Unstructured Data for Internal Control

Esperanza Huerta
San José State University

ABSTRACT

This paper proposes a framework for studying the connections between unstructured data and internal control. The framework is organized in three stages. Stage 1 defines the goals of using unstructured data for internal control. Stage 2 describes the types of unstructured data to be accessed and the analyses to be done. Stage 3 defines the outcomes that can be achieved. Patterns in unstructured data are relevant for internal control, but using unstructured data poses unique challenges in its analysis. This paper discusses some challenges, including structurization, variability, veracity, cleaning, tagging, privacy, digitization, bias, and explainability. Informed by public sources, the proposed framework describes the use of unstructured data in the Airbus' corruption investigation conducted by the enforcement agencies of three countries.

Keywords: unstructured data, internal control, data analytics, research framework

Copyright statement: Authors retain the copyright to the manuscripts published in AABRI journals. Please see the AABRI Copyright Policy at http://www.aabri.com/copyright.html

INTRODUCTION

Organizations can use many kinds of data to implement controls and identify control breaches. Data for designing and testing internal controls can be drawn from the large amount of structured data produced by accounting information systems and the even larger amount of unstructured data created inside and outside organizations. However, limited research exists on how unstructured data can support the design and testing of controls. This paper proposes a framework for studying the connections between unstructured data and internal controls.

Frameworks list factors for phenomena of interest (Nagarajan & Overbeek, 2018) and relate concepts and theories to advance and systematize knowledge (Rocco & Plakhotnik, 2009). Frameworks amount to “organized structures of ideas and concepts” that illustrate “relationships between relevant dimensions of a given phenomenon” (Mwansa, 2015, p. 109). The proposed framework identifies types of unstructured data and analysis, highlighting the challenges of using unstructured data in designing and testing internal controls.

The consequences and the number of frauds and material misstatements fuel research on internal controls. Although providing an accurate amount is impossible, the estimated global fraud loss in 2024 is nearly $5 trillion (ACFE, 2024). In addition to designing controls for fraud prevention and detection, testing controls can provide evidence of policies, procedures, regulations, and laws being violated or ignored.

Internal controls have traditionally relied on structured data analysis—identifying unusual patterns in amounts or dates of transactions, for instance. Although unstructured data has been used to support forensic investigations—the Enron investigation, for example, analyzed emails for evidence of fraudulent intent—technological advances have only recently made it easier to exploit large amounts of unstructured data.

In 2020, approximately 64.2 zettabytes were created, with an estimated compound annual growth rate from 2020 to 2025 of 23% (Rydning & Shirer, 2021). The fastest-growing data segments are unstructured data from the Internet of Things, video, and social media. The growth of unstructured data has implications for designing and testing controls. Unstructured data is vital because "most of the data that moves markets are inherently unstructured" (Stewart, 2015, p. 114). News on newspapers, radio, websites, and other outlets about central bank announcements, geopolitical developments, product releases, research breakthroughs, natural disasters, or weather phenomena is unstructured and delivered as free text, speech, video, or images (Stewart, 2015). Although some data can be acquired later in a structured format from data brokers, the announcements immediately impacting markets are initially unstructured. There are “no reasonable limits on sources of data, but there are great limits on what data an organization can store and make useful” (Bumgarner & Vasarhelyi, 2015, p. 22).

This paper contributes to the literature in three different ways. First, it discusses current and potential applications of unstructured data for internal controls. It assists accountants, auditors, and regulators by giving guidance on incorporating unstructured data into internal controls and audit procedures. Second, it outlines accountants' challenges in adopting and exploiting unstructured data. Third, it proposes a framework to foster a research agenda on using unstructured data for internal control.

The remainder of the paper is organized as follows: Section 2 defines unstructured data and states its characteristics. Section 3 discusses the relationship between unstructured data and internal controls. Section 4 summarizes how unstructured data can be analyzed. Section 5 presents the challenges of using unstructured data. Section 6 puts forward the framework

developed for unstructured data and internal control. Section 7 uses the framework to illustrate the use of unstructured data in Airbus' corruption investigation. Finally, section 8 presents the conclusions.

UNSTRUCTURED DATA

Data can be classified using different criteria ("Data" is used as a mass noun, as is commonplace in data analytics writing, acknowledging that some people prefer "data are" in their writing.) Classifications are not mutually exclusive; each provides a lens through which data may be viewed. Standard categorizations include data representation (digital/non-digital), data creator (human/machine), data event (main-event/circumstantial), and data arrangement (structured/unstructured). Structured data has a predefined arrangement. For instance, a list of employees in an organization might be structured to consist of employee ID, last name, first name, middle name, and other information. Unstructured data, as the name implies, does not have a predefined arrangement.

Images are considered unstructured data because the meaning of an image is not inferred merely from the bits it holds. Text is also regarded as unstructured data because the content expressed is more than a collection of characters. Audio and video are also considered unstructured data. The meaning of recordings must be gathered from "seeing" and "listening" to the recording, making sense of it, not from noting the file structure in which they are stored.

Dichotomizing data into structured and unstructured can help describe its characteristics. However, the degree of structure is more of a continuum than a dichotomy—unstructured data usually includes some degree of organization. Digital images, for instance, contain some data organized in a predefined pattern, such as file name and extension, date, size, and other metadata associated with the image. While acknowledging the structure continuum in the data, the data is classified as structured and unstructured, as is customary in data analytics.

In addition to the commonly known forms of unstructured data (text, audio, photos, videos), organizations have creatively used several others (sensor data and images captured by drones or satellites). Table 1 in the Appendix lists various types of unstructured data.

The types of unstructured data listed in Table 1 reflect the classic five human senses: vision for text, image, and video, audition for video and audio, olfaction for scent, gustation for taste, and touch for texture. Although humans naturally process these types of unstructured data, computers, with appropriate input devices, can capture and process the same data in even greater ranges than humans. Specialized sensors can capture and process unstructured data, such as infrared or ultraviolet waves, magnetic fields, and ultrasonics and infrasounds, that humans cannot detect.

INTERNAL CONTROL AND UNSTRUCTURED DATA

Internal controls are policies and procedures implemented to reasonably ensure the reliability of the information, the safeguarding of assets, efficiency of operations, and compliance with laws and regulations. Until recently, auditors and accountants have heavily relied on transactional data to design and test internal controls. Data is typically drawn from the structured databases of accounting information systems. Analyses aim to identify unusual patterns of transactions—anomalies—that can result from errors, fraud, bribery, money laundering, or other illegal activities.

To identify anomalies, auditors establish a benchmark regarding quantities, prices, dates, and potentially other pieces of structured information. Auditors then compare the results of their analyses with the established benchmark to identify anomalies. Anomalies are investigated further because they might represent fraud or illegal activities, but they may also represent unusual but legal events.

Despite the reliance on structured data, unstructured data is increasingly being used to design internal controls and to provide evidence of breaches to controls in fraud cases. For instance, the U.S. Securities and Exchange Commission (SEC) used satellite images to demonstrate that a construction company recognized revenue for buildings that had not been built at all (SEC, 2017). In addition to providing evidence, unstructured data can be analyzed to evaluate the strength of controls. Auditors should seek to verify transactions, not just with an invoice and receipt, but multi-modal evidence that a transaction took place. Photo, video, GPS location, and other metadata could accompany transaction data" (Moffitt & Vasarhelyi, 2013, p. 9). Although not explicitly geared towards testing controls, there is evidence of using unstructured data to verify events. Thasos Group, for instance, used the number of cell phone signals going in and out of a Tesla factory to determine whether the company was ramping up production as promised (Dezember, 2018).

Text data can be processed "to extract textual features such as part of speech, readability, cohesion, tone, certainty, tf-idf scores, and other statistical measures" (Moffitt & Vasarhelyi, 2013, p. 5). The SEC, for instance, analyzes text disclosures, computing "tonality" indexes, which reflect the positive or negative tone used in the written discussion of the results (Baugess, 2016). Tonality indexes are compared with the structured data analysis (data from the financial statements). The expectation is that the tonality of text disclosures should match the structured data analysis; unfavorable results should align with negative tonality, and favorable results should align with positive tonality. The divergence between structured and unstructured data analysis would raise a flag for further investigation.

Similarly, text data from transactions can be analyzed along with their structured data. Internal control commonly requires accounting entries to have a written description of the concept originating the transaction. In the Airbus case described in Section 7, the corruption investigation analyzed the text on the accounting entries to identify unusual patterns that may reflect bribes. Beyond text data, audio and video conversations can also identify collusion or bribery. Audio and video provide richer information than text data because subtle features, such as irony or jokes, can be inferred from the pace and tone of the conversation. For instance, the vocal signatures of managers have been analyzed to identify positive or negative vocal cues in conference calls (Mayew & Venkatachalam, 2013). In addition to single conversations, a relational analysis of communications can identify who is related to whom and uncover unknown patterns. A well-known application of relational analysis is the Panama Papers, in which relational patterns identified the players in the money laundering and tax evasion schemes (Eifrem, 2017).

Unstructured data can also be generated as a byproduct of technologies used for asset protection. For example, radio frequency identification (RFID) chips affixed to goods enable real-time inventory tracking, yielding data on their location and movement (Canelón et al., 2020). Similarly, videos can supply information regarding the movement of inventory. Because videos capture the entire surroundings, they can reveal details about individuals interacting with items (Canelón et al., 2020). For instance, videos in Amazon self-service stores trace customers' movements and identify the goods customers select for automatic checkout (Nishihara, 2018)

Video and images do not need to come from fixed cameras. Drones can take video to automatically scan inventory items in warehouses or outdoor locations. Drones have also been used to take aerial images of houses to determine eligibility for insurance renewal (Eaglesham, 2024). Audio can be used to protect assets. Budweiser, for instance, compares the audio of its equipment to a benchmark of equipment functioning normally to determine when maintenance is needed before the equipment breaks down, thus preventing factory downtimes (Castellanos, 2019).

Haptic data has been used to optimize production processes and to provide robots with a sense of touch, enabling them to manipulate objects and navigate an environment (Bartlett, 2023). This technology can predict material feel before manufacturing and help e-commerce by recommending products based on tactile preferences (Bartlett, 2023). Brands like Hanes and Adidas utilize haptic technology to develop fabrics and materials that are comfortable, breathable, and perform well (Bartlett, 2023).

Unstructured data concerning scents and tastes can also be captured. Devices known as e-noses will detect disease by the smell of someone's breath or excrement. Recent work at Brown University has produced a device, TruffleBot, that sniffs aromas, sucking up vapors and moving them past special sensors. Additionally, the device measures temperature and pressure changes, which greatly help identify smells (Scudellari, 2018).

ANALYZING UNSTRUCTURED DATA

A variety of techniques exist to analyze unstructured data. These techniques could be placed on a continuum from manual techniques (conducted by humans), such as content analysis, to automated techniques, mainly derived from artificial intelligence (conducted by computers). Although analyzing large amounts of unstructured data benefits from automated approaches, manual methods are still relevant for internal controls, particularly when unstructured data is used to provide evidence of a control breach and to train and evaluate the performance of automated approaches.

The analysis selected—manual or automated, or more likely a combination of both—should be aligned with the scope of the intended use of unstructured data. Scope refers to the boundaries of the project. A narrow scope uses a limited quantity of unstructured data for a specific event and a short period. A broad scope uses much unstructured data for a vaguely defined event and a considerable period. A project with a narrow scope aligns with manual data analysis; a project with a broad scope aligns with automated data analysis.

When the SEC used satellite images as evidence in a fraud case (SEC, 2017), the data was gathered from preselected locations where the company had disclosed construction. Humans could interpret the limited number of images collected. Similarly, an insurance claim included a video as evidence of car damage by a bear (Correal, 2024). Ironically, the same video was analyzed by experts and determined to be fraudulent insurance as the bear was a person in a bear costume (Correal, 2024). These are examples of a narrow scope: the unstructured data was limited, the event was specific, and the time frame was short.

In contrast, a project with a broad scope using large amounts of unstructured data can analyze data automatically faster and cheaper than a manual analysis. Although manually analyzing large amounts of unstructured data is time-consuming and costly, manual analysis can and has been conducted in the past. In 1978, lawyers and paralegals analyzed six million

documents for months for an antitrust lawsuit against CBS. The cost was over 2.2 million dollars (Markoff, 2011). Artificial intelligence enables automating at least some part of the data analysis.

The literature shows multiple automated methods that have been used for analyzing text, including word frequency counts for measuring tone in text (Henry & Leone, 2016), unsupervised naïve Bayesian classification algorithms to classify news articles (Van den Bogaerd & Aerts, 2011), and more recently generative AI (Huang et al., 2023).

Unstructured data is multifaceted; a unit of unstructured data can have different facets (Balducci & Marinova, 2018). An image, for instance, can be described based on facts (location or the number and type of objects in the picture) or inferred meanings (happy, sad, or neutral feelings), in addition to more mundane descriptors such as file size and set of pixels.

The method selected to analyze unstructured data, whether manually or automatically, must identify the facet of the data that is the focus of attention. For instance, when analyzing text, the user must decide whether the focus of attention is the tone, the sentiment, or any other facet. When analyzing audio, the user must decide whether the focus of attention is the volume, the intonation, or even the background noise.

Considering the vast array of automated approaches to analyzing unstructured data, selecting a method that aligns with the data type and the focus of attention is essential. Not all automated methods require the use of artificial intelligence. Bag of words, for instance, is an automated method based on word frequency that uses general-purpose or custom dictionaries of words and phrases associated with meaningful characteristics. However, other automated methods rely heavily on artificial intelligence. As a field of study, artificial intelligence aims to develop systems capable of performing tasks typically associated with human intelligence. Artificial intelligence can use machine learning and non-machine learning methods. Machine learning allows computers to learn patterns not explicitly specified by the programmer. Some techniques require more human guidance (supervised learning), while others require less (unsupervised learning). Non-machine learning methods allow computers to decide well-defined tasks based on rules, optimization, or simulations.

In the Airbus case described in Section 7, the investigation used a combination of manual and automated analyses. People manually identified the text of entries recording bribe payments. These entries were used to train a machine learning algorithm that, once trained, automatically analyzed the text on other accounting entries. The entries flagged by the automated analysis were further analyzed manually.

Although the Airbus probe is the most extensive international corruption investigation in scope (Department of Justice, 2020), at least until 2020, AI has been used before. In a Deferred Prosecution Agreement by Rolls-Royce, AI was used to automatically analyze thirty million documents with unstructured data to identify material relevant to the investigation (Murgia, 2017).

CHALLENGES TO USING UNSTRUCTURED DATA

At a high level of abstraction, analyzing unstructured data follows the steps of any information system: input, process, and output. However, the characteristics of unstructured data add unique challenges in each step. This section focuses on challenges exclusive to unstructured data (e.g., structurization) or that differ from structured data (e.g., data cleaning). The nine challenges discussed are structurization, variability, veracity, cleaning, tagging, privacy, digitization, bias, and explainability.

Structurization

Before a quantitative analysis, unstructured data is assigned numeric values manually or automatically. Structurization refers to reducing unstructured data to numerical values that computers can manipulate. For instance, Spotify assigns values to audio data in fourteen attributes, such as danceability. The structurization of the audio determines the number of dimensions and the processes to convert the audio to numeric values in each dimension.

Although some types of unstructured data, such as image recognition, can be analyzed without structuring the data, higher levels of abstraction require structurization. Many features can be extracted from images beyond objects. For example, features of the mood depicted in the picture, like aggressiveness, friendliness, or excitement, would require determining the dimensions to be extracted.

Variability

Variability refers to the peaks and troughs of the flow of the data (Gandomi & Haider, 2015). Although structured data experiences variability from seasonality or fashion, the variability of unstructured data is triggered by events difficult or impossible to predict, such as natural events. Variability imposes restrictions on how much data systems can handle, forcing systems to flexibly allocate resources or dump data that cannot be managed. Capturing unstructured data for testing controls in real time would require allocating resources to handle different degrees of variability.

Veracity

Veracity refers to the accuracy and truthfulness of the data. Although veracity is also relevant for structured data, the ability of current technology to produce fake video or audio (deepfakes) presents a significant challenge for unstructured data. As the old computer acronym GIGO (garbage-in, garbage-out) indicates, summarizing data and reporting it as facts without using verified data leads to misinformation. The recent Google blunder of reporting people as dead exemplifies the errors of automatically processing large amounts of unstructured data without a verification process (Ramachandran, 2019).

Deepfakes also pose a challenge for internal controls based on unstructured data. Controls have failed to detect deepfakes to authenticate users (Bousquette, 2024). During an experiment, Chase Bank's controls could not detect a deepfake voice generated with AI (Bousquette, 2024).

The veracity of unstructured data generated outside organizations should be systematically evaluated. For instance, an auditor analyzing social media to determine whether an increase in revenue is explained by consumers' acceptance of a new product could verify the veracity of the data. Good performance should be aligned with positive reviews. However, social media postings can be manipulated; companies sell reviews, likes, and tweets that can be purchased to manipulate people's perceptions. Auditors could conduct a relational analysis to determine whether the company being audited has paid companies to manipulate data on social media. The SEC has long been aware of the ability of social media to manipulate opinions; it continuously investigates whether postings are from people with no conflict of interest or are created to manipulate share prices (SEC, 2022).

Cleaning

Cleaning refers to the process of transforming data to be ready for analysis. Unstructured data shares all the challenges structured data has, from unformatted forms to unexpected missing and noisy data. However, unstructured data has unique challenges that make cleaning and integration more difficult. Moreover, cleaning techniques are tailored to the type of unstructured data. For instance, cleaning text data includes standardizing language, stemming, and lemmatization. Image cleaning might include centering and cropping, and audio cleaning might consist of noise reduction and normalization of loudness.

Tagging

Tagging refers to attaching a label to the unstructured data used in supervised machine learning. Large amounts of unstructured data tagged by humans are required to train a machine-learning algorithm. The resources needed to tag unstructured data are not the only difficulty. Humans can introduce their personal bias by incorrectly tagging content based on their cultural background (Wakefield, 2021).

Tagging can be completed by the general public, when it does not require specialized knowledge. For instance, users tagged text and images in CAPTCHA images and games such as ESP (Chocano, 2023). However, tagging unstructured data for internal control requires accounting knowledge. Moreover, the data that needs to be tagged is usually proprietary and must be kept confidential. In the Airbus corruption investigation described in Section 7, investigators tagged entries known as bribe payments.

Data privacy

Privacy refers to the proper handling of personal data. Although privacy is relevant for all types of data, the ubiquity of unstructured data, such as video surveillance and location data, imposes additional challenges. Regulations like Europe's General Data Protection Regulation (GDPR) (GDPR.eu, 2020), California's Consumer Data Privacy Act (CDPA) (CCPA, 2020), and more recently, Europe's Artificial Intelligence Act (AI Act) (European Union, 2024) establish limits on the use of people's data. A key element for using unstructured data without compromising the analysis has been anonymizing the data. Anonymization should make it challenging to correlate the data with other external data.

Digitization

Digitization (or digitalization) refers to converting non-digital data to a digital format. Archival data needs to be digitized before it can be used on computers. For instance, the New York Times has been scanning photos from its archives into digital form and then using various Google tools to create a catalog that includes not only the images but information recorded with them (e.g., on their backs), including text, handwriting, and dates, performing at least some automatic categorization of information (Greenfield, 2018). In addition to archival data, non-digital unstructured data still exists. Financial wrongdoing could be reported by phone, chat, email, or an old-fashioned paper note dropped in a secure box.

Bias

Bias refers to systematic errors in the data or in the analysis. Amazon, for instance, developed an algorithm for analyzing text data resumes to identify promising applicants. The initial data set contained a larger proportion of men than women, as is common in technology jobs, resulting in a biased algorithm that evaluated men more favorably than women. After unsuccessful attempts to debias the algorithm, Amazon discontinued the project (Tugend, 2019).

Explainability

The explainability of algorithms refers to the inability of an algorithm to define the criteria used to reach a decision (Knight, 2017). Regulations might limit the use of unstructured data because of the lack of explainability in the algorithms used. The GDPR, for instance, requires explicit explanations on how decisions are reached when they significantly impact people's lives. Denials of mortgage loans or jobs must explicitly indicate the factors for denial so people have an opportunity to improve.

FRAMEWORK

Figure 1 in the Appendix shows the proposed framework for using unstructured data for internal control. The framework includes three stages. Stage 1 establishes the goals of using unstructured data for internal controls. Stage 2 identifies the types of unstructured data that can be accessed or created and the data analysis techniques. Finally, Stage 3 reports the outcomes achieved.

The control goals established in Stage 1 indicate whether unstructured data is used to design and implement an internal control or is used to investigate a control breach. When using unstructured data as a recurring control, a system should be in place to capture and process the unstructured data to identify anomalies. For instance, when Budweiser used audio to determine whether the equipment was functioning normally (Castellanos, 2019), the company required a system to capture the audio and compare it to a benchmark.

In contrast, when the goal is to use unstructured data to investigate breaches of internal control, the scope for collecting and analyzing unstructured data is limited to the investigation. Although a company may decide to implement a recurring control using unstructured data following an investigation, using unstructured data to an inquiry is considered a one-time event.

In Stage 2, the unstructured data and analysis techniques are selected. When unstructured data is used to design and implement an internal control, unstructured data is the input from which patterns are identified. Unstructured data may already exist, or it might need to be captured. If unstructured data does not exist, the company should define the process to acquire the data systematically. For instance, Budweiser needed to determine multiple aspects of the control, from the frequency of recording the audio and the quality of the audio to specifying data storage and management (Castellanos, 2019).

However, when unstructured data is used to investigate breaches of control, unstructured data is drawn from existing data on an ad-hoc basis, depending on the investigation. Unstructured data is limited to the historical data captured before the investigation was initiated. Identifying the type of unstructured data that provides the most value could assist accountants,

auditors, law enforcement, or forensic accountants in allocating resources to the data with the highest potential for value.

Figure 1 shows a link between types of unstructured data and analysis approaches because the analysis depends on the type of unstructured data. Unstructured data does not necessarily imply large amounts of data; the SEC used a limited number of satellite images as evidence of construction not completed in a fraud investigation. However, unstructured data is usually massive and requires automated approaches for analysis.

Machine learning algorithms need much unstructured data, usually tagged by humans. Part of the analysis requires implementing procedures to ensure the accuracy of the tagging. Also, testing the veracity of unstructured data to detect deepfake audio or videos before they are analyzed involves sifting through the data, either manually or automatically. The task of structuring unstructured data is probably the most challenging part. Determining the data dimensions and the methodology to measure them limits the analyses that can be conducted.

Stage 3 reports the outcomes. Using unstructured data is expected to produce positive outcomes, that is, strengthening controls or identifying evidence in an investigation. For instance, unstructured data from contracts has been successfully structured to improve the efficiency and effectiveness of gathering cost information (Beaulieu, 2020). However, using unstructured data may contribute nothing, or even negatively, to improving internal controls or an investigation. For instance, bias in the training sample will result in biased results. Potential pitfalls may be avoided by conducting a traditional cost-benefit analysis before using unstructured data.

The proposed framework can guide researchers and practitioners in different roles, including auditors, law enforcement officers, and forensic accountants. What methods could be more helpful in using unstructured data for control? Which methods are more resilient to bias? Which methods provide more explainability? What is the best way to initiate using unstructured data in these processes? How can existing processes be improved?

Using unstructured data for control purposes requires access to the data. Auditors and law enforcement officers usually have the authority to request access to unstructured data. However, researchers face the common challenge of gathering data. Although public unstructured data is readily available, organizational unstructured data is limited for academic research. Unstructured data available for research, like the Enron corpus, is an exception.

UNSTRUCTURED DATA IN THE AIRBUS INVESTIGATION

The proposed framework identifies the stages in the investigation leading to the settlement between Airbus and the enforcement agencies of France, the United Kingdom, and the United States. The information was drawn from publicly available sources. Figure 2 in the Appendix presents the proposed framework on the left with the example of Airbus on the right.

In 2016, the Serious Fraud Office of the United Kingdom started a probe into Airbus' corruption practices (Booth, 2019). Later, France (Wall, 2017) and the United States joined the probe (Department of Justice, 2020). Airbus used intermediaries to bribe government officials and non-governmental airline executives to obtain or retain businesses from 2008 until at least 2015 (Department of Justice, 2020). Airbus entered a Deferred Prosecution Agreement, in which, among other issues, Airbus agreed to open its operations to law enforcement to determine the extent of transactions resulting from bribery and to impose an appropriate fine (Beioley, 2020).

Stage 1 of the framework identifies the goal of using unstructured data for control. For Airbus, the goal was to investigate the extent to which transactions in the revenue cycle breached

anticorruption controls, ultimately estimating the sales that resulted from bribery. Relying exclusively on structured data would be a serious drawback in a corruption investigation that spanned multiple countries and years and was disguised with payments through intermediaries. Unstructured data, such as text documents from emails, presentations, and other communications, could provide evidence not present in structured data.

Stage 2 of the framework identifies the types of unstructured data that can be used and the data analysis techniques. An initial collection of 500 million documents and transactions was identified. The investigation on Airbus considered only internal documents as evidence (Beioley, 2020). The investigation combined structured data from the accounting information systems (entries) with millions of unstructured data items from documents and communications (memos, contracts, spreadsheets, presentations, emails).

Before the analysis, the data was cleaned, duplicates and irrelevant material were eliminated, and the number of documents was reduced from 500 million to 60 million (Dempsey, 2021). In addition, the data had to be secured. Although all business data is private, analyzing data from Airbus required additional steps to protect sensitive military information, such as using computers with no internet connection (air-gapped) and redacting documents (Dempsey, 2021).

Analyzing such a vast number of documents required an automated approach. A machine learning algorithm was trained to identify instances of documents in which potential bribes were paid. Training the algorithm required tagged data from documents where corruption was present (Dempsey, 2021). Although the specific data used for training the algorithm was not publicly disclosed, the U.S. Department of Justice stated that Airbus had provided documents that business partners had used to conceal payments in China with indirect payments to an account in Hong Kong (Department of Justice, 2020).

The analysis defined the dimensions to structure the data. Artificial intelligence was used to extract metadata from each document. (Dempsey, 2021). The algorithm returned scores in different dimensions; the scores were added to determine an overall score. The documents above a defined threshold, about five percent of the total documents, were further investigated by people (Dempsey, 2021). The analysis was determined based on the types of unstructured data available. Regulators approved using machine learning in unstructured data to narrow the investigation (Dempsey, 2021).

The algorithm identified events that seemed out of context, such as payments for sports sponsorships (Dempsey, 2021). Between 2013 and 2015, Airbus paid $50 million to sponsor a team owned by airline company executives (Katz, 2020a). The machine learning algorithm identified communications reflecting bribery that could not be identified with other techniques because the communications used aliases and codes, such as prescriptions from a doctor or paintings from Van Gogh (Katz, 2020b).

Stage 3 reports the outcomes achieved based on the analysis. Using unstructured data enabled investigators to estimate the sales resulting from bribery in the years investigated. The investigation, conducted by more than seventy experts, concluded with a settlement in January 2020. It took four years to complete and resulted in a 3.9 billion-dollar fine for Airbus (Dempsey, 2021).

CONCLUSION

Unstructured data can be used to identify patterns relevant to internal control. The characteristics of unstructured data add unique challenges for its analysis and management. This paper highlights some of the challenges of using unstructured data and proposes a framework for studying the connections between unstructured data and internal controls. The challenges discussed include structurization, variability, veracity, cleaning, tagging, privacy, digitization, bias, and explainability.

The first stage of the framework identifies the goal for using unstructured data for internal control. The goal can be the design and implementation of internal controls or the investigation of internal control breaches. The second stage identifies the unstructured data and the analyses to be conducted. The third stage identifies the outcomes of using unstructured data for control purposes.

Since analyzing unstructured data requires substantial resources, researchers could investigate the extent to which the investment justifies the implementation costs. They could also examine operations' efficiency or effectiveness changes compared to using structured data for internal control.

Unstructured data can also support audits in risk assessment, control testing, and substantive testing. Researchers could investigate whether the analysis of public data regarding the industry and the company is in alignment with the assessment conducted by the auditor and identify the reasons for discrepancies. Regarding the test of controls, researchers could investigate whether the analysis of unstructured data from business manuals, memorandums, or minutes could assist them in identifying relevant changes to controls in a given period. Regarding substantive testing, researchers could also investigate whether the analysis of unstructured data from contracts for sales and expenses agrees with the study of structured data from the accounting information system.

The contributions of the proposed framework should be considered along with its limitations. First, the framework described using unstructured data in the Airbus corruption investigation, which was drawn from secondary sources. The framework has not been tested with a case study using primary sources. Additional research is needed to demonstrate the framework's validity and propose further refinements. Second, although the framework provides a generic list of unstructured data that can be used, the framework may be adapted for particular uses of unstructured data. Despite these limitations, the proposed framework can assist accountants in designing internal controls and investigating control breaches. In addition, the framework can spur research on using unstructured data for internal control.

REFERENCES

ACFE. (2024). *Occupational Fraud 2024: A Report to the Nations*. https://www.acfe.com/-/media/files/acfe/pdfs/rttn/2024/2024-report-to-the-nations.pdf

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, *46*(4), 557-590. https://doi.org/https://doi.org/10.1007/s11747-018-0581-x

Bartlett, H. (2023). *Synthetic Touch: Haptic Data is Revolutionizing Consumer Products*. USC Viterbi. Retrieved November 19, 2024, from https://viterbischool.usc.edu/news/2023/01/synthetic-touch-haptic-data-is-revolutionizing-consumer-products/

Baugess, S. W. (2016). *Has Big Data made us lazy?* Retrieved September 20, 2020, from https://www.sec.gov/news/speech/bauguess-american-accounting-association-102116.html

Beaulieu, P. R. (2020). Contract-Based Cost Analytics. *Journal of Emerging Technologies in Accounting*, *17*(1), 11-19. https://doi.org/https://doi.org/10.2308/jeta-52718

Beioley, K. (2020, February 18). The Airbus case reflects France's changed ways on corruption: Bribery legislation. *Financial Times*, 4.

Booth, J. (2019, May 23). SFO probe into Airbus could be settled this year. *City A.M.*, 6.

Bousquette, I. (2024, Apr 04). Deepfakes Are a New Threat To Finance. *Wall Street Journal*.

Bumgarner, N., & Vasarhelyi, M. A. (2015). Continuous Auditing—A New View. In *Audit Analytics and Continuous Audit: Looking toward the Future* (pp. 3-52). AICPA. https://www.aicpa.org/interestareas/frc/assuranceadvisoryservices/downloadabledocuments/auditanalytics_lookingtowardfuture.pdf

Canelón, J., Huerta, E., Leal, N., Ryan, T., (2020) Unstructured Data for Cybersecurity and Internal Control. Proceedings of the 53rd Hawaii International Conference on System Sciences. 5411-5420. https://core.ac.uk/reader/326835900

Castellanos, S. (2019, Jan 23). Budweiser Maker Uses Machine Learning to Keep Beverages Flowing; AB InBev uses smart sensors to predict equipment malfunctions, reducing factory downtimes. *Wall Street Journal (Online)*.

CCPA. (2020). *California Consumer Privacy Act*. Retrieved September 20, 2020, from https://www.oag.ca.gov/privacy/ccpa

Chocano, C. (2023). The language game. *The New Yorker*, (April 24 & May 1). https://www.newyorker.com/magazine/2023/04/24/how-much-can-duolingo-teach-us

Correal, A. (2024, Nov 19). Insurers Seek To Claw Back $140,000 Paid To Claimants. *New York Times*.

Dempsey, M. (2021). *How to investigate a firm with 60 million documents*. BBC. Retrieved March 1st, 2021 from https://www.bbc.com/news/business-55306139

Department of Justice. (2020). *Airbus Agrees to Pay over $3.9 Billion in Global Penalties to Resolve Foreign Bribery and ITAR Case* https://www.justice.gov/opa/pr/airbus-agrees-pay-over-39-billion-global-penalties-resolve-foreign-bribery-and-itar-case

Dezember, R. (2018, Nov 02). Your Smartphone's Location Data Is Worth Big Money to Wall Street; The phone in your pocket is dishing info on where you spend your time and, likely, money. *Wall Street Journal (Online)*.

Eaglesham, J. (2024, Apr 08). Home Insurers Take to the Sky To Spot Policyholders' Risks. *Wall Street Journal*.

Eifrem, E. (2017). *The Panama Papers: What business can learn from data-driven scoops.* CMSWiRE. Retrieved September 20, 2020, from https://www.cmswire.com/big-data/the-panama-papers-what-business-can-learn-from-data-driven-scoops/

Artificial Intelligence Act (2024). https://artificialintelligenceact.eu/

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137-144. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2014.10.007

GDPR.eu. (2020). *Complete guide to GDPR compliance*. Retrieved September 20, 2020 from https://gdpr.eu/

Greenfield, S. (2018). *New York Times is using Google Cloud to find untold stories in millions of archived photos*. Retrieved September 20, 2020 from https://cloud.google.com/blog/products/ai-machine-learning/how-the-new-york-times-is-using-google-cloud-to-find-untold-stories-in-millions-of-archived-photos

Henry, E., & Leone, A. J. (2016). Measuring Qualitative Information in Capital Markets Research: Comparison of Alternative Methodologies to Measure Disclosure Tone. *Accounting Review*, *91*(1), 153-178. https://doi.org/https://doi.org/10.2308/accr-51161

Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research*, *40*(2), 806-841. https://doi.org/https://doi.org/10.1111/1911-3846.12832

Katz, B. (2020a, Jan 28). Airbus Faces Nearly $4 Billion in Penalties From Corruption Probes; European plane maker's contract dealings have been under investigation by the U.S., U.K. and France. *Wall Street Journal (Online)*.

Katz, B. (2020b, Feb 01). Airbus Settles Bribery Probe for $4 Billion. *Wall Street Journal*.

Knight, W. (2017). The dark secret at the heart of AI. *Technology Review*, *120*(3), 54-61. https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/

Markoff, J. (2011, Mar 05). Armies of Expensive Lawyers, Replaced by Cheaper Software. *New York Times*, 2.

Mayew, W. J., & Venkatachalam, M. (2013). Speech Analysis in Financial Markets. *Foundations and Trends® in Accounting*, *7*(2), 73-130. https://doi.org/10.1561/1400000024

Moffitt, K. C., & Vasarhelyi, M. A. (2013). AIS in an age of Big Data. *Journal of Information Systems*, *27*(2), 1-19. https://doi.org/https://doi.org/10.2308/isys-10372

Murgia, M. (2017, Feb 13). AI robot sleuth helps SFO crack crime: Technology. *Financial Times*, 19.

Mwansa, G. (2015). *Exploring the development of a framework for agile methodologies to promote the adoption and use of cloud computing services in South Africa* University of South Africa].

Nagarajan, A. D., & Overbeek, S. J. (2018). A DevOps Implementation Framework for Large Agile-Based Financial Organizations. In H. Panetto, C. Debruyne, H. A. Proper, C. A. Ardagna, D. Roman, & R. Meersman, *On the Move to Meaningful Internet Systems. OTM 2018 Conferences* Cham.

Nishihara, N. (2018). *"Just Walk Out" with Amazon and the Internet of Thinking*. Retrieved September 20, 2020 from https://www.accenture.com/us-en/blogs/technology-innovation/nishihara-future-retail-internet-of-thinking

Ramachandran, S. (2019, May 30). What If Google's 'Knowledge Panels' Insist You're Dead? Or Married? Or French? People who find themselves in 'knowledge panels' search for ways

to fix errors; 'They have more authority over my life story than I do.'. *Wall Street Journal (Online)*.

Rocco, T. S., & Plakhotnik, M. S. (2009). Literature Reviews, Conceptual Frameworks, and Theoretical Frameworks: Terms, Functions, and Distinctions. *Human Resource Development Review*, *8*(1), 120-130. https://doi.org/https://doi.org/10.1177/1534484309332617

Rydning, J., & Shirer, M. (2021, Mar 24). Data Creation and Replication Will Grow at a Faster Rate Than Installed Storage Capacity, According to the IDC Global DataSphere and StorageSphere Forecasts. *Business Wire*. https://www.businesswire.com/news/home/20210324005175/en/Data-Creation-and-Replication-Will-Grow-at-a-Faster-Rate-Than-Installed-Storage-Capacity-According-to-the-IDC-Global-DataSphere-and-StorageSphere-Forecasts

Scudellari, M. (2018). Meet the E-Nose That Actually Sniffs. *IEEE Spectrum*. Retrieved April 5, 2021, from https://spectrum.ieee.org/the-human-os/biomedical/devices/meet-the-enose-that-actually-sniffs

SEC. (2017). *SEC Charges Mexico-Based Homebuilder in $3.3 Billion Accounting Fraud*. Retrieved September 20, 2020 from https://www.sec.gov/news/pressrelease/2017-60.html

SEC. (2022, Dec 14). *SEC Charges Eight Social Media Influencers in $100 Million Stock Manipulation Scheme Promoted on Discord and Twitter* https://www.sec.gov/newsroom/press-releases/2022-221

Stewart, T. R. (2015). Data Analytics for Financial Statement Audits. In *Audit Analytics and Continuous Audit: Looking toward the Future* (pp. 105-128). AICPA. https://www.aicpa.org/interestareas/frc/assuranceadvisoryservices/downloadabledocuments/auditanalytics_lookingtowardfuture.pdf

Tugend, A. (2019). Exposing the Bias Embedded in Tech. *New York Times (Online)*. https://www.nytimes.com/2019/06/17/business/artificial-intelligence-bias-tech.html

Van den Bogaerd, M., & Aerts, W. (2011). Applying machine learning in accounting research. *Expert Systems with Applications*, *38*(10), 13414-13424. https://doi.org/https://doi.org/10.1016/j.eswa.2011.04.172

Wakefield, J. (2021). *AI: Ghost workers demand to be seen and heard*. BBC. Retrieved March 28 from https://www.bbc.com/news/technology-56414491

Wall, R. (2017, Mar 16). Airbus Says French Fraud Investigators Open Probe Into Possible Misconduct; Company says France is investigating alleged missteps already under scrutiny by U.K. *Wall Street Journal (Online)*.

APPENDIX

Type of unstructured data	Definition. Data that represents …
Text	Semantic meaning in writing
Image	A visual representation of something real or imagined
Video	A combination of successive images and sounds
Sound (audio)	Vibration waves of noise or speech

Scent	Airborne chemicals that evaporate from a source
Taste	Chemical composition of a substance
Texture	Softness, roughness, and other tactile sensory characteristic
Other physical phenomena	Air pressure, temperature, movement, magnetic fields, and other signals captured by sensor devices

Table 1: Types of unstructured data

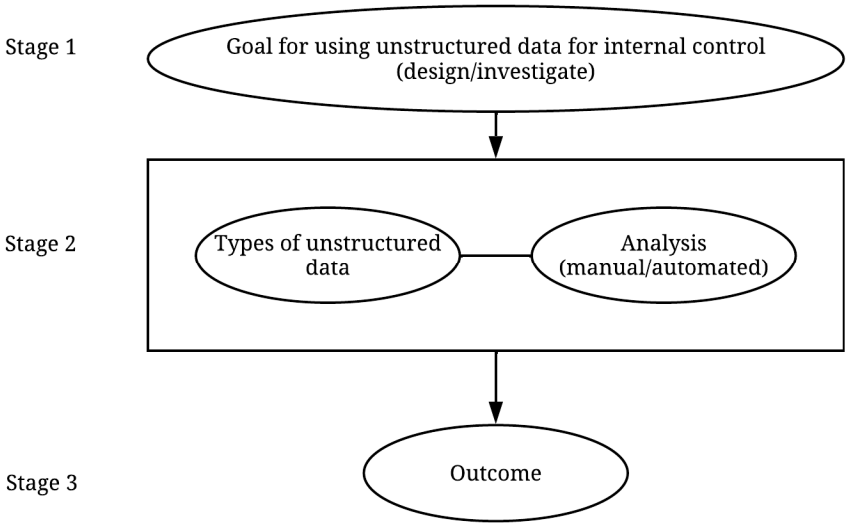

Figure 1: Proposed framework for the use of unstructured data for internal control

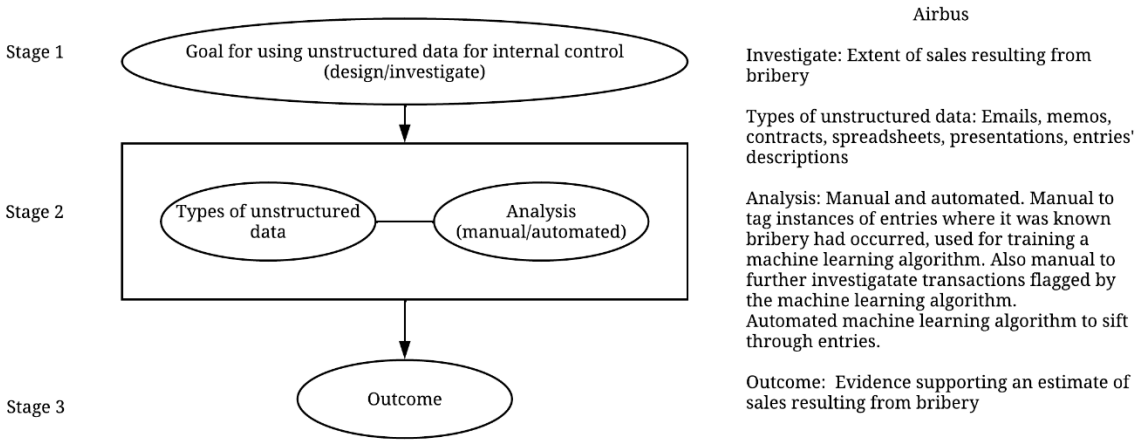

Figure 2 Proposed framework identifying the stages for the use of unstructured data in the Airbus corruption investigation

T
R