# Predicting Unethical Physician Behavior At Scale: A Distributed Computing Framework

Anastasia Quinn Keck*, Miguel Romero*, Robert Sandor*, Diane Myung-kyung Woodbridge, Paul Intrevado
{akeck,mromerocalvo,risandor,dwoodbridge,pintrevado}@usfca.edu
Data Science Program
University of San Francisco

***Abstract*—As the amount of publicly shared data increases, developing a robust pipeline to stream, store and process data is critical, as the casual user often lacks the technology, hardware and/or skills needed to work with such voluminous data. In this research, the authors employ Amazon EC2 and EMR, MongoDB, and Spark MLlib to explore 28.5 gigabytes of CMS Open Payments data in an attempt to identify physicians who may have a high propensity to act unethically, owing to significant transfers of wealth from medical companies. A Random Forest Classifier is employed to predict the top decile of physicians who have the highest risk of unethical behavior in the following year, resulting in an F-Score of 91%. The data is also analyzed by an anomaly detection algorithm that correctly identified a high-profile case of a physician leaving his prestigious position, as he failed to disclose anomalously-large transfers of wealth from medical companies.**

***Index Terms*—Distributed computing, Machine learning, Anomaly detection, Random Forest Classifier, Medical payments**

## I. Introduction

Sectors that deal with vast amounts of public data, such as healthcare, have long held the potential to unlock untold mysteries about the populations they serve. Until recently, the amount of data available for analysis far outstripped the abilities of both the technology and machine learning algorithms necessary to extract actionable information. Very recent advances in data and computational science have allowed researchers to tap into and identify patterns and relationships hidden in this sea of data. In healthcare, this leap has facilitated the identification of issues, both clinical and administrative, throughout the healthcare continuum.

Often times, medical research is focused on the clinical, owing to the high salaries of physicians, significant costs of procedures for patients, the costly operation of medical facilities, and the relatively limited amount of data required for well-scoped medical studies, e.g., a study on hypertension. However, recent advances have enabled researchers to comb through the vast amounts of data associated with medical administration.

One salient facet of healthcare administration is the interconnected nature of the companies who provide medical supplies, devices and drugs to the physicians who use and/or prescribe the aforementioned products. There are several examples supporting the hypothesis that a physician receiving disproportionately large transfers of wealth or value from an organization, may be more inclined, persuaded, or outright fraudulent in concluding that certain medication, procedures, or medical devices are more effective than they truly are. Such payments or transfers of value have been formally linked to unethical physician and/or institutional behavior [1], [2], [3]. However, the ability to apply machine learning algorithms at scale to analyze all physicians receiving transfers of wealth has been elusive. The authors have therefore mined the Open Payments data from the Center for Medicare & Medicaid Services (CMS) to analyze all payments or other transfers of value from group purchasing organizations (GPOs) and device and drug manufacturers to physicians or research institutions.

With an impressive 28.5GB of available data from 2013–2017, the authors tested the boundaries of a distributed computing framework for this data set, identifying a systems configuration—after many iterations—sufficiently computationally powerful and robust to store, manage, process, and analyze data at this scale. Once established, the authors explored the computational performance of a Random Forest classifier to identify those physicians who are predicted to rank in the top decile of all those receiving payments or other transfers of value from purchasing organizations (GPOs) and device and drug manufacturers in the aggregate. This *Audit List* identifies a subset of physicians—based on mean transfers of wealth—who warrant additional scrutiny, similar to how the IRS audit a small fraction of yearly tax returns [4].

Using anomaly detection techniques, the authors also identified individual transfers of wealth to physicians or research organizations that were anomalously high, warranting perhaps the highest level of scrutiny. In fact, the anomaly detection technique employed identified a physician that recently left his position, and was cited by the New York Times for potential conflicts of interest owing to his receiving disproportionate payments or other transfers of value.

Machine learning models were tested across several distributed computing system configurations using Amazon Web Services (AWS) Simple Storage Services (S3) and Apache Spark on Elastic Map Reduce (EMR).

## II. Background

Publicly available government data sets can provide various social and economic benefits. Many government organizations have begun to publish more data every year. Publishing data

* These authors contributed equally.

helps governments to crowwdsource and improve upon the quality of public services, and their outcomes [5]. It also eases interaction and cooperation between public and private sectors, in an effort to improve economic growth [6]. The United States government releases and maintains many publicly available data sets; the breadth and volume of these data sets increases yearly. For instance, `data.gov`—the United States government responsible for publishing many public data sets—currently maintains over 246,000 publicly available data sets; it only had 312 publicly available data sets in 2009 [7], [8].

In order to collect and process CMS Open Payments and related data sets, it is essential to develop a sustainable data pipeline. As the volume of data—which is already very large—continues to grow, it is particularly important to develop a scalable, time- and cost-efficient data storage and processing system.

*1) Cloud Computing:* Cloud computing utilizes storage and computing resources in multiple data centers connected via a network, and provides services on demand to their users. Cloud computing is highly scalable and user-friendly, reacting to user needs dynamically by scaling resources based on needs, and providing IT infrastructure and maintenance [9]. Minimizing costs by providing shared hardware resources and maintenance services, cloud computing has become a powerful tool for individuals and organizations to store, manage and process a large volumes of data [10], [11], [12].

*2) MapReduce and Apache Spark:* Hadoop MapReduce, introduced by Google, is a programming paradigm for processing a large volume of data in parallel by dividing a task into a set of subtasks, and processing them in parallel. MapReduce is designed to run on a single machine, or on a cluster with multiple machines to efficiently process data. For MapReduce, users have to design map and reduce steps, where a map function, such as filtering, processes key-value pairs in parallel. A reduce function takes the outputs of the map function as input from multiple machines and executes a summary operation, returning a single answer to a driver [13]. MapReduce is a highly efficient model, and it and its variations are actively used in both research and industry [14], [15], [16].

Apache Spark adopts the MapReduce model, but executes a task 100 times faster than MapReduce by processing data in memory. Also, Spark uses efficient job scheduling and recovery model using directed acyclic graph representation, and still runs 10 times faster in disk than MapReduce [17], [18], [19].

*3) Amazon Web Services:* Amazon Web Services (AWS) is the largest public cloud service provider, providing various services including data storage, management, computing, analytics, etc. [20]. Cloud storage is a type of cloud computing where data is stored in multiple servers throughout multiple locations, but its virtualization capabilities makes its distributed nature seamless. Additionally, cloud storage provides high durability by maintaining duplicated copies in different machines [21]. AWS Simple Storage Services (S3) is a cloud storage system, adopting a "pay-as-you-go" charging model, offering infinite storage capacity, data redundancies, 99.99% availability, and low data access latency [22].

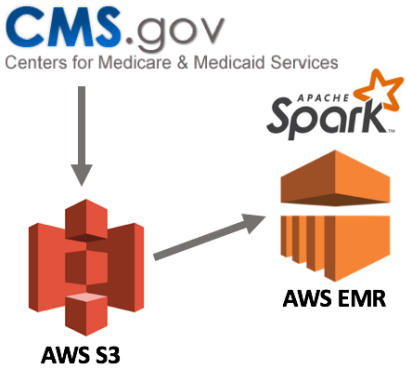

Fig. 1: Data science pipeline

AWS Elastic Map Reduce (EMR) provides a fully managed Hadoop framework hosted on Amazon Elastic Compute Cloud (EC2) instances [23]. EC2 is an AWS cloud computing environment where users can configure the operating system, CPU, memory, disk, network, etc. [24]. Based on Iosups study, EC2 demonstrates better performance and lower overall costs compared to other competing cloud computing services suitable for a large scale scientific studies [25]. AWS EMR provides cluster maintenance and big data processing tools including Spark, Hadoop MapReduce, HBase, Flink, etc., and allows users to easily configure and launch a cluster, which is easily resizable [26]. In this research, we configured various AWS EMR clusters with Apache Spark installed.

## III. System Overview

### A. System Workflow

The data science pipeline was designed around scalability as well as the ability to efficiently run machine learning algorithms, leveraging cloud resources and distributed methods. Technologies were selected to build a scalable data ingestion, audit generation, and anomaly detection system. We therefore selected Amazon Web Services (AWS) as the primary platform to host storage, data extraction, transform and load (ETL) processes and machine learning tasks. The data pipeline, including data and distributed storage and computing services is depicted in Figure 1.

*1) Data Sources:* The publicly available Open Payments data was obtained from the Center for Medicare & Medicaid Services (CMS) [27]. This data set includes payments or transfers of value to physicians and research organizations from group purchasing organisations (GPOs), medical device and drug manufacturers. It also contains data on physician ownership and investment interests. Additional details about the data can be found in § IV-A.

*2) Data Storage:* The Open Payment CMS data was stored in AWS Simple Storage Services (S3) to ensure scalablitiy and data integrity, while minimizing time and cost for server maintenance.

*3) Data Processing:* Feature engineering and machine learning was performed on an EMR cluster by processing data read from S3. Accessing data stored in S3 from EMR allows users avoid storing multiple copies of data in each cluster node, by looking up data in a central storage system. EMR also provides a built-in function to load and write data to/from S3 efficiently.

### B. Algorithms

Variations on a Random Forest Classifier were used to generate both the *Audit List*, as well as to detect individual anomalous transfers of wealth to physicians or research organizations.

*1) The Audit List:* The objective of the *Audit List* is to predict which physicians or research organizations will receive sufficiently large *mean* transfers of wealth from GPOs, medical device and drug manufacturers, that may be indicative potentially unethical behavior. The measure of interest, mean transfers of wealth, is computed by taking the sum of all transfers of wealth to a physician or research organization across all years, and dividing that sum by the total number of transfers of wealth to said physician or research organization.

To identify physicians or research organizations potential at risk of unethical behavior, aggregate CMS data from 2013–2017 (inclusive) was input into a Random Forest Classifier to generate predictions. An 80/20 train/test cross validation split across years was used to train the Random Forest Classifier.

A Random Forest is an ensemble of decision trees, where the output is either a mean prediction (regression setting) or the mode of the classes (classification setting) of the constituent trees. Each individual tree is exposed to a (potentially bootstrapped) subset of the rows and columns and consists of a series of binary splits that look to optimize a loss function. The compartmentalized nature of the algorithm with independent trees is especially well-suited to a distributed framework.

Tree depth is an important hyper-parameter of a random forest algorithm. Although increasing tree depth improves a model's predictive accuracy, it requires a longer training time. Moreover, it may cause overfitting, a scenario where the model is overly sensitive to training data. Spark's MLlib provides a method to specify the maximal depth of any individual tree within the forest, allowing us to analyze the relationship between tree depth and the size of data used to train the random forest.

*2) Anomaly Detection:* Whereas the *Audit List* seeks to analyze mean transfers of wealth in the aggregate, anomaly detection seeks to detect singularly large, outlying transfers of wealth that may be indicative of potentially unethical behavior. Although the ability to detect anomalous data is a skill relevant across all fields that employ data, is has been historically used very heavily in the field of cyber-security [28], [29], [30], [31]. The authors employed an anomaly detection technique similar to the one outlined by Shi et al. [32]—with a case study provided by Dan Mallinger [33]—effectively turning a supervised machine learning technique into an unsupervised learning technique.

Supervised machine learning trains an algorithm on data that contains both inputs and an target variable. E.g., trying to predict house prices (target variable) based on several features such as square feet, zip code, number of bathrooms. A data scientist may use linear regression—a supervised learning technique—to train the model using both input data ($x_1$ = square feet, $x_2$ = zip code, $x_3$ = number of bathrooms) and target data ($y$ = known sales prices). Once the model is trained, it can be used to make predictions about future home sales prices ($\hat{y}$) given $x_1$, $x_2$, and $x_3$.

Unsupervised machine learning trains an algorithm on data that contains exclusively input variables. E.g., one may be interested in finding out which customers are *similar*, given a set of input features such as age ($x_1$), job title ($x_2$), years of higher education ($x_3$), etc. The term *similar* here is intentionally nebulous: loosely speaking, the algorithm is left to its own devices to discover interesting structures in the data, as there is no output data ($y$) provided to the model on which to train.

The struggle with applying anomaly detection to the CMS data to identify individual physicians or research organizations that have an anomalously-high propensity for unethical behavior, is that the data is not labelled, i.e., there is no feature in the CMS data set that identifies individual physicians or research institutions that have behaved unethically in the past. With over half a billion individual physicians and research organizations that have received at least one transfer of wealth in the past five years, it is certainly infeasible to manually identify unethical behavior, and extremely difficult to do so by scraping information from the internet.

When applying machine learning to anomaly detection, the authors selected a Random Forest Classifier—a supervised learning technique—as the algorithm of choice, with a clever modification. This modification resulted in a pseudo-Unsupervised Random Forest Classifier. To achieve this modification, an additional binary indicator variable (column) is added to the existing unsupervised data. All observations in the original data are labeled as *non-anomalous*. This label is effectively a target ($y$) for the previously unlabeled data, allowing the data to be used in a supervised machine learning algorithm.

To intentionally create anomalous data, a copy of the original CMS data is taken, and the values (cells) in the data set are randomly permuted across rows, i.e., the cells were randomly shuffled. These random permutations create a data set whose individual values are non-anomalous, but whose relationships across columns *are* anomalous, resulting in *anomalous* data. An additional binary indicator variable is also added to the intentionally-anomalous data, and those observations are labelled as *anomalous*, and can now similarly be used as input to a supervised machine learning algorithm. Finally, the original CMS data—with the newly appended *non-anomalous* labelled column—is concatenated row-wise with the new *anomalous* data. This newly created data set now has twice as many observations (rows) as the original data, and one additional variable (column), the indicator as to whether

or not the observation is anomalous.

The pseudo-Unsupervised Random Forest Classifier is subsequently trained on the modified data, and predictions generated based on *individual* transfers or wealth made to physicians or research organizations. Alternate machine learning algorithms were tested, with mixed results. Certain classification algorithms generated results inferior to those generated by the Random Forest Classifier. Other algorithms, such as XGBoost—a scalable tree boosting algorithm—took almost a full day to run, far longer than the Random Forest Classifier. Owing to the high costs associated with running and maintaining the large and complex distributed computing framework associated with this research and data, the authors lacked the funds to run an extensive comparative analysis of algorithms, run times, and predictive ability. Moreover, in the interest of uniformity, the ability to use a Random Forest Classifier for both the analysis of aggregate as well as individual anomalous physicians payments was preferred.

## IV. Experimental Output

### A. Data

As of 2013, the Center for Medicare & Medicaid Services (CMS) publishes yearly data documenting payments or transfers of value to physicians and research organizations, as well as physician ownership and investment interests [27]. All physicians and/or research organizations who receive payments or transfers of value from group purchasing organizations (GPOs) and device and drug manufacturers are required to report those transactions to the CMS. The physician disciplines included in this data set are: medical doctors, podiatrists, osteopaths, dentists, ophthalmologists, and chiropractors. All published CMS data was employed in this research, spanning the years 2013–2017, totalling 28.5 gigabytes of information. The raw data contains 58 variables (columns), and 52,992,403 observations.

As this research seeks to answer two different hypotheses, the data was used differently to answer each question. To identify the top decile of physicians who we predict may operate in a potentially unethical manner, the CMS data was subset to exclude research organizations, so as only to focus on physicians, resulting in a data set that had 49,026,626 observations (i.e., 92.5% of the original CMS data was retained). After some initial exploratory data analysis and feature engineering, eleven features were derived from the original 58 variables, and are outlined in Table I. This was also aggregated at the physician level, including only those physicians who received one or more transfers of wealth over the five year period spanning 2013–2017, resulting in 976,208 aggregate observations.

In an effort to identify single, anomalous transfers of wealth to physicians or research organizations, the entirety of the 52,992,403 observations of CMS data was employed. The anomaly detection algorithm was run separately on the physician data (98,053,252 observations = 2 $\times$ 49,026,626 observations) and on the research organization data (7,931,554 observations = 2 $\times$ 3,965,777), as running the entire CMS dataset through the algorithm was infeasible owing to the size of the data and computational complexity of the algorithm.

TABLE I: Random Forest Features

| Feature | Definition |
|---|---|
| 1 | The sum total number of years a physician received transfers of wealth |
| 2 | A tally of the total number of individual transfers of wealth to a a physician |
| 3, 4 | Mean of all (3) cash payments of a physician's type or (4) alternative transfers of wealth |
| 5 | The fraction of transfers of wealth that were reported to the CMS with a delay |
| 6 | Fraction of transfers of wealth that were in cash |
| 7 | The city in which physician practices |
| 8 | Primary type of medicine practiced by the physician: medical doctor, podiatrist, osteopath, dentist, ophthalmologist, or chiropractor |
| 9 | Physician's specialty, selected from standardized provider taxonomy [34] |
| 10 | The state in which physician practices |
| 11 | Number of research of payments the physician received |
| 12 | The number of unique companies transferring wealth to the physician |

TABLE II: System Configurations for Audit List

| Master/Slave Config | Slaves | Memory | Cores | Run Time |
|---|---|---|---|---|
| c3.8xlarge | 4 | 60GB | 32 | 22m23s |
| c3.8xlarge | 2 | 60GB | 32 | 37m12s |
| m4.xlarge | 8 | 16GB | 4 | 37m43s |
| m4.xlarge | 4 | 16GB | 4 | 64m13s |
| m4.xlarge | 2 | 16GB | 4 | `n/a` |
| local machine | `n/a` | 16GB | 4 | `n/a` |

### B. Results

Perhaps the most daunting task was to establish a distributed computational infrastructure that would reliably process such a large amount of data. Table II provides a summary of the AWS EC2 configurations, including the number of slaves, memory (same across both master and all slaves), the number of cores (same across both master and all slaves), and the total run time required to process all data (inclusive of running the Random Forest Classifier on the aggregate CMS data).

The final two rows of Table II indicate run times of `n/a`, signifying that the data processing and/or Random Forest Classifier algorithm failed to compute, owing to memory errors.

Figure 2 breaks down the time required by each configuration to both process the data and run the Random Forest Classifier. Note that $x$-axis category titles are the concatenation of several configuration features, e.g., `c3.8xlarge_60GB_32c_4s` implies a `c3.8xlarge` configuration with 60GB of memory, 32 cores and 4 slaves.

With a reliable distributed computing system identified, the authors quantified the quality of the Random Forest Classifier to generate the *Audit List* on the aggregate CMS data. For classification algorithms, a traditionally reported metric of success is the F-Score (also known as an $F_1$-Score or F-Measure). The value of an F-score falls between 0 and 1, with 0 being the worst possible score, and 1 being the best. An F-Score is an aggregate measure, the harmonic mean of two other measures, *precision* and *recall*:

$$\text{F-Score} = 2\Big(\frac{precision \cdot recall}{precision + recall}\Big) \tag{1}$$

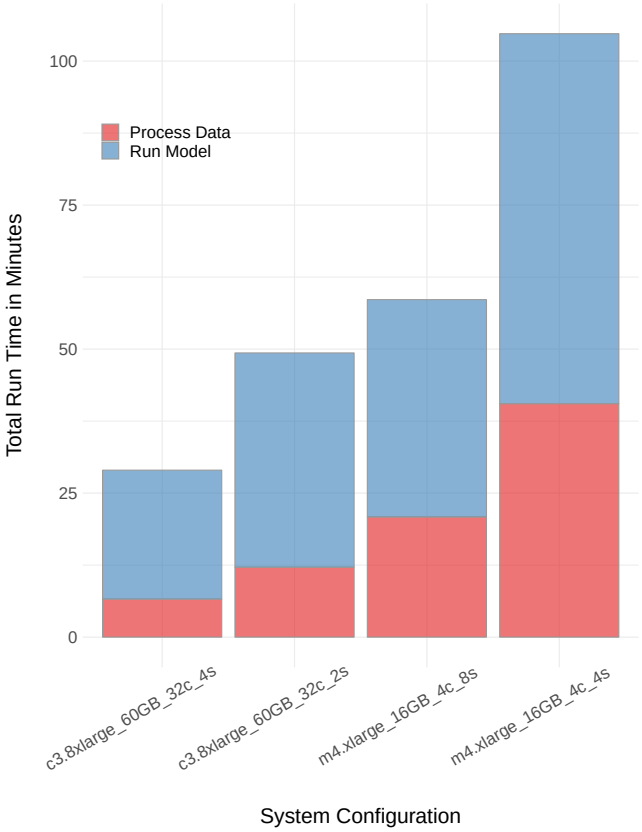


Fig. 2: Run Times vs. System Configurations for Audit List

where *precision* and *recall* are defined as follows:

$$precision = \frac{\text{true positives}}{\text{true positives + false positives}} \tag{2}$$

$$recall = \frac{\text{true positives}}{\text{true positives + false negatives}} \tag{3}$$

Running the Random Forest Classifier on the aggregate CMS physician data resulted in an F-Score 91%, a very strong result when predicting those physicians that are expected to be in the top decile of those receiving transfers of wealth, and therefore placed on the *Audit List*. The *Audit List* contains roughly 60,000 physicians. Although this may seem like a large number and/or fraction of total physicians to audit, in 2017, the IRS audited 1.1 million tax returns [35].

Also of note is that the top decile—or 90th percentile—of all mean transfers of wealth to physicians is only $258 US Dollars. This means that on average, any physician who received yearly average transfers of wealth that exceeded $258 warrant additional scrutiny. Owing to the trivial amount of $258, the practical implications are that very few physicians regularly receive transfers of wealth. Those physicians who do *regularly* receive virtually any transfer of wealth from group purchasing organizations (GPOs) or devices or drug manufacturers on a regular basis warrant additional scrutiny.

The second task was the identification of anomalous individual transfers of wealth to either physicians or to research institutions, that would be indicative of unethical behavior. Finding a stable and sufficiently powerful system configuration for this task was painstaking, owing to the significant computational requirements involved in the pre-processing of the data—duplicating, permuting, and finally concatenating up to 100 million records and 59 variables—as well as running the anomaly detection algorithm on such a large data set. Table III details the AWS EC2 configurations, including the number of slaves, memory (same across both master and all slaves), the number of cores (same across both master and all slaves), and the total run time required to process all data, inclusive of running the anomaly detection algorithm.

TABLE III: System Configurations for Audit List

| Master/Slave Config | Slaves | Memory | Cores | Run Time | Data |
|---|---|---|---|---|---|
| c3.8xlarge | 10 | 60GB | 32 | 81m | Research |
| c3.8xlarge | 6 | 60GB | 32 | 118m | Research |
| c3.8xlarge | 2 | 60GB | 32 | 165m | Research |
| c3.8xlarge | 10 | 60GB | 32 | 1201m | Physician |

In order for the pre-processing and algorithmic analysis not to fail, the CMS data was split into two separate data sets: one exclusively for research organizations, and another solely for individual physicians. Recall that, as a result of the required data duplication for the anomaly detection (see §III-B), data associated with transfers of wealth to research organizations contained 7,931,554 observations ($2 \times 3{,}965{,}777$) and 59 variables, totaling 1.3GB of information. CMS data associated with transfers of wealth to individual physicians was an astonishingly large 98,053,252 observations ($2 \times 49{,}026{,}626$ observations) and 59 variables, totaling 47.3GB of data.

Figure 3 breaks down the time required by each configuration to both process the data and run the anomaly detection algorithm for the smaller CMS *research organization* data. Note that $x$-axis category titles are the concatenation of several configuration features, e.g., `c3.8xlarge_60GB_32c_10s` implies a `c3.8xlarge` configuration with 60GB of memory, 32 cores and 10 slaves.

CMS data associated with transfers of wealth to individual physicians was not tested on various machine configurations, nor was it run more than once, on account of the significant amount of cost and time required. When it was successfully run however, 270 minutes were required to pre-process the data, and 931 minutes were required to run the anomaly detection algorithm.

There is no ground truth by which the authors can verify the veracity of the algorithmic results. To the best of our knowledge, there exists no single federal list of provably unethical behavior by specific physicians which can be mapped to the results generated by the anomaly detection algorithm. There was, however, a timely and notable validation of the authors' research: in September 2018, Dr. José Baselga, the former chief medical officer of Memorial Sloan Kettering Cancer Center—**and a physician flagged by the anomaly detection algorithm**—resigned his position in connection to

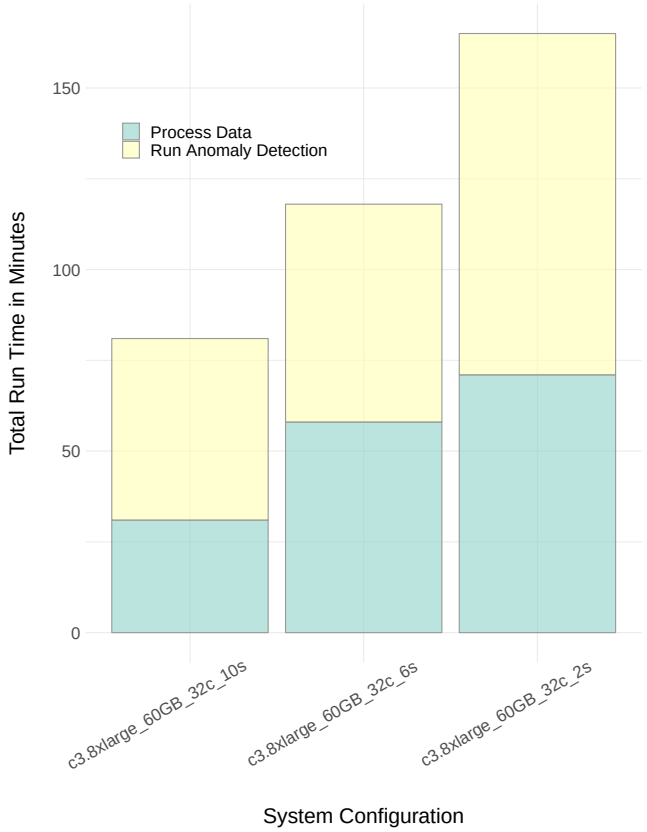


Fig. 3: Run Times vs. System Configurations for Anomaly Detection on Research Institutions

his failures to disclose millions of dollars in payments from GPOs, medical device and pharmaceutical companies [36].

## V. Conclusion

The vast amount of data being publicly shared by U.S. government institutions has the potential to pull back at least some of the opaque layers surrounding large, government-run institutions and operations. With CMS Open Payments data, it is now possible for an individual patient to sift through the data and examine whether or not a particular physician or organization is receiving any notable transfers of wealth, from whom, and with what frequency; this task is, however, not for the uninitiated.

Additionally useful is the ability to compare and contrast how different physicians and research organizations are related to each other with respect to their connections to group purchasing organizations (GPOs), medical device and drug manufacturers. This seemingly benign exploratory data analysis task takes on a whole new dimension when one realizes they will have to process almost 28.5 gigabytes of data, far too large for any single machine. The scale of the data leaves comprehensive analysis of this data just out of reach of the average individual.

The authors have therefore developed a distributed computing framework using several Amazon Web Services as well as Apache Spark to ingest, process, and distill the data into a smaller, more manageable aggregate data set. Moreover, using *mean* transfer of wealth to each physician over the last five years as a target variable, the authors used a Random Forest Classifier to identify the top decile of all physicians who demonstrate the highest propensity for exhibiting potentially unethical behavior in the coming year, resulting in an F-Score of 91%.

Using anomaly detection techniques, the authors also processed up to 100 million records to identify anomalously large *individual* transfers of wealth to physicians. Although no ground truth exists on which to validate the model's results, a prominent physician who recently left his prestigious role as the chief medical officer owing to his failure to disclose millions of dollars in payments from health care companies, was identified by the anomaly detection algorithm as receiving highly suspect individual transfers of wealth.

## References


[1] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 1, p. 18, 2018.

[2] S. E. Hadland, M. Cerdá, Y. Li, M. S. Krieger, and B. D. Marshall, "Association of pharmaceutical industry marketing of opioid products to physicians with subsequent opioid prescribing," *JAMA internal medicine*, vol. 178, no. 6, pp. 861–863, 2018.

[3] Emma Ockerman, Vice News. (2019) The more drugmakers wowed doctors with gifts and lunches, the more people died of drug overdoses, study shows. [Online]. Available: https://news.vice.com/en_us/article/59xm4a/the-more-drugmakers-wowed-doctors-with-gifts-and-lunches-the-more-people-died-of-drug-overdoses-study-shows?utm_source=reddit.com

[4] United States Government Accountability Office. (2018) Tax fraud and noncompliance: Irs could further leverage the return review program to strengthen tax enforcement. [Online]. Available: https://www.gao.gov/assets/700/693374.pdf

[5] D. Analytics, "Open data: Driving growth, ingenuity and innovation," 2011.

[6] T. M. Harrison, S. Guerrero, G. B. Burke, M. Cook, A. Cresswell, N. Helbig, J. Hrdinová, and T. Pardo, "Open government and e-government: Democratic challenges from a public value perspective," *Information Polity*, vol. 17, no. 2, pp. 83–97, 2012.

[7] (2019) Data.gov. Data.gov. [Online]. Available: https://www.data.gov/

[8] L. Ding, D. DiFranzo, A. Graves, J. R. Michaelis, X. Li, D. L. McGuinness, and J. Hendler, "Data-gov wiki: Towards linking government data," in *2010 AAAI Spring Symposium Series*, 2010.

[9] T. C. Chieu, A. Mohindra, A. A. Karve, and A. Segal, "Dynamic scaling of web applications in a virtualized cloud computing environment," in *2009 IEEE International Conference on e-Business Engineering*. IEEE, 2009, pp. 281–286.

[10] B. Furht and A. Escalante, *Handbook of cloud computing*. Springer, 2010, vol. 3.

[11] Z. Chaczko, V. Mahadevan, S. Aslanzadeh, and C. Mcdermid, "Availability and load balancing in cloud computing," in *International Conference on Computer and Software Modeling, Singapore*, vol. 14, 2011.

[12] D. Zissis and D. Lekkas, "Addressing cloud computing security issues," *Future Generation computer systems*, vol. 28, no. 3, pp. 583–592, 2012.

[13] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[14] A. Howard, T. Lee, S. Mahar, P. Intrevado, and D. Woodbridge, "Distributed data analytics framework for smart transportation," in *IEEE 16th International Conference on Smart City*. IEEE, 2018, pp. 1374–1380.

[15] J. Ma, A. Ovalle, and D. M.-k. Woodbridge, "Medhere: A smartwatch-based medication adherence monitoring system using machine learning and distributed computing," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 4945–4948.

[16] C. Dong, L. Du, F. Ji, Z. Song, Y. Zheng, A. Howard, P. Intrevado, and D. Woodbridge, "Forecasting smart meter energy usage using distributed systems and machine learning," in *IEEE 16th International Conference on Smart City*. IEEE, 2018, pp. 1293–1298.

[17] L. Gu and H. Li, "Memory or time: Performance evaluation for iterative operation on hadoop and spark," in *High Performance Computing and Communications & 2013 IEEE International Conference on Embedded and Ubiquitous Computing (HPCC_EUC), 2013 IEEE 10th International Conference on*. IEEE, 2013, pp. 721–727.

[18] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets." *HotCloud*, vol. 10, no. 10-10, p. 95, 2010.

[19] Apache Spark, "Apache spark: Lightning-fast cluster computing," 2019. [Online]. Available: http://spark.apache.org

[20] Amazon Web Services. (2019) Amazon web services (aws). [Online]. Available: https://aws.amazon.com/

[21] P. Gupta, A. Seetharaman, and J. R. Raj, "The usage and adoption of cloud computing by small and medium businesses," *International Journal of Information Management*, vol. 33, no. 5, pp. 861–874, 2013.

[22] Amazon Web Services . (2019) Amazon s3. [Online]. Available: https://aws.amazon.com/s3/

[23] P. Deyhim, "Best practices for amazon emr," *Technical report*, 2013.

[24] Amazon Web Services. (2019) Amazon ec2. [Online]. Available: https://aws.amazon.com/ec2/

[25] A. Iosup, S. Ostermann, M. N. Yigitbasi, R. Prodan, T. Fahringer, and D. Epema, "Performance analysis of cloud computing services for many-tasks scientific computing," *IEEE Transactions on Parallel and Distributed systems*, vol. 22, no. 6, pp. 931–945, 2011.

[26] Amazon Web Services. (2019) Overview of amazon emr architecture. [Online]. Available: https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-overview-arch.html

[27] (2019) Open payments dataset. Centers for Medicare and Medicaid Services. [Online]. Available: https://www.cms.gov/OpenPayments/Explore-the-Data/Dataset-Downloads.html

[28] E. Eskin, "Anomaly detection over noisy data using learned probability distributions," 2000.

[29] S. Omar, A. Ngadi, and H. H. Jebur, "Machine learning techniques for anomaly detection: an overview," *International Journal of Computer Applications*, vol. 79, no. 2, 2013.

[30] T. Shon and J. Moon, "A hybrid machine learning approach to network anomaly detection," *Information Sciences*, vol. 177, no. 18, pp. 3799–3821, 2007.

[31] P. K. Chan, M. V. Mahoney, and M. H. Arshad, "A machine learning approach to anomaly detection," Tech. Rep., 2003.

[32] T. Shi and S. Horvath, "Unsupervised learning with random forest predictors," *Journal of Computational and Graphical Statistics*, vol. 15, no. 1, pp. 118–138, 2006.

[33] D. Mallinger. (2015) Unsupervised anomaly detection with spark. [Online]. Available: https://mapr.com/ebooks/spark/08-unsupervised-anomaly-detection-apache-spark.html

[34] Center for Medicare and Medicaid Services. (2019) Crosswalk medicare provider/supplier to healthcare provider taxonomy. [Online]. Available: https://data.cms.gov/Medicare-Enrollment/CROSSWALK-MEDICARE-PROVIDER-SUPPLIER-to-HEALTHCARE/j75i-rw8y

[35] United States Internal Revenue Service. (2019) Enforcement: Examinations. [Online]. Available: https://www.irs.gov/statistics/enforcement-examinations

[36] Katie Thomas and Charles Ornstein, New York Times. (2018) Top sloan kettering cancer doctor resigns after failing to disclose industry ties. [Online]. Available: https://www.nytimes.com/2018/09/13/health/jose-baselga-cancer-memorial-sloan-kettering.html