INL/CON-21-62716-Revision-0



# INTEGRATION OF DATA ANALYTICS WITH SYSTEM HEALTH PROGRAMS

September 2025

Diego Mandelli, Congjian Wang, Steve Hess, Rosmary Sugrue, David Morton , Ivilina Popova, Chad Pope, Daniel Cole

Changing the World's Energy Future

Idaho National Laboratory

*INL is a U.S. Department of Energy National Laboratory operated by Battelle Energy Alliance, LLC*

INL/CON-21-62716-Revision-0

# INTEGRATION OF DATA ANALYTICS WITH SYSTEM HEALTH PROGRAMS

**Diego Mandelli, Congjian Wang, Steve Hess, Rosmary Sugrue, David Morton , Ivilina Popova, Chad Pope, Daniel Cole**

**September 2025**

**Idaho National Laboratory**
**Idaho Falls, Idaho 83415**

**http://www.inl.gov**

**Prepared for the**
**U.S. Department of Energy**
**Under DOE Idaho Operations Office**
**Contract DE-AC07-05ID14517**

# INTEGRATION OF DATA ANALYTICS WITH SYSTEM HEALTH PROGRAMS

**D. Mandelli**[†] **and C. Wang**
Idaho National Laboratory
{diego.mandelli, congjian.wang}@inl.gov

**S. Hess and R. Sugrue**
Jensen Hughes
{shess, rsugrue}@jensenhughes.com

**D. Morton**
Northwestern University
david.morton@northwestern.edu

**I. Popova**
Texas State University
ip12@txstate.edu

**C. Pope, J. Miller, and S. Ercanbrack**
Idaho State University
{popechad, jadenmiller, spencerercanbrack}@isu.edu

**D. Cole and J. Yurko**
University of Pittsburgh
{dgcole, jyurko}@pitt.edu

[Digital Object Identifier (DOI) placeholder]

## ABSTRACT

Industry equipment reliability and asset management programs are essential elements that help ensure the safe and economical operation of nuclear power plants. The effectiveness of these programs is addressed in several industry-developed and regulatory programs. However, these programs have proven to be labor intensive and expensive. The goal of this paper is to provide effective and efficient analytical methods and tools to support risk-informed decisions for the equipment reliability and asset management programs at nuclear power plants. This is accomplished by creating a direct bridge between component health/lifecycle data and decision making (e.g., maintenance scheduling and project prioritization). Here we are supporting typical system engineer decisions regarding maintenance activity scheduling and component aging management. This is performed in a risk-informed context where the term "risk" is broadly constructed to include both plant reliability and economics. This framework combines data analytics tools to analyze equipment reliability data with risk-informed methods designed to support system engineer decisions (e.g., maintenance and replacement schedules, optimal maintenance posture) in a customizable workflow.

Key Words: data analytics, health management, decision making

## 1 INTRODUCTION

The Risk-Informed Systems Analysis (RISA) Pathway [1] of the United States Department of Energy Light Water Reactor Sustainability (LWRS) [2] Program is conducting collaborative research that applies risk-informed technology to assist operating nuclear power plants (NPPs) to reduce costs and support their adaptation to the changing economic and generating environment. The pathway is being performed within the framework of specific use cases, which are intended to provide a mechanism to achieve rapid technology development, deployment, and dissemination throughout the operating U.S. NPP fleet to address issues of pressing economic, operational, or safety significance.

One area of research in the RISA pathway is focusing on the development of methods and tools being designed to optimize plant operations (e.g., maintenance/replacement schedules, optimal maintenance postures for plant system structure and components [SSCs]) in a manner that is more cost effective than current approaches and makes better use of available SSC health and cost data. This is accomplished by creating a direct bridge between component health/lifecycle data and decision making (e.g., maintenance scheduling and project prioritization). Here we are supporting typical system engineer decisions regarding maintenance activity scheduling and component aging management. The second direction focuses on linking equipment reliability data (e.g., maintenance/failure reports, component monitoring data) directly to system reliability models using a margin management framework rather than

[†] Corresponding author: diego.mandelli@inl.gov

one that utilizes a probability-based language. The main advantage of a margin-based language approach is that it can provide responsible plant engineers and their management (i.e., decision makers) a more tangible and comprehensive set of information on SSC health and predict how system/plant level performance is likely to change in the future.

The overall workflow is partitioned in three main tasks. The first task focuses on the analysis of equipment reliability data with a particular emphasis on condition-based data, such as test/surveillance reports and component monitoring data (see Sections 2 and 3). Section 4 provides additional health assessment methods that integrate measured data with simulation models when available. The second task focuses on the integration of equipment reliability and simulated data into system/plant reliability models to determine system/plant health and identify the components that are critical to maintain an operational system (see Section 5). Lastly, the third task manages plant resources, such as maintenance activities (MAs) and replacement scheduling through the use of optimization methods (see Section 6).

## 2 EQUIPMENT RELIABILITY DATA TAXONOMY

Typically, a single component SSC is part of a system of components (see Figure 1 [left]) where such system is designed to provide a designed function, that is, *emergence* (such as electric power generation for a power plant). Each component contributes to the system emergence by providing a specified functionality that is being used by other components through a set of connections where *operands* (e.g., mass, energy, or data) are exchanged. The goal of a system health program is to monitor not only the correct operation of each component but also health parameters, such as aging and degradation (indicated as $\underline{F}(t)$ in Figure 1 [right]). In addition, a system health program is designed to perform appropriate actions to assure component functionality (indicated as $\underline{T}(t)$ in Figure 1 [right]). In this paper, $\underline{T}(t)$ includes all the external stressors that contribute to altering component aging and degradation.

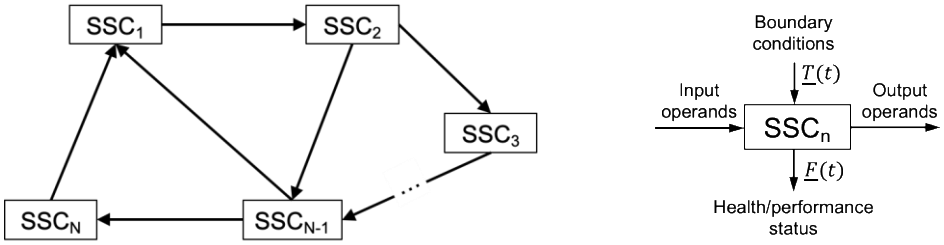


**Figure 1. System (left) and component (right) representation.**

### 2.1 System Engineer Perspective

When moving in more detail to the component level, it is vital to understand the relationship between monitoring/testing data, MAs, and failure modes (FMs). Figure 2 provides a detailed functional/form description of a generic SSC by employing an object process methodology (OPM) diagram [15]. An SSC OPM diagram provides an essential description of the SSC from both a form and functional perspective. This diagram explicitly indicates how SSC internal functions ($Func_f, f = 1, \dots, F$) process and act upon operands and how the elemental components ($ssc_r, r = 1, \dots, R$) support these functions.

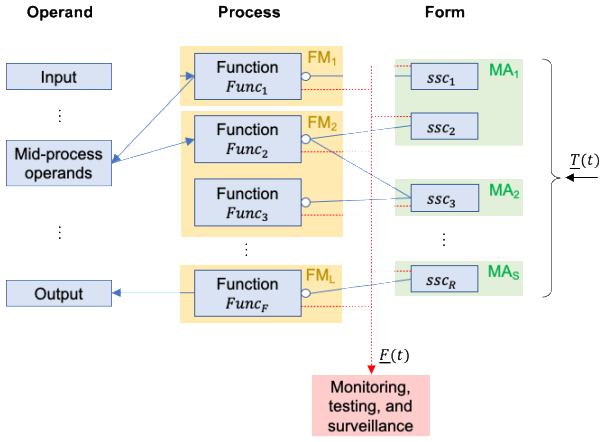


**Figure 2. SSC representation through an OPM diagram.**

From an equipment reliability perspective, monitoring/testing activities (i.e., $\underline{F}(t)$) act on both SSC functions (e.g., rpm recorded for an induction motor) and form (e.g., blade corrosion of centrifugal pump) elements. Degradation processes (i.e., $\underline{T}(t)$) directly alter the form-related elements of the component (i.e., $ssc_r$) that consequently affect SSC functional elements (i.e., $Func_f$,). Typically, from a reliability perspective, component FMs are described in term of loss of function, and, hence, in the OPM diagram, FMs are only directly linked to the functional elements of the component (i.e., $Func_f$,). Lastly, note that MAs (such as component replacement, refurbishment, or reconditioning), indicated as $MA$ in Figure 2, act on the form elements of components (i.e., $ssc_r$).

For the scope of this article, the OPM diagram of a component represents the key point to automatically understand and analyze health data $\underline{F}(t)$. In particular, it clearly links monitored/recorded data with FMs that might affect component performance and MAs that would restore component functionality. We are employing model-based data analysis methods with the goal of linking component models with data rather than using on machine-learning methods, which solely rely on the available data in order to perform diagnostic/prognostic operations. Note that an OPM diagram extends failure modes and effect analysis tables by providing a form and functional description of the considered system in a graphical form.

## 2.2 Data Scientist Perspective

The next step is to characterize a generic component SSC from a data scientist point of view. This is shown in Figure 3 where three levels are identified: the component level (which would correspond to what is shown in Figure 2), a sensor/monitoring level (which retrieves and records portions of $\underline{T}(t)$ and $\underline{F}(t)$ in digital form), and data level. Data retrieved from $\underline{T}(t)$ (i.e., $\underline{\theta}(t)$) can be either textual (e.g., work orders) or numeric (e.g., environment temperature). We indicate here with "num" the portion of $\underline{\theta}(t)$ that is numeric while we indicate with "NL" the portion of $\underline{\theta}(t)$ that is textual (NL here stands for natural language). Data retrieved from $\underline{F}(t)$ has been portioned in two portions, component health and performance monitoring ($\underline{\varrho}(t)$ and $\underline{\gamma}(t)$), which can be numeric or textual in nature as well.

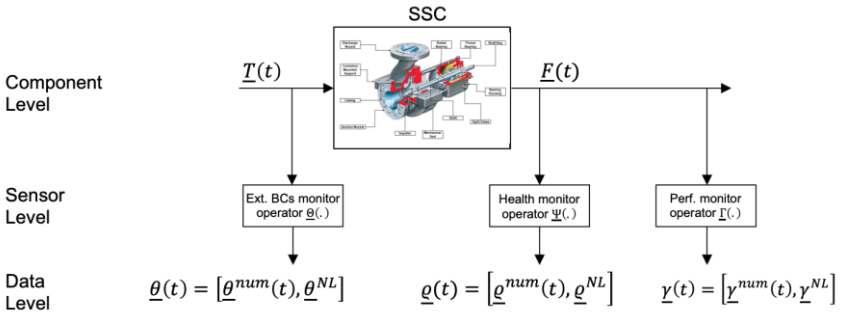


**Figure 3. System health program: component representation from a data point of view.**

# 3 PROCESSING AND ANALYSIS OF EQUIPMENT RELIABILITY DATA

As indicated in Section 2.2, equipment reliability data can be of different formats (i.e., numeric and textual). In addition, the events/logs that are recorded in $\underline{\theta}^{NL}(t)$ or $\underline{\gamma}^{NL}(t)$ can be defined over an interval or a single time instant. These two observations lead to a challenge when we analyze equipment reliability data: identify a common data structure that can be employed to represent numeric and textual data and events defined over time instants and time intervals. The advantage of having a common data structure is that it considerably simplifies the causal representation of events and monitoring data for condition-based monitoring applications.

This challenge has been here resolved by representing all elements of $\underline{\theta}(t)$, $\underline{\varrho}(t)$, and $\underline{\gamma}(t)$ (numeric and textual) in symbolic form (i.e., a series of symbols). This approach has the advantage that it simplifies the integration of numeric data with recorded events to identify patterns and outliers. In more detail, the method is structured in the following four steps:

1. Symbolic conversion of numerical time series. This is performed using the SAX method [3]. Data preprocessing (e.g., identification of anomalous behavior) may be required depending on the actual situation.
2. Symbolic representation of textual data. This is performed by characterizing events and logs into a graph form using natural language processing (NLP) methods [4]. A graph form has the advantage that it easily captures the structural relationship among text objects (i.e., nouns, verbs).

3. Combine data from Steps 1 and 2 into a common symbolic data structure. In our case, this is performed by creating a multivariate symbolic time series.
4. Apply model-based causal inference methods on the structure generated in Steps 3 by coupling data analysis methods with component OPM diagrams to infer component health, its FMs, and related maintenance activity that should be performed.

### 3.1 Preprocessing of Numerical Time Series

Depending on the situation at hand, the numeric monitoring data (in form of a time series) might require preprocessing. Typical operations are Z-normalization (mean subtraction and standard deviation scaling), filtering (e.g., through smoothing kernel), and moment decomposition. If required, the time series can be also preprocessed by extracting its residual: the difference between the actual signal and its reconstruction using data analysis or machine-learning methods. In this paper, this process is performed using the auto-associative kernel regression (AAKR) [5]. This method relies solely on data observed during normal conditions, and it uses such training data to estimate a reconstructed signal based on the evolution of the observed (i.e., real) signal and identify anomalous behaviors.

### 3.2 Symbolic Representation of Numerical Time Series

The next step is to convert the time series into symbolic form using the SAX [3] algorithm. SAX is an algorithm that allows the user to represent continuous time series $S$ as a series of $n$ symbols $\bar{\bar{S}} = \bar{\bar{s}}_1, \bar{\bar{s}}_1, \dots, \bar{\bar{s}}_n$, where $\bar{\bar{s}}_i$ is a symbol. This is performed by discretizing the time scale into $n$ intervals and the range of variability of the time series into $m$ intervals. While the temporal discretization is performed by partitioning the time axis into $n$ intervals having the same length, the discretization of the variability range of the time series is typically performed by dividing the range of the time series into $m$ equi-probable regions. Each region has a character $\bar{\bar{s}}$ associated with it and the alphabet size has cardinality $m$. The resulting conversion generates a time series of length $n$ and an alphabet size equal to $m$.

### 3.3 Symbolic Representation of Textual Data

Most methods found in the literature [6,7] perform processing of text reports using supervised learning [8] in order to predict the report nature (e.g., failure, operating). In this paper, we are following a different path, where the goal is to analyze the sentence structure of logs and reports, organize information in a structured form, and create a structural relationship among text objects (i.e., understand who/what did what, when, why, where). This is being accomplished by employing NLP methods[1] to perform two main tasks, syntactic and semantic analysis tasks, as follows:

1. *Syntactic analysis*
   - 1.1. Sentence segmentation and word tokenization: each sentence is translated into a list of strings
   - 1.2. Part of speech tagging: identification of grammatic elements of each string (e.g., nouns, verbs)
   - 1.3. Named entity recognition: classify text entities (names, dates, events) and identify them (component ID, type of event [e.g., maintenance or surveillance], event occurrence time)
   - 1.4. Relation extraction: create a knowledge graph where entities identified in Step 3 are linked together in a graph that reflects the structure of the original sentence
2. *Semantic analysis*
   - 2.1. Identify elements of graph in the OPM model (operands, objects or processes)
   - 2.2. Translate portions of OPM diagram identified in Step 2.1 into text and into graph
   - 2.3. Characterize event by measuring distance between two graphs of Steps 2.1 and 2.2
     - Objects: degradation, failure, test pass, restoration, abnormal state
     - Processes: degradation, loss of function, test pass, restoration, abnormal behavior

### 3.4 Construction of Common Data Structure

At this point, the numeric (see Section 3.2) and text data (see Section 3.3) need to be "merged" together in a single time series. This is performed by following the same philosophy behind the SAX algorithm (see Section 3.2), where the recorded events (which are graph structures) are inserted in the corresponding cells of the time series symbolic array. As an example, in Figure 4 we are focusing on a time series where three events are recorded from text data: events $E_1$ and $E_3$ are defined over a time instant while $E_2$ is defined over an interval. The graph structure

[1] In this work, we are employing two main `Python` libraries: `NLTK` (nltk.org) and `SPACY` (https://spacy.io).

for each event (i.e., type of event, component ID, date) is then associated with the corresponding cell generated by the SAX algorithm.

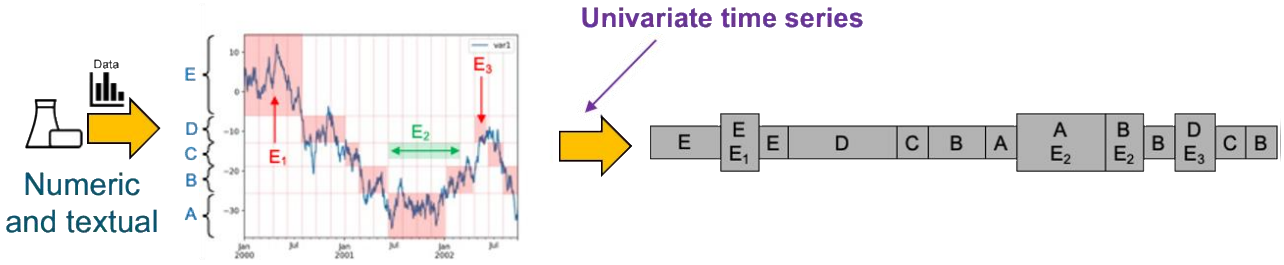


**Figure 4. Symbolic conversion of numerical (time series in the plot on the left) and textual data (events $E_1$, $E_2$, and $E_3$) into a single data structure.**

### 3.5 Data Analysis Methods

At this point, the data structure shown in Section 3.4 is fairly flexible to analyze by employing fairly standard methods [9] (such as Markov models, decision trees, and suffix trees) but also more advanced symbolic analysis methods [17]. However, the most relevant is based on subseries clustering: the symbolic time series (as generated in Section 3.4) can be partitioned into subseries and can identify those subseries that occur frequently (i.e., motif discovery), those that have never been recorded in the past (i.e., anomaly detection), and those that are close to a newly recorded subseries (data forecast). As a final comment, note that the memory requirements for the data structure shown in Section 3.4 is very small; numeric values occupy more memory than a string or a char. This guarantees fast performances for diagnosis and prognosis applications.

## 4 INTEGRATION OF SIMULATIONS TO DETERMINE COMPONENT HEALTH

An important element in health monitoring is that simulation models might provide additional information about component aging mechanisms and forecast degradation behavior. This section briefly describes methods and computational tools to integrate operational data and simulation models to forecast the degradation of SSCs for prognostic health management. Of particular interest for this article is the degradation of reactor internals during the life of a reactor and from outage to outage. Due to the large flow velocities in a reactor, there can be considerable internal vibrations. In this research, we are specifically interested in the core barrel and its support structures [15, 16]. The vibrations of the reactor internals are measured by fluctuations in the ex-core neutron detectors, which presents its own challenges since this is an indirect measure of the vibrations. Over time, the nature and contact between the reactor internals and the reactor vessel changes. A gradual decline in contact frequency as the unit ages is a result of a reduction in contact and an increase in the space between the vessel and internals (e.g., wear in the radial key). The challenge when integrating model-based and data-driven techniques is that there must be a coherent framework that combines the known physics-based model and the as yet unknown data-driven model of the degradation mechanism. We have used a full Bayesian approach [10, 11] for inferring the unknown contact model by combining the neutron noise measurements with computer simulations. The computer simulations provide the context, and the full Bayesian approach represents the unknown states as probability distributions.

Rather than trying to learn a general discrepancy function between the measurements and the physical simulation, we are using machine learning to represent specific physical processes. Machine learning is therefore still capturing difficult-to-model dynamics, but prior information about the behavior of such processes can be directly applied to regularize and constrain the machine-learning models. Here, the machine-learning approximated physical behavior can then be separated from the physics-based simulations. We explicitly model the feedback interaction between the nominal system and the degradation mechanism. Figure 5 illustrates this process for our specific problem, which accounts for the effect of contact mechanics on a vibrational structure. This configuration is a common framework in systems and control theory, and the problem can be transformed into a standard model-matching problem where the objective is to determine, using data, the degradation mechanism that minimizes the error between the true system and the integrated model.

Regarding the contact model we assume a nonlinear friction force that is viscous (proportional to velocity) for small displacements from equilibrium and is constant for large displacements. Such a model is characterized by two unknowns: $\alpha$ (which controls the asymptotic friction at large velocities) and $\beta$ (which defines the characteristic velocity). These parameters control the behavior of the contact force, which impacts the evolution of displacement

under an external load. The displacement impacts the ex-core measurements. The problem is estimating the parameters $\alpha$ and $\beta$ given the noisy measurements $z$.

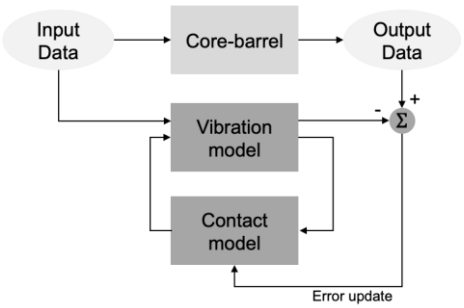


**Figure 5. The interconnection of a vibrational system (model) and a contact mechanism is a feedback process. The contact mechanism can be identified using machine learning in real time using model matching. Its change over time can then be used to estimate and predict component health.**

The inference of the unknown machine-learning parameters, $\alpha$ and $\beta$, in a full Bayesian setting requires specifying the likelihood between the physics model and the measurements as well as the prior on the unknowns. The prior may help prevent unphysical behavior by ruling out extreme values. The updated belief or posterior distribution on the unknown machine-learning parameters is a compromise between the data and the prior. The main steps are the following:

1. Specify a grid of candidate parameter values, $(\alpha, \beta)_m$, where $m = 1, \ldots, M$
2. For each of the $m = 1, \ldots, M$ grid points:
   a. Integrate the physics simulation forward in time using the specific set of machine-learning parameters, $\varphi_n((\alpha, \beta)_m)$, where $n = 1, \ldots, N$
   b. Calculate the posterior, $p((\alpha, \beta)_m | z)$
3. Visualize the posterior surface over the candidate parameter grid (see Figure 7).

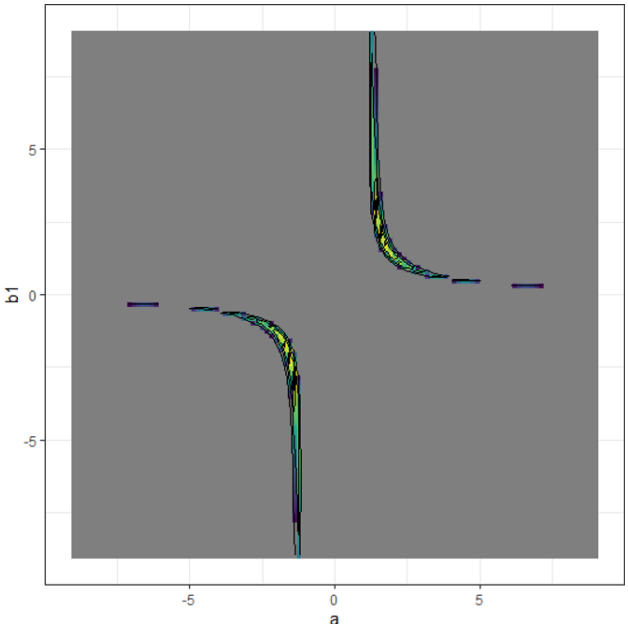


**Figure 6. Joint posterior surface graphically solved by the grid approximation for the unknown machine-learning parameters $\alpha$ and $\beta$.**

## 5 MARGIN-BASED RELIABILITY CALCULATION

Current reliability models are based on Boolean logic structures [12] (e.g., fault trees), which describe the deterministic functional relationship between SSCs and human interventions. Each basic event in a reliability model represents a specific elemental occurrence (e.g., failure of a component, failure to perform an action by the plant

operators, recovery of a safety system, etc.), and a probability value is associated with each basic event, which represents the probability that the basic event can occur. However, maintenance and surveillance operations are typically not completely integrated into a PRA structure. In addition, a probability value associated with an event is thus an integral representation of the past operational experience for such an event, and it does not incorporate information on the present health status of SSCs (e.g., from diagnostic and condition-based data) and health projections (when available from prognostic data) on anticipated changes in SSC condition and performance in the near future.

A possible alternate path can start by redefining the word "reliability" to encompass a broader meaning that better reflects the needs of a system health and asset management decision-making process. Rather than focusing on how likely an event is to occur (in probabilistic terms), we think in terms of how far this event is from occurring [12]. This new interpretation of risk transforms the concept from one that focuses on the probability of occurrence to one that focuses on assessing how far away (or close) an SSC is to an unacceptable level of performance or failure. This transformation has the advantage that it provides a direct link between the SSC health evaluation process and standard plant processes used to manage plant performance (e.g., the plant maintenance and budgeting processes). The transformation also places the question into a form that is more familiar and readily understandable to plant system engineers and decision makers. When dealing with condition-based data (actual and past/archived data), margin $\widetilde{M}$ is defined here as the distance between actual SSC observed past conditions (e.g., oil temperature, vibration spectrum) that lead to failure (see Figure 7).


**Figure 7. Margin in a condition based maintenance context: evolution of an SSC condition as a function of time and margin definition.**

Consider now two components ($A$ and $B$). The $\widetilde{M}$ for both components can be visualized in a 2-dimensional space, as shown in Figure 8. Starting with brand-new components (i.e., $\widetilde{M}_A, \widetilde{M}_B = 1$), aging degradation that affects both can be represented by the blue line of Figure 8, which parametrically represents the combination of the normalized margins $(\widetilde{M}_A(t), \widetilde{M}_B(t))$ as at a point in time $t$. Note that if no maintenance (whether preventive or corrective) was ever performed on either component, this path would move from the coordinates $(1,1)$, components $A$ and $B$ at the beginning of life, to the coordinates $(0,0)$ where both components had failed. We can identify these regions in Figure 8: the occurrence of both events where $\widetilde{M}_A = 0$ and $\widetilde{M}_B = 0$ and the occurrence of either event when $\widetilde{M}_A = 0$ or $\widetilde{M}_B = 0$. Now we can calculate the $\widetilde{M}$ for the events listed above. This is accomplished by following the definition of margin: by measuring the distance between the actual condition of components $A$ and $B$ and $\widetilde{M}$ conditions identified by the event under consideration (e.g., the occurrence of both or either events):

$$\begin{aligned} \widetilde{M}(A\ AND\ B) &= dist[(\widetilde{M}_A, \widetilde{M}_B), (0,0)] \\ \widetilde{M}(A\ OR\ B) &= min\left(\widetilde{M}_A, \widetilde{M}_B\right) \end{aligned} \tag{1}$$

The function $dist[X, Y]$ is designed to calculate the Euclidean distance between points $X$ and $Y$.

Hence, exact solutions can be obtained extremely fast. More precisely, reliability calculations using $\widetilde{M}$ based data can be performed by completing these four steps:

1. Construct the fault tree (FT); at this point, an FT contains only deterministic information about the architecture of the system under consideration (i.e., it simply models how the basic events are related to each other from a functional perspective).
2. Generate the minimal cut-sets (MCSs) from the FT; as also indicated in Step 1, an MCS still represents the minimal combinations of basic events that lead to the TE.
3. Assign $\widetilde{M}$ to each basic event.
4. Calculate the $\widetilde{M}$ of the union of the MCSs.

As part of system reliability modeling, it is always important to determine the importance of each basic event. In a PRA setting, this is performed by relying on risk importance measures [12], such as Birnbaum or Fussell-Vesely. Given the different nature of $\widetilde{M}$, it is possible to perform a risk importance ranking by relying on a classical sensitivity

measure (derivative based) for each basic event $BE$ defined as: $S_{BE} = \frac{\partial \widetilde{M}(TE)}{\partial \widetilde{M}(BE)}$. In other words, $S_{BE}$ indicates how a small variation of $\widetilde{M}(BE)$ directly affects $\widetilde{M}(TE)$.

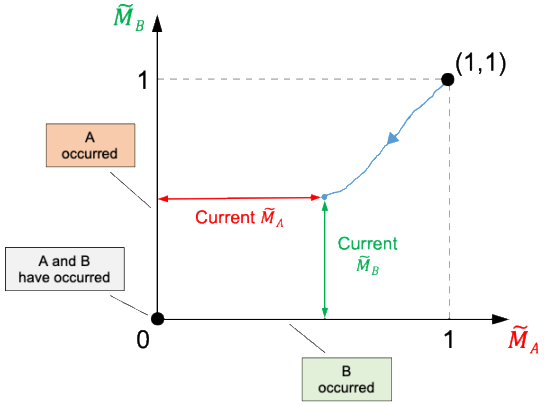


**Figure 8. Graphical representation of event occurrences based on a margin framework.**

## 6 RESOURCES OPTMIZATION METHODS

The last step is to manage plant resources based on the system health analysis indicated in Sections 4 and 5.

### 6.1 Project Prioritization

The FMs with higher $S_{BE}$ (see Section 5) are the ones selected as candidates to be subject to MAs (see Figure 2). A list of possible options to address each failure mode is available where costs (e.g., procurement costs for a new or refurbished component) and benefits (e.g., increased margin for loss of production) are readily available or can be numerically determined. Given the candidate MAs and their options, we can now identify the best set of activities and options that give "the most bang for the buck."

This is accomplished by identifying the Pareto frontier [13] out of the all the possible MAs and options. Let's assume that a decision can be taken from a set of options by considering the utility and cost of each option. Using a graphical representation (see Figure 9) it is possible to plot each option as a point in a 2-dimensional space, cost vs. utility[2]:

- *Cost*: this axis represents the cost associated with each option ranging from 0 (i.e., cheapest option) to a maximum value $C_{max}$ (i.e., the most expensive option)
- *Utility*: this axis represents the added value (or the performance) associated with each option ranging from 0 (i.e., lowest performance option) to a maximum value $U_{max}$ (i.e., option with highest performance).

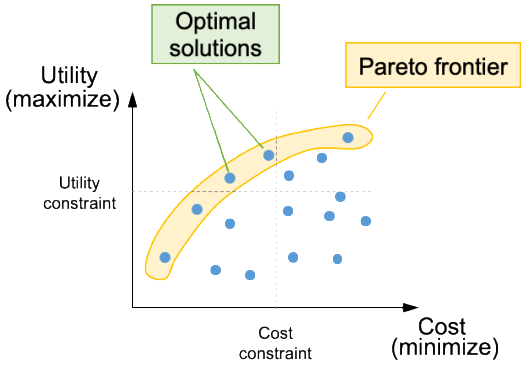


**Figure 9. Pareto frontier obtained from a set of options plotted in a cost vs. utility space and imposition of cost and utility constraints (right).**

[2] As indicated earlier, the number of attributes considered in complex settings can be $N > 2$. Thus, in such cases, the space would be $N$-dimensional.

Once the complete set of options have been generated and the utility and cost values have been determined for each option, the next step is the determination of the Pareto optimal frontier, which is fundamentally an envelope of options that dominates (in terms of both utility and cost) the set of remaining options (see Figure 9).

### 6.2 Long-Term Decisions: Project Scheduling Given Budget Constraints

The method described in Section 6.1 does not explicitly take into consideration project actuation scheduling but instead focuses on the optimal subset of projects that provide higher value through a multi-objective optimization lens. In practical settings, project scheduling is done in phases (e.g., monthly, quarterly) wherein each phase budget is allocated and the goal now is to choose the optimal project actuation schedule that minimize costs and satisfy budget constraints [14]. The notation and formulation are as follows:

*Indices and sets:*

| | |
|---|---|
| $i, i' \in N$ | candidate projects |
| $d \in D$ | types of resources (e.g., capital funds, O&M funds, labor-hours, time during outages) |
| $M \subseteq I$ | must-do projects (e.g., for safety reasons, even if their NPVs are negative) |
| $\omega \in \Omega$ | scenarios |

*Data:*

| | |
|---|---|
| $p_i$ | NPV of investment $i$ |
| $c_d$ | available budget for a type $d$ resource |
| $w_{i,d}$ | consumption of a type $d$ resource if investment $i$ is selected |
| $p_i^\omega$ | profit of investment $i$ under scenario ω (NPV) |
| $c_d^\omega$ | available budget for a type $d$ resource under scenario ω |
| $w_{i,d}^\omega$ | consumption of a type $d$ resource if investment $i$ is selected under scenario ω |
| $q^\omega$ | probability of scenario ω |

*Decision variables:*

$$x_i = \begin{cases} 1 & \text{if project } i \text{ is selected} \\ 0 & \text{otherwise} \end{cases}$$

*Model:*

$$\max \sum_{i \in N} p_i x_i$$

$$s.t. \sum_{i \in N} w_{i,d} x_i \leq c_d, d \in D$$

$$\sum_{i \in N} x_i = 1, i \in M$$

### 6.3 Short-Term Decisions: Maintenance Activity Scheduling Given Personnel Constraints

Once the project schedule has been finalized (see Section 6.2), each project is decomposed into tasks where each task is characterized by a set of parameters (e.g., duration, number of personnel required, list of required tasks that need to be performed prior to start this task, skill set required, completion deadline). Plant resources are now constituted by a set of crews where each crew is characterized by the number of people, schedule availability, and available skill set. The goal is to minimize the time to complete all jobs and tasks provided the constraints that the order of tasks for each job must be preserved and one task can be assigned to each crew.

## 7 CONCLUSIONS

In this paper, we have presented a series of methods and models designed to create a direct bridge between equipment reliability data and equipment reliability related decisions. Even though this type of bridging is not new, we are here presenting a different structure for such a bridge. First, we have introduced a novel approach to analyze ER data that integrate logs and events data with numeric data available from plant monitoring and diagnostic centers in a common data structure (symbolic in nature). Rather than focusing on machine-learning heuristics, the system view of the component (through a OPM diagram) provides required knowledge to our data analysis methods to extract knowledge from text data retrieved by logs or workorders. The processed data can then be integrated into classical reliability models (e.g., fault trees) that are solved not using a probability-based but a margin-based language. The main advantage of this method is that it allows a much better use of

ER data, and it provides more adequate risk importance ranking of the FMs for the considered set of SSCs. Lastly, the decision-making step is carried through by determining the set of projects/operations that provide "the most bang for the buck" (i.e., the Pareto frontier), prioritizing the actuation schedule for the selected projects (medium-term decisions), and identifying the optimal schedule that minimizes the completion time of the required maintenance tasks (short-term decisions). This is performed in a risk-informed context where the term "risk" is broadly constructed to include both plant reliability and economics.

## 8 ACKNOWLEDGEMENTS

This work of authorship was prepared as an account of work sponsored by Idaho National Laboratory (under Contract DE-AC07-05ID14517), an agency of the U.S. Government. Neither the U.S. Government, nor any agency thereof, nor any of their employees makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

## REFERENCES

[1] R. Szilard, H. Zhang, S. Hess, J. Gaertner, D. Mandelli, S. Prescott, and M. Farmer, "RISA Industry Use Case Analysis," Idaho National Laboratory Technical Report, INL/EXT-18-51012 (2018).

[2] U.S. Department of Energy, "Light Water Reactor Sustainability Program Integrated Program Plan," Idaho National Laboratory Technical Report, INL/EXT-11-23452 (2020).

[3] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms," *8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, San Diego, June, pp.2-11 (2003).

[4] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*, O'Reilly Media (2009).

[5] P. Baraldi, F. Di Maio, P. Turati, and E. Zio, "Robust Signal Reconstruction for Condition Monitoring of Industrial Components via a Modified Auto Associative Kernel Regression Method," *Mechanical Systems and Signal Processing*, **60**, pp.29-44 (2015).

[6] R. Sipos, D. Fradkin, F. Moerchen, and Z. Wang, "Log-Based Predictive Maintenance," *KDD 14: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, August, pp.1867-1876 (2014).

[7] M. Zhou, N. Duan, S. Liu, and H.-Y. Shum, "Progress in Neural NLP: Modeling, Learning, and Reasoning," Engineering, **6**, no. 3, pp.275-290 (2020).

[8] S. Marsland, *Machine Learning: An Algorithmic Perspective*, Chapman & Hall/CRC Machine Learning & Pattern Recognition (2009).

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer ed. (2009).

[10] M. C. Kennedy and A. O'Hagan, "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, no. 3, pp.425-464 (2001).

[11] D. Mandelli, C. Wang, J. Cogliati, C. Smith, S. Hess, R. Sugrue, C. Pope, J. Miller, S. Ercanbrack, D. Cole, J. Yurko, "Integration of Data Analytics with Plant System Health Program," Idaho National Laboratory Technical Report, INL/EXT-20-59928 (2020).

[12] J. Lee and N. J. McCormick, *Risk and Safety Analysis of Nuclear Systems*, Wiley edition (2011).

[13] R. K. Sarin, *Multi-Attribute Utility Theory, Encyclopedia of Operations Research and Management Science*, Springer, Boston, MA (2013).

[14] D. Mandelli, C. Wang, M. Abdo, A. Alfonsi, P. Talbot, J. Cogliati, C. Smith, D. Morton, I. Popova, and S. Hess, "Development and Application of a Risk Analysis Toolkit for Plant Resources Optimization," Idaho National Laboratory Technical Report, INL/EXT-20-59942 (2020).

[15] I. Pazsit, J. Karlsson, and N. S. Garis, "Some Developments in Core-Barrel Vibration Diagnostics," *Annals of Nuclear Energy*, **25**, no. 13, pp.1079-1093 (1998).

[16] I. Pazsit, H. Nylen, and C. Montalvo Martin, "Refined Method for Surveillance and Diagnostics of The Core Barrel Vibrations of the Ringhals PWRs," *Proceedings of PHYSOR 2014 - The Role of Reactor Physics Toward a Sustainable Future*, Sept. 28–Oct. 3 (2014).

[17] F. Morchen, "Time Series Knowledge Mining," PhD Thesis, Philipps-University Marburg, Germany (2006).