World Bank
December 2, 2020

SHAPING THE FUTURE IN BUSINESS
MEASUREMENT AND ASSURANCE:
EMBRACING TECHNOLOGY AND CHANGE

Creating a Digital Strategy
Outline

• Introduction
  – The CarLab

• Some of our projects
  – GASB – PIR
  – Exogenous Data
  – PIOB – what is public interest?
  – Machine Learning for accounting estimates
  – Big data analytics
  – Cooperation with the Volcker Alliance
  – Continuous assurance of medication procurement of a state
  – NYC cleanliness with Tweet text mining
  – Continuous pandemic monitoring
INTRODUCTION

The CarLab
All academic Accounting programs around the world are ranked annually by BYU. For many years now, the Accounting Information Systems (AIS) group at RBS has led the world in the application of information technology to the audit profession. We are very proud to announce that the just-released BYU rankings for 2019 confirm again the continued success of Rutgers Business School in both AIS and audit research:
Updating Dashboard with document links

http://raw.rutgers.edu/CAR%20Lab%20Directory/Sign-in.html

PASSWORD: RARC777
GASB
Post-Implementation Review Project

Ben Yoon
Huaxia Li
Kevin Moffitt

Rutgers CarLab
July 2020
This project will build a dynamic information system that
1) automatically captures the CAFRs from different governmental entities,
2) parses relevant items from the CAFRs, and
3) converts them into a structured data

The structured data be easily used by the GASB to perform the post-implementation review (PIR) of the new GASB pension standards.

* In 2012, the GASB announced new pension standards (No. 67 and 68).
4 Steps of This Project

- This project consists of 4 steps.

- Rutgers has conducted initial pilot tests.
  - Step1: Collecting 36,676 CAFRs from 3 repositories
  - Step2: Converting PDF documents
  - Step3: Extracting 8 items from the CAFRs
  - Step4: Report with Excel format
Exogenous data analytics for Auditing

Miklos A. Vasarhelyi
Helen Brown Liburd

Rutgers Business School
Some sources

- Amazon sales
- Google searches
- Apps used
- Calls made
- GPS or JEEP location
- Sites accessed
- Car license plates photographed
- Pictures of parking lots
- Face recognition pictures
- Site clickpaths
Exogenous Data

- Social Media
- Internet of Things
- Click data
- Locational data
- Search Data
- Economic data
- Weather data
- Internet sales data

ED may be of easier access
ED is likely less tamperable
ED relationships will be stochastic

ED is a form of confirmation
ED may complement many current procedures
ED may create many new procedures
IFAC / PIOB project

Kevin Moffitt
Ben Yoon
Hiaxia Li
Rutgers/ PIOB / IFAC Project

• The Public Interest Oversight Board (PIOB) is the global independent oversight body that seeks to improve the quality and public interest focus of the international audit and assurance, and ethics standards formulated by the Standard Setting Boards supported by the International Federation of Accountants (IFAC).

• The Rutgers team will automatically identify public interest regarding auditing from investor, regulator, and professional accounting websites.
Problem: Automatically identify public interest from investors, regulators, etc…

- Collected and cleaned 7159 documents from 5/40 identified organizations
  - ESMA - 4145
  - IOSCO - 1164
  - SIFMA - 959
  - THEIA - 731
  - ICGN – 160
- Searched 30 topics identified by PIOB

<table>
<thead>
<tr>
<th></th>
<th>conflict of interest</th>
<th>fee dependency</th>
<th>objectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>audit deficiencies</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>audit quality</td>
<td>critical mindset</td>
<td>fraud</td>
<td>professional skepticism</td>
</tr>
<tr>
<td>auditor independence</td>
<td>ethical behavior</td>
<td>going concern</td>
<td>…and more…</td>
</tr>
<tr>
<td>Topic</td>
<td>Source</td>
<td>Text</td>
<td></td>
</tr>
<tr>
<td>------------------------------</td>
<td>-------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Audit Quality</td>
<td><a href="http://www.IOSCO.org">www.IOSCO.org</a> 2018</td>
<td>“this may provide an effective safeguard that a decision is not unduly influenced by a low audit fee in circumstances where audit quality may be compromised.”</td>
<td></td>
</tr>
<tr>
<td>Auditor Independence</td>
<td><a href="http://www.ICGN.org">www.ICGN.org</a> 2018</td>
<td>“so, as you can see, while things have changed since the passage of the sarbanes-oxley act, it appears that new threats to auditor independence have emerged, and that others have reappeared.”</td>
<td></td>
</tr>
<tr>
<td>Going Concern</td>
<td><a href="http://www.ESMA.Europa.eu">www.ESMA.Europa.eu</a> 2019</td>
<td>“accordingly, the fair value of the land should be determined based on the current use of the land in view of the going concern principle.”</td>
<td></td>
</tr>
<tr>
<td>Fraud</td>
<td><a href="http://www.THEIA.org">www.THEIA.org</a> 2019</td>
<td>“would it be possible to devise a 'reasonable person' test in assessing the auditors work in relation to fraud detection?”</td>
<td></td>
</tr>
<tr>
<td>Professional Skepticism</td>
<td><a href="http://www.THEIA.org">www.THEIA.org</a> 2019</td>
<td>“annually it should assert why it believes the auditor has been challenging and exercised professional skepticism.”</td>
<td></td>
</tr>
</tbody>
</table>
Machine Learning Improves Accounting Estimates

Kexing Ding¹
Baruch Lev²
Xuan Peng³
Ting Sun⁴
Miklos A. Vasarhelyi⁵

¹Southwestern University of Finance and Economics; and Rutgers, the State University of New Jersey
²Stern School of Business, New York University
³Southwestern University of Finance and Economics; and Rutgers, the State University of New Jersey
⁴The College of New Jersey
⁵Rutgers, the State University of New Jersey
Accounting Estimates

• Accounting estimates are highly uncertain and are sometimes manipulated.
• Accounting estimates are difficult to audit and impossible for investors to evaluate.
• Accounting estimates are ubiquitous in financial reporting.
  – Example: account receivables, insurance loss reserves, revenues from contracts, and pension and warranty expenses.
• Researchers have made several proposals:
  – Financial statements disclose which accounts are subject to extreme uncertainty (Christensen et al. 2012).
  – Firms report ex-post realization of critical estimates (e.g., Lundholm 1999).
  – Managers restate earnings in case of large deviations (Lev et al. 2008).
Machine Learning Algorithms

- Linear Regression
- Random Forest
- Gradient Boosting Machine (Gradient Tree Boosting)
- Artificial Neural Networks
## Main Results Summarized

Table 5 Cross-validation results

<table>
<thead>
<tr>
<th>Business line</th>
<th>Training/Validation Sample</th>
<th>Obs</th>
<th>Managers’ estimates</th>
<th>Machine learning without manager estimates</th>
<th>Machine learning with manager estimates</th>
<th>MAE</th>
<th>RMSE</th>
<th>MAE</th>
<th>RMSE</th>
<th>Accuracy edge (MAE)</th>
<th>Accuracy edge (RMSE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Private Passenger Auto Liability</td>
<td>1996-2005</td>
<td>5,949</td>
<td>9,461</td>
<td>37,494</td>
<td>8,213</td>
<td>34,687</td>
<td>13%</td>
<td>7%</td>
<td>7,258</td>
<td>18%</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>1996-2006</td>
<td>6,298</td>
<td>9,793</td>
<td>38,266</td>
<td>7,848</td>
<td>34,547</td>
<td>20%</td>
<td>10%</td>
<td>7,220</td>
<td>26%</td>
<td>21%</td>
</tr>
<tr>
<td></td>
<td>1996-2007</td>
<td>6,602</td>
<td>9,575</td>
<td>37,940</td>
<td>7,869</td>
<td>35,047</td>
<td>18%</td>
<td>8%</td>
<td>6,902</td>
<td>28%</td>
<td>20%</td>
</tr>
<tr>
<td>Commercial Auto Liability</td>
<td>1996-2005</td>
<td>5,383</td>
<td>4,209</td>
<td>18,562</td>
<td>3,565</td>
<td>14,051</td>
<td>15%</td>
<td>24%</td>
<td>3,446</td>
<td>18%</td>
<td>27%</td>
</tr>
<tr>
<td></td>
<td>1996-2006</td>
<td>5,661</td>
<td>4,155</td>
<td>18,375</td>
<td>3,520</td>
<td>13,881</td>
<td>15%</td>
<td>24%</td>
<td>3,266</td>
<td>21%</td>
<td>26%</td>
</tr>
<tr>
<td></td>
<td>1996-2007</td>
<td>5,957</td>
<td>4,338</td>
<td>19,175</td>
<td>3,575</td>
<td>13,671</td>
<td>18%</td>
<td>29%</td>
<td>3,322</td>
<td>23%</td>
<td>32%</td>
</tr>
<tr>
<td>Workers’ Compensation</td>
<td>1996-2005</td>
<td>4,183</td>
<td>11,547</td>
<td>43,652</td>
<td>7,518</td>
<td>29,418</td>
<td>35%</td>
<td>33%</td>
<td>7,144</td>
<td>38%</td>
<td>34%</td>
</tr>
<tr>
<td></td>
<td>1996-2006</td>
<td>4,398</td>
<td>12,360</td>
<td>44,187</td>
<td>7,434</td>
<td>29,387</td>
<td>40%</td>
<td>33%</td>
<td>6,988</td>
<td>43%</td>
<td>39%</td>
</tr>
<tr>
<td></td>
<td>1996-2007</td>
<td>4,645</td>
<td>13,214</td>
<td>47,541</td>
<td>7,298</td>
<td>29,468</td>
<td>45%</td>
<td>38%</td>
<td>6,861</td>
<td>48%</td>
<td>44%</td>
</tr>
<tr>
<td>Commercial Multi-Peril</td>
<td>1996-2005</td>
<td>5,235</td>
<td>5,737</td>
<td>27,615</td>
<td>5,103</td>
<td>22,060</td>
<td>11%</td>
<td>20%</td>
<td>4,854</td>
<td>15%</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>1996-2006</td>
<td>5,457</td>
<td>5,871</td>
<td>27,931</td>
<td>5,151</td>
<td>23,404</td>
<td>12%</td>
<td>16%</td>
<td>4,968</td>
<td>15%</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>1996-2007</td>
<td>5,846</td>
<td>6,017</td>
<td>28,349</td>
<td>4,963</td>
<td>22,556</td>
<td>18%</td>
<td>20%</td>
<td>4,534</td>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>Homeowner/Farmowner</td>
<td>1996-2005</td>
<td>6,121</td>
<td>3,905</td>
<td>16,789</td>
<td>5,674</td>
<td>22,069</td>
<td>-5%</td>
<td>-31%</td>
<td>4,402</td>
<td>16%</td>
<td>3%</td>
</tr>
<tr>
<td></td>
<td>1996-2006</td>
<td>6,544</td>
<td>3,878</td>
<td>16,611</td>
<td>5,687</td>
<td>21,070</td>
<td>-4%</td>
<td>-27%</td>
<td>4,203</td>
<td>16%</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>1996-2007</td>
<td>6,946</td>
<td>3,962</td>
<td>16,826</td>
<td>5,548</td>
<td>21,269</td>
<td>-4%</td>
<td>-26%</td>
<td>4,321</td>
<td>16%</td>
<td>1%</td>
</tr>
</tbody>
</table>
Concluding Remarks

- There is an urgent need to enhance the quality of accounting estimates and auditors’ ability to independently evaluate the reliability of these estimates.

- Machine learning can generate accounting estimates useful for auditors to evaluate managers’ estimates, and for managers to generate original estimates.

- These findings should be of value for consideration of the value of machine learning models to standard setters.

- More research is needed to generalize the application of machine learning in other accounting settings.
Big data and algorithmic trading against periodic and tangible asset reporting: the need for U-XBRL

Dr. Miklos A. Vasarhelyi
KPMG Distinguished Professor Rutgers Business School - Newark & New Brunswick
Director, Rutgers Accounting Research Center & Continuous Auditing & Reporting Lab

Duo (Selina) Pei
PhD Student Rutgers Business School - Newark & New Brunswick
Some other phenomena observed

Exogenous Data
- Mobility Data
- Weather
- Sales
- Economic

Locational Data
- Internet of Things
- Social Media

Internal Data
- Intangibles
- Human Resources
- Peer-to-Peer
- Order-to-Cash
- Supply Chain
- Inventory
- General Ledger
- Internet Protocol

Interlinkages
We need a reporting schema that also integrates well with current advances in auditing.
So what kind of information may be missing?

- SASB metrics
  - Intellectual property and data privacy (Industry-specific)
  - Integrity and recruitment/retention programs (Overall)
- “The relationships between [an organization’s] various operating and functional units and the capitals that the organization uses or affects” in the International Integrated Reporting Council Framework (2013)
APPLICATIONS OF DATA ANALYTICS: VISUALIZATION AND CLUSTER ANALYSIS OF GOVERNMENTAL DATA – TWO CASE STUDIES

ESSAY 2: COOPERATION WITH THE VOLCKER ALLIANCE
OBJECTIVES

• Since data analytics is one way to explore the data and to help uncover hidden relationships
  – In these case studies we plan to explore the literature for the use of emerging data mining techniques in auditing
    ✓ In particular, cluster analysis & visualization techniques as supportive tools to gain more insights into data.

• Conduct two case studies:
  1) Rutgers AICPA Data Analytics Research Initiative (RADAR): A Case Study.
    ✓ Facilitate the integration of different data analytics tools and techniques into the audit process.
  2) Visualization and Clustering Analytics of U.S. states’ on budgeting.
    ✓ Information on U.S. States.

CONTRIBUTION

• We show how visualization and data clustering techniques could be used on governmental data and to help gain more information about financial statements & budgeting.
CONT’D: Moody’s Ratings
CONT’D: Clustering Results
Continuous Monitoring and Audit Methodology for Medication Procurement

Wenru Wang – Rutgers University
Miklos A. Vasarhelyi – Rutgers University
Overview

• Continuous monitoring and audit system for exception and anomaly detections.
Continuous Monitoring Dashboard

Monitoring dashboard - 2017

Medications over R$2,000,000 total value alarm - 2017

<table>
<thead>
<tr>
<th>Item Name (standard)</th>
<th>Description (Standard)</th>
<th>Total Acquisition Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIRA, DETERMINACAO..</td>
<td>SANGUE FAIXA DETECC....</td>
<td>16,400,000</td>
</tr>
<tr>
<td>HIDRALAZINA, CLORID..</td>
<td>25MG COMPRIMIDO</td>
<td>4,231,718</td>
</tr>
<tr>
<td>METFORMINA</td>
<td>850MG COMPRIMIDO</td>
<td>3,630,217</td>
</tr>
<tr>
<td>AMOXICILINA + CLAVULANATO, POTA..</td>
<td></td>
<td>3,545,250</td>
</tr>
<tr>
<td>LOSARTANA POTASSI..</td>
<td>LOSARTANA POTASSIC..</td>
<td>3,297,540</td>
</tr>
<tr>
<td>SINVASTATINA</td>
<td>20MG COMPRIMIDO</td>
<td>3,155,962</td>
</tr>
<tr>
<td>ISOSSORBIDA, MONON..</td>
<td>20MG COMPRIMIDO</td>
<td>2,935,223</td>
</tr>
<tr>
<td>CIPROFLOXACINO SIS..</td>
<td>CLORIDRATO, 2MG/ML</td>
<td>2,774,536</td>
</tr>
<tr>
<td>CLOREDO, SODIO</td>
<td></td>
<td>2,669,204</td>
</tr>
</tbody>
</table>

% of Total Acquisition Value

- Secretaria Municipal de Saúde: 53.61%

Medications over R$2,000,000 unit price alarm

<table>
<thead>
<tr>
<th>Material Code</th>
<th>Item Name (standard)</th>
</tr>
</thead>
<tbody>
<tr>
<td>65054207155</td>
<td>12.25</td>
</tr>
<tr>
<td>65054600671</td>
<td>14.25</td>
</tr>
<tr>
<td>65050820464</td>
<td>9.72</td>
</tr>
<tr>
<td>6505090735</td>
<td>5.90</td>
</tr>
<tr>
<td>65050805317</td>
<td>10.00</td>
</tr>
<tr>
<td>65050811200</td>
<td></td>
</tr>
</tbody>
</table>

% of Total Acquisition Value

- PRATI DONADUZZI & ..: 100%
- ROCHE DIAGNOSTICA..: 80%
- JRG DISTRIBUIDORA..: 60%
- CRISTALIA PRODUTO.: 40%
- COSTA CAMARGO CO.: 20%
- FRESENIUS KABI BRA.: 0%
- EXFARMA LTDA EPP  
New York City Street Cleanliness: Apply Text Mining Techniques to Social Media Information

Huijue Kelly Duan
Mauricio Codesso
Zamil Alzamil

1Rutgers, the State University of New Jersey
2Northeastern University
3Majmaah University
Motivation

- NYC government performs a cleanliness inspection every year, the method has not changed for nearly 50 years.

- NYC districts receive ratings of 90% or higher; therefore, NYC government rates majority of its streets as acceptably clean.

- NYC residents increasingly contact DSNY via 311 about missing trash pickups, overflowing litter baskets, and dirty conditions.

### Monthly SCORECARD Community Board Report - July 2019

**Percent of Acceptably Clean Streets (Citywide Totals) - 2019 / 07**

<table>
<thead>
<tr>
<th>Borough</th>
<th>Acceptable Streets %</th>
<th>Acceptable Streets % - Previous Month</th>
<th>Acceptable Streets % - Year Ago</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manhattan</td>
<td>96.4</td>
<td></td>
<td>96.6</td>
</tr>
<tr>
<td>Bronx</td>
<td>97.3</td>
<td></td>
<td>94.6</td>
</tr>
<tr>
<td>Brooklyn</td>
<td>98.3</td>
<td></td>
<td>94.0</td>
</tr>
<tr>
<td>Queens</td>
<td>98.7</td>
<td></td>
<td>97.7</td>
</tr>
<tr>
<td>Staten Island</td>
<td>100.0</td>
<td></td>
<td>98.0</td>
</tr>
<tr>
<td><strong>Citywide Total</strong></td>
<td>98.4</td>
<td></td>
<td><strong>96.3</strong></td>
</tr>
</tbody>
</table>
Objective

• Examine the social media information
  ➢ to identify temporal trends and patterns of the cleanliness of NYC streets
  ➢ to analyze whether crowdsourcing information is consistent with NYC cleanliness ratings
  ➢ to assess the performance of municipal services via sentiment analysis

@nyc311 @NYCSanitation HARLEM 116th b/t 7th & 8th street is trash & rodent ridden. When is this going to stop? #nowyouknow #getittogether #foodbank #petopia #ctown #kingston

when I was out walking tonight a rat jumped in front of me and I accidentally kicked it .... it was ok but I think @NYC_DOT should study rat crosswalks between street trash collection sites and buildings to avoid future injuries #streetsafety

Im from NYC; Where once u find parking, U cant go out for the remainder of the day 😂😭😭
Workflow

Data Collection

• Twitter API (Streaming API)
• 6.8M collected (8/27/2018 - 5/22/2019)

Data Preparation

• Data Cleaning
• Variable Selection and Aggregation
• Data Aggregation

Relevancy Determination

• Apply Keyword List
• Data Preprocessing
• Supervised Machine Learning
  - Naïve Bayes
  - Random Forest
  - XG Boost

Sentiment Analysis

• Negative
• Positive
• Neutral
Twitter Dashboard

Overall Sentiment

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>Count of Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>62%</td>
</tr>
<tr>
<td>Neutral</td>
<td>28%</td>
</tr>
<tr>
<td>Positive</td>
<td>10%</td>
</tr>
</tbody>
</table>

Descriptive Statistics - Category

<table>
<thead>
<tr>
<th>Sentiment</th>
<th>Homeless</th>
<th>Parking</th>
<th>Street</th>
<th>Subway</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative</td>
<td>10%</td>
<td>16%</td>
<td>31%</td>
<td>5%</td>
<td>62%</td>
</tr>
<tr>
<td>Neutral</td>
<td>3%</td>
<td>7%</td>
<td>15%</td>
<td>2%</td>
<td>28%</td>
</tr>
<tr>
<td>Positive</td>
<td>2%</td>
<td>3%</td>
<td>4%</td>
<td>1%</td>
<td>10%</td>
</tr>
</tbody>
</table>

Sentiment by Category

<table>
<thead>
<tr>
<th>Category</th>
<th>Count of Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homeless</td>
<td>10%</td>
</tr>
<tr>
<td>Parking</td>
<td>16%</td>
</tr>
<tr>
<td>Street</td>
<td>15%</td>
</tr>
<tr>
<td>Subway</td>
<td>5%</td>
</tr>
</tbody>
</table>

Time Series Analysis

- **Homeless**: 2%
- **Parking**: 2%
- **Street**: 2%
- **Subway**: 2%
Continuous Intelligent Pandemic Monitoring (CIPM)

Huijue Kelly Duan
Hanxin Hu
Miklos Vasarhelyi

Accounting Information System
Rutgers, the State University of New Jersey
Research Objective

• Use measurement science (accounting), assurance science (auditing) to enhance pandemic responses

• This study aims to establish a Continuous Intelligent Pandemic Monitoring system (CIPM)
  – Validate the epidemic related numbers
  – Provide guidance to policymakers so that sufficient resources can be allocated to the upcoming high risky areas
Data Collection

- Collect relevant exogenous and endogenous data sources
  - Total confirmed cases, Total deaths, Daily confirmed cases, Daily deaths, Total test, number of positive tests, etc.
  - Demographic density, Industrial jobs, Industrial establishments, % of urban population, Territorial extension of the municipality, list of municipalities by region, Volume of passengers carried, Transported cargo flow, GDP, etc.
  - Apple mobility, Google trends, Official announcements, Tweets, Unemployment claims filed, etc.

Model Construction

- Establish a systematic and continuous COVID-19 monitoring model
  - Use time series and machine learning algorithms to perform predictive analytics
  - Apply clustering approach to perform cross sectional analysis
  - Simulate the Epidemic models

Alert

- Incorporate audit risk assessment to establish an alert system
  - The number reasonableness
  - Disease severity
  - Regional severity
  - The policy sufficiency

Action Recommendations

- Present guidance to policymakers based on the simulation results
  - Number validation
  - Peer groups evaluation
  - Policy simulations

Figure 1. Continuous Intelligent Pandemic Monitoring Framework
Figure 2-1: Predicted positive test ratio vs Actual positive ratio

Figure 2-2: Predicted confirmed cases vs Actual confirmed cases

Figure 2-3: Predicted death cases vs Actual death cases

Figure 2-4: Predicted hospitalizations vs Actual hospitalizations
Figure 3: Use SEIQHRF model to simulate the impacts of different social interventions policies assuming the total number of population is equal to 10000. When enforcing self-isolation and social distancing, we can better control the transmission of COVID-19.
A Machine learning approach To measuring audit quality With surprise scores:
Evidence from China

Authors:
Hanxin Hu (Rutgers),
Ting Sun (TCNJ),
Miklos A. Vasarhelyi (Rutgers),
Min Zhang(Renmin University, China)
For each dependent variable:
- Net income adj
- Total assets adj
- Total liability adj
- Stockholders' equity adj
- Income before income tax adj
- Income tax adj
- nonclean opinion
- Restatement (misstatement)

**Independent variables (example: using net income adjustment as the dependent variable):**
Companies' characteristics (27 variables)
Audit firms' characteristics (28 variables), e.g., revenue, subsidiaries, net assets
Audit partners' characteristics (15 variables), e.g., education, age, gender, birthplace, title

**Data sources:** Chinese Ministry of Finance, CICPA (Chinese Institute of CPAs), CSMAR (China Stock Market & Accounting Research Database)

**Research period:** 2010-2017
**Data size:** 11574
**Data splitting:** training (6626 observations) / test (1325 observations) / application (3546 observations)

---

**Machine learning algorithms**

- Random Forest
- SVM
- Gradient boosting
- XGBoosting
- Deep neural networks
- Logistic regression (the benchmark algorithm)
- RUSboosting
- Balanced Random Forest
Prediction Results

Figure 1: ROC AUC for varying algorithms when “nonclean opinion” is the target variable

Figure 2: ROC AUC for varying algorithms when “net income adjustments” is the target variable
## Application: Results

1. **Nonclean audit opinion; Nonclean opinion surprise score; The aggressiveness score**

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Estimated coefficient</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonclean audit opinion</td>
<td>-27.78</td>
<td>0.960933</td>
</tr>
<tr>
<td>Nonclean opinion surprise score (ML)</td>
<td>3.467</td>
<td>7.84e-13 ***</td>
</tr>
<tr>
<td>The aggressiveness score (Logistic)</td>
<td>0.3958</td>
<td>0.008126 **</td>
</tr>
</tbody>
</table>

2. **Net income adjustment; Net income adjustment surprise score; The aggressiveness score**

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Estimated coefficient</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Net income adjustment</td>
<td>-0.01700</td>
<td>0.853533</td>
</tr>
<tr>
<td>Net income adjustment surprise score (ML)</td>
<td>0.01105</td>
<td>0.911363</td>
</tr>
<tr>
<td>The aggressiveness score (Logistic)</td>
<td>0.02298</td>
<td>0.884880</td>
</tr>
</tbody>
</table>

Note: the regression results for other audit adjustment variables are similar to those for net income adjustment.
Multidimensional Clustering for audit fault detection

Sutapat Thiprungsri
Miklos A. Vasarhelyi
Metlife:

- Data stream of over 200K wire transfers
- Data only currently available for the wires and the records possess little information
- Little context knowledge of the major feeding streams
- No fraud training data available
- Worked during the audit supplementing the audit team work
- Developed a series of data filters relating to specific conditions and trends
- Working on an aggregate weighting model
- Need in the field verification of picked data
Visualizing combination of attributes, we will be able to see similarity and differences among claims.
Analyzing individual variables, we will be able to see clearly that some claims have rare values.
Conclusions

• A wide range of analytic methodologies exists to treat any data rich problem
• Change is very slow rule to anachronistic regulations and embedded interests, as well as lack of training within organizations
• CarLab develops an approach for each project it does
  – If you know what you are doing you are not doing research (Albert Einstein)
• The World Bank has the scope and nature to be an ideal location for experimental analytics using, big data, exogenous variables, machine learning, and a set of out-of-the-box sensing and measurement methods