Contents

[Top 5 Predictive Analytics Models and Algorithms 1](#_Toc63678591)

[Top 5 Predictive Analytics Models 1](#_Toc63678592)

[Classification Model 1](#_Toc63678593)

[Clustering Model 2](#_Toc63678594)

[Forecast Model 2](#_Toc63678595)

[Outliers Model 2](#_Toc63678596)

[Time Series Model 2](#_Toc63678597)

[Common Predictive Algorithms 3](#_Toc63678598)

[Random Forest 3](#_Toc63678599)

[Generalized Linear Model (GLM) for Two Values 4](#_Toc63678600)

[Gradient Boosted Model (GBM) 4](#_Toc63678601)

[K-Means 5](#_Toc63678602)

[Prophet 5](#_Toc63678603)

[Summary 6](#_Toc63678604)

[Top 10 Areas Where Data Analysis Adds the Most Value 6](#_Toc63678605)

[1. Acounts Payable (AP) / Accounts Receivable (AR) 6](#_Toc63678606)

[2. Duplicate Detection 7](#_Toc63678607)

[3. Sampling 7](#_Toc63678608)

[4. Data Imports/Extractions & Analysis of Large Data Sets 7](#_Toc63678609)

[5. Continuous Auditing & Monitoring 7](#_Toc63678610)

[6. Fraud Detection & Forensic Auditing 8](#_Toc63678611)

[7. Analysis of P-Cards 8](#_Toc63678612)

[8. Payroll & Time Sheets 8](#_Toc63678613)

[9. Joins & Comparisons 8](#_Toc63678614)

[10. Inventory Audits 8](#_Toc63678615)

[Honorable Mentions from Survey Findings 9](#_Toc63678616)

[Introduction to Data Analysis for Auditors and Accountants 10](#_Toc63678617)

[EXHIBIT 1 Applying Analytics Step by Step 10](#_Toc63678618)

[The Steps in the Process 10](#_Toc63678619)

[***Flowcharting the process.*** 10](#_Toc63678620)

[EXHIBIT 2 Sample Flowchart of an Insurance Claim Process 11](#_Toc63678621)

[Choosing and extracting the data. 11](#_Toc63678622)

[Understanding the population. 11](#_Toc63678623)

[Understanding the fields with descriptive statistics. 11](#_Toc63678624)

[Exploratory data analysis. 11](#_Toc63678625)

[Choice of analytic methods and alternative approaches. 11](#_Toc63678626)

[EXHIBIT 3 Analytic Methods and Tools 12](#_Toc63678627)

[Confirmatory data analysis and finding outliers. 12](#_Toc63678628)

[Evaluating results evaluation and integrating with traditional findings. 12](#_Toc63678629)

[EXHIBIT 4 Audit by Exception (ABE) Approach 13](#_Toc63678630)

[EXHIBIT 5 Outliers Identified to Be Examined 13](#_Toc63678631)

[Emerging Approaches 14](#_Toc63678632)

[Tools and Information Sources 14](#_Toc63678633)

[A Growing Phenomenon 15](#_Toc63678634)

[FOR FURTHER WATCHING 15](#_Toc63678635)

[Three Ways to Enhance Audit Quality with Audit Data Analytics 15](#_Toc63678636)

[Why Apply Audit Data Analytics? 15](#_Toc63678637)

[Define Audit Data Analytics 16](#_Toc63678638)

[Audit Data Analytics Examples 16](#_Toc63678639)

[Call to action 17](#_Toc63678640)

# Top 5 Predictive Analytics Models and Algorithms

BY SRIRAM PARTHASARATHY

Predictive analytics tools are powered by several different models and algorithms that can be applied to wide range of use cases. Determining what predictive modeling techniques are best for your company is key to getting the most out of a [predictive analytics solution](https://www.logianalytics.com/predictive-analytics/the-4-common-challenges-of-predictive-analytics/) and leveraging data to make insightful decisions. For example, consider a retailer looking to reduce customer churn. They might not be served by the same predictive analytics models used by a hospital predicting the volume of patients admitted to the emergency room in the next ten days. What are the most common predictive analytics models? And what predictive algorithms are most helpful to fuel them? In this post, we give an overview of the most popular types of predictive models and algorithms that are being used to solve business problems today.

## Top 5 Predictive Analytics Models

### Classification Model

The classification model is, in some ways, the simplest of the several types of predictive analytics models we’re going to cover. It puts data in categories based on what it learns from historical data.

Classification models are best to answer yes or no questions, providing broad analysis that’s helpful for guiding decisive action. These models can answer questions such as:

* For a retailer, “Is this customer about to churn?”
* For a loan provider, “Will this loan be approved?” or “Is this applicant likely to default?”
* For an online banking provider, “Is this a fraudulent transaction?”

The breadth of possibilities with the classification model—and the ease by which it can be retrained with new data—means it can be applied to many different industries.

### Clustering Model

The clustering model sorts data into separate, nested smart groups based on similar attributes. If an ecommerce shoe company is looking to implement targeted marketing campaigns for their customers, they could go through the hundreds of thousands of records to create a tailored strategy for each individual. But is this the most efficient use of time? Probably not. Using the clustering model, they can quickly separate customers into similar groups based on common characteristics and devise strategies for each group at a larger scale.

Other use cases of this predictive modeling technique might include grouping loan applicants into “smart buckets” based on loan attributes, identifying areas in a city with a high volume of crime, and benchmarking SaaS customer data into groups to identify global patterns of use.

### Forecast Model

One of the most widely used [predictive analytics models](https://www.logianalytics.com/predictive-analytics/what-is-predictive-analytics/), the forecast model deals in metric value prediction, estimating numeric value for new data based on learnings from historical data.

This model can be applied wherever historical numerical data is available. Scenarios include:

* A SaaS company can estimate how many customers they are likely to convert within a given week.
* A call center can predict how many support calls they will receive per hour.
* A shoe store can calculate how much inventory they should keep on hand in order to meet demand during a particular sales period.

The forecast model also considers multiple input parameters. If a restaurant owner wants to predict the number of customers she is likely to receive in the following week, the model will take into account factors that could impact this, such as: Is there an event close by? What is the weather forecast? Is there an illness going around?

### Outliers Model

The outliers model is oriented around anomalous data entries within a dataset. It can identify anomalous figures either by themselves or in conjunction with other numbers and categories.

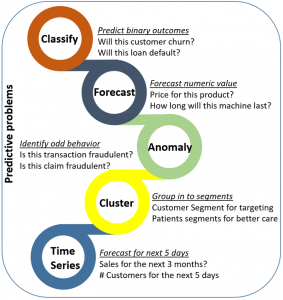
* Recording a spike in support calls, which could indicate a product failure that might lead to a recall
* Finding anomalous data within transactions, or in insurance claims, to identify fraud
* Finding unusual information in your NetOps logs and noticing the signs of impending unplanned downtime

The outlier model is particularly useful for predictive analytics in retail and finance. For example, when identifying fraudulent transactions, the model can assess not only amount, but also location, time, purchase history and the nature of a purchase (i.e., a $1000 purchase on electronics is not as likely to be fraudulent as a purchase of the same amount on books or common utilities).

### Time Series Model

The [time series model](https://www.logianalytics.com/logi-news/logi-predict-product-update-key-enhancements/) comprises a sequence of data points captured, using time as the input parameter. It uses the last year of data to develop a numerical metric and predicts the next three to six weeks of data using that metric. Use cases for this model includes the number of daily calls received in the past three months, sales for the past 20 quarters, or the number of patients who showed up at a given hospital in the past six weeks. It is a potent means of understanding the way a singular metric is developing over time with a level of accuracy beyond simple averages. It also takes into account seasons of the year or events that could impact the metric.

If the owner of a salon wishes to predict how many people are likely to visit his business, he might turn to the crude method of averaging the total number of visitors over the past 90 days. However, growth is not always static or linear, and the time series model can better model exponential growth and better align the model to a company’s trend. It can also forecast for multiple projects or multiple regions at the same time instead of just one at a time.

[](https://www.logianalytics.com/wp-content/uploads/2019/07/Image-1.png)

## Common Predictive Algorithms

Overall, predictive analytics algorithms can be separated into two groups: machine learning and deep learning.

* **Machine learning** involves structural data that we see in a table. Algorithms for this comprise both linear and nonlinear varieties. Linear algorithms train more quickly, while nonlinear are better optimized for the problems they are likely to face (which are often nonlinear).
* **Deep learning** is a subset of machine learning that is more popular to deal with audio, video, text, and images.

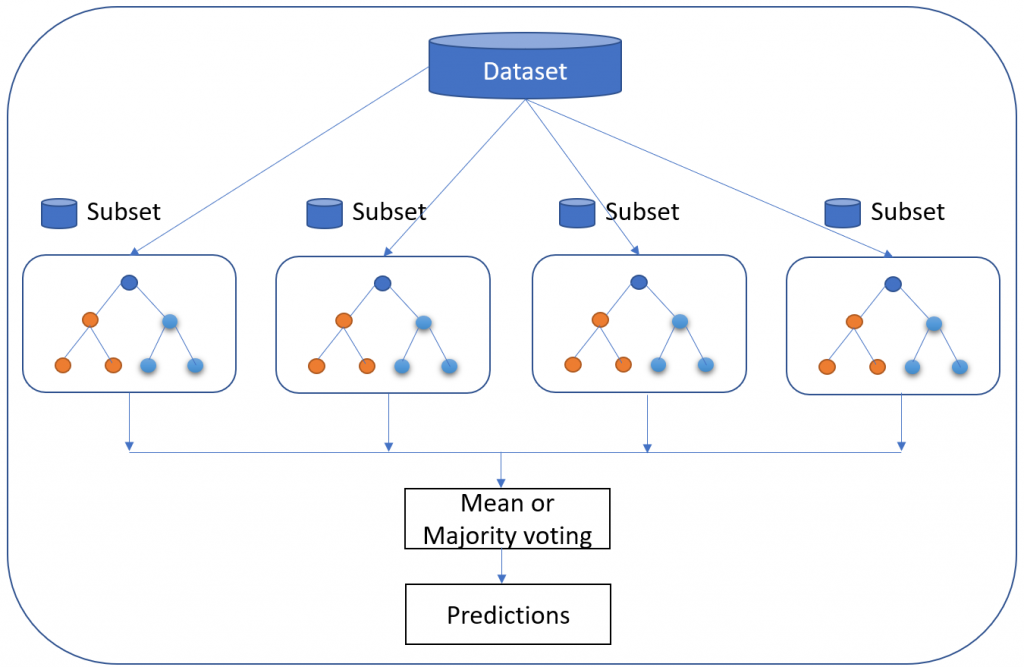
With machine learning predictive modeling, there are several different algorithms that can be applied. Below are some of the most common algorithms that are being used to power the predictive analytics models described above.

### Random Forest

Random Forest is perhaps the most popular classification algorithm, capable of both classification and regression. It can accurately classify large volumes of data.

The name “Random Forest” is derived from the fact that the algorithm is a combination of decision trees. Each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the “forest.” Each one is grown to the largest extent possible.

Predictive analytics algorithms try to achieve the lowest error possible by either using “boosting” (a technique which adjusts the weight of an observation based on the last classification) or “bagging” (which creates subsets of data from training samples, chosen randomly with replacement). Random Forest uses bagging. If you have a lot of sample data, instead of training with all of them, you can take a subset and train on that, and take another subset and train on that (overlap is allowed). All of this can be done in parallel. Multiple samples are taken from your data to create an average.

[](https://www.logianalytics.com/wp-content/uploads/2019/07/Image-2.png)

While individual trees might be “weak learners,” the principle of Random Forest is that together they can comprise a single “strong learner.”

The popularity of the Random Forest model is explained by its various advantages:

* Accurate and efficient when running on large databases
* Multiple trees reduce the variance and bias of a smaller set or single tree
* Resistant to overfitting
* Can handle thousands of input variables without variable deletion
* Can estimate what variables are important in classification
* Provides effective methods for estimating missing data
* Maintains accuracy when a large proportion of the data is missing

### Generalized Linear Model (GLM) for Two Values

The Generalized Linear Model (GLM) is a more complex variant of the General Linear Model. It takes the latter model’s comparison of the effects of multiple variables on continuous variables before drawing from an array of different distributions to find the “best fit” model.

Let’s say you are interested in learning customer purchase behavior for winter coats. A regular linear regression might reveal that for every negative degree difference in temperature, an additional 300 winter coats are purchased. While it seems logical that another 2,100 coats might be sold if the temperature goes from 9 degrees to 3, it seems less logical that if it goes down to -20, we’ll see the number increase to the exact same degree.

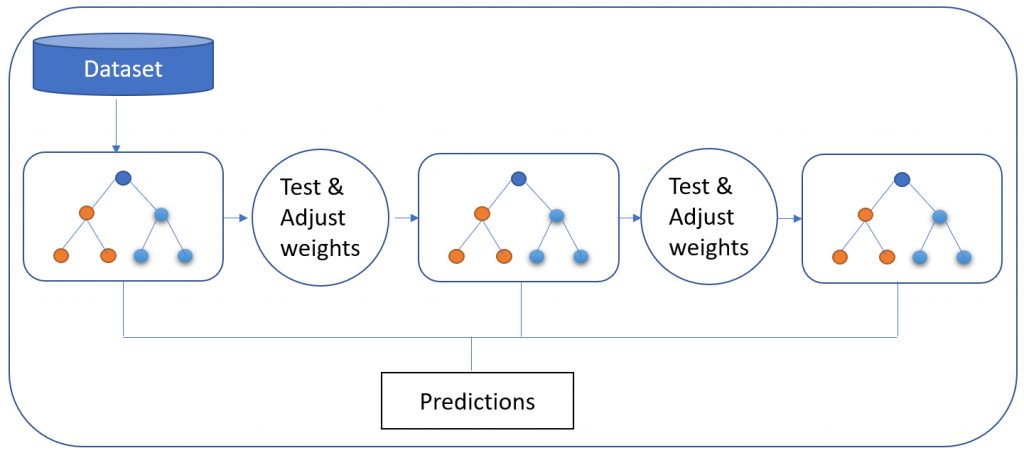
The Generalized Linear Model would narrow down the list of variables, likely suggesting that there is an increase in sales beyond a certain temperature and a decrease or flattening in sales once another temperature is reached.

The advantage of this algorithm is that it trains very quickly. The response variable can have any form of exponential distribution type. The Generalized Linear Model is also able to deal with categorical predictors, while being relatively straightforward to interpret. On top of this, it provides a clear understanding of how each of the predictors is influencing the outcome, and is fairly resistant to overfitting. However, it requires relatively large data sets and is susceptible to outliers

### Gradient Boosted Model (GBM)

The Gradient Boosted Model produces a prediction model composed of an ensemble of decision trees (each one of them a “weak learner,” as was the case with Random Forest), before generalizing. As its name suggests, it uses the “boosted” machine learning technique, as opposed to the bagging used by Random Forest. It is used for the classification model.

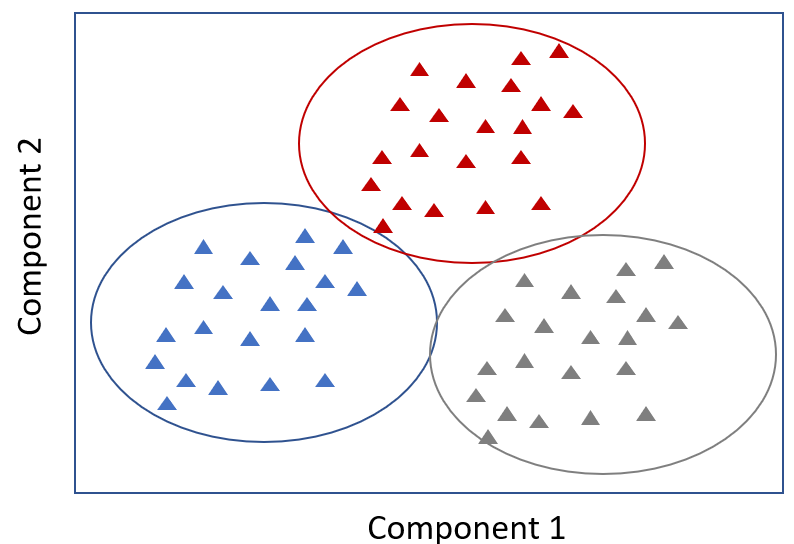
The distinguishing characteristic of the GBM is that it builds its trees one tree at a time. Each new tree helps to correct errors made by the previously trained tree⁠—unlike in the Random Forest model, in which the trees bear no relation. It is very often used in machine-learned ranking, as in the search engines Yahoo and Yandex.

[](https://www.logianalytics.com/wp-content/uploads/2019/07/Image-3.png)

Via the GBM approach, data is more expressive, and benchmarked results show that the GBM method is preferable in terms of the overall thoroughness of the data. However, as it builds each tree sequentially, it also takes longer. That said, its slower performance is considered to lead to better generalization.

### K-Means

A highly popular, high-speed algorithm, K-means involves placing unlabeled data points in separate groups based on similarities. This algorithm is used for the clustering model. For example, Tom and Rebecca are in group one and John and Henry are in group two. Tom and Rebecca have very similar characteristics but Rebecca and John have very different characteristics. K-means tries to figure out what the common characteristics are for individuals and groups them together. This is particularly helpful when you have a large data set and are looking to implement a personalized plan—this is very difficult to do with one million people.

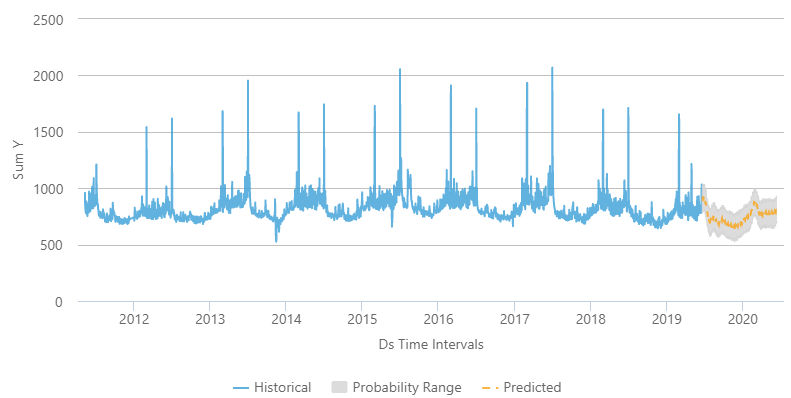
[](https://www.logianalytics.com/wp-content/uploads/2019/07/Image-4.png)

In the context of [predictive analytics for healthcare](https://www.logianalytics.com/predictive-analytics/predictive-analytics-for-healthcare-predicting-missed-appointments/), a sample size of patients might be placed into five separate clusters by the algorithm. One particular group shares multiple characteristics: they don’t exercise, they have an increasing hospital attendance record (three times one year and then ten times the next year), and they are all at risk for diabetes. Based on the similarities, we can proactively recommend a diet and exercise plan for this group.

### Prophet

The Prophet algorithm is used in the time series and forecast models. It is an open-source algorithm developed by Facebook, used internally by the company for forecasting.

The Prophet algorithm is of great use in capacity planning, such as allocating resources and setting sales goals. Owing to the inconsistent level of performance of fully automated forecasting algorithms, and their inflexibility, successfully automating this process has been difficult. On the other hand, manual forecasting requires hours of labor by highly experienced analysts.

[](https://www.logianalytics.com/wp-content/uploads/2019/07/image-5.png)

Prophet isn’t just automatic; it’s also flexible enough to incorporate heuristics and useful assumptions. The algorithm’s speed, reliability and robustness when dealing with messy data have made it a popular alternative algorithm choice for the time series and forecasting analytics models. Both expert analysts and those less experienced with forecasting find it valuable.

## Summary

How do you determine which predictive analytics model is best for your needs? You need to start by identifying [what predictive questions you are looking to answer](https://www.logianalytics.com/predictive-analytics/are-you-asking-the-right-predictive-questions/), and more importantly, what you are looking to do with that information. Consider the strengths of each model, as well as how each of them can be optimized with different predictive analytics algorithms, to decide how to best use them for your organization.

# Top 10 Areas Where Data Analysis Adds the Most Value

Survey Results from Audit & Accounting Professionals. We surveyed our extensive database of accounting, auditing and financial professionals to find out what [data analysis](https://www.audimation.com/data-analysis) project or technique has brought them the most value. Results:

## 1. Acounts Payable (AP) / Accounts Receivable (AR)

Given the sheer amount of data embedded within Accounts Payable (AP) and Accounts Receivable (AR), data analysis is adding the most value in these areas. Many noted they are reviewing data within the vendor master file, while others mentioned these specific areas:

* Purchasing to payment analysis
* Review of operating disbursements, with primary focus on AP
* Validation of payments to vendors
* Analyzing AR open order data for system deficiencies within processes
* Evaluating AR aging files for internal and external financial reporting
* AR testing and searching for duplicate payments

**Notable Quote**

“Currently we have been able to add value from a fraud perspective by performing audits of AP vendor master maintenance and AP transactions. We have also added value via our journal entry analysis.”

## 2. Duplicate Detection

If you’re looking for recovery dollars, fraud, errors, anomalies or outliers, duplicate detection is a proven technique used to find them. Survey respondents indicated value in use of testing, analysis and review of expenses. Common areas for duplicate detection searches include payments, invoices, transactions and vendor returns for credit.

## 3. Sampling

Sampling is a fundamental part of audit work, and with so many ways to sample, this frontrunner is no surprise. Multiple survey respondents indicated they use IDEA’s statistical sampling methods for their sampling work to set the scope, providing defensible and valuable insights when results are extrapolated against the population. One internal auditor noted they use the statistical functions, summarizations, and stratifications to allow management to analyze data to make projections and evaluate past performance.

Some survey respondents provided details, such as using stratified sampling to examine gains and losses on backdated trades, stratify invoices and disbursements, and perform random samples. Data analysis is also used to sample and perform keyword searches. One of the many benefits data analysis technology offers is the ability to analyze an entire population of data, as opposed to sampling. However, data analysis also allows you to join, sort and summarize areas of data to analyze smaller data sets.

**Notable Quotes**

“We use IDEA to help select sample sizes and perform audit tasks, which has greatly reduced audit time and manpower requirements.” “The ability to sort through large data populations of GL data has helped us greatly. Being able to combine databases has also been a help since it is hard to have IT create one database at times.”

## 4. Data Imports/Extractions & Analysis of Large Data Sets

Most auditors or accounting professionals will tell you their biggest challenges are knowing where to apply data analysis, getting the data they need and importing it into an analysis tool. Data analysis software developers have made great strides in developing new technology to simplify data imports, such as the addition of PDF converters, drag-and-drop capabilities, and the expansion of data capacity limits. Survey respondents indicated one of the greatest values data analysis offers them is the ability to convert PDF files, text data, and information stored in spreadsheets into an analysis tool. Many responded they were using data analysis to perform data extractions, which would be nearly impossible for them otherwise, and extracting data based on certain criteria such as by name or customer codes. One individual stated their overall use of IDEA and ODBC instead of MS query and SQL to download, analyze and summarize data had saved them hundreds of hours. Those surveyed said they value the ability to import and handle large amounts of data to be analyzed for different criteria, and setting up different worksheets within a file to perform various audit steps. Auditors are using data analysis to perform big downloads of financial transactions during the audit period, combined with summaries and analysis, to help scope and test during the audit. An internal auditor within the health care industry stated they are using data analysis to obtain entire payment data from an ERP system to conduct analysis on areas such as payment dates on weekends, and search for duplicates and vendor classification coding errors.

**Notable Quotes**

“Being able to import various types of reports (text, PDF, Excel, etc.) into the software has been most beneficial when attempting to compare data from various repositories. This process has been used to compare active employee records against payroll registers and building access (security) reports.”

“The most useful aspect of IDEA is the amount of information the program can hold. We do many audits that involve server data, which can be quite large. Without IDEA, we would not be able to analyze this data.”

## 5. Continuous Auditing & Monitoring

Data analysis continues to be used to automate manual processes and testing by implementing continuous auditing. Several internal auditors responded they value the ability to develop continuous audit scripts to look for errors in account entry and the use of continuous monitoring for travel and entertainment (T&E), financial statements, p-cards, human resources, and accounts payable.

While continuous monitoring can provide near real-time information to identify control breaches, several survey responses indicated they were using continuous monitoring monthly or quarterly. One respondent noted they use continuous auditing in their audit department to monitor credit data, market feeds, credit and financial information to track and flag significant changes. Another stated they use continuous monitoring for Foreign Corrupt Practices Act (FCPA) compliance.

**Notable Quote**

“Continuous monitoring of store transactions where we score each unit and research the outliers.”

## 6. Fraud Detection & Forensic Auditing

While you may stumble upon fraud by searching for outliers or anomalies, many auditors and accountants are using [data analytics](https://www.audimation.com/data-analysis/) technology as a forensic tool to search for fraud and schemes. Data analysis tools allow you to look at the data from different angles to get to the root cause of fraud, which is what many of our survey respondents found most valuable. Some of the ways data analysis is being used to search for fraud include trend and pattern analysis to look for indications of diversion of funds or theft, behavioral analysis and monitoring of spending trends.

## 7. Analysis of P-Cards

Procurement or purchasing cards (p-cards) have simplified the amount of administrative work organizations handle to process small purchases, but also increased risks. While p-card abuse has made headlines, particularly in government departments, they can be managed by implementing proper internal controls to reduce misuse and waste. It’s naturally an area accountants and auditors must analyze to search for evidence of red flags, review transactions and track trends.

**Notable Quote**

“The ability to script IDEA tests has enabled our company to automatically identify potentially anomalous procurement card transactions with maximum efficiency.”

## 8. Payroll & Time Sheets

Hunting for ghost employees, falsifying wage claims and tampering with timesheets are all key areas where data analysis is adding value, according to those surveyed. One respondent indicated they conduct “payroll weekly dollar or total reasonableness testing,” while another stated they use IDEA to perform 30+ analysis procedures on a quarterly basis covering purchase-to-pay, revenue, journal entries and payroll. Data analysis is also bringing value by enabling review of electronic time entry records for compliance with existing policies, procedures and labor regulations.

**Notable Quote**

“Through the use of IDEA our payroll department has reduced the time required to review payroll prior to their weekly transmission to ADP from several hours to less than 30 minutes.”

## 9. Joins & Comparisons

Few will argue that one of the biggest values data analysis brings is the ability to join databases and files to sort, summarize and analyze data – again, enabling users to look at the data from different angles to search for outliers, fraud, errors and other information. Those surveyed provided some applications of using joins to check for anomalies, create inventory turnover or perform open to sell analysis, and extract data from PDFs, then join it together for analysis. One survey participant stated they join databases to save hours reconciling outstanding checklists and another stated they benefit from joining large databases to replace a manual comparison process. Areas were data analysis is being used to compare sets of data include:

* Active application user accounts against a list of terminated employees
* Physician scheduling system to the billing system
* Vendor terms including days paid early and aged receivables
* Missing invoice numbers
* Employee/HR to vendor address
* Tax ID numbers and conflict of interest
* Review of billing and pricing agreements against contract agreements

**Notable Quote**

“Comparing the HR files to the AP vendor files provided information leading to employee fraud of valuable assets and other services. This technique provided the link needed to connect the employee’s address to the AP vendor delivery address, thus giving the needed data for further investigation.”

## 10. Inventory Audits

Given the massive size of some inventories, data analysis is being used to conduct inventory audits. It is being used to identify potential or obsolete inventory, slow-moving inventory, reconciling the inventory count with the general ledger, and performing sampling, valuations, and testing on inventory data. One person surveyed noted they use data analysis as a business intelligence tool to better understand both long- and short-term inventory risk, and another indicated they analyze (oil and gas) well, property and cost center data.

Honorable Mentions from Survey Findings

While reviewing some of the survey comments, we found a few creative, or somewhat unexpected uses of data analysis we thought were worth mentioning:

Centralized Accounting & Data Mining

“Centralized accounting and related services for processing company transactions. There is more volume in terms of transactions, but because it is done via centralized accounting, the data is easier to obtain and analyze.”

Compliance

Several survey respondents noted they find data analysis valuable in meeting industry and documentation standards. One individual stated their internal audit department uses data analysis for planning, executing and document SOX procedures as well as audits of other areas and processes.

Data Warehousing

“A growing number of data warehousing efforts with key data being extracted from legacy systems spread across the enterprise from various product lines. A long-overdue process that is showing good results for data analysis on many fronts.”

Population Testing

“We try to incorporate data analysis in every audit we do. The biggest win for us has been performing a lot more population testing, instead of doing sampling. It takes the same amount of time for us, but provides much more clarity and value to the client.”

Summarization

“Summarization of data in IDEA (tool) that couldn’t otherwise be summarized by company personnel. This gave us confidence that a process was working as expected.”

Teaching Tool for College & Universities

IDEA is taught by more than 160+ colleges and universities throughout the U.S. to help educators bring data analysis techniques and technology into the classroom. One professor stated: “I teach fraud examination and forensic accounting at the university, so IDEA fraud analytics has been of the most value for exposing my students to data analysis capabilities in fraud prevention and detection.”

Reconciliation

“The most important project involving IDEA at this organization has been reconciling government data to the organization’s mainframe. This is a requirement of the government contract.”

Revenue Reporting

Many of those surveyed stated they use data analysis for reporting purposes including the distribution of company-wide reports used to analyze financial results. “IDEA has enabled our company to analyze data for balancing and reporting purposes more quickly and efficiently. The program builds on techniques that other software utilizes.” Another survey respondent noted, “Revenue recognition with IDEA that was not feasible to be performed by any other means.”

Creative Uses:

 Hourly energy bidding analysis

* Analyzing customer behavior around service renewal periods
* Analyzing customer surveys
* Measuring the effectiveness of a promotional program
* Analysis of financial and student enrollment data
* Direct billing review project
* FCPA monitoring
* Evaluation of fixed asset balances
* Integration of student information system with online learning management systems
* Risk assessment
* Medical bad debt projects
* Testing information in various system logs
* Banking – Analysis of where wires were being split to avoid limits
* Loan grade and portfolio trending

Our survey of nearly 600 accountants, auditors and financial professionals working in various industries provided us with some interesting insights about how data analysis was being used, and which areas they believe were most valuable to their respective organizations. While data analysis has been around for more than 25 years, it continues to evolve in its technological advancements, use and significance. As one of our survey respondents noted, “There’s just an immense amount of ways the data can be analyzed.” Using data analysis technology to help you perform better, more in-depth, analysis in less time with minimal effort is not only effective but worthwhile.

*Survey conducted by Audimation Services, Inc. and analyzed by a third-party marketing consultant.*

# Introduction to Data Analysis for Auditors and Accountants

By  [Alexander Kogan, PhD](https://www.cpajournal.com/author/alexander-kogan-phd/), [Miklos A. Vasarhelyi, PhD](https://www.cpajournal.com/author/miklos-a-vasarhelyi-phd/) and [Deniz Appelbaum, PhD](https://www.cpajournal.com/author/denizappelbaum/)

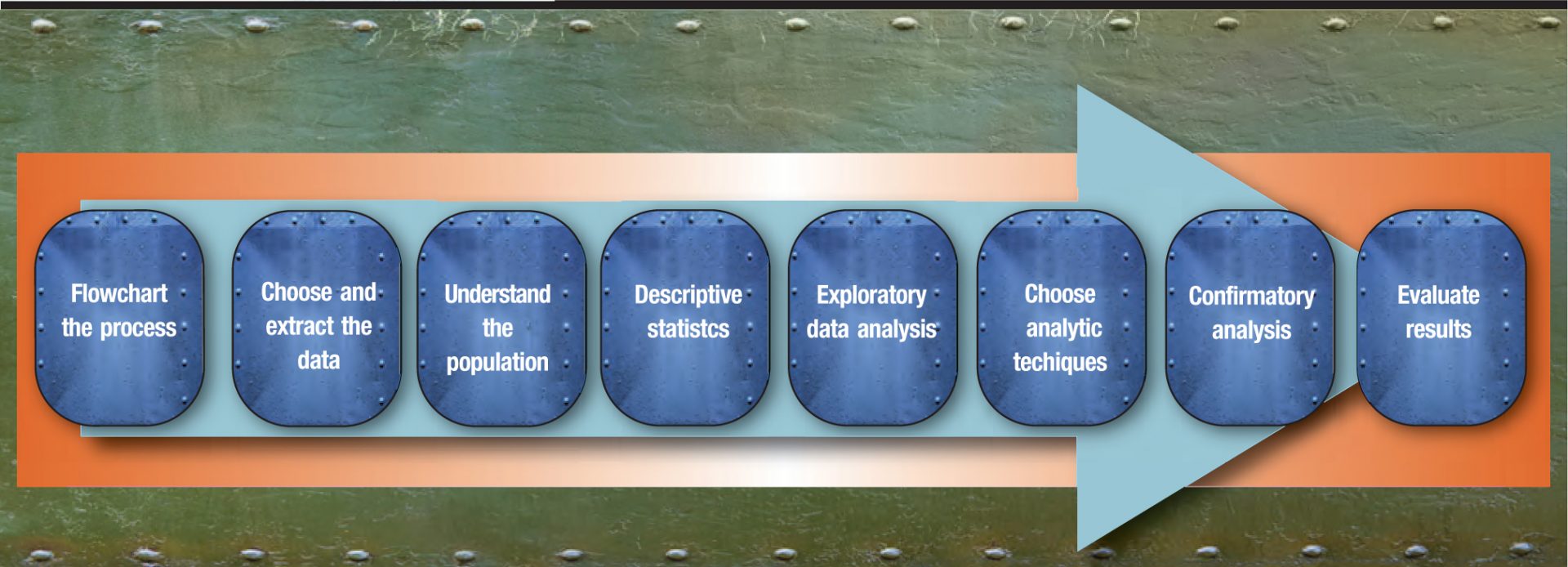
**In Brief**

The audit world is changing. Technology has transformed business processes and created a wealth of data that can be leveraged by accountants and auditors with the requisite mindset. Data analysis can enable auditors to focus on outliers and exceptions, identifying the riskiest areas of the audit. The authors introduce the process, with a review of some emerging approaches and compilation of useful resources for auditors new to the topic.

\* \* \*

The advent of inexpensive computational power and storage, as well as the progressive computerization of organizational systems, is creating a new environment in which accountants and auditors must practice. This article aims at introducing basic data analysis concepts to enable accounting professionals to understand how to navigate within this new environment. Specifically, the focus will not be on auditing and accounting standards and their current required procedures, but rather on what the profession can progressively achieve with data analytics. Most analytical procedures, in the right circumstances, may be applicable to the entire audit process, from risk assessment to test of details. What follows is a step-by-step overview ([***Exhibit 1***](https://www.nysscpa.org/news/publications/the-cpa-journal/article-detail?ArticleID=12261#F1)) of best practices for the process of applying analytics, with an emphasis on audit by exception (ABE).

### EXHIBIT 1 Applying Analytics Step by Step

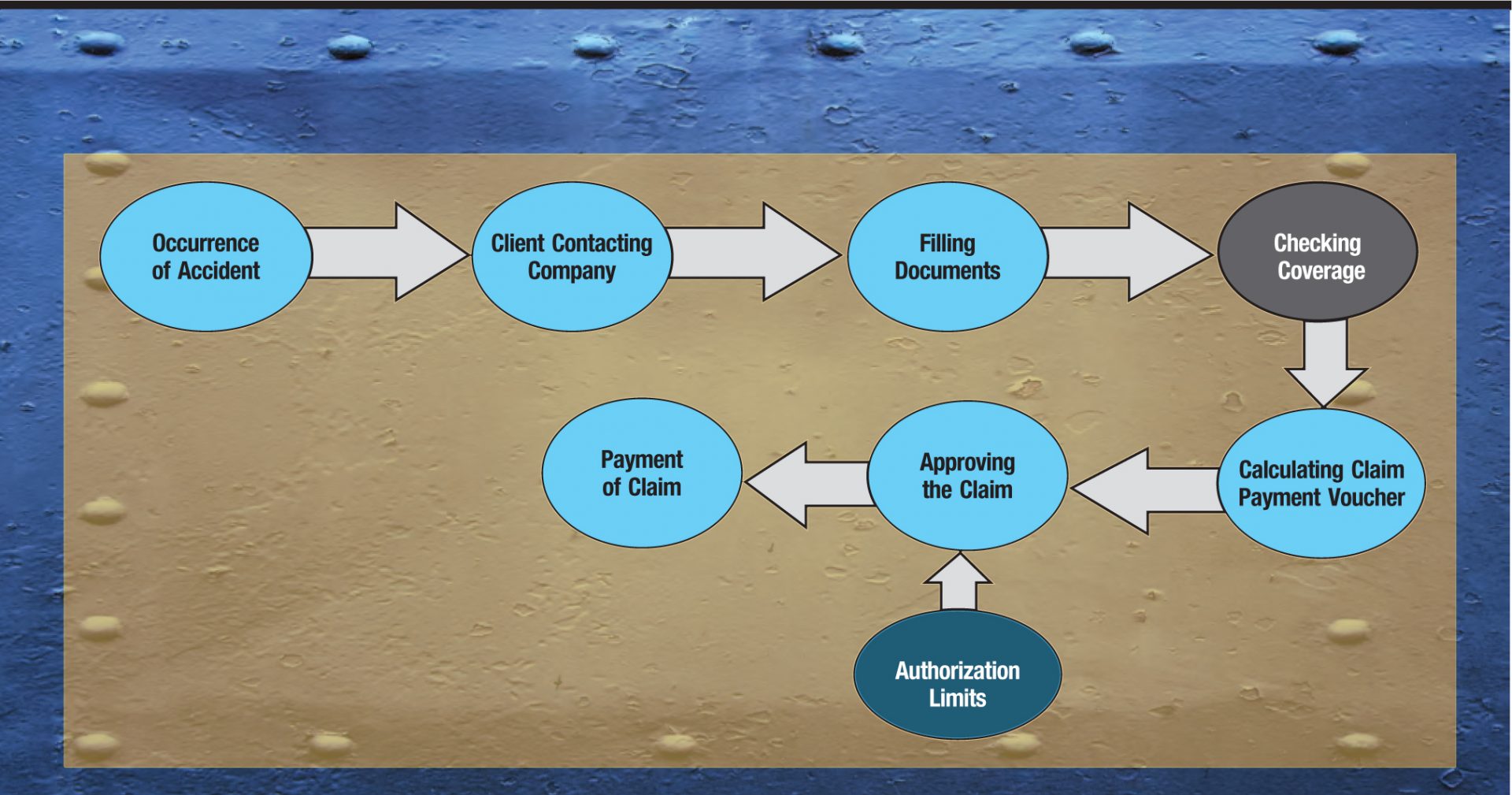
**[](https://www.cpajournal.com/wp-content/uploads/2017/02/CPA.2017.87.2.032.f001.jpg)**

## The Steps in the Process

***Flowcharting the process.***

Understanding the elements of a certain cycle or application is essential for selecting data and understanding risk. Many tools are available for flowcharting, such as Tableau Public, QlickSense, and RapidMiner, all of which are free. Flowcharting is also possible in Microsoft Excel or PowerPoint. [***Exhibit 2***](https://www.nysscpa.org/news/publications/the-cpa-journal/article-detail?ArticleID=12261#F2) shows a sample flowcharting process taken from an insurance company.

### EXHIBIT 2 Sample Flowchart of an Insurance Claim Process

**[](https://www.cpajournal.com/wp-content/uploads/2017/02/CPA.2017.87.2.032.f002.jpg)**

## Choosing and extracting the data.

With the risks in mind, the next step is to choose the data fields to be extracted and examined. This type of analysis is not very different from what would be done on a traditional audit. A progressively increasing number of audit apps are being sold or shared that can serve to simplify the audit task (e.g., [**http://www.capterra.com/audit-software/**](http://www.capterra.com/audit-software/)). Unfortunately, providers have not yet standardized around the AICPA’s Audit Data Standards (ADS) or any other common standard. Nevertheless, many audit software providers (e.g., ACL and CaseWare) have extensive libraries of scripts that can be adapted to various data formats, as well as extraction software that allows for access to traditional data and enterprise resource planning (ERP) systems (e.g., SAP and Oracle).

### Understanding the population.

It is very important for the sake of completeness to understand the nature, distribution, and limitations of the population to be tested. Understanding the scope and limitations of the data is imperative, as it enables an accountant to choose the most appropriate and effective analytical technique.

### Understanding the fields with descriptive statistics.

The examination of key fields for their characteristics and statistical parameters (e.g., maximum, minimum, median, variance) and data availability (e.g., missing values) is probably the most important initial task, but one that is often underappreciated or even neglected.

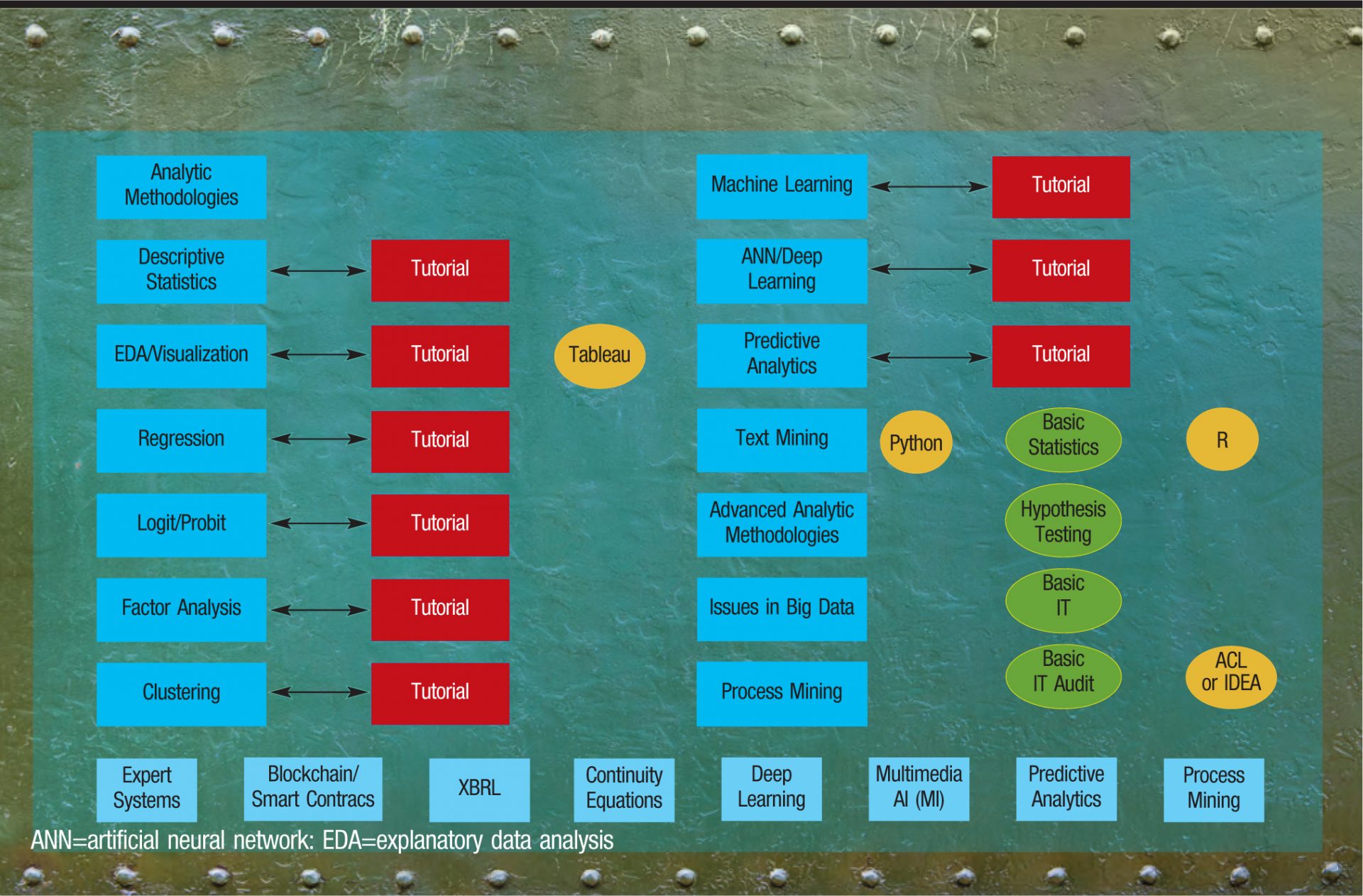
### Exploratory data analysis.

Modern tools of visualization (e.g., Tableau or Excel) allow for data exploration that helps auditors carefully choose where to place their analytic efforts and which assertions to test. Auditors can focus more extensive testing on the areas highlighted as highest risk.

### Choice of analytic methods and alternative approaches.

A great number of analytic methods have been applied to audits in a research mode (Deniz Appelbaum, Alexander Kogan, and Miklos Vasarhelyi, *Analytics for External Auditing: A Literature Review*, Rutgers CARLab, 2016) and are being progressively adopted by CPA firms. [***Exhibit 3***](https://www.nysscpa.org/news/publications/the-cpa-journal/article-detail?ArticleID=12261#F3) provides examples of several analytic methods. Given this variety of choices, auditors need to know the data as intimately as possible, as well as understand the specific analytic task, in order to reduce the pool of potential analytical methods.

### EXHIBIT 3 Analytic Methods and Tools

**[](https://www.cpajournal.com/wp-content/uploads/2017/02/CPA.2017.87.2.032.f003.jpg)**

*Any significant deviations should be investigated by auditors.*

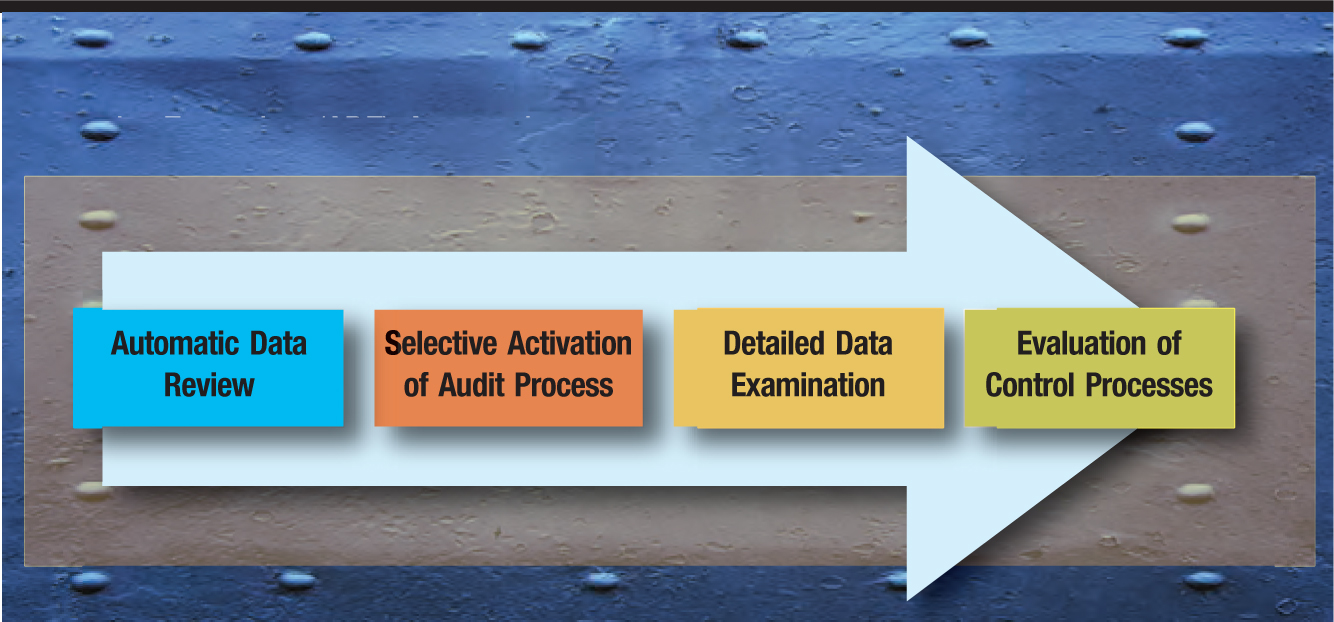
### Confirmatory data analysis and finding outliers.

Having identified the riskiest areas of the audit, an auditor should next use some of the techniques discussed above to evaluate the data. These techniques are used first to infer analytic models to provide audit benchmarks or expectations; the actual values are then compared with the benchmarks. Any significant deviations should be investigated by auditors. For example, regression analysis can be used to derive a model for the revenue account based on archival data. The values calculated by this model should be compared against the actual revenue amounts, and any significant differences investigated.

### Evaluating results evaluation and integrating with traditional findings.

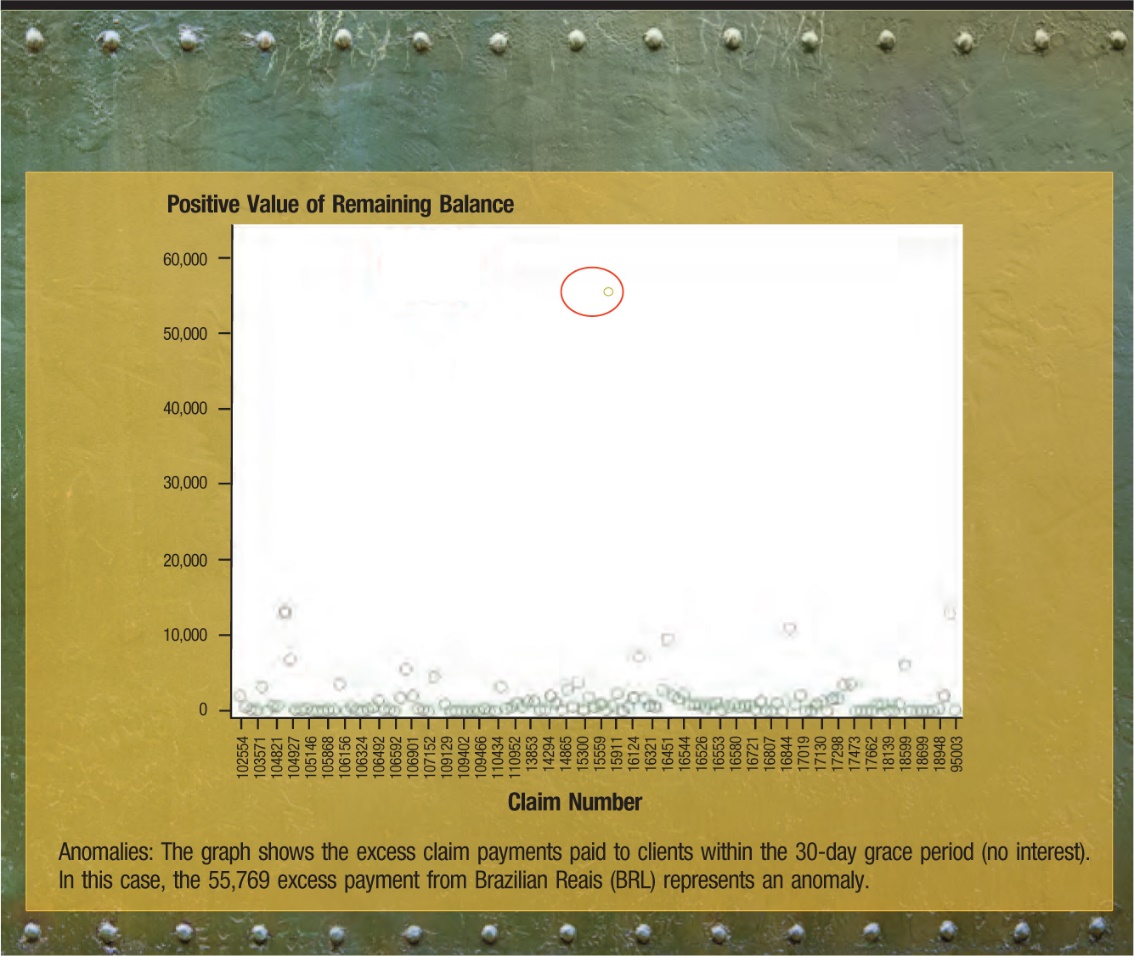
Ideally, the outliers should be segregated from the population for more detailed audit examination, as discussed above. In such an audit by exception (ABE) approach ([***Exhibit 4***](https://www.nysscpa.org/news/publications/the-cpa-journal/article-detail?ArticleID=12261#F4)), an auditor’s attention is more focused on the problematic transactions rather than a traditional sample pool (which may or may not identify problematic transactions). Theoretically, ABE provides a more efficient and effective approach for identifying questionable numbers.

### EXHIBIT 4 Audit by Exception (ABE) Approach

**[](https://www.cpajournal.com/wp-content/uploads/2017/02/CPA.2017.87.2.032.f004.jpg)**

Because this examination process is not sample-based but exception-based, it represents a significant departure from the currently prevalent audit practice of statistical sampling. The main difference between the ABE and a sample-driven audit is how the subset to be examined is obtained. Both approaches start with the entire population, but an ABE tests every transaction and ultimately focuses only on those transactions that present problems ([***Exhibit 5***](https://www.nysscpa.org/news/publications/the-cpa-journal/article-detail?ArticleID=12261#F5)), whereas a statistical sample does not test every transaction, as the sample purportedly represents the diversity and content of the entire population. If, however, the error-prone transactions as determined by the ABE tests represent, for example, less than .15% of the population, a sample of 60 transactions may or may not include even one data point that is significantly deviant, whereas every one of these .15% outlier transactions would be flagged for detailed testing by an ABE.

### EXHIBIT 5 Outliers Identified to Be Examined

**[](https://www.cpajournal.com/wp-content/uploads/2017/02/CPA.2017.87.2.032.f005.jpg)**

Nevertheless, many auditors and accountants may not initially feel comfortable with conducting an ABE of 100% of the population, unless this ABE examination were to be accompanied by a traditional statistical sample. The results of the ABE would then be examined in detail, just as currently the samples pulled are tested, with the findings compared and reported.

It is worth remembering that sampling became an accepted audit practice during a time when data sets were expanding in size but auditors were still examining transactions *manually*. Detailed examinations of entire datasets were infeasible at that time. Now that automated audit software capable of testing datasets rapidly with minimal manual involvement from the auditor exists, this obstacle is no longer an issue.

## Emerging Approaches

Although many of them have not yet been included auditors’ daily repertoire nor codified in audit standards, there are many emerging data analytics approaches that could assist with the audit process. Some of these are shown in [***Exhibit 3***](https://www.nysscpa.org/news/publications/the-cpa-journal/article-detail?ArticleID=12261#F3). The most promising of these approaches are described below.

***Predictive analytics.***

Carefully validated and highly accurate predictive analytic models for aggregated accounting numbers can be used by auditors to reduce the time-consuming effort of disaggre-gated testing if the predicted values and the values of management assertions are sufficiently close.

***Deep learning.***

The large audit firms are investing significant resources into the use of artificial intelligence to take advantage of their past experiences and industry knowledge. For example, data from working papers can be used to create automatic protocols for certain audit judgments, such as bad debt estimation, lease classification, and identification of abnormal contracts. Deep learning uses this knowledge in tandem with more advanced methods, such as neural networks, to represent the deeper structure of events and conditions in multiple layers of the neural network. Another term associated with deep learning is “cognitive computing,” a blend of automation and human interpretation. Deep learning requires tremendous computational storage and power, however, since the learning occurs by combining human expertise with enormous amounts of data. Many businesses outsource deep learning projects to contractors and research centers, such as IBM Watson. It is conceivable that in the near future an “Auditor Watson” could exist that would assist accounting firms with financial and operational audits.

***Blockchain/Smart contracts.***

The recent development of the virtual currency Bitcoin has been facilitated by a technology known as blockchain that can keep data public and replicates many transactions in a network using encryption methods. This methodology may presage a fundamental change in methods of data storage and validation. Smart contracts associated with blockchain might be able to automatically execute contract features without human intervention. For example, the contract between the auditor and the firm may dictate that if an outlier is larger than 100% of the median value of the transactions, it must be stopped and examined by human eyes; blockchain could theoretically flag such outliers and refer them to an auditor.

***Text mining.***

The emergence of big data, and the mixing of large corporate datasets and external, unstructured data, allows for highly promising machine understanding of text that may one day provide great validation for management-supplied numbers and support new audit products, such as continuous auditing and monitoring from external data. Of note is the fact that three of the largest audit firms have employed legal discovery tools or developed methods to text mine information from converted PDF documents to create deep learning inputs.

## Tools and Information Sources

More than 700 firms audit public companies, and many more audit or examine other entities. Smaller firms do not have the extensive financial and human resources that larger ones have, and thus may not be able to leverage data analytics technology to the same extent. There are, however, many sources of free software and educational materials that are currently available. A selection of these resources, in addition to commercially available tools, is listed below.

The open source R software has one of the largest library of applications available. Free software such as R and Weka are used nationwide in university courses and by some research and technology firms, but are somewhat frowned upon by accounting firms because they are not validated. These concerns are not without merit, since open source software can be clumsier and less user friendly than proprietary software, but their utility should not be ignored. In addition, while a basic knowledge of statistics and information technology is becoming essential for all accountants, other, more specialized functions can be contracted to other experts, perhaps online.

Proprietary tools such as Audit Command Language (ACL) and Interactive Data Extraction and Analysis (IDEA), as well as generic statistical software such as Statistical Analysis System (SAS) and Statistical Package for the Social Sciences (SPSS), are frequently used by large businesses and large firms. Furthermore, the capabilities and scope of these packages are constantly evolving, requiring that accountants and auditors have sufficient knowledge of analytics.

Large firms typically retrain their professionals through internal courses about their own approaches to auditing and are progressively trying to introduce audit analytics into this process. Four decades ago, each one of the then–Big Eight had its own IT audit packages, but today the Big Four use vendor-provided software such as ACL and IDEA. This convergence will likely also take place with the emerging statistical and visualization toolsets being developed.

A major difference in today’s environment is the power of group sourcing and the diffusion of the Internet. Powerful education mechanisms are emerging, ranging from free public resources to online Masters of Accountancy programs in audit analytics, some of which are financed by major firms (“KPMG, Villanova, Ohio State Launch First-Of-Its-Kind Data and Analytics Master’s Degree to Prep Data-Age Auditors,” KPMG, Aug. 4, 2016, [**http://bit.ly/2jWihzN**](https://bit.ly/2jWihzN)).

*The advent of data analytics and big data is not a fad; it is a real phenomenon driven by new technologies being adopted by many businesses.*

## A Growing Phenomenon

The advent of data analytics and big data is not a fad; it is a real phenomenon driven by new technologies being adopted by many businesses. Accountants and auditors are currently very far behind the curve. The profession will inevitably be forced to modernize audit approaches by corporate processes that are not auditable by traditional methods, accounting packages that can perform without manual intervention, and pressure from clients for more value in the audit engagement. This article provides a general introduction to modern analytic methods and sources of information and education for accountants. Further resources can be found at [**http://raw.rutgers.edu/CPAjrefs.html**](http://raw.rutgers.edu/CPAjrefs.html).

## FOR FURTHER WATCHING

Introduction to Audit Analytics:

[**https://www.youtube.com/playlist?list=PLauepKFT6DK8nsUG3EXi6lYVX0CPHUngj**](https://www.youtube.com/playlist?list=PLauepKFT6DK8nsUG3EXi6lYVX0CPHUngj)

Special Topics in Audit Analytics:

[**https://www.youtube.com/playlist?list=PLauepKFT6DK-PpuseJtSMlIy-YBhaV4TH**](https://www.youtube.com/playlist?list=PLauepKFT6DK-PpuseJtSMlIy-YBhaV4TH)

Information Risk Management:

[**https://www.youtube.com/playlist?list=PLauepKFT6DK8uxePhPCoHjDf8\_DlhRtGS**](https://www.youtube.com/playlist?list=PLauepKFT6DK8uxePhPCoHjDf8_DlhRtGS)

Tutorials for Risk Management:

[**https://www.youtube.com/playlist?list=PLauepKFT6DK9Grq8J67NMyGpYh1AsBb–PhD**](https://www.youtube.com/playlist?list=PLauepKFT6DK9Grq8J67NMyGpYh1AsBb--PhD)

Recordings:

[**https://www.youtube.com/playlist?list=PLauepKFT6DK9vKn7-eKxzmxBegpe8v8xw**](https://www.youtube.com/playlist?list=PLauepKFT6DK9vKn7-eKxzmxBegpe8v8xw)

Deniz Appelbaum is a PhD student in the department of accounting and information systems at Rutgers University, Newark, N.J.

Alexander Kogan, PhD is a professor of accounting and information systems in the department of accounting and information systems at Rutgers University, Newark, N.J.

Miklos A. Vasarhelyi, PhD is the KPMG Distinguished Professor of Accounting Information Systems and Director of Rutgers Accounting Research Center and Continuous Auditing & Reporting Lab in the department of accounting and information systems at Rutgers University, Newark, N.J.

# Three Ways to Enhance Audit Quality with Audit Data Analytics

April 4, 2018 [*Cathy Rowe*](http://news.cchgroup.com/2018/04/04/three-ways-enhance-audit-quality-audit-data-analytics/accounting-audit/a-a-hot-topics/#grve-about-author)

## Why Apply Audit Data Analytics?

Audit data analytics is a hot topic. In fact, there has been a loud call to action recently. That call to action is this. We must transition to applying audit data analytics (ADA) in performing our audit engagements. By applying audit data analytics, it allows those in the profession to enhance audit quality by working more effectively with large data sets. This in turn enhances the quality of the audit for three main reasons.

1. **There will be a stronger understanding of the entity’s operation and associated risks, including the risk of fraud.**
2. **Applying audit data analytics will help detect material misstatements by allowing 100% of items in a population to be examined thereby effectively managing sampling risk.**
3. **It will improve the communication of those charged with governance of audited entities because visualization dashboards offer an effective way to discuss insights or key matters, which are important in understanding why events may have occurred and the possible implication for control, financial reporting, or governance processes.**

Additionally, applying audit data analytics is a giant step towards meeting the objectives of the AICPA RADAR initiative. The desired outcome of RADAR for the auditing profession is improved audit effectiveness through the integration of data analytics and related technologies into everyday practice.

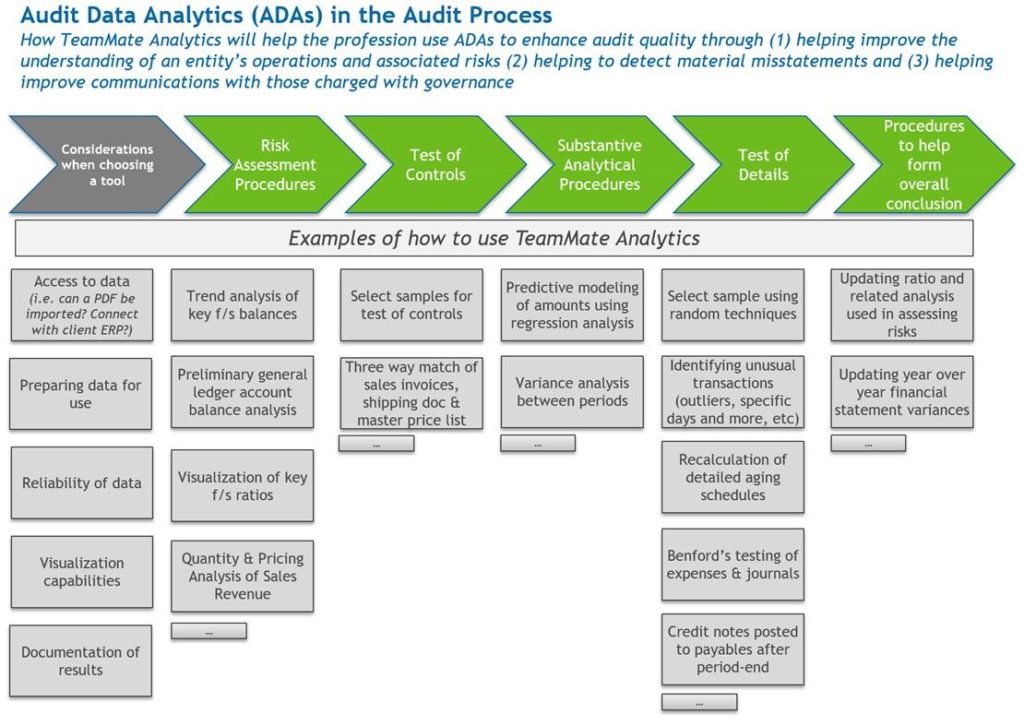
## Define Audit Data Analytics

Before we go any further, what exactly is audit data analytics (ADA)? ADAs are defined as “the science and art of discovering and analyzing patterns, identifying anomalies, and extracting other useful information in data underlying or related to the subject matter of an audit through analysis, modeling, and visualization for the purpose of planning or performing the audit”.(1)

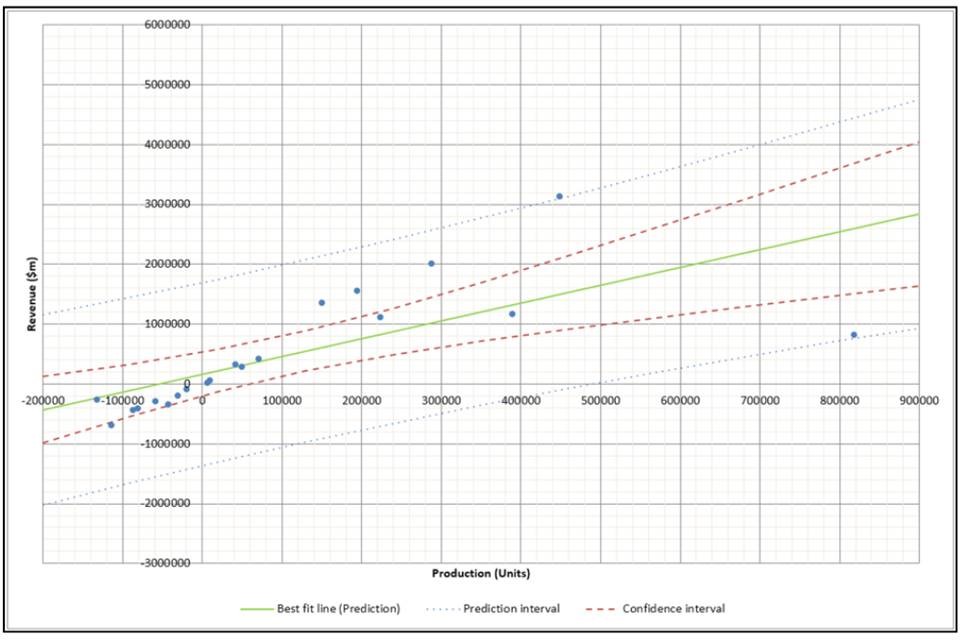
ADAs are similar to CAATs. However, audit data analytics have evolved and now enable auditors to use techniques to visualize the data and use it throughout the entire audit.

## Audit Data Analytics Examples

As can be seen in the below chart, there are significant opportunities to use ADAs throughout the entire audit process. Start with risk assessment procedures and go all the way through procedures to help form the overall conclusion. Check out the following three examples.



* **Example 1:**  Perform a preliminary “general ledger account balance analysis” as part of the risk assessment of the audit. This will allow you to identify unusual changes, unexpected trends, and risky transactions. It is important for the auditor to define risk transaction scores and have a full audit trail. Another way to accomplish this goal is to visualize key financial statement ratios.  This will help indicate areas where there is a higher risk of material misstatement.
  + *Tip: If you can access the client’s general ledger system throughout the year, there is an opportunity for continuous auditing and timely conversations about unusual transactions vs. waiting until the end of the year.*
* **Example 2:**Select a “sample for test of controls” and “test of details”. Instead of manually picking your sample items – use technology to create a true random sample! That eliminates any potential auditor bias in the selection process. Many auditors have historically just picked “25”. Now though, using audit data analytics, you can pick a true sample.
* **Example 3:**Use an ADA as part of completing substantive analytical procedures. As defined in AU-C section 520, analytical procedures evaluate financial information through analysis of plausible relationships among both financial and non-financial data. Before you perform the analytical, remember it is critical to develop the expectation of the results so you can compare the results and investigate the difference. One way to accomplish this is by using predictive modeling, such as a regression analysis, to predict amounts for future periods to help identify the potential for material misstatement. This analysis will give you a precise method for forming an opinion. This opinion will then allow you to include a larger number of variables in the analysis.
  + *Tip: Consider using the visualization of these trends for more effective conversations with your client. This in turn will help them understand how they can better run their business.*



## Call to action

In closing, keep in mind that the above are just a few examples of how to use technology to enhance the quality of your audits. There are hundreds of tests available in [TeamMate Analytics](http://engagetax.wolterskluwer.com/l/339101/2018-03-29/2w349?_ga=2.255958344.219707104.1612808826-1522942568.1612808826). The solution’s Test Library comes pre-packaged in customizable modules.  These modules stand ready for your firm to run with guidance to help interpret the results. In addition, coming soon in [CCH ProSystem fx Knowledge Coach](http://engagetax.wolterskluwer.com/l/339101/2018-03-29/2w34h), there will be a visual indicator added to the substantive audit programs to indicate when (and how) to use TeamMate Analytics to complete a substantive procedure. Wolters Kluwer’s [Integrated Audit Approach](http://engagetax.wolterskluwer.com/l/339101/2018-03-29/2w34r) is continuing to expand to support the evolution of the audit and allow you to challenge your status quo. [Learn more about TeamMate Analytics](http://engagetax.wolterskluwer.com/l/339101/2018-03-29/2w349) today.

*(1) Byrnes, Paul; Criste, Tom; Stewart, Trevor; and Vasarhelyi, Miklos. “Reimaging Auditing in a Wired World.”*