**Fighting Auto and Property Insurance Fraud through Big Data and Machine Learning**

Capstone Seminar

for

Master of Applied Science

Information and Communication Technology

Frederick McGovern

University of Denver University College

August 5, 2015

Faculty: Timothy E. Leddy, MBA

Director: Thomas Tierney, PhD

Dean: Michael J. McGuire, MLS

**Abstract**

Individuals depend on auto and property insurance to help minimize financial loss in the occurrence of a catastrophic event. For insurance companies to remain competitive, it is critical that they establish an adequate rate compared to the overall potential for loss. However, insurance fraud continues to be an important factor in higher rates. According to the FBI, the total cost of auto and home insurance fraud is estimated to exceed more than 40 billion dollars per year. Insurance companies pass these costs onto consumers such that the typical family sees a rate increase of up to 700 hundred dollars per year over rates based solely on risk characteristics (FBI 2010). By increasing the amount of data collected and by analyzing how the financial, telecommunication, and healthcare industries utilize machine learning algorithms to combat other types of fraud, insurance companies should be able to decrease significantly the costs of insurance fraud.

Table of Contents

[Background 1](#_Toc426582089)

[Reasons Why Insurance Fraud is Committed 2](#_Toc426582090)

[Types of Property and Causality Insurance Fraud 3](#_Toc426582091)

[Approach 5](#_Toc426582092)

[Literature Review 7](#_Toc426582093)

[History of data analysis 7](#_Toc426582094)

[Machine Learning Overview 10](#_Toc426582095)

[Unsupervised Learning 11](#_Toc426582096)

[Supervised Learning 12](#_Toc426582097)

[Machine Learning Examples 13](#_Toc426582098)

[Healthcare Industry 13](#_Toc426582099)

[Telecommunication Industry 17](#_Toc426582100)

[Financial Industry 17](#_Toc426582101)

[Solution 19](#_Toc426582102)

[Discussion 21](#_Toc426582103)

[Types of data available for analysis and potential usage 21](#_Toc426582104)

[Internal Data 21](#_Toc426582105)

[External 22](#_Toc426582106)

[Best method of data analysis 25](#_Toc426582107)

[Concerns with the utilization of external data and machine learning algorithms 28](#_Toc426582108)

[Recommendations 31](#_Toc426582109)

[Conclusion 32](#_Toc426582110)

[References 34](#_Toc426582111)

# Background

The purpose of insurance is to help mitigate the financial loss on the part of an individual or corporation should a catastrophic event occur. Lenders require property insurance to minimize their exposure to loss. Most states require auto insurance, also known as casualty insurance, as a prerequisite for driving so as to reduce the cost associated with property damage and personal injury if an automobile accident were to occur. It is critical that an insurance company establishes an adequate rate compared to the overall potential for loss to ensure continued solvency.

Unfortunately, insurance companies do not establish rates solely based on an insured’s risk characteristics. Every day, the insurance industry faces an increased exposure to loss due to insurance fraud. Insurance companies then pass along this increase in risk to the consumer in the form of higher rates. The exact cost of insurance fraud is unknown due to the high number of undiscovered fraudulent claims, however, according to the Insurance Fraud Division of the FBI, the total cost of auto and property insurance fraud is estimated to exceed more than 40 billion dollars per year. As a result, a typical family sees a rate increase between 400 hundred and 700 hundred dollars per year over rates based solely on risk characteristics (FBI 2010). However, there are additional costs as well. For example, companies also pay higher premiums due to fraud and these companies in turn pass those costs onto the consumer in the form of higher prices. In fact according to the Texas Department of Insurance, when these additional costs are added to the rate increase due to fraud, the typical family in reality pays up to 1,000 dollars per year due to auto and property insurance fraud (Texas Department of Insurance 2015).

## Reasons Why Insurance Fraud is Committed

People engage in property and casualty insurance fraud for a variety of reasons. The potential for monetary gain drives the vast majority of individuals to commit insurance fraud. Others are reluctant to commit fraud but due to the requirements of many states and lenders to carry insurance; they engage in misrepresentation so as to reduce the overall premium to something that the individual can afford. Still others commit fraud due to the high cost of auto repairs hoping to get repairs done that they otherwise could not afford (Coalition Against Insurance Fraud 2007).

Interestingly, a new reason has recently been gaining traction, that being that people who commit fraud do so in retaliation due to either a poor experience with the insurance company or a negative perception of the insurance industry in general. Many individuals know that insurers pass along the cost of fraud to their insureds in the form of higher premiums and so these individuals engage in fraud in order to “balance the equation” (Lesch and Byars 2013). Research conducted by the Coalition Against Insurance Fraud seems to support this believe by finding that 40% of Americans believe that fraud is a direct result of insurance industry unfair practices (Coalition Against Insurance Fraud 2015a). The Insurance Research Council (IRC), in a 2013 online poll, found that 24% of respondents felt it was acceptable to state a higher damage amount, known as claim padding, following an accident in order to cover the deductible they had to pay. Another IRC report found that 23% of males aged 18-34 think that claim padding is also acceptable to recoup prior paid premiums and that 10% of Americans think that “insurance fraud doesn’t hurt anyone” (Insurance Information Institute 2015). A 2010 survey by Accenture found that 68% of people believe that fraud primarily occurs because people think they can get away with it (Coalition Against Insurance Fraud 2015b). These findings seem to suggest that many people believe that insurance fraud is a low-risk crime, and therefore it empowers more individuals to engage in fraudulent activity.

## Types of Property and Causality Insurance Fraud

The list of various scams that those engaged in insurance fraud commit is lengthy with new schemes appearing each week. While there exist many methods to commit fraud, each method can be classified into one of two categories, those being soft and hard fraud. The insurance industry considers soft fraud to be more opportunistic in nature and is the most prevalent type of insurance fraud. This kind of fraud is either committed by the insured individual or by insurance agents. Examples of soft fraud include:

* Misrepresentation on the application to get a lower premium such as listing another address than where the vehicle is regularly garaged or leaving a driver off of the application.
* Padding of a legitimate damage claim to obtain additional reimbursement.
* Claiming injuries are more extensive than they are.
* Individuals decide to exaggerate the value of stolen items.
* Agents adding extra coverage not requested by the insured to increase the agent’s commissions.
* Illegally obtaining insurance for an individual who does not possess a driver’s license.

Soft fraud is considered to be a “victimless fraud” with fraudsters believing that insurance companies can easily afford to absorb the costs of the fraud due to the vast amounts of insurance premium being collected (Levin 2013).

Hard fraud is the deliberate act of destroying or the engaging of an individual in action to create an accident to submit false claims. In this case, individuals plan the fraudulent act out ahead of time. Frequently they enlist the assistance of other parties, which can include other drivers, insurance agents, medical providers, tow truck operators, auto repairmen and lawyers (Levin 2013). Organized crime rings that engage in hard fraud are common in large metropolitan cities across the United States. Examples of hard fraud include:

* Arson
* Phony injury claims on someone else’s property i.e. falls
* Disposal of a vehicle by insured and then filing a fraudulent theft claim
* Replacing deployed airbags with poor quality airbags or trash
* Premium diversion by the insurance agent where the agent keeps premiums instead of forwarding it to the insurance company.
* Staged Accidents

Staged Accidents are an especially common form of hard fraud that is increasing in frequency. The National Insurance Crime Bureau reported an increase of 46% from 2007 to 2009 in suspected staged accidents (Scafidi 2010). From 2008 to 2011, that percentage increased to 102% (Tidball 2015). There are many different variations of this scheme with a common example being the drive down. While trying to take a left-hand turn, the defrauder waves on the insured to proceed. Once the insured starts to pull out, the driver of the other car speeds up and crashes into the insured. When questioned by police, the fraudster denies every waving out the insured putting the blame for the accident solely on the victim (Estrin 2013).

Insurance carriers need to do a better job in identifying insurance fraud. Analysts need to create new processes that are faster and cheaper than the current methods used to detect fraud. These processes will require the utilization of a wide variety of internally as well as externally available data, much of which may not directly correlate to insurance. One possible solution is to look at the latest trends in machine learning. Machine learning is a relatively new area of computer science that focuses on creating algorithms that allow the computer to make predictions based on the data that it is given. By analyzing the latest machine learning algorithms and how data analysts apply these algorithms in other industries to combat fraud, insurance carriers may be able to apply those same techniques to insurance related and disparate data sources, such as social media, to significantly decrease the amount of insurance fraud.

# Approach

Research for this project began with an overview of the internal data that is typically available to insurance carriers. As this information is proprietary in nature, the author conducted a review of the applications utilized by auto and property insurers to determine data points that appear to be consistent from one insurance carrier to another. Many of the insurance carriers reviewed use the industry standard auto and homeowner insurance applications developed by the Association of Cooperative Operations Research and Development (ACORD).

Insurance companies also consider some of the external data that they utilize to be proprietary in nature. Information provided via email from a supervisor at a Fortune 500 property and casualty insurance company, however, allowed for the creation of an initial list of external data vendors. Various national and state organizations have been set up to combat fraud, and, therefore, a review of these organizations websites also provided a list of potential data suppliers. Further research identified other potential data points, such as public voter records to cross check against the submitted insurance application’s stated residential address.

The review of websites belonging to private organizations such as the National Insurance Crime Bureau and the Coalition Against Insurance Fraud as well as governmental agencies such as the FBI and individual state insurance fraud bureaus provided valuable data. These sites allowed for the both the identification of the scope of insurance fraud and to understand the different types of insurance fraud.

The author performed a substantial literature review to determine how insurance companies are currently conducting insurance fraud detection. The exact methodologies that each insurance carrier utilizes are confidential, as the publication of such data would provide those committing fraud to avoid easy detection. However, there was plenty of available research on the utilization of general principles such as generalized linear models, Bayesian networks, and decision trees. This study also indicated that insurance carriers and researchers interest in machine learning algorithms to identify fraud, while low, is starting to increase. Research into the data techniques used in the past to identify fraud was considered unnecessary in part due to the authors prior experience in the insurance industry.

In order to show that fraud is not unique to the insurance industry and that the potential solution to insurance fraud may exist through an analysis of other industries effort in fraud detection, a comprehensive literature review was carried out. This review allowed for the compilation of data on the cost of fraud in these industries. An attempt to determine the data analysis techniques these industries utilize followed the determination of fraud costs. Details of the actual algorithms specific companies employ were not available as expected, though some companies did indicate the employment of general machine learning techniques such as Support Vector Machines, Artificial Neural Networks, and Deep Learning to identify fraud. Researchers developed most of the advanced machine learning techniques to combat financial, telecommunication and healthcare fraud that the author discusses in this paper. Many of these techniques resulted in significant improvement in fraud detection utilizing sample data.

# Literature Review

In order to make a determination on whether or not the incorporation of vast quantities of data, in conjunction with an established machine learning algorithm developed by another industry, can reduce the overall rate of insurance fraud, the following three questions should be addressed.

1. How have data scientists conducted data analysis in the past?
2. What is Machine Learning?
3. What other industries have utilized Machine Learning to combat fraud?

## History of data analysis

Actuaries have utilized a type of linear regression analysis, known as the General Linear Model (GLM), for decades mainly as predictors of potential risk. Insurance companies also use the GLM for the development of insurance rates and fraud detection. Figure 1 on the following page displays a graphical representation of a linear regression analysis. The GLM has many limitations, an example being that missing values in the data detrimentally affected the model. Another limitation is that the model does not perform well if there exist data points outside the normal exponential distribution path (Kolyshkina, Wong, and Lim 2005).

The GLM is also not well suited to handle data sets that have a large number of categories (Kolyshkina, Wong, and Lim 2005). Developers can gain some flexibility through the use of additional tools such as the Multivariate Adaptive Regress Splines (MARS) to build regression models that analysts then provided as input into a GLM. This hybrid approach does have the benefit of having slightly better accuracy and less development time when compared to other approaches such as binary recursive partitioning (Kolyshkina and Brookes 2002).

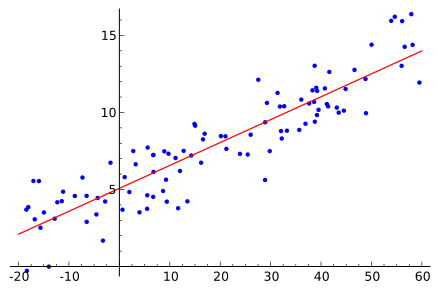


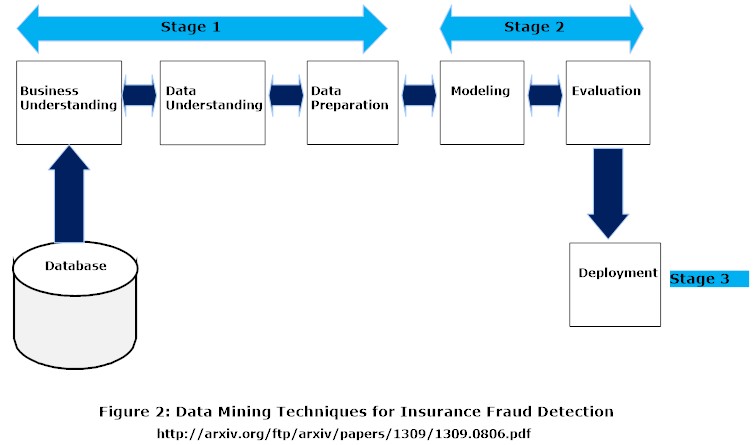
Figure 1: Example of Linear Regression Analysis

https://upload.wikimedia.org/wikipedia/commons/thumb/3/3a/Linear\_regression.svg/438px-Linear\_regression.svg.png

The disadvantages of the GLM model, including its inability to handle enormous data sets, has led researchers to augment or replace the GLM with data mining methods (Kolyshkina, Wong and Lim 2005). Data mining is a process by which analysts examine a large amount of data to identify patterns that exist in the data. Analysts compare these patterns against new subsets of data to determine if the pattern remains. If it does, then analysts can use this model to make predictions about future data. Data mining consists of three stages (Data Mining Techniques 2015).

1. Exploration of the data
2. Building of the model
3. Deploy the model so as to begin generating predictions

As can be seen in figure 2 below, the first stage of data mining consists of three steps. The business and the data analysts both need to understand what they are looking for and come to an agreement as to what the data represents. The data preparation step would include removing any bad or inconsistent data and integrating it with other data sources. The application transforms data as needed so as to be in a form that allows for data mining. The second stage consists of building the model and includes performing an evaluation of the completed model. The completion of the third stage, deployment, occurs after this validation is finished (Sithic and Balasubramanian 2013).



The last fifteen years has seen the development of many data mining techniques. Two of the most common utilized methods in insurance are the Bayesian network and the decision tree (Bhowmik 2008). Both of these methods attempt to identify class membership. A Bayesian network is a graphical representation of variables and the relationships that exist between these variable. Utilizing node probability tables, an analyst can determine the likelihood of an event occurring if another event occurs. A decision tree expresses a relationship between dependent and independent variables based on a set of rules as a set of IF-THEN statements. Figure 3, below, shows an example of a simple decision tree.

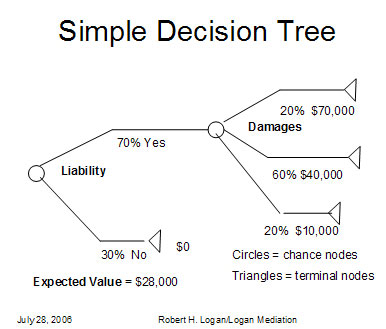


Figure 3: Simple Decision Tree

https://lh3.googleusercontent.com/-xlMgwfuJC9GYiog8rr0lvo1mUjbq2va7LV0qV9bQhvSQ7Kvez8-dlMulw6X\_l2HWD8ISQ=s100

Bayesian networks tend to have better accuracy then decision trees. However, the introduction of redundant data results in a decrease in the Bayesian network accuracy. Decision trees have noticeable performance decreases as the data set is increased (Bhowmik 2008).

## Machine Learning Overview

There exists a plethora of private and public data which the industry has since defined as Big Data. Combining all of this Big Data would overwhelm the capabilities of traditional databases and data processing applications. Earlier data mining algorithms such as decisions trees, suffer significant performance degradation as data set size is increased (Bhowmik 2008). To handle these large data sets, data scientists have turned to new methods of performing data analysis.

Programmers typically develop algorithms to solve a particular problem. However, with many complex issues the solution to the problem is not as obvious due to the overwhelming amount of data needing to be analyzed. Machine learning methods pass the responsibility of developing the algorithm or identification of a pattern over to the application. This self-learning has allowed us to find solutions to problems in the area of vision, speech recognition and robotics (Alpaydin 2014, 1-4).

Other utilizations of machine learning include basket analysis, which is the process of making associations in the data. For example, according to researchers there is a strong relationship between beer sales and diaper sales, especially on Thursdays and Fridays (Marakas 2003, 84-94). The ability to determine associations between products has allowed the marketing industry to improve sales through targeted promotions, bundling products, and the modifying of store layouts.

To find the patterns and the relationships in the data and to create a model, developers need to provide the application with a data set. Following the creation of the initial model, the model is continuously refined through the introduction of new data. There are two methods, known as unsupervised and supervised learning, with which the application can learn to create new models.

### Unsupervised Learning

An unsupervised learning method is one in which the application receives no feedback from the application designer in regards to whether the classification or decision made was a correct one. Instead the application clusters data together, making an initial guess based on the inputs and the importance (weight) of that data point as initially determined by the developer. The application utilizes this information in the creation of various training sets. The application then compares the produced output to the input to see if the patterns of both data sets approximate each other (Marakas 2003, 135-136).

The application then calculates a percent error between the two data sets and automatically adjusted the weight setting assigned to each data point. The process continues to repeat itself until reaching a particular error threshold value. An advantage of the unsupervised learning approach is that there are no preconceived patterns already defined. It is, therefore, an excellent tool for finding hidden or new relationships in the data that were previously unknown (Rawte and Anuradha 2015).

### Supervised Learning

Supervised learning utilizes training sets provided by the application developer to determine if the output produced by the application is a correct one. Backpropagation is a standard technique used to compare the output to the input. Similar to unsupervised learning, the application can then modify any of the data weights and rerun with the goal of having a higher percentages of matches (Marakas 2003, 136).

An example of this might be to determine the classification of a car. The input provided includes the vehicle make, model, engine displacement, and the number of doors. The application attempts to classify the data into categories, for instance, family cars or sport utility vehicles. The application compares the training set to the output provided to determine which classifications were correct. The application uses the compared results to adjust the input weights and then reruns. The process continues until a certain number of cycles have run, or the process reaches an acceptable error rate (Alpaydin 2014, 21-27).

## Machine Learning Examples

Several industries currently utilize many different types of machine learning applications to identify fraud. Since many of the details of each method are overly complex, as well as the fact that complex mathematical theory is the basis for each method, the following only provides an overview of these techniques including any metrics that the reader can utilize to make a comparison.

### Healthcare Industry

The National Healthcare Anti-Fraud Association estimates that losses due to health care fraud are somewhere between 3% - 10% of the total US annual healthcare costs. With spending expecting to increase over 2 trillion dollars over 2010 levels, advances in machine learning detection of healthcare fraud are being researched in the hopes of reducing these costs (Johnson and Nagarur 2014).

In relation to health care, both the supervised and unsupervised learning methods have particular advantages and disadvantages (Rawte and Anuradha 2015). For example, data analysts can use supervised learning to establish pattern classification. If a submitted claim maps to a previously determined valid classification pattern, the algorithm will deem the claim to be legitimate. However, the process of establishing each class is very time-consuming. Additionally, the developer cannot ascertain whether or not every claim in the training set is, in fact, valid and as a result, this can lead to a skewed output. Unsupervised learning, on the other hand, focuses on behavior and thus analysts can utilize this technique to identify not only existing patterns of fraud but new ones not yet identified. A disadvantage is that unsupervised learning method does not work that well within the parameters of a specific problem. Another issue is that the algorithm deems many unknown patterns as fraudulent when in fact they are legitimate (Rawte and Anuradha 2015).

Since each method has its strengths and weaknesses, one model that researchers have proposed is a hybrid model. In this model, an unsupervised approach is used to first cluster data. This application then utilizes the clustered data as input to a supervised learning application. Researchers, utilizing this hybrid model, successfully developed an application that grouped inputted claims into categories such as disease type using what the researchers call an Evolving Cluster Method. Once the algorithm clustered the claims, they were processed into a supervised learning Support Vector Machine (SVM) to identify any duplicate claims. The analyst provides a training set for the SVM application to improve the classification of a claim as either legitimate or fraudulent recursively. According to the developers, this method should be very helpful in identification of health insurance fraud (Rawte and Anuradha 2015).

Researchers from the University of New York at Binghamton did a significant investigation into the type of data mining techniques that healthcare industry analysts were researching to combat health care fraud. This investigation led to the identification of gaps in the overall research and led the researchers to develop a hybrid approach to healthcare fraud detection. Researchers divided this methodology into the following six stages that cover both an analysis of provider fraud and claim fraud.

1. Provider Profiling
2. Demographics Screening
3. Claim Amount Screening
4. Risk Quantification
5. Risk Threshold Determination
6. Comparison of Risk with Threshold

In the first stage, the researchers attempted to determine provider encounter metrics by creating a model that tracked such things as how many times and for how long a doctor visited with a patient. In the second stage, the researchers created a model based on patient demographic data. Since there is a relationship between an individual’s gender and age to certain types of diseases, this model helped researchers determine if the provider diagnosis and the subsequent service provided were, in fact, appropriate for the individual. The third stage had researchers creating a model to map the expected claim costs against actual claim costs. Similar patient diagnosis and services should have similar costs so any claim cost that is outside of the expected range may indicate fraudulent behavior by the provider.

In stage four, the researchers developed a mathematical formula to quantify the risks associated with the combined input data from the previous models. The outputs of the models created in the first three stages establish the parameters used in this calculation. For example, the researchers utilize the provider profiling stage output to calculate the distance variable ∆p and the density variable δp used in the fraud risk quantification formula. In stage five, they established a unique threshold value for each claim through the utilization of a decision tree. There were four possible outcomes from the decision tree those being true positive and false negative if the algorithm deemed the claim to be fraudulent, or false positive and true negative if the algorithm considers the claim to be legitimate. The final stage compares the individual risk values calculated in stage four to the overall calculated risk threshold from stage five. If the risk values are below the calculated risk threshold, then the algorithm considers the claim to be valid. Researchers conducted an application of this algorithm on a test data set, provided by a healthcare insurer, consisting of 878,691 claims of which the insurance company had found 27% of the claims to be fraudulent. The resultant was an 86% accuracy rate that exceeds the accuracy rate for traditional neural network approaches to fraud identification (Johnson and Nagarur 2014).

Another technique that has shown some promise is the utilization of a method known as outlier detection. Outlier detection is typically an unsupervised learning method that attempts to determine if a particular situation deviates from normal behavior. Analysts frequently utilize this technique to identify data input errors, credit card fraud and network intrusion attempts (Alpaydin 2014 199-200).

Researchers in California, Germany, and the Netherlands employed the outlier detection method as part of a multistep iterative process to identify fraud in a sample of 650,000 dental claims. The researchers began by determining the metrics that they would use in their data analysis. They defined the metrics through an analysis of parameters established by the FBI as well as other industry metrics existing in the current literature (Thorton et al. 2014). They started their analysis by cleaning and filtering the data to remove records with missing values. Following the completion of data scrubbing, the researchers grouped providers based on similar characteristics with those below a certain claim count being discarded. Utilizing a variety of outlier detection methods, the researchers identified 14 predictive flags of potential fraud. Of the 360 providers analyzed with these methods, 17 providers were found to have hit at least three flags. Further review by the researchers identified that 12 out of the 17 providers were extremely likely to have engaged in fraudulent activities. The researchers referred this information to government officials for review Thorton et al. 2014).

### Telecommunication Industry

In an attempt to identify telecommunication fraud, researchers utilized a technique known as clustering to analyze user logs. Every phone call made, or web page visited results in the creation of a user log. A combination of these logs results in the creation of a user profile that can identify a person’s habits or preferences. By utilizing a variety of clustering techniques, such as K-means or Agglomerative clustering, researchers analyzed a sample of user profiles to determine if individuals were perpetrating fraud against the provider. Unfortunately, while the researchers found that clustering techniques are promising, the volatility in user behavior resulted in a high degree of fraud misclassification. The researchers of this study also point out that they did not have access to many private data points such as the number that the individual called or the location that the person was when the individual placed the call. The authors of this study suggest that the addition of these private data points may improve the overall detection capabilities through clustering techniques. The authors also believe that other clustering techniques such as subtractive clustering or other approaches such as social network analysis might improve upon the overall results (Hilas, Mastorocostas, and Rekanos, 2015).

### Financial Industry

The GLM is also highly utilized by the financial sector in the detection of corporate fraud. However, there are other techniques being used such as decision trees, Bayesian networks, and neural networks, with neural networks being the most utilized methods of fraud identification following the GLM (Ngai et al. 2011).

Researchers are investigating and using several innovative machine learning methods. One such method is the Artificial Immune System (AIS algorithm). The idea behind this approach is that fraud detection is very similar to how the body’s immune system works. Our immune system is designed to seek out anything that’s not normal. The human body considers a virus to be abnormal, and our immune system would set out to detect it. Corporations likewise found fraud to be an abnormal event, so this led to the creation of the AIS algorithm that simulates the functions performed by an immune system, those being Negative Selection and Clonal Selection. Researchers have made an improvement to the AIS algorithm through the implementation of an Apache Hadoop framework in conjunction with an implementation of MapReduce. MapReduce is software used to process multi-terabyte data sets in parallel to cut down on the time needed to train the algorithm. The algorithm has had several other modifications made to it, such as improving the distance function as well as tweaking the scoring value that determines if a transaction is fraudulent. These combined changes have resulted in an increase in fraud detection of 25% while reducing the overall performance time by 40% when compared to a baseline AIS Algorithm (Soltani Halvaiee and Akbari 2014).

Another novel machine learning technique is a modification of a standard supervised learning method known as Fisher Discriminant Analysis (FDA). A few speech and facial recognition applications have successfully utilized the FDA method (Mahmoudi and Duman 2015). The researchers who developed this technique point out the there is no model that can catch all fraud activity. There are both fraudulent transactions that the model classified as valid (false negative) as well as legitimate transactions classified as fraudulent (false positives). While many studies equate the two as equally detrimental, the authors argue that when determining a model to use from a profit standpoint it is better to classify correctly “beneficial transaction” over other transaction. To minimize false negatives, the researchers attempted to improve model cost sensitivity through an introduction of a weighted average into the FDA. The modified FDA, when compared to other machine learning and data mining techniques such as an unmodified FDA, Artificial Neural Networks, and decision trees, was able to show a higher level of profit than other methods (Mahmoudi and Duman 2015).

Another example of a company’s attempt to identify financial fraud involves the use of machine learning methodology known as Deep Learning. This technique has recently shown excellent results in both visual and speech pattern recognition (Noda et al. 2014). About three years ago, PayPal decided to combat financial fraud by adapting this technique. Hui Wang, PayPal’s senior director of Global Risk Sciences, stated that the typical “if someone does X, then the result is Y” isn’t seen as much in fraud. Wang points out that fraud patterns are growing increasing more complicated. The deep learning algorithm that PayPal developed can analyze tens of thousands of variables and detect variations of existing schemes. When compared to other top performing fraud detection algorithms used by PayPal, the deep learning algorithm improves fraud detection by 10%. The company hopes to someday be able use the model as transactions are processed so as to identify the fraud quickly as well as allowing for the retraining of the algorithm based on extremely current data (Harris 2015).

# Solution

Insurance companies do in fact want to identify fraud and employ hundreds of employees and utilize numerous private investigators in conjunction with each state's insurance bureau. However, the processes are tedious and require significant manual work to identify the fraud and to document it sufficiently for prosecution. Research by the Coalition Against Insurance Fraud, which this author’s employer is a member of, showed that 75% of large insurance companies spent over 800 dollars per fraud investigation (Coalition of Insurance 2014). Additionally, estimates of automated Special Investigative Unit (SIU) referrals through predictive analysis currently stands at only 15% (Coalition Against Insurance Fraud 2014) establishing the fact that the usage of machine learning by the insurance industry is limited.

The solution to reducing the overall insurance fraud rate is a two part process. It first requires that insurance companies embrace the concept of Big Data by dramatically increasing the amount of data that they collect. This solution would include non-insurance related data and textual data such as found on social media. Currently, investigators manually analyze much of this data but only after a claim representative assigned to the claim initiates the request. The automation of this analysis will reduce the manual workload of the SIU resulting in a reduction in both costs and time to prosecution. Additionally, with the inclusion of more automated referrals, insurance carriers will quickly be able to identify more individuals engaged in fraud and stop their activities sooner resulting in additional cost reductions.

The acceptance of Big Data, however, will require a paradigm shift in the way data scientists perform data analysis. Traditionally methods of analysis simply cannot process the vast amounts of data insurance companies will collect. As a result, more and more insurance carriers will need to turn to machine learning algorithms to detect fraud. These algorithms can process more data, and have better prediction accuracy than traditional methods and can utilize unsupervised learning to identify new patterns of fraudulent activities quickly. Fraud is prevalent across many industries including financial, telecommunications and healthcare. Companies in these industries have already conducted research on the usage of machine learning algorithms to detect the various forms of fraud that exist in these industries. This research has shown that machine learning algorithms can dramatically improve the capability of businesses to identify fraud.

As a result, insurance carriers need to adopt the methodologies used so successfully by these other industries in combatting fraud. Through an analysis of what worked and did not work for these other industries, insurance carriers can reduce the normal learning curve that accompanies the usage of new technologies and methodologies.

The resultant solution will, therefore, be cost-effective, be faster than existing data analysis in identifying fraud and will identify a higher percentage of fraudulent activity. This reduction in fraudulent behavior will result in a decrease in overall premiums paid by the average family as well as a decline in the number of people victimized by those engaged in hard fraud.

# Discussion

For the reader to better understand the need to incorporate additional data into fraud detection, the author of this paper conducted an analysis of the availability of both internal and external data. Additionally, the author identified the usage of this data in the area of fraud detection based on either actual fraud analysis conducted by the author for his employer or through extrapolation based on his prior experience.

## Types of data available for analysis and potential usage

### Internal Data

Internal data is data that an insurance carrier collects as part of the normal activities engaged in by an insurance carrier. For example, insurers collect and store all data that an individual submits on an insurance application such as address, vehicles, coverages selected, drivers, and property details. Insurance carriers would classify additional information such as billing history, method of payment, underwriting notes, and submitted claims would also as internal data. Insurance companies would consider most of this information to be proprietary data.

Data analysis on internal data is limited. The proprietary nature of the data limits the use of third party applications for data analysis. Typically, the frequency of claims is a good indicator of potential fraud. Residential and mailing addresses are also an excellent data source for fraud analysis. For example, the author was able to identify an individual who was providing insurance for over 20 individuals that didn’t have insurance by aggregating counts based on residential address. Unfortunately, internal data alone will identify a very small percentage of fraud due to the low number of relationships that exist in the limited data.

### External

External data is data that a vendor service supplies to the insurance company. There are many organizations that provide potential sources of external data, including national and state level organizations, to combat insurance fraud. The following list is a small sample of some of the external data that is available to all insurance companies.

Insurance Credit Scores (ICS) - The insurance industry has long ago identified ICS as being predictive of the propensity for an insured to file an insurance claim. The algorithm used to calculate the ICS score utilizes about thirty of the elements that make up a traditional credit score (Hennelly 2015). The higher an individual’s credit score is, the less likely they are to file an insurance claim. Because of this the majority of the automobile and homeowners insurers use it in their rating algorithms when allowed by state law (NAIC 2015). A lower credit score may indicate that an insured is facing financial hardship, and thus under more pressure to commit fraud.

ISO Claim Search – Contains data that the NICB formerly administered as well as the Property Insurance Loss Register. Information contained in these databases includes vehicle thefts and theft recoveries, salvage titles, auto physical damage claims. The SIU utilizes non-insurance related tools in their investigations such as being able to identify individuals who are delinquent in child support or have a prior criminal background (Verisk 2011).

Lexis Nexis – Insurance companies order Comprehensive Loss Underwriting Exchange (C.L.U.E) reports as part of the underwriting process as it provides an extensive history of both auto and property claims made by an insured. C.L.U.E reports provide property claim data sorted not only by the individual who is making the claim but also by the location of the claim. The higher the number of previously filed claims then the higher the potential exists that an insured will make a future claim (Kissell 2012). Higher claims can also indicate the existence of a fraudulent claim.

Melissa Data – As address misrepresentation is a very common type of soft fraud, any external data that can potentially validate an address provided by an insured on an application can be a tremendous tool in fraud prevention. These tools allow an analyst to enter an address, and it will return a potentially corrected address and an indication of the status of that address. Status indicators can include whether the address is valid, has an incorrect town or zip code assignment, is missing information such as a suite number, is a mail drop box, or is vacant. It also has the capability of identifying whether or not an individual stated on the application lives at the address stated (Melissa Data 2015).

Merit Rating Board – From this data, an insurance company can identify individuals with a suspended or expired license. The presence of driver history also provides a key fraud indicator in the identification of prior accidents (Massachusetts Registry of Motor Vehicles 2014).

Social Media - This is one of the SIU’s best tools. Investigators can visit sites like Facebook and Pinterest looking for visual evidence that someone is faking an injury. Insurance carriers can conduct text mining from sites like Twitter to identify fraudulent activity that is taking place. As more tools enter the market, that provide automated searches of this data, the manual process of identification of fraud will be significantly reduced (Coalition Against Insurance Fraud 2014).

Vehicle Inspection Services – For fraud identification the most important data point provided by this service is the odometer reading. (Automobile Insurance Board 2015). Many insurance companies offer an annual mileage discount. This data allows for the comparison of the mileage stated by the insured to the exact vehicle mileage to determine if a provided discount is legitimate.

Voter Registration - The New York Times reports that last year an auto insurance company filed a patent for a system that analyzed voter registration as a predictor of risk. Currently, this usage would not be legal in the determination of an insured’s rate, as state regulators have not approved its use (Tugend 2014). However, this data would include the insured registered voting address and as such insurance companies can utilize this information as part of an address verification process.

The addition of external data dramatically improves the identification of fraud. For example, the author recently conducted an analysis of 160,000 addresses to determine the percentage of policies with potential application misrepresentation due to invalid addresses. Through a comparison with data provided by Melissa Data, it was determined that approximately 2.5% of the submitted addresses were invalid. The author provided this data to each affected insurance agent for review with the policyholder so that the agent could correct these addresses.

The Insurance Information Institute states:

“Rate evasion, where policyholders misrepresent facts on applications, includes the use of a false Social Security number to avoid showing a bad credit score, misrepresenting the major use of a vehicle and giving a false address where rates are cheaper. Industry observers estimate that this type of fraud costs auto insurers about $16 billion a year.”

Insurance carriers can dramatically reduce rate evasion, a type of soft fraud, at the point of sale by utilizing external data to validate the data that the insured provides on the insurance application. As previously stated, 68% of people believe that fraud primarily occurs because people think they can get away with it (Coalition Against Insurance Fraud 2015b). By using external data to identify soft fraud quickly, individuals will no longer perceive insurance fraud as a low-risk crime, and therefore it will reduce the number of individuals engaged in both soft and hard insurance fraud.

## Best method of data analysis

The GLM is an excellent analytical tool when presented with data points that are linearly separable; that is the classification of the data points produces results that can be separated by a line. The GLM can be used to determine the relationship between a dependent variable and one or more independent variables (Ngai et al. 2011). As previously discussed, however, the General Linear Model is ill suited to perform data analysis on Big Data sets. The main reason for this is the fact that as the amount of data being analyzed increases, many of the relationships between the data that does exist will not be identified due to the fact that the relationship is nonlinear. Additionally, as data increases the likelihood of incomplete data also increases which dramatically affects the overall model performance (Kolyshkina, Wong and Lim 2005). At best, this approach may give companies a baseline number to compare against additional methodologies.

Another potential method is to utilize Bayesian networks. The Bayesian network is a type of graphical model in which analysts can identify relationships between data points, also known as nodes, by the presence of an arc. For example, the graphical model in figure 4 below shows that rain is a cause of making the grass wet (Alpaydin 2014, pg. 388)



Figure 4 - Example of Bayesian Graphical Network

An advantage of the Bayesian network is that unlike the GLM, missing data points do not dramatically affect the Bayesian network performance or accuracy (Bhowmik 2008). When used in conjunction with Big Data sets, however, Bayesian networks begin to have issues. For example, the increase in external data gathered will ultimately result in some data being redundant in nature. The Bayesian network approach does show better accuracy than decision trees. However, overall accuracy decreases as redundant data in introduced into the data set (Bhowmik 2008).

Another issue with Bayesian networks is the restriction that every node in the graph can have a single parent node and a single child node. Decision trees resolve this issue by allowing each node to have multiple child nodes (Alpaydin 2014, pg. 402). As a result, several insurance fraud detection systems do employ this technique (Ngai et al. 2011). The problem with this approach is that performance is inversely proportional to the amount of data being analyzed (Bhowmik 2008). The new solution that insurance company will develop needs to be fast, in hopes of identifying fraud quickly. The ultimate goal should be to detect fraud at the point in time that the insured submits the application. With the massive increase in the size of data needing to be processed the Bayesian network approach is just too time consuming.

Additionally, as pointed out by Hui Wang, PayPal’s senior director of global risk sciences, the data analysts are not seeing the scenario of X resulting in Y as much anymore (Harris 2015). As decision trees are rule-based and depend on X causing Y, it would seem that decision trees are no longer the best methodologies for fraud detection.

These would appear to indicate that machine learning techniques such as neural networks, support vector machines, and deep learning would be the strongest candidates for use in insurance fraud detection. Neural networks are currently the second most utilized method of financial fraud detection following the GLM (Ngai et al. 2011). Neural networks have several advantages over traditional GLM approaches such as a decrease need to have formalized statistical training, the ability to identify non-linear relationships and the fact that these models can identify all possible relationships between the variables (Tu 1996). A disadvantage is that neural networks are prone to overfitting. Overfitting is the identification of noise due to data errors into the relationship identified thus producing an incorrect assumption (Alpaydin 2014, 39). Support Vector Machines, like neural networks, can also identify non-linear relationships. Additionally, Support Vector Machines have the advantages of being less prone to overfitting (Sahin and Duman 2011).

There is no correct model to utilize in every instance. Each model has its strengths and weaknesses with is why, as previously indicated, some researchers have developed hybrid approaches. The question that insurance carriers need to ask isn’t whether or not machine learning algorithms can identify fraud but whether the insurance carriers can incorporate the work done by other industries in fraud detection for use in insurance fraud detection. Evidence would seem to show that it certainly can. For example, researchers in Australia documented the application of unsupervised and supervised fraud detection techniques to efforts to detect terrorist activity, money laundering, insider trading and in spam detection (Phua et al. 2010).

## Concerns with the utilization of external data and machine learning algorithms

Misuse of data

The main argument with regards to the collection of public data seem to revolve around concerns that potential data points fall disproportionately, affecting low-income groups and minorities far more than affluent groups (NAIC 2015). The often quoted example is that of credit score. In regards to the utilization of credit score to determine the insured’s premium, the author does agree with their concerns. However, in the context of identifying fraud, the usage of this data point is only affecting any individual engaged in fraudulent behavior. Additionally, a singular data point is not a good predictor of fraud. Only through the analysis of hundreds to upwards of thousands of variables can data analysts develop patterns that provide a strong indicator of fraud.

A stronger case can be made that insurance companies can potential misuse this data to engage in a process called price optimization. Price optimization as defined by the Consumer Federation of America is “an underwriting practice of using information unrelated to risk to set rates for a particular insured – information like supply and demand, or competition” (Jergler 2014). In this case, the central data point causing the issue is the length of time with the current insurance carrier. Insurance carriers have seemingly developed a methodology that determines an individual’s propensity to stay with an existing carrier and thus tweak individual rates slightly higher for those people less likely to seek out comparable rates.

While no concrete proof seems to exist that any insurance carrier has implemented the practice of price optimization, the NAIC is currently investigating and will issue a report later this year. Some states such as Massachusetts and Maryland have issued official statements to insurance carriers deeming this practice illegal if it, in fact, is being utilized (Jergler 2014).

The bottom line is that unless a specific data point is determined to be an indicator of risk and subsequently approved for use by state regulators insurance companies cannot use it in calculating the premium. No data that is gathered via social data mining by an insurance company could be used in rating without a lengthy regulatory approval, and numerous consumer groups would vehemently oppose its use. The collection of the vast majority of additional public data by insurance companies would be for the identification of insurance fraud and hence should be of limited concern to consumers.

Cost of Infrastructure and Lack of Available Expertise

The biggest concern for insurance companies is the overall cost to implement the infrastructure needed to store the necessary data that the machine learning algorithms need for analysis. Databases, such as Hadoop that can hold hundreds of terabytes of data are expensive to buy or rent. Therefore, the only way that an insurance carrier can make a cost justification for such technology is for each insurance company to conduct a return on investment analysis. The Coalition Against Insurance Fraud engaged SAS, a leading business analytics company, to carry out a study on the usage of anti- fraud detection tools. According to SAS, only 8 percent of insurance companies have stated that, “a lack of ROI was a challenge to implementing anti-fraud weapons” compared to 36% of companies in 2012 (Coalition Against Insurance Fraud 2015c). This survey does not specifically state that these companies utilize machine learning algorithms, but it did establish that 43% of companies have begun investing in predictive analysis and text mining. Cost is certainly a concern, but more and more companies realize that the future cost savings justify the upfront costs.

Another potential concern is the lack of available expertise to implement a machine learning solution. One of the advantages of the proposed solution is that it leverages prior work done by an individual in other industries. Data analysts do not need to develop an initial concept to implement but can instead choose from dozens of implementations conceived by other developers. Data analysts will always need to have a particular skill set. Individuals will need extensive backgrounds in mathematics as well as computer science. However, there exist various training programs to help these developers get the additional skills they need to create machine language algorithms. For example, Stanford University offers a free 11-week online course in Machine Learning including such topics as Support Vector Machines, learning algorithm evaluation and anomaly detection (Coursera 2015). Additionally, companies such as Allstate and Liberty Mutual have begun sponsoring public competitions on websites such as Kaggle to have the creation of the machine learning algorithm outsourced to individuals with the necessary skill sets. Insurance companies many also consider utilizing consultants to get systems up and running and to train employees within the confines of an established budget.

# Recommendations

For the last fifteen years, this author has worked in the insurance industry with a focus on the Massachusetts Property and Casualty market. During these years, he has spent time working as an underwriter, an actuary, a database designer and as a software developer. In many of these positions, he occasionally has had the opportunity to work with his companies Special Investigation Unit in combating various types of insurance fraud. Much of the data that insurance companies analyze is internal in nature or limited to a few vendors who provide information about prior claims. His experience has shown that fraud detection does improve as insurance companies incorporate additional external data into their company’s internal data. Therefore, to better identify fraudulent activities, insurance companies should expand external data collection to capture disparate data sources such as voting registrations and social media.

Fraud is not unique to just the insurance industry. Many other industries, such as the financial, telecommunication and healthcare industries also suffer from various types of fraud. For example, the Communications Fraud Control Association estimates telecommunication fraud to be 46.3 billion dollars per year. However, this translates to only 2.09% of global telecommunication revenues and is very near a five-year low (CFCA 2013). While there are many reasons for this, one factor to consider is that the telecommunication industries, as well as other industries, are developing advanced data mining algorithms to detect fraud. While not every technique developed has been a success, the vast majority of research indicates that the application of machine learning detection in other industries has had a tremendous impact in reducing overall fraud. For example, a major US Bank contracted with IBM Research’s Machine Technologies Group in 2011. The algorithms developed by IBM for this bank resulted in a 15% increase in fraud detection, a 50% reduction in classifying a legitimate claim as fraudulent, and overall savings of 60% (IBM 2011). Yarra Goldschmidt, the manager of IBM’s Machine Learning Technology Group, stated:

"The triple combination of prevention, catching more incidents of actual fraud, and reducing the number of false positives results in maximum savings with minimal hassle. In essence, we are able to apply complicated logic that is outside the realm of human analysis to huge quantities of streaming data."

Traditional methods of fraud identification through data analytics, such as General Linear Models, are not capable of keeping up with the increased rate of fraud. Additionally, these methods are focused on identifying those existing schemes currently being employed by fraudsters with limited ability to quickly identify new and more complex patterns of fraud. Adding to these concerns is that the need to analyze a tremendous amount of data requires that new methods of data analysis must be considered. A logical approach would be to examine both the successes and failures of machine learning applications in regards to fraud detection in other industries. Researchers could then apply these findings to the development of machine learning algorithms to reduce the incidents and severity of insurance fraud.

# Conclusion

The insurance industry estimates that 10% of the incurred losses from claims are due to fraud (Insurance Information Institute 2015). This translates to increased costs to insured of hundreds of dollars per year. For the majority of individuals, these costs would more than justify consumer action to insist insurance companies develop better methods to identify fraud. However, there would still be a large population of individuals that believe that insurance fraud is indeed a victimless crime and that insurers would raise rates even if fraud were nonexistent.

These individuals need to understand that fraud, especially hard fraud, is not victimless and that they at any time could themselves be a victim. People die in staged accidents. Two-year-old Joanna Lopez burned to death in a staged accident in 1997 (Coalition Against Insurance Fraud 2015a) while Altagracia Arias, a sixty-five-year-old grandmother, died from head injuries sustained in a staged accident in 2003 (Insurance Journal 2008).

Whether due to costs or public safety, one thing is clear – insurance companies need to do a better job in reducing all forms of insurance fraud. Research and the author’s experience in fraud detection have shown that fraud detection improves with the amount of data collected. Therefore, the development of a comprehensive solution for increasing external data collection is the first step needed to reduce overall fraud. This increase in data, however, requires a rethinking of the current methods of data analysis as these methods do not perform adequately as the data volume is increased. An answer seems to exist in the utilization by other industries of machine learning applications to identify fraud. By performing a comprehensive analysis of these machine learning algorithms, it should be possible to recreate these algorithms for use in identifying insurance fraud with the same degree of success seen in other industries.

# References

Alpaydin, Ethem. 2014. *Introduction to Machine Learning*. 3rd ed. Cambridge, Mass: MIT Press.

Automobile Insurance Board. 2015. “VIS Odometer Data” Accessed July 16, 2015. https://www.‌aib.‌org‌/ContentPages/PPAuto/VISOdometerData.aspx

Bhowmik, Rekha. 2008. "Data Mining Techniques in Fraud Detection." *Journal of Digital Forensics, Security and Law,* Vol 3, no. 2. Accessed June 13, 2015. http://ojs.jdfsl.o‌r‌g/‌index.php/jdfsl/article/viewFile/180/108.

CFCA. October 10, 2013. “Global Telecom Fraud Increases by 0.21% from 2011, still near 5-year low”. Communications Fraud Control Association. Accessed July 11. 2015. http://www.‌cfca.‌org/pdf/survey/CFCA2013GlobalFraudLossSurvey-pressrelease.pdf

Coalition Against Insurance Fraud. 2007. “Four Faces: Why Americans do— and don't— tolerate insurance fraud.” Accessed July 9, 2015. http://www.insurancefraud.org/four-faces-‌

stud‌y.htm#.VaqlyPlVhBf

———. 2014. “The Power of Partnering – Bridging the Divides: 2013 Annual Report.” Accessed July 10, 2015. http://www.insurancefraud.org/downloads/annualReport13.pdf

———. 2015a. “Insurance fraud backgrounder – Insurance fraud: the crime you pay for.” Accessed July 9, 2015. http://www.insurancefraud.org/fraud-backgrounder.‌htm#.‌VaqyGPlVhBc

———. 2015b. “Fraud statistics.” Accessed July 9, 2015. http://www.insurancefraud.org/‌statistics.htm#.VareI\_lVhBe

———. 2015c. “**Survey: Anti-Fraud Tech Plays Growing Role And Offers ROI, More Insurers Say**.” Accessed July 21, 2015. http://www.insurancefraud.org/news-release-detail.htm?recid=3373#.VbSfJvlVhBf

Coursera. 2015. “Stanford University Machine Learning.” Accessed July 16, 2015. https://www‌.coursera.org/learn/machine-learning/home/info

Data Mining Techniques. May 18, 2015. "Statistics – Textbook." Accessed July 16, 2015. http://documents.software.dell.com/Statistics/Textbook/Data-Mining-Techniques

Estrin, Michael. April 3, 2013. "6 Shadiest Auto Insurance Fraud Schemes" 6 Shadiest

Auto Insurance Fraud Schemes | Bankrate.com. Accessed June 20, 2015. http://‌www.‌bankrate.com/finance/insurance/auto-insurance-fraud-schemes-1.aspx

FBI. March 17, 2010. "Insurance Fraud." Accessed June 13, 2015. https://www.fbi.gov/‌stats-‌

ser‌vices/publications/insurance-fraud

Harris, Derrick. March 6, 2015. "How PayPal Uses Deep Learning and Detective Work to Fight Fraud." Gigaom. Accessed June 14, 2015. https://gigaom.com/2015/03/06/‌how-‌paypal-uses-deep-learning-and-detective-work-to-fight-fraud/

Hennelly, Robert. July 30, 2015. “Why use credit scores to set car insurance rates?” CBS Money Watch. Accessed August 5, 2015. http://www.cbsnews.com/news/why-use-credit-‌‌scores‌‌-to-‌set-car-insurance-rates/

Hilas, Constantinos S., Paris A. Mastorocostas, and Ioannis T. Rekanos. 2015 "Clustering of Telecommunications User Profiles for Fraud Detection and Security Enhancement in Large Corporate Networks: A case Study." *Appl. Math* 9, no. 4: 1709-1718.

IBM. August 26, 2011. “Using machine learning and stream computing to detect financial fraud.” IBM Research. Accessed July 11, 2015. http://www.research.ibm.com/foiling-financial-fraud.shtml

Insurance Information Institute. March 2015. “Insurance Fraud.” Accessed July 9, 2013. http://www.iii.org/issue-update/insurance-fraud

Insurance Journal. October 3, 2008. “Massachusetts Crackdown on Insurance Fraud Sets Milestones.” Accessed July 9, 2015. http://www.insurancejournal.com/news/east‌/2008/10/03/94293.htm

Jergler, Don. December 18, 2014. “Price Optimization Allegations Challenged, NAIC Investigating Practice.” Accessed July 21, 2015. http://www.insurancejournal.‌com/news/national/2014/12/18/350630.htm#

Johnson, Marina Evrim, and Nagen Nagarur. 2015. “Multi-stage methodology to detect health insurance claim fraud.” *Health care management science*

Kissell, Chris. November 26, 2012. “What is a CLUE report – and why should you care” InsWeb.com. Accessed July 16, 2015. http://www.insweb.com/news-features/clue-report.html

Kolyshkina, Inna, Sylvia Wong, and Steven Lim. 2005 "Enhancing Generalised Linear Models

with Data Mining." Casualty Actuarial Society. Accessed June 14, 2015. http://‌casualtyactuarialsociety.com/pubs/dpp/dpp04/04dpp279.pdf

Kolyshkina, Inna, and Richard Brookes. Oct 22, 2002. "Data Mining Approaches to Modeling Insurance Risk." PricewaterhouseCoopers. Accessed June 14, 2015. http://docs.salford-systems.com/insurance4211.pdf

Lesch, William C., and Bruce W. Byars. January 9, 2013. "See No Evil, Speak No Evil: Why Consumers Don’t Report Fraud." Articles on Insurance Fraud. Accessed July 10, 2015. http://www.insura‌ncefraud.org/article.htm?RecID=3220#.Vaqas\_lVhBc

Levin, Tatyana. June 12, 2013. "The Difference Between Soft and Hard Insurance Fraud." Accessed July 9, 2015. http://www.pinow.com/articles/1666/the-difference-between-soft-and-hard-insurance-fraud

Marakas, George M. 2003. "Machines That Can Learn." *In Modern Data Warehousing, Mining, and Visualization: Core Concepts*, 123-154. Upper Saddle River, NJ: Prentice Hall.

Massachusetts Registry of Motor Vehicles. 2014. “Merit Rating Board.” Accessed July 16, 2015. https://www.massrmv.com/MeritRatingBoard.aspx

Mahmoudi, Nader, and Ekrem Duman. 2015. “Detecting credit card fraud by Modified Fisher Discriminant Analysis.” *Expert Systems with Applications* 42, no. 5 (April): 2510-16.

NAIC. February 27, 2015. "Credit-Based Insurance Scores." Accessed June 13, 2015. http://www.naic.org/cipr\_topics/topic\_credit\_based\_insurance\_score.htm.

Ngai, E.W.T., Yong Hu, Y.H. Wong, Yijun Chen, and Xin Sun. 2011. “The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature.” *Decision Support Systems* 50, no. 3 (February): 559-69.

Noda, Kuniaki, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G. Okuno, and Tetsuya Ogata. 2014.

"Audio-visual Speech Recognition Using Deep Learning." *Applied Intelligence Appl Intell*, 722-37.

Phua, Clifton, Vincent Lee, Kate Smith, and Ross Gayler. 2010. "A Comprehensive Survey of Data Mining-based Fraud Detection Research." Accessed June 14, 2015. http://arxiv.org/ftp/‌arxiv‌/papers/1009/1009.6119.pdf

Rawte, Vipula, and G. Anuradha. 2015. “Fraud detection in health insurance using data mining techniques”. 2015 International Conference on Communication, Information & Computing Technology (ICCICT): 1-5

Sahin, Y., and E. Duman. 2011. “Detecting credit card fraud by decision trees and support vector

machines.” Lecture Notes in Engineering and Computer Science 2188, (1): 442-447

Scafidi, Frank. May 10, 2010. "Staged Accident Questionable Claims Up." National Insurance Crime Bureau. Accessed July 10, 2015. https://www.nicb.org/newsroom/news-releases/staged-accident-questionable-claims-up.

Sithic, H. Lookman, and T.Balasubramanian. 2013. “Survey of insurance fraud detection using‌

data mining techniques.” *International Journal of Innovative Technology and Exploring Engineering* 2, (3): 62-65

Soltani Halvaiee, Neda, and Mohammad Kazem Akbari. 2014. “A novel model for credit card fraud detection using Artificial Immune Systems.”*Applied Soft Computing* 24 (November): 40-49.

Texas Department of Insurance. June 26, 2015. “Facts and Frequently Asked Questions.” TDI.Texas.Gov. Accessed July 10, 2015. http://www.tdi.texas.gov/fraud/faq.html

Thornton, Dallas, G. C. van Capelleveen, Mannes Poel, Jos van Hillegersberg, and Roland M. Müller. 2014 . "Outlier-based Health Insurance Fraud Detection for US Medicaid Data." Accessed June 15, 2015. http://eprints.eemcs.utwente.nl/24984/01/49861.pdf

Tidball, Christopher. April 15, 2015. “Counting the cost of America’s insurance fraud epidemic.” Property Casualty 360⁰. Accessed July 10, 2013. http://www.propertycasualty360.com/‌2015‌/04/15/counting-the-cost-of-americas-insurance-fraud-epid

Tu, Jack V. November 1996. "Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes." *Journal of Clinical Epidemiology* 49, no. 11: 1225-231.

Tugend, Alina. April 18, 2014. “As Data About Drivers Proliferates, Auto Insurers Look To Adjust Rates.” New York Times. Accessed July 16, 2015. http://www.nytimes.com/2014‌/04/19/‌your-money/as-data-about-drivers-proliferates-auto-insurers-look-to-adjust-rates.html?smid=tw-share&\_r=0

Verisk. August 1, 2011. "ISO ClaimSearch." System and Services Overview. Accessed July 21, ‌

2015. http://www.verisk.com/images/stories/documents/cls/pdf/System-and-Services-Overview.pdf.