Research towards a systematic signature discovery process

(Submitted to IEEE ISI 2013)

Nathan A. Baker*†, Jonathan L. Barr†, George T. Bonheyo‡, Cliff A. Joslyn†, Kannan Krishnaswami†,
Mark E. Oxley§, Rich Quadrel†, Landon H. Sego†, Mark F. Tardiff†, Adam S. Wynne†

*To whom correspondence should be addressed
†National Security Directorate, Pacific Northwest National Laboratory, Richland, WA 99352, USA
Email: {firstname.lastname}@pnnl.gov
‡Energy and Environmental Directorate, Pacific Northwest National Laboratory, Sequim, WA 98382, USA
Email: george.bonheyo@pnnl.gov
§Department of Mathematics and Statistics, Wright-Patterson Air Force Base, WPAFB, OH 45433, USA
Email: mark.oxley@afit.edu

***Abstract*—In its most general form, a signature is a unique or distinguishing measurement, pattern, or collection of data that identifies a phenomenon (object, action, or behavior) of interest. The discovery of signatures is an important aspect of a wide range of disciplines from basic science to national security for the rapid and efficient detection and/or prediction of phenomena. Current practice in signature discovery is typically accomplished by asking domain experts to characterize and/or model individual phenomena to identify what might compose a useful signature. What is lacking is an approach that can be applied across a broad spectrum of domains and information sources to efficiently and robustly construct candidate signatures, validate their reliability, measure their quality, and overcome the challenge of detection – all in the face of dynamic conditions, measurement obfuscation, and noisy data environments. Our research has focused on the identification of common elements of signature discovery across application domains and the synthesis of those elements into a systematic process for more robust and efficient signature development. In this way, a systematic signature discovery process lays the groundwork for leveraging knowledge obtained from signatures to a particular domain or problem area, and, more generally, to problems outside that domain. This paper presents the initial results of this research by discussing a mathematical framework for representing signatures and placing that framework in the context of a systematic signature discovery process. Additionally, the basic steps of this process are described with details about the methods available to support the different stages of signature discovery, development, and deployment.**

I. Introduction

A signature is a unique or distinguishing measurement, pattern, or collection of data that detects, characterizes, or predicts a target phenomenon (state, object, action, or behavior) of interest. By definition, a signature has a reliable correlation to its target and consequently becomes extremely useful for anticipating future events by detecting precursor signatures, such as circumstances that may lead to a cascading power failure in an electrical grid; diagnosing current conditions by matching observations against known signatures, such as the monitoring of a computer network for security events; or analyzing past events by examining signatures left behind, such as the forensic analysis of pathogens or other biological threat agents. Such analyses can contribute to larger signature libraries which in turn serve as a resource for future anticipation and characterization.

Because of their value, significant effort has been expended to discover reliable signatures for specific applications. This discovery is typically accomplished by asking domain experts to characterize and/or model the phenomenon to identify salient features that might compose a useful signature. Such inquiry is generally undertaken for a specific problem domain and the resulting signatures are typically chosen and characterized by trial and error. We believe that current methods are inefficient, and can frequently produce sub-optimal results. More importantly, these "one-off" signature development efforts may overlook novel and unconventional features that could be used to form highly effective signatures.

This paper describes our work towards the generalization and formalization of the signature discovery process. Development of this process is based on broad literature research, user interviews, and hands-on experience through our Signature Discovery Initiative research investment (http://signatures.pnnl.gov). The following sections present a brief introduction to the concept of a signature, an overview of our generalized process for signature discovery, and examples of the signature discovery process to real-world applications.

II. Signature systems

In order to develop a generalizable process for signature discovery, we require a robust and generalizable definition of a signature. This section outlines our framework for signatures as transformations from events to measurements to features to categorical labels and associated uncertainties.

Let $\mathcal{E}$ denote a collection of events of a certain problem domain of interest. These events are instances of interest that could possibly occur in nature; e.g., the multiple expression levels of genes, the varying concentration levels of impurities in a chemical substance, the wide range of packet contents found in cyber networks, etc. In general, we can think of $\mathcal{E}$

978-1-4673-6213-9/13/$31.00 ©2013 IEEE

and the other sets as consisting of a cross-product of certain dimensions, or some phenomenon of nature which can be represented or recorded: a quantity, quality, or some other state. Dimensions have mathematical **types**; e.g., boolean, categorical, or scalar. We will refer to the cross-products in $\mathcal{E}$ as **dimensional sets**, although they can also be thought of as **data tensors**. Note that while $\mathbb{R}^n$ is a perfectly acceptable dimensional set, a dimensional set can also represent a highly heterogeneous collection of variables of various mathematical types.

The particular signature problem domain determines the purpose of the signature and the type of its output, defined as $\mathcal{L}$, which denotes a set of labels as illustrated in the following examples. This set $\mathcal{L}$ is typically discrete, with finite $|\mathcal{L}| = K \in \mathbb{N}$; however, this discreteness is not a requirement. For the simplest binary signature used in threat detection problems, $\mathcal{L}$ is Boolean with $\mathcal{L} = \{l_1, l_2\} = \{\text{threat, no threat}\}$. If the problem is a classification problem (with many finite classes), then $\mathcal{L}$ will be finite, categorical (or numerical), and have a possible partial or total ordering. However, some problems (e.g., signatures for the prediction of vessel direction) require continuous probability density function outputs. In such cases, then $\mathcal{L}$ is a collection of functions with probability properties. Given a label $\ell \in \mathcal{L}$, we presume there exists a set of events that truly corresponds to ℓ. To simplify the discussion, we will assume that the label set is chosen such that there exists a truth function τ defined on $\mathcal{E}$. Therefore, for every label $\ell \in \mathcal{L}$ there exists a subset $\mathcal{E}_\ell = \{e \in \mathcal{E} : \tau(e) = \ell\}$, and $\mathcal{E}_\ell \cap \mathcal{E}_{\ell'} = \varnothing$, the empty set, for every $\ell, \ell' \in \mathcal{L}$ that satisfies $\ell \neq \ell'$. For example, in an explosives detection signature, if $p =$ explosive material present, and $n =$ explosive material not present, so that $\mathcal{L} = \{p, n\}$, then τ is the truth detection function. In most applications, the truth function is unknown to us and we seek to discover it through the processes described below.

Given these definitions, a **signature system** is a collection of mappings that, when combined together, will equal or approximate the truth mapping. The first mapping μ is a process that observes an event and then produces a measurement, i.e., a random variable. The mapping μ can arise from a sensor, a measurement device, or any other type of data collector that maps events into raw data. Let $\mathcal{M}$ denote the **measurement** dimensional set of the raw data, and thus $\mu : \mathcal{E} \longrightarrow \mathcal{M}$. In many applications, analysis begins with the identification of $\mathcal{M}$, since a measurement device or observation channel is provided *a priori*.

In general, raw data are not used directly in the discovery of the signature. Instead, we seek another transformation η that processes the raw data by extracting salient, informative data (called **features**) from the raw measurement data. Let $\mathcal{F}$ denote the dimensional set of features that η produces. Thus, $\eta : \mathcal{M} \longrightarrow \mathcal{F}$. The final mapping is a classifier $\delta : \mathcal{F} \longrightarrow \mathcal{L}$, which maps features to a label. The composition of these mappings is designed to approximate the truth mapping; i.e., $\hat{\tau} \equiv \delta \circ \eta \circ \mu \approx \tau$.

Rather than map directly to $\mathcal{L}$, it may be preferable for the classifier δ to map to an **uncertainty space**, $\mathcal{P} = [0,1]^K$, that is associated with the label set, so that an element $\vec{p} \in \mathcal{P}$ is a vector of numbers each in $[0,1]$. It is common, but not necessary, to interpret the vectors elements as probabilities of the corresponding label $\ell \in \mathcal{L}$, so that they sum to one. Hence, the classifier δ is modified to output a vector of pairs $((\ell_1, p_1), (\ell_2, p_2), \ldots, (\ell_K, p_K))$ such that each $\ell \in \mathcal{L}$ has a corresponding score or probability, $p \in \mathcal{P}$. Let $\delta_{\mathcal{P}} : \mathcal{F} \rightarrow \mathcal{L}^K \times \mathcal{P}$ denote this type of classifier.

In summary, the set of mappings that describe a signature system can illustrated by the diagram:

$$\text{Events} \xrightarrow{\mu} \text{Measurements} \xrightarrow{\eta} \text{Features} \xrightarrow{\delta} (\text{Labels}, \text{Probabilities})$$

The generality of the framework defined here provides value in abstracting processes of arbitrary complexity from a variety of problem domains. For example, as the dimensional sets and spaces are mapped into each other, the number of dimensions of a subsequent set is not determined by those of the prior. Thus, “multi-INT” [1] signatures can be developed through processes wherein five distinct measurements can result in two features, which in turn map to three labels. Furthermore, this generalizable framework offers the reuse of transformations across domains; e.g., processes from biology can be mapped into processes from cyber security by identifying their common mathematical structure, independent of the different nature of their observables, labels, and analytical processes.

III. The Signature Discovery Process

Our research has focused on the identification of common elements of signature discovery across application domains and the synthesis of those elements into a systematic process for more robust and efficient development of the signature system components described in the previous section. A systematic signature discovery process lays the groundwork for leveraging knowledge obtained from signatures to a particular domain or problem area, and, more generally, to problems outside that domain.

Fig. 1 illustrates our proposed systematic process for signature discovery. This process was developed by surveying the signature development literature. The literature survey yielded over 1000 peer-reviewed articles with a signature discovery focus; a subset of 100 of these papers were selected for more in-depth analysis. The papers from this subset fell predominantly into the following domains: disease identification and prognosis, chemical or biological forensics, cyber-related signatures, and semantic-based signatures. After inspection of the complete survey list, it was possible to prioritize analysis of articles based on the level of detail presented and focus on the signature development process as opposed to application of the developed signature.

To date we have reviewed the literature in each of the aforementioned domains and at three levels of detail (example references provided): standard signature development practices [2]–[5], original scientific research related to signature development and discovery [6]–[11], and detailed interaction with researchers to document nuances of specific signature discovery process. Works that provided sufficient detail to their process allowed us to break down each step documented into a context diagram that detailed inputs, outputs, and activities/processes. Through this process, it was possible to identify, group, and order common steps into a notional signature discovery process. For those works that did not provide a detailed account of the approach taken by the researchers, it was possible to map the activities that were

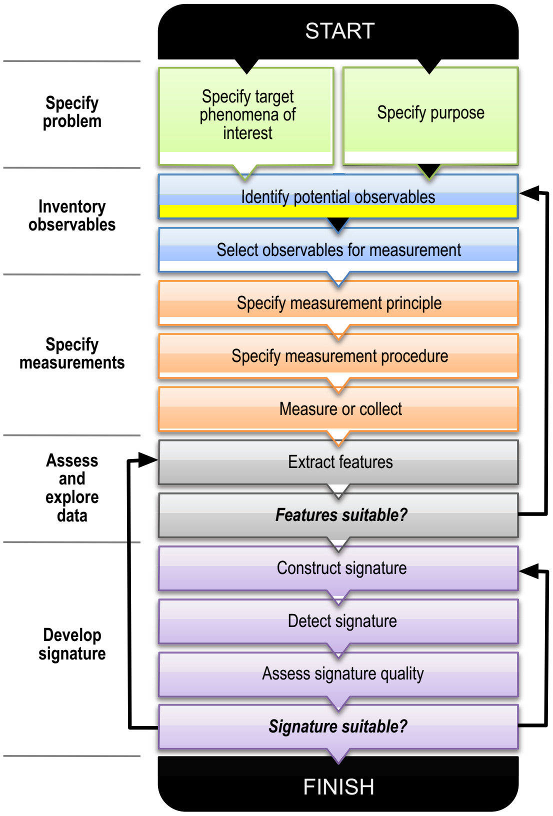

Fig. 1. A schematic representation of the proposed systematic and generalizable signature discovery process. See text for more information on each step of this process.

detailed to the signature discovery process and determine if there were activities or steps not identified in the process. The following sections describe each step of this process and briefly provide highlights from the state-of-the-art in methods for implementing these steps.

A. *Specify problem*

1) Specify target phenomenon of interest: The process of signature discovery begins by specifying a particular target phenomenon of interest. While this step may seem obvious, the explicit specification of a target is important because it determines the event space $\mathcal{E}$ that defines the domain of our signature system mappings. Targets can be divided according to the top-level structure of the Basic Formal Ontology [12] into continuants, objects that exist at a snapshot in time, and occurrents, processes that unfold over an interval of time. Example continuant targets include objects such as a type of tumor, a particular class of vessel or vehicle, a type of explosive or explosive residue, etc. Example occurrent targets include activities such as an intrusion on a computer network, trafficking of drugs, nuclear proliferation processes, etc.

2) Specify purpose: The purpose of a signature is defined by two characteristics: the size of the label set $\mathcal{L}$ and the relationship in time between the target and the detection of the signature. The simplest signature purpose is **detection** where $\mathcal{L}$ includes two mutually exclusive labels such as the binary signature example used earlier where $\mathcal{L} = \{\text{threat}, \text{no threat}\}$. More complicated **characterization** signature purposes involve $\mathcal{L}$ with multiple labels, including the possibility of a continuum of labels; e.g., for characterizing the concentration of a component in a complex chemical mixture. The purpose of a signature is further defined by the relationship in time between the signature target and the detection of the signature. A signature purpose is **prognostic** if the signature is used to detect or characterize targets in the future; e.g., signatures to determine the outcome of a disease or the likelihood of a power failure. Conversely, signatures are **forensic** if the signature is used to detect or characterize targets in the past; e.g., the attribution of origin for a biological threat agent or the analysis of computer network events that lead to a loss of service. Finally, a signature purpose is **diagnostic** if the signature is used to detect or characterize targets in the present; e.g., characterizing the current state of a power grid or determining the presence or absence of a tumor.

B. *Identify and select observables*

Once the problem has been specified, the next steps are to define variables and parameters surrounding the problem and the required solution. Key issues include the nature of the signature purpose (e.g., to detect, characterize, or predict within a given amount of time, cost, and certainty) and the nature of the target phenomenon (e.g., the temporal and spatial location of the phenomenon, the environment, and the dynamics of these elements). Exploration of the target phenomenon and signature purpose establishes the parameters under which a set of possible observables may then be identified.

The best practices for facilitating this creative ideation process are not well understood. For several problem types, it is highly desirable to bring together a group of experts representing diverse backgrounds, fully describe the problem to them, and let them identify and select possible solutions. The most prevalent group ideation method is brainstorming, a principle first codified by Osborne in 1953 [13]. The brainstorming process is based upon the idea that a small group (five to seven people) working on a problem should outperform an individual by producing more diverse or alternative perspectives that lead to a greater variety of ideas. Additionally, brainstorming assumes that "social facilitation," which is the process of intellectual stimulation via group interaction, should result in individuals performing better and therefore improve overall group productivity. However, other group processes introduce several challenges to the brainstorming process [14]–[16]. The first challenge is **production blocking**, where individuals forget or withhold their ideas while others take turns speaking or where a few people dominate discussions and limit the participation and contribution of others. The second is **evaluation apprehension**, which is the fear of how ones ideas will be perceived by the group. The third challenge is **social loafing**, which entails reduced effort by some members within a group and/or the lack of individual accountability. The fourth challenge is **performance matching**, described both as pressures for conformity and/or a tendency to match ones level of performance with others in the group, usually resulting in lower performance. Several studies, critiques, and variations on the original Osborn method have emerged over the years to address these problems. However, only a few studies have examined group ideation and creativity focusing on real problems with real stakeholders in the context of multidisciplinary teams and scientific examination or discovery [16]–[22]. Recently, electronic brainstorming tools have emerged to alleviate some of the negative behaviors described above, though they introduce a different set of challenges and have had mixed results [22]–[24]. Despite ongoing advances in brainstorming and related techniques, stimulation and facilitation of the creative process leading to the identification and selection of observables remains a very challenging step in the signature discovery process.

In addition to brainstorming methods, experimental design [25], [26] can be a useful technique for identifying and selecting observables that may be related to the phenomenon of interest. Many signature development problems have measurement spaces with high dimensionality. Developing the signature for such problems requires thoroughly exploring the feature space to cover the intended bounds of the signature and exhaustive exploration can often be infeasible. Experimental design methods such as fractional factorial designs [27] and Plackett-Burman methods [28] can be used to simultaneously obtain adequate data coverage based upon the signature development objectives and manage factorial explosion of the signature development parameter space. Variable selection techniques [29], [30] can be used to identify variables in existing datasets when further experimentation is not possible or relevant.

C. Specify measurements

The identification of observables determine the types of events $e \in \mathcal{E}$ to be included in the signature system. The next step in the signature system is to identify the measurement principles and instruments μ that allow these events to be mapped into data or measurements. Selection of a measurement system is not a simple process and many possible strategies could be applied to a given observable with specific choices guided by operational or cost constraints. The multiplicity of choices and the importance of other constraints motivates the iterative process outlined in Fig. 1, with signature quality assessment (Sec. III-E3) playing an essential role in the process. After the measurement process has been identified for each observable variable, the **observational unit** must be clearly identified. The observational unit is the entity that is measured (i.e., observed) whose data and resulting signature is used to detect, predict, or identify the phenomenon of interest. For example, if the goal of the signature were to predict the onset of a disease in humans, each patient under study would be an observational unit.

The strength of the signature depends on the quality of the sample $\mathcal{S} \subset \mathcal{E}$ of observational units and the extent to which it generalizes to the population of interest, which for our purposes, is the same as $\mathcal{E}$. The choice of $\mathcal{E}$, in large part, is determined by the purpose of the signature. In situations where a representative sample is desired, traditional statistical sampling techniques such as simple random, systematic, stratified, or clustered sampling [31] may be appropriate.

D. Assess and explore data

For each observational unit or instance in the sample $e \in \mathcal{S}$, the measurement process μ will produce data $\chi \in \mathcal{X} \subset \mathcal{M}$, i.e., $\chi = \mu(e)$. The signature system transforms these data into features through the mapping η. Careful exploratory data analysis [32] is an important step in identifying features. When the measurement space $\mathcal{M}$ has a high dimension, data and dimensionality reduction processes can be very important. For example, with continuous components, dimensionality reduction techniques such as principal components analysis [33] can be employed to identify potential features and structure within the data. When one or more of the components of $\mathcal{M}$ are not continuous, subsetting the data according to the levels of the categorical types and making plots that are conditioned on those levels [34] can be especially helpful. Clustering [33] is often a useful technique for uncovering patterns and relationships in data.

Feature extraction, *i.e.* constructing η, is arguably the most crucial and difficult aspect of the signature construction process. Typically, η will reduce the dimensionality [35] of the data, which inevitably sacrifices some information. Consequently, constructing η is especially challenging when signature events are rare, as it would be easy to discard the essential information required to detect the phenomenon of interest. The objective in constructing η is analogous to the concepts of statistical sufficiency and completeness [36]: we want $\eta(\mathcal{X})$ to contain as much of the salient information about τ as it possibly can (sufficiency) while avoiding superfluous or unneeded information (completeness). Data mining techniques [29], [37] can be useful in developing η. If the events in $\mathcal{E}$ are time-dependent, smoothing methods like splines [38] or local regression [39] can be useful. For data with a variety of mathematical types, generalized linear [40] and nonlinear [41] models may also be applicable, as they can accommodate responses and predictors of virtually any mathematical type.

E. Develop signature

The final steps of the signature discovery process involve the development and application of the signature, iterating based on decisions about the signature quality and suitability for deployment.

1) Construct signature: Having developed a suitable feature extraction process η and set of features, the next step is to construct a classifier δ or $\delta_{\mathcal{P}}$. There is a wealth of classification and machine learning techniques that have been developed simultaneously in the statistics [29], [42] and computer science [43] communities, respectively, which may be used to construct δ. We have observed that Bayesian networks [44] can be very useful in developing classifiers when many or all of the components of $\mathcal{E}$ are not continuous. Bayesian networks have the additional advantage of high user interpretability, which assists in the adoption and application of the signature system by end users.

The process of constructing δ is an interactive one, where the classifier is *trained* using a training set $\mathcal{T} \subset \mathcal{S}$ and then tested using a preferably distinct testing set, $\mathcal{T}' \subset \mathcal{S}$; i.e., $\mathcal{T} \cap \mathcal{T}' = \varnothing$. Training typically entails estimating, or learning, the parameters that govern the functional form of δ. There are various approaches to learning, including supervised, unsupervised, semi-supervised, and active [45]. Having trained the classifier δ, the estimate of truth relation, $\hat{\tau}$ is complete and the full signature system is available for validation and application to new measurements.

2) Detect signature: Signature detection is the application of the signature system to actual problem datasets: the series of transformations from events to labels and probabilities. Many issues associated with the detection process have been addressed in the steps above. However, the "real world" application of signature systems often raises additional challenges that need to be considered during the development process. Measurement data produced in and used by the signature system can vary in its quantity and complexity. A popular example is the "big data" problem [46], [47] where applications present data with challenges in "volume, velocity, and variety" [48]. Such challenges often require deliberate choices for feature extract η and classification δ algorithms that can handle diverse, large, and high-throughput datasets. For example, signatures for power grid characterization can involve millions of data streams and require scalable algorithms. Some applications offer different challenges in signature detection: measurement data may be sparse and the application of the signature must be robust to missing or incomplete information. For example, forensic signatures must often reach conclusions based on limited data and require algorithms and approaches that can infer or impute missing data [49], [50]. Finally, the security domain often faces unique challenges in data obfuscation and falsification and requires signature systems that can, at a minimum, detect such activities and, ideally be robust to such interference.

3) Assess signature quality: Our objective in assessing signature quality is to measure the extent to which a signature system achieves its intended purpose. To that end, we have developed a formalism based on decision theory [51] and multi-attribute decision science [52], [53]. We explain our approach to assessing signature quality in terms of fidelity, risk, cost, and any other attributes of importance related to the deployment and use of the signature system.

Fidelity refers to how well the signature system detects, predicts, or characterizes the phenomenon of interest. It includes metrics such as sensitivity, specificity, positive predictive value, precision, accuracy, and receiver operating characteristic (ROC) curves [42], [54], [55]. These fidelity metrics attempt to assess how well $\hat{\tau}$ approximates τ.

Risk refers to the assessment of likelihoods and consequences associated with decision errors which may arise during signature detection. For instance, when the phenomenon of interest is binary (e.g., threat, no threat), calculating risk helps evaluators balance the tradeoff between false positives, which may impede the flow of traffic or commerce, and false negatives, which increase exposure to threats.

Cost refers to the resources expended to develop, deploy, and/or utilize the signature system. Examples include the cost of signature systems, training, maintenance, consumable reagents, and labor.

Other attributes include any other factors or criteria that may distinguish one signature system from another that are not already accounted for by fidelity, cost, or risk. Examples include the time required to collect, process, and analyze samples, human safety, ease of use, system portability, policy considerations, etc.

These four components of signature quality provide a basis for guiding investigators in their assessment of signature systems. Not all of them may be relevant for a particular assessment; only the attributes that are most appropriate should be selected to evaluate a given system. Suggested criteria to use when selecting the set of attributes can be found in [56].

The assessment of signature quality begins by listing (or calculating) the possible outcomes of the signature systems in terms of attributes of interest for each system, followed by a comparison of the system in terms of its performance with respect to each attribute [52]. For example, suppose some signature system (call it $\hat{\tau}_1$) is better than another system, $\hat{\tau}_2$, for at least one attribute, and that $\hat{\tau}_1$ is at least as good as $\hat{\tau}_2$ with respect to the remaining attributes. Then $\hat{\tau}_2$ would be considered inferior to $\hat{\tau}_1$ and would be removed from future consideration. This process can be repeated for each signature system until all inferior systems are identified. Those signature systems that remain constitute the "Pareto" [57] or "efficient" frontier [53] and form the optimal set of signature systems to choose from for a given application.

Once the single attribute utility functions are identified, they can be aggregated in a linear or multiplicative fashion to form an overall utility function that measures the quality of the signature system. This utility function can then be used to calculate the expected utility for each system under comparison. It may also be used to assess the error associated with the choice of $\mathcal{S}$. Specifically, the overall utility function may be used instead of the traditional loss function when performing cross-validation or bootstrapping [29]. An example of this is provided by Sego et al. [58]. While cross-validation is a powerful statistical technique, caution must be exercised in its application to ensure that realistic comparisons are made between the test and training data. For example, in the validation of signatures with temporal components, cross-validation should not be performed in reverse time; the training data should only include events prior to the test data.

F. Iteration and convergence

The signature discovery process described above is designed to be iterative, converging to a final signature (and by extension, its corresponding signature system) if it achieves predetermined, measurable criteria. These criteria or attributes are determined, in part, by the technology readiness level (TRL) [59] of the signature system. For TRL 1 to 3, the criteria may be less rigorous and focused primarily on the fidelity of the signature system. For higher TRLs, operational considerations informed by stakeholders become increasingly important and would be reflected in the criteria. If there are two or more signature systems under consideration, the extent to which we value one criterion or attribute over another will be required to determine which system is *good enough.* A useful tool for quantifying the extent to which we value one criterion over another is multi-attribute utility theory [53], a cornerstone of signature quality metrics methods described above (see Sec. III-E3).

IV. Example applications

Through a review described above (Sec. III), we identified several articles from the literature that illustrate the signature discovery process shown in Fig. 1.

A. Prognostic breast cancer signatures

The first example application is based on signatures for prediction of clinical outcomes for breast cancer through gene expression profiling by van't Veer et al. [11] to determine the best treatment for patients by predicting clinical outcomes of non-nodal breast cancer. The process detailed by van't Veer et al. is outlined below and closely matches other genomic analysis process examined in our literature review [7], [10].

1) Specify problem: This study focused on the development of prognostic signatures to determine the best treatment for patients by predicting clinical outcomes of non-nodal breast cancer.

2) Inventory observables: Gene expression was identified as the primary observable and selected for this study. Additional patient-specific information such as tumor type, patient age, and patient outcomes were required observables for signature creation and validation.

3) Specify measurements: Measurements of gene expression profiles were performed on a subset of tumor samples with and without non-nodal breast cancer using microarray methods.

4) Assess and explore data: Data were processed for background correction and normalization and gene expression features were extracted using agglomerative hierarchical clustering.

5) Develop signature: The microarray signature was constructed with unsupervised and supervised clustering techniques resulting in a collection of 70 gene expression levels. Patient outcome was predicted through calculation of an odds ratio that utilizes a multivariate model based in part on the correlation with the microarray signature along with clinical parameter correlations. Quality was assessed through leave-one-out cross-validation and analysis of external data not used in the signature construction.

The resulting signature showed good classification accuracy and will allow physicians to pursue courses of action appropriate to the predicted outcome.

B. Chemical forensics signatures

The second example application is based on trace contamination signatures in chemical agents and their precursors for the purpose of forensic attribution [8], [9], [60]. The steps detailed by Fraga et al. [8] for developing such signatures are outlined below and closely follows the signature discovery process outlined in Fig. 1.

1) Specify problem: This study focused on the development of forensic signatures to assist in the attribution of origin for nerve agent chemical precursors.

2) Inventory observables: The impurities present in chemical samples were identified as the primary observables and selected for this study, documentation of suppliers of the unique stocks were also required for signature creation and validation.

3) Specify measurements: The mass and retention time of impurities, measured by liquid chromatography mass spectrometry, were used for the forensic signature.

4) Assess and explore data: Spectra generated in the measurement step were analyzed for peaks utilizing a matched filter tool. An initial data reduction step using the XCMS mass spectrometry metabolite profiling software [61] reduced the number of peaks by selecting those appearing in more than 90% of analyzed samples. Impurity peak features were extracted by hierarchical cluster analysis followed by non-negative matrix factorization to identify the most discriminating impurity profiles for sample matching.

5) Develop signature: The signature was constructed via k-nearest-neighbor clustering techniques that were used to match unknown samples to the impurity profile features. Quality was assessed by matching against signal-averaged test samples.

The resulting signature showed high accuracy in classification of chemical sample origin and will serve as an important tool in forensic analysis of chemical agents of unknown origins. The examination of this work provided a unique opportunity to examine a signature discovery process from peer reviewed literature and then validate the mapping to the generic process through interviews with the researchers.

C. Signatures for computer executable file forensics

The final example is constructed with the primary source being direct input from researcher interviews about methods described in another paper in this workshop [62]. The focus of this work is the examination of signature creation for sequence-based phenomena. While the underlying sequence-based signature development technique is broadly applicable to a variety of such phenomena, the specific application under examination was the identification computer executable file types. This signature development concept is based on transforming instructions (opcodes) in each executable to an alphabet scheme that can take advantage of the well-developed sequence alignment and identification tools used in the genomics and proteomics domains. The signature discovery process steps used in this project are outlined below.

1) Specify problem: This study focused on the development of signatures for the characterization and identification of executable files.

2) Inventory observables: The instruction set from executable files were identified as the primary observables. Additional observable information included known identities and version histories of executable files for testing.

3) Specify measurements: The instruction path for each function in an executable is translated to a protein representation. These protein representations are then compared to one another using a local alignment algorithm (BLAST) that has been encoded into a high-performance version called ScalaBLAST [63]. ScalaBLAST produces similarity scores and confidence measures for the alignments of any two executable file protein representation. These similarity scores form the measurements used for signature development.

4) Assess and explore data: Average linkage (hierarchical) clustering is performed on the output of the ScalaBLAST calculations and analysis is performed to determine the appropriate clustering parameters and to confirm that each cluster is well defined. Each resulting cluster is labeled as a **family** of similar functions that forms the features used for signature construction.

5) Develop signature: The final step in the signature development process is to determine consensus regions of each family by performing multiple alignments on each family using the MAFFT algorithm [64]. It is possible to identify more than one potential consensus region for each functional family and therefore construct more than one signature for each family. Quality of the signature is assessed by using the ScalaBLAST local alignment algorithms to determine the uniqueness and strength of each signature.

V. Conclusion

In addition to the specific steps of the signature discovery process described in this paper, our research has also been guided by a few key postulates. First, we believe that the most useful signatures are human-interpretable such that the logic of the signature and its classification output can be understood by the analyst or decision-maker. "Black box" signatures without such interpretability can provide good statistical fidelity; however, they often lack the confidence of the end-user and can be difficult to troubleshoot or analyze when performance deteriorates. On the other hand, signatures that are amenable to human analysis are more likely to be adopted and can be useful in a wider range of activities; e.g., for the identification of additional useful observables and the guidance of additional data collection. Second, we believe that the most useful signatures are generally "multi-INT" in nature; i.e., composed of features from multiple observables. Signatures built from multiple observables (and associated measurements) often have the advantage of higher statistical fidelity and robustness to noise, interference, and missing data. Of course, the inclusion of multiple measurements often implies additional cost; however, approaches like the signature quality metrics methods (Sec. III-E3) can help determine the right balance between the number of observables, measurements, cost, and the desired signature fidelity.

This paper provides an overview of the domain-independent signature development process developed through the Signature Discovery Initiative (http://signatures.pnnl.gov) along with some simple examples to illustrate its use in real-world problems. Several companion papers in this workshop provide specific details on the various methods developed to support the steps of this process. Our primary objective with the development of this process and its supporting methodologies is to improve signature discovery by making it more reproducible and robust. Additionally, the generalizable nature of this process across several domains offers the opportunity for the reuse of the signature discovery/development methodologies by identifying the steps common across disciplines. Ultimately, the goal of this research, and the larger Signature Discovery Initiative, is to reduce the time and cost for the discovery and deployment of new signatures in whatever domain the signatures may be required.

Acknowledgments

The research described in this paper is part of the Signature Discovery Initiative (http://signatures.pnnl.gov) at Pacific Northwest National Laboratory (PNNL). It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy. The authors acknowledge Janine Anderson for her help with a succinct graphical representation of the Signature Discovery Process (Fig. 1), Mary Zalesny for her contributions to the brainstorming research, Elena Peterson for information about computer file forensics signatures, and Alejandro Heredia-Langner and Mary Lancaster for their review of this document. The authors would like to thank the anonymous reviewer for their helpful feedback.

References

[1] P. Deignan, M. Wong, and A. Douglass, "Low-level multi-INT sensor fusion using entropic measures of dependence," in *Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on*, July 2011, pp. 1–7.

[2] W. Dubitzky, M. Granzow, and D. P. Berrar, Eds., *Fundamentals of Data Mining in Genomics and Proteomics*, 2007th ed. Springer, Dec. 2006.

[3] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recogn. Lett.*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[4] O. Mason and M. H. A. Verwoerd, "Graph theory and networks in biology," *Systems Biology, IET*, vol. 1, no. 2, pp. 89–119, Mar. 2007.

[5] T. D. Veenstra and J. R. Yates, Eds., *Proteomics for Biological Discovery*. Hoboken, NJ, USA: John Wiley & Sons, Inc., May 2006.

[6] B. I. A. Barry and H. A. Chan, "Syntax, and semantics-based signature database for hybrid intrusion detection systems," *Security Comm. Networks*, vol. 2, no. 6, pp. 457–475, Nov. 2009.

[7] A. Dupuy and R. M. Simon, "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting," *Journal of the National Cancer Institute*, vol. 99, no. 2, pp. 147–157, Jan. 2007.

[8] C. G. Fraga, B. H. Clowers, R. J. Moore, and E. M. Zink, "Signature-discovery approach for sample matching of a nerve-agent precursor using liquid chromatography-mass spectrometry, XCMS, and chemometrics." *Analytical Chemistry*, vol. 82, no. 10, pp. 4165–4173, May 2010.

[9] C. G. Fraga, O. T. Farmer, and A. J. Carman, "Anionic forensic signatures for sample matching of potassium cyanide using high performance ion chromatography and chemometrics." *Talanta*, vol. 83, no. 4, pp. 1166–1172, Jan. 2011.

[10] H. Hernandez-Vargas, M.-P. P. Lambert, F. Le Calvez-Kelm, G. Gouysse, S. McKay-Chopin, S. V. Tavtigian, J.-Y. Y. Scoazec, and Z. Herceg, "Hepatocellular carcinoma displays distinct DNA methylation signatures with potential as clinical predictors." *PLoS ONE*, vol. 5, no. 3, 2010.

[11] L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, "Gene expression profiling predicts clinical outcome of breast cancer." *Nature*, vol. 415, no. 6871, pp. 530–536, Jan. 2002.

[12] P. Grenon and B. Smith, "SNAP and SPAN: Towards dynamic spatial ontology," in *Spatial Cognition and Computation*, 2004, pp. 69–103.

[13] A. F. Osborn, *Applied Imagination: Principles and Procedures of Creative Problem Solving*. New York: Charles Scribner's Sons, 1953.

[14] R. Ziegler, M. Diehl, and G. Zijlstra, "Idea production in nominal and virtual groups: Does Computer-Mediated communication improve group brainstorming?" *Group Processes & Intergroup Relations*, vol. 3, no. 2, pp. 141–158, Apr. 2000.

[15] M. Diehl and W. Stroebe, "Productivity loss in brainstorming groups: Toward the solution of a riddle," *Journal of Personality and Social Psychology*, vol. 53, no. 3, pp. 497–509, Sep. 1987.

[16] K. L. Dugosh and P. B. Paulus, "Cognitive and social comparison processes in brainstorming," *Journal of Experimental Social Psychology*, vol. 41, pp. 313–320, 2005.

[17] P. B. Paulus and H.-C. Yang, "Idea generation in groups: A basis for creativity in organizations," *Organizational Behavior and Human Decision Processes*, vol. 82, no. 1, pp. 76–87, May 2000.

[18] A. Furnham, "The brainstorming myth," *Business Strategy Review*, vol. 11, no. 4, pp. 21–28, 2000.

[19] C. D. Schunn, P. B. Paulus, J. Cagan, and K. Wood, "Final report from the NSF innovation and discovery workshop: The scientific basis of individual and team innovation and discovery," National Science Foundation, Tech. Rep., August 2006 2006.

[20] K. Girotra, C. Terwiesch, and K. T. Ulrich, "Idea generation and the quality of the best idea," *Management Science*, vol. 56, no. 4, pp. 591–605, 2010.

[21] V. R. Brown and P. B. Paulus, "Making group brainstorming more effective: Recommendations from an associative memory perspective," *Current Directions in Psychological Science*, vol. 11, no. 6, pp. 208–212, 2002.

[22] N. Michinov and C. Primois, "Improving productivity and creativity in online groups through social comparison process: New evidence for asynchronous electronic brainstorming," *Computers in Human Behavior*, vol. 21, no. 1, pp. 11–28, 2005.

[23] H. Barki and A. Pinsonneault, "Small group brainstorming and idea quality: Is electronic brainstorming the most effective approach?" *Small Group Research*, vol. 32, no. 2, pp. 158–205, 2001.

[24] P. B. Lowry, T. L. Roberts, J. Nicholas C. Romano, P. D. Cheney, and R. T. Hightower, "The impact of group size and social presence on small-group communication: Does computer-mediated communication make a difference?" *Small Group Research*, vol. 37, no. 6, pp. 631–661, 2006.

[25] G. P. Box, J. S. Hunter, and W. G. Hunter, *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd ed. John Wiley & Sons, Inc., 2005.

[26] A. Dean and D. Voss, *Design and Analysis of Experiments*. Springer-Verlag New York, Inc., 1999.

[27] R. H. Myers and D. C. Montgomery, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 2nd ed. John Wiley & Sons, Inc., 2002.

[28] R. L. Plackett and J. P. Burman, "The design of optimum multifactorial experiments," *Biometrika*, vol. 33, no. 4, pp. 305–325, 1946.

[29] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer-Verlag, 2001.

[30] R. R. Hocking, *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. John Wiley & Sons, Inc., 1996.

[31] R. L. Scheaffer, I. William Mendenhall, R. L. Ott, and K. G. Gerow, *Elementary Survey Sampling*, 7th ed. Brooks/Cole, Cengage Learning, 2012.

[32] J. W. Tukey, *Exploratory Data Analysis*. Addison-Wesley, 1977.

[33] A. C. Rencher, *Methods of Multivariate Analysis*, 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc., 2002.

[34] D. Sarkar, *Lattice: Multivariate Data Visualization with R*. Springer, 2008.

[35] I. K. Fodor, "A survey of dimension reduction techniques," Lawrence Livermore National Laboratory, Tech. Rep., 2002.

[36] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Pacific Grove, CA: Duxbury, 2002.

[37] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "The KDD process for extracting useful knowledge from volumes of data," *Communications of the ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[38] C. De Boor, *A Practical Guide to Splines*. Springer, Nov. 2001.

[39] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1988.

[40] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. Chapman & Hall/CRC, 1989.

[41] E. F. Vonesh, *Generalized Linear and Nonlinear Models for Correlated Data: Theory and Applications Using SAS*. SAS Institute Inc., 2012.

[42] D. J. Hand, *Construction and Assessment of Classification Rules*. John Wiley & Sons, Inc., 1997.

[43] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[44] F. V. Jensen and T. D. Nielsen, *Bayesian Networks and Decision Graphs*, 2nd ed. Springer-Verlag, 2007.

[45] V. Cherkassky and F. M. Mulier, *Learning from Data: Concepts, Theory, and Methods (Adaptive and Learning Systems for Signal Processing, Communications and Control Series)*, 1st ed. Wiley-IEEE Press, Mar. 1998.

[46] I. Gorton and D. K. Gracio, Eds., *Data-Intensive Computing: Architectures, Algorithms, and Applications*. Cambridge University Press, Oct. 2012.

[47] T. Hey, S. Tansley, and K. Tolle, Eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Oct. 2009.

[48] D. Laney, "3D data management: Controlling data volume, velocity, and variety," META Group, Stamford, CT, Tech. Rep., February 2001.

[49] B.-J. J. Webb-Robertson, H. Kreuzer, G. Hart, J. Ehleringer, J. West, G. Gill, and D. Duckworth, "Bayesian integration of isotope ratio for geographic sourcing of castor beans." *Journal of Biomedicine & Biotechnology*, vol. 2012, 2012.

[50] B.-J. Webb-Robertson, C. Corley, L. A. McCue, K. Wahl, and H. Kreuzer, "Fusion of laboratory and textual data for investigative bioforensics," *Forensic Science International*, Jan. 2013.

[51] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, 1985.

[52] W. Edwards, R. Miles, and D. von Winterfeldt, *Advances in Decision Analysis: From Foundations to Applications*. Cambridge University Press, 2007.

[53] R. L. Keeney and H. Raiffa, *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: John Wiley & Sons, Inc., 1976.

[54] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, Jul. 2009.

[55] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. New York, NY, USA: Cambridge University Press, 2011.

[56] R. L. Keeney, *Value-Focused Thinking. A Path to Creative Decision Making*. Cambridge: Harvard University Press, 1992.

[57] J. Henderson and R. Quandt, *Microeconomic Theory: a Mathematical Approach*, ser. Economics Handbook Series. McGraw-Hill, 1980.

[58] L. H. Sego, A. E. Holmes, L. J. Gosink, B.-J. M. Webb-Roberton, H. W. Kreuzer, R. M. Anderson, A. J. Brothers, C. D. Corley, and M. R. Tardiff, "Assessing the quality of bioforensic signatures," in *Proceedings of the 2013 IEEE Intelligence and Security Informatics Conference, June 4-7*, 2013.

[59] Assistant Secretary of Defense for Research and Engineering, "Technology readiness assessment (TRA) guidance," United States Department of Defense, Tech. Rep., 2011.

[60] C. G. Fraga, L. H. Sego, J. C. Hoggard, G. P. A. Acosta, E. A. Viglino, J. H. Wahl, and R. E. Synovec, "Preliminary effects of real-world factors on the recovery and exploitation of forensic impurity profiles of a nerve-agent simulant from office media." *Journal of Chromatography A*, vol. 1270, pp. 269–282, Dec. 2012.

[61] C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan, and G. Siuzdak, "XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification," *Anal. Chem.*, vol. 78, no. 3, pp. 779–787, Jan. 2006.

[62] E. S. Peterson, D. S. Curtis, A. R. Phillips, J. R. Teuton, and C. S. Oehmen, "A generalized bio-inspired method for discovering sequence-based signatures," in *Intelligence and Security Informatics: IEEE International Conference on Intelligence and Security Informatics*, Seattle, WA, June Submitted.

[63] C. S. Oehmen and D. J. Baxter, "ScalaBLAST 2.0: rapid and robust BLAST calculations on multiprocessor systems," *Bioinformatics*, vol. 29, no. 6, pp. 797–798, Mar. 2013.

[64] K. Katoh and D. M. Standley, "MAFFT multiple sequence alignment software version 7: Improvements in performance and usability," *Molecular Biology and Evolution*, vol. 30, no. 4, pp. 772–780, Apr. 2013.