INTELLIGENCE INSIGHTS

PRACTITIONER VIEWS ON INTELLIGENCE
PROGRAM DESIGN AND IMPLEMENTATION

JULY 2018

An Assessment of Data Analytics Techniques for Insider Threat Programs

PRESENTED BY INSA'S INSIDER THREAT SUBCOMMITTEE

EXECUTIVE SUMMARY

A wide array of data analytics methods, tools, and techniques exist to improve the detection and mitigation of insider threats – trusted employees who seek to steal an organization's data or intellectual property or to harm an organization or its staff. Determining which data analytics methods and software tools are best for an organization, however, depends on the quality and comprehensiveness of data, the clarity of rules, the organization's risk tolerance, and other factors.

This paper presents a framework which plots the continuum of data analytics techniques used to solve specific insider threat problems. Decision-makers and insider threat program managers in both government and industry can leverage this framework to evaluate the merits of different analytic techniques and then choose or develop tools to address their most pressing needs.

To identify the type of tools that would be most beneficial to their organizations, insider threat managers should take four steps: (1) integrate data analytics into their risk management methodologies; (2) assess which analytic techniques are likely to be most effective given the available data, their organizational structure and culture, and their levels of risk tolerance; (3) evaluate the myriad software tools that evaluate data using the chosen approach; and (4) assess the human and financial resources needed to launch a data analytics program, including the expense of software tools and the training and time needed to structure data, apply tools, and execute a data analytics initiative over time.

INTRODUCTION

A wide array of data analytics methods, tools, and techniques exist to improve the detection and mitigation of insider threats – trusted employees who seek to steal an organization's data or intellectual property or to harm an organization or its staff. Insider threat program managers do not have a standard reference document to consult when determining which of these techniques and tools best address the challenges specific to their organizations. In the absence of such a guide, organizations often move forward without a strategic approach, which leads to ineffective and/or unnecessary spending.

INSA designed a framework which plots the continuum of data analytics techniques used to solve specific insider threat problems. Decision-makers and insider threat program managers can leverage this framework to evaluate the merits of different analytic techniques and then choose or develop tools to address their most pressing needs.

OUR APPROACH

A working group of INSA's Insider Threats Subcommittee looked across the spectrum of data analytics techniques currently used in insider threat programs. Many data analytics techniques are subsets of others. For example, clustering algorithms (such as random forests) and methods to extract patterns in unstructured text (such as natural language processing) are subsets of machine learning. Our assessment was agnostic regarding specific products; we do not seek to endorse any particular product or tool developer.[1]

We determined that there are six primary techniques used in insider threat programs:

1. Rules-based engines
2. Correlation and regression statistics
3. Bayesian Inference/ Bayesian Networks
4. Machine Learning (Unsupervised)
5. Machine Learning (Supervised)
6. Cognitive/Neural Networks/Deep Learning

We then established that there are two lenses through which one can view these techniques:

1. LENS 1: *Descriptive vs. Predictive*
   Does the technique answer "what *has* happened," or does it try to answer "what *could* happen?"
2. LENS 2: *Traceable vs. Black Box*
   Does the technique allow the user to understand exactly why it produced the outputs it did, or are the inputs hidden and not easily explained?

Figure 1 shows a 2 x 2 matrix of these lenses, into which we grouped the five principal techniques under consideration.

Because it is somewhat of a contradiction in terms to discuss a descriptive technique that conceals data inputs and analytical methods, no techniques fall into the lower right quadrant.

It is important to consider that the best tools and programs often use combinations of these techniques. As an example, probabilistic models are often best combined with rules-based systems and machine learning algorithms.

FIGURE 1

MATRIX FOR ORGANIZING ANALYTICS TECHNIQUES

	TRACEABLE	"BLACK BOX"
PREDICTIVE	Bayesian Inference	Cognitive/ Neural Networks Supervised and Unsupervised Machine Learning
DESCRIPTIVE	Rules-Based Simple Correlations	

[1] Gartner and Forrester Research provide extensive reviews of many technology providers. Federally Funded Research and Development Centers (FFRDCs), which conduct direct testing and evaluation of products, and are often used by the Government to determine the best tools for their needs.

TECHNIQUE ASSESSMENTS

Figure 2 captures the advantages and disadvantages of the six analytic techniques widely used in insider threat programs, each of which is described in detail below.

FIGURE 2

SUMMARY OF DATA ANALYTICS TECHNIQUES

	DESCRIPTION	*ADVANTAGES*	*DISADVANTAGES*	*INSIDER THREAT APPLICATIONS*	*EXAMPLES*
RULES-BASED	Uses "if-then" statements to determine actions	Easy to understand Characterizes data in light of rules and policies Represents expert judgment on simple or complicated subjects Cause and effect relationships are transparent	Doesn't handle incomplete information well Data that does not have an associated rule will be ignored Can't handle continuous variables	Flag whether subject matches conditions requiring review	DoD's Automated Continuous Evaluation System *(Mirador)*
CORRELATION	Analyzes how data is related to other data within the data set	Automates the matching and linking of data records (structured or unstructured) Uncovers non-obvious relationships to multiple degrees Reduces false positives and negatives	Heavily reliant on data (accuracy increases as data records increase) Business-specific matching algorithms need to be developed	Assess potential security implications of data (e.g., likelihood that overdue debts indicate financial problems)	Entity/Identity Resolution Anomaly detection Link Analysis Vetting Watchlisting
BAYESIAN INFERENCE	Infers likelihood of something occurring given available evidence	Solves complex problems in areas without much data or history (e.g., insider threats) Automates reasoning of subject matter experts Cause and effect relationships are transparent Requires less data Very powerful when combined with hybrid systems	Can require significant expertise to build Not always intuitive	Assess risk levels even when data is incomplete Quantify degree of uncertainty in analysis so adjudicator can make informed decision	DoD's Continuous Evaluation Risk Rating Tool Terrorism risk models Spam filters
MACHINE LEARNING (UNSUPERVISED)	Data-driven approach that doesn't use/rely on labeled examples	Automates analysis of big data Finds previously unseen patterns	Heavily reliant on data Cause-and-effect relationships not transparent *("black box")*	Comparing subjects' traits to those of known malicious insiders	Google Search User Activity Monitoring Anomaly Detection Pattern recognition
MACHINE LEARNING (SUPERVISED)	Data-driven approach that uses/ relies on labeled examples	Automates analysis of big data Finds previously unseen patterns More transparent and "tweakable" than unsupervised learning	Heavily reliant on data May miss non-obvious relationships	Extrapolating presence of risk from existence of known risks in similar subjects	Government Risk Rating Tool Update taxonomies Identify psychoanalytic motivations
COGNITIVE	Mimics the human brain by using multiple data analysis techniques to learn, adapt and reason from the data it ingests	Creates deeper human engagement Scales and Elevates Expertise Enables Cognitive Processes Enhances exploration and discovery	Heavily reliant on data Need to initially train the system	Interpreting data in light of adjudication guidelines Comparing initial adjudications of subjects and individuals known to have become insider threats Identifying anomalies in behavior and communications	Psycholinguists Behavioral Analytics Emotion Detection Event Detection Query and Answer

RULES-BASED[2]

Rules-based systems use "if-then" rules to derive actions. They are simple and direct, automating policies and procedures that already exist. For example, a policy might state that someone with debts that are more than 120 days overdue should be reviewed by security clearance adjudicators. In a rules-based system, as soon as someone is over 120 days delinquent, an alert is generated with instructions for review.

Rules-based systems are relatively easy to understand and explain, and can be built to represent expert judgment on simple or complicated subjects. Cause and effect triggers are transparent—there is no "black box." Even though the if-then reasoning can become complex, a domain expert can verify the rule base and make adjustments when necessary.

That said, there are three critical weaknesses to rules-based systems:

- Rules engines often become almost as complicated as the problem the system is trying to solve. Unlike machine learning, rules cannot be learned and must instead be added manually. As a result, rules-based systems become difficult to understand in the aggregate. The more knowledge (rules) a user adds, the more complex and opaque the system can become.
- Rules-based systems do not handle incomplete or incorrect information very well, and data that does not relate to a rule will be disregarded. This means that rules-based systems are particularly bad at detecting 'unknown unknowns' like unreported cash income or unreported foreign travel.
- Rules-based systems do not know what to do with variables that have an infinite number of possible legitimate values. For example, a computer might raise a red flag if a person spends an undue amount of time in a sensitive database, but it cannot know how much time the employee might legitimately need to spend perusing the database to perform his or her job. It is difficult to derive a rules-based system from subjective factors, such as the length of time it takes to do a task, the hours one spends in the office, or the size of a project team.

Arbitrarily converting these continuous variables into discrete variables means potentially missing patterns or deriving false patterns. For example, a rule flagging someone who is more than 120 days delinquent on a single debt payment may miss someone else who has missed payments for years but was never more than 119 days late for any account. The chronic delinquent is likely a higher risk than the one-time offender. Similarly, while a credit score above a designated threshold might make an employee seem more stable and conscientious than a lower credit score, a high credit score in decline might indicate emerging financial problems, whereas a low score on the rise might indicate fiscal responsibility.

Rules-based systems can best be used as foundations for probabilistic and machine learning systems. For example, in all models, there are hard-rules (e.g., you must be a U.S. citizen to obtain a clearance), which can be implemented as policy regardless of the amount of risk they may indicate. Rules-based systems can also be used to sort or highlight information after it has been analyzed. For example, rules can be used to dictate what to display on dashboards or how to organize work-flows.

There are many examples of rules-based systems. The Department of Defense's (DoD) continuous evaluation program, Mirador, reflects the federal investigative standards and adjudicative guidelines. Mirador is useful for automating the existing manual process for flagging cleared persons according to established rules. However, Mirador does not assess a person's risk level. These subjective evaluations are calculated by humans (adjudicators) who decide whether, based on the available data, to grant or sustain a clearance eligibility.

CORRELATION

A **Correlation Analysis** is the statistical tool used to study the possibility and closeness of the relationship between two or more variables. The variables are said to be correlated when the movement of one variable is accompanied by the movement of another variable. It is particularly valuable in:

- Determining whether and to what degree a relation exists (The measure of correlation is called the *Coefficient of Correlation*);
- Testing the significance of the relationship;
- Establishing the existence, if any, of a cause-and-effect relationship.

[2] Adapted from Tom Read, "Three Weaknesses of Rules-Based Systems," Haystax Technology Blog, September 6, 2016. At https://haystax.com/blog/2016/09/06/three-weaknesses-of-rules-based-systems/.

In a correlation analysis, there are two types of variables: **Dependent and Independent.** The purpose of such analysis is to find out if any change in the independent variable results in the change in the dependent variable or not. Several variables may show some kind of relationship, such as income and expenditure, demand and sales, etc. With the help of correlation analysis, the degree of the relationship between two variables can be measured in one figure. As an example, the financial services industry uses correlation to gain maximum insights into consumers' risk trajectory toward bankruptcy and collections; banks will examine the prevalence of bills more than 90 days past due to assess a client's risk before extending a loan. These tools can be repurposed for security risk insights as well. Analyzing an employee's payment history on all outstanding loans will help security personnel understand which employees are heading toward financial distress before they reach a crisis point.

Once the closeness of variables is determined, one can use regression analysis to estimate the value of an unknown variable, provided the value of another variable is given.

STATISTICAL INFERENCE (BAYESIAN)

A Bayesian network is a statistical model for reasoning about complex problem domains. Bayesian networks provide a way to make inferences even when evidence is missing, incomplete, or inconsistent. Note that the inference results will depend on the strength, completeness and consistency of the evidence; strong consistent evidence will yield a strong Bayesian network inference, giving a clear answer. If evidence is sparse, weak, or inconsistent, the Bayesian network will reflect the uncertainty inherent in the knowledge base. For example, if a subject's spouse's income is unknown, a comparison of known household debt levels to total household income will yield uncertain results. A Bayesian model can identify and quantify the uncertainty present in the resulting analysis.

Bayesian networks can advance insider threat assessments in several ways:

- Bayesian networks provide a statistically sound way to combine heterogeneous data even when the data is incomplete, uncertain, or contradictory.
- The degree of uncertainty is made explicit, enabling an adjudicator to assign risk.
- Evidence is statistically weighted, so that even weak evidence can contribute to an overall risk assessment.
- Bayesian networks can be learned from data, constructed from the knowledge of experts, or a combination of both.
- Bayesian networks are an explicit and meaningful representation of known information, which can facilitate communication between experts, decision makers, and users.

The primary shortcoming of Bayesian networks is that many modeling experts do not have the required domain knowledge and thus must extract the information from an organization's internal experts, policy documents, knowledge bases and/or other sources. This knowledge elicitation is a time-consuming and expensive process, tying up domain experts and delaying implementation of the solution.

Results from Bayesian models can be used in other evaluation techniques. For example, the Commercial Risk Rating Tool used by the Army and the Department of Defense is a Bayesian-based algorithm that ranks cleared persons by riskiness according to the likelihood that they are trustworthy. The results are fed into the DoD's rules-based Mirador system.

MACHINE LEARNING

Machine learning is an application of artificial intelligence that enables a machine to identify patterns in data and interpret their significance or meaning without being programmed to do so, as well as to improve its performance through an analysis of its previous experience. Unlike Bayesian inference, machine learning requires data —and the more data the better.

There are two fundamental types of machine learning: supervised learning and unsupervised learning.

- In **supervised learning** the data are labeled. For example, insider threat programs that use supervised learning rely on past information on insider incidents to train the model to discern similar patterns of behavior. The labels that assign 'threat' or 'not a threat' to each of the individuals in such a model is the 'supervision' part of supervised learning. Supervised learning works well when a great deal of labeled data is available, especially when there are a number of examples of every possible kind of 'threat'. But supervised learning does not work well in the absence of extensive labeled data or where available examples of threats do not reflect the full range of threats.

- **Unsupervised learning** does not start with labeled data. It clusters individuals into similar groups. The challenge is the word 'similar', and many statistical methods can be employed to define or estimate similarity. If a known threat is present in one of the clusters, the model could determine that everyone in that cluster (who is 'similar' to the threat) is also a threat, which could yield many false positive judgements. An unsupervised approach would be more accurate when a great deal of data is available *and* if it was believed that most of the threats have similar characteristics.

In practice, a machine learning approach follows five steps:

1. **Identify the response variable:** This is the variable to be analyzed (e.g., 'subject traveled to a country of concern, or not').
2. **Select features relevant to the response variable:** Features are a particular subset of data, or calculations from data, that are hypothesized to be predictive. For instance, words like "flight," "hotel," or "visa" in email messages could indicate that the sender is planning foreign travel.
3. **Choose a machine-learning algorithm:** For classification problems (e.g., 'subject traveled to a country of concern, or not'), examples of classification machine-learning algorithms are: logistic regression, naïve Bayes, neural network and support vector machines. If the response variable is continuous rather than an either/or option (e.g., a credit score, which can vary along a scale), algorithms such as a regularized form of linear regression might be appropriate.
4. **Tune hyper-parameters:** In this stage, the machine learning algorithm will generally have parameters that need tuning. These are often selected by a technique such as cross validation.
5. **Test:** It is important to test the performance of the algorithm. Usually, a technique like k-fold cross validation is used to compute metrics such as error rates, precision, recall, etc.

Machine learning's ability to cope with massive quantities of data is offset by the fact that it is entirely dependent on data, and thus is unable to offer solutions in cases where data is scarce.

Insider threat programs use many different types of machine learning. Most common are the algorithms used to detect anomalous behaviors/patterns from network activity. User Activity Monitoring (UAM), Data Loss Prevention (DLP), and other network tools use machine learning algorithms because they are comparing current behavior to known baselines in an attempt to identify anomalous activities – for example, if a user spends atypical amounts of time in a sensitive area of a network or prints unusually large amounts of materials.

COGNITIVE

Cognitive behavioral computing mimics the human brain by using multiple data analysis techniques to understand, reason, and learn from the data it ingests and then provides a capability to interact with humans. Using behavioral models based on predetermined personality traits, rule sets can be developed to determine collection methodology.

Cognitive computing enables the analysis of data in the following ways:

- **Understanding data,** both structured and unstructured, and both text-based or sensory, in context and meaning, at astonishing speeds and volumes.
- **Reasoning,** principally by forming hypotheses, making considered arguments, and prioritizing recommendations to help humans make better decisions.
- **Learning from previous data analysis,** a skill enabled by experts who train (rather than program) the system to enhance, scale and accelerate its expertise. As a result, the system improves its performance over time.
- **Interacting with users,** responding and communicating with people in a natural way.

Unlike traditional programmable systems that are deterministic and thrive in structured data, cognitive systems are probabilistic and thrive in unstructured data while also reasoning and offering hypotheses based on their behavioral models. Cognitive systems also extend the types of unstructured data beyond text to visual (images and video) and audio. They can understand what they "see"; they can find a plane in an image, identify atypical situations (such as a package left unattended on the subway), or notice abnormal behavior in video (such as driving on the wrong side of the road). Cognitive systems can also understand the meaning of recorded audio through the use of natural language processing.

Cognitive learning, when integrated with behavioral modeling and psycholinguist technologies, can make complex tasks possible. For example, by changing speech to text and leveraging the power of natural language processing, organizations can scan calls, emails, and social media and chat applications to drive insight and better understand a person's wants, needs, emotions, and intentions.

This capability has two clear applications for insider threat assessments:

1. **Adjudication:** Cognitive behavioral computing can be used to aid the adjudication of applicants by not only ingesting data on an applicant, but also by interpreting the data in light of the key underlying personality factors imbedded in the adjudicative guidelines. Moreover, in interpreting data on an individual, the computer can learn from and mimic adjudicators' previous use of those guidelines to make decisions. As cognitive systems learn about past decisions, they can then apply this reasoning to new cases, thus providing consistent standardized adjudicative determinations.
2. **Risk Assessment & Continuous Evaluation:** Cognitive computing allows insider threat programs to continually compare current employees' behaviors and expressed attitudes to those of other employees known to have been insider threats and to models of insider threat behavior.[3] Furthermore, by examining the initial adjudication of employees who later became malicious insiders, cognitive computing can develop a model that uses initial adjudicative data to identify the risk that a given employee could also become an insider threat.

RECOMMENDATIONS FOR DECISION-MAKERS AND PROGRAM MANAGERS

Managers of insider threat programs in both government and industry can take several steps to identify the types of tools that would be most beneficial to their organizations:

1. **Integrate data analytics into the risk management methodology they use to rationalize decision-making.** The methodical analysis of available data can help organizations better identify, weigh, and assess the factors that could increase the likelihood a trusted insider will act maliciously.
2. **Assess which techniques explored in this paper are most likely to be effective given the available data, their organizational culture, and their levels of risk tolerance.** Program managers should assess different combinations of techniques, as the availability of data may make some combinations more effective than others at identifying and calculating risk.
3. **Evaluate the myriad software tools available that most effectively evaluate data using the preferred approach.** [An evaluation of individual commercial products is beyond the scope of this paper.]
4. **Assess the human and financial resources needed to launch a data analytics program,** including the expense of software tools, the training and time needed to structure data and apply tools, and a clear definition of the skills that program staff need in order to develop, maintain, and execute a data analytics initiative over time.

[3] *See Intelligence and National Security Alliance (INSA), Assessing the Mind of the Malicious Insider: Using a Behavioral Model and Data Analytics to Improve Continuous Evaluation, white paper, April 2017. At https://www.insaonline.org/wp-content/uploads/2017/04/INSA_WP_Mind_Insider_FIN.pdf.*

INTELLIGENCE AND
NATIONAL SECURITY
ALLIANCE

ABOUT INSA

The Intelligence and National Security Alliance (INSA) is a nonpartisan, nonprofit forum for advancing intelligence and national security priorities through public-private partnerships. INSA strives to identify, develop, and promote collaborative approaches to national security challenges, and it works to make government more effective and efficient through the application of industry expertise and commercial best practices. INSA has more than 160 organizations in its membership and enjoys extensive participation from leaders, senior executives, and intelligence experts in government, industry, and academia. Learn more at www.INSAonline.org.

ABOUT INSA'S INSIDER THREAT SUBCOMMITTEE

INSA's Insider Threat Subcommittee researches, discusses, analyzes, and assesses counterintelligence and insider threat issues that affect government agencies (particularly, but not only, those working on intelligence and national security issues), cleared contractors, and other public and private sector organizations. The objective of the Subcommittee's work is to enhance the effectiveness, efficiency, and security of both government agencies and their industry partners, as well as to foster more effective and secure partnerships between the public, private and academic sectors.

ACKNOWLEDGEMENTS

This paper was developed and written by a Data Analytics Working Group operating under the auspices of INSA's Insider Threat Subcommittee.

Thanks are due to:

Tom Read, *Haystax Technology; Working group lead*

Sandy MacIsaac, *Deloitte; Subcommittee chair*

Vincent Corsi, *IBM; Subcommittee vice chair*

Rob Campbell, *ForcePoint*

Thomas Crawford, *General Dynamics Mission Systems*

Cory Davenport, *University of Maryland*

Bryan Denson, *TransUnion*

Daniel McGarvey, *Alion Science and Technology*

Doug Thomas, *Lockheed Martin*

Appreciation is also due to INSA staff members who contributed to the publication:

Chuck Alsup, *President*

Larry Hanauer, *Vice President for Policy*

Ryan Pretzer, *Senior Manager,*
Policy and Public Relations

Ehrlich Alba, *Digital Marketing Manager*

Colin Blowers, *Intern*

Bill Harley, *Intern*