Carnegie Mellon University
Software Engineering Institute

# THE PROCESSES OF INSIDER THREAT ANALYSIS

*Angela Horneman*
*Robert M. Ditmore*
*Derrick Spooner*

March 2021

## Introduction

Insider threat analysis comes in many shapes and forms, each with its own use cases, prerequisites, and pros and cons. The purpose of this paper is to lay out a framework that will help organizations understand and examine different analytical techniques, applications, and use cases so they can select and apply the most effective methods to improve their capabilities for analysis.

Analysis that supports the detection, response, and investigation of insider threat risks and events forms a major part of any insider threat program. For this paper, we use the term "analysis" to refer to the process of using data, context, and techniques to answer questions or to develop theories or hypotheses. This process allows practitioners to detect and vet potential insider risk events and to set operational risk tolerances.

Throughout this paper we also use two other phrases that are important to define up-front: analysis capability and analytic technique. We use analysis capability—or capability for analysis—to refer to the knowledge and resources required to apply analysis techniques or overall processes for analysis, which we also refer to as the analysis process. When we use the phrase analytic technique, we mean a method that is applied to data to extract information for interpretation. This term contrasts with "analysis," which we use to mean the full process. To clarify the relationship between the terms, analysis capabilities are used to apply analytic techniques within a process to perform analysis.

The paper is organized into four main sections. The first section outlines various process constructs that you can apply to insider threat analysis. In the second section, we discuss the process of collecting data and analyzing it to get observables. The second section also covers how to obtain indicators from observables, and behaviors from indicators. The third section examines how these concepts apply to several common insider threat analysis goals. Finally, the last section discusses how to measure the effectiveness of the insider threat analysis you use at your organization.

## Process Constructs for Insider Threat Analysis

Although there are many ways to classify analysis processes, for the purposes of this paper, we organize our initial discussion around the "five W and one H" (5W1H) questions: who, what, when, where, why, and how.

These questions are useful for our discussion because the answers determine the data and tooling required to implement an analysis for a given use case. In addition, the information provided by the answers as well as the determination of the data and tooling requirements will also help understand the assumptions that managers, analysts, and tool designers make when applying or designing analyses. This information will give you a better idea of the biases that exist in your analysis and that you will need to mitigate.

It is worthwhile to clarify that the discussion of the 5W1H questions focuses on *atomic* analysis processes. An *atomic* analysis process takes input data, performs a processing function, and obtains an output (see Figure 1). In many cases, the analyses used in insider threat programs are composed of multiple atomic analysis processes, where outputs of one or more atomic processes feed into others (see Figure 2), with each step progressing towards an end goal. The analyses performed in each of these steps may have different answers to the 5W1H questions.

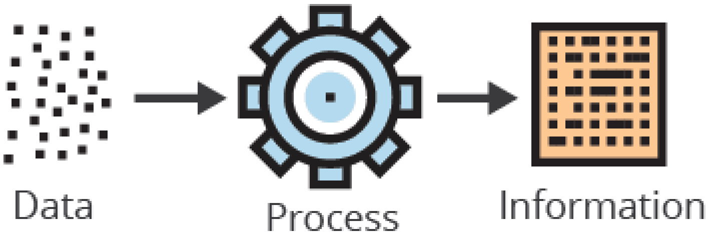

*Figure 1: Atomic Analysis Process*

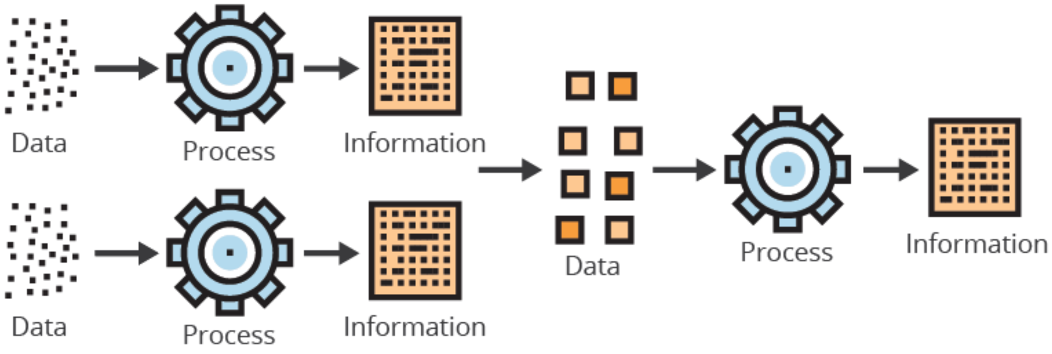

*Figure 2: Composite Analysis Process*

In the following sections, we discuss each of the six questions in the 5W1H framework. The aim of this discussion is to establish how get useful answers for each question to help decide how to perform the analysis required for a given use case.

## *Who* Will Perform the Analysis?

The first W question, "who," specifies the entity required to perform the analysis for a given use case. This "who" is not necessarily, or even usually, a human. Rather, it is any entity that can perform or execute an analysis. The potential answers to the "who" question are as follows:

- **Analyst**: A human analyst that applies personal discretion and expertise to performance of analysis. He or she may construct analysis processes or execute playbooks using standard analysis tools (e.g., spreadsheets, scripts, or generic analysis programs), which he or she either manually executes or automates with something like scheduled tasks. The analyst directly interprets results. Often, these analyses are for tasks such as threat hunting, forensic investigation, and incident response.
- **Machine**: An application that performs a specific task or set of tasks and executes analysis without (human) analyst intervention, often on a continual basis. If the machine acts on results automatically, discretion and decision making is encoded in the application. Otherwise, specific results can be referred to analysts as alerts or stored for later use by analysts. Often, these analyses apply to tasks such as threat detection.
- **Analyst and Machine Team**: An application that allows direct analyst interaction to improve the analysis process or influence the decisions made on results. For example, an application might allow analysts to provide feedback on alerts or decisions which is incorporated automatically into future analyses. This process allows for the benefits of automation, and often machine learning, while providing analysts the ability to apply some level of discretion and expertise.

## *What* Form of Analysis Will We Apply to Get the End Result We Want?

The second W question, "what," asks about the logical goal for the analysis that we are trying to apply to a particular use case. The answer to the "what" question will indicate the end result of the analysis—generically, what kind of insight we are trying to gain from the information we have. This question is similar to the "why" question, but with the point of providing an idea of how the analyst needs to transform the inputs to reach the goal. Possible answers to the "what" question are as follows:

- **Mapping**: Analyses that assign evidence to actions of violation. For example, an alert generated by a data-loss prevention system or finding a list of unauthorized users that attempted to access a sensitive file.
- **Associating**: Analyses that combine related evidence. For example, associating a series of failed password login attempts to password cracking or credential theft, or relating retrieval of proprietary information to a subsequent data upload to an external server.
- **Descriptive**: Analyses that describe a set of data, such as summary statistics or pattern identification.

## *When* Will the Analysis Need to be Executed?

The third W question, "when," specifies the time when you will need to execute an analysis given the use case. Answers to the "when" question will indicate the point in time in the incident-mitigation process when analysis is meant to occur in relation to some reference, rather than in relation to clock time.

For insider threat, the reference point can be the occurrence of an insider event. With this reference point, possible answers are as follows:

- **Pre-Event**: Analysis of data generated before a violation or threat activity.
- **Mid-Attempt**: Analysis of data generated during an event, in real time.
- **Post-Event**: Analysis of data related to violation or threat activity for event detection or response after the fact.

## *Where* Will the Analysis be Performed?

The fourth W question, "where," establishes the location where the analysis will occur. You can determine the location based on both the use case and the tools used to gather data and implement the analysis. Potential answers are as follows:

- **Endpoint**: Analysis takes place on the device where data is generated, such as forensic investigation.
- **Tool Specific Hub**: An application collects data from endpoints and federates it to a centralized server for analysis, such as UAM.
- **Centralized Hub**: The location where data is aggregated from multiple applications (applications themselves may process data from the centralized hub or only feed data to it). Permits analysis across a range of data sources for multiple users, such as a SIEM or UEBA tool might do.
- **Analyst Specific Workspace**: An area where an analyst pulls data on an as-needed basis. Depending on the use case and situation, this could be a system, application, or even a physical workspace.

## *Why* Are We Performing the Analysis?

The fifth W question, "why," establishes the purpose for conducting an analysis. In contrast to the "what" question, the answer to the "why" question is based on the domain-specific goal of the use case. Possible answers could be as follows:

- **Threat or Violation Detection**: Analysis identifies occurrences of violations or breaches (successful or attempted).
- **Risk Evaluation**: Analysis determines a risk score for a set of evidence.
- **Prediction**: Analysis assigns a likelihood of a future event.
- **Decision Support**: Analysis recommends a response or recommendations for next steps based on inputs.

## *How* Are We Going to Construct the Analysis Process?

The answer to the H question, "how," establishes the method for conducting the analysis required for the given use case. It provides information to define the process of analysis and about process repeatability. Possible answers are as follows:

- **Ad-Hoc**: An analyst applies analysis as seems best at the time.
- **Defined Process**: An analyst or machine performs analysis as a set of defined steps or criteria.
- **Learned**: A machine is "taught" a model (either on a periodic or on-going basis) and evaluates new data based on the learned model.

### More On "How" Methods

The methods of the "how" analysis deserve further commentary as they will factor heavily into later discussion on use cases and efficacy.

**Ad-hoc analyses** do not have a defined process, and analysts might execute steps differently and in different sequences at given times. Consequently, they are always performed by humans. In insider threat programs, it is often necessary to enable ad-hoc analysis by analysts to answer one-off questions. Ad-hoc analysis can also serve as part of situations such as an incident inquiry or investigation in which the analyst might not know what is needed to find the answers at the beginning of the process. However, insider threat programs where the majority of analyses are ad-hoc are likely fairly immature in their capabilities. Ad-hoc analysis can be difficult, if not impossible to scale, replicate, and transfer to other analysts. Compared to defined processes, they can also be more prone to discrimination that arises from individual analysts' own biases.

**Defined analyses** have concrete steps that are documented, for example, in a standard operating procedure like a playbook. Such documentation mitigates the issues that occur with ad-hoc analysis. It is much easier to replicate these analyses and transfer the process to other analysts. In addition, while defined analyses can be executed by human analysts, they are also candidates for automation if appropriate technology exists. Automation alleviates scaling issues.

**Learned analyses** use models that are built with machine learning. You can view these analyses as having two distinct pieces—the build of the model (the machine learning part), and the use of the model for analysis. Depending on the use case and the tools you use, an insider threat program may or may not be involved in the building of the model. Application of a model for analysis is almost always done with some automated method, usually from a commercial or in-house application.

There are numerous considerations involved in building a model. We can begin by setting aside the diversity of machine learning methods and focus instead on considering whether to apply the methods using either an ad-hoc or a defined process. Ad-hoc model creation is usually accomplished with in-house applications or scripts as developers work out the solution. These models may be put to use in the insider threat program, but because all models require periodic, if not continual, updates and performance monitoring, organizations should strive to define a repeatable process that covers everything from data collection and preprocessing to model building and deploying the model to "production." This defined process should be automated as much as is feasibly possible.

## 5W1H Examples

Now that we have covered the 5W1H questions with respect to insider threat analysis, the following sections provide a few examples of how possible answers to each question could help you define required analysis for several insider threat scenarios.

### Exfiltration after Solicitation from Competitor

Exfiltration after solicitation from a competitor occurs when an insider in your organization is approached by a competitor for the purpose of information gathering (e.g., corporate espionage). The insider then removes sensitive data from your organization's systems with the purpose of sending it to the competitor. To detect exfiltration after solicitation, we need a way to identify the solicitation and the transfer of information.

*Who will perform the analysis?* For this use case, one possible answer is that analysis could be performed by an analyst as part of an investigation or threat hunting exercise. Another possible answer is that it could be performed by a machine through the implementation of data-leak prevention tools and by monitoring communications for specific contents.

*What form of analysis will we apply to get the end result we want?* The detection of solicitation and exfiltration are mapping analyses. Linking them together as part of a single chain of events is an associating analysis.

*When will the analysis need to be executed?* For this use case, all analyses occur post-event for detection of a violation.

*Where will the analysis be performed?* The answer to this question will depend on context. In a common scenario, it is likely that exfiltration is caught on the endpoint or in a tool-specific hub, and that an analyst will manually link solicitation and exfiltration in a centralized repository or in their local analysis workspace.

*Why are we doing the analysis?* For this use case, analyses are performed to detect policy violations.

*How are we going to construct the analysis process?* The answer for this use case will depend on whether the analysis is performed by security software or by analysts directly. Security applications will execute defined processes for exfiltration and solicitation detection. It is possible that they may also use some form of learning to define their detection criteria. Analysts may use either an ad-hoc or a defined process for analysis.

### Mass File Deletion

Mass file deletion occurs when many files on a single system are deleted or when a single user deletes a large number of files across multiple systems in a short period of time. Such an event might signal the possibility that a user is attempting to sabotage a system or destroy evidence.

*Who will perform the analysis?* One possible answer for this use case is that the analysis could be performed manually by an analyst as part of an investigation or threat hunting exercise. Another possible answer is that the analysis can be performed by a machine using threat detection tools.

*What form of analysis will we apply to get the end result we want?* Detection of mass file deletion involves an associating analysis. It is performed by linking together multiple file deletion events into a mass deletion event.

*When will the analysis need to be executed?* Mass file deletion will normally involve a post-event analysis. Even analysis meant to detect an in-progress mass file deletion event will only detect it after some threshold of deletions have already occurred.

*Where will the analysis be performed?* The answer to this question depends on context. Two common answers are on the endpoint with threat detection tools, or in a centralized hub or analyst workspace by manual analysis.

*Why are we doing the analysis?* The likely answer to this question is that we are performing the analysis for threat detection. However, another answer could be that we are attempting to predict the threat. For example, a tool might be able to assess whether an event that is underway, or that is about to start, fits the pattern of a mass deletion event. In such a case, the tool could attempt to interrupt the event as it is happening.

*How are we going to construct the analysis process?* If using a security application, the answer is that you will execute a defined process to determine if a mass deletion has occurred. It is possible that security applications may also use some form of learning to adjust the threshold for what defines a mass deletion. If analysts are performing the analysis, the answer may be that the process is either ad-hoc or defined.

### Sabotage via Denial of Service

Denial of service (DoS) occurs when an attacker makes an asset unavailable for use or degrades its performance. This attack is commonly executed by stopping or crashing a service, flooding an application with requests, or overwhelming a network component, such as a router, with traffic. The goal of such an attack is to cripple business processes, communications, or public facing services (such as websites).

*Who will perform the analysis?* The answer to this question is that the analysis is usually performed by a human analyst. Denial of service is often detected from user reporting. Analysts may also look for evidence of unsuccessful denial of service attacks as part of hunting activities. The answer could be that a machine performs the analysis if you are using a DoS monitoring tool.

*What form of analysis will we apply to get the end result we want?* Depending on the type of DoS attack, the answer may be that the analysis involves mapping, or both associating and mapping. If the DoS attack occurs through stopping or crashing a service, then mapping analysis is required. Flood-type attacks, however, require associating various requests or traffic flows into an event.

*When will the analysis need to be executed?* For automated DoS detection, analysis is usually implemented with the goal of detecting the attack mid-event to allow mitigation before the attack is successful. For manual DoS, analysis looks for evidence of an attack or of an attempted attack post-event.

*Where will the analysis be performed?* For automated DoS detection, analysis will usually be implemented at the endpoint or at a tool-specific hub. Manual analysis may occur anywhere.

*Why are we doing the analysis?* For automated DoS, analysis can be implemented to detect an actual threat or to predict whether an attack is occurring. For manual DoS, analysis is to detect threats or violations.

*How are we going to construct the analysis process?* The answer for commercial DoS flooding detection tools is by defined processes, often with an aspect of learning to define thresholds. Defined processes are also used for analyses used to detect a DoS attack triggered by something other than flooding that shuts down or crashes a service. The answer for analyst-executed analysis could be either defined processes or ad-hoc analyses.

# Turning Data into Observables, Indicators, and Behaviors

In an insider threat program, processes of analysis are used to describe data that represent events and behaviors, and to reason through them. These analytical fact-finding and inference-generating exercises can be used directly to answer questions, as well as to feed subsequent threat models that continuously gauge relative risk and imminent indicators of potential insider events. To describe events, indicators, and behaviors, we base our approach on Frank Greitzer's model-based classification work, which describes how to use data to infer behavioral patterns [Greitzer 2009].

## Explaining Data, Observation, Indicator, and Behavior

The model-based classification framework defines data as "directly available information." Another way to think about data is as a record of the individual characteristics of events, states, assets, etc., that are captured through some sort of monitor (human or machine). In practice, data can be almost indistinguishable from observations, which, according to the model, are "inferences from data that reflects a specific state" [Greitzer 2009, p. 7]. Many of the monitors—human and machine—process data to make observations, which are what get recorded as evidence or as artifacts available for further analysis.

For example, a monitor might see data elements such as a user requesting authentication and getting a "failed to authenticate" response. The monitor, such as Windows Event or syslog logging, would record the observation as a failed authentication. Similarly, a monitor might observe that a user inserted a portable USB drive into their computer, attempted to copy a document to the drive, and was denied because the document metadata was marked as company sensitive. The data loss prevention (DLP) tool might record the observation as a failed attempt to copy sensitive information to a portable drive.

Indicators, defined as "actions/events as evidence of precursor to inferred behavior" [Greitzer 2009, p. 7], are obtained by adding meaning to observations. In other words, they represent the conditions and actions that must occur for a behavior to arise. It may take more than one observation to infer an indicator, and the observations could be of the same or various types. In addition, the meanings assigned to the observations could be intrinsic to them or inferred.

An example of an indicator intrinsically tied to the meaning of an observation would be an authentication attempt to a forbidden service (or physical location). The meaning of the observation is intrinsic because such an action is a violation of policy. Information harvesting is an example of an indicator that is inferred from multiple observations. Actions associated with information harvesting could possibly be inferred as an indicator if the subject of an investigation was observed referring to a directory listing on several servers, followed by retrieval and printing of all documents related to a specific topic.

Note that the definition of indicators does not include a statement about purpose. That is where behavior comes in. According to the model, a behavior is a "sequence of actions associated with a purpose" [Greitzer 2009, p. 7]. Indicators are about what is happening; behaviors are about why. Behaviors are derived from a set of indicators of different types, that, taken together, can be used to infer a goal. For example, the goal of "information leak to competitor" might be inferred from a "forbidden authentication" indicator followed by an "information harvest" indicator, and, finally, by a "large attachment email to competitor" indicator. It is worth noting that indicators and behaviors are not necessarily always malicious. Indicators and behaviors can also be used to provide mitigating evidence that a usually suspicious activity is actually legitimate.

In this model, as you move from data to behavior, more and more inference comes into play. Inference is not necessary when recording data. However, inference may come into play during the transformation of data to observation. For example, a category could be assigned to a website based on characteristics found in the website's data, but the data may not be exclusive to that category and could therefore result in an incorrect categorization of the data [1]. Inference is heavily applied to obtain indicators from observations; and behavior—which determines purpose—is almost exclusively inferred. Because of this strong reliance on inference, it is important to note several caveats.

First, and perhaps most importantly, is that, because inferences are not definitive and have varying levels of accuracy, you must consider the possibility that the inference is wrong. In addition, your organization's response to the insider's activity based on the use of observations, indicators, and behaviors must be acceptable from a legal, privacy, civil liberties, ethics, and organizational perspective. You should also consider the possibility that the response itself may trigger or accelerate a malicious insider event, and you should consider adjusting the response with this possibility in mind.

Finally, for inferences to be effective, there must be supporting contextual information. For example, if you want to infer that something is in violation of a policy, first a policy must exist. However, having a policy is not enough. The policy must be detailed enough to both understand what constitutes a violation and what types and characteristics of evidence would support identification of a violation.

Another thing to note about this model is that not all use cases for insider threat analysis require transforming data all the way from observables to behaviors. Looking at observables or indicators can be enough to answer certain types of questions, such as questions about whether a certain event happened or how often it happened.

[1] An example of this is a breast cancer information website that gets labeled as adult content or pornography.

## Getting from Data to Behavior

At its most basic level, the process of going from data to behavior involves taking a set of related information from one level, applying some sort of transformation to it, and getting back more meaningful information. In the Greitzer paper mentioned above, this process is described as putting together a puzzle. In completing a set of puzzles where all the pieces are mixed up, you might first sort out the pieces into categories—e.g., parts of trees or buildings. Then, you might take those categories and piece them together into their instances. Finally, you can take all the various instances of the different categories and see which of those fit together.

This process can also be illustrated with the simple flow chart in Figure 3.

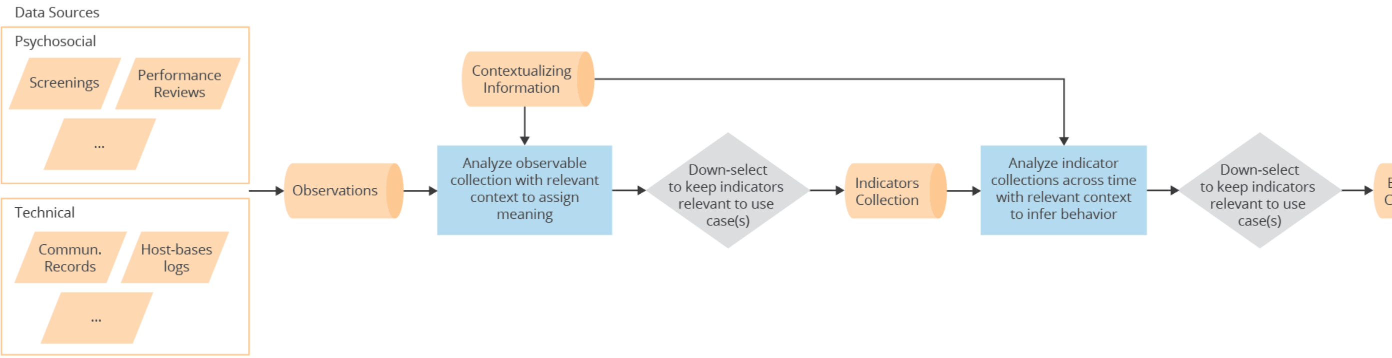
Data Sources
Psychosocial
Screenings
Performance Reviews
...
Technical
Commun. Records
Host-bases logs
...
Observations
Contextualizing Information
Analyze observable collection with relevant context to assign meaning
Down-select to keep indicators relevant to use case(s)
Indicators Collection
Analyze indicator collections across time with relevant context to infer behavior
Down-select to keep indicators relevant to use case(s)

*Figure 3: Data to Behavior Flow Chart*

## Data to Behavior as Part of a Greater Insider Threat Analysis Process

The simplified flow chart in Figure 3 illustrates how to get from data to behaviors. However, the transformation is not the goal in and of itself. Rather, obtaining behaviors is part of the larger process of insider threat risk evaluation and incident detection. As such, a slightly more detailed flow chart is useful for this discussion of insider threat analysis.

The flow chart below, broken into three parts for readability, illustrates a conceptual way of integrating data through behavioral transformations into analysis for detection and risk evaluation.

To read the flowcharts, note the following points about the image conventions:

- Peach cylinders (database symbol) represent data that is stored for further and future analysis. Data persists after one “round” of analysis, and new observables, indicators, and behaviors are added to those that already exist from previous rounds. In addition, each step of the analysis can apply data that existed previously—the process is not limited to only using “new” information. Though stored data is depicted in each step, these sets of stored data do not need to be physically or logically separate, which would result in a lot of duplicate data. In practice, each step may just update new tables or even just add new information to the processed records.

- The green rectangles represent processes that may contain multiple steps, with analysis for multiple purposes.
- The blue rectangles represent processes for a single purpose. That purpose, in this case, is the transformation between the components of the model-based classification concept.
- Peach circles represent the connections between different parts of the flow chart.

Figure 4 depicts part A of the analysis process. Part A starts with data taken from various data sources that are relevant to insider threats. The tools or people collecting the data have most likely already processed the data into observations. These observations can fall into three different classes: non-relevant, relevant, and indeterminate. You can discard non-relevant observations from the process (but you should keep them in your data storage in case you need them for something else).

Some observations are *definitive* evidence that a violation of policy occurred. An example could be an observation of an unknown device connected to the corporate network. You should escalate such evidence for investigation and any necessary mitigation. As a reminder, these may not be *insider*-perpetrated events. Rather, they could be an indication of external compromise or possibly of a gap in an internal process or policy.

*Indeterminate* events are those observations that are not known to be noise or are potential evidence of violation that you will need to process further. Note that, depending on the use case, direct, concrete evidence of a violation also may be further analyzed. For example, such evidence might point to poor security behavior.

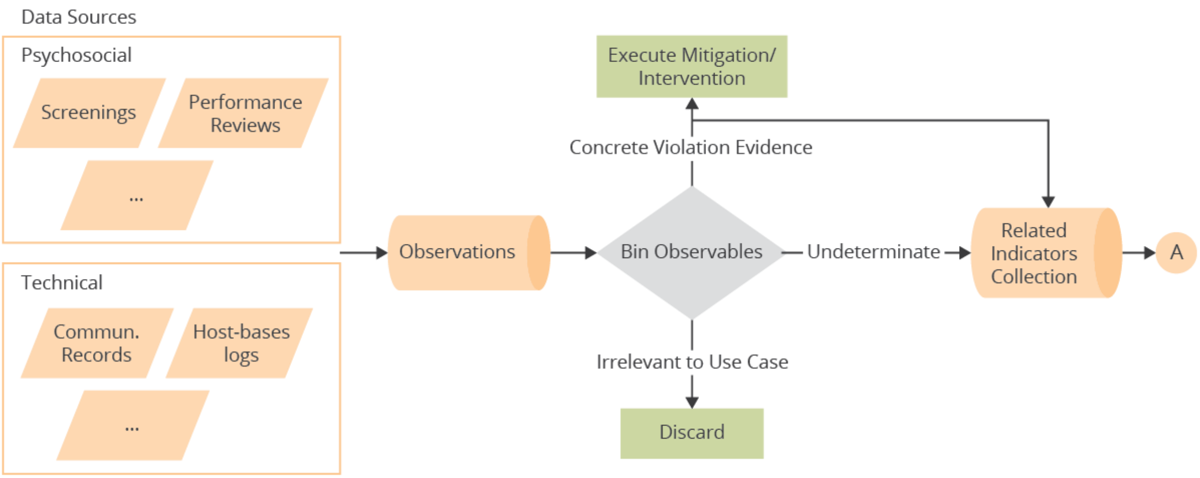

*Figure 4: Insider Threat Analysis Flow Chart, Part A*

Figure 5 depicts part B of the process. In part B, you analyze the observations in conjunction with contextual information to give each observation, or set of related observations, some sort of meaning. It is possible for an element to have multiple meanings, depending on the use cases under consideration. Note that relevant meanings do not just map to potentially malicious activities, but also to mitigating factors. Whether specific meanings will map to indicators (or, rather, make an observation or set of observations into an indicator) will depend on insider threat analysis use cases. Observations or observation sets that have meaning but do not map to indicators may be irrelevant for further analysis. In many cases, you can discard them. However, they sometimes might have the potential to be informative and can be useful for either updating or creating new contextual information.

Indicators can be definitive evidence of a violation, and when they are they should be escalated. An example of such an indicator could be the detection of PII (personally identifiable information) in an email. Again, remember that these indicators may not have been perpetrated by an insider, but you should still investigate them.

Like observations and meanings, indicators could be irrelevant to use cases of interest. However, you are likely to filter out irrelevant indicators after meaning assignment. The other indicators will be stored for further processing.

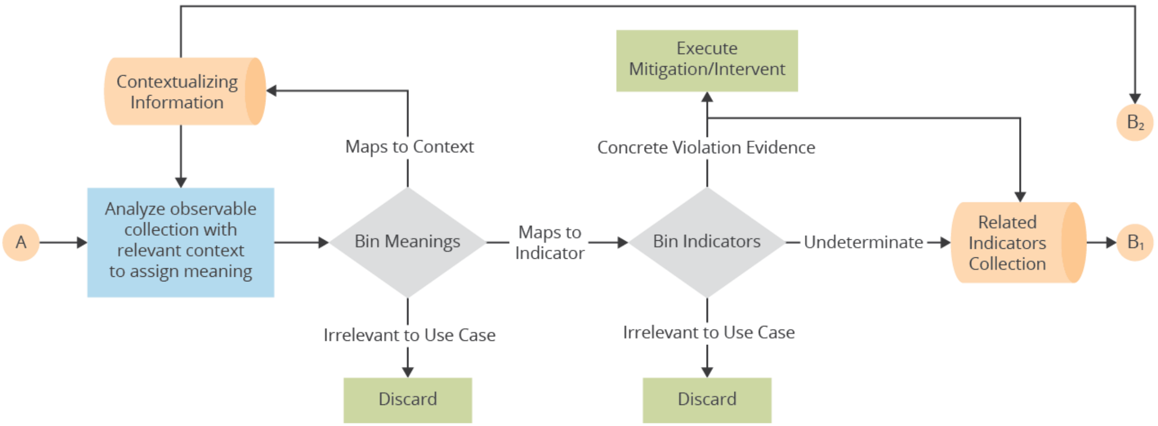
Contextualizing Information
Execute Mitigation/Intervent
B₂
Maps to Context
Concrete Violation Evidence
A
Analyze observable collection with relevant context to assign meaning
Bin Meanings
Maps to Indicator
Bin Indicators
Undeterminate
Related Indicators Collection
B₁
Irrelevant to Use Case
Irrelevant to Use Case
Discard
Discard

*Figure 5: Insider Threat Analysis Flow Chart, Part B*

Part C of the process (shown in Figure 6) illustrates how you process the indicators with contextual information to infer behaviors that are relevant to the insider threat analysis use cases. Relevant behaviors should include behaviors that infer malicious or accidental violation activity as well as behaviors that provide mitigating evidence. As with other elements of the process, behaviors you identify could be irrelevant to the use cases—and can therefore be discarded—or they might also provide definitive evidence of a violation and be escalated for investigation. Because behaviors are at least partially inferred, they are unlikely to indicate unambiguous, concrete evidence of a violation. Rather, you must assess behaviors for risk.

You can evaluate risk for either individual behaviors or for all behaviors associated with an individual. Either way, the evaluation will take context into account and will determine some sort of risk factor. If warranted by the risk factor value, organizational policy and processes will dictate that you trigger a response, such as further investigation, enhanced monitoring, employee support or counseling, or even incident response.

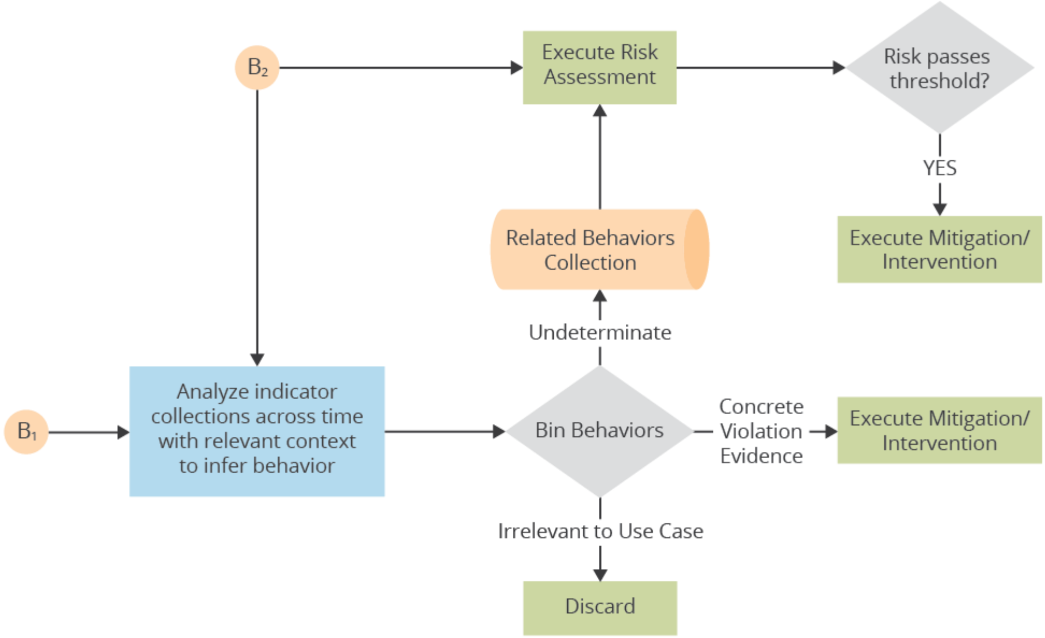
B₂
Execute Risk Assessment
Risk passes threshold?
YES
Related Behaviors Collection
Execute Mitigation/ Intervention
Undeterminate
Analyze indicator collections across time with relevant context to infer behavior
B₁
Bin Behaviors
Concrete Violation Evidence
Execute Mitigation/ Intervention
Irrelevant to Use Case
Discard

*Figure 6: Insider Threat Analysis Flow Chart, Part C*

## Analyses in the Insider Threat Analysis Process

Now that we have outlined the flow and components of the analysis process, we turn to a more in-depth discussion of analysis for the various components. For each relevant component, we provide a general discussion, and then we explore opportunities for automation and machine learning.

### Data Collection

Data collection and observables storage do not specifically involve recording information about events or states of being. As shown in the three-part flow chart above, we can think of data as being of two types: technical (i.e., that which comes from monitoring the use of technical and communication systems and devices), and non-technical (sometimes also referred to as psychosocial) [CERT 2018].

For technical data, collection and initial processing into observations are almost always automated and done by a monitoring tool, such as UAM or DLP. You can collect non-technical data with a tool, but non-technical data often has a human documentation component and may require manual conversion or entry.

**Automation:** In addition to collection and storage, automation can be utilized during the data-collection phase as part of transforming data into observables. Specifically, it can be used for preprocessing to accomplish tasks such as the following:

- transforming unstructured data to structured
- feature extraction
- data summary

**Learning:** Machine learning could potentially be used to inform data collection by aiding in determining what data fields should be collected or which data records should be kept or discarded.

### Binning

The various binning components of the flow chart would likely use multiple types of analyses. Analysis to determine if an observation, indicator, or behavior is definitive evidence of a violation that must be escalated is likely to use rule-based criteria.

**Automation:** Ideally, detection of observations, indicators, and behaviors that are direct evidence of a violation that requires immediate escalation should be automated. In practice, automation is probably easiest to implement for observations (with an UAM tool, IDS, or SIEM) and indicators (with a SIEM).

**Learning:** In theory, supervised learning could be applied to create the models of direct evidence, though doing so is likely overkill for a true direct-evidence model where the suitable evidence should be well defined. In general, common learning techniques applied here will not result in a direct-evidence model.

Determining if an observation, meaning, indicator, or behavior are noise—i.e., irrelevant to the use cases—may also involve a rule-based process. However, it’s also possible to use advanced analysis methods to determine potential utility. For example, you could assess the correlation of data sources or evaluate information gained by including specific types of data in future analysis.

**Automation:** Filtering out noise should be automated, ideally in the collection and processing systems. However, you could also implement it as a separate process executed against the stored data.

**Learning:** You can potentially use various learning techniques to identify noise in data. In the field of insider threat, it is unlikely that you could implement such a solution in a system of analysis. However, if you create the overall system with learning in mind, it is possible that you could apply machine-learning processes could to data collected from previous analyses (e.g., captured information about adjudication or usefulness of the data at the end of the process for insider threat analysis).

### Observations to Indicators

Analyses that transform observations into indicators will have either a mapping component or will include both an associating and mapping component. Both are likely to be rule based, though there may be room for learning or other advanced analytic techniques with specific use cases.

**Automation:** You can automate defined analysis processes as long as you can access all required data by the chosen automation method (e.g., a commercial tool, an in-house solution, or an analyst script).

**Learning:** You can implement machine learning (supervised or semi-supervised) to build mapping models for specific use cases.

### Indicators to Behaviors

Analyses that transform indicators to behaviors will involve both associating and mapping components. This is an area that needs much more research across all fronts of analysis.

**Automation:** Transformation of indicators to behaviors may be possible for specific use cases where the analysis is well defined and where you can access all required data by the chosen automation method.

**Learning:** You could potentially use supervised and unsupervised (for clustering) machine learning as a first pass for specific use cases. It is unlikely that any results could be definitive, but techniques might be useful for narrowing down what analysts must look at or for clustering similar behaviors together.

### Risk Evaluation

Analyses that are part of risk evaluation involve understanding how suspicious and mitigating behaviors relate; how behaviors change over time; how context influences interpretation of behaviors, changes, and associations; and how to make some sort of risk determination. As such, risk evaluation may include associating, mapping, and descriptive components. Parts of the process may be performed by a system, but there is likely also a lot of human, ad-hoc analysis that must be performed.

Because this is a complex process that requires inference that is not based on obvious, well-defined rules, this portion of insider threat analysis is where machine learning would likely provide the greatest value, both in the near and long term.

**Automation:** Automation applicability will depend on specific evaluation tasks.

**Learning:** Any potential learning will depend on specific evaluation tasks. You could potentially use supervised and unsupervised (for clustering or anomaly detection) machine learning for specific use cases.

## Considerations for Insider Threat Analysis

In this section we will describe approaches to mitigating common insider threat use cases. The threat scenarios we will cover include all of the following:

- data exfiltration or leak via cyber communication mechanisms (email, network traffic)
- data theft via accessory device (printing, removable media)
- data theft via steganography

- unauthorized account use
- unauthorized data deletion
- unauthorized modification
- insider risk analysis

For each of the points above, we will discuss data sources, scope and scale constraints, and considerations regarding techniques. Before we begin our discussion, however, we want to offer a few caveats and considerations about data sources, scope, and scale that apply to each threat scenario we outline in the sections below.

**Data Sources**: The “Common Sense Guide to Mitigating Insider Threats, Sixth Edition” contains a section that discusses data sources [CERT 2018]. Based on the advice outlined there in Best Practice 12, for insider threat analysis, analysts will need a combination of technical and non-technical data. The exact combination of data that analysts need will be determined by the use cases you are pursuing. The goal is to get a combination of data that allows analysts to understand what an insider is doing with and to assets, and why they are doing it. In most cases, it is impossible to collect information on all assets that insiders use or may use. This difficulty makes it necessary to prioritize assets to ensure that you will be able to catch and possibly prevent the most detrimental events that could happen to your organization. To have a successful analysis for an insider threat use case, however, it is not sufficient to obtain only monitoring data related to the critical assets. At a minimum, analysts must also have information that can help them determine the actual perpetrator (because actors can masquerade as other users). Analysts must also have data that can help them determine whether activities involve security problems and, if so, whether those activities are malicious or accidental. Non-technical data can help with these cases, but so can technical sources, such as communication logs.

**Scope and Scale**: The scope and scale of a given insider threat analysis has major implications on not only how the analysis can be performed, but also on whether it can be performed at all.

Scope relates to the breadth of use cases covered by an analysis. We can think of breadth in terms of the number of insiders and types of concerns that analysts must detect or mitigate. For instance, an analysis can work on a single individual at a time or on a group. It can apply for the detection of specific malicious behavior or for identifying individuals exhibiting signs of acting out against the organization, managers, or colleagues. It is important to realize that there is no single process for insider threat analysis that can cover all potential insider threat use cases, nor will we likely arrive at such a process in the foreseeable future. The better you define the scope for a particular analysis, the more likely it is to provide effective results.

Scale is a matter of the volumes of data that you must analyze. Individual users can generate thousands of events a day. This data adds up quickly over time and increases with the number of users you are monitoring. For an analysis to be successful, you must collect data, move it to the analysis location, and process it with available computing power in a reasonable time and at an acceptable monetary cost. The volume of data can have significant impact on timeliness and cost, and you must make careful plans and considerations to avoid building systems that either cannot function or are too costly to operate at full scale.

A note of caution: cloud services have the potential to mitigate some of the scale issues, such as computing power and storage, but the costs of moving data around—especially anything that you need to pull back from the cloud service or moved between services—can be significant [Siegert 2020; Shacklett 2019; Saran 2020].

## Data Exfiltration via Electronic Communications

You can detect data exfiltration through email or other electronic communication methods using direct observation or inference. Direct observation occurs when you can inspect the actual contents of the communications. Inference occurs when you can't monitor the contents of communication, but communication volumes or other patterns imply data transfer that is suspicious for some reason.

Direct observational analyses require a source of the data content. For email, you can analyze the contents from the email gateway or on the endpoint with the email client. In addition to the contents of the communication, these types of analyses require contextual information that identifies suspicious data and any mitigating circumstances.

Inference-based analyses require that metadata about communications be available. Such metadata should include information about source and destinations, communication protocols used, volumes, time stamps, and communication direction. The information could come from endpoint logs, but also from network-based monitoring tools. In addition to the communication metadata, these types of analyses require contextual information to understand the identities and purposes of the asset source, the asset destination, and service (e.g., desktop versus server and payroll processing system versus email gateway).

### Scope

In addition to the time of occurrence, the scope of electronic data exfiltration analysis identifies the person whose communications are being analyzed; determines the specific type of concerning information; and specifies the applications you need to watch. The target for analysis can be a single individual, a subset of individuals, or the whole organization. You can adjust the scope of information to specific topics or sensitivity levels. The scope related to communication mechanisms determines what applications the analysis examines, such as email, chat, FTP file transfer, DNS used as a tunnel, etc. Covering the full scope of how exfiltration can occur in an organization will likely require multiple components for the various communication mechanisms.

### Scale

Individual users can generate a lot of communications, meaning that scale quickly becomes a factor for any analytic technique that does not occur in "real time." Even in-line tools can run into scaling issues if they must handle communications for all users in an organization. You can mitigate scaling issues for out-of-band tools by adjusting the scope of analysis, usually by shortening the time frame of the volume of data processed or number of users whose data you want to process at once. Ultimately, however, organizations should ensure that their tools scale to their current needs and expected growth and that analysts have access to systems that can process and store many gigabytes, if not terabytes, of data at a time.

### Techniques

Many techniques are available for the detection of cyber communication data exfiltration. Whether you can reasonably apply them for a task depends on the scope, scale, data, and tools available.

For analyses that involve direct observation to identify explicit violations, it is most reasonable to have defined processes—preferably automated, real-time tools—that look for instances of data communications that violate policy, including communications that go to specific destinations, that have specific contents, or a combination of both. If an organization has a good set of criteria, these analyses are highly valuable and relatively affordable.

Of course, there are always complications, such as encrypted communications, dynamic addressing, or a poor static set of criteria. In such cases, most exfiltration detection comes down to some level of inference [Wojtasiak 2020; Pepper 2020; SecureCircle 2020]. Ideally, you should create defined processes to perform inference techniques and automate them where possible. They can be based on various anomaly detection methods, or may be based on some reasonable, defined criteria. Simple anomaly detection and defined criteria methods are also relatively affordable to implement. Unfortunately, they tend to provide lower value results, which often incorrectly identify normal actions as suspicious, or miss suspicious actions that should have been identified [Center for Internet Security 2021; Caltagirone 2018].

One way to mitigate some of these inference problems is to increase the contextual information for a given analysis. However, many current security tools do not support much, if any, contextualizing information. In addition, determining—and then collecting—the contextual information that would help at a large scale requires a lot of organizational and threat-related insight.

Learning techniques are another possible way to mitigate some of these issues by improving criteria to identify a greater number of legitimate problems while decreasing the number of issues incorrectly identified as suspicious. This process usually requires gathering many known examples to learn from and may also require collecting contextualizing information. Accomplishing both of these tasks and applying the expertise required to implement, or even just tune, the learning-based system usually results in a costly implementation. In addition, it is unclear how much of an improvement these techniques currently provide and how long they remain useful after implementation.

## Data Exfiltration via an Accessory Device

Data exfiltration via an accessory device occurs when information is exfiltrated off of an authorized system via an accessory device such as a printer or a removeable media device (e.g., a thumb drive or DVD). There are a number of requirements to successfully analyze possible instances of this activity. Detecting suspicious printing requires print logs, and use of removeable media requires that all endpoints are configured to log connected devices. You also need good contextual information, such as well-defined policies that explain when and why your organization allows printing and the use of removeable media. It is also helpful to understand the circumstances under which a user printed information or used removeable media. Further information about other user activity, such as file transfers and file access, is necessary to infer more about what the user printed or copied if the act of printing or copying was not, in and of itself, an outright violation of policy.

### Scope

The scope for mechanical data theft detection comprises two components: (1) the devices you are monitoring and analyzing for printing or (2) removeable media use and the associated events that provide insight into what and why a user printed or copied something to removeable media. To address the first of the two components, you must determine which devices to monitor. The second component requires more thought. It needs to list which users, which types of events (e.g., local file accesses, network communications, emails), and possibly which devices those events occurred on (i.e., you must determine whether you want to monitor or analyze only the device used for the print or copy in question, or any device used by the user within some time frame).

### Scale

With respect to scale, the analysis of printing and removeable media use is likely to be relatively small. However, analyses that try to determine what was going on around a specific print job or use of removeable media requires much more data, and the size and complexity of the data can increase rapidly.

### Techniques

Logging print jobs and removeable media usage and triggering alerts in specific circumstances is relatively simple, and possibly would require only some additional log configuration. Understanding what was going on around the alerts to determine if they represent data theft is where the analysis can get complicated. Such a task is likely to proceed in an ad-hoc manner, in part due to the variability from case to case with regards to the information that might prove valuable in learning about what the files in question are and why a user printed or copied them. However, creating defined processes for working with the different potential data sources would make the analysis more efficient and possibly allow most of it to be automated.

Current learning capabilities are unlikely to be helpful in this situation, beyond potentially detecting anomalies in printing or removable media usage. The overall tasks, criteria, and processes for evaluating the events for data theft are too variable and ill-defined to allow analysts and data scientists to create an effective solution. To train an effective model, the model engineers must know both what data is necessary and how to combine it to be processed by the learning algorithms.

## Steganographic Data Theft

Steganographic data theft occurs when a user hides sensitive or valuable information in what appears to be an innocuous file, which the user then transfers to a location where it should not be or to a person who should not have it. To detect steganographic data theft, an analyst would need to be able to find the steganographic changes in a file or infer that a user applied a steganographic tool on a file. From a practical perspective, this analysis will most likely involve monitoring systems for use of known steganographic tools, so it requires information about downloaded files, installed programs, executed scripts, and possibly running processes.

The other method for detecting steganography is to attempt to detect it from the potential steganographic file. Doing so, of course, requires having an actual copy of the file. It remains difficult to detect a steganographic file through analysis of the file, though tools and techniques do exist that can make an attempt. Because of the limits for detection through file analysis, this method is unlikely to provide much value if applied broadly, though it may be appropriate to try detection tools or techniques on certain types of files that originate from targeted locations.

Detection of steganographic tools is likely to use some sort of allow or deny list capability for downloads, installations, and executions. If there is ever a legitimate use for steganography in the organization, you can limit false alarms by restricting where the tool resides and by implementing practices to monitor and limit how employees use the tools.

## Unauthorized Account Use

Unauthorized account use refers either to a user gaining access to an account that he or she is not authorized to use, or misusing an account to which the user has legitimate access.

### Scope

Analyzing unauthorized account use requires data about user activities and the accounts in which a user conducted them. There are data sources such as system event logs, user activity monitoring alerts, and application audit logs that you can use for this analysis. Additionally, security logs, employee time logs, information about time off and normal work hours, and biometric measurements (e.g., typing speed and cadence, mouse usage) can help infer whether the account user is the legitimate owner of the account or an imposter.

### Scale

You should monitor all accounts regularly for masquerading attempts and policy violations, which you cannot not accomplish using ad-hoc analyses unless the group you are monitoring is very small. Even for very small groups, you cannot perform ad-hoc analyses on a continuous basis because doing so increases the likelihood that you will miss events or not be able to detect them quickly enough to fully mitigate them. Consequently, you should automate these analyses with emphasis on monitoring high-value (e.g., administrator) accounts.

### Techniques

Unauthorized account use requires different techniques for different aspects of the problem. You can detect certain unauthorized access simply by looking for the occurrence of certain events in an event log, such as attempted access to forbidden resources. However, most tasks require more inference to determine activities that could be unauthorized or the result of masquerading. Inference is based on looking for events that are out of the ordinary for a given user or account or for items that contain characteristics that are known to be suspicious.

You can use various methods to identify events that need further analysis. Rule-based analyses will look for events or sets of events that meet specified criteria, such as non-maintenance related account activity outside of business hours or several failed authentication attempts for an account. Deviation-type anomaly methods look for significant changes from a baseline or average behavior and are often used to find abnormal volumetric activity, such as suspicious copying of large amounts of data. Novelty-type methods look for activity that has not been seen before, such as an account connecting to or being used on a new system.

All of these approaches can be defined directly by analysts. They also have the potential of being learned using various machine learning techniques. Learning rule-based models requires that the training data consist of records where the type of event corresponding to the record is known—in the terminology of data science, the data must be labeled. When building the rule-based system in-house, getting this data can be a problem, as labeling data is time consuming and requires that the labeler has sufficient information to know what labels to assign to each record. Consequently, labeling data is expensive. Some machine-learning methods, such as semi-supervised [Algorithmia 2020], reinforcement learning [Ribeiro 2020], and few shot learning [Ozsubasi 2020], have the potential to mitigate the labeling issues, but the trade-off is that they have other prerequisites, such as additional expert-knowledge requirements or pretrained models.

Anomaly-based models do not require labeled data because their goal is to find activities that differ from other activities in some manner, and not to find activities that correspond to a specific category. While machine-learning-based anomaly detection has the potential to find new and different types of anomalies that other methods might not find, these methods currently have the same problems as anomaly detection in general: they find an arbitrary number of anomalies (depending on how they are tuned), most of which are not relevant. Often, even if they are tuned to find a greater number of relevant anomalies, they also miss a lot of anomalies that should be detected [Salinas 2020; Perry 2020].Better incorporation of organizational context, as well as information about how system processes work, can relieve these issues. However, incorporating context remains a difficult task and analyses currently seldom have much, if any, contextual support.

## Unauthorized Data Deletion or Modification

Unauthorized data deletion or modification occurs when someone with no legitimate access or ownership deletes or modifies data from a file, database, etc. It can also occur when a data owner deletes or modifies data with malicious intent, such as to circumvent data discovery or to corrupt a document or system.

### Scope

To detect unauthorized data deletion or modification, analysts need information about what data should exist on a system, whether it does, and, if not, who deleted or modified it. In practice, information about deletions usually come from event or audit logs that capture deletion and modification events for the data of interest. Other potential data that could be useful in some circumstances are lists of expected directory contents, current and reference file hashes, trash contents, or system backups. However, unless an unauthorized user succeeded in deleting data, none of these sources provide direct insight into whether a deletion was legitimate or not.

While the listed data can potentially be used to infer suspicious patterns—such as a large number of deletions that occur in a short period of time or a user attempting to wipe a database table or full directory— that suggest a deletion is unauthorized, analysts will have difficulty detecting more subtle malicious acts using such data. It is not feasible to provide constant awareness for all organizational data, so organizations should ensure that they have robust backup procedures. Organizations should also ensure that they have notified analysts about any critical data that should raise a red flag if deleted. For this type of data, a change control system can help ensure that analyses can quickly identify unauthorized deletions.

### Scale

Monitoring all data for modifications and deletions is not likely to be cost effect or to provide a very good return on investment due to the sheer volume and types of data that exist and how often it changes or users delete it legitimately. Consequently, organizations must prioritize what they monitor and analyze for unauthorized changes.

### Techniques

A security tool should perform the initial analysis of deletions or modifications to flag specific instances that match given criteria (e.g., a modification of a file in a system directory on a desktop computer), that match a suspicious pattern (e.g., mass deletion of files on a server), or that are otherwise anomalous (e.g., modifications performed outside of a user's work hours). You can configure these analyses using explicit criteria defined by analysts or policy makers, or you could use statistical or machine-learning methods that identify anomalies. The standard caveat about anomaly detection of either type—standard or machine learning—applies here: anomalies do not imply malicious intent, and they are not, in fact, usually malicious. Contextualizing information, such as information from a change management system, can help filter anomalies, but incorporating contextual information into the automated analyses can be difficult. This difficulty is due both to tool limitations, as well as to problems with capturing the contextual information to begin with.

## Insider Risk Analysis

Insider risk analyses are different from the preceding use cases because they are not about finding or investigating insider events. Rather, insider risk analyses are processes that determine the level of risk an individual poses of becoming an insider threat. You can conduct these analyses on periodic or ongoing bases. They include activities like pre-employment background checks, security clearance investigations, ongoing insider risk monitoring, and continuous evaluation for security clearances [ODNI 2021].

### Scope and Scale

You can conduct insider-threat analyses using only psycho-social data or using a combination of psycho-social and technical data, depending on the time and purpose of the analysis. The information used will vary based on the goal, as well as legal requirements, and privacy and ethics considerations. Generally, information is needed to provide insight into potential motivation, capability, or opportunity.

The following are examples of information used to infer potential motivation:

- financial records to indicate if someone might be motivated by financial gain
- performance reports or formal complaints to indicate disgruntlement

- social media activity to infer possible hacktivism tendencies
- social contacts to infer possible espionage

Examples of information used to infer capability include the following:

- permissions and accesses granted on systems and to other assets, like buildings or rooms
- software installed on they systems the insider uses

Examples of information used to infer opportunity include the following:

- intrusion detection or user activity monitoring alerts
- communication logs
- activity logs
- internal reporting

When looking at information to help determine insider risk, it is important to remember to include not only information that can reflect adversely on individuals, but information that negates or explains away aspects of risk.

Precisely defining scope is important for any analysis, but it is especially important for analyses designed to determine insider risk. The scope for insider risk must consider not only technical needs and ability, but legal requirements and privacy implications. Because of the psycho-social data required for these analyses, they are intrusive beyond just monitoring work activities and looking for actual policy violations. Organizations must consult with various experts, including their legal counsel, to set policies about the allowable uses for the results, as well as what data sources and information are permissible to include in the analysis.

### Techniques

Implementing insider risk analysis is complicated. While you might have clearly defined thresholds and criteria to examine, there is usually also a human judgment aspect as well. This means that the analyses may completely involve ad-hoc processes or a combination of defined processes—some of which may be automated—and ad-hoc analyses. Such a mix makes sense because it is currently impossible to enumerate everything that can reflect poorly on risk or to mitigate something that would normally increase insider risk.

When insider risk determination is a regular part of an insider threat program, organizations should work to define and automate as much of the analyses as possible. Many existing threat defense products, such as SIEMs, have capabilities that can support this process. However, analysts should always thoughtfully consider the results of those analyses and whether or not there is information that was not considered that would change a determination.

Application of learning-type methods may be a reasonable goal for pieces of insider risk analysis. All the caveats and requirements for machine learning that apply to the other topics also apply here, but you must take extra care due to the social aspect (e.g., potential impact to reputation of subject, analyst, and organization; possible discrimination—deliberate or accidental) of these analyses. Furthermore, we caution strongly about learning systems that assign overall risk to people. Machine learning has well-documented problems with learning and perpetuating existing social biases (e.g., racial, gender, religious, economic class) [AJL 2021]. These biases can arise in learned models, even when protected characteristics are removed from the data.[2] Furthermore, the information used to train the models may result in spurious correlations or in models that do not reflect reality, regardless of whether the training data is chosen carefully or poorly [D'Amour 2020].

Identifying these problems can be difficult and requires careful attention to collecting, selecting, and cleaning the data used for testing and evaluation. It also requires robust testing of results and conscious probing of the system for problematic biases. As these systems' models are periodically, if not continually, updated, vigilance must be ongoing.

## Effectiveness

It is good business practice to evaluate the effectiveness and utility of processes and tools. With insider risk analyses, robust evaluation tends to be difficult. Part of the reason for this difficulty is the subjectivity and context dependencies of the topic. Another, and arguably larger, reason is that insider risk analyses deal to a large extent with known unknowns (e.g., what factors actually correspond to a risk and how much; what incidents might insiders actually trigger) and unknown unknowns (e.g., threats we cannot currently imagine). The best way to evaluate these analyses and, by extension, the tools that perform them is open for debate and an active area for research.

Evaluation of effectiveness begins with planning for implementation of the new analysis and continues throughout its whole lifecycle. For purposes of this framework, we will explain evaluation of effectiveness as having four components.

1. **Purpose definition**: identifying why the analysis is needed, what you expect it to provide, and performance constraints.
2. **Side-effect identification**: enumerating unintended consequences (both potential and unintended) from the analysis—either directly or from how the results are used.
3. **Utility assessment**: identifying the benefits of the analysis, both intrinsically and within the context of the organization and in relation to other available analyses.
4. **Cost analysis**: identifying and evaluating the direct and indirect costs that arise from all aspects of the analysis, including implementing, running and maintaining, and utilizing results.

[2] One reason this can happen is because a single characteristic, such as gender or race, may be strongly correlated with some other field in the data, such as college attendance (If you attended a women's college, you are likely female; if you attended a historically black university, you are likely non-white.) or degree (82% of computer science degree recipients are men) [ComputerScience.org 2020].

Each of these should form part of the planning phase for major analysis development projects or as part of an acquisition process. You should update each of the components periodically with new insights and needs to ensure that the analysis continues to meet expectations and that benefits continue to outweigh costs.

## Purpose Definition

The idea that an analysis does what it is supposed to implies that a concrete purpose has been defined, along with factors of "success" that can be evaluated. Defining purpose and success factors are perhaps the easiest part about evaluating effectiveness, but it is only easy when the goals are well understood, then well defined.

To elaborate, in general business terms, mitigating potential threats and detecting actual threats are not goals so much as vision or purpose statements. To measure effectiveness, we need more concrete expectations for an analysis, which considers a use case, context, and desired result. In a sense, these actions are akin to goal setting for an analysis. As such, we can apply the concept of SMART goals for the insider threat program and individual analyses.

SMART goals are specific, measurable, achievable, relevant, and time-based [Boogaard 2020; CDC 2021]. Analysts often have SMART goals unconsciously in mind when they execute ad-hoc analyses. Below are short examples of how an analyst might think through each item of the SMART framework.

- **Specific**: Analysts have a scope and context in mind and usually try to answer a specific question, such as, "Is there evidence of exfiltration from this set of machines?"
- **Measurable**: Analysts have criteria about what they need to look for and the type of information that is required to answer their question.
- **Achievable**: Analysts adapt their methods to the tools and data they have available and know the limitations of their analysis capabilities and adjust their expectations for and interpretations of the results accordingly.
- **Relevant**: Analysts' questions are usually tied to a specific ask or a task of their job function.
- **Time-based**: Analysts understand the urgency of and time limitations for answering their questions, and they adjust their analytic techniques, breadth of inquiry, and scope of analysis as needed to meet deadlines and to be productive in their overall job responsibilities.

Defining SMART goals for insider threat programs can be more difficult than for other types of programs because insider threat processes can be more complex and often involve myriad analyses. To define useful goals, you should outline not only the analyses used in your program, but also the people, policies, and materials involved in executing it, including employees, events, mitigations, regulations, privacy, ethics, and more.

It is also important to mention that it is often more difficult to define goals for analytic techniques that are performed with commercial tools or developed by a development team than those performed by analysts. For techniques performed by tools or developed by development teams, organizations must consider both the "ideal" goals—what the analysts actually need to accomplish—and the "is possible" goals—what the tool is capable of helping the analysts accomplish.

## Side-effect Identification

Implementing an analysis produces effects, both intended and unintended. We think of the intended effects as being benefits, for example, improved capability and decreased incident risk. Unintended effects are costs (usually non-monetary, though they can incur monetary costs as well) and increased risk. Some unintended effects are obvious after implementation, such as having to vet large amounts of irrelevant or incorrect results (i.e., "false positives") or managing resource monopolization (e.g., memory or computing power) by the analysis process. More insidious consequences may be harder to identify. These include analyses that are

- biased against certain protected characteristics, like race, gender, age, or religion.
- misinterpreted or that produce results that encourage inappropriate actions or responses, such as confrontation or unwarranted escalation.
- prioritized incorrectly and take up time and resources that should be spent on higher priorities.

All side effects should feed into the cost analysis. In addition, you must mitigate any potential negative side effects as much as possible, and you must actively accept the remaining effects.

To mitigate the negative consequences, you must first detect them. Doing so is not easy.

How to detect bias is a question that is currently under investigation by a broad amount of research, but all of this research has not yet been successful in completely solving the problem. One common method for detecting bias is to compare statistical accuracy measures that are calculated on the various subgroups of the protected characteristic. For example, some studies have attempted to mitigate gender bias by calculating whether analyses provide information more accurately when the subjects being analyzed are male as opposed to female. In the field of insider threat, this means that detection and risk analysis could be evaluated for accuracy separately on subgroups (male and female or majority and minority) to assure that the rules or criteria apply equally. If there is a statistically significant difference, the implication is that there is a bias somewhere in the process—either in the data used to build the model, in the criteria encoded in the analysis process, or with the analyst's application of an analysis. If bias exists, the analysis could be considering gender somewhere it should not, or it might not be including it somewhere it should.[3]

[3] Context determines whether the inclusion or exclusion of a characteristic creates bias. To cite a non-insider threat example, much research has shown that not accounting for gender (and including both males and females) in health research results in men having better outcomes for certain conditions than women because the research assumed that men and women would respond similarly to the same treatment [Paulsen 2020]. In other cases, bias can also occur when gender is included in an analysis. For instance, one study showed that gender in advertising models resulted in women being shown fewer high-paying postings then their male counterparts with similar backgrounds [Gibbs 2015].

This method of detecting bias is not sufficient in all circumstances. For example, cases where "accuracy" itself is hard to measure makes the detection of bias difficult. In such cases, critical evaluation of all steps and assumptions in an analysis process is currently the best method to detect bias.[4]

A complication with interpreting results is that the meaning of the results is not always straightforward. Appropriate interpretation of results requires critical evaluation of steps and assumptions. For insider threat programs that apply commercial products, it may be impossible to evaluate the steps and assumptions the products use. It is therefore important that analysts have sufficient training on the use of the product to understand how it is meant to be used and its limitations. If the product's use and limitations are not well defined, you should request further documentation or clarification from the vendor.

Evaluating how well aligned your organization is with its priorities requires that you have clearly identified your priorities and that you have accurately mapped analyses and other processes to them. A matrix with the analysis and process components on one axis and the insider threat program scenarios in order of priority on the other axis is one way to accomplish the mapping and will help ensure that all scenarios are covered. Tracking effort by priority helps ensure that it is aligned appropriately with the priority of the scenarios. While it is not the case that the highest priority items should necessarily get the most effort, tracking work can help ensure that all priorities receive attention. Tracking effort could mean recording actual time spent on a priority, but it can also be as simple as listing what an analyst worked on in a day or week.

## Utility Assessment

Organizations do not, or at least should not, implement analyses for their own sake. Rather, the purpose of implementing an analysis is to fulfill a need and to provide a benefit to the organization. In economics, such benefits correlate to the concept of utility.[5] Utility refers to the overall usefulness and value of a product or service, both alone and in relation to other available products and services. Utility can take any of the following four types: form,[6] time, place, and possession [Lewis 2015].

Each of the four types of utility can be defined as follows:

- **Form utility**: refers to the characteristics of a product or service that give it value. For analytic products, this type of utility could include benefits such as capacity, speed, and structure and content of output. The ability to scale to a needed capacity, integrate with other existing tools, execute in a timely fashion with available computing power, and provide human consumable output are examples of form factors that help determine the utility of products or processes for insider threat analysis.

[4] The "Towards Data Science" blog provides an introduction to this concept in the post "How to Detect Bias in AI" [Grootendorst 2020]. Research in bias and ethics for predictive policing is directly relevant here [Crawford 2019].

[5] This nuanced view of utility is often discussed in marketing and behavioral economics.

[6] Sometimes, as in the paper on utility by Lewis [Lewis 2015], form is broken into form and task utility.

- **Time utility**: refers to the availability of a product or service when a customer needs it. Maintenance needs, licensing availability, and availability of required data are some of the factors related to the utility of time.
- **Place utility**: refers to the context, not necessarily the geography, where a product or service has value. For insider threat analytic products and processes, place could refer to contexts like organizational structure dependencies or to whether an employee has a mental disability. Place can also refer to the product and service ecosystem where the new product or service will be applied. The existence of other products or services with similar capabilities decreases the utility of a new one.
- **Possession utility**: refers to the ability and time span of actually having a product or service and being able to use it. Licensing can play a role here too, as can things like the ability to invoke and allow time-consuming processes to run.

Because utility is a subjective concept, there is no direct way to measure it. Economists evaluate it by discussing utility either as cardinal or ordinal. Economists score items they deem to be cardinal using numeric values. While, overall, this approach is still subjective—because different people may assign values differently—utility could be assigned value based on criteria set by an individual or organization. Assigning value in this way provides the opportunity to compare which items provide more utility, and allows people to judge how much more or less utility an item provides. When organizations have a list of weighted requirements that they use to compare products, they are applying this type of utility to decide.

Ordinal utility does not assign a numeric value, but rather a rank or ordering between different options. It is used for comparing options based on less concrete criteria. This approach can be sufficient for comparing items where the choice comes down to personal preference, such as the scripting or query language a product uses. When people use ranked choice voting to choose an option, they are being asked to assign this type of utility to the options.

## Cost Analysis

When analyzing the cost of a system, it is necessary to consider the direct financial costs, as well as indirect financial costs and costs that can be hard to assign a monetary value. Common direct costs for insider threat analysis development or acquisition include

- the monetary cost of hardware, software, and communication infrastructure required to collect data and execute the analysis. These costs need to account for acquisition, maintenance, and possibly decommissioning.
- the effort of the insider threat analyst for data pre-processing, execution, and interpretation.
- the effort of the business risk analyst to adjust perception of risk based on data collected and the analyses and processes utilized.

Examples of indirect financial costs that could arise include increased cloud service costs, as well as training needs and the time needed for analysts, developers, and maintainers to learn how to effectively use the new technology and techniques. For both direct and indirect financial costs, you should take care to identify which of the costs are fixed and which are variable. As part of the cost analysis, an organization should consider best case, worst case, and most-likely case cost estimates.

Costs that can be hard to assign a monetary value include

- the potential degradation to organizational, managerial, or analyst reputation.
- emotional effect on employees.

You should not evaluate the cost of a product on an individual basis. You should consider the cost of a product in comparison with similar products and in relation to the overall program budget. The cost for a single analytic component or product may seem reasonable, but if you need multiple products to achieve a goal, the combined cost may not be. In addition, it is important to not only weigh the cost with respect to the budget, but in relation to the utility provided—in other words, you should conduct a cost-benefit analysis.

Actual cost measurement involves estimating expected costs and tracking actual costs. To conduct estimations, you can

- estimate the cost of systems using life-cycle cost methods [Wall Street Mojo 2021.
- estimate insider threat and business risk analysts' effort using effort-estimating methods commonly used for project management [Dummies 2021].
- estimate reputation and emotional impact using enterprise risk management methods [Sickler 2019] or surveys that gauge organizational engagement or sentiment.

To track actual costs, you should periodically compare the costs your organization incurs to the estimates you made. For monetary costs, this comparison is as simple as comparing the values of what you spent with the values you estimated. You should evaluate costs that are significantly over expectations (as defined by the organization) to determine whether the additional cost is acceptable for the continued use of the resource. Organizations should be careful to not fall into the sunken cost fallacy during this evaluation [Mohammed 2019].

Measuring non-monetary costs is a more difficult task. While updating risk estimates, surveys or other engagement and sentiment analysis can provide information into how things like reputation are changing. However, attributing those changes to a specific cause is difficult. Formal or informal sentiment analysis is currently the most likely method to provide the necessary insight into these costs. While formal sentiment analysis or engagement surveys can be valuable, managers and HR should also regularly be engaging in discussion with employees about how they believe things are going in the organization and with their work. From an external reputation perspective, organizations should monitor any press coverage and jobsite feedback on the company, at minimum.

## Open Problems in Efficacy of Insider Threat Analysis

So far, we have only discussed individual aspects of measuring the efficacy of an analysis process or product. For these aspects, there are open questions about measuring the accuracy and utility of analytical results, as well as for accounting for cost, especially for those analyses and products whose monetary values are difficult to estimate. In addition, developing formally defined processes could help in holistically accounting for all aspects of effectiveness in decision making.

## Conclusion

Analysis capability is one of the defining factors of an insider threat program. In this paper, we discussed a view of analysis specific to insider threat, with the goal of helping managers of analysts better understand their organization's capability.

There are many other ways to think about analysis that could be helpful for insider threat teams that this paper does not cover. For methods geared towards conducting analysis, we recommend learning more about intelligence analysis. Specifically, we recommend that analysts and managers both read "The Psychology of Intelligence Analysis" [Heuer 1999] and "A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis" [CIA 2009].

## References

**[Algorithmia 2020]**

Algorithmia. Semi-supervised learning. *Algorithmia*. August 11, 2020. https://algorithmia.com/blog/semi-supervised-learning

**[AJL 2021]**

Our Mission. *Algorithmic Justice League*. February 24, 2021 [accessed]. https://www.ajl.org/about

**[Boogaard 2020]**

Boogaard, Kat. How to write SMART goals. *Atlassian Work Life*. November 15, 2020. https://www.atlassian.com/blog/productivity/how-to-write-smart-goals

**[Caltagirone 2018]**

Caltagirone, Sergio & Lee, Robert M. The Four Types of Threat Detection With Case-Studies in Industrial Control Systems (ICS). *Dragos*. July 13, 2018. https://www.dragos.com/resource/the-four-types-of-threat-detection-with-case-studies-in-industrial-control-systems-ics/

**[CDC 2021]**

Develop SMART Objectives. *Centers for Disease Control and Prevention*. February 15, 2021 [accessed]. https://www.cdc.gov/phcommunities/resourcekit/evaluate/smart_objectives.html

**[Center for Internet Security 2021]**

Cybersecurity Spotlight – Signature-Based vs Anomaly-Based Detection. *Center for Internet Security*. February 15, 2021 [accessed]. https://www.cisecurity.org/spotlight/cybersecurity-spotlight-signature-based-vs-anomaly-based-detection/

**[CERT 2018]**

CERT National Insider Threat Center. *Common Sense Guide to Mitigating Insider Threats, Sixth Edition*. CMU/SEI-2018-TR-010. Software Engineering Institute, Carnegie Mellon University. 2018. https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=540644

**[CIA 2009]**

Central Intelligence Agency. *A Tradecraft Primer: Structured Analytic Techniques for Improving Intelligence Analysis*. U.S. Central Intelligence Agency, Center for the Study of Intelligence. 2009.

**[ComputerScience.org 2020]**

ComputerScience.org. Women in Computer Science: Getting Involved in STEM. *ComputerScience.org*. November 23, 2020. https://www.computerscience.org/resources/women-in-computer-science/

**[Crawford 2019]**

Crawford, Kate; Dobbe, Roel; Dryer, Theodora; Fried, Genevieve; Green, Ben; Kaziunas, Elizabeth; Kak, Amba; Mathur, Varoon; McElroy, Erin; Sánchez, Andrea Nill; Raji, Deborah; Rankin, Joy Lisi; Richardson, Rashida; Schultz, Jason; West, Sarah Myers; & Whittaker, Meredith. *AI Now 2019 Report*. AI Now Institute, New York University. 2019. https://ainowinstitute.org/AI_Now_2019_Report.html

**[D'Amour 2020]**

D'Amour, Alexander; Heller, Katherine; Moldovan, Dan; Adlam, Ben; Alipanahi, Babak; Beutel, Alex; Chen, Christina ; Deaton, Jonathan; Eisenstein, Jacob; Hoffman, Matthew D.; Hormozdiar, Farhad; Houlsby, Neil; Hou, Shaobo; Jerfel, Ghassen; Karthikesalingam, Alan; Lucic, Mario; Ma, Yian; McLean, Cory; Mincu, Diana; Mitani, Akinori; Montanari, Andrea; Nado, Zachary; Natarajan, Vivek; Nielson, Christopher; Osborne, Thomas F.; Raman, Rajiv; Ramasamy, Kim; Sayres, Rory; Schrouff, Jessica; Seneviratne, Martin; Sequeira, Shannon; Suresh, Harini; Veitch, Victor; Vladymyrov, Max; Wang, Xuezhi; Webster, Kellie; Yadlowsky, Steve; Yun, Taedong; Zhai, Xiaohua; & Sculley. D. "Underspecification Presents Challenges for Credibility in Modern Machine Learning." *arXiv*. November 2020. https://arxiv.org/pdf/2011.03395.pdf

**[Dummies 2021]**

Dummies. Estimating Required Work Effort. *Dummies a Wiley Board*. February 15, 2015 [accessed]. https://www.dummies.com/careers/project-management/estimating-required-work-effort/

**[Gibbs 2015]**

Gibbs, Samuel. Women less likely to be shown ads for high-paid jobs on Google, study shows. *The Guardian*. July 8, 2015. https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

**[Greitzer 2009]**

Greitzer, Frank L.; Paulson, Patrick R.; Kangas, Lars J.; Franklin, Lyndsey R.; Edgar, Thomas W.; & Frincke, Deborah A. *Predictive Modeling for Insider Threat Mitigation*. United States Department of Energy. April 2009.

**[Grootendorst 2020]**

Grootendorst, Maarten. How to Detect Bias in AI: Detecting common (cognitive) biases in your data. *Towards Data Science*. January 2020. https://towardsdatascience.com/how-to-detect-bias-in-ai-872d04ce4efd

**[Heuer 1999]**

Heuer, Jr., Richards J. *The Psychology of Intelligence Analysis*. Pherson Associates, LLC. 2007. ISBN 978-0979888007.

**[Lewis 2015]**

Lewis, B.R. Utility. *Wiley Encyclopedia of Management*. Volume 9, Marketing. January 21, 2015. https://doi.org/10.1002/9781118785317.weom090695

**[Mohammed 2019]**

Mohammed, Shah. Sunk Cost Bias, Example & Decision-Making in Business. *Medium.* March 19, 2019. https://shahmm.medium.com/sunk-cost-bias-decision-making-in-business-and-intels-exit-to-microprocessors-ba0b99a69991

**[ODNI 2021]**

Continuous Evaluation – Overview. *Office of the Director of National Intelligence*. February 15, 2021 [accessed]. https://www.dni.gov/index.php/ncsc-how-we-work/ncsc-security-executive-agent/ncsc-continuous-evaluation-overview

**[Ozsubasi 2020]**

Ozsubasi, Izgi Arda. Few-Shot Learning (FSL): What it is & its Applications. *AI Multiple*. November 5, 2020. https://research.aimultiple.com/few-shot-learning/

**[Paulsen 2020]**

Paulsen, Emily. Recognizing, Addressing Unintended Gender Bias in Patient Care. *Duke Health*. January 14, 2020. https://physicians.dukehealth.org/articles/recognizing-addressing-unintended-gender-bias-patient-care

**[Pepper 2020]**

Pepper, Tony. DLP has failed you – and here’s what you need to do now. *Egress*. July 24, 2020. https://www.egress.com/en-us/blog/dlp-has-failed-you

**[Perry 2020]**

Perry, Christopher. 4 Machine Learning Challenges for Threat Detection. *InformationWeek*. May 4, 2020. https://www.informationweek.com/strategic-cio/security-and-risk-strategy/4-machine-learning-challenges-for-threat-detection/a/d-id/1337639?

**[Ribeiro 2020]**

Ribeiro, Jair. Reinforcement Learning and 9 examples of what you can do with it. *Towards Data Science*. October 23, 2020. https://towardsdatascience.com/about-reinforcement-learning-2ff0dafe9b75

**[Salinas 2020]**

Salinas, Stephen. How Traditional Machine Learning Is Holding Cybersecurity Back. *Info Security*. August 3, 2020. https://www.infosecurity-magazine.com/opinions/traditional-machine-learning/

**[Saran 2020]**

Saran, Cliff. Hidden cost of cloud puts brakes on migration projects. *ComputerWeekly.com*. February 27, 2020. https://www.computerweekly.com/news/252479237/Hidden-cost-of-cloud-puts-brakes-on-migration-projects

**[SecureCircle 2020]**

SecureCircle. Why Isn't DLP Preventing Data Breaches and Data Leakage? *SecureCircle*. January 17, 2020. https://www.securecircle.com/blog/why-isnt-dlp-preventing-daily-data-breaches

**[Shacklett 2019]**

Shacklett, Mary. 6 tips for controlling cloud costs. *Tech Republic*. June 7, 2019. https://www.techrepublic.com/article/6-tips-for-controlling-cloud-costs/

**[Sickler 2019]**

Sickler, Jonas. What is Reputational Risk and How to Manage it. *ReputationManagement.com*. February 8, 2019. https://www.reputationmanagement.com/blog/reputational-risk/

**[Siegert 2020]**

Siegert, Zack. Common Culprits for Unexpected AWS, Azure, and GCP Service Cost Spikes. *CloudHealth*. September 24, 2020. https://www.cloudhealthtech.com/blog/common-culprits-unexpected-aws-azure-and-gcp-service-cost-spikes

**[Wall Street Mojo 2021]**

Life-cycle Cost Analysis. *Wall Street Mojo*. February 15, 2021 [accessed]. https://www.wallstreetmojo.com/life-cycle-cost-analysis/

**[Wojtasiak 2020]**

Wojtasiak, Mark. Security Product Versus Practitioner – Something's Gotta Give. *Code42*. June 9, 2020. https://www.code42.com/blog/security-product-versus-practitioner-somethings-gotta-give/

## Contact Us

Software Engineering Institute
4500 Fifth Avenue, Pittsburgh, PA 15213-2612

**Phone**: 412/268.5800 | 888.201.4479
**Web**: www.sei.cmu.edu
**Email**: info@sei.cmu.edu

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

DM21-0066