Continuous Process Mining Monitoring

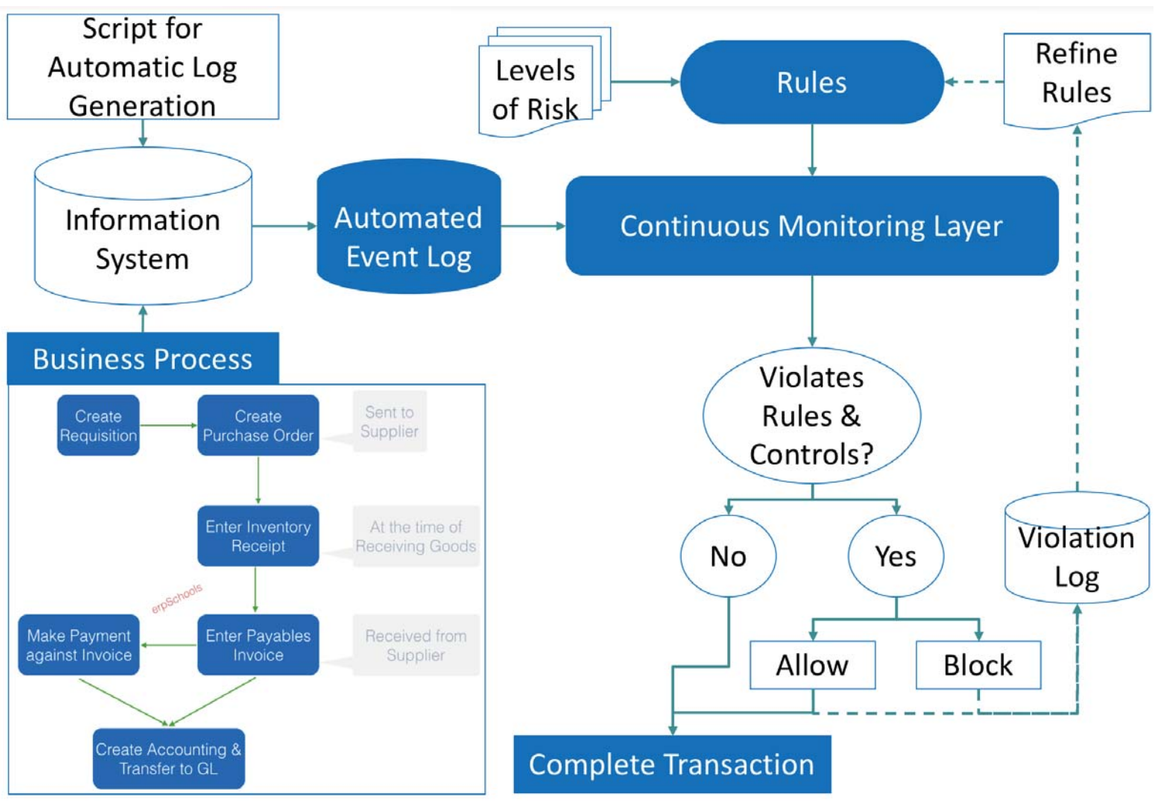
Abdulrahman Alrefai and Miklos Vasarhelyi

Introduction

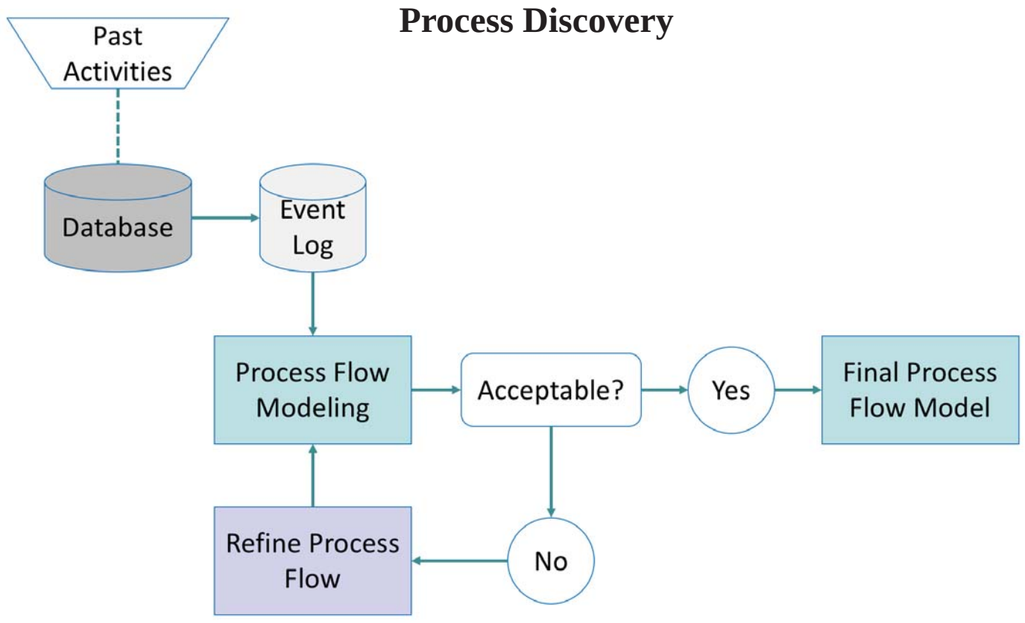
Traditionally, the testing of controls has been performed on a retrospective and cyclical basis, often many months after business activities have occurred. The testing procedures have often been based on a sampling approach and included activities, such as reviews of policies, procedures, approvals, and reconciliations. With today's real-time economy and the advancements in technology, it is recognized that this approach only offers auditors a narrow scope of evaluation, and is often too late to be of a real value to business performance or regulatory compliance.

The aim and contribution of this paper are developing a novel approach for assurance that combines the advantages of continuous monitoring with those of process mining. Auditors can actively detect and investigate deviations and exceptions as they occur along the transaction process by continuously monitoring business process controls and testing transactions, rather than react after the exceptions have long occurred. Any transaction that violates a set of business rules would be intercepted or flagged by the system until investigated by an auditor. This continuous monitoring using rule-based process mining approach provides a high level of assurance about the operating effectiveness of controls throughout a business process. Basically, this study is attempting to answer the research question of how can the time delay between the occurrence of a business operation related event and its analysis be reduced.

General Framework



Process Discovery



Methodology

The architectural methodology that is used for implementing continuous monitoring with process mining techniques is based on implementing an abstracted layer on top of the business process that would continuously monitor the activities throughout a transaction and prevent or flag any violations (Vasarhelyi et al. 2004). The event stream of the information system is used as an input for the monitoring layer, which consists of an adapted rule-based process mining technique. So, instead of relying on "after the fact" process mining techniques, the system would flag any transaction that does not conform with the approved model for that business process. Hence, logs can be automatically generated and process instances are automatically mined on a continuous basis for deviations and assuring compliance.

Framework

1) Process Discovery

Prior to the implantation of a continuous monitoring layer, there needs to be a clear understanding of how transactions of the underlying business process are being conducted. Auditors examine all past transactions of a current business process to establish the path, and establish the final process flow model.

2) Automatic Log Generation

To insure a continuous flow of data into the continuous monitoring layer, a script is written that would extract the required information from the already determined tables in the information system as the activities are being performed for a transaction, and combine them to automatically and continuously generate the log. Continuous data collection is a necessity for continuous auditing and without a continuous feed of data, continuous auditing cannot be achieved.

3) Relevant Rules Identification

Rules have to be defined to compare each activity throughout a transaction against it. They need to cover many different risk scenarios that a company might face. In addition, rules are based on different levels for risk assessment, such as industry risk, company specific risk, and business process risk

4) Continuous Monitoring Layer

The design of the system architecture is based on an abstracted layer placed on top of the business process. The layer stores the ideal model for the business process and the rules already defined. This later tests each transaction against prescribed model and defined rules to catch violations.

Rule Pattern	Example from the P2P process
An activity of type *a1* must be performed at least once	A *sign* activity must be performed at least once
If an activity of type *a1* is performed then an activity of type *a2* must be performed	If a *goods receipt* activity is performed then an *invoice receipt* activity must be performed
An activity of type *a1* must be started/ completed before/ after/on *t* time units	A *Sign* activity must be started before date of *goods receipt*
The value of event data type *a1* is equal to the value of event data type *a2* and *a3*	The values of *Purchase order*, *goods receipt*, and *invoice receipt* must match before the corresponding invoice can be paid

RUTGERS
Rutgers Business School
Newark and New Brunswick

Audit Firms' Reliance on Internal Controls: Two Case Studies from Multinational Firms

Ahmad AlQassar and Gerard Brennan

Abstract

This essay explores external auditors' utilization of continuous auditing tools provided by the internal audit department via two recent case studies. Organizations continuously enhance their processes via their systems to achieve targets more efficiently and effectively. Investments in organizations' systems resulted in the production of real time financial information and sophisticated internal controls to monitor them. Nonetheless, the external auditing processes did not witness a similar development. Even though there is no regulatory preclusion from leveraging automated audit tools, the approaches and techniques used in current engagement are considered relatively outdated (AICPA, 2012; M. G. Alles, 2015; Manson, McCartney, & Sherer, 1997). The cases and discussion highlight the probable barriers to realizing the full potential of technology in an audit environment and propose ways to move the profession forward.

Objective and Motivation

The objective of this paper is to explore how audit firms incorporate evidence generated from internal controls in their audit procedures. We also aim to discuss some barriers that preclude change and propose ways to move forward

Barriers

In this section we aim to explore some barriers that impede the reliance on internal control evidence in audit procedures. The barriers are many and singling or prioritizing one over the other is nontrivial. Nonetheless, we are concerned with presenting the various barriers based on literature and practice. The next section presents some barriers that may be contributing to the current situation. The points are sorted based on the three major players: Auditors, Auditees, and Standard Setters.

Audit Firms

- IT-related activities are sophisticated
- Dilemma of exposing overlooked cases in the past
- Profitability of the firm might be effected
- Techniques learned and investments made may not be replicated in other engagements

Auditees

- Protective of their data
- The driver of technology utilization is the demand for it rather than the supply of technology

Standard Setters

- No professional auditing guidance on both the theory and practice of advanced methods in auditing like data analytics and CA/CM (Byrnes et al. 2015)
- The vagueness of standards and guidance in that area might dissuade both auditors and auditees from moving forward

Case Study 1

The first case involves a large multinational company which has an extensive financial services arm in support of sales and internal financing. They developed a capable continuous monitoring solution that provides assurance and monitoring for more than 250 controls related to operation and compliance on a continuous basis. The continuous monitoring tool was fully accredited by both the internal audit staff and the external auditors for all key IT general controls (ITGC), which helped assure that IT application controls, analytics, and monitoring frequency could not be compromised. Thus, auditors and the company could rely on the assurance provided by the tool. The external auditors proceeded to ask for non-statistical samples from the control system even though the system reports documented that the 250+ controls ran during the exposure period of the audit and that all identified anomalies where remediated and documented.

Case Study 2

The second case involves a large multinational IT service provider and their external audit provider delivering a Third Party Assurance Type II audit. The service provider had three consecutive years of qualified SSAE-16 reports for failures identified via non-statistical sampling. The deficiencies identified were different each year but were mostly in the areas of missed security updates, patching, and network level version upgrades on servers in some of their data centers. Recognizing that using manual identification and remediation methods to identify and update more than 7000 servers is nearly impossible, the service provider developed and purchased a set of CA/CM tools with analytics that monitor all 7000+ servers continuously and automatically install updates and patches for all servers as required. However, the external auditors were unwilling to leverage the tools the service provider already had in place and that were fully accredited .

Conclusion

It is evident that technologies such as CA/CM and analytics can provide a greater level of assurance and deliver it at a faster rate. It is essential for both the audit standard setters and practitioners to keep up with the new business landscape. The need for a change in standards goes beyond replacing sampling and encouraging population-based monitoring. Standards need to incorporate agile, robust, quantitative, and qualitative audit processes that are able to detect more anomalies and deficiencies. Furthermore, standards need to assure that appropriate management judgements are made to remediate and report such issues.

While our paper provides some insight, further research is definitely needed in order to fully explore and explain the presented phenomenon. Future research may look into more definitive answers as to why this phenomenon is so entrenched in the modern accounting practice, and how practitioners can alter this behavior.

RUTGERS
Rutgers Business School
Newark and New Brunswick

Duplicate Payment Prioritization in the Government Sector: A Case Study for a U.S. County

Andrea Rozario, Hussein Issa

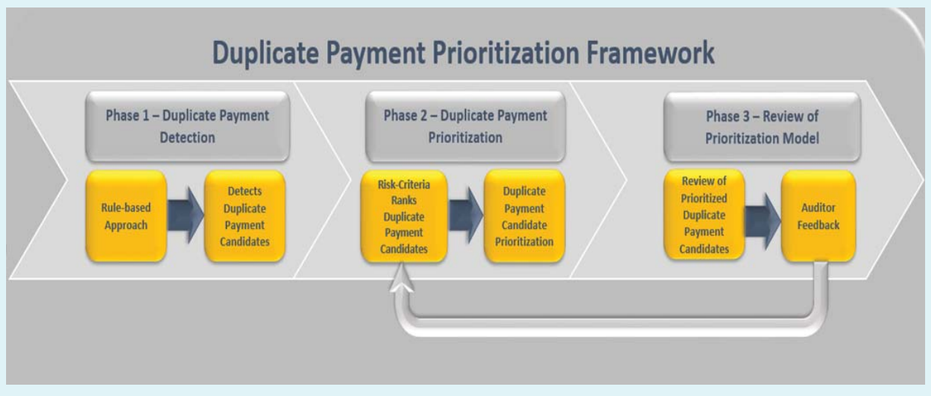
Introduction

The problem of duplicate payments has yet to be examined in the governmental accounting literature. Despite the importance of this problem in the public sector, the extant literature is surprisingly scarce. Federal agencies reported duplicate payment amounts for various government funded programs in fiscal year 2015 and 2016 to be more than $120 billion (Payment Accuracy, 2016 and 2017). The U.S. government has suffered significant financial losses due to the problem of duplicate payments. Hence, because there is an increasing trend of duplicate payments that could be the result of fraud or unintentional error, it is crucial for government auditors to detect these payments in a timely manner.

There have been numerous research studies that develop and implement continuous auditing (CA) methodologies that improve audit efficiency and assist in the timely identification and detection of exceptions (e.g. Vasarhelyi and Halper, 1991; Alles, Brennan, Kogan, and Vasarhelyi, 2006; Alles, Kogan, and Vasarhelyi, 2008), however, these studies have emphasized the implementation of CA systems by internal auditors in the private sector (Vasarhelyi, Alles, Kuenkaikaew, and Littley, 2012). As government entities have been pressured to provide timely and reliable information (Citizant, 2014; Kozlowski, 2016), it is reasonable for them to report government data that could facilitate armchair audits on a continuous basis. Therefore, with the digitization of business processes across public and private entities, it is imperative to examine the benefits CA can provide in the public sector. The purpose of this study is to apply a CA methodology that can effectively prioritize duplicate payments by ranking more suspicious duplicate payment candidates according to risk-based criteria.

Methodology

The figure below illustrates the conceptual prioritization framework. The framework consists of three phases. In Phase 1 a continuous auditing application performs the duplicate detection test by performing a 2 way match (same vendor, same amount). In Phase 2, the identified duplicate instances are prioritized using risk-based criteria including “frequency of user”, “frequency of vendor” , “materiality”, “off-business hours”, “multiple invoices per week” and “size of duplicate set”. Subsequent to the computation of the relative/absolute values and weights for the prioritization criteria presented above, the summation of the values for each criterion is computed to calculate the composite score of the duplicate candidate. The composite score for the duplicate candidate set is equal to the average composite score for each of the duplicate candidates of that same set. . Finally, in Phase 3, the prioritized instances are reviewed by the internal auditor to determine whether the model captures true duplicate payments. Based on the auditor’s feedback, the model may be re-trained to adapt new criteria or modify existing criteria in order to increase its predictive power.



Conclusion

With the digitization of business processes in the ‘now economy’ and the increased demand for transparency and timely reporting by constituents, it is logical to examine whether CA can provide benefits to audits in the public sector. In this paper, a duplicate payment prioritization framework was described and applied to a semi-labeled imbalanced dataset. The prioritization framework implemented a rule-based duplicate detection technique and utilized risk-based criteria to predict the riskier, duplicate payment candidates. Finally, the framework relies on expert knowledge to improve its predictive prioritization ability. The results of the trained model provide evidence that duplicate payments can be effectively prioritized; hence, providing the auditor with a technique that can help enhance audit efficiency and at the same time, meet demands for timely and reliable government data.

Literature Review

Internal Audit in the Government Sector: The Internal Audit Function helps entities’ achieve its objectives by providing an unbiased and objective view. With respect to the public sector, auditing sets forth the foundation for good public sector governance (IIA, 2012). Hence, internal audit in the government sector is responsible for assuring that public funds are received and spent appropriately and that actions in the public sector are ethical and legal (IIA, 2012; Diamond, 2002). Moreover, auditing in the public sector differs from auditing in the private sector in other aspects. The IIA refers to sufficient funding as a key element of an effective public sector audit activity (IIA, 2012). Because government entities are financed by the taxpayers, resources tend to be scarce, and budgets allocated to the internal audit function may be lower than budgets that are typically allocated to the internal audit function in the private sector (Mebratu, 2015; Sterck and Bouckaert, 2006). Lower audit budgets may result in lower pay scales and limitations of skilled staff (Sterck and Bouckaert, 2006; Alzeban and Sawan, 2013).

Continuous Auditing: The paradigm of continuous auditing (CA) has existed for more than two decades. Pioneered by Vasarhelyi and Halper (1991), the theory of continuous auditing is based on the premise of the “audit by exception”, which enables the real-time monitoring and analysis of the entire population of records (Vasarhelyi and Halper, 1991). The entire population is examined based on a specific threshold, if that threshold is exceeded, then an exception would be identified. However, CA methodologies generally identify a large amount of exceptions which results in the problem of information overload. This problem has two main consequences. First, the more exceptions identified, the more time the auditor would have to spend investigating and explaining those exceptions, this would result in audit inefficiencies (Alles et al., 2006). Second, it is difficult for humans to process large amounts of information which could result in less than optimal decisions and have an impact on audit effectiveness (Kleinmuntz, 1990; Iselin, 1988). For these reasons, this study seeks to apply a prioritization framework that can discriminate the more suspicious duplicate payments.

Results

Phase 1: Duplicate Payment Detection Algorithm

The attributes of interest, vendor name and invoice amount, follow the rule to match instances that have the same vendor name and the same invoice amount. The results of the test reveal that from approximately 76,000 observations about 40,000 are detected as duplicate candidates, a decrease of 53%. Of course, it would be rather ineffective for the audit team to investigate 40,000 duplicate candidates in one year.

Phase 2: Duplicate Payment Prioritization

The risk-criteria of frequency of user, frequency of vendor, materiality, off-business hours, multiple invoices per week and size of duplicate set are applied to the duplicate payment candidates from Phase 1. Calculations are performed for each criterion on a candidate by candidate basis. Weights for the criteria, which were derived from the expert knowledge are also applied. The composite score for each duplicate set is the average of the composite scores for the duplicate candidates in the same duplicate set. Once the average composite scores per duplicate candidate set are calculated, a ranking system proceeds to apply lower ranking to the riskier duplicate payment candidates. Following the path of duplicate candidate ID 5112016933, set 1, the composite score is equal to 1.48614 (ranking 1251). The composite score for the second duplicate set is 1.60377 (ranking 1017). Hence, duplicate set 2 is deemed to be riskier (more suspicious) than duplicate set 1. The results of the framework application to the labeled duplicate payments that were used to train the model suggest that 4 of the 7 duplicate candidate sets, have a ranking below 800 , meaning that the auditor would have to review and investigate at most, 800 duplicate payment candidate sets, instead of 6,980 sets. Essentially, the current prioritization framework has the ability to detect duplicate payments by reviewing 12% of the duplicate sets (in comparison to 53% as in Phase 1), or 0.01% of all duplicate payment records, this represents a significant reduction of duplicate sets that would require investigation by the auditor.

		Prioritization Criteria							
Duplicate Set	Record ID	Frequency of user	Frequency of vendor	Materiality	Off-business hours	Multiple Invoices per week	Size of duplicate set	Composite score	Rank
1	5112016933	0.20404	0.13889	0.35154	0.00000	1.00000	0.12500	1.48614	1251
	5112019310	0.20404	0.13889	0.35154	0.00000	0.00000	0.12500	1.48614	
	5112026920	0.20404	0.13889	0.35154	0.00000	1.00000	0.12500	1.48614	
2	5113011889	0.06818	0.00893	0.44332	1.00000	1.00000	0.08333	1.60377	1017
	5113012439	0.06818	0.00893	0.44332	0.00000	0.00000	0.08333	1.60377	
3	5113005218	0.12963	0.18692	0.86034	0.00000	0.00000	0.29167	1.89712	594
	5113005219	0.12963	0.18692	0.86034	0.00000	0.00000	0.29167	1.89712	
	5113005220	0.12963	0.18692	0.86034	0.00000	0.00000	0.29167	1.89712	
	5113007263	0.12963	0.18692	0.86034	1.00000	0.00000	0.29167	1.89712	
	5113009779	0.12963	0.18692	0.86034	0.00000	0.00000	0.29167	1.89712	
	5113013138	0.12963	0.18692	0.86034	0.00000	1.00000	0.29167	1.89712	
	5113013377	0.12963	0.18692	0.86034	0.00000	1.00000	0.29167	1.89712	
4	5112015089	0.23391	0.25000	0.44370	0.00000	1.00000	0.54167	2.39235	248
	5112018187	0.23391	0.25000	0.44370	0.00000	1.00000	0.54167	2.39235	
	5112020179	0.23391	0.25000	0.44370	0.00000	1.00000	0.54167	2.39235	
	5112022364	0.23391	0.25000	0.44370	0.00000	1.00000	0.54167	2.39235	
	5112025354	0.23391	0.25000	0.44370	0.00000	1.00000	0.54167	2.39235	
	5112027611	0.23391	0.25000	0.44370	0.00000	1.00000	0.54167	2.39235	
	5113000496	0.23391	0.25000	0.44370	0.00000	1.00000	0.54167	2.39235	
	5113002735	0.23391	0.25000	0.44370	0.00000	1.00000	0.54167	2.39235	
	5113006241	0.23391	0.25000	0.44370	0.00000	1.00000	0.54167	2.39235	
	5113007409	0.23391	0.25000	0.44370	0.00000	1.00000	0.54167	2.39235	
	5113009008	0.23391	0.25000	0.44370	0.00000	1.00000	0.54167	2.39235	
	5113010531	0.23391	0.25000	0.44370	0.00000	0.00000	0.54167	2.39235	
	5113012227	0.23391	0.25000	0.44370	0.00000	1.00000	0.54167	2.39235	

Phase 3: Review of Prioritization Model

In Phase 3, the internal auditor would review the prioritized duplicate sets and provide feedback to the research team. Expert feedback to assess the applicability of the framework is necessary in order to improve the predictive ability of the model.

RUTGERS
Rutgers Business School
Newark and New Brunswick

Regtech: Upgrading the system of U.S. SEC for its Regulation

Ben Yoon

Background

- Data analytic techniques have been widely applied in all industry areas. This is especially so in the financial industry. Fintech such as Algorithmic trading, Robo-advisor, etc., are some examples of these current trends.
- Such techniques are not only used in the private sector but are being developed for use by regulators as well. Data analytic technology is being tailored for more specific market regulatory purposes.
- The capital market regulator of the US markets, the SEC, has started using the data analytic skills to further its stated mission. The adaptation of public sector technology trends has been branded Regtech (Regulatory + Technology)[1)] by the SEC.
- The US SEC actively using these data analytic techniques since mid-2013. Although there is no public official announcements for the main reason of recent increased actions, the financial reporting actions — one of the areas into which regtech is applied— is dramatically increased in recent years.

Number of SEC Actions[2)3)]

	2013	2014	2015	2016
Financial Reporting Actions	**68**	**98**	**134**	**N/A[4)]**
Independent Actions	341	413	507	548
Total Actions	676	755	807	868

1) SEC (Scott W. Bauguess, Acting Chief Economist), June 2017

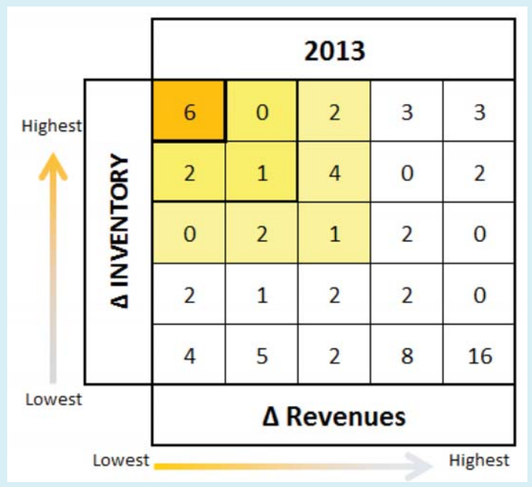
2) Wall Street Journal, "SEC Doubles Number of Financial Reporting", Oct. 22. 2015

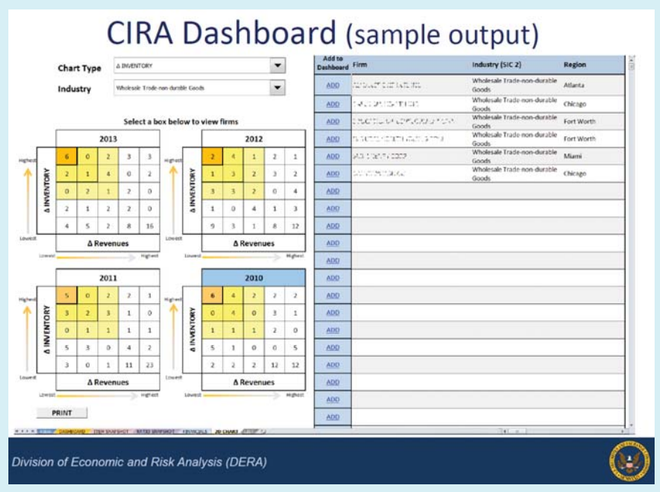
3) SEC Press Release 2015-245, Oct. 22. 2015, 2016-212, Oct. 11. 2016

4) Not Announced

Current SEC CIRA system

- The US SEC uses a data analytic system named Corporate Issuer Risk Assessment (CIRA) system for assessing the risk of the market, more specifically on the issuer's financial reporting.
- Due to the fact it is a system for monitoring market misconduct, much about the system is not known to the public. However, some of the key functions along with examples have been presented to the public. Based on this information, some inference can be drawn on its internal mechanisms.
- One aspect of the CIRA system is the use of discretionary accrual concepts (Modified Jones model) for predicting fraud or misconduct.
- Another part of CIRA system, which is known to public, is the CIRA Dashboard system[5)]. It displays the distribution of financial variables of the companies among the peer groups. Furthermore, it uses the combinations of financial variables. These combinations help detect outlier companies within peer groups.





5) Company Dashboard [SEC (Scott W. Bauguess), March 2015]

Previous Literature on predicting problematic companies

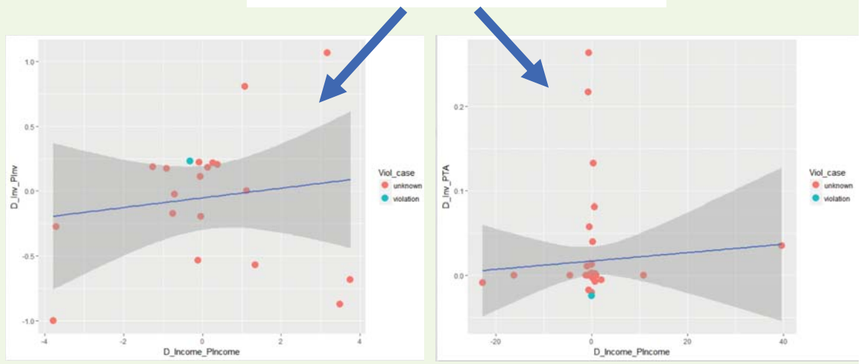
- Prediction of misconduct is a traditional topic.
- Dechow et al.(1995) developed discretionary accrual model (Modified Jones model) with Net-income, CFO, Δrevenue, Δreceivables, PPE, etc. by regression.
- Beneish (1999) studied predicting accounting standard violation with days sales in receivables, gross margin, Δrevenue, asset quality by probit or WESML.
- Dechow et al. (2002) developed another discretionary accruals approach with time series of Cash flow (CFO).
- Cecchini et al. (2010) studied a new model with sales, preferred stock price, short-term investment with support vector machine (SVM).
- Dechow (2011) introduced the F-score for predicting misstatement with Δreceivables, Δinventory, Δcash sales, Δnumber of employees, etc. by logistic regression.
- Roychowdhury (2006) developed **specific accounts predicting model** called Real Activity Manipulation (RAM)

  - SALES account: CFO/total asset = a + b (Sales/total asset) + c(Δsales/total asset) + e

Research Idea

- Predicting violation at a company level is useful. However, if the model can predict misconduct at specific account level, it would be much more useful especially for the regulation perspectives.
- The SEC system emphasizes on the inflated area, which is upper parts of matrix in dash board. However, the lower parts can be suspicious, too, because the variables are not the absolute values but the changes.
- In addition, I figured out that the current SEC system divides the matrix area unevenly. It indicates the left upper area as containing the most suspicious companies. However, depending on the dividing criteria, the other upper areas, or the other areas on the left can contain larger outliers than the left upper area.

Δ Inventory	2010 ~ 2013					Total	(Ratio)
	19	8	7	9	7	50	17%
	6	10	9	7	5	37	13%
	4	7	5	5	5	26	9%
	13	5	6	7	10	41	14%
	18	10	6	39	63	136	47%
	Δ Revenue						
Total	60	40	33	67	90	290	100%
(Ratio)	21%	14%	11%	23%	31%	100%	



Pilot test with Korean accounting violation cases

- I pilot tested my new methods with Korean accounting violation cases of 2015~2017 (July).
- I tested cases with violations involving Inventory, accounts receivable, or allowance.
- In some cases, the new model produces better results.
- I will apply this model to accounting violation cases for US firms for evaluation.

Pilot test results (smaller is better)

All companies in the same industry

Type	Current SEC	New Model
Inventory	61.9%	26.6%
A/R	50.2%	52.6%

Narrowing to Listed companies

Type	Current SEC	New Model
Inventory	- (N/A)	- (N/A)
A/R	67.8%	48.8%

RUTGERS
Rutgers Business School
Newark and New Brunswick

An Architecture Design for the Data Preparation of Continuous Risk Monitoring and Assessment System

Jiahua (Edward) Zhou

Backgrounds of the design

Continuous Risk Monitoring and Assessment (CRMA) is a Continuous Audit (CA) methodology to monitor an organization's business risks, identify its uncontrolled significant risks, and prioritize audit and risk management procedures for the timely mitigation of such risks (Vasarhelyi et al., 2010, Moon, 2016). Under this framework, the data preparation needs to interconnect real-time heterogeneous data from different sources, like CRM, ERP and supply chain systems, and also to integrate the legacy data into the current system to explore potential risks. As Kogan et al. (1999) stated, CA systems need elaborate data-capture mechanisms that supply the enterprise data for auditing. This is a challenging task and we have to transfer data from different sources with different data types. In this process, it is important to have a deep understanding of how, from where, and from whom the data are originated and how the data are processed. The proposed architecture is to offer an integrated solution for this data capturing mechanism. With this standardized data platform, it is achievable to integrate CRMA with COSO's Enterprise Risk Management (ERM) framework and offer a platform to help external auditors to give a fair judgment about the effectiveness of risk monitoring.

Related Works and Literature

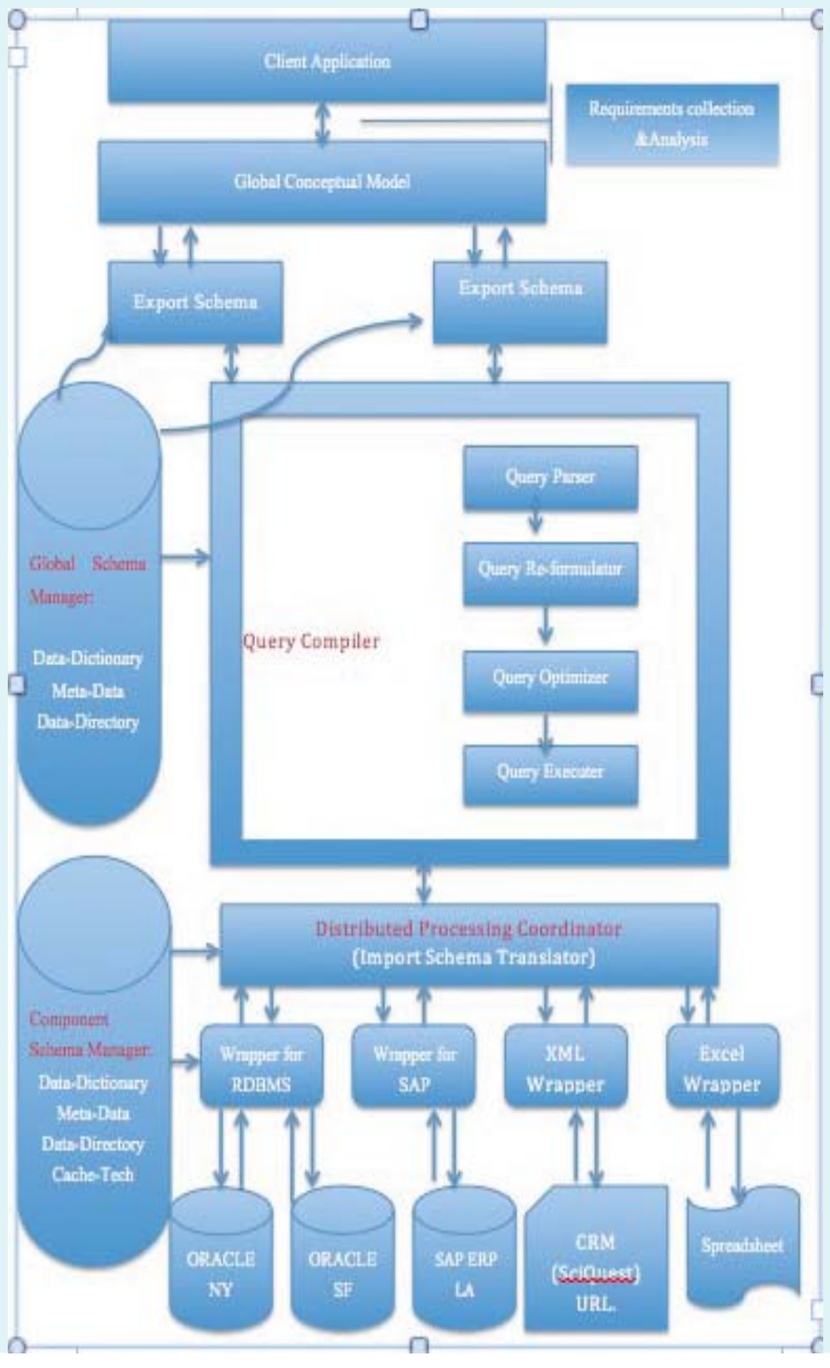
- Heimbigner, et al. (1985) introduced the federated architecture to construct a virtual database that standardizes different data types and data sources
- Scheuermann et al. (1990) had proposed a multi-databases solution to use centralized database administrators to develop translators that hide differences in query languages and database structures
- Noble, et al. (1997) designed an FIM architecture that uses a smart data dictionary (SDD), a data information manager (DIM), and a plurality of local information managers (LIMs) to integrate heterogeneous data
- Bouguettaya et al., (1997) proposed a Federated ObjectBase System (FOBS)
- Roscheisen et al. (1998) designed digital libraries to collect results from multiple different data sources in response to user's request (example: Stanford's InfoBus architecture)
- Chorafas (2001) proposed how to integrate ERP, CRM and supplier chain with smart materials.
- Kashyap et al, (2002) demonstrated a brokering mechanism to interconnect heterogeneous data sources by a federation database
- Doan, et al (2012) proposed a series of techniques to integrate sufficient and relevant unstructured big data in *Principles of Data Integration*

Basic components and working mechanisms

Figure 1 shows the designing process of the project-oriented virtual (non-materialized) federation database system to prepare a standardized data for the application of CRMA.

1. Global Conceptual Model (GCM): GCM is used to conceptualize companies' strategy, internal control principles and legislation with an applicable model
2. Export Schema: Transforming the GCM into a Mediated Schema with the following mechanisms: 1) schema mapping; 2) declare related data sources; 3) declare the related integrity constraints and functional dependency of the mediated schema
3. Query Compiler: 1) creating a global query over the mediated schema by query parser mechanism; 2) transforming the global query into unfolding sub-queries over data source by query re-formulator mechanism (meta-data match, semantic mapping and data-level homogenization); 3) query optimizer
4. Distributing Processing Coordinator: Construct a temporary RDBMS with ODBC driver and Talend software. Two reasons for this arrangement: 1) ODBC driver can take advantage of the virtues of SQL and relational data model; 2) Talend is a coordinator platform with high-degree function to interconnect relational database, semi-structured XML data, and Big Data
5. Global Schema Manager and Component Schema Manager are two small database repositories to record related schema transformation, meta-data mapping, data dictionary, related data-director, and act as a knowledge base to guide the data preparation in the following routine CRMA works

Figure 1: The Federation Database Architecture for the



Demonstration with Real Business Scenario

To demonstrate the mechanism of the integrated model the paper assume ABC, a kitchenware company, wants to construct a CRMA system. The company has an Oracle ERP system for the financial and operating department, and the market department uses the MARKETO system, and the procurement department uses SciQuest system by Website. The system is designed to monitor how customers' transaction behavior and customers' preference impact profit.

- Step 1: GCM: sales > cost
- Step 2: Mediated Schema: SELECT the information elements concerned FROM available data sources.
- The information elements include transaction price, transaction date, amounts, product item, material cost, direct labor, advertisement fee (*directly from RDBS)*; and also include customers' complaints, products' repair rate, customers' suggestions *(indirectly from Web system*).
- Step 3: Reformulate the query into unfolding sub-queries *The following is one SQL example:*

CREATE VIEW Transaction (location_ID, Customer_ID, Tran_date, Tran_ID, Tran_Unit_Price, Tran_Amount, Item_ID)

AS SELECT location_ID, Customer_ID, Tran_date, Tran_ID, Tran_Unit_Price, Tran_Amount, Item_ID

FROM ny.Tansactions **UNION ALL**

SELECT location_ID, Customer_ID, Tran_date, Tran_ID, Tran_Unit_Price, Tran_Amount, Item_ID

FROM sf.Tansactions

Step 4: Distributed Processing Coordinator (construct a temporary database to meet the data requirement of the system)

- Step 5: Constructing some wrappers to extract data from different data sources. The following is a syntax sample of Java API.

```
Query.Dataset
ds = query.addDataset ("customer_complaints");
ds.addFilter ("complainttype", "service_quality");
ds.addAttribute ("customer_id");
ds.addAttribute("product_id");
ds.addAttribute ("description');
ds.addAttribute("complaint_category");
ds.addAttribute ("solutions_for_complaints");
ds.addAttribute ("responsibility_attribute");
```

Contributions

This paper designs a project-oriented virtual (non-materialized) federation database system for the data preparation of CRMA. It offers a solution to interconnect heterogeneous data sources and legacy data to feed CRMA a standardized data platform. It allows leveraging the power of goal-oriented semi-structured data extraction to do for further analytics.

RUTGERS
Rutgers Business School
Newark and New Brunswick

Automating the General Ledger Audit: A New Approach to Manual Entry Sampling

Jamie Freiman, Yongbum Kim, and Miklos Vasarhelyi

Introduction

- Current audit standards permit less than 100% sampling of large sets of transactional data. (AU350)
- Traditionally this has proven to be effective and the most logical method of testing important accounts during an audit.
- Modern technology has brought with it large changes to the business environment. Primarily it has resulted in the exponential increase in the volume of transactions and related records.
- Current audit procedures have not adapted to effectively address this increase in data. In the case presented, internal auditors admitted to sampling about 50 records of the 12 billion entered per year.
- Fortunately modern technology provides us with the ability to apply analytic techniques to the whole population of these records.
- To this end, this project aims to develop a system with a set of analytics that can be applied to audit a total population of big data at a near continuous basis. A suspicion scoring system will be used to highlight the most egregious outliers for further examination.

Project Goals

Use a case study and gather information from internal auditors and academic literature to develop a system that can:

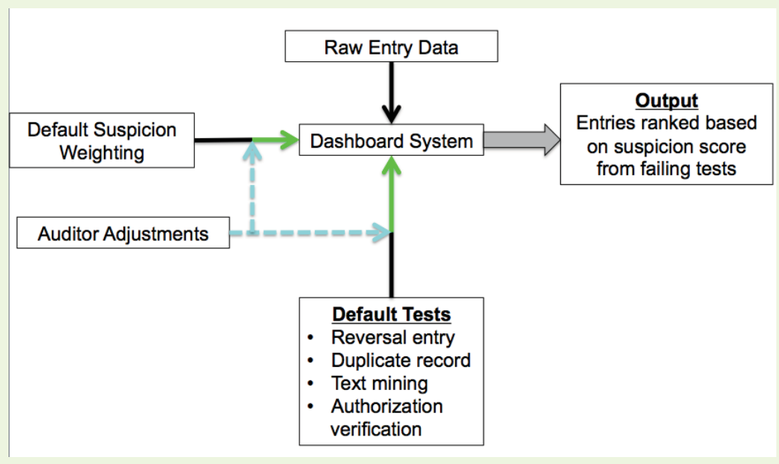
- Apply analytic tests to an entire population of big data.
- Score entries based on degree of suspicion to handle the large number of exceptions.
- Operate on a continuous basis if needed.
- Provide a level of customization for auditors.

Model Development

1. Develop preliminary control rules based on current audit practice, academic literature, and data mining techniques.
2. Test the preliminary controls with an internal audit team.
3. Add/revise/delete rules based on results from step 2.
4. Repeat steps 1-3 until a satisfactory model is developed.

Dashboard System Overview

- The proposed dashboard system will be made available for internal or external auditors to audit an entire population of big data on a continuous basis.
- A finalized list of analytics will be made available to the auditors to select from.
- A suspicion weighting system is being developed based on input from industry auditors and results gathered from case study testing.
- Weights and tests can be adjusted from default settings by auditors using the system to tailor it to their individual audit needs.
- The system will output suspicious entries that will be ranked in order of suspicion level based on the customized or default suspicion function.



Case Study Background

- The case study to which our methodology is applied involves the manual entries for a large multinational bank.
- The bank has provided transactional records from several different types of systems each of which contain different data structures. This will aid in the simulation of the dynamic variety of business environments for our model in real world application.
- Our initial sample consisted of 19,321 manual entries but has since grown to include the most recent years information.
- The data contains a sample of entries from several years allowing for intertemporal studies and analytics to be applied to the dataset.
- The dataset provided currently contains only manual entries. This subset of data is believed to be at the greatest risk of fraud or errors. These entries have already been "audited" but only using traditional methods.
- By working with an organization we are able to verify that our methods are detecting problematic entries, as well as hone our testing procedures based on industry input.

Initial Testing Procedures

Date Matching: For every journal entry a credit and debit should be made on the same date. We have found this is not the case

Reversal Entry Matching: This test is designed to focus on account manipulation. First we match originating and reversal entries. Then a detailed analysis and pattern is determined. For example, are all the proper authorizations in place, or is the reversal occurring in an unusual account?

Comment Text Mining: Manual entries require some descriptive comment. By mining these comments we can determine patterns of entries made by various users. This can be used to discover anomalies such as entries made by unauthorized parties.

Rules Mining: This more customized test examines corporate rules associated with entries. For example if a company only allows certain departments to generate specific entries.

Additional tests include:

- Basic statistical testing to determine trends, patterns and statistical deviations.
- Duplicate entry detection
- Benford's Law testing for account values

Preliminary Results

So far our testing on the case study audited data has garnered a host of significant results. Below is a selection:

- Reversal entry testing resulted in the discovery of several thousand suspicious entries made by an employee that was fired.
- Text mining comments resulted in the discovery of 6 unapproved entries totaling over $10 million that should have required some approval.
- In one system alone 365 suspicious duplicate entries were detected over a 6 month period.
- A large number of entries were discovered with the same credit and debit accounts which should not occur indicating inter-departmental irregularities.
- Within one system it was discovered that a majority of entries had different dates on credits and debits which is highly unusual.

These results clearly indicate some merit to our procedure. We are currently working on applying further tests as well as a suspicion system to deal with the large number of exceptions that is being generated by our methodology.

RUTGERS
Rutgers Business School
Newark and New Brunswick

Implementation of Data Analytics in Audit Process: Model for Individual Adoption

Lu Zhang

Research Objective

- In the big data era, data analytics, normally referred to as Big Data Analytics (BDA) (Russom, 2011), is defined as a holistic process to help answer questions or solve problems for decision makers by analyzing big data. (Cao, Chychyla and Stewart 2015, Schneider, Gary., et al 2015, Bose and Ranjit 2009, Akter, Shahriar, and Wamba 2016, Kiron, David, Prentice and Ferguson 2012).
- According to Gartner's 2017 Business Intelligence report based on the Hype cycle theory, Big data analytic techniques such as Predictive analytics, Prescriptive analytics, Text analytics and Visual data discovery, are considered emerging technologies and many companies have already started using them to gain business benefits (Bitterer 2017).
- However, auditing is lagging behind the other research streams in the adoption of big data analytics (Gepp et al 2017). This is even slower than adoption by other areas of public accounting firms such as consulting practice and tax (Earley, 2015). In the context of auditing, using big data analytics technique, is still at the technology trigger stage of the Gartner's hype cycle (O'Leary 2008 and Thomas, et al 2017).
- This research highlights the obstacles of adopting data analytics in audit process and proposes to use UTAUT model to predict the individual auditor's adoption of Big Data Analytics in the audit process.

Literature Review: Data Analytics

- Data analytics have been around for a long time. The notion of data analytics, defined as the process of deriving insights from data, is as old as the field of statistics, dating back to the 18th century (Davenport and Dyche 2013, Agarwal and Dhar, 2014).
- Data analytics tools have been used in business since the mid-1950s. Davenport and Dyche called the initial era of data analytics the analytics 1.0 era during which data analytics, majority of which is descriptive analytics (Sivarajah, et al 2017) are conducted based on small and structure data sets from internal sources, sometimes even following manual batch process. (Davenport and Dyche 2013)
- Nowadays, the notion of data analytics is normally referred to as Big Data Analytics (BDA) (Russom, 2011). It is quite different from the initial era in the following ways: 1. Data was often externally-sourced (Web based, Mobile based or sensor based data); 2. Volume of the datasets is large; 3. Both structured and unstructured data (text data, log data or visual data) are analyzed; 4. Data is stored and processed rapidly; 5. Advanced data mining, machine learning or data visualization tools are used; 6. Not only descriptive analytics (report on the past) but also predictive (analyzing the past data to predict the future) and prescriptive analytics (using models to optimize business actions) are used in the analytic process. (Chen et al, 2012, Davenport and Dyche 2013, Cao, Chychyla and Stewart 2015, Schneider, Gary., et al 2015, Bose and Ranjit 2009, Akter, Shahriar, and Wamba 2016, Kiron, David, Prentice and Ferguson 2012, Sivarajah, et al 2017)

Literature Review: United Theory of Acceptance and Use of Technology

- The United Theory of Acceptance and Use of Technology Model (UTAUT) (Venkatesh et al., 2003) was a unified model developed based on eight different technology acceptance models to predict behavior across many settings. It can also be applied to the adoption of Big Data Analytics by auditors in the audit process.
- According to Venkatesh et al., 2003, The UTAUT model suggested that performance expectance, effort expectance, social influence and facilitating conditions can be used to predict user acceptance of innovation. The brief explanations of the four constructs can be listed as follows:
- (1) Performance expectancy: "the degree to which an individual believes that using the system will help him or her to attain gains in job performance."
- (2) Effort Expectance: "the degree of ease associated with the use of the system."
- (3) Social influence: "the degree to which an individual perceives that important others believe he or she should use the new system."
- (4) Facilitating conditions: "the degree to which an individual believes that an organizational and technical infrastructure exists to support use of the system."

Apply UTAUT to predict auditors' adopting Big Data Analytics in the Auditing Process

I propose to use UTAUT model to predict auditors' adopting Big Data Analytics in the Auditing Process. The UTAUT model suggested that performance expectance, effort expectance, social influence and facilitating conditions can be used to predict auditor's acceptance of big data analytics in the audit process (Figure 1). Based on the model survey and instruments will be designed to test the model. The explanation of the key constructs can be listed as follows:

- (1) Performance expectancy: the degree to which an auditor believes that using the big data analytics in the audit process will help him or her to gain the better audit judgment.
- (2) Effort Expectance: the degree of ease associated with the use of the big data analytics in the audit process, evaluated by the auditor based on the characteristics of the task, analytic tools and his or her own skill set.
- (3) Social influence: the degree to which an auditor perceives that the related parties such as partners, managers, clients, audit team members, PCAOB and financial statement users believe he or she should use the new system.
- (4) Facilitating conditions: the degree to which an auditor believes that there are sufficient financial and technology resource in the audit firm as well appropriate regulations exist to support use of the system.

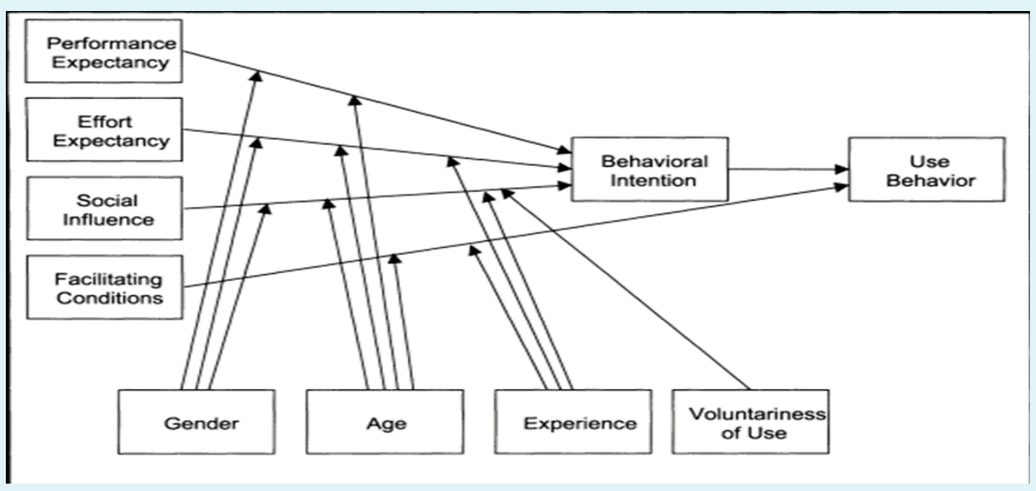


Figure 1 taken from Venkatesh et al., 2002

Obstacles for auditors in using Big Data Analytics in the audit process

- Lack of Appropriate Regulations: Auditing standards need to be changed to facilitate the use of more advanced technology, making sampling techniques obsolete. (Moffitt and Vasarhelyi 2013, William 2013, Whithouse 2014, Alles 2015, Vasarhelyi et al. 2015)
- Data standards are needed. (Alles and Vasarhelyi 2013, Vasarhelyi 2014, Whithouse 2014)
- Fear of litigation risk (Whithouse 2014, Cao, Chychyla, and Stewart 2015)
- Clients fear losing information privacy (Yoon et al, 2015, Cao, Chychyla, and Stewart 2015)
- Initial investment is a barrier to use (Thomas et al. 2017 working paper)
- Auditor lack of required skills for big data analytics. (Russom 2011, Brown-Liburd, Issa and Lombardi 2015). According to the interview results of Thomas et al, 2017, most of the audit firms are hiring specialized data scientists or analytic groups to help run the necessary analytics for the audit. (Thomas et al, 2017)

RUTGERS
Rutgers Business School
Newark and New Brunswick

Cognitive Assistant for Audit Plan Brainstorming Sessions

Qiao Li and Miklos A. Vasarhelyi

Motivation

- The brainstorming meeting for audit risk assessment allows the engagement team to identify risks and initiate a free flow of ideas about how a material misstatement could occur whether fraudulent or erroneous (AICPA, 2016)
- Most common method used is a discussion topic checklist
- More powerful tools should be developed to assist auditors in making better judgments in risk assessment

Contribution

- Introduce cognitive computing into auditing domain by proposing a Cognitive Assistant (Intelligent Personal Assistant)
- Develop a knowledge base which potentially includes numerous auditors' knowledge and experience, with the objective to improv audit risk judgments during brainstorming

Cognitive Computing

Cognitive computing refers to a system that learns at scale, reasons with purpose, and interacts with humans naturally.

- Learn, reason, and improve the "knowledge" of the machine by interacting with humans (IBM Research, 2016)
- Mimics the functioning of the human brain and helps human to make better decision (Wang, 2009d, Wang et al., 2009, Terdiman 2014, Knight 2011, Hamill 2013, Denning, 2014, Ludwig 2013)
- Involve technologies such as data mining, pattern recognition and natural language processing

Research Methodology

- Piggybacking on public domain modules (Modules used in current Cognitive Assistant such as Apple's Siri, Google Now)

Future of Cognitive Computing in Auditing

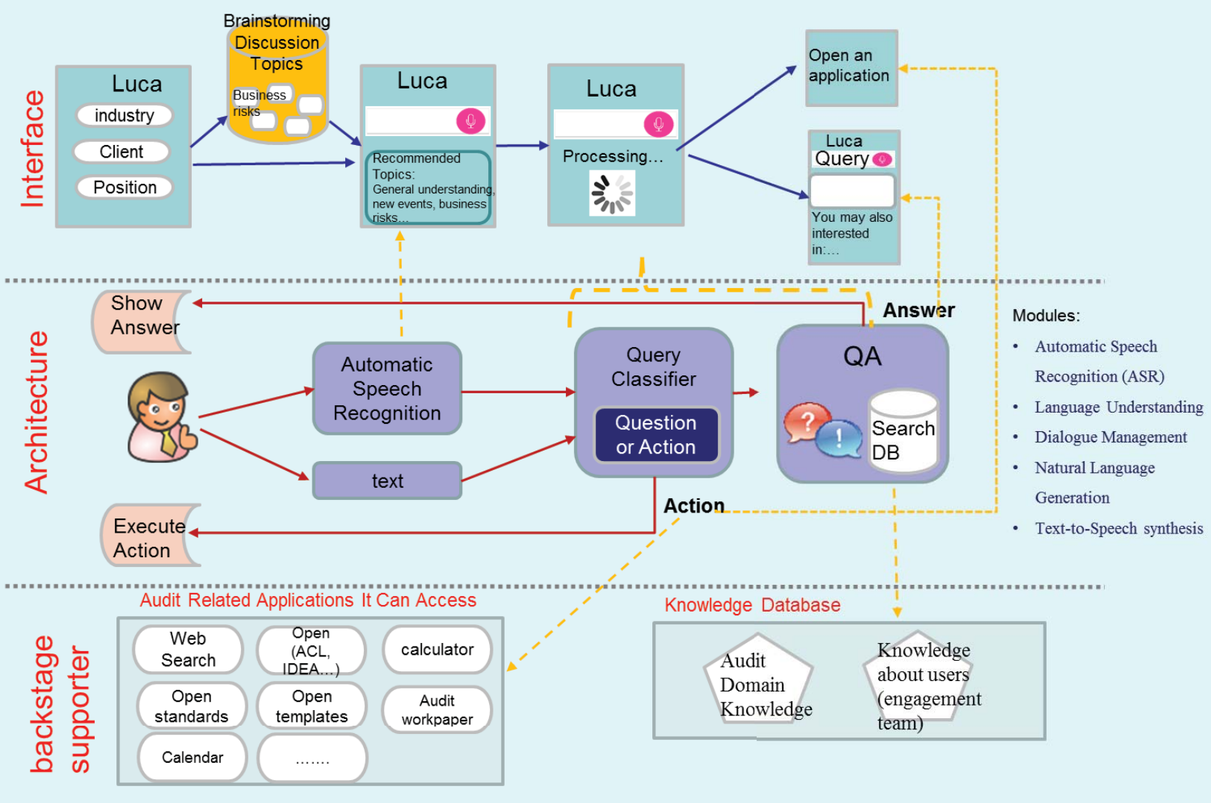
1. Possible to support human decision making by processing large data within a short time
2. Many accounting firms have already started adding cognitive technologies to their businesses and more firms will join (Kokina and Davenport 2017)
   - KPMG signed a brad agreement with IBM to apply Watson to a series of audit processes (Lee 2016)
   - Deloitte is trying to assemble different cognitive capabilities from various vendors and integrate them to support audit process, such as document review and predictive risk analytics (Raphael 2016)
   - PwC and EY are increasing their use of audit platforms and predictive analytics (Kokina and Davenport 2017)

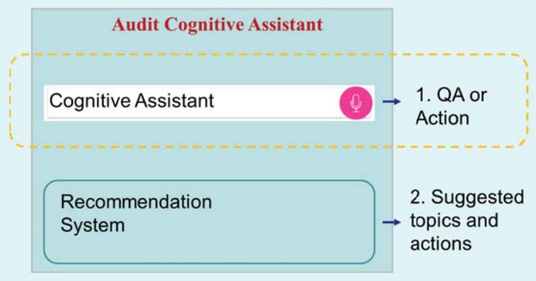
Further application in audit planning:

- Retrieve huge amount of information

Architecture of the Proposed Audit Cognitive Assistant

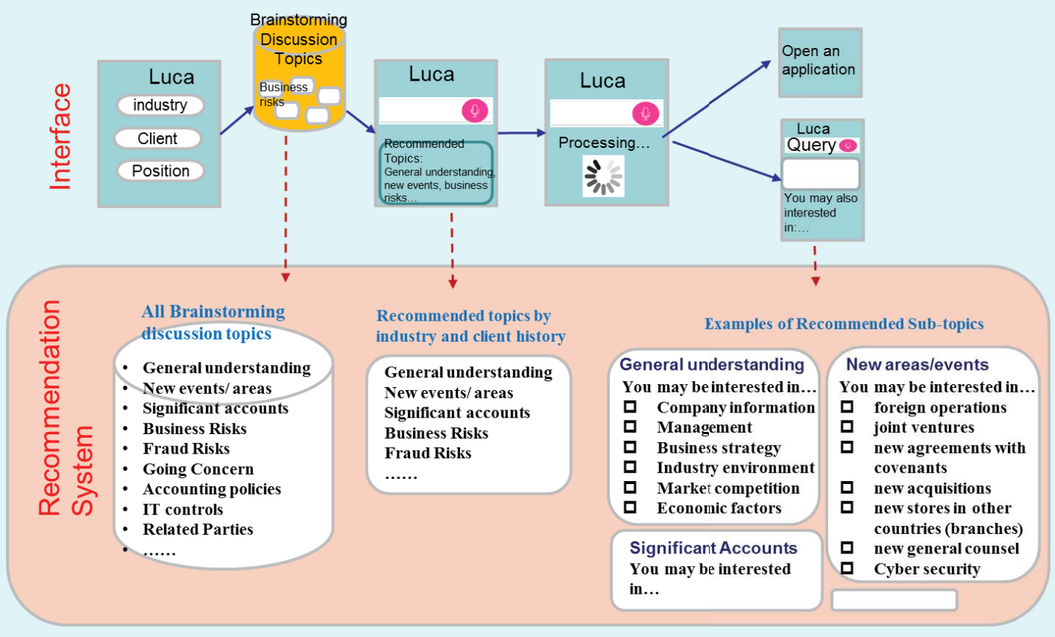
Components of



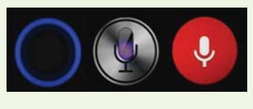


Cognitive Assistant *(Left)*

Recommended Topics *(Below)*



Cognitive Assistant (Intelligent Personal Assistant)



CA is an application that uses inputs such as the user's voice, vision, and contextual information.

- Properly answers User's questions in natural language, making recommendations, and performing actions (Hauswald et al. 2015)
- Provide information retrieval and support users through recommender systems (Garrido et al. 2010)
- Adapt itself by learning users' preferences over a wide range of functions within the system (Myers et al., 2007).

RUTGERS
Rutgers Business School
Newark and New Brunswick

A Framework of Applying Process Mining for Fraud Scheme Detection

Tiffany Chiu, Yunsen Wang, and Miklos A. Vasarhelyi

Introduction

Process mining is an analytical technique that is used to analyze an entity's business processes based on event logs that have been automatically recorded in the accounting information systems prior to the analysis. To detect and prevent corporate fraud is one of the major objectives of audit practice. Corporate fraud includes 1) intentional embezzlements of corporate resource, 2) corruption and bribery, and 3) intentional misstatement of financial statement to misguide the stakeholders (i.e., financial statement fraud). It refers to an entity's management improperly uses accounting schemes to falsify and report misleading financial statement in order to meet or beat the analysts forecast.

Process mining can be applied as non-financial information in the prediction of financial statement fraud because it enables whole population of event logs and has potential of adding value to auditing. The purpose of this paper is to provide a framework on how process mining can be applied to identify fraud schemes and assess the riskiness of business process. Specifically, the proposed framework captures how the suspicious patterns in process mining can be used to detect potential fraudulent transactions.

The contribution of this paper is three-fold. First, this paper proposes a framework that link notable variants/ activities in process mining with corresponding fraud schemes to detect potential fraudulent transactions. Second, the proposed framework can be applied to build a continuous fraud monitoring system that using suspicious patterns and risk level as filters to detect financial statement fraud. Third, process mining enables auditors with full population of event logs that can be used as non-financial information to detect fraud.

Literature Review

Applying process mining in accounting or auditing research fields is still in its infancy. For example, prior research applied process mining to examine bottlenecks, identify business process deviations and monitor authorization rules (Van der Aalst et al., 2003; Rozinat and Van der Aalst, 2005; Rozinat and Van der Aalst, 2008; Bukhsh and Weigand, 2012). Prior literature on applying process mining of event logs in internal control framework stressed that using business process focused information in the structure of internal control could improve the evaluation of internal control effectiveness (Kopp and O'Donnell, 2005). In addition, Chiu et al. (2017) indicated that by adopting process mining to evaluate the effectiveness of internal control, auditors would be able to utilize the results from process mining analysis in the audit procedure.

Accounting research on financial statement fraud and Accounting and Auditing Enforcement Releases (AAERs) includes testing hypotheses grounded in the literature of earnings management (Summers and Sweeney, 1998; Beneish, 1999; Sharma, 2004) and corporate governance (e.g., Beasley, 1996). The early research of financial statement fraud dates back to 1980s (Elliott and Willingham, 1980). Feroz, Park, and Pastena Feroz et al. (1991) documented the AAERs affecting stock price. Beasley Beasley (1996) examined the association between board of the director composition and financial statement fraud.

To provide additional non-financial information for fraud detection, event logs that are created automatically by accounting information systems could be extracted and analyzed by process mining techniques (Chiu et al. 2017).

Data Preprocessing

The fraud financial statement sample is collected from WRDS Restatement Database that contains 279 restatements related to fraud. After dropping the fraud observations with short period restatement (less than 350 days) and merging with Compustat data from 1992 to 2017, there are 202 fraud firms and 513 fraud firm-year observations.

Fraud Types and Fraud Category

Fraud Category	Frequency	Percentage
Revenue recognition issues	174	37.02%
Foreign, related party, affiliated, or subsidiary issues	150	31.91%
Liabilities, payables, reserves and accrual estimate failures	114	24.26%
Accounts/loans receivable, investments & cash issues	107	22.77%
Inventory, vendor and/or cost of sales issues	107	22.77%
Foreign, subsidiary only issues (subcategory)	97	20.64%
Expense (payroll, SGA, other) recording issues	90	19.15%
PPE intangible or fixed asset (value/diminution) issues	44	9.36%
Deferred, stock-based and/or executive comp issues	35	7.45%
Acquisitions, mergers, disposals, re-org acct issues	34	7.23%
Tax expense/benefit/deferral/other (FAS 109) issues	31	6.60%
Intercompany, investment in subs./affiliate issues	30	6.38%
Fin Statement, footnote & segment disclosure issues	30	6.38%
Debt, quasi-debt, warrants & equity (BCF) security issues	25	5.32%
Lease, SFAS 5, legal, contingency and commitment issues	24	5.11%
Capitalization of expenditures issues	21	4.47%
Unspecified (amounts or accounts) restatement adjustments	21	4.47%
Acquisitions, mergers, only (subcategory) acct issues	16	3.40%
PPE issues - Intangible assets, goodwill only (subcategory)	15	3.19%
Consolidation issues included Fin 46 variable interest & off-B/S	13	2.77%
Intercompany, only, (subcategory) - accounting issues	13	2.77%

Framework

To detect corporate fraud using process mining, it is necessary to understand the standard business process for accounting cycles. For example, the standard business process for "order-to-cash" cycle is: Order Created → Goods Issue → Invoice Created → Invoice Posted → Payment Received → Invoice Cleared, and the standard process pattern for "procure-to-pay" cycle is: Create Purchase Order → Sign → Release → Goods Receipt → Invoice Receipt → Payment.

An example of the mapping notable variants into fraud categories is that once process mining detects a sales order missing activity "goods issue" or missing activity "payment received," there is a risk that this order (transaction) could be fictitious or turn out to be a "bill-and-hold" fraud scheme, which will result in revenue recognition issue.

Based on the most common corporate fraud schemes and the activities and variants in the event logs of an ERP system, this study identifies suspicious patterns or activities for each fraud scheme and assigns the risk levels.

Accounting Fraud Schemes and Suspicious Process Patterns

Accounting Cycle	Fraud Scheme	Notable Activity	Suspicious Pattern Example	Risk Level
Order-to-Cash	Altering Documentation	1. Order Adjusted: Goods Issue Date 2. Invoice Adjusted	Frequent occurrence of order adjusted and/or invoice adjusted activities without approval process during fiscal year end period.	High
Order-to-Cash	Bill and Hold	3. Goods Issue 4. Payment Received	Missing goods issue and/or payment received.	High
Order-to-Cash	Channel Stuffing	5. Order Adjusted: Order Return 6. invoice adjusted: invoice credit note	Frequent occurrence of order return or invoice credit note immediately after fiscal year end without approval process.	High
Order-to-Cash	Up-Front Fees	7. Payment Received 8. Good Issue 9. Order Adjusted: Change Goods Issue Date	Payment received occurs before goods issue or invoice created. Order adjusted without approval process.	Low
Order-to-Cash	Failure to Record Sale Allowances	10. Payment Received	Missing payment received or incomplete payment	Medium
Order-to-Cash	Inflating the Value of Inventory	11. Order Adjusted: Net Price	Order adjusted without approval process Putting in improper price comparing to market value	Medium
Procure-to-Pay	Off-site or Fictitious Inventory	12. Goods Receipt	Abnormal goods receipt records: missing goods receipt and/or have duplicate or more than one goods receipt in one purchase order	High
Others	Fraudulent Audit Confirmation	13. All Activities	Matching trading partners corresponding event logs	/
Others	Refresh Receivables	14. Invoice Adjusted	Invoices adjusted occurs for many transactions without approval process	High
Others	Promotional Allowance Manipulation	15. Invoice Adjusted: Cash Discount	/	Medium
Others	Intercompany Manipulations	16. Invoice Posted: Revenue Intercompany	/	Medium
Others	Bribery and Corruption	17. All Activities	Using resource information in event logs to identify potential violation of segregation of duty controls	Medium

Conclusion

The application of process mining in financial statement fraud detection can assist auditors in detecting potential fraud by examining the potential fraudulent process patterns. The framework proposed in this study indicates that process mining can be a powerful fraud detection tool when auditors include the potential fraudulent patterns in their fraud detection process.

Future research could extend the current framework by incorporating more fraud schemes and other accounting cycles when discussing how process mining can be used in fraud detection. Furthermore, a proof-of-work (e.g. prototype) can be built to simulate the application of the proposed framework to detect certain types of fraud schemes. This prototype can be designed to incorporate clustering algorithms to automatically assign risk levels to all transactions.

RUTGERS
Rutgers Business School
Newark and New Brunswick

The Incremental Informativeness of Management Sentiment for the Prediction of Internal Control Material Weakness (ICMW) : An application of deep learning to textual analysis for conference calls

Ting Sun

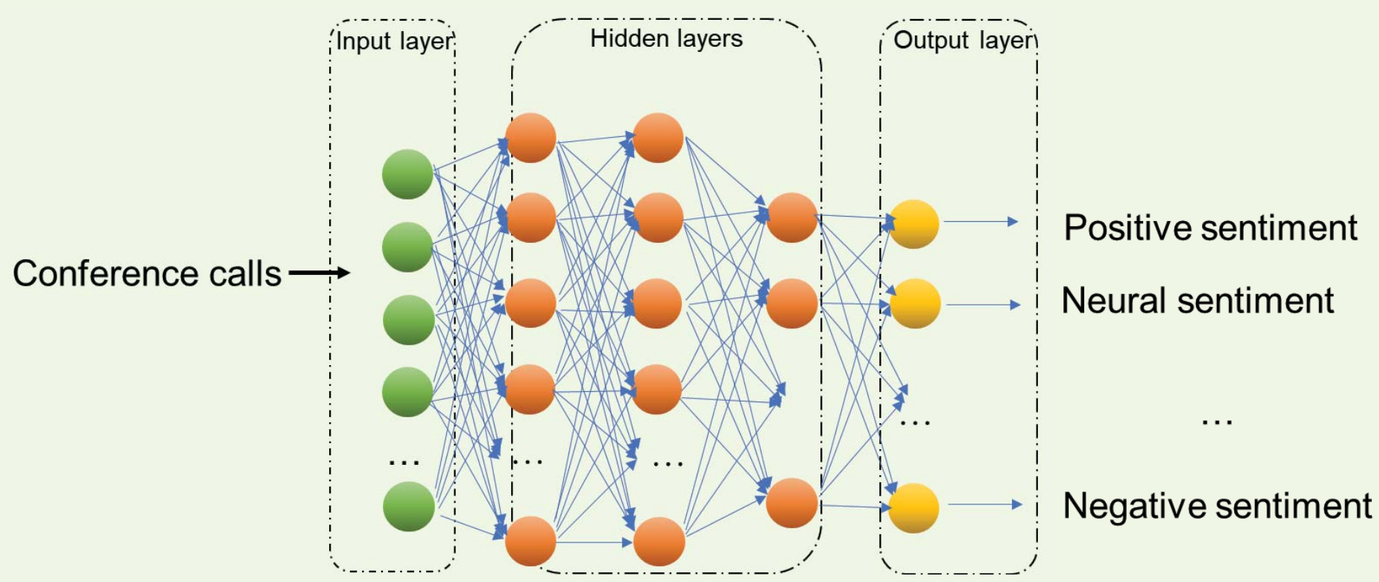
Research Objectives

- Examining the relationship between the sentiment features of management from conference calls and the likelihood of ICMW
- Investigating whether the sentiment features contain incremental information for the prediction of ICMW
- Demonstrating the effectiveness and efficiency of deep learning

Hypotheses

- H1: The sentiment features of conference calls are significantly associated with the likelihood of ICMW
- H2: The explanatory ability of the model that incorporates sentiment features of conference calls along with the major financial determinants is superior to that of the model that merely uses the financial determinants.

Deep Neural Network for Sentiment Analysis



Sentiment Features

- **Overall sentiment score**
  Measures the sentiment strength of the document, ranged from -1 to 1
- **Joy score**
  The value ranges from 0 to 1, which represents the confidence level indicating the probability that the emotion of joy is implied by the sample text.

Data

- The size of the final conference call transcripts is 1758 corresponding to fiscal year from 2004 to 2014, among which 201 firm-years are related to ICMW.
- All conference calls come from SeekiNF

Regression

$$ICW = \beta_0 + \alpha_1 Sentiment + \alpha_2 Joy + \beta_1 Marketvalue + \beta_2 Aggregateloss + \beta_3 Distress + \beta_4 Segments + \beta_5 Foreign + \beta_6 Inventroy + \beta_7 Restructure + \beta_8 Acquisition + \beta_9 Resign + \beta_{10} Big4 + \beta_{11} Litigation + \sum IndustryFE + \varepsilon$$

Results

- There is a significant negative relationship between the joy score and ICMW.
- With the incorporation of the sentiment features, especially the score of joy, the explanatory ability of the model improves significantly over the baseline model that merely utilizes the major ICW determinants suggested by prior literature.

RUTGERS
Rutgers Business School
Newark and New Brunswick

The Performance of Sentiment Features of MD&As for Financial Misstatement Prediction:

A Comparison of Deep Learning and "Bag-of-Words" Approach

Ting Sun, Yue Liu, and Miklos A. Vasarhelyi

Research Questions

1. Do sentiment features add information for financial misreporting prediction?
2. If they do, are they effective only for fraud prediction or for misstatements including both fraud and error?
3. How effective is the model using deep learning based sentiment features is as compared to the model using sentiment feature obtained by bag of words approach?

Sentiment analysis approaches

	Deep learning approach	Bag of words approach
Description of the technique	Emerging technique employing deep hierarchical neural network and trained with a large amount of text files	Prevalent technique using various predefined word lists, with each one representing a particular sentiment feature
Rationale	"Understand" the meaning of a text file	Count the frequency of the words originated from a specific dictionary
Output Sentiment feature	Sentiment scores: Sentiment_DL Emotion scores: JOY	Sentiment scores: Sentiment_TM
Prior literature in accounting & auditing domain	Limited	Yes
Tool	IBM Watson	Loughran and McDonald (2011)
Classification as a finance-specific tool	No. But includes finance-related content	Yes (10-K)
Required text document	HTML/text document and webpage	HTML/text document
Data preprocessing requirement	No	Yes

Data

- All MD&As are from SeekiNF
- 31466 MD&As of 10-K filings for fiscal years from 2006 to 2015
- With deep learning approach, we obtained Sentiment_DL and Joy
- With bag of words approach, we obtained Sentiment_TM

Sentiment Features

- Overall sentiment score: Measures the sentiment strength of the document, ranged from -1 to 1
- Joy score: the value ranges from 0 to 1, which represents the confidence level indicating the probability that the emotion of joy is implied by the sample text.

Classification Models

The structure of initial proposed models

Panel A: misstatement prediction

		Baseline model	Model 1	Model 2
Dependent variable		MISSTATEMENT	MISSTATEMENT	MISSTATEMENT
Independent variables	Sentiment measures	N/A	SENTIMENT_TM	SENTIMENT_DL JOY
	Other candidate predictors	35 variables related to misstatement	35 variables related to misstatement	35 variables related to misstatement

Panel B: fraud prediction

		Baseline model	Model 1	Model 2
Dependent variable		FRAUD	FRAUD	FRAUD
Independent variables	Sentiment measures	N/A	SENTIMENT_TM	SENTIMENT_DL JOY
	Other candidate predictors	35 variables related to misstatement	35 variables related to misstatement	35 variables related to misstatement

Panel C: error prediction

		Baseline model	Model 1	Model 2
Dependent variable		ERROR	ERROR	ERROR
Independent variables	Sentiment measures	N/A	SENTIMENT_TM	SENTIMENT_DL JOY
	Other candidate predictors	35 variables related to misstatement	35 variables related to misstatement	35 variables related to misstatement

Prediction Results

M: Misstatement
F: Fraud
E: Error

	Baseline Model (no sentiment features)			Model 1 (bag of words based sentiment)			Model 2 (deep learning based sentiment)		
Metrics	(A) M	(B) F	(C) E	(D) M	(E) F	(F) E	(G) M	(H) F	(I) E
Type □ error rate	43.22%	44.68%	51.46%	46.48%	43.49%	51.87%	46.51%	39.53%	53.11%
Type □ error rate	42.04%	24.59%	33.64%	36.94%	29.51%	32.71%	37.58%	31.15%	26.17%
AUC	0.596	0.68	0.598	0.619	0.658	0.602	0.61	0.704	0.616

Answers to Research Questions

1. Yes
2. Fraud prediction only
3. Improvement of effectiveness in terms of Accuracy, AUC, and false positive rates.

RUTGERS
Rutgers Business School
Newark and New Brunswick

Predicting Public Procurement Irregularities

Ting Sun and Leonardo J. Sales

Research Objectives

- Developing a system for CGU (the Comptroller General of the Union, Brazil) using the characteristics of the bidding company to predict the risk of public procurement irregularities
- Ensuring that the identified companies are actually with high risk of public procurement irregularity
- Allowing a sufficient number of contractors to participate in the bidding process

Methodology

Traditional Artificial Neural Network (ANN)

Deep Neural Network (DNN)

Logistic Regression

Discriminant Analysis

Dataset Name	Description	Time period	Risk Dimension
RAIS	Employee information	2011-2014	Operational Capabilities
RFB	Partners information	2011-2014	
SIAFI	Public spending information	2011-2014	
SIASG	Historical Biddings information	2011-2014	Profile of Participation in Biddings / History of Punishments and Findings
SIAPE	Public servant information	2011-2014	Conflict of Interests
SICONV	Amount transferred to the bidding company by a non-governmental organization (NGO)	2011-2014	
TSE	Information of the involvement of the company in elections	2011-2014	Political Bonding

Predictive Performance

	TNN	DNN	Logistic Regression	Discriminant
Overall Accuracy	0.8383	0.9157	0.7789	0.7948
F_1	0.3259	0.4677	0.3005	0.3251
F_2	0.4381	0.5205	0.4545	0.4844
$F_{0.5}$	0.2594	0.4246	0.2244	0.2447
Precision	0.2283	0.4000	0.1920	0.2101
Recall	0.5683	0.5630	0.6906	0.7194
Specificity	0.8582	0.9405	0.7854	0.8003
True Negatives	1616	1803	1479	1507
True Positives	79	76	96	100
False Negatives	60	59	43	39
False Positives	267	114	404	376
False Negative Rate	0.4317	0.4370	0.3094	0.2806
False Positive Rate	0.1418	0.0595	0.2146	0.1997
AUC	0.82	0.8780	0.8190	0.817

Z Test

Z Test for Differences between Proportions (DNN vs. TNN)

$H_0 : P_{TNN} \geq P_{DNN}$ versus $H_a : P_{TNN} < P_{DNN}$

n	DNN	TNN	z	P value
For all firms: Percentage of hits (AUC)				
10186	87.80	82.00	11.60***	0.001
For all firms: Percentage of hits (accuracy)				
10186	91.57	77.89	27.64***	0.001
For firms with irregularities: Percentage of hits (type two hit)				
744	56.29	56.83	-0.21	0.4168
For firms without irregularities: Percentage of hits (type one hit)				
9442	94.05	85.82	18.97***	0.001

Conclusions

- DNN has much less false positives but slightly more false negatives
- DNN outperforms other algorithms in terms of all F scores, suggesting that DNN has better predictive accuracy when considering precision and recall together
- The Z test suggests that DNN is significantly more powerful than other algorithms in terms of accuracy, AUC, and type I hit. With regards to the type II hit, TNN is more effective than DNN, but the difference of the type II hit between TNN and DNN is insignificant. For the same metric, Discriminant Function Analysis has the best result and the difference is significant, but the Z score (which is

RUTGERS
Rutgers Business School
Newark and New Brunswick

Applying Deep Learning to Audit Procedures: An Illustrative Framework

Ting Sun

Capabilities of Deep Learning

- Text understanding
- Speech recognition
- Visual recognition
- Structured data analysis

Metrics of Deep Learning Functions

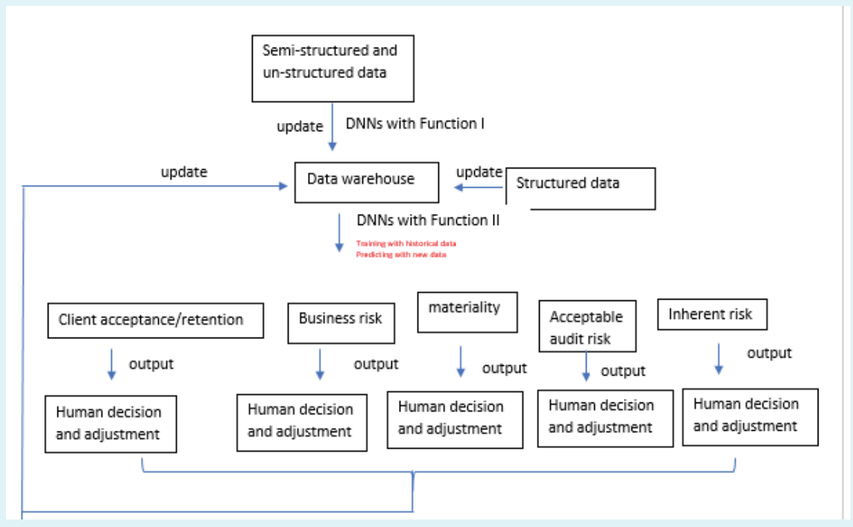
Audit \ Functions	Function I Information identifier	Function II Judgment supporter
Structured	Applicable	Applicable
Semi-structured	Not Directly Applicable	Applicable
Unstructured	Not Directly Applicable	Applicable

Functions of Deep Learning for Audit Data Analytics

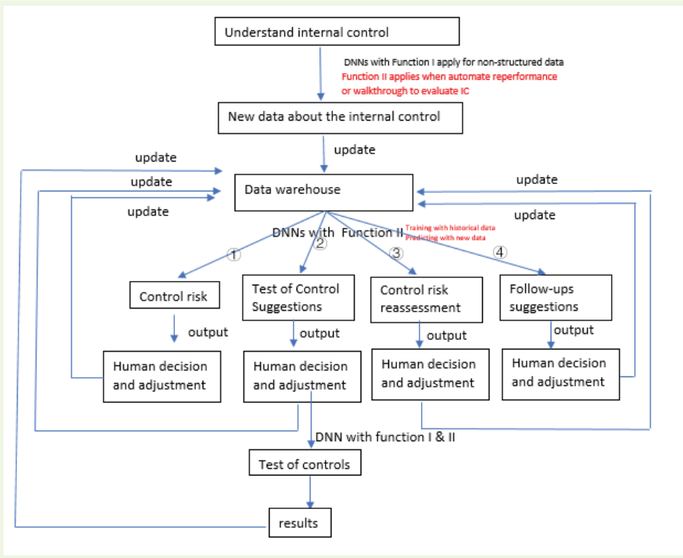
- Function I: Information Identifier
- Function II: Judgment Supporter

	Text	Audio	Video	Image
Data Source	· Contracts conference call transcripts · Trading publications · Tweets, and Facebook posts · Customer reviews · News articles · Analyst reviews and emails	· Phone calls · Speeches · Presentations	· Surveillance videos in stores, ware-houses, and offices	· Pictures · Scanned documents · Handwrites
Deep Learning Model (function I)	Use pre-trained DNN like Watson Analytics for data in general topics Develop our own DNNs with finance-specific data			
Output	Sentiment and emotion score for each topics, keywords, and concepts (entities involved related items)	Speaker's topics, keywords, Sentiments, and emotions	Human face Objects Quantity/Quality of the object Human's behavior	Face, Object, and Text
Deep Learning Model (function II)	Data (evidence) warehouse combined with traditional numerical data Use deep neural network as the final classification model to support complex audit judgment			

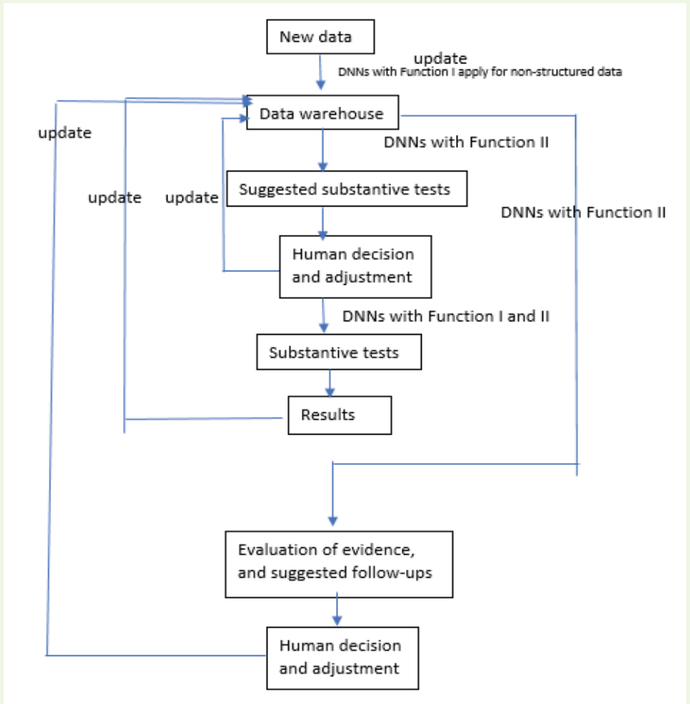
Phase 1. Audit Planning



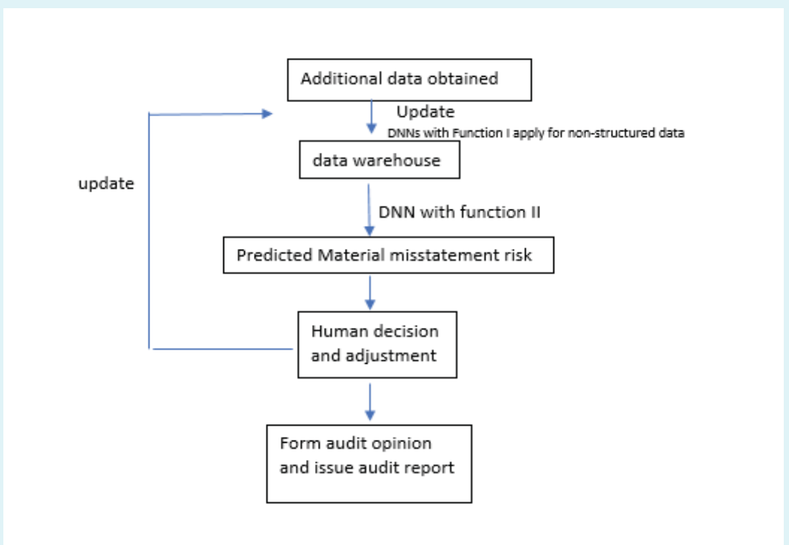
Phase 2. Initial Internal Control Evaluation



Phase 3. Substantive test



Phase 4. Completion



RUTGERS
Rutgers Business School
Newark and New Brunswick

Geographic Industry Clusters and Audit Quality

Cheng Yin, Alexander Kogan, and Miklos Vasarhelyi

Introduction

- Using a large sample of audit client firms, from 2000 to 2015, this paper examines whether there is a difference in audit quality between firms within "geographic industry clusters" and those firms outside clusters.
- We define the geographic industry clusters as the agglomerations of firms from the same industry, located in the same metropolitan statistical area (MSA). Further, based on a significant negative effect of geographic industry clusters on audit quality, we also investigate the reasons that foster such quality gap.
- In addition, we also find that the geographic industry clusters have a positive effect on audit pricing and the existence of local connection intensifies such impact.
- Overall, our evidence suggests that due to the lower communication cost in the geographic industry clusters, clients are more likely to learn questionable accounting practices and form alliances to negotiate with auditors and convince them to accept questionable accounting practices. For a fear of losing clients, auditors charge clients within the clusters higher audit fees to compensate the raising litigation risks, especially those clients with local connections.

Literature Review and Hypotheses 1/4

- As argued in prior literature, firms within clusters behave in ways similar to their local peers. Examles include similar investment patterns (Dougal Parsons, and Titman 2015), strong co-movement in stock returns (Pirinsky and Wang 2006), and a great degree of co-movement in fundamentals (Engelberg et al. 2013). The similarities of local/neighboring firms allow auditors to collect more useful, relevant and timely information to generate effective benchmarks in audit analytical procedures, resulting in a better audit quality.
- On the other hand, as argued by extant research (Kedia et al. 2007), the perceived cost of adopting aggressive accounting practice by a certain firm is largely influenced by its' neighboring firms. The probability of a firm adopting/using aggressive accounting practices is positively associated with the increasing number of wrongdoing neighboring firms. Thus, within an industry cluster, the spread of aggressive accounting practices may provide auditors biased or unreliable accounting information, which weakens the effectiveness of audit analytics based on accounting numbers.
- Hypothesis 1: There is a negative effect of the geographic industry cluster on audit quality.

Literature Review and Hypotheses 2/4

- As the agglomeration of companies within the same industry facilitates a face-to-face communication between local industry companies (Choi et al. 2012), we believe such connection imposes an insidious plot in persuading auditors to forsake strict inspections and hold back qualified opinions on questionable industry practice.
- In this paper, we treat the number of local industry companies sharing the same auditor as a measure of connection between local companies.
- All in all, we expect the degree of local "connection" between local industry competitors to moderate the effects of the geographic industry clusters on audit quality. To provide empirical evidence of this prediction, we test the following hypothesis.
- Hypothesis 2: The negative effect of geographic industry clusters on audit quality is more pronounced for clients with more local industry competitors, whom they share the same auditors with, all else equal.

Literature Review and Hypotheses 3/4

- Prior studies documented that audit fees are mainly determined by the input efforts and the risk exposure. (Alan I. Blankley et al. 2009) On one hand, the agglomeration of companies within the same industry allows auditors to profit from economies of scale. This is because a more similar and relevant pool of local peers can afford auditors information advantages, saving on costs by reducing repetitive procedures such as generating industry benchmark and gathering local information. The reduced input efforts lead to lower audit fees.
- On the other hand, the uncertainty caused by the acceptance of questionable industry practices may leave a higher litigation risk. To hedge against these potential losses, auditors may charge higher audit fees.
- Following our previous hypothesis, we believe the contaminated accounting information environment and the decreasing audit quality within the geographic industry clusters will lead the auditors to charge higher audit fees to compensate for their extra efforts and excessive litigation risks.
- Hypothesis 3: The auditors will charge the clients within the geographic industry clusters higher audit fees than those clients outside the clusters.

Literature Review and Hypotheses 4/4

- In a similar vein, our fourth hypothesis is going to investigate the moderating effects of the local "connections" on the association between geographic industry clusters and audit fees. Unlike using the number of local industry competitors whom a client share the same auditor with as the measure of local "connection", we use the existence of local industry competitors as the measure of local "connection" .
- Consistent with our second hypothesis, we anticipate that the existence of local "connection" has possibilities to herd clients towards offering higher audit fees on their own.
- Specifically, when a client successfully negotiates with an auditor on an industry questionable practice by ceding the power of asking for a lower audit fee, other local industry competitors can easily learn such behaviors and offer higher audit fees and capitalize on the opportunity continue using questionable accounting practice.
- Hypothesis 4: A premium will be charged by auditors to clients with local connections within geographic industry clusters, when compared to fees paid to other clients.

Empirical Results

- Geographic Industry Cluster and Audit Quality

	DA_1	DA_1	DA_1	DA_2	DA_2	DA_2
CMV	-0.008***			-0.006***		
	(-5.05)			(-3.91)		
DUM		-0.006***			-0.006***	
		(-3.10)			(-2.69)	
ROF			-0.010**			-0.012**
			(-2.15)			(-2.19)

- Local Connection and Audit Quality

	DA_1	DA_1	DA_1	DA_2	DA_2	DA_2
CMV*LCONNECTION	-0.002*			-0.003*		
	(-1.65)			(-1.67)		
DUM*LCONNECTION		-0.005***			-0.006***	
		(-2.79)			(-3.08)	
ROF*LCONNECTION			-0.007*			-0.010**
			(-1.67)			(-2.20)

	LAF	LAF	LAF
CMV	0.105***		
	(7.11)		
DUM		0.135***	
		(8.13)	
ROF			0.434***

- Geographic Industry and Audit Fee

	LAF	LAF	LAF
CMV*CONNECTION_DUM	0.060**		
	(2.05)		
DUM*CONNECTION_DUM		0.086**	
		(2.44)	
ROF*CONNECTION_DUM			0.047
			(0.72)

RUTGERS
Rutgers Business School
Newark and New Brunswick

The Information Content of Risk Disclosures: Evidence from Credit Ratings

Michael Chin, Yue Liu and Kevin Moffitt

Introduction

- Beginning in December 1, 2005, public firms are required by the Securities and Exchange Commission (SEC) to disclose risk factors in a separate section - Item 1A, in their annual reports. Risk factor section is required to discuss factors that make an investment in the registrant's securities speculative or risky and should be concise and organized logically. SEC encourages firms to prioritize important risk factors in their disclosure and states on the official website that "companies generally list the risk factors in order of their importance." (SEC 2011).
- This paper is the first to consider whether the order in which disclosures are presented reflects the relative severity of the risks.
- We find relative position, or rank, is the most informative attribute of the credit risk disclosures. Risk order is significantly associated with firms' credit rating levels, and rank increases predict future credit rating downgrades. The predictive power of credit risk rank persists, even when controlling for changes in bond spreads.

Related Literature

- Prior studies find evidence that textual risk disclosures are informative and can enhance the risk perceptions of investors (Kravet and Muslu 2013; Campbell et al. 2014; Bao and Datta 2014; Hope, Hu, and Lu 2016; Gaulin 2017; Filzen 2015; Campbell et al. 2016).
- Campbell et al. (2014) suggest future research on the informativeness of risk factor disclosure in debt markets.
- Recent studies find the information placement in earnings announcement provides incremental information for investors (e.g., Elliott 2006; Huang, Nekrasov, and Teoh 2013; Bowen, Davis, and Matsumoto 2005; Allee and DeAngelis 2015; Cheng, Roulstone, and Van Buskirk 2017).
- This paper responds to Campbell et al. (2014) and studies the information content of risk factor disclosure in terms of credit risks. In addition, the paper extends the examination of information placement to risk factor disclosure and finds evidence that firms meaningfully rank risk factors to reflect their importance.

Data and Risk Disclosure Measures

- Annual reports are downloaded from EDGAR website. For each annual report, Item 1A section is extracted and split into individual risk factors in the same method described in (Campbell et al. 2014). At the same time, the order of each risk factor and the total number of risk factors disclosed in Item 1A are recorded.
- Using a list of key words, a program is used to search within the summaries of risk factors to identify credit risks. The key word list includes the following words and phrases: bankrupt, bankruptcy, downgrade, bonds, loan(s), default, distress, debt (s), obligation(s), credit rating, covenant violation (s), line of credit(s), credit line(s), rating agency, S&P, Standard & Poor, Standard and Poor, Moody's, Fitch. .
- Examples of successfully identified credit risk summaries:
  - **Example 1.** Our substantial indebtedness could adversely affect our operations, including our ability to perform our **obligations** under the notes and our other **debt obligations**.
  - **Example 2.** Our access to liquidity may be negatively impacted if disruptions in credit markets occur, if **credit rating** downgrades occur or if we fail to meet certain covenants. Funding costs may increase, leading to reduced earnings.

- For the identified credit risks, we calculate the length and specificity of the risk factors.
- Credit rating data and other related financial data come from COMPUSTAT and CRSP databases. The final merged dataset contains a sample of firm-year observations during fiscal year 2005 to 2016.
- In order to quantify the importance of identified credit risks in Item 1A, two measures of risk order are employed: CREDIT_RISK_RANK and CREDIT_RISKDECILE. CREDIT_RISK_RANK is calculated by 1 minus the original order of the credit risk divided by the total number of risks in the Item 1A, and set to 0 if there is no credit risk disclosed. CREDIT_RISKDECILE equals 1 if the credit risk is in the 10th decile of the Item 1A, 2 if the credit risk is in the 9th decile of the Item 1A, and so on, and CRDIT_RISKDECILE is set to 1 when no credit risk is disclosed in the Item 1A.

Research Method

- The following models are used to examine the association between credit risk rank and credit ratings:

$$SPRATENUM_t = \beta_0 + \beta_1 CREDIT\_RISK\_RANK_t(CREDIT\_RISKDECILE_t) + \beta_2 CREDIT\_RISK_t + \beta_3 TOTALRISKS_t + \beta_4 SPECIFICITY_t + \beta_5 LENGTH_t + \beta_6 INTEREST\_COVERAGE_t + \beta_7 OPINC\_SALE_t + \beta_8 LD\_AT_t + \beta_9 DT\_AT_t + \beta_{10} SIZE_t + \beta_{11} BETA_t + \beta_{12} SE_t + \beta_{13} R\&D\_EXPENSE_t + \beta_{14} TANGIBILITY_t + \varepsilon_t \quad (1)$$

$$SPRATENUM_t = \beta_0 + \beta_1 CREDIT\_RISK\_RANK_{t-1}(CREDIT\_RISKDECILE_{t-1}) + \beta_2 CREDIT\_RISK_{t-1} + \beta_3 TOTALRISKS_{t-1} + \beta_4 SPECIFICITY_{t-1} + \beta_5 LENGTH_{t-1} + \beta_6 INTEREST\_COVERAGE_t + \beta_7 OPINC\_SALE_t + \beta_8 LD\_AT_{t-1} + \beta_9 DT\_AT_t + \beta_{10} SIZE_t + \beta_{11} BETA_t + \beta_{12} SE_t + \beta_{13} R\&D\_EXPENSE_t + \beta_{14} TANGIBILITY_t + \varepsilon_t \quad (2)$$

$$DOWNGRADE_t(UPGRADE_t) = \beta_0 + \beta_1 MOVEUP_t(MOVEDOWN_t) + \beta_2 SPRATENUM_{t-1} + \beta_3 DOWNGRADE_{t-1}(UPGRADE_{t-1}) + \beta_4 \Delta CREDIT\_RISK_t + \beta_5 \Delta TOTALRISKS_t + \beta_6 \Delta SPECIFICITY_t + \beta_7 \Delta LENGTH_t + \beta_8 \Delta INTEREST\_COVERAGE_t + \beta_9 \Delta OPINC\_SALE_t + \beta_{10} \Delta LD\_AT_t + \beta_{11} \Delta DT\_AT_t + \beta_{12} \Delta SIZE_t + \beta_{13} \Delta BETA_t + \beta_{14} \Delta SE_t + \beta_{15} \Delta R\&D\_EXPENSE_t + \beta_{16} \Delta TANGIBILITY_t + year\_dummy + industry\_dummy + \varepsilon_t \quad (3)$$

$$DOWNGRADE_t(UPGRADE_t) = \beta_0 + \beta_1 MOVEUP_{t-1}(MOVEDOWN_{t-1}) + \beta_2 SPRATENUM_{t-1} + \beta_3 DOWNGRADE_{t-1}(UPGRADE_{t-1}) + \beta_4 \Delta CREDIT\_RISK_{t-1} + \beta_5 \Delta TOTALRISKS_{t-1} + \beta_6 \Delta SPECIFICITY_{t-1} + \beta_7 \Delta LENGTH_{t-1} + \beta_8 \Delta INTEREST\_COVERAGE_t + \beta_9 \Delta OPINC\_SALE_t + \beta_{10} \Delta LD\_AT_t + \beta_{11} \Delta DT\_AT_t + \beta_{12} \Delta SIZE_t + \beta_{13} \Delta BETA_t + \beta_{14} \Delta SE_t + \beta_{15} \Delta R\&D\_EXPENSE_t + \beta_{16} \Delta TANGIBILITY_t + year\_dummy + industry\_dummy + \varepsilon_t \quad (4)$$

Results

- Level models (firm-fixed effect regression)

VARIABLES	(1) SPRATENUM	(2) SPRATENUM	(3) SPRATENUM	(4) SPRATENUM
CREDIT_RISK_RANK	0.4712***			
	(4.32)			
CREDIT_RISKDECILE		0.0368***		
		(3.93)		
$CREDIT\_RISK\_RANK_{t-1}$			0.4224***	
			(4.02)	
$CREDIT\_RISKDECILE_{t-1}$				0.0350***
				(4.04)
CONTROL VARIABLES	controlled	controlled	controlled	controlled
N	9,595	9,595	8,320	8,320
R-squared	0.9464	0.9463	0.9517	0.9517
Fixed effects	Firm & Year	Firm & Year	Firm & Year	Firm & Year
Cluster	Firm & Year	Firm & Year	Firm & Year	Firm & Year

- Change models

VARIABLES	(1) DOWNGRADE	(2) DOWNGRADE	(3) DOWNGRADE
MOVEUP1	0.4490***		
	(5.37)		
MOVEUP2		0.2330**	
		(2.31)	
MOVEUP3			0.3791***
			(3.20)
CONTROL VARIABLES	controlled	controlled	controlled
Observations	6,553	6,553	6,553
Pseudo R-square	0.1986	0.1950	0.1963
Fixed effects	Industry & Year	Industry & Year	Industry & Year
Cluster	Firm & Year	Firm & Year	Firm & Year

VARIABLES	(1) DOWNGRADE	(2) DOWNGRADE	(3) DOWNGRADE
$MOVEUP1_{t-1}$	0.3379***		
	(3.25)		
$MOVEUP2_{t-1}$		0.4799***	
		(3.38)	
$MOVEUP3_{t-1}$			0.5603***
			(6.06)
CONTROL VARIABLES	controlled	controlled	controlled
Observations	6,433	6,433	6,433
Pseudo R-square	0.1976	0.1985	0.1997
Fixed effects	Industry & Year	Industry & Year	Industry & Year
Cluster	Firm & Year	Firm & Year	Firm & Year

Additional Test and Conclusion

- Three additional tests are performed to test the robustness of the results.
  1. There may be concern that the association between the MOVEUP1 (MOVEUP2) and DOWNGRADE can be driven by the changes in the total number of risks. Therefore, a new variable MOVEUP3 is generated which equals 1 if the order of credit risk moves toward the top of the disclosure and 0 otherwise, regardless of the total number of risks in Item 1A. Using MOVEUP3 for the change models, the results are robust.
  2. We test whether the explanation power of risk order will be taken away by bond spread, and results show that the predictive power of credit risk order persists, even when controlling for changes in bond spreads.
  3. Using ordered logistic regression for level models, the association between credit risk order and credit rating is robust.
- This paper provides evidence that disclosures of credit risks can be used as an indicator of credit rating changes: the order of credit risks provides incremental information about current as well as future credit rating. The finding also indicates that firms do meaningfully rank risk factors to reflect their importance.

RUTGERS
Rutgers Business School
Newark and New Brunswick

Cloud-based In-memory Columnar Database Architecture for Continuous Audit Analytics

Yunsen Wang and Alexander Kogan

Introduction

- In the era of big data, audit profession is starting to leverage the emerging data analytic techniques (e.g., deep learning, process mining) to examine financial data, evaluate internal control effectiveness, and detect fraudulent transactions.
- In order to apply audit analytics to examine a client's business data, an auditor needs to periodically extract the full population of transactions (e.g., purchase orders, invoice receipts) from the client's Enterprise Resource Planning (ERP) system.
- To examine such high volume transactions, continuous auditing systems (Vasarhelyi and Halper 1991) are designed to systematically extract transaction data from ERP systems, test every transaction entry, and report exceptions or anomalies in close to real time (Alles, Kogan, and Vasarhelyi 2008; Kogan et al. 2014).
- In-memory columnar database system is such an infrastructure that supports high-speed data analytics using main memory as the primary storage (Garcia-Molina and Salem 1992; Plattner 2009).
- This study introduces the architecture of the modern in-memory columnar database system and proposes a design of applying the new database for high-speed continuous audit analytics.

Technology Breakthroughs

In-Memory Database System

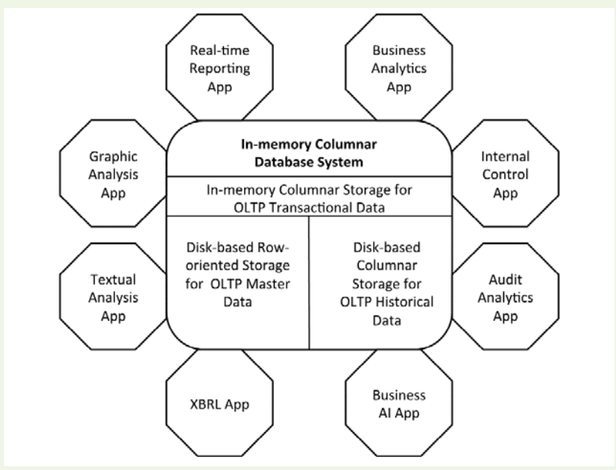
- In a conventional DBMS, data resides permanently in hard disk and will be loaded into main memory when needed, while in the modern in-memory database system (IMDB), data resides permanently in main physical memory. As data in main memory can be directly accessed by multi-core CPUs, IMDB has better response time and transaction throughputs (Plattner 2009).

Row-oriented Storage and Columnar Storage

- In contrast to conventional row-oriented storage, the values for each attribute are stored contiguously in the column-oriented storage; therefore, its compression efficiency is usually 4 to 5 times that of row-oriented storage (Abadi, Madden and Ferreira 2006). Moreover, a complex analytical query could be fast responded to as data aggregation in columnar storage outperforms row-oriented storage, especially in the case of large number of data items. By applying columnar storage schemas to IMDB, the novel in-memory columnar database would be superior to row-oriented IMDB with regards to memory consumption.

Design: In-Memory Columnar Database for Continuous Audit Analytics

- As all operational data resides in main memory, conclusion materialization can be performed on the fly immediately with high efficiency. In this case, no data warehouse needs to be created separately from operational databases because selection, aggregation and analysis can be performed very efficiently in main memory.

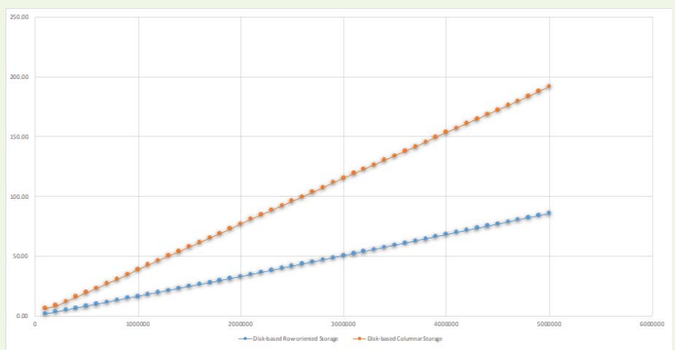


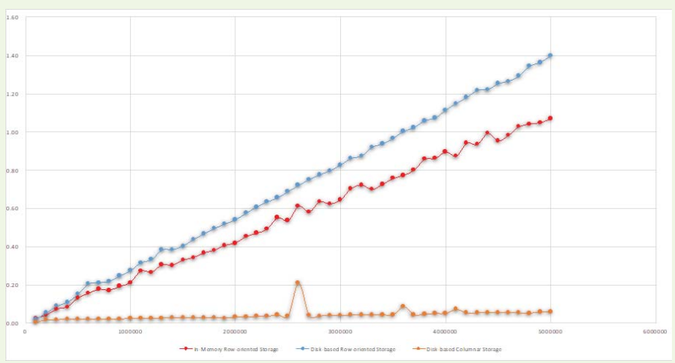
In-memory Columnar Database Performance Evaluation

- This study creates a number of artificial transactions to be stored in those three databases. The number of artificial transactions varies from 100,000 to 5,000,000 and each transaction is represented by five numeric attributes. After creating the tables storing the transaction data, the tables will be immediately stored on disk or in memory. This study measures the sizes and durations of two groups of databases on disk (i.e., disk-based row-oriented database and disk-based columnar database).

Computational Complexity Level of Audit analytics

	Time (s)			Size (MB)		
# of transactions	In-Memory Row-oriented Storage	Disk-based Row-oriented Storage	Disk-based Columnar Storage	In-Memory Row-oriented Storage	Disk-based Row-oriented Storage	Disk-based Columnar Storage
100000	0.03	0.02	0.01	\	1.63	6.58
200000	0.04	0.05	0.02	\	3.28	8.42
300000	0.07	0.09	0.02	\	4.93	12.17
400000	0.09	0.11	0.02	\	6.58	15.92
500000	0.13	0.15	0.02	\	8.23	19.67
600000	0.16	0.21	0.02	\	9.88	23.42
700000	0.18	0.21	0.02	\	11.53	27.17
800000	0.17	0.22	0.02	\	13.18	30.92
900000	0.19	0.25	0.02	\	14.83	34.67
1000000	0.21	0.28	0.03	\	16.47	39.05
1100000	0.27	0.32	0.03	\	18.12	42.80
1200000	0.27	0.33	0.03	\	19.77	46.55
1300000	0.31	0.38	0.03	\	21.42	50.30
1400000	0.30	0.39	0.03	\	23.07	54.05
1500000	0.33	0.40	0.03	\	24.71	57.80
1600000	0.34	0.44	0.03	\	26.36	61.55
1700000	0.37	0.47	0.03	\	28.01	65.30
1800000	0.38	0.50	0.03	\	29.66	69.05
1900000	0.41	0.52	0.03	\	31.31	72.80
2000000	0.42	0.54	0.03	\	32.96	77.17





- The disk-based columnar oriented databases perform much better as it increases much slower than the other two databases, which shows greater efficiency in terms of aggregation queries. There are a number of spikes in the computation timeline, which may be due to the fact that the experiment was conducted on a shared server, therefore the CPUs might not be able to serve only this analysis.

Deployment on Cloud Computing

- Cloud computing offers an economical way for many companies to deploy the in-memory columnar database for ERP. This research proposes implementing in-memory columnar databases on cloud computing to facilitate automatic and continuous audit analytics.
- Due to the difficulty of overcoming the hurdle of significant upfront fixed cost, the deployment on the cloud could be cost-effective for a firm to gain fast access to the in-memory database systems, which is budget affordable and easy to implement.
- A cloud solution is maintained by many IMDB specialists for operation and update. It will save expenditures for hiring IMDB specialists on site. For example, a company with $1 billion in revenue is likely to have 50-plus applications running at a time. With cloud access, the company relies on the cloud infrastructure and service to manage the business data. Deployment on the cloud would be the best solution for Small and Medium Size Enterprises (SMEs) or large firms starting investing in IMDB.

RUTGERS
Rutgers Business School
Newark and New Brunswick

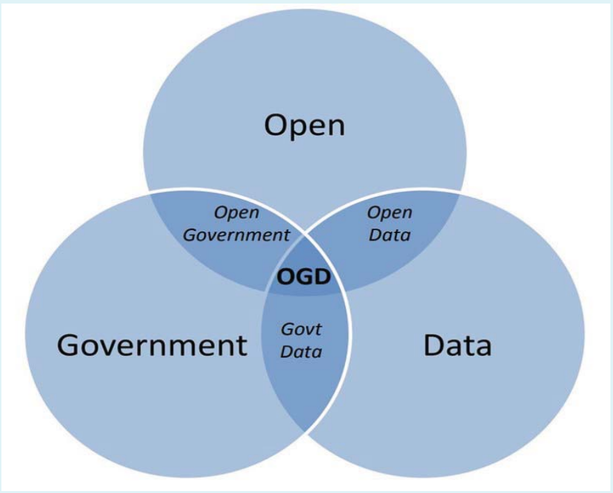
Financial Reporting and Open Government Data Initiatives:

A Comparison Between Brazil and Saudi Arabia

Zamil S. Alzamil, Deniz Appelbaum, and Miklos Vasarhelyi

Introduction and Background

- What is Open Government Data (OGD)?
- Open Data as defined by the Open Definition: "Open data and content that can be **freely used, modified**, and **shared** by **anyone** for **any purpose.**"
- Open Government Data (OGD) is described by Maude, F. (2012) as "Public Sector Information that has been made available to the public as Open Data."
- Growing practice of open government data (OGD): from the origin of Freedom of Information (FOIA), June 1966, in the U.S., up to the present day – a particular expansion during the 2010s.
- Achieving transparency as a goal for OGD can be very challenging. Transparency can play a main role in government decision making. By making government data available to the public, citizens, professionals, and other interest groups can access the data to help monitor public spending, and increase overall participation. Thus, enabling the government officials to make better decisions.



Research Objectives

- This paper considers the process and the use of open government data (OGD) initiatives by focusing on financial reporting with a comparison of procurement contracts of the Federal Republic of Brazil and Saudi Arabia. Comparing the Republic of Brazil OGD initiatives with Saudi Arabia's OGD is because Brazil is part of the Open Government Partnership. Furthermore, Brazilian OGD are considered advanced, and thus, comparing them to Saudi Arabia's OGD would be of great interest.
- It pursues two main arguments regarding:
  - I. the possibility of disseminating more financial data in Saudi Arabia, especially procurement tenders, and how this could transform relationships between different levels of government and citizens;
  - II. and also it assesses the level of data transparency based of the definition of open data, and more importantly, the paper suggests new dimensions to the open data concept when utilized by governments.
- In addition to the data availability, openness, and access that defines open data (Group, 2017), we add two more dimensions which are data analytics, and applications within governments.

Proposed Model

- Here, we will discuss the level of data transparency of government procurement contracts (GPC) between the Republic of Brazil and the Kingdom of Saudi Arabia based on a 4-condition model:

1. Data availability of government's procurement contracts (GPC).
2. Data Openness: once the data is available, into what extent the data is open.
3. Data Analytics: if the data is available and open to an extent, is it analyzable.

Level of Data Transparency of Procurement Tenders

- The following table shows the level of transparency of procurement tenders presented at the Global Open Data Index which measures open data around the world, where the green color denotes 'Yes', the red color denotes 'No', and the blue color denotes 'Unsure':

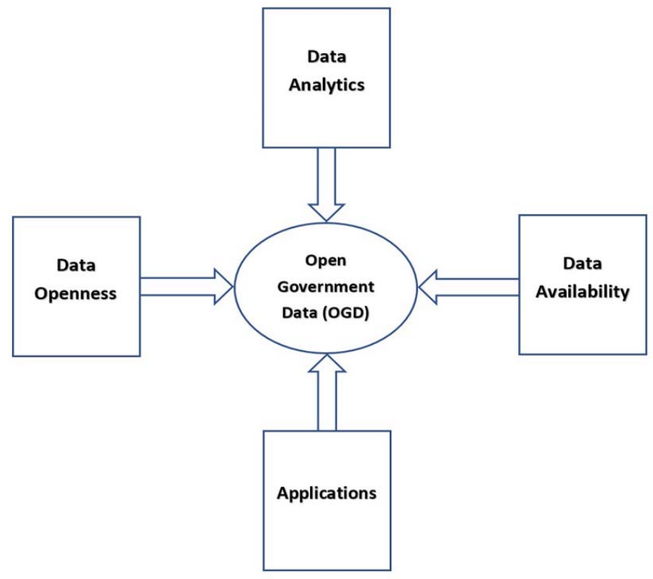
Place	Score	Breakdown*	Year	Location (URL)
United States	100%	*	2015	https://catalog.data.gov
Australia	100%		2015	http://data.gov.au
United Kingdom	90%		2015	https://www.gov.uk/contracts-finder
Spain	90%		2015	https://contrataciondelestado.es
Brazil	80%		2015	http://dados.gov.br
Denmark	45%		2015	http://envs.au.dk
Saudi Arabia	35%		2015	http://www.csc.org.sa

***Breakdown (Data Availability)**

Openly Licensed · Available for free · Machine readable · Available in bulk · Up to date · Publicly available · In digital format · Available online · Does the data exist

Table 1: Procurement Tenders Evaluation [13]

Proposed Model for Effective and Efficient Open Governments Data (OGD).



Continued

Data Availability & Openness

- In this section, we will briefly describe the 4-condition model:
- In Saudi Arabia: the government's procurement contracts are limited in availability to the public.
- There isn't yet a central place for the GPC.
- Some contracts could be found at the Council of Saudi Champers website at: http://www.csc.org.sa; others could be found at the Ministry of Finance portal at: https://www.mof.gov.sa.
- The all-time total number of government procurement contracts presented are around 12,534 only (HTML).
- Most contracts are missing contract number, class, signature date, and effective start and end dates.
- While Brazil provides a central place for all OGD includes GPC with 470,683 published procurement contracts prepared in ready-to-use format.
- This paper collect a sample of procurement/ contract data from the Council of Saudi Champers website and presents it in a machine-readable format. And show the potential of analysis.
- In Brazil, http://www.dados.gov.br, the open data portal, provides some ready-to-use visual analysis tools that could be utilized by the public.
- In Saudi Arabia and Brazil, there isn't any existence for applications or centralized place that facilitate communication between government officials and citizens.

Data Analytics & Application

- Using Saudi Arabia's open procurement/ contract data, the data is neither analyzed nor prepared in ready-to-use format.
- This paper collect a sample of procurement/ contract data from the Council of Saudi Champers website and presents it in a machine-readable format. And show the potential of analysis.
- In Brazil, http://www.dados.gov.br, the open data portal, provides some ready-to-use visual analysis tools that could be utilized by the public.
- In Saudi Arabia and Brazil, there isn't any existence for applications or centralized place that facilitate communication between government officials and citizens.

Contribution and Further Research

- This paper discusses open government data among different countries, focusing on the cases of the Republic of Brazil and Saudi Arabia open government data initiatives. It first compares the different financial reporting and auditing systems in Brazil and Saudi Arabia. Second, the paper gives an overview of OGD initiatives in different countries.
- In addition, it assesses the level of data transparency based on the definition of open data, and more importantly, this study proposes a new model which expands the open data definition to include its potential to encourage a better feasible decision making by governments.
- The assessment of the proposed attributes for data transparency has been conducted using procurement contracts available at the Republic of Brazil and Saudi Arabia's open data portals. We found that OGD initiatives in Saudi Arabia lack appropriate datasets, formats, analytical tools, and applications. This situation could be improved dramatically by enhancing the government efficiency; whereas the Brazilian OGD initiatives lack applications.
- We found out that there is a gap between data analytics and applications for reporting and interacting with public officials. Open government data should be analyzable, and the subsequent results should be actionable, for the evolution of better governments. Our long-term goal of this research is to develop a unified framework that applies our proposed 4-condition model for better government data transparency.

RUTGERS
Rutgers Business School
Newark and New Brunswick