Science helps auditors take on the data challenge

By Professor Jan Scholtes, ZyLAB and University of Maastricht, Youri van der Zee, University of Amsterdam, and Marcel Westerhoud, Ebben Partners

Source: geralt/Pixabay

Digitalisation is a key aspect the 2021-25 ECA Strategy deals with. Digitalisation links up to all three strategic goals and some key enablers, including making enhanced use of data and IT tools and technologies and is essential for strategy *implementation*. The audit profession finds itself at a crossroads in auditing by humans and the use of machine learning techniques and artificial intelligence to prevent errors and combat fraud. Public auditors can join hands with scientists to utilise advanced digital techniques to optimise the audit work and increase its impact. This will sometimes require diving in at the deep end when it comes to the techniques that can be used. Professor Jan Scholtes from the Department of Data Science and AI of the University of Maastricht and Chairman of ZyLAB, Youri van der Zee from the University of Amsterdam and Marcel Westerhoud from Ebben Partners, look at the example set by fraud investigations to show how the audit sector could benefit from AI and achieve some strategic goals.[1]

The data conundrum

In today's world, auditors, compliance officers and fraud investigators face an overwhelming amount of digital information that can be reviewed. In the majority of cases, they do not know beforehand what exactly they are looking for, nor where to find it. In addition, individuals or groups may use different forms of deception to hide their behaviour and intentions, varying from using complex digital formats[2], rare languages[3] or by using code words.[4] Effectively, this means fraud investigators are looking for a needle in the haystack without knowing what the needle looks like.

1 The authors are grateful for the extensive support obtained for this research from ZyLAB Technologies BV and Ebben Partners BV, both based in the Netherlands.

2 Such as an email with a ZIP attachment that contains non-searchable TIFF or PDF documents or even audio recordings.

3 Google translate makes it very easy to translate messages into rare languages, or even into artificial languages such as Star-Trek's Klingon, thereby effectively hiding the content for tooling that only searches for words in more common languages.

4 Van der Zee, Y., Scholtes, J. C., Westerhoud, M., and Rossi, J., *Code Word Detection in Fraud Investigations using a Deep-Learning Approach*, arXiv e-prints, art. arXiv:2103.09606, March 2021.

Using technology is essential to address the – hopefully – high strategic ambitions auditors and fraud investigators have regarding such large digital data collections. The main problem with using such technology is to balance finding what is really suspicious from finding too many false positives, which would create too much work for auditors or victimise innocent individuals.[5]

In today's digital world, both auditors and fraud investigators have to sift through an ever-increasing mass of unstructured data when looking for valuable information or even direct evidence. To do so, one of the frequently used tools is *eDiscovery*. Many such AI-techniques are primarily aimed at isolated topics, such as sentiment and emotion analysis, assisted review (searching using machine learning), Named Entity Recognition (NER), or community detection, to organise data for better anomaly detection and to help auditors and investigators find answers to common questions more efficiently.

However, the application of such – rather promising – AI techniques is often ad hoc and not guided by an overall strategy or vision, and, where it is, it is rather focused on the 'what' in more abstract terms, and less on the 'how' in more concrete terms. To remedy this, we propose a model that gives such AI-techniques a more logical and organised role in audits and fraud investigations. Let us first have a look at how a typical auditor or investigator approaches a case. This we can do by examining three building blocks that provide a basis where we can `plug in' an AI-technique and use the outcome as a diagnostic variable in the investigated case.

These building blocks are:

- the Fraud Triangle;[6]
- the six `golden' investigation questions;
- the Theory of the Analysis of Competing Hypotheses.[7]

These blocks allow us to deconstruct a (partial) investigation question into a number of tasks that can each be executed by a specific search, text mining or a machine learning algorithm. To explain what these three building blocks are exactly, how they can be combined, and how AI-techniques can be used in a more structured manner using this overall framework, we should first look at the deep learning algorithms. More and more, algorithms have become a digital tool in many areas and thereby become more and more part of the auditor's realm. Also in natural language processing, they have created revolutionary breakthroughs.

Deep learning for Natural Language Processing (NLP)

The ability to model the context of text is vital to avoid finding too many false positives in audits and fraud investigations. Algorithms that enable us to properly understand such context have greatly advanced in recent years due to progress in using deep learning algorithms for highly context-sensitive Natural Language Processing (NLP) tasks, such as machine translation, human-machine dialogues, named entity recognition, sentiment detection, emotion detection or even complex linguistic tasks such as co-reference and pronoun resolution.

The above-mentioned progress comes from the development of what is known as transformer architecture. Transformer models are based on large pre-trained recurrent neural networks that already embed significant amounts of linguistic knowledge and which can be fine-tuned for specific tasks requiring a relatively small amount of additional training.

5 What is known as Bonferroni's Principle is interesting in this context, which states that if you look for certain types of data, you will certainly find such patterns, even if their occurrence is caused by chance. In large data sets, one has a higher probability of finding such suspicious patterns, which may in fact occur less frequently than chance would dictate. This will then lead to wrong conclusions.

6 Cressey, D., *Why do trusted persons commit fraud? A social-psychological study of defalcators,* Journal of Accountancy, 92:576, 1951.

7 Heuer, R. J. Psychology of intelligence analysis. Center for the Study of Intelligence, 1999.

A fundamental benefit of transformer architecture is the ability to perform Transfer Learning.[8] Traditionally, deep learning models require a large amount of task-specific training data in order to achieve desirable performance (billions of data points required to fine tune hundreds-of-millions of neural interconnections). However, for most tasks, we do not have the amount of labelled training data required to train these networks. By pre-training with large sets of natural text, the model learns a significant amount of task-invariant information on how language is constructed. With all this information already contained in these models, we can focus our training process on learning the patterns that are specific to the task in hand. We will still require more data points than required in most statistical models (typically 50-100k based on our experience in earlier NLP deep learning projects), but not as much as the billions required, should we start the training of the deep learning models from scratch.

Transformers are able to model a wide scope of linguistic context, both depending on previous words, but also on (expected) future words. They are, so to speak, more context sensitive than models that can only take past context into consideration. In addition, this context is included in the embedding vectors, which allows for a richer representation and more complex linguistic tasks.

Currently, the Bidirectional Encoder Representations from Transformers (BERT), released by Google AI Language is considered to be the state-of-the-art language representation model. Another successful application of transformers can be found in OpenAI's Generative Pre-trained Transformer 3 (GPT-3) project, based on 175 billion machine learning parameters. The quality of GPT-3 is so high, that it is almost impossible to distinguish text written by GPT-3 from text written by humans.[9]

For many linguistic tasks, both GPT-3 and BERT outperform humans both in speed, scalability but also in quality. This progress allows us to use these new models to analyse large volumes of textual information in audits and investigations and identify sentences and paragraphs that provide relevant information.

Organising extracted information for auditors and investigators

How can extracted linguistic patterns be organised to be useful to auditors and investigators? This is where the fraud triangle, Golden W questions and the analysis of competing hypotheses come in, and which are relevant for both auditors and fraud investigators: for the latter in view of their detecting capabilities, for the first in view of their systemic assessment of whether a system has enough preventive elements built in to prevent fraud from happening in the first place.

Fraud Triangle

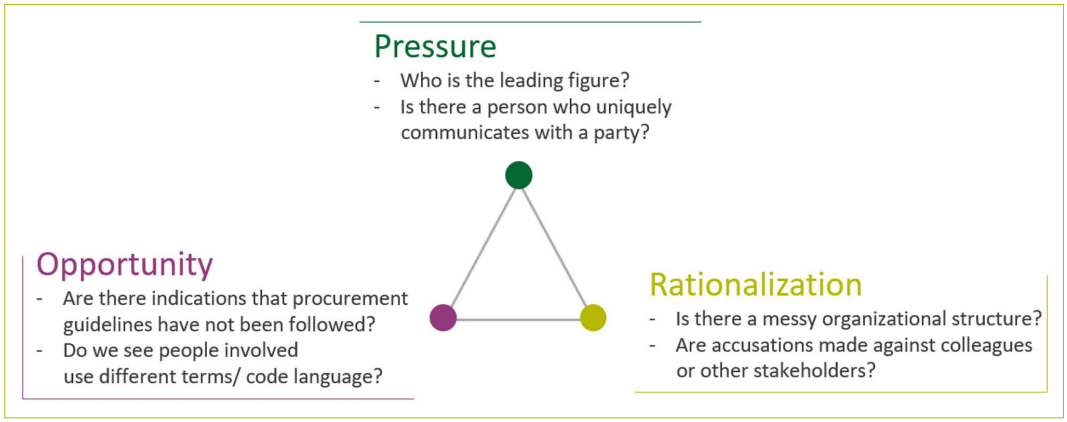
A widely used method to model organisational fraud risk is the fraud triangle (see **Figure 1**). Just as fire requires fuel, oxygen and a spark, in the case of a fraud there are also three ingredients which are essential: the perpetrator must have a motive to commit fraud, the situation must provide an opportunity, and the fraudster must find a way for himself/herself to rationalise his/her dishonesty. Motives can vary from perverse financial incentives to personal problems, such as financial need or addiction. All these can be referred to as pressure. The opportunity is often related to the control environment of the victim organisation: weak controls and tone at the top. Finally, the rationalisation relates to the perceived relation between the fraudster and his environment. This relation provides the internal justification of a fraud: `I was mistreated', `everybody does it', etc.[10]

8 Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805, 2018.

9 Floridi, L. and Chiriatti, M., *Gpt-3: Its nature, scope, limits, and consequences. Minds and Machines*, 30:681{694, 2020.

10 Kassem, R., and Higson, A., *The new fraud triangle model*, Journal of emerging trends in economics and management sciences, 3(3):191{195, 2012.

Figure 1 - the Fraud Triangle

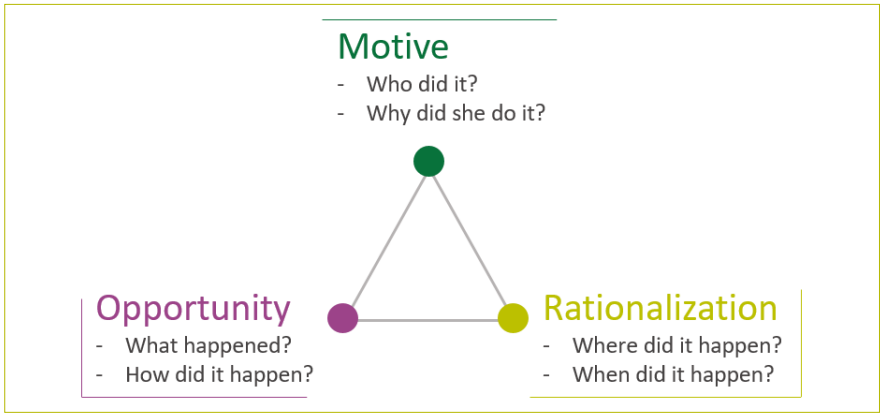


Text mining technology, in particular machine learning, can be used to detect text sentences that indicate one of these three components of the fraud triangle. For example, by showing a machine learning algorithm such as BERT several thousand sentences related to Pressure, Opportunity or Rationalisation, it can automatically recognise similar language in other contexts.[11].

Six golden investigation questions

Usually the fraud triangle is used as a risk tool. But we can also use the model as part of our investigation framework. To do this, we propose a relationship between the three edges of the fraud triangle and the six golden questions that lie at the basis of almost every fraud investigation: who, why, what, how, when and where. Answering these questions will almost automatically lead to the construction of a possible fraud scenario and fill the elements of an evidence matrix. If one needs to know what the motives of a fraudster are, one needs to know *who* did it and *why*. If one needs to know about possible fraud opportunities, questions about the *what* and *how* need to be answered. And finally, for the rationalisation component of the fraud triangle, situational variables are important, in particular: *where* and *when* (see **Figure 2**).

Figure 2 – Combining the Fraud Triangle with the golden investigation questions



Answers to (variations of) these questions produce evidence items that can populate elements of the evidence matrix. The Who questions can be addressed by a well-established technique such as Named-Entity Recognition to detect Person, Company, Organisation; the Where can be answered using the same technique detecting Localities such as City, Country, Continent, etc. When can be extracted by detecting time notions such as Date, Time, Month, Year, Holiday, etc.[12]

11 Soares, L. B., FitzGerald, N., Ling, J., and Kwiatkowski, T., *Matching the blanks: Distributional similarity for relation learning*, 2019.

12 Ehrmann, M., Romanello, M., Fluckiger, A., and Clematide, S., *Extended overview of clefhipe 2020: named entity processing on historical newspapers;* Cappellato, L., Eickho, C., Ferro, N., Neveol, A. (eds.) *CLEF 2020 Working Notes*, Conference and Labs of the Evaluation Forum CEUR-WS, 2020.

Detecting answers to the Why question is harder, but empirical data has shown that the answer to this question can often be found by detecting communication with high levels of sentiments or emotions. Using a similar approach to detect the elements of the fraud triangle, sentiments and emotions can be identified deploying a deep learning approach.[13] A corresponding empirical approach can be used to extract information on the How and What questions, using methods such as Topic Modeling,[14] by deriving communities,[15] or by combining the above mentioned extracted information in more complex analysis such as Who-Why, What-When, etc.[16]

As mentioned earlier, while deep learning can provide assistance in allocating linguistic patterns to the right context, it cannot prevent the generation of many false-positives, which causes enormous amounts of irrelevant work. A few false positives are acceptable, especially in the light of the need not to overlook irregularities, but an overload of thousands of false positives is a professional nightmare for every auditor or investigator, as nothing is more frustrating than having to chase thousands of false leads, let alone that we do not have the time or capacity for this. Intelligence services have long struggled with this problem as well. In the 1970s, the Central Intelligence Agency (CIA) developed the Analysis of Competing Hypotheses (ACH) to address this problem, which will be explained in the subsequent paragraph.

Analysis of Competing Hypotheses (ACH)

For each type of crime, what is called an evidence matrix can be constructed holding key items to be proved. For instance, in the case of a murder one needs a victim, a murder weapon, a motive, a crime scene, intent, etc. These items relate to the above-mentioned Golden Investigation Questions. Instead of using a simple numeration of such items, we can use a more advanced model of an evidence matrix as developed in the 1970s by Richard Heuer.[17] This methodology was named 'Analysis of Competing Hypotheses' (ACH). It is based on the evaluation of various competing hypotheses, given a set of information items (i.e. evidence). This involves the following step-by-step approach as presented in **Table 1**.

13 Gerolemou, Z. and Scholtes, J., *Target-based sentiment analysis as a sequence-tagging task*, Benelux Articial Intelligence Conference, Brussels, November 2019.

14 Tannenbaum, M., Fischer, A., and Scholtes, J. C., *Dynamic topic detection and tracking using nonnegative matric factorization*, Benelux Articial Intelligence Conference (BNAIC), Hasselt, Belgium, November 5-6, 2015.

15 Helling, T., Takes, F., and Scholtes, J.C., *A community-aware approach for identifying node anomalies in complex networks,* The 7th International Conference on Complex Networks and Their Applications December 11-13, 2018, Cambridge, United Kingdom, 2018.

16 An overview and examples of such techniques can be found in Smeets, J., Scholtes, J., Rasterfo, C., and Schravemaker, M. Smtp, *Stedelijk museum text mining project,* Digital Humanities Benelux (DHBenelux), Luxemburg, June, 2016; Scholtes, J. C., *Text-mining and ediscovery for big-data audits*, ECA Journal. No 1, pp. 133 -140, 2020.

17 Heuer, R. J., *Psychology of intelligence analysis*, Center for the Study of Intelligence, 1999.

Table 1 – Step-by step outline of Analysis of Competing Hypotheses (ACH)

STEP	ACTION
Identify the possible hypotheses to be considered	Use a group of analysts with different perspectives to brainstorm the possibilities
Listing pros and contras	Make a list of significant evidence and arguments for and against each hypothesis
Prepare a matrix with hypotheses across the top and evidence down the side	Analyse the 'diagnosticity' of the evidence and arguments{that is, identify which items are most helpful in judging the relative likelihood of the hypotheses
Refine the matrix	Reconsider the hypotheses and delete evidence and arguments that have no diagnostic value
Draw tentative conclusions about the relative likelihood of each hypothesis	Proceed by trying to disprove the hypotheses rather than prove them
Analyse how sensitive your conclusion is to a few critical items of evidence	Consider the consequences for your analysis if that evidence were wrong, misleading, or subject to a different interpretation
Report conclusions	Discuss the relative likelihood of all the hypotheses, not just the most likely one
Hooks for future observations	Identify milestones for future observation that may indicate events are taking a different course than expected

Source: J. Scholtes and others

The `weighted inconsistency score' (see **Table 2**) provides a measure for the plausibility of a specific hypothesis, given a set of evidence items in terms of credibility and relevance. Lower values of the scores correspond with a lower plausibility of the hypothesis. The numerical values are determined based on a simple lookup table. These initial values do not represent probabilities, but they can be normalised towards a [0-1] range, giving a normalised confidence score. Combining confidence scores can be done by multiplication. There are obvious issues with this approach, as the use of multiplication in the calculations presumes complete independence of the underlying hypothesis, which is off course not always the case. In addition, the values are manually assigned, which leads to bias risks. But for now, this is what is used.[18]

Table 2 – A weighted inconsistency score

	Type	Credibility	Relevance	H: 1	H: 2	H: 3
				Hypothesis1	Hypothesis2	Hypothesis3
Weighted Inconsistency Score				-1,707	-2,828	-1,414
Evidence item1		LOW	HIGH	I	C	C C
Evidence item2		HIGH	MEDIUM	NA	I	I
Evidence item3		MEDIUM	LOW	I	I I	C C

Source: J. Scholtes and others

I Inconsistent, II Strongly inconsistent, C Consistent, CC Strongly consistent, NA Not applicable.

18 See this document for a technical discussion on the use of 'weighted inconsistency score.'

Now we have a conceptual model we can systematically inject the results of a various set of AI-methods into, for example in a case of the investigation of a possible purchasing scheme. Typical for this scheme is the incidence of collusion between perpetrators.

Several of the evidence components listed above, can now be filled automatically with possible candidates using the text mining techniques we referred to earlier. Named Entity Extraction in combination with 'inconsistency scores' with basic linguistic contextual analysis can provide candidates for the Who, Where, and When questions. Sentiment and emotion mining can identify the textual sections containing and providing valuable insights into Why something is done and Who is driving the actions. Topic modelling can be used for the What question, and combinations of the above and the extraction of more complex (dedicated) patterns can answer the How questions.

Examples of typical investigation questions, relevant AI techniques that can identify potential answers to such questions and more detailed facts are listed in **Table 3**.

Table 3 – Examples of W-Questions that can be used to validate competing hypotheses

QUESTION	AI METHOD	RESULT
Who is likely to be involved in the scheme?	Detect Who's and Community Detection	Close relation between employee A, employee B and supplier C
Where did it happen?	Detect Where's and Geomapping	Location of Meetings and Transactions C
When Did it Happen?	Detect When's and Time Lines	Time of Meetings and Transactions
Why and How Did it Happen?	Detect Sentiments and Emotions	Motivations, Responsible Individuals, Modus Operandi
What happened?	Topic Modelling	Overview of all activity

Source: J. Scholtes and others

This information can then be used to construct elements of the scenario when added to the ACH-matrix: instead of using the actual extracted sentences, it is better to use a straightforward quantitative analysis, such as the total number of occurrences of certain relations, the above or below average percentage, or the nature of the polarity of the emotions and sentiments.

The extraction of the above entering this result as an evidence item in an ACH-matrix would look like what is shown in **Table 4**.

Table 4 –ACH-matrix of competing hypotheses

	Credibility	Relevance	H1: A is acting alone	H2: A and B are colluding
145 emails scheduling meetings between A and B	Medium	Medium	II	CC
87 agenda items indicating meetings	High	High	II	CC
237 Phone Records indicating direct conversations	High	Medium	II	CC
90 more email between A and B than average	Medium	Low	I	C
Friendly emotions and sentiments detected in all emails between A and B	Medium	High	I	C
B consistently bcc-ed in communication between A and others	High	High	II	CC

Source: J. Scholtes and others

I = Inconsistent, II = Strongly inconsistent, C = Consistent, CC == Strongly consistent, NA = Not applicable.

This is an example of how the ACH-matrix can be used to create a complete mapping of the golden W questions initially to hypotheses and ultimately to evidence items that can be generated from AI-techniques. In the example above this mapping relates to the Who-question - in this case: 'is A acting alone, or colluding with B?' The first step is to establish a meaningful relationship between a relevant item of evidence (in relation to the hypotheses) and the output of the AI-process. 145 emails related to scheduling meetings and 87 meetings found in the mutual agendas (extractions of email communication is medium credible or relevant, as meetings can also be arranged by assistants), lead to a strong inconsistency with the hypothesis that A is acting alone and to a strong consistency that A and B are colluding. Many direct phone records, friendly emotions in email exchange and the consistent use of blind carbon copies (bcc) to B when A emails C, are all indications that A and B are colluding.

AI deep learning techniques create new challenges, not only technical ones

We have described how new deep learning techniques are able to capture richer contextual representations which can be used in audits and fraud investigations. With our proposed framework, we aim to bring structure to the search space auditors and fraud investigators have to explore for anomalies and irregularities. With the employment of machine learning techniques, this search space is reduced and made insightful, and hopefully helpful for public audit institutions, such as the ECA, to address the 'how' of reaching its strategic objectives, in particular in relation to fraud prevention and detection.

At the same time, our proposed framework offers cohesion to the collection, classification and weighting of evidence that is collected via AI-methods. We think it is possible to automatically organise investigative data, so it is easier for auditors and investigators to find answers to typical investigative questions, without being overwhelmed with non-relevant information or false signals. In the near future, further research is planned to identify more relevant evidence items which have a discriminatory relation to a fraud scenario and which can be obtained by an appropriate AI method. In our proposed model, these evidence items all are formatted as an answer to one or more variants of the six golden investigation questions. With an adequate amount of these 'triples' (scenario-evidence-AI-method) we expect that many investigations can benefit significantly in terms of efficiency and quality. Another topic for further development is the automation of applying weightings in terms of relevance and credibility to the output of the AI-method and subsequently inserting consistency values into the ACH-matrix.

Many of the algorithms used are language dependent. In a European context, that means that one of them should support 30+ languages in order to be useful for European auditors or investigators, and even up to 40 if Chinese, Russian, Turkish, Arabic, Japanese, Korean and Hindi are included. The future will therefore call for automatic techniques to transfer algorithms and classifiers automatically from one language to others. This is also a major topic of interest to the research community.

Now that we live in a post COVID-19 world, the need to collaborate on large case files in a 'working from home' situation has complicated the audit and investigation process significantly. How to share such large case files in a secure way has become a daily challenge. Especially in the light of a strict General Data Protection Regulation, there are issues relating to privacy, data protection and cyber security. Without a secure digital platform and a variety of tooling, sharing is not possible. Once such a digital platform is in place, why not benefit to the maximum from all possibilities? This is a question auditors and investigators should also be asking themselves.