Combining Learning and Model-Based Reasoning to Reduce Uncertainties in Cloud Security and Compliance Auditing

Uttam Thakore
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
thakore1@illinois.edu

Rohit Ranchal
IBM
Cambridge, MA
ranchal@us.ibm.com

Yi-Hsiu Wei and Harigovind V. Ramasamy
IBM
Austin, TX 78758
{*ywei, hvramasa*}*@us.ibm.com*

***Abstract*— Security and compliance auditing is expensive, time-consuming, and error-prone for cloud service providers operating in multiple domains. Existing approaches predominantly use formal logic and domain-specific languages to facilitate collection and validation of evidence needed for compliance certification. Such approaches do not sufficiently account for the uncertainties and challenges caused by human involvement, which are a major contributor to inefficiencies and mistakes in the audit process. We propose that *hybrid* approaches, in which formal, model-based approaches are combined with machine learning techniques to reason about evidence and historical audit data, are necessary to address such uncertainties. Such approaches can help both auditors and service providers better deal with uncertainties and reduce costs, errors, and the manual effort required to identify evidence needed for compliance certification. We present a taxonomic framework for understanding the causes of and potential solutions to uncertainty in the audit process. We identify areas within evidence collection and validation in which machine learning can augment model-based techniques to reduce uncertainties. We provide some examples of hybrid approaches that we are exploring and discuss the need for more work in this area.**

***Keywords*-cloud computing, security, compliance, audit, machine learning, formal models**

I. Introduction

As the use of computing has expanded into sensitive domains (e.g., healthcare, finance, military, and critical infrastructure), the need for security assurance of the underlying information technology (IT) infrastructures has increased. One way enterprises and IT service providers provide such assurances is through certifications of compliance with mandatory regional or sector-specific regulations and optional global industrial and commercial standards. However, with the emergence of numerous, unique IT security standards in different domains, compliance auditing has become increasingly expensive, time-consuming, and error-prone for service providers that operate in multiple domains, such as cloud service providers (CSPs).

Popular CSPs such as Amazon Web Services (AWS), Google Cloud, IBM Cloud, and Microsoft Azure, are currently certified with tens of certifications [1]–[4]. These certifications require periodic renewals, so cloud providers need to go through several audits every year. Failure to achieve certification for a standard may have legal and financial consequences for a CSP and its customers.

The compliance audit process remains expensive for service providers that support clients in multiple industry verticals because it involves many technical and nontechnical challenges. On the technical side, the challenges include:

- The heterogeneity of data demanded by auditors as acceptable evidence of compliance even within a given regulation, not to mention, across diverse standards,
- The scale of evidence collection and validation,
- Limited time availability for evidence collection, and
- Varying levels of automated evidence collection supported by systems from different vendors, which run the gamut from manual screenshots of command outputs to full API support for automated queries.

On the nontechnical side, the challenges arise from human factors. Uncertainties and unpredictability in the audit process stem from the high level of subjectivity involved in interpreting the requirements for a given compliance control. That may lead to an incomplete or incorrect understanding of the evidence needed and the type of validation that would be acceptable. Those factors are compounded by the reality that auditors have varying levels of skill and experience even within the same auditing company. The sheer number of controls (even in a single standard like FedRAMP [5]) and the amount of evidence to sift through make the audit process prone to human error both on the provider side and the auditor side.

Despite the challenges and inadequacies, compliance auditing and certifications serve an essential purpose: demonstrating that service providers or enterprises meet a common minimum set of security requirements as laid out in specific standards. Thus, industry and academia alike have focused on addressing challenges in compliance auditing.

Existing solutions are predominantly model-based; that is, they use models to represent a subset of controls, and verify the evidence gathered (either manually or automatically) against the representation and reason about its adequacy. However, model-based solutions by and large cannot solve the nontechnical challenges of compliance auditing, as they are dependent on the assumptions made in constructing the models, which themselves may be inaccurate due to subjectivity in interpretation of controls and disagreement between different parties involved in the audit process.

In this paper, we propose that *hybrid* approaches should be used instead, wherein model-based approaches are augmented with machine learning approaches to reduce uncertainties. Our work makes the following four contributions:

1) A discussion of the challenges present in evidence collection for compliance auditing, especially in large-scale IT systems (such as public clouds), with anecdotal illustration of said challenges. We specifically focus on the uncertainty and unpredictability introduced by humans.
2) A taxonomic framework for understanding the causes of and potential solutions to uncertainty in the audit process. We classify uncertainties by the stages of audit in which they arise, party or parties involved, target, and impact of ambiguities in language. We also explain the underlying causes of uncertainties and describe why existing model-driven approaches are ill-suited to addressing them.
3) Our proposal to use machine learning in conjunction with model-based approaches to understand the uncertainties and to reduce the manual effort of identifying meaningful evidence within large corpora of data.
4) Example scenarios to illustrate potential application of learning techniques to reduce the specific uncertainties we have identified in the audit process.

II. Background and Motivation

Compliance auditing is a systematic, independent, and documented process for obtaining objective evidence and evaluating the evidence objectively to determine the extent to which audit criteria are fulfilled [6]. Enterprises go through various types of auditing, such as financial, tax, environmental, and fraud investigations.

Our focus in this paper is on information system auditing that involves assessment of security- and reliability-related controls relevant to the IT systems of an organization. Furthermore, we consider only external third-party auditing because it is the predominant mechanism for certification with regulatory and industrial standards, and is therefore of major importance to operators and providers of large IT infrastructures like CSPs. An external third-party audit involves examination of an organization by an independent entity to produce an unbiased report in accordance with a reporting framework, such as FedRAMP for CSPs offering services to the US government.

A. Description of the compliance audit process

Here, we provide a primer on the compliance audit process that frames the rest of our analysis. We identify the different parties involved in audit and describe the stages of the audit process in more detail.

Parties involved. Audit involves the following *parties*:

- **Auditor**: The organization or individuals, internal or external, who perform the audit.
- **Compliance SME**: The team internal to the service provider (the auditee) that consists of subject matter experts (SMEs) on compliance audit and liaises with both the auditor and system owners.
- **System owner**: The team(s) internal to the service provider that operate and manage the system and have the technical expertise to collect evidence and change system configurations to remediate noncompliance.
- **Tool vendors**: Third-party vendors of tools that are designed to help system owners automate audit evidence collection and validation.

Stages of audit. We separate the compliance audit process into three main stages: **initial evidence collection**, **audit interview**, and **reporting and remediation** of noncompliance.

1) Initial evidence collection: In the first stage, the service provider collects evidence from its system (during system operation) based on its own compliance SMEs' understanding of what is sufficient to prove compliance. Often, compliance SMEs will use an audit checklist provided by the auditor as a guide for what needs to be collected. The evidence is collected using a combination of manual collection by the system owners, third-party automated audit evidence collection tools, and software developed in-house to collect evidence automatically. During the evidence collection process, the compliance SMEs evaluate the evidence based on their prior experience with audits and their knowledge of the standards and the service provider's systems. Evidence collection is an iterative process. If the service provider's compliance SMEs believe the evidence to be insufficient, they request additional evidence from the system owners.

2) Audit interview: Once the evidence has been collected, the service provider, usually represented by its compliance SMEs, then submits the evidence to the auditor for review. During the audit interview phase, the auditor evaluates the evidence against the standard in question, and may request clarifications or additional evidence from the service provider. The audit interview process takes place within an allotted duration, so it is important for both the auditor and service provider that evidence is collected and presented in a timely manner. The audit interview itself can be an iterative process in which the auditor and compliance SMEs may go back and forth until the auditor is satisfied that the service provider has shown sufficient evidence or the audit deadline has passed.

3) Reporting and remediation: Once the allotted audit duration has passed, the auditors compile their findings into an audit report, which they give to the service provider. The report contains a list of inadequacies and instances of noncompliance that should be remediated. During the remediation phase, the service provider is given a fixed amount of time to rectify the issues laid out in the audit report and demonstrate compliance by providing additional evidence to the auditors. If additional evidence must be collected, it is collected in the same iterative manner (between the compliance SMEs and system owners) as in the previous phases. Finally, in the case of external audits, the auditors make their recommendations to a regulator, which makes final decisions on certification.

III. Examining Uncertainties in Compliance Audit

We define *uncertainties* in the audit process as any respects in which a party involved in the audit has an insufficient interpretation or understanding (of the system, evidence, or other parties in the audit process) that can lead to inefficiencies or errors.

A. A taxonomy of uncertainty in compliance audit

We present a taxonomy for examining and classifying uncertainties in the audit process. We identified the four parties involved in audit in Section II-A, namely, the auditor, the compliance SME, the system owner, and the tool vendor. In our taxonomy, we use those parties to identify the source(s) of a given type of uncertainty and the stakeholder(s) who would desire to reduce it.

We further classify uncertainties in two ways related to human involvement, which is the source of much of the uncertainty in auditing. First, we distinguish between two kinds of *cross-party involvement*, i.e., whether the uncertainty involves interactions between different parties (i.e., **multi-party**) or is internal to one party (i.e., **single-party**). Second, we look at the role that *human interaction* plays; that is, whether the uncertainty arises because of **interpersonal** interactions (whether single-party or multi-party) or if it is due to some **inherent** ambiguity, vagueness, or challenge that is irrespective of the individuals involved.

We also attribute uncertainties to the *stage* in the compliance audit process in which they arise. Recall from Section II-A that the three stages are initial evidence collection, audit interview, and reporting and remediation. In addition, we identify five *targets* for uncertainty that describe what, within a cloud system, an individual can be uncertain about:

1) **System compliance**: Parties may be uncertain about whether a particular control has been satisfied by the current *state* of the system, which refers to policies and processes in place, configuration of system components, data being collected, and the monitoring setup of the system, among other elements.
2) **Evidence necessity**: Parties may be uncertain about what evidence they are required to collect for compliance with a particular control.
3) **Evidence availability**: Parties may be uncertain about what evidence is available in the system that could potentially be used to prove compliance with a control.
4) **Evidence utility**: Parties may be uncertain about how to use the evidence available in the system to prove compliance with a particular control.
5) **Control interpretation**: Parties may be uncertain about whether the evidence collected will be considered sufficient by another party involved in the audit.

Each of the aforementioned targets of uncertainty affects each of the parties involved in an audit in different ways because of the parties' different responsibilities. The first, second, and third targets of uncertainty are predominantly felt by the parties internal to the service provider (namely, the compliance SMEs and system owners), whereas the rest of the above targets of uncertainty affect all parties.

Finally, for aspects of audit involving natural language, such as regulatory controls, policy documents, and communications between parties, we identify *language-derived* classes of uncertainty. While others have developed more granular classes for language-related uncertainties in policies and regulations [7], we boil them down to two basic classes: **ambiguity** and **vagueness**. *Ambiguity* arises when a statement or concept has multiple but finitely many interpretations. More generally, ambiguity can be caused if words have multiple meanings or statements have multiple grammatically correct but semantically distinct readings [8]. *Vagueness* arises when a statement provides insufficient detail to convey its meaning or can have borderline cases. Consider the NIST SP 800-53 control AU-6 [9] which specifies actions for "organization-defined inappropriate or unusual activity." Every individual within an organization may have his/her own definition of "inappropriate or unusual activity," with some cases being dependent on context and/or an arbitrarily-specified threshold (e.g., too many failed authentication attempts).

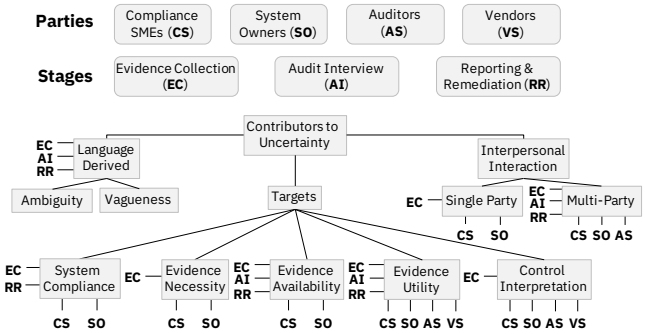

Figure 1: A taxonomy of uncertainties. Each subtree describes a classification of uncertainty, and leaf nodes denote classes. Parties affected are annotated underneath each class, and stages of audit affected are annotated to the left.

We summarize our taxonomy of the types of uncertainty in Fig. 1. Classifying uncertainties in the way we have done above is important because it enables us to *localize* uncertainties experienced in practice to a particular set of parties or individuals and stage of audit, and thereby guide mitigation approaches.

B. Underlying causes of uncertainty in compliance audit

Auditor domain perspective. Third-party auditors in different domains often have different perspectives that guide their interpretations of the standards and hence the evidence they deem sufficient/insufficient. Automated scanning and compliance testing tools like Chef InSpec [10], which specify relationships between tests and the controls they validate by using handcrafted mappings written in a domain-specific language, reduce interpretation-related uncertainties because the mappings they specify can easily be vetted by auditors and are therefore more likely to be accepted by auditors in different domains. However, there is still a need for more work in this area.

Variability of technical knowledge. Of course, individuals involved in compliance audit have varying levels of technical knowledge. That influences their abilities to identify, for example, what evidence is necessary to demonstrate compliance and whether controls can be circumvented in ways that call

for additional evidence. Automated scanning and compliance testing tools can reduce the amount of uncertainty introduced by variability of knowledge between parties by reducing the amount of discretion involved in collecting or interpreting evidence, but their scope is limited to a subset of controls.

Diversity and scale of evidence collection. A large CSP may have thousands of system components that may be managed internally by separate teams (e.g., network, compute, and storage). As a result, during compliance audit, a CSP's compliance SMEs may need to coordinate the efforts of dozens of individuals (system owners, auditors, and other compliance SMEs) to collect evidence in a timely manner and verify that the evidence collected is correct and sufficient. That introduces interpersonal, multi-party uncertainties, as there is increased room for duplicated effort, miscommunication, and oversight that can lead to noncompliance findings.

RSA Archer [11] addresses some of the multi-party, interpersonal uncertainties in coordinating evidence collection and identifying potential oversights by assisting with audit management. Formal modeling-based approaches (e.g., Amazon's Zelkova [12], works by Majumdar et al. [13] [14], and work by Stephanow et al. [15]) make strong assumptions that the formal models of controls and evidence that they manually construct are authoritative. They are therefore still susceptible to control interpretation issues. Furthermore, the test cases and case studies used by each approach are from control families that are relatively easy to model formally, so it is likely that the approaches would not generalize to the larger array of controls present in standards.

Dependencies between cloud service levels. When audit is done for a CSP at a particular level, the CSP is responsible for evidence collection for its own components. However, a CSP may also itself utilize cloud services from providers at lower levels and depend on evidence collected by the lower-level provider for the CSP's own compliance auditing. When the dependencies span organizational boundaries (e.g, a SaaS provider using a public IaaS infrastructure), uncertainties may be introduced in terms of evidence utility towards audit and evidence availability, as the dependent provider may not have visibility into the underlying provider's evidence collection procedures. The adoption of the multi-cloud paradigm introduces another avenue for uncertainty, as the different kinds of evidence provided by underlying providers must also be coordinated/consolidated during evidence collection. Uncertainties caused by dependencies between cloud service levels can be partially mitigated by clear delineation of threats and compliance responsibilities between the different CSPs

IV. Tackling Uncertainty using Machine Learning

The majority of solutions proposed in the literature for dealing with uncertainties and issues in evidence collection and validation rely on formal models. Logically, they can be grouped into: automated compliance-scanning tools that use domain-specific languages, such as Chef InSpec [10]; formal logic-based compliance policy verification approaches, such as Zelkova [12] and PVSC [14]; and tools to interpret natural language controls into formal compliance requirements, such as those proposed by Breaux [16] and Massey [8].

Model-based approaches are good at representing policies in unambiguous ways and verifying system compliance against them. They are suitable for reasoning about security properties of systems that have a large number of components of a small number of unique types, as models for such systems can be handled easily by existing model-checking and SAT solver tools. However, model-based approaches rely on accurate model construction, which needs manual verification, so the results produced based on those models are dependent on the assumptions made when creating them. The interpretation of controls and their mappings to evidence used by the modelers—who are often compliance SMEs or tool vendors—may contradict with those of the auditors (as described in Section III-B), thereby reducing their utility. Furthermore, it is difficult to examine the diverse data available in large CSPs, as each message format within each data type must be manually translated to the model language [14].

Many of the issues present in model-based approaches can be addressed by combining them with machine learning-based approaches. Machine learning (ML) techniques are well-suited to problems involving uncertainty and noise, and have been widely applied to problems in data mining, natural language processing, image recognition, and decision-making.

For compliance audit, ML is well-suited to solving problems in evidence discovery (i.e., those affecting evidence necessity, availability, and utility) and in identifying discrepancies between interpretations of controls when observed audit outcomes deviate from model-driven expectations. Below, we propose some potential applications of ML techniques in compliance audit that would help reduce specific kinds of uncertainty and improve audit efficiency. We describe existing literature and identify areas where more work is needed. We then present some applications of how ML can be used to augment model-based approaches to create what we call *hybrid* approaches. We do not claim that this list is comprehensive, but we try to highlight the areas of ML that would be most impactful in solving audit-related problems.

A. Applications of machine learning

Evidence Discovery. A key challenge in compliance audit is *evidence discovery*, i.e., identifying data sources relevant to audit from within the large array of sources available to CSPs. Evidence can take many forms. Log messages, scan results, configuration files, and alerting rules are well-studied and covered by existing tools. Additionally, CSPs must also provide policy documents, communication proofs, user activity documentation, and other unstructured or human-generated evidence to be certified for most standards.

In our taxonomy, improving evidence discovery can reduce uncertainties about evidence availability and utility. Evidence discovery is particularly important to compliance SMEs and system owners to optimize evidence collection. It is also

relevant to tool vendors as there is a large market for tools that aid in evidence discovery. Evidence discovery can help deal with uncertainties caused by the diversity and scale of evidence and dependencies between cloud layers. We explore ML solutions for evidence discovery in natural language documents, multi-modal documents, and structured data.

1) Natural language documents: These are documents that contain human-written text semantically relevant to audit, e.g., policy and process documentation, e-mails, meeting recordings/transcriptions, and presentation slides. These documents often comprise crucial evidence, but their collection is usually time-consuming because they are scattered throughout the organization. It is also challenging for compliance SMEs and system owners to determine which documents are relevant for compliance from metadata alone.

Natural language processing (NLP) and ML are promising evidence discovery techniques for natural language documents. Text mining algorithms have been used extensively in the legal and business intelligence domains to perform document summarization, entity-relationship modeling, topic modeling, and document relevance tagging; they could be used to identify documents that are topically related to specific controls or to extract semantic information from documents necessary to prove compliance. Weak supervision techniques (e.g., Snorkel [17]) are powerful for automating extraction of domain-specific information from text; such techniques can be applied to process- or policy-related documents.

2) Multimodal documents: These are documents that contain more than one form of data (text, images, audio, etc.) that can be used as evidence. Multimodal evidence predominantly includes structured text (e.g., titles, column headers), unstructured text, images, and figures or charts.

While the text component of multimodal evidence can be analyzed as described above, the visual components must be processed using different techniques. A potential solution is multimodal learning [18] that combines image recognition with NLP algorithms to jointly learn semantics from multiple modes of data present in a single data source. Multimodal learning can be particularly useful for classifying, identifying, and decomposing multimodal evidence sources, such as diagrams and flowcharts, to identify those pertaining to controls. Symbolic reasoning [19] is another approach in which concepts and semantics are learned from multimodal data, in a semi-supervised or unsupervised manner for answering semantic questions about the data. While it has only been demonstrated for simple queries, we believe that the high amount of structure (like flowcharts) in process- and policy-related documents make them a prime candidate for symbolic reasoning techniques. Finally, weak supervision in combination with the above approaches, may be useful in automatic extraction of useful parts of multimodal evidence.

3) Structured textual evidence: This category of evidence includes source code, configuration files, event logs, and alerting rules. Considerable work has been expended in understanding how to use structured evidence in audit. Many unsupervised techniques have been developed to identify log message formats and use them for anomaly detection [20]; these could be applied to reduce manual effort required by automated compliance approaches [21]. NLP techniques may also be useful in semantic analysis of cloud log data [22]. Services like Amazon's Macie [23] can analyze structured evidence using ML techniques; as the area is quite saturated, we do not explore it further here.

Extracting Requirements from Standards. Some of the ML approaches we describe above can also be applied for extracting requirements from natural language standards. Determining unambiguous requirements is important to auditors and compliance SMEs, who face uncertainties in control interpretation and evidence utility that may be caused by auditor domain perspective, variability of technical knowledge, and language-related ambiguities.

Massey et al. [24] use text mining and topic modeling to identify and classify ambiguities in the text of regulatory standards. Riaz et al. [25] use a combination of semantic analysis NLP techniques and unsupervised clustering to extract requirements specifications from standards. Both these approaches could be applied during audit interview to identify and address potential control interpretation differences.

B. Hybrid ML and model-based approaches

We propose combining the power of ML with model-based approaches to maximize benefit for compliance audit. We demonstrate the proposed hybrid approach with two examples.

Evidence Mining. Due to the scale of CSP systems, it is infeasible to collect and store all possible evidence in preparation for audit. CSPs must *mine* their evidence to understand which parts of the evidence are relevant to each control in a given standard and thus necessary to retain.

We propose an evidence mining framework that uses NLP and ML techniques to construct semantic models for each evidence source used in audit based on labeling from historical audits. We would use the semantic models to label each line of the evidence with its meaning in the context of compliance controls. We would then combine the labeled historical evidence with existing model-based mappings between controls and evidence types to identify the specific lines of evidence relevant to each control type. Different types of evidence may require different techniques for identifying relevant evidence, as we describe in Section IV-A. We envision that the semantic models for multiple evidence sources of similar types could be trained as a cohort so as to create a semantic translation model across those sources. Such an approach would facilitate greater automated evidence collection and validation, and potentially enable support for previously unseen standards and system components.

Dynamic, End-to-end Evidence Collection and Validation. The extent to which previous audit results educate future audits is dependent entirely on the manual effort (or lack thereof) of individual compliance SMEs and system owners. We envision that ML can be integrated into the audit process

to create a feedback loop from auditor validation of evidence to internal evidence collection and validation. We propose the use of a graphical model to represent the mappings between evidence, controls, and standards. The model could be trained using historical audit records to establish a baseline understanding of how evidence relates to controls. During audit, a CSP could use the graphical model to decide what evidence to collect and present to auditors, and use both positive and negative compliance findings as labels for supervised learning to refine the model as the audit is taking place. In this way, CSPs could continuously improve their models for evidence collection and validation using audit results as feedback to reduces uncertainties and inefficiencies over multiple self-audit and formal audit cycles.

V. Conclusions and Future Work

We presented a taxonomic framework for understanding uncertainties in the audit process created by human involvement. Using that foundation, we identified the causes of and potential solutions to uncertainties. We explained why existing approaches based on formal logic and domain-specific languages are not sufficient to address challenges created by those uncertainties. We proposed hybrid approaches that combine machine learning with formal model-based solutions, and presented examples of their applicability. Such approaches can be useful not only to better deal with uncertainties, but also to reduce costs, errors, and the manual effort of identifying evidence needed for compliance certification. Clearly, there is need for more work in this area. Many of the uncertainties identified in our taxonomy affect security and compliance audit in domains beyond cloud, such as autonomous vehicles and critical infrastructures. Thus, progress made in addressing the uncertainties for cloud audit will have broad application and impact.

Acknowledgments

We thank Brian Cram and Sandhya Narayan for their valuable input and Jenny Applequist for her editorial assistance.

References

[1] Compliance Programs – Amazon Web Services (AWS). [Online]. Available: https://aws.amazon.com/compliance/programs/

[2] Cloud Compliance – Regulations & Certifications. [Online]. Available: https://cloud.google.com/security/compliance/

[3] Compliance on the IBM Cloud. [Online]. Available: https://www.ibm.com/cloud/compliance

[4] Azure Compliance. [Online]. Available: https://azure.microsoft.com/en-us/overview/trusted-cloud/compliance/

[5] Fedramp.gov. [Online]. Available: http://www.fedramp.gov/

[6] *Quality management systems – Fundamentals and vocabulary*, ISO Std. 9000:2015(en), Sep. 2015.

[7] A. K. Massey, R. L. Rutledge, A. I. Anton, and P. P. Swire, "Identifying and classifying ambiguity for regulatory requirements," in *IEEE 22nd International Requirements Engineering Conference*, Aug. 2014, pp. 83–92.

[8] A. K. Massey, R. L. Rutledge, A. I. Antn, J. D. Hemmings, and P. P. Swire, "A Strategy for Addressing Ambiguity in Regulatory Requirements," Georgia Institute of Technology, Technical Report, 2015.

[9] "Security and Privacy Controls for Federal Information Systems and Organizations," NIST, SP 800-53, Apr. 2013, doi: 10.6028/NIST.SP.800-53r4.

[10] Chef InSpec - Audit and Automated Testing Framework. [Online]. Available: https://www.inspec.io/

[11] RSA Archer. [Online]. Available: https://www.rsa.com/en-us/products/integrated-risk-management/audit-management

[12] J. Backes, P. Bolignano, B. Cook, C. Dodge, A. Gacek, K. Luckow, N. Rungta, O. Tkachuk, and C. Varming, "Semantic-based Automated Reasoning for AWS Access Policies using SMT," in *Formal Methods in Computer Aided Design (FMCAD)*, Austin, TX, Oct. 2018, pp. 1–9.

[13] S. Majumdar, T. Madi, Y. Wang, Y. Jarraya, M. Pourzandi, L. Wang, and M. Debbabi, "Security Compliance Auditing of Identity and Access Management in the Cloud: Application to OpenStack," in *IEEE 7th International Conference on Cloud Computing Technology and Science*, Nov. 2015, pp. 58–65.

[14] S. Majumdar, Y. Jarraya, T. Madi, A. Alimohammadifar, M. Pourzandi, L. Wang, and M. Debbabi, "Proactive Verification of Security Compliance for Clouds Through Precomputation: Application to OpenStack," in *Computer Security ESORICS 2016*, vol. 9878. Springer, 2016, pp. 47–66.

[15] P. Stephanow and C. Banse, "Evaluating the Performance of Continuous Test-Based Cloud Service Certification," in *17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, May 2017, pp. 1117–1126.

[16] T. D. Breaux, "A method to acquire compliance monitors from regulations," in *Third International Workshop on Requirements Engineering and Law*, Sydney, Australia, Sep. 2010, pp. 17–26.

[17] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. R, "Snorkel: Rapid Training Data Creation with Weak Supervision," *VLDB Endowment*, vol. 11, pp. 269–282, 2017.

[18] T. Baltruaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *arXiv:1705.09406 [cs]*, May 2017, arXiv: 1705.09406.

[19] T. R. Besold, A. d. Garcez, S. Bader, H. Bowman, P. Domingos, P. Hitzler, K.-U. Kuehnberger, L. C. Lamb, D. Lowd, P. M. V. Lima, L. de Penning, G. Pinkas, H. Poon, and G. Zaverucha, "Neural-Symbolic Learning and Reasoning: A Survey and Interpretation," Nov. 2017. [Online]. Available: https://arxiv.org/abs/1711.03902v1

[20] J. Zhu, S. He, J. Liu, P. He, Q. Xie, Z. Zheng, and M. R. Lyu, "Tools and Benchmarks for Automated Log Parsing," in *Proc. of the 41st International Conference on Software Engineering: Software Engineering in Practice*, ser. ICSE-SEIP '10, Piscataway, NJ, USA, 2019, pp. 121–130.

[21] S. Majumdar, A. Tabiban, Y. Jarraya, M. Oqaily, A. Alimohammadifar, M. Pourzandi, L. Wang, and M. Debbabi, "Learning probabilistic dependencies among events for proactive security auditing in clouds," *Journal of Computer Security*, vol. 27, no. 2, pp. 165–202, Mar. 2019.

[22] C. Bertero, M. Roy, C. Sauvanaud, and G. Tredan, "Experience Report: Log Mining Using Natural Language Processing and Application to Anomaly Detection," in *IEEE 28th International Symposium on Software Reliability Engineering (ISSRE)*, Oct. 2017, pp. 351–360.

[23] Amazon Macie. [Online]. Available: https://aws.amazon.com/macie

[24] A. K. Massey, J. Eisenstein, A. I. Antn, and P. P. Swire, "Automated text mining for requirements analysis of policy documents," in *21st IEEE International Requirements Engineering Conference*, Jul. 2013, pp. 4–13.

[25] M. Riaz, J. King, J. Slankas, and L. Williams, "Hidden in plain sight: Automatically identifying security requirements from natural language artifacts," in *IEEE 22nd International Requirements Engineering Conf.*, Aug. 2014, pp. 183–192.