EPRI | ELECTRIC POWER RESEARCH INSTITUTE

I4Gen Project 2: Data Analytics—Key Learnings, Update 1

3002014148

I4Gen Project 2: Data Analytics—Key Learnings, Update 1

3002014148

Technical Update, June 2018

EPRI Project Manager

S. Seachman

ELECTRIC POWER RESEARCH INSTITUTE
3420 Hillview Avenue, Palo Alto, California 94304-1338 ▪ PO Box 10412, Palo Alto, California 94303-0813 ▪ USA
800.313.3774 ▪ 650.855.2121 ▪ askepri@epri.com ▪ www.epri.com

DISCLAIMER OF WARRANTIES AND LIMITATION OF LIABILITIES

THE FOLLOWING ORGANIZATION, UNDER CONTRACT TO EPRI, PREPARED THIS REPORT:

M&S Consulting

This is an EPRI Technical Update report. A Technical Update report is intended as an informal report of continuing research, a meeting, or a topical study. It is not a final EPRI technical report.

NOTE

ACKNOWLEDGMENTS

The following organization, under contract to the Electric Power Research Institute (EPRI), prepared this report:

M&S Consulting
235 High Street, Suite 515
Morgantown, WV 26505

Principal Investigator
J. Mason

This report describes research sponsored by EPRI.

The contributions of L. Shelton, the analytics specialist for the work presented in this report, are gratefully acknowledged.

This publication is a corporate document that should be cited in the literature in the following manner:

I4Gen Project 2: Data Analytics—Key Learnings, Update 1. EPRI, Palo Alto, CA: 2018. 3002014148.

ABSTRACT

This report summarizes the key learnings from the first technical update in a series of reports being developed by the Electric Power Research Institute's (EPRI's) Insight through the Integration of Information for Intelligent Generation (I4Gen) initiative, which addresses topics related to analytics tools and techniques. In this report, advanced analytics techniques from the disciplines of artificial intelligence and machine learning are described in the context of applications for predictive maintenance in power generation facilities. Specific analytics software applications are described, as are common terms and techniques that will be helpful to understand when evaluating such products. Future updates to this report will cover additional analytics software and techniques and will assess these applications using the criteria presented herein.

Keywords
Artificial intelligence
I4Gen
Machine learning
Neural networks
Predictive analytics
Predictive maintenance

Deliverable Number: 3002014148

Product Type: Technical Update

Product Title: I4Gen Project 2: Data Analytics—Key Learnings, Update 1

PRIMARY AUDIENCE: Plant personnel and monitoring and diagnostics (M&D) center personnel who are exploring options for analytics techniques used for predictive anomaly detection, diagnosis, and prognosis

KEY RESEARCH QUESTION

The data analytics tools now in widespread use for M&D tasks have a limited ability to analyze complex data sets. However, with the increasing demands that flexible operations place on plant assets, utilities are beginning to explore advanced data analytics tools that can manage disparate and complex data sets. The utilization and maturity of these data-driven tools in the power industry are currently at very low levels, and their adoption and benefits cannot be fully realized unless the tools can "learn" from the numerous existing component configurations and provide prognostics, including information about remaining useful life. Therefore, determining the most valuable data analytic approaches for targeted application at power generation facilities is paramount for demonstrating the most impactful elements of the I4Gen vision.

RESEARCH OVERVIEW

The objective of the research presented in this report is to augment M&D for utilities using data analytics tools focused specifically on the power industry. This project seeks to evaluate several technologies in the area of predictive analytics to provide early detection, diagnosis, and prognosis of faults on a specific gas turbine type. The project will use historical data provided by utilities to complete an initial evaluation of the technology.

KEY FINDINGS

- Advanced analytics techniques from the disciplines of artificial intelligence and machine learning are described in the context of applications for predictive maintenance in power generation facilities.
- Specific analytics software applications are described, as are common terms and techniques that will be helpful to understand when evaluating these products.
- Future updates to this report will cover other analytics software and techniques and will assess these applications using the criteria presented in this report.

WHY THIS MATTERS

This report is one deliverable of the broader I4Gen (Insight through the Integration of Information for Intelligent Generation) initiative, the objective of which is to define the components and techniques required for a digitally connected, dynamically optimized power plant. The desired output of the solutions targeted in this study is the quantified likelihood of specific failure modes of major assets—specifically, gas turbines at multiple participating utilities—over a given timeframe, along with the associated degree of confidence. In addition to assessing the predictive performance of the software solutions, the study is evaluating them for ease of deployment, execution speed, usability, support, and other criteria.

HOW TO APPLY RESULTS

The I4Gen project will ultimately evaluate several technologies and their ability to predict, identify, and diagnose a component's faults. If successful, the process will provide the opportunity to fully understand the technologies available and open further opportunities to apply the techniques to additional components. The results of the project will also help utilities understand the process involved in implementing the technologies. Anyone who is starting to explore analytics to improve their business would see benefit from this report.

LEARNING AND ENGAGEMENT OPPORTUNITIES

- *I4Gen Supplemental Suite* (EPRI product 3002009264)
- *I4Gen Project 1—Technology Fact Sheets* (EPRI product 3002012001)
- *Analytics for Predictive Maintenance in the Power Generation Industry* (EPRI product 3002011017)

EPRI CONTACT: Steven Seachman, Senior Technical Leader, sseachman@epri.com

PROGRAM: Instrumentation, Controls, and Automation, P68

Electric Power Research Institute
3420 Hillview Avenue, Palo Alto, California 94304-1338 • PO Box 10412, Palo Alto, California 94303-0813 USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1
INTRODUCTION

Data analytics is the process of preparing and studying data to discover patterns that can provide actionable information. *Data analytics techniques* are employed to derive meaning and value from vast and complex data sets. The domain of *predictive analytics* involves the application of pattern and anomaly detection techniques to predict future events. These techniques include the practical application of statistical methods and machine learning algorithms. Related, overlapping areas of interest include artificial intelligence, forecasting, and predictive modeling. For power utilities, these techniques are used for prediction of asset or component failure as part of a predictive or condition-based maintenance program. Successful deployment of predictive maintenance leads to lower operation and maintenance costs due to longer equipment lifespans. Tools in common use for power generation monitoring and diagnostics (M&D) are limited in their ability to analyze complex data sets. However, with the increasing demands of flexible operations on plant assets, some utilities are beginning to explore the use of advanced data analytics tools that can manage disparate and complex data sets to augment their M&D programs. In addition, it is desirable for these analytics tools to identify operation states that have not been previously recognized by subject matter experts (SMEs). Moreover, the utilization and maturity of these tools in the power industry are very low. Potential adoption and benefits of these data-driven tools cannot be fully realized unless the tools can "learn" from the numerous component configurations that exist in power plants and provide prognostics and remaining useful life information for plant assets. The targeted application of data analytics for power generation is currently limited. Determining the most valuable data analytic approaches is paramount for demonstrating the most impactful aspects of the EPRI I4Gen vision.

Breakdowns of capital-intensive equipment lead to safety risks and lost generation capacity. When preventive maintenance can be scheduled in advance, costly damage can be avoided, thus maximizing equipment uptime. However, asset health is difficult to predict, especially under the varying operating conditions encountered in real-world power plants. In this study, various advanced analytics techniques and tools are described and assessed.

This study is a component of a broader I4Gen initiative. The objective of I4Gen is to define the components and techniques required for a digitally connected and dynamically optimized power plant. The key objectives of this study are as follows:

- Create and describe the process in which data can be transferred, filtered, and normalized to make plant data useful for data analytics tool sets currently used and validated in other industrial settings
- Evaluate several data analytics methodologies and their usefulness to M&D and plant asset health

- Demonstrate the benefits of data analytics, which may include event detection during transient conditions
- Evaluate the use of data analytics tools using historical plant data provided by participating members
- Engage SMEs and existing databases to develop a data-driven condition assessment from observed events

This report contains an evaluation of several predictive maintenance software solutions. The desired output of these solutions is the likelihood of specific failure modes of major assets, specifically gas turbines at multiple participating utilities, over a given timeframe, along with degree of confidence. In addition to evaluating predictive performance, the software solutions are evaluated for ease of deployment, execution speed, usability, support, and a number of other criteria as described in the Section 4 of the report.

2
ANALYTICS WORKFLOW

Current and historical data are the primary inputs to a predictive maintenance system. Prior to data ingestion, the raw data must first be cleansed and standardized. Data governance and master data management techniques can be applied to ensure the level of quality required for accurate predictions. The output of predictive analytics can consist of warnings, calculations of the likelihood of asset failure, and other useful information such as estimates of remaining useful life. These outputs can be used to make informed decisions for maintenance prioritization and planning. Collectively, these steps form a data collection-to-analysis-to-action process designed to meet the condition-based maintenance needs of a power generation fleet. Consumers of the information include engineers, management, and operators in the field. Figure 2-1 provides a simplified overview of the data analytics process.

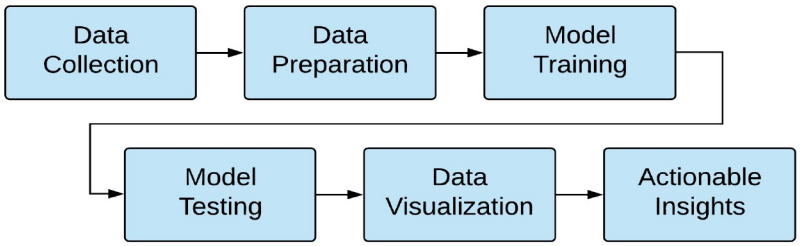

Figure 2-1
High-level analytics workflow

Predictive analytics is the process of using data analytics in conjunction with statistical and machine learning techniques to identify patterns and make predictions based on a large set of measurements, or *observations*. The aim is to apply statistical or machine learning techniques to create a quantitative prediction of the likelihood of specific future events, such as failure modes of plant equipment. The process utilizes large heterogeneous data sets to prepare *models* that can generate actionable outcomes to support business decisions. Models consist of *algorithms*, meaning a series of computational and logical steps, and *parameters*, which are numeric values that can be tuned to optimize the performance of the model. The more accurate a model is at predicting the likelihood of future events, the better the model.

As shown in Figure 2-1, the predictive analytics workflow includes several steps in compiling the final results:

1. **Data collection.** Acquire data from distributed control systems (DCSs), sensors, plant information (PI) historians, or other sources.
2. **Data preparation.** Filter, clean, enrich, and prepare data and make it accessible for analysis.
3. **Model training.** Determine the algorithms and parameters that yield accurate predictions.
4. **Model testing.** Apply the model and determine prediction accuracy for new samples.
5. **Data visualization.** Generate and interact with graphs, reports, and dashboards.
6. **Actionable insights.** Determine what needs to be done and how best to do it.

A brief overview of these steps as they relate to advanced analytics tools and techniques is presented in the sections that follow. For more information on these concepts, as well as predictive maintenance and related analytics technologies, refer to EPRI report 3002011017, *Analytics for Condition-Based Maintenance in the Power Industry*.

Data Collection

The first step in an industrial analytics workflow is typically *data collection*, whereby data is acquired from various data sources and stored in an operational data repository for further downstream processing. In most utilities, the operational, measurement, and utilization data is collected from a DCS or directly from sensors and plant equipment. Data can also be collected from software application databases or online data services. Data sources can include enterprise applications, such as asset data from an accounting system, or work order data from a work management system, or external data feeds, such as weather or market data.

Data Preparation

In addition to data stored in an operational historian, or PI database, data that can be helpful for detecting potential problems is usually distributed across multiple systems and stored in many different data formats. Once data is collected from equipment and sensors, the next step in the analytics workflow is to extract the raw data from the PI systems, enterprise applications, and external data sources and store the data in a location where it can be explored using query tools and machine learning programs.

Preparation of data involves standardizing, formatting, and cleansing data from various source systems to prepare it for use by predictive maintenance software. Any necessary advanced data transformations can be applied at this step. One or more applications may determine which extract, transform, load (ETL) path the data will follow. The data can be transferred from the plant to the technology provider as a batch, through real-time streaming, or via a combination of both. The data may also be sent to more than one downstream system.

Fast-moving data such as sensor data can be streamed real-time for immediate use but also to the data warehouse for in-depth analysis and inclusion into the data repository. The transferred data may be preprocessed to identify and correct or remove data spikes, missing data, or anomalous data points.

Power utilities represent critical infrastructure and as such fall under extensive compliance requirements. Accordingly, utilities have valid concerns about loading data to external partners or hosting providers, such as cloud-based asset health monitoring and predictive maintenance services. *Data masking techniques* can be applied to mitigate risks when leveraging services from these external entities. Data masking can also protect sensitive information from internal users who do not have a valid “need to know” the original unmasked data values.

Data can be captured and stored in a variety of formats, depending on application. For instance, *structured data* has traditionally been stored in relational databases. *Unstructured data* may be stored in file formats such as audio, spatial, image, or video. Free-format text data is often stored in application memo fields or document files, such as incident reports, on shared network drives. In addition to these data formats, *semi-structured data* files, such as log files, XML files, or other formatted text files, are popular formats for transferring machine-generated data.

Data quality issues should be addressed in the data preparation step. Incorrect data and missing observations are often the natural result of sensors working in harsh industrial environments. Machine learning techniques can be particularly sensitive to data quality issues, lending credence to the old adage, “garbage in, garbage out.” Poor data quality can lead to significant increases in false alarms, or to worse impacts, such as incorrect conclusions that have significant adverse consequences.

Although sensor malfunctions or technical issues could be root causes of data quality issues, often such problems are evidence of a larger problem with the operational process that is being monitored. It is important to have procedures in place to automatically detect anomalies and potential data quality issues, and to alert appropriate personnel to investigate such issues. Raw data can be given more context, or *enriched*, by many different techniques, such as combining data from multiple sources, applying standardized formats, or replacing obscure lookup codes with meaningful descriptions. Some data quality issues can be mitigated by enriching the data by applying *imputation* techniques, whereby anomalous values in a data set are replaced by average or default values, or other calculated values. For example, it may be common to replace null pressure measurement values with a zero, thus significantly altering statistical values (e.g., mean, standard deviation) used by machine learning algorithms. By applying imputation techniques where appropriate, prediction accuracy can be meaningfully improved.

Model Training

Once the data required for a particular problem of interest has been prepared, model training can begin. A *model* is a combination of variables, known as *features*, algorithms, and parameters. The parameters can be tuned to optimize the prediction accuracy or ability to detect anomalies. When sample data is made available to train and test a model, the data set is often split into training samples and test samples. The majority of observations, typically 50 to 90 percent, are generally used for training purposes. In general, the more data available for training, the better.

Model Testing

Once the models have been tuned during the training process, the models can then be evaluated using the test data set. Accuracy of predictions is measured by determining the frequency of incorrect predictions, which occur in the form of *false positives* and *false negatives*. In the event of a false positive, the model indicates there is an issue that a plant operator needs to investigate, and it turns out to be a false alarm. The over-abundance of false positives is a common problem encountered with advanced pattern recognition (APR) systems and other traditional methods. By contrast, false negatives occur when an issue in the plant goes undetected. As with human health in cases when a diagnostic measure fails to detect an illness and the condition is left untreated, false negatives can lead to serious problems when assessing asset health within the power plant, especially when potentially disastrous maintenance problems are overlooked, or not noticed until it is too late.

Data Visualization

Once developed and tested, predictive models can be executed, and results presented for end users to explore or visualize. The model can be employed to trigger alerts that are to be investigated by engineers or operators. Analytics tools provide reports and dashboards that are interactive, with parameters and filters that can be modified to drill down into specific subsets

of the source data and predicted outcomes. New analytical reports can be developed and presented to the user community to gain further insights. Analytics tools also provide the ability to export results including charts, annotations, and raw data to file for sharing with stakeholders for further review.

Subject matter experts (SMEs) can use data visualization tools to better understand the incorrect predictions and anomalies in the data samples. Their observations can be used to further refine the predictive models, a process known as *retraining*. Data visualizations are also used to deliver insights to management, engineers, operators, and other personnel so that they can make more informed decisions.

Actionable Insights

When the likelihood of a specific component failure mode exceeds a predetermined threshold, the next step is to automatically trigger an action to occur. Typical actions include alarms and workflows. Modern industrial analytics suites are integrated with enterprise systems to support such business functions as work order management, inventory control, and maintenance scheduling. These capabilities are not tested in this study, but such functionality is described for each software suite where applicable.

Alarms and Warnings

Effective analytics techniques will minimize the frequency of false alarms. The acceptable level of false alarms may vary according to the risk tolerance of specific equipment failure modes. Providing the ability to adjust alarm thresholds, frequency, and time windows is an essential control feature of predictive maintenance software. For example, a rule can be defined that says to ignore an excessive low temperature reading unless the preset limit is exceeded five times within a 5-minute period.

The techniques and systems evaluated in this study are not intended for triggering alarms for emergencies or managing the day to day, ongoing operations of the plant. Rather, these systems are intended to be utilized for planning purposes to schedule maintenance, repair, or replacement of plant assets—typically days, weeks, or even months in advance. The objective of predictive maintenance is to trigger alerts as accurately and as early in advance as possible, thus allowing planning time for determining and scheduling the best course of action.

Enterprise Systems Integration

Integration of analytics solutions with back-end enterprise systems can improve the overall efficiency of maintenance and repair operations. These enterprise systems include work management, supply chain, and accounting systems. For example, predictions of the likelihood of various failure modes can be very helpful inputs for cost optimization of maintenance schedules. These failure mode predictions are even more reliable when expected load and asset configuration settings are taken into account, as these factors can significantly impact equipment degradation rates. Accurate predictions of remaining useful life are also helpful for planning and scheduling purposes. Over time, the feedback from integrated analytics systems can be used to optimize the centrally managed maintenance policies.

In order to make fully informed decisions involving asset repair, replacement, and maintenance, the costs of these activities must be made accessible to decision makers. Enterprise asset management and cost accounting systems are necessary data feeds for assessing tradeoffs and prioritizing tasks. By integrating cost and inventory systems, adaptive service logistics can lead to a reduction in spare parts inventories. Parts can be ordered automatically by integrated back-end systems based on predicted demand.

Prescriptive Maintenance

Predictive maintenance indicates that a problem is likely to occur. *Prescriptive maintenance* goes a major step further in preventing unplanned downtime by prescribing a solution and identifying the tasks to be performed, along with relevant information needed to address the issue that has been detected. For example, predictive maintenance may detect that bearing temperature for a pump is increasing steadily and the pump is likely to fail in 3 to 5 weeks. With prescriptive maintenance, recommendations may be automatically generated that say to reduce throughput by 10%, in which case maintenance can be deferred another month and performed during a scheduled maintenance window.

From this example, it becomes apparent that information from sensors needs to be combined with data from enterprise systems, such as a maintenance scheduling system and asset management systems, to determine relative costs of repair versus replacement. With proper system integration, work orders can automatically be generated, providing the technician with specific instructions for repair along with links to product documentation specific to fixing the indicated problem.

3
PREDICTIVE ANALYTICS TECHNIQUES

Modern software solutions employ a blend of established, best practice predictive techniques and proprietary "black box" methods. The solutions evaluated in this study all detect anomalies, identify previously unknown correlations, and make predictions of asset failure and remaining useful life. To differentiate their solutions in a rapidly growing market, there is a virtual flood of techniques, technologies, and buzzwords that vendors use in describing their respective products. The more common terms and techniques are described in this section, so that an evaluator or prospective buyer of such predictive analytics solutions can make a more informed assessment.

Advanced Pattern Recognition

A key capability of predictive analytics solutions is the automated detection of hidden patterns. Advanced pattern recognition (APR) is a method that uses algorithms to detect anomalies, subtle changes, and decreased performance in real time. APR uses historical data from sensors to create models and calculate tolerances. APR techniques have been applied for use at some power plants for decades, but are generally considered to be most useful for triggering alerts in steady-state plant operations. The software vendors in this study strive to differentiate their solutions from APR systems by emphasizing their abilities to assess transient states and integrate with enterprise systems to provide more prescriptive capabilities.

Predictive Modeling

Predictive modeling is a process that employs data mining and statistical techniques to forecast outcomes. The model is composed of many variables that would likely influence future results and may utilize simple linear equations or complex neural networks to make the predictions. Predictive models can be developed based on the laws of physics, or they can be derived from large sets of data samples. Physics-based models are known as *first principle models*. The systems in this study primarily apply data-driven techniques, though the models can be seeded based on first-principle analysis and refined using sample data for like assets from throughout the organization or even across organizations via shared data agreements. The goal of data-driven models is to find relationships between the system state variables (input and output) without explicit knowledge of the physical behavior of the system.

Predictive models can be offline or online. Offline models are created and refined using batch processing techniques. Online models can be revised, deployed, and executed in near real time. The technology solutions in this study support both offline and online models.

Fault Diagnosis

The systems and techniques in this study employ intelligent fault diagnosis techniques to identify potential failures and classify them according to the most likely failure mode. Fault detection algorithms aid in the analysis of condition data in real time in order to detect failures as soon as they become visible in the condition data. The algorithms aim at identifying the location of fault and type of fault in order to schedule maintenance and, if necessary, identify and procure the parts and supplies needed for repair.

Artificial Intelligence

Artificial intelligence (AI) is the set of techniques that allow machines to be able to perform tasks in an "intelligent way." The futuristic promises of AI seen in science fiction movies, known as *general AI*, are years or decades away from being of practical use for industrial applications. However, *narrow AI* enables machines to perform human-like tasks in specific problem domains by learning from existing data and adjusting to new observations. AI applications often make use natural language processing (NLP) and machine learning techniques, which are described further in the sections that follow.

Natural Language Processing

Natural language processing (NLP) is a subset of artificial intelligence, focusing on functionality that helps computers analyze, interpret, and generate human language. It allows for a natural interaction between humans and computers using regular everyday language, either written or verbal. NLP uses a variety of techniques to interpret human language, including machine learning methods and other sophisticated algorithms. The process starts by breaking down language into smaller parts, or *tokens*, and then determines patterns and interactions between the tokens as compared to predefined templates that describe common language structures. NLP capabilities include content categorization, sentiment analysis, speech-to-text and text-to-speech conversion, document summarization, automated storytelling, and language translation.

Machine Learning

Machine learning is the application of artificial intelligence whereby machines can apply generalized algorithms to automatically learn from data samples, without being explicitly programmed to follow predefined rules. Data-driven machine learning has proven to be effective and relatively simple to prepare and validate compared to a large customized rule base. Machine learning has been successfully applied in a variety of industries, such as in credit scoring and fraud detection in the financial services sector. APR techniques follow the machine learning workflow and apply some common machine learning algorithms.

Machine learning utilizes both supervised and unsupervised learning techniques. *Supervised learning* is used when the outcomes (or output) for each observation are known. A supervised machine learning model therefore uses sets of previously evaluated input and output data to train the model, as is done when defining APR models. For example, data sets of measurement values for a gas turbine are collected, and each set of observations is labeled to identify whether the measurements indicate a real-world problem, and if so, the type of problem is identified. The model can then be applied to new data sets for testing and refinement. Supervised learning uses classification for categorical responses and regression for continuous numerical responses. Such outputs can be combined with simple rules to define numerical ranges or thresholds that can also be used for categorical purposes—for example, in setting alarm limits.

Unsupervised learning is used when there is training data available that does not have labeled responses, so there is no error or correction signal to evaluate the conclusions. The algorithms learn to identify patterns without any guidance. Clustering is the most frequent method used for unsupervised learning—for example, where closely associated groups, or clusters, of problem types automatically emerge from a large data set.

Kernel Methods

A *kernel* is a mathematical calculation that determines the similarity between two points of data. *Kernel methods* use kernel functions to solve machine learning and statistical problems. For example, data points where the kernel determines a high degree of similarity can be clustered together. Kernel methods have been shown to be effective for time-series data, as is common with sensor measurements, text analysis, and image classification. Support vector machines and kernel principal component analysis techniques utilize kernel methods.

There are a variety of kernels used in advanced analytics software, including the following:

- Polynomial kernel
- Fisher kernel
- Graph kernels
- Kernel smoother
- Radial basis function kernel (RBF)
- String kernels

Support Vector Machines

Support vector machines (SVMs) are supervised learning models used for classification. The goal is to find a dividing line, or *hyperplane*, that separates data points into two classes, making SVMs effective for classification problems. SVM approaches identify the widest possible gap between the two classes. If a perfect hyperplane does not exist, a few points are considered "misclassified," and a tuning parameter controls the width of the margin on either side of the hyperplane. In this case, if there are many points within a wide margin, there is low variance and high bias. See Figure 3-1 below, where Class A may indicate a problem with a turbine blade that needs to be addressed, and Class B represents normal operating conditions. The diagram shows a highly simplified example with two dimensions, or feature values, that are plotted on the x- and y-axes. The true power of SVMs is that they can support any number of dimensions, which is why the "hyperplane" is used to separate classes of data values. It is not uncommon to see SVMs used to analyze hundreds or thousands of features.

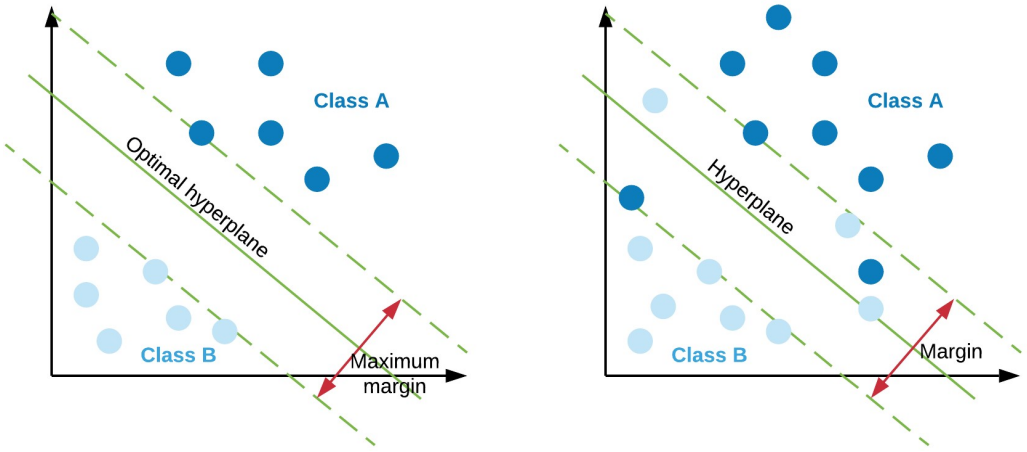

Figure 3-1
SVM vs. SVM with misclassification

Principal Component Analysis

Principal component analysis (PCA) is a machine learning technique that uses unsupervised learning to identify the factors that have the greatest impact on the successful operation or failure of an asset. PCA performs dimensionality reduction on complex data, so that trivial or redundant features can be omitted from further analysis. For example, PCA can be used to preprocess data, pruning a large data set down to a core subset of features so that a neural network can make accurate predictions more efficiently.

Conventional PCA does not evaluate records where any of the measurement values are missing. In large data sets with many features, it is common to have many null values, so conventional PCA may unnecessarily rule out large sections of the data set, including rows of data that may yield the best insights. To address this problem, probabilistic PCA uses imputation techniques to replace the null values with default or average values to standardize the data, allowing all of the data set to be fully utilized in the analysis.

Kernel principal component analysis (kernel PCA) is a variation of PCA that uses kernel methods, extending conventional PCA to support high-dimensional feature spaces. It allows the ability to extract nonlinear principal components without expensive computations.

Neural Networks

A *neural network*, also known as an *artificial neural network*, is a supervised machine learning architecture designed loosely based on the most basic structures of the human brain, with neuron-like *nodes* and synapse-like *connections*. Neural networks consist of few components but perform iterative, data-driven computations that perform complex classifications in a relatively short period of time. Neural networks are designed to be content agnostic. They can be applied to recognize patterns in text, images or other digital artifacts. Neural networks take large data sets as inputs as for training and tune the parameters of a model to optimize predictive accuracy. Neural networks can consist of thousands of interconnected nodes, which are stacked sequentially in layers, forming millions of connections for relaying information. Layers are usually interconnected with the layer of nodes before and after them in the information flow via inputs and outputs. (See Figure 3-2.)

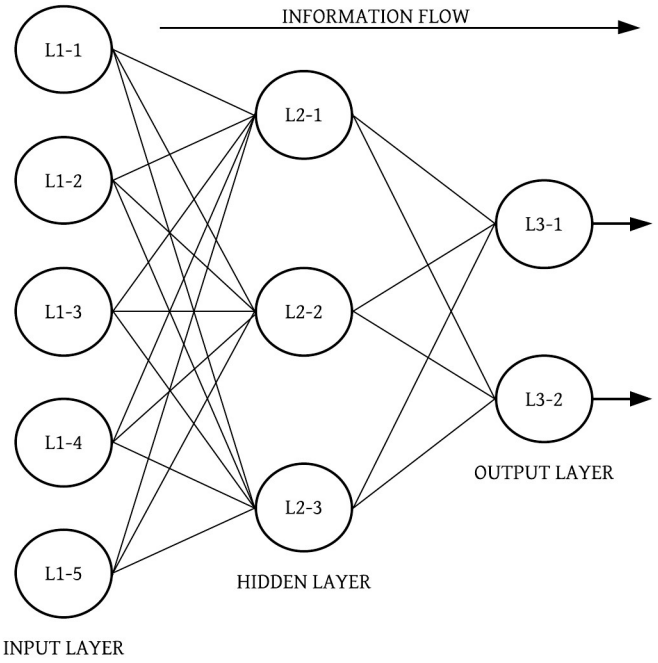

Figure 3-2
Neural network

Neural networks come in a tremendous variety of forms. Many neural networks apply supervised learning methods, whereby the neural network is fed large sets of training data that is already classified or labeled. The model attempts to predict the response, and if the prediction is wrong then the inner layer parameters are adjusted. The process is repeated until the model is able to identify responses with an acceptable degree of accuracy.

Deep learning is a technique that is growing in popularity. The majority of deep learning models are based on an artificial neural network. Put simply, deep learning can be considered simply as a neural network with more than one layer. Deep learning is frequently used for image and speech recognition.

Neural networks are sometimes referred to as a *black box*. Neural networks, or "universal approximators," can closely approximate the behavior of any function, However, it can be very difficult or even impossible to understand what specific logic is applied in the hidden layers between the input and output layers, especially when deep learning is used where more than one hidden layer is involved. Studying the hidden layers does not bring enough insights on the structure of the function being estimated. Therefore, full *interpretability* is not possible, meaning the specific rules that the neural network applies cannot be accurately documented for human understanding. The machine "just knows" the answer, based on the black box computations. Figure 3-3 provides a visual representation of how artificial intelligence, machine learning, and deep learning are related.

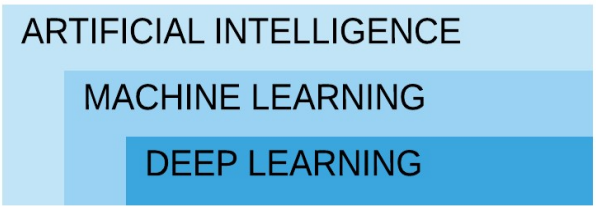

Figure 3-3
Relationship between AI, ML, and deep learning

Confusion Matrix

A confusion matrix is a classification table used for model assessment. (See Table 3-1.) Accurate classifications are True Positives and True Negatives. False alarms are False Positives.

Conditions that should have triggered an alert, but did not, are False Negatives. These categories are used to determine the *accuracy* of a predictive model for comparison purposes.

Table 3-1
Confusion matrix example

	Prediction = 1	**Prediction = 0**
True Value = 1	True Positive (TP)	False Negative (FN)
True Value = 0	False Positive (FP)	True Negative (TN)

Algorithms

Advanced analytics solutions make use of many different *algorithms*, or predefined sequences of computer instructions. Decision trees, genetic algorithms, and fuzzy algorithms are three types of algorithms used as differentiators by the tools evaluated in this study. These algorithm types are described further in the sections below.

Decision Trees

A *decision tree* is used to bring visibility and interpretability to a given decision-making problem. At the top of the decision tree is the root. The internal nodes are the conditions for splitting up data observations into categories based on the condition, with the tree splitting into branches known as *edges*. When there are no splits left, the end node is known as the *leaf,* where the final decision outcome is made available. To improve and simplify a decision tree so that accurate predictions can be made more quickly, the tree can be *pruned* by removing the less important branches.

The response variable can be discrete in a classification tree or continuous in a regression tree, which is why decision tree algorithms are known as *classification and regression trees* (CART). Some advantages of CART include easy to understand feature selection, and they require little data preparation. Some disadvantages include overfitting and instability due to high variance.

Genetic Algorithm

The genetic algorithm is based on natural selection. Natural selection picks out the fittest individuals from a population, and only those individuals reproduce. The genetic algorithm uses this idea and selects the best solution out of a set of solutions for a problem. There are five stages in a genetic algorithm:

- **Initial population.** The initial population is the set of solutions to the problem. Each solution is defined by parameters (variables). The solutions are represented by a string, usually 0s and 1s.
- **Fitness function.** The fitness function determines how fit the solution is. Each solution is given a fitness score.
- **Selection.** Pairs (known as parents) are selected based on fitness scores—the higher the better.
- **Crossover.** Crossover is the most important step in the process. For each pair, a crossover point is chosen at random. Offspring are created by swapping the data between the two parents until the crossover point is reached.
- **Mutation.** Offspring are subjected to mutation, which applies random changes to individual parents to maintain diversity.

Once the algorithm does not produce offspring that are significantly different from the previous generation, then the genetic algorithm has provided a set of solutions, or best model methods to solve the given problem.

Reference: towardsdatascience.com/introduction-to-genetic-algorithms

Fuzzy Algorithms

Fuzzy algorithms use "*degree of truth*" logic. Boolean logic uses a 0 for false and 1 for true. Fuzzy logic is different in that depending on the "degree of truth," the result could be any number between 0 and 1. The process begins with *fuzzification* of all input values. Then, the fuzzy output functions are computed and *defuzzification* is used to find the clear output values. (See Figure 3-4.)

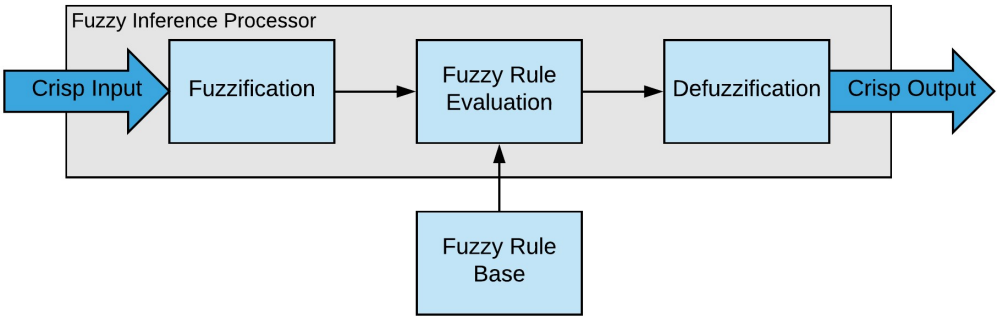

Figure 3-4
Fuzzy process

Key concepts related to fuzzification as shown in Figure 3-4 include the following:

- **Crisp input:** the real value measured by sensors, such as temperature or pressure
- **Fuzzification:** the process of using transformation functions to change a real value into a fuzzy value
- **Defuzzification:** the process of transposing the fuzzy outputs back to crisp, real values
  Reference: whatis.techtarget.com/definition/fuzzy-logic

4
SOFTWARE EVALUATION PROCESS

In this study, multiple analytics technology solutions are evaluated and compared. Standard data sets are processed by these software applications in a controlled network environment. Predictions of component failure related to one or more gas turbines at each participating utility are evaluated for accuracy as well as runtime performance. Additional functionality provided by each solution is also discussed for cases in which the functionality is potentially beneficial to the predictive maintenance program within a connected plant.

In-Scope Software Solutions

Several software solutions designed for predictive maintenance in the connected power plant were considered for this study. Each application takes data samples as input and, at a minimum, predicts the likelihood of failure within a given timeframe. Another key requirement is for the software to perform batch analysis of historical data for model training purposes. In some cases, participating utilities identified the technology solutions that they would like to see evaluated. Ultimately, predictive analytics solutions from DecisionIQ, SparkCognition, and STEAG were chosen to be evaluated in this study. It is likely that additional technologies will be assessed in an upcoming revision to this report.

Data Collection from Utilities

Each utility participating in this study is providing at least one year of historical data for the study. In each case, the data was extracted from a plant intelligence (PI) historian database and filtered down for a specific gas turbine. Data extraction from source systems is often a difficult, time-consuming process for power companies. A significant portion of data is stored in operational historian systems. These systems are designed for rapid data storage and perhaps some pattern matching at the time of data collection, but they are not designed for efficient data extraction for analytical purposes.

Data Content

Each participating utility is providing 12 to 15 months of historical data from its PI system. The data will contain all information available for one gas turbine per utility. Each of the technologies in the study is capable of processing this volume of data. In general, the more granular the data, and the more tags provided, the better.

Actual data is preferable to simulated data. However, in cases where failures have never occurred, or are extremely rare, simulated data may be provided. It is preferable that at least two or three instances of any anomaly or failure type be included within the data sets. Some rare failure modes may not have a sufficient number of historical observations to enable prediction, and other failure modes may not have occurred within the data samples at all.

For security purposes, utilities may prefer to mask data prior to making it available for analysis. The masking may be performed by the utilities or by EPRI. In addition, dates will be modified slightly, thus further masking such real-world occurrences. Technology vendors will not know which utility they are processing during any evaluation sessions.

Data Format

Data can be provided in a number of formats. These include comma separated values (CSV) text files, which are a commonly used method for PI exports. Other formats include database exports that are provided by the source system's database technology. For instance, MySQL dump files are a preferred method of export for MySQL-based historian systems. Other acceptable file formats include XML and JSON text files.

File Transfer

To transfer data from the utilities to EPRI for evaluation, one of the two following methods is utilized in each case:

- **Physical file transfer (Sneakernet).** Using this approach, a device such as a Drobo unit is connected to a computer within the utility. Files are copied to the device, and the device is carried or shipped to EPRI. Such units are capable of storing data in an encrypted format.
- **Secure FTP (SFTP).** The utility may prefer to make files available via SFTP. With this approach, external network access must be provided to EPRI. Uploads and downloads may take several hours or longer. The advantage of this approach is that physical devices do not need to be connected to systems on utility networks, so this approach may be preferable for some utilities.

Data Export Challenges

Exporting data from PI systems can be a challenge. Operational historian databases are designed for rapid ingestion of observation records, and not for bulk query extracts. However, most modern historian software applications do have export utilities to facilitate the process. Also, the volume of PI system exports, especially when exporting a year of historical data or more, can lead to very large export files. At the discretion of each utility, smaller exports can be performed (e.g., monthly or quarterly) and sent to EPRI as a set of multiple files.

EPRI Advanced Analytics Assessment Lab

EPRI has prepared an advanced analytics assessment lab that is hosted within a secure data center. (See Figure 4-1.) Data from each utility will be imported to an instance of OSISoft PI System running in its own virtual machine (VM). Similarly, each software application being evaluated will be installed in its own VM. All software application VMs will contain the same system resources in terms of memory size and computation power, though there can be variations in operating systems (Windows or Linux) and operating system configuration. The goal is to create a standard infrastructure setup that allows performance comparisons to be as fair as they can be, within reason.

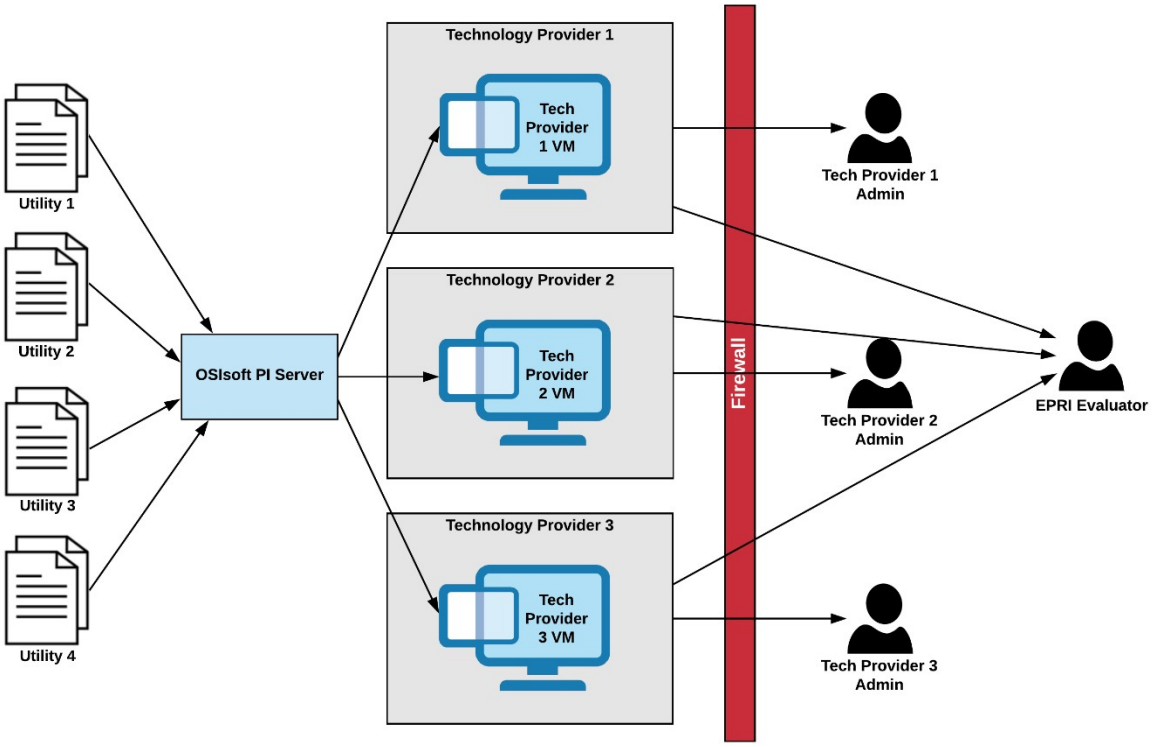

Figure 4-1
EPRI Analytics Assessment Lab

Evaluation Criteria

This study aims to document objective performance results and evaluate the technology products given the data sets provided by the participating utility companies. Test plans are created and executed to evaluate the performance on the following criteria:

1. Data Requirements
2. Functionality
3. Quality of Anomaly Detection
4. Quality of Predictions
5. Speed of Execution
6. Scalability
7. Batch and Streaming Data Feeds
8. Usability
9. Implementation
10. Software Maintenance and Administration

Data Requirements

For each technology solution, required data sources and formats are described. The data source review documents any new instrumentation that is needed. It is expected that at a bare minimum, text file exports from operational historian systems will be supported by each technology solution evaluated. Any necessary new infrastructure or hardware is described. Where applicable, the recommended approach for general data management is defined.

Functionality

The technology providers have been evaluated based on the functionality provided by their respective products. Such functionality includes the types of faults supported, anomaly detection, remaining useful life (RUL) predictions, and diagnostic capabilities. The assessment also seeks to evaluate the model training capabilities of each product in terms of whether the model requires supervised or unsupervised training, what parameters can be user-defined and to what extent, the amount of subject matter expert (SME) engagement (from vendor, utility, EPRI, or other thirdparty provider) that will be required through the setup and use of the technology, and flexibility regarding whether the technology will be delivered with utility-specific solutions or as a platform. Test plans in this study are designed to verify any functionality that the technology providers identify as being unique to their respective products.

Quality of Anomaly Detection

Anomalies such as failure events occur at various times within each batch data set used for evaluation. Such anomalies can be used to determine that a component has a potential problem worthy of further investigation. The approach to anomaly detection used by each technology solution is described, to the extent that each vendor is willing to divulge this information. Accuracy, false positives, and false negatives are reported and compared.

Quality of Predictions

The technology solutions are evaluated in terms of their ability to predict the likelihood of machine or component failures over a given timeframe. The accuracy of failure predictions is evaluated for each technology. Warnings are verified by comparing with actual results and input from subject matter experts. The timing of predictions is also evaluated. The earlier an accurate prediction is made, the better. If it is determined that a component is likely to fail within the next 60 to 90 days, it may be possible to replace the component during an upcoming scheduled outage.

The duration between the current time and the time of machine breakdown is referred to as the *remaining useful life (RUL)* of the machine. Extending the productive life of an asset can lead to significant savings in capital. A calculation of RUL is a prediction based on several factors, such as expected workload and operating conditions.

Speed of Execution

Each technology provider is given access to standard data sets containing PI data from each utility, hosted within the EPRI network. In addition, each solution is deployed on a predefined, standardized system architecture, where the same processors, memory, and network configuration are deployed for each test. Any required modifications to the standard data sets or hardware will be noted. System requirements for each technology solution will be described and evaluated, data loading performance will be timed, and the execution time required for each technology solution to process each utility data set will be recorded and evaluated.

Scalability

The volume of data generated by environment and condition sensors, as well as operational output from equipment and processes within a utility, can be significant. It is important that any predictive analytics systems and approaches are capable of scaling to meet the quantity of devices to be managed. Architectural and cost implications of adding turbines or plants are discussed. In addition, the implications of increasing the frequency or granularity of sensor measurements are considered.

Batch and Streaming Data Analysis

This assessment documents the ability of each technology solution to be able to analyze batch historical data as well as streaming live data. For batch loads, frequency and volume considerations are described. For streaming data feeds, the supported underlying streaming technologies are identified, but streaming functionality will not be tested as part of this study.

Usability

It is important that an analytics system be practical to use by its intended audience. In an industrial setting, user roles include engineers, operators, and system administrators. Usability leads to efficiency and productivity. An easily usable system can help to address the skills gap within an organization if the tool reduces the need for technical and data science expertise required to effectively conduct a data-driven maintenance program. The look and feel of multiple analytics systems are qualitatively assessed and features are compared in this study.

Implementation

This study will provide information to help utilities understand the additional effort and costs related to implementing each technology solution. The computational resources required by each technology provider are documented, including details pertaining to any server nodes and cluster requirements, new infrastructure that may be required, server location, server access, and security protocols followed. Support for cloud and on-premises deployments are evaluated. In addition, the network bandwidth and capacity required by each technology solution as well as any commonly deployed variations in system architecture will be discussed. Features that simplify the implementation process, or any related difficulties encountered, will also be described.

The tools reviewed in this study are components of larger software suites that provide additional capabilities that would be of interest to decision makers considering an advanced analytics solution for predictive maintenance. For instance, this report evaluates predictive maintenance technologies based on batch processing of historical measurement data. A separate, related study under the EPRI I4Gen program is planned for evaluation of analytics tools that use streaming analytics technologies to trigger alerts in near real time.

Software Maintenance and Administration

The analytics solutions in this study can be significant investments in terms of software licensing, setup, and ongoing administration. These latter costs tend to be less understood when a software procurement decision is being made. Therefore, approaches to common maintenance tasks are described. Such tasks include patching, software updates, and configuration tuning. Documentation quality and support provided by the vendor are also evaluated.

5
ADVANCED ANALYTICS SOLUTIONS

Multiple advanced analytics solutions are evaluated in this study, including technologies from DecisionIQ, SparkCognition, and STEAG Energy services. Each solution takes historical operational and maintenance data as input and predicts the likelihood of equipment failure and triggers alerts when significant problems or anomalies are detected that may need attention. To make such predictions, each technology solution takes a *black box* approach, utilizing proprietary algorithms tuned using parameters known only to the software provider. Some solutions are fully data-driven, and others apply rules learned from physics or from assessment of particular industrial equipment types. The differences in algorithms and their turning parameters lead to prediction performance variations across the software solutions in this study.

Table 5-1 presents a brief overview of the advanced analytics vendors and their respective solutions evaluated in this study.

Table 5-1
Advanced analytics solutions overview

	DECISIONIQ	sparkcognition™	steag
Founded	2014	2013	1937*
Product	Genesis Platform	SparkPredict	SR::Suite
Industry Focus	Manufacturing, field service, and energy	Energy, oil and gas, manufacturing, finance, aerospace, and security	Renewable and conventional energies, nuclear technologies
Standard Deployment	Cloud	On-premises, hybrid, or fully on cloud	On-premises

*Analytics software developed in recent years

Export Control Restrictions

Access to and use of EPRI Intellectual Property is granted with the specific understanding and requirement that responsibility for ensuring full compliance with all applicable U.S. and foreign export laws and regulations is being undertaken by you and your company. This includes an obligation to ensure that any individual receiving access hereunder who is not a U.S. citizen or permanent U.S. resident is permitted access under applicable U.S. and foreign export laws and regulations. In the event you are uncertain whether you or your company may lawfully obtain access to this EPRI Intellectual Property, you acknowledge that it is your obligation to consult with your company's legal counsel to determine whether this access is lawful. Although EPRI may make available on a case-by-case basis an informal assessment of the applicable U.S. export classification for specific EPRI Intellectual Property, you and your company acknowledge that this assessment is solely for informational purposes and not for reliance purposes. You and your company acknowledge that it is still the obligation of you and your company to make your own assessment of the applicable U.S. export classification and ensure compliance accordingly. You and your company understand and acknowledge your obligations to make a prompt report to EPRI and the appropriate authorities regarding any access to or use of EPRI Intellectual Property hereunder that may be in violation of applicable U.S. or foreign export laws or regulations.

The Electric Power Research Institute, Inc. (EPRI, www.epri.com) conducts research and development relating to the generation, delivery and use of electricity for the benefit of the public. An independent, nonprofit organization, EPRI brings together its scientists and engineers as well as experts from academia and industry to help address challenges in electricity, including reliability, efficiency, affordability, health, safety and the environment. EPRI members represent 90% of the electric utility revenue in the United States with international participation in 35 countries. EPRI's principal offices and laboratories are located in Palo Alto, Calif.; Charlotte, N.C.; Knoxville, Tenn.; and Lenox, Mass.

Together…Shaping the Future of Electricity

3002014148

Electric Power Research Institute
3420 Hillview Avenue, Palo Alto, California 94304-1338 • PO Box 10412, Palo Alto, California 94303-0813 • USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com