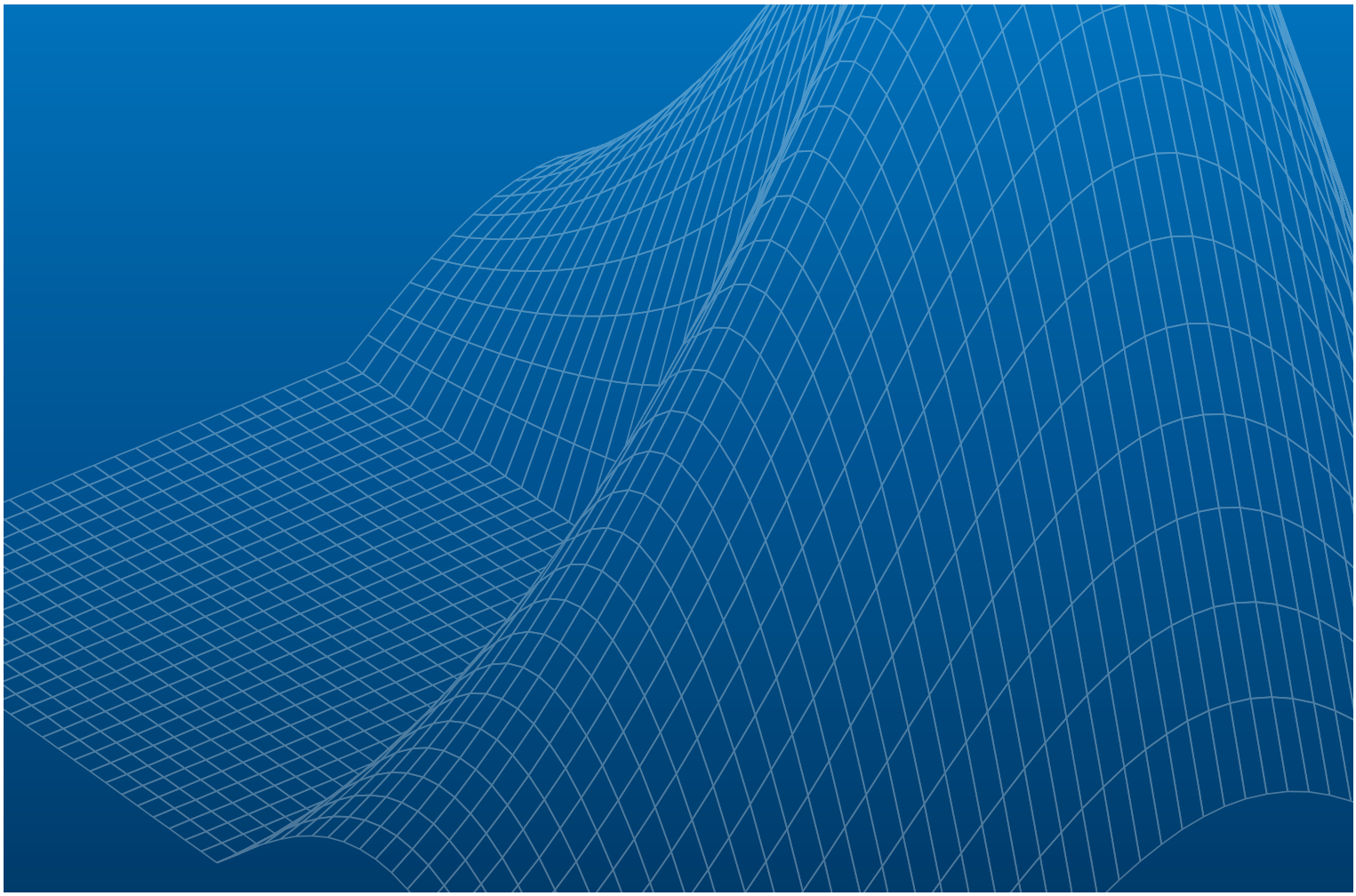

Finding Fraud in Big Data Haystack Saves Billions

Fraud Detection Through Automated Data Analytics

WHITE PAPER

MathWorks®

Introduction

Fraud is a global concern for all types and sizes of organizations and institutions. It is seen across all sectors: financial services firms, government agencies, and public administration organizations.

The scale of losses attributed to fraud is enormous. For example, the 2014 Global Fraud Study of the Association of Certified Fraud Examiners (ACFE) highlights the losses of one particular type, occupational fraud, which it defines as:

The use of one's occupation for personal enrichment through the deliberate misuse or misapplication of the employing organization's resources or assets. (ACFE)

The study estimates that a typical organization loses 5% of revenues each year due to occupational fraud. For 2014, this translates to a projected global loss of nearly $3.7 trillion. Occupational fraud is one of the biggest loss categories of fraud to date, although the newest category, cyber fraud, is increasing at an alarming rate. In 2014, McAfee estimated over $400 billion in losses from cybercrime, which is 5 times the 2012 estimates of $72 billion.

In the banking industry, fraud is considered part of the operational risk framework as defined within the Basel framework, categorizing internal and external sources. External sources include hacking damage and cyber theft of information, as well as more traditional third party theft, such as forgery. External sources of fraud often account for less than 10% of total bank operational risk exposures, but more than 40% of the total operational risk events. (As an example, see Barclays operational risk profile, page 100, in their PLC report.) Detecting external fraud can be compared to finding a needle in a haystack: the fraudulent event is the needle; to find it requires sifting through a heterogeneous data universe – the haystack.

Other industries also suffer significant fraud-related losses. Based on insurance industry data from 2009 to 2013, about 10% of property/casualty insurance losses are due to fraud. This equates to about $32 billion each year.

The longer it takes to detect fraud, the more financial damage is caused. Consequently, proactive fraud detection techniques should seek to catch fraudulent activity early to limit losses.

This paper provides an overview of the challenges associated with fraud detection and shows how organizations can use MATLAB® to address these challenges.

Big Data and the Needle in a Haystack: Challenges Associated with Fraud Detection

Systematic fraud detection presents several challenges:

- analyzing vast amounts of data
- identifying the best modeling and analysis techniques
- understanding the context in which the models are used

In the past, traditional data analysis techniques focused on first formulating a hypothesis and then

gathering and organizing data to test the hypothesis. This technique approaches modeling as a calibration problem. (I have a model; let's figure out the parameters.) With information easily available online, and the cost of data storage diminishing (leading to vast repositories of heterogeneous data), computational intensive modeling approaches have changed the paradigm from data analysis to data analytics.

These elements, distinguishing data analytics from traditional data analysis, are depicted in figure 1. The primary goal of data analytics is to leverage existing data sources to develop predictive models. This is accomplished using advanced computationally intensive statistical or machine learning algorithms. A recent article, How PayPal uses deep learning and detective work to fight fraud, describes how PayPal uses machine learning to detect and combat fraud and demonstrates how fast the market is adopting the data analytics approach.

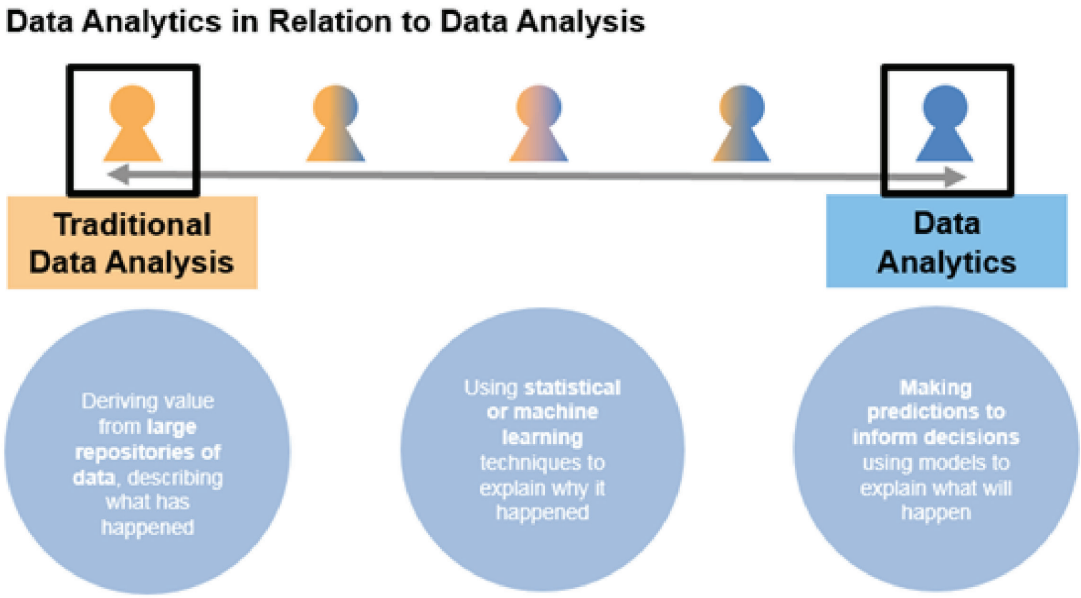

Figure 1. Differentiators between traditional data analysis and modern data analytics.

Data Related Challenges

Fraud detection methods require complex investigations that involve processing large amounts of heterogeneous data. This data is derived from multiple sources and crosses multiple knowledge domains, including finance, economics, business, and law.

Furthermore, fraud is "a needle in a haystack" problem because only a very small fraction of the data will likely represent a fraudulent case. The vast quantity of regular data—that is, data produced from non-fraudulent sources—tends to minimize cases of real fraud. Therefore, identifying the few cases of fraud, the needle in the haystack, is not trivial.

Exponentially increasing amounts of collected data can make it increasingly difficult to identify occasional suspicious patterns and relevant information. So far, professionals tasked with anomaly detection have had to rely on their experience and good instincts to find fraud. However, gathering and processing this data manually is prohibitively time-consuming as well as error-prone. This volume is the first of the 4 Vs associated with fraud detection on big data (Figure 2).

The second V, variety, relates to the different heterogeneous sources of data used to monitor and detect fraudulent behavior. Data collected and used for fraud detection is pulled from many sources that are pieced together to characterize the difference between normal and fraudulent behavior.

Time is of the essence. Detecting fraudulent activity and subsequently mitigating or preventing losses needs to be done in real-time. This demands increasing the velocity of data (the third V) acquisition, capture, and analysis to near real-time.

The fourth V, veracity, revives the needle-in-a-haystack problem once more. Not only does fraudulent behavior account for a small fraction of the data, but data errors, faulty data, and misclassified data can lead to false positives for detecting fraud, reducing reliability in the predictive ability of these models. Data must be scrubbed for outliers, missing data points, and erroneous data to ensure it is of high quality.

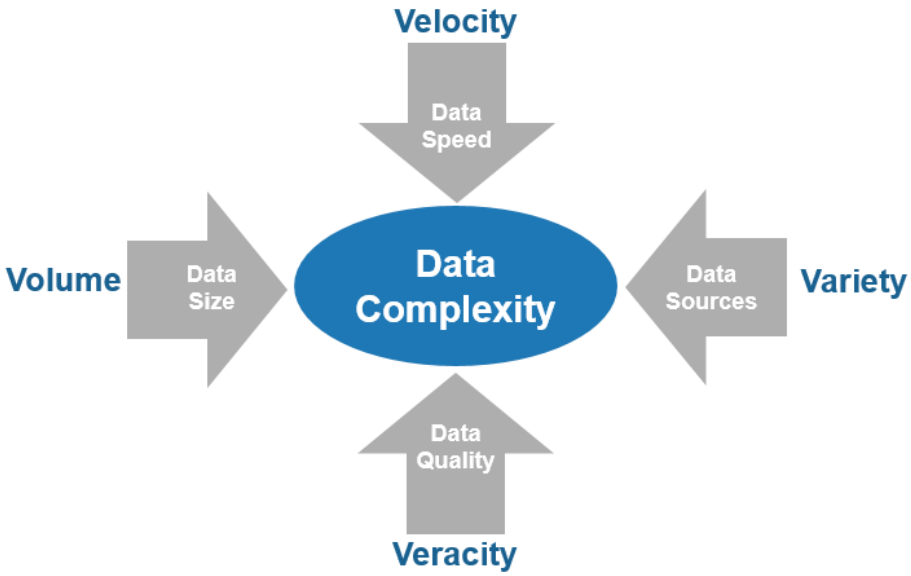

Figure 2. Factors influencing data complexity. The 4 Vs of big data include volume, variety, velocity, and veracity.

Technique Related Challenges

As the volume, variety, velocity, and veracity of data grow, associated task complexities increase as well. The analytical models designed to work on this data need to be nimble. The velocity at which data comes in or changes requires automating where possible and performing continuous real-time analysis. Increasing volumes of data demands scalable solutions.

Furthermore, the analysis outcomes should enable the user to permanently monitor the data on a high level of abstraction. Therefore, the analysis techniques need to be able to glean meaning from the data and visualize or summarize the findings. Only then will domain experts be able to understand fraud patterns and assert countermeasures.

Keep in mind that fraudsters are continually evolving their methods, thus changing the patterns to be detected. Detectors should be aware and dynamically adapt to such transformative patterns and respond to them flexibly, ideally in an automated way.

New and improved algorithms need to be quickly developed, deployed, and integrated into the system in order to keep up with changing fraudster behavior.

Context Related Challenges

Developing the analytics is likely a collaborative activity, with inputs from multiple individuals and teams. If so, care should be lavished on how model development is documented and existing models maintained. Ideally, development tools need to support collaborative workflows.

Resulting models may be shared across groups, perhaps integrated into enterprise-wide production systems, requiring additional flexibility in terms of reporting capabilities and implementation options.

Fraud Detection Workflow

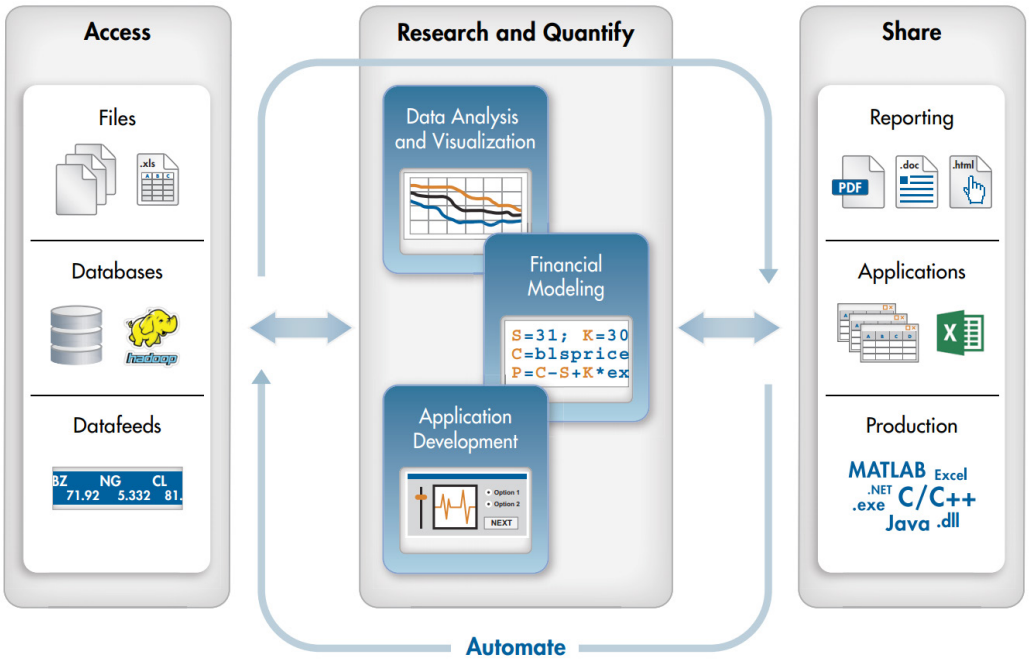

Figure 3. Data analytics workflow, the basis for a sound fraud detection approach.

In a data analytics workflow (Figure 3), the first challenge is accessing available data. Data analysts often access heterogeneous data – numeric, text, instrument-derived signals, and images – from multiple sources. Heterogeneous data is usually unstructured and stored in multiple formats. It needs to be filtered, transformed, and aggregated to be consistent and enable insightful analysis. In addition, faulty data needs to be accounted for and filtered out.

Data analysts research and quantify the problem. A model might be developed based on these data-informed insights and the data analysts' skill and intuition, in combination with appropriate mathematical assumptions. These models are refined and revised repeatedly until their accuracy is deemed sufficient.

Once models are developed, they need to be validated and shared with decision makers to derive insight-oriented value. During the development process, data scientists may collaborate with others in the form of an initial report, to test and challenge experiential assumptions. In addition, a strong yet agile testing process can facilitate assurance. Model cross validation, for example, tests the performance of the model.

It is important to select a set of tools and techniques that will enable rapid iteration, repeatability, maintainability, and reuse through automation to maximize the ROI.

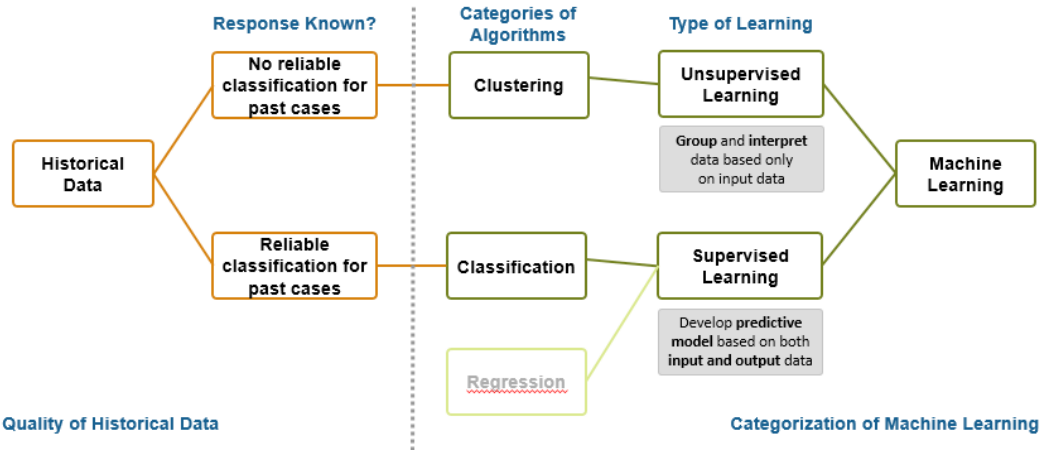

Figure 4. Relationship between available data and machine learning categories.

Due to the data complexity, analysts are attracted by machine learning methods. However, before selecting the techniques to apply, analysts should determine whether historical data contains reliable signals about past cases of fraud (Figure 4). If data samples include cases in which fraud has been identified and verified, analysts can apply supervised learning methods, typically advanced classification algorithms. In such cases, the models can be trained with the historical data in a predictive modeling workflow, increasing the likelihood of detection.

If the data does not include reliable information about past cases of fraud, analysts might turn to unsupervised learning methods such as clustering, techniques that can identify irregularities or anomalies within datasets. This may identify fraudulent cases or behavior that were not previously known to exist in the data.

Once a fraud detection model is finalized, it is applied to new data or out-of-sample data for additional testing and cross validation. If the model has satisfactory performance it is then used with live operating data in production. For the analysis to be useful across the business, however, the models must be integrated into the operational environment, to be accessed and used by end users and decision makers. This may entail integrating the predictive model into a desktop application, reporting platform, or web/database server based solution.

Using MATLAB Based Data Analytics for Fraud Detection

With big data analytics capabilities in MATLAB, analysts can replace manually defined red flags with a data-driven approach. This enables a comprehensive analysis of the data landscape for suspicious patterns and reduces risks of undetected cases, in contrast to a sample based analysis. Detection algorithms can be fully automated without the need for manual interactions or visual checks. Thus, results are available quickly and analysts can react early to emerging issues.

Automatically generated reports can further summarize and visualize analysis outcomes. Thorough reports enable analysts to better understand the underlying patterns of the detected cases of fraud. Instead of applying black-box analysis, analysts gain insight into the data, enabling refinement and adjustment of detection processes.

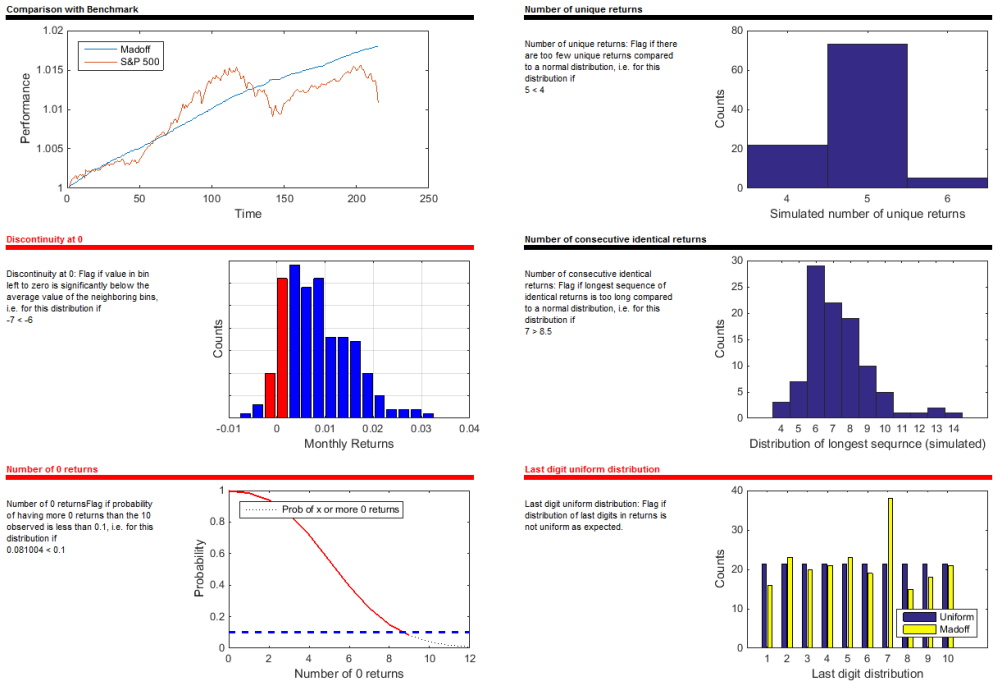

Figure 5. Example automated reporting of key fraud detection metrics. Metrics with red titles include data anomalies, suggesting possible fraudulent behavior.

MATLAB has a reliable, easy-to-use interactive interface. In combination with quality-assured and fully-documented advanced analysis routines, MATLAB enables rapid and flexible prototyping of algorithms. This helps analysts quickly adapt fraud detection analytics to their needs, and proactively respond to the evolution of the techniques of those committing fraud. (For an example, see the article "Systematic Fraud Detection Through Automated Data Analytics in MATLAB" on hedge fund fraud detection.)

Enabling real-time fraud detection by scaling computations in MATLAB to big data on a cluster, the cloud, or on Hadoop is straightforward. Scaling up requires no knowledge of distributed or parallel computing paradigms. Developers write MATLAB code as they would for a serial application, and run it in parallel by adding a few configuration options, instructing MATLAB what resources to use. Real-time processing lets analysts detect fraudulent behavior as it happens, not after the fact.

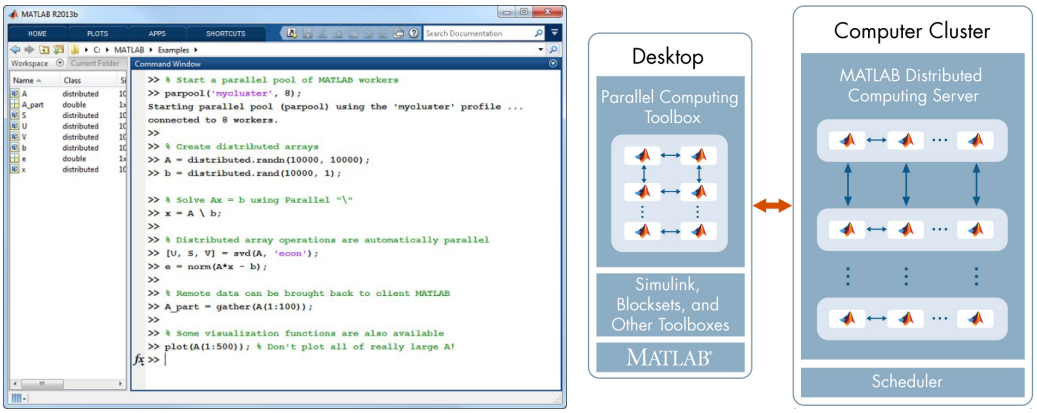

Figure 6. Scaling up computations in MATLAB. Analysts process big data using parallel processing on the desktop for prototyping and then move to a compute cluster without rewriting the model.

MATLAB can also scale out to multiple end systems. This enables developers to create a fraud detection algorithm once, and reuse the same model implementation across databases, Excel spreadsheets, web interfaces, and other applications written in .NET, JAVA, C/C++, or Python. Fraud analytics can be deployed where needed to control increasing amounts of data, users, or requests.

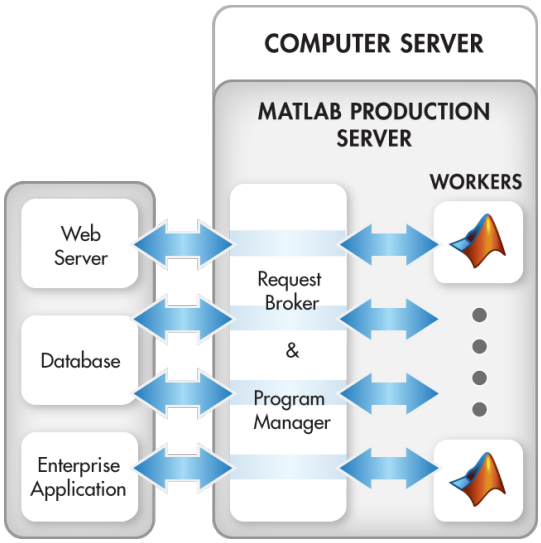

Figure 7. Scaling out access using MATLAB Production Server™. Multiple end systems and users can access multiple analytics managed by the server.

The combination of flexibility and scalability makes MATLAB the standard development environment in the financial services industry. It's used at the top 15 asset management companies, 9 of the top 10 U.S. commercial banks, 12 of the top 15 hedge funds, the reserve banks of all OECD member countries, and the top 3 credit rating agencies.

For fraud detection, risk management, and other core financial services strategies, MATLAB enables organizations to rapidly develop and test advanced solutions and deploy them across the enterprise.

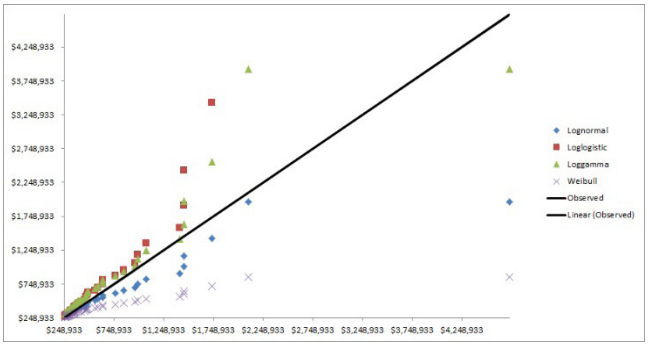

Figure 8. Plot from Change of Measure (COM) operational risk model, developed by Wolters Kluwer. See video and article.

Conclusion

Losses related to fraud are estimated to represent 5% of an organization's revenues. Such losses can be mitigated by implementing an automated data analytics process for detecting irregularities, and reporting them in a timely fashion. But implementing such a system presents challenges in managing data, algorithms, collaboration, and reporting.

MATLAB provides an ideal interactive development environment that enables users to move ideas into production in a timely and cost effective manner. MATLAB can be used from end to end: performing research, designing algorithms, and testing and implementing in an automated analytics system. Analysts can turn the latest research idea into a tested algorithm to run in production in a matter of hours, not days. This enables the development and adaptation of fraud detection algorithms and approaches in a fraction of the time it would take using traditional programming languages, closed platforms, database systems, or spreadsheets.

Appendix

Training

MATLAB is intuitive and easy to learn. However, teams may be able to exploit its full potential and shrink the learning curve with MathWorks professional training, including:

- Public training (available worldwide)
- Onsite training with standard or customized courses
- Online instructor-led live training courses
- Self-paced interactive online training

MathWorks Training Services provides more than 30 course offerings. Specialized courses in financial analysis, parallel computing, code generation, platform deployment, and other areas are available, along with introductory and intermediate training on MATLAB.

Consulting

In addition to training, teams can also engage a global team of experts to support adoption and integration of MATLAB based solutions within an organization, based on the team's specific needs and requirements.