensor. While large power transformers have a clear value proposition for specific monitoring and measurement of condition through techniques such as dissolved gas analysis, distribution asset monitoring is a field which does not benefit from the economies of scale in the same regard. Each component is a magnitude smaller at least,

and for every large power transformer, there may be thousands of distribution level transformers. At present, condition monitoring and maintenance in the distribution system is based upon a run to failure, and age based approach. Often the first sign of a distribution transformer failure is an outage for a number of customers, detected via smart metering, or a customer call to indicate a component with a visible failure, i.e. smoke.

In the event of a prediction of incipient failure, there is potential for delaying catastrophic failure with preventative action such as unloading the device, or a deeper dive analysis to find the root cause in the data. In doing this there is potential for utilizing the future DER and building resources available to operators. Some synergistic work in different fields for predicting behavior has been performed. [5] created a predictive model with data gathered from vehicle drivers to generate individualized driver models. This model uses the prior two seconds of driver pose to estimate the future four seconds. The approach showed that driver state (attentive, distracted) can increase the accuracy of prediction. [6] leverages a new data source, µPMUs (micro phasor measurement units) to help distribution planners accurately anticipate and control risks and opportunities for improvement on the distribution grid instead of relying on reactive operations (Fig. 2). The method illustrated in [6] uses event clustering to interpret events and predict issues such as anomalous tap change detection, arc flash, and capacitor bank switching. At the building to grid interface, the ability of customers to transact or exchange resources is being considered as a new operational paradigm, both individually and in clustered aggregate systems with the grid. At the heart of this structure is controllable load, demand response (DR) and distributed energy resources (DER).

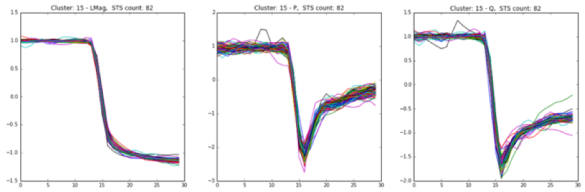

Fig. 2. Clustered repetitive behavior can be utilized in incipient failure detection in tap changers

A new promising approach currently under development assumes the stochastic linear dynamics of the device, and that the associated dynamic parameters that represent the normal behavior can be extracted from data using the state of the art regression-based algorithms with sparsity regularizations [7][8]. The amplitude of the failure related noise can be learned from the time series data. These parameters are then used to predict the near-term evolution and rigorously identify anomalies that significantly deviate from the predicted behavior. The long-term changes in parameters will serve as an indication of an incipient failure.

Most approaches considered in these realms could be decentralized or distributed with only key data being communicated. A centralized analytics approach to fault analysis is limited by the requirement that a human in the loop make decisions, which in itself is an inefficiency. Centralized approaches for fault analysis rely on whether smart meter reporting status is "on" or "off," a key example of the limitations of centralized analytics. The bandwidth for communicating status and other useful information from smart meters has been tied to vendor and proprietary data structures, and is, in turn, limited by customer communications. Decentralized architectures based on machine learning can propel this information to a new level. Incipient failure detection also requires detailed interpretation of high-fidelity sources, which is computationally intensive, so decentralization and communication of "mean time to failure" form a more efficient approach. New automated, online streaming algorithms, in the distributed environment will enable a) Classification of anomalies into key types that will inform a range of conditioned-based maintenance including vegetation management and transformer and switch replacement over a longer time period and b) Reduced loads on communications networks due to the distributed and local nature of the algorithms to be deployed.

C. Topology and Parameter Identification

Distribution utilities typically do not model the network structure below the medium-voltage distribution lines and hence do not have full observability of their network topology or detailed models of the thousands of low-voltage distribution components in play. Additionally, it can be difficult to track changes that occur to the medium-voltage distribution system, so the utility may not know the current state of the system. Lack of accurate knowledge of the current topology is problematic in a planning environment but especially challenging following extreme damage events, e.g. hurricanes. Further, accurate distribution topology and parameter estimation is necessary for improved situational awareness, control and optimization of DERs and Electric Vehicles and validation of line parameters. As placement of meters specifically for topology estimation is expensive, it is imperative to use the available power system data for identifying the network. Learning in this regime also needs to be amenable to prevalent availability of data, where sensors for voltage and power injections are placed only at a subset of the buses in the grid. In secondary circuit modeling for example a large portion of the per-unit voltage drop/raise occurs over the service transformers and lines that have large impedances, and accurate accounting for these is essential for an efficient environment. Significant portions of these errors can be attributed to the GIS system, which is typically corrected with manual inspection, requiring considerable personnel hours and resources. This can also be challenging to perform with underground wiring and inside customer owned buildings [9].

The current state-of-the art is to assume that the topology in the distribution grid is correct for planning studies, but also assumes a conservative large margin of error in the studies. This can lead to an increased interconnection cost as the conservative approach breaches more requirements and requires mitigation. In transformer modeling at the low voltage side, the models assume a fixed voltage drop or include a typical model.

[10] have demonstrated that theoretically correlations in nodal voltage magnitudes can be used to design greedy algorithms that provably learns the operating topology of radial grids. One benefit of this effort is its easy extension to the case where voltage and injections measurements are available only

from a subset of the grid nodes, provided missing nodes are separated by two or more hops in the operational tree. The learning algorithm is based on trends in fluctuations in voltages that arise from fluctuations in loads and get propagated by the power flow relations (Fig. 3). Proof-of-concept algorithms have also been demonstrated using simplistic real datasets for parameter estimation and topology detection [9].

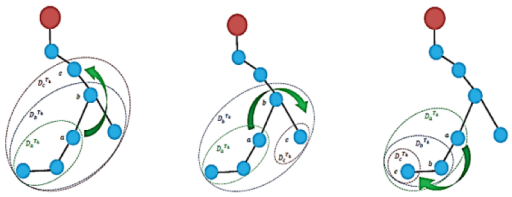

Fig. 3. Topology detection algorithms deployed on the distribution grid

D. Summary of Case Studies

At the building to grid interface, the ability of customers to transact or exchange resources is being considered as a new operational paradigm, both individually and in clustered aggregate systems with the grid. At the heart of this structure is controllable load, demand response (DR) and DER. Other new machine learning approaches to parameter estimation could include self organizing maps for outlier and bad data detection, random forest for topology identification and robust regression for grid parameter estimation all being developed through the Grid Modernization Initiative work. In all these methods, the accuracy of the metering and sensing is a challenge, which learning based methods distributed through multi-variate sources can seek to improve significantly. There is thus a benefit to utilizing new techniques which predict and prescribe.

To summarize, an automated, streaming algorithm for learning distribution network topology will enable a) improvements in control and optimization of distribution grid resources b) validation of topology switching action by system operator and c) increased situational awareness and system restoration following extreme damage events. Consumer benefits will include improved power quality, and less outage time during extreme events, lesser interconnection costs through enablement of better control actions for stability and voltage regulations.

VI. Benefits and Stakeholders

This paper has presented a vision of the future distribution grid and buildings operation as a cohesive environment where customers are rewarded for utilizing behind-the-meter distributed energy resources, and the distribution grid is enabled as an automated, reliable, safe, resilient energy-transport vehicle with high penetration of DER. The rewards to customers and other end users include energy cost savings and enhanced grid reliability and resiliency. These benefits are interlinked; for example, enhanced reliability could reduce customer energy cost by reducing unplanned solar-photovoltaic (PV) outages caused by voltage variability.

For the benefits assessment, we identified some of the areas in which grid and building applications would benefit from enhanced information derived from local data provided at the distributed level. We also identified where distributed analytics would benefit both the building and grid versus a centralized approach. These included availability of markets-based analytics for millions of resources, dynamic behind the meter controls to provide ancillary services.

In all of the case studies presented a combination of customer or ratepayer resources, and grid resources will be able to be coordinated and optimized. This will reduce the overall cost of energy to the consumer while increasing system reliability and reducing customer outage time. At the same time, advanced analytics will give grid operators useful visibility into the performance and controllability of these resources. The current lack of visibility of these behind-the-meter resources is a key obstacle to their participation in new and existing markets. If the operator cannot see a resource, the operator cannot use it [11]. The benefit to the grid must be co-optimized with the benefit to the customer (reduced energy cost). Co-optimizing these benefits is a challenge that new power-flow-linked local machine-learning-driven analytics are ideally suited to solve. The interoperable, data-vendor-agnostic approach of these analytics allows for a low-economic-barrier entry into the market for individual end users wishing to utilize their systems' flexibility to reduce their electricity bills. The customer themselves will not require detailed knowledge of the performance of the analytics, but more consider a framing of their needs for resources such as driving their car for a planned trip, or scheduling laundry. The benefits are not limited to customers who participate in behind-the-meter and transactive markets. There are significant numbers of customers who may not participate in these markets but will be affected by the development of these markets. For example, if the cost of electricity increases because of additional maintenance or integration costs associated with high penetration of DER, all customers will be directly affected. Therefore, we consider here the benefits to the ratepayer of enhanced reliability and resiliency and therefore reduced or at a minimum stabilized cost of being served by a particular utility [12].

Ratepayers reap substantial savings from grid and grid-to-building interface performance enhancements and new markets because: (1) Visibility into the instantaneous generation of intermittent renewable sources can assist in optimal operation of both the grid, buildings and the renewable resources. Variability can be tracked by observing the localized performance of PV units, allowing scheduling of charging/discharging of grid storage and EVs in an automated fashion that accounts for transmission and distribution constraints as well as opportunities for participation to achieve wider system objectives. (2) Communication to the control center of information on the aggregate controllability of a node relieves system operators of a computational burden and facilitates seamless integration of this information into operators' current decision-making process, reducing requirements for significant information technology services that might otherwise be required in a data-rich utility environment. (3) Aggregation of resources at a node can facilitate scheduling of those resources in a complementary manner and thus reduce stress on the grid, enabling benefits from condition-based maintenance rather than the "run to failure" approach. A potential failure could be predicted and DER utilized to extend system operating lifetime or to reduce load and enable reconfiguration to avoid failure. (4) Operators

of larger commercial buildings have an increased choice of markets in which they can participate.

Examples of grid regulation schemes that are improved with enhanced application of grid informed machine learning include a) Providing reactive power in addition to kilowatt-hour-driven schemes, reducing need for traditional regulation equipment such as local capacitors b) Controlling voltage and providing accurate knowledge of voltage, which enables local participation in voltage regulation c) Improving voltage and power quality at the building, enabling more up time for generation resources d) Saving energy for the building owner by allowing participation in new markets and better quantification of resource availability at all grid levels.

VII. SUMMARY

Generic "big data black box" machine-learning approaches can only be a starting point for this work. New, innovative machine-learning approaches are needed that incorporate complex constraints imposed by engineering, physical, communication, security, and other principles unique to power systems. Power system data are being collected from a variety of sources, including Phasor Measurement Units (PMUs), meters, outage-management systems, supervisory control and acquisition data recorders, weather, and DER. The diversity in volume and time "stamps" of these data sets are important to the findings derived from the data. Additionally, many utilities are tapping into other data sets, including smart meters, call centers, social media, billing systems, and mobile apps to support grid planning and operations. The volume of data from these and additional data sets is expected to grow exponentially in the next few years. Machine-learning analytics would support each area of grid modernization by using this growing volume of data to improve detection of normally invisible phenomena, learn grid topology, and support security applications including detection of physical or cyber-based attacks. Machine-learning analytics will also enable resilience and reliability applications, for example predictive models for responding to hazards or models for reconstructing events. Lack of useful operational visibility and information from extensive disparate data sources will drive the need for forward-thinking integrated approaches. The "brute force" approach to data collection is to analyze every node with bigger and better computing, which is often not available to utilities and customer-facing industries. A more efficient approach is to combine data sources with metadata; aggregate and fuse the data; incorporate buildings physics (for example air-flows); and develop sophisticated, relevant, and computationally efficient analytics.

In summary, the overall goals of this activity within the Grid Modernization Initiative are to make use of new and improved machine learning-based analytics to improve the state of the art in the fields of building and grid science. This will be completeed by evaluating operator and customer information that would benefit from a distributed, learning-based approach rather than centralized analytics, and with minimum investment, improve the visibility, observability, verification, and validation of the building-to-grid resource.

ACKNOWLEDGMENT

This document is an overview and review of current technology and potential gaps for application of machine learning techniques to the interface of building and power distribution systems. The current version was developed as a collaborative effort across the DOE national laboratory system through support provided under the Department of Energy Grid Modernization Initiative. More specifically, the effort is an early product of a project awarded under the DOE Grid Modernization Laboratory Consortium Laboratory Call issued in 2015 and announced in January of 2016. The project team includes the following national laboratories: Lawrence Berkeley National Laboratory (LBNL), Los Alamos National Laboratory (LANL), National Renewable Energy Laboratory (NREL), Argonne National Laboratory (ANL), Oak Ridge National Laboratory (ORNL), Sandia National Laboratory (SNL), Lawrence Livermore National Laboratory (LLNL). The authors gratefully acknowledge the input and guidance from Joseph Hagerman and Christopher Irwin in developing the ideas in this work.

REFERENCES

[1] Department of Energy. 2015. "Grid Modernization Multi-Year Program Plan". November. Available Online < https://energy.gov/downloads/grid-modernization-multi-year-program-plan-mypp>

[2] E.M. Stewart, E. Kara, C.M Roberts. "Contextually Supervised Generation State Estimation (CSGSE)" Provisional U.S. Patent 62/311,319, Submitted March 21, 2016

[3] A. Aswani, N. Master, J. Taneja, D. Culler, C. Tomlin. 2012. "Reducing transient and steady state electricity consumption in HVAC using learning-based model-predictive control." Proceedings of the IEEE, pp 240-253.

[4] M. Behl, A. Jain., R. Mangharam, 2016. "Data-driven modeling, control and tools for cyber-physical energy systems." 2016 ACM/IEEE 7th International Conference on Cyber-Physical Systems (ICCPS) 11 April

[5] D. Sadigh, K. Driggs-Campbell, A. Puggelli, W. Li, V. Shia, R. Bajcsy, et al. 2014. Data-driven probabilistic modeling and verification of human driver behavior. AAAI Spring Symposium - Technical Report, SS-14-02, 56 - 61. UC Berkeley: 701392

[6] E.M. Stewart, C.M. Roberts, A. von Meier, A. McEachern. O. Arkadian, 2016. "Predictive Distribution Component Health Monitoring with Distribution Phasor Measurement Units" Submitted to Renewable and Sustainable Energy Reviews.

[7] S. Misra, M. Vuffray A. Lokhov, M. Chertkov, 2017. "Towards Optimal Sparse Inverse Covariance Selection through Non-Convex Optimization", https://arxiv.org/abs/1703.04886, submitted to ICML (International Conference on Machine Learning)

[8] J. Bento, J. Pereira, M. Ibrahimi, A. and Montanari. 2010. Learning networks of stochastic differential equations. In Advances in Neural Information Processing Systems, pp 172-180

[9] J. Peppanen, M. Reno, R. Broderick, S. Grijalva, S., 2016 "Distribution System Model Calibration with Big Data from AMI and PV Inverters,"IEEE Transactions on Smart Grid. September pp 2497-2506

[10] D. Deka, M. Chertkov S. Backhaus. 2016. "Learning topology of the power distribution grid with and without missing data," in European Control Conference (ECC), June. pp 313-320.

[11] E.M Stewart, J. MacPherson. T. Aukai, 2013, Analysis of High - Penetration Levels of Photovoltaics into the Distribution Grid on Oahu, Hawaii. Detailed Analysis of HECO Feeder WF1. National Renewable Energy Laboratory

[12] K. Eber, D. Corbus. 2013. Hawaii Solar Integration Study: Executive Summary, National Renewable Energy Laboratory.