



# Get Analytics Right from the Start

**Neil Raden**

Hired Brains Research

February, 2010

Sponsored by

**SYBASE®**



## **ABOUT THE AUTHOR**

Neil Raden, based in Santa Barbara, CA, is an active consultant and widely published author and speaker and also the founder of Hired Brains, Inc., <http://www.hiredbrains.com>. Hired Brains provides consulting, systems integration and implementation services in Business Intelligence, Decision Automation and Advanced Analytics for clients worldwide. Hired Brains Research provides consulting, market research, product marketing and advisory services to the software industry

Neil was a contributing author to one of the first [1995] books on designing data warehouses and he is more recently the co-author with James Taylor of *Smart (Enough) Systems: How to Deliver Competitive Advantage by Automating Hidden Decisions*, Prentice-Hall, 2007.

He welcomes your comments at [nraden@hiredbrains.com](mailto:nraden@hiredbrains.com) or at his blog at Intelligent Enterprise magazine at: <http://www.intelligententerprise.com/blog/nraden.html>

## TABLE OF CONTENTS

Executive Summary	1
Types of Advanced Analytics	3
Predictive analysis	3
Optimization	5
Decision Services	5
Data	5
Sources	5
Use Cases	6
Retail	6
Banking	6
Telecommunications: Self-Repairing Network	6
Insurance	7
Marketing Services	7
Organization	8
Resources/Skills	8
Getting Started	9
Conclusion	9

## Executive Summary

A decade ago, businesses, governments, even charitable foundations spent considerable time and effort in rationalizing and replacing the disparate legacy operational systems that consumed a major portion of the IT budget. Despite the enormous cost these efforts represented as well as the risk of delay, disruption and even outright failure, today most organizations have at least partially migrated to an integrated operational platform. These enterprise systems, Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM), for example, have succeeded in streamlining operations therefore eliminating much redundancy in support of the running business processes. However, they failed to deliver much analytical insight into operations. Today, organizations are turning their attention to finding insight not just efficiency and are exploring ways to leverage those investments to get a better grip on what is happening and what is likely to happen.

A few years ago, Davenport and Harris released an influential book<sup>1</sup>, “Competing on Analytics,” which described how a dozen or so companies used “analytics” to not only advise decision-makers, but to play a major role in the development of strategy and implementation of business initiatives. The book found a huge following and was a best-seller on the business book lists. It certainly placed the word “analytics” in the top of mind of many decision makers. James Taylor and I released another in 2007<sup>2</sup>, “Smart (Enough) Systems,” which took a deeper look into how and why to deliver “smarter” systems and “decision services.” Other books related to the topic include “The Black Swan<sup>3</sup>,” “Outliers,<sup>4</sup>” and “Super Crunchers.<sup>5</sup>”

But what is analytics? For many years, the term used for analyzing information after-the-fact and reporting and distributing it was Business Intelligence (BI). If it involved any sort of math or statistics, it was called Operations Research or just “rocket science” for short. A dozen years or more ago, another term crept in, Knowledge Discovery in Databases (KDD) or, more simply, data mining. While it is true that definitions or boundaries for most terms in information technology are a little soft, in BI and analytics, terms are very imprecise. Terms like analytics, advanced analytics, descriptive analytics and predictive analytics can be used interchangeably (though they shouldn’t). The term predictive analytics (PA) is used without much precision at all. Technically, for something to be predictive, it has to have the capability of making some assertions about what it likely to happen in the future. To do that requires some insight into causality.

---


<sup>1</sup>Thomas Davenport and Jeanne Harris, *Competing on Decisions: The New Science of Winning* (Cambridge: Harvard Business School Publishing Corporation, 2007)

<sup>2</sup>James Taylor and Neil Raden, *Smart (Enough Systems): How to Deliver Competitive Advantage by Automating Hidden Decisions* (Boston: Prentice Hall, 2007)

<sup>3</sup>Naseem Taleb, *The Black Swan: The Impact of the Highly Improbable* (New York, random House, 2007)

<sup>4</sup>Malcolm Gladwell, *Outliers: The Story of Success* (New York, Little Brown and Company, 2008)

<sup>5</sup>Ian Ayres, *Supercrunchers: Why Thinking-by-Numbers Is the New Way to be Smart* (New York, Bantam, 2007)



The confusion is made worse because data mining is often one stage in developing a predictive model, and BI reporting and dashboard tools are invaluable in monitoring the results of a predictive model. In fact, one could say that analytics has to rest on the shoulders of the entire data management and BI effort or there would be no reliable data to use and there would be no effective way to disseminate the results.

In classical OLAP, a person builds reports from a combination of filtering conditions, the development of measures or metrics and attributes that define the grain and dimensionality of the data. In a true OLAP system, it is possible to perform nearly limitless navigations on this output by traversing various relationships, especially hierarchies (aggregations) and drill-down into underlying detail. There is certainly no question that this function of BI is a kind of analytics, but it isn't predictive, except insofar as the operator can draw conclusions from the manipulations.

There isn't much debate that analysis using quantitative methods such as statistics, mathematical algorithms, stochastic processes like Monte Carlo simulation and even optimization, are "advanced analytics," but not all are predictive. At the other end of the spectrum, are reports and dashboards that use basic statistics such mean, median and standard deviation advanced analytics? To a statistician, the answer would clearly be no. But given that most people do not understand the difference between two central tendencies, mean and median, much less the influence of left- or right-skew in a median, in the context of business, any use of statistics should be considered "advanced analytics."

Whether or not analytics should become an integral part of your organization's planning and decision-making seems to be beyond question now. However, at what level, and for what purpose, are questions that each organization needs to answer for itself. In addition, there are details that have to be addressed such as how to deploy analytics, finding the expertise, selecting the right tools and technology, setting reasonable goals and mapping out a three- to five-year progression of expanding techniques and results. These questions are the focus of this paper.

## Types of Advanced Analytics

Analytics emerge in an organization in a number of different forms: Bespoke, generic, embedded and packaged. In the first case, bespoke, people with advanced skill and understanding of the business apply mathematical techniques to profile data, search for meaning and develop models to report and predict business conditions, outcomes and forecasts.

On the other hand, there are many “generic” models that organizations use that are created from often extremely complicated and arcane models, but they are so widely used and such faith is placed in their usefulness, that organizations adopt them without any real understanding of their essential nature. The FICO score is an example of this. It may be used directly as a driver for credit evaluation or even employment decisions, or it may be used as one component of more customized models, such as a large bank’s mortgage underwriting model. In this case, the generic model is “embedded.”

The last case is the packaged analytic model. Here, analytical models are part of a broader application package, employing perhaps both generic and bespoke models, but it is all transparent to the company that uses the package. Shipping logistics, supply chain optimization and consumer recommendation engines are all examples of this type.

At the functional level, there are three types of actual advanced analytics methods: descriptive, predictive and optimization.

### Descriptive Analytics/Data Mining

One could say that all quantitative methods that do not involve prediction or optimization are data mining, though some would argue with that. Crisp definitions are not as important as understanding the breadth of analytics as a whole. Interactively navigating through data, either visually or using OLAP tools, is a sort of cognitive data mining, but in general, data mining uses mathematical and statistical techniques to understand typically large volumes of data, such as from a data warehouse, though there are many other data sources that are used routinely. The routines attempt to add some intelligence to the data, either in a directed way by an operator (also known as a knowledge engineer, but we don’t like this term), or in a purely mechanical way.

For instance, sorting a list of people into 10-year age brackets is simple, provided the list includes their age or date of birth and the list isn’t too large. But suppose you needed to derive their

approximate age from other criteria such as their date of high school graduation or first drivers’ license. This is called “categorization,” because the data mining routine is given the categories for grouping. In contrast, “classification” uses clustering algorithms to figure out groupings of things based on many criteria and their relationships to others in the group.

There are a multitude of applications for this technology, especially gaining insight into how two things are related in ways that have escaped attention, leading to the development of association rules, likelihood scores, and even trending. However, data mining is not a crystal ball. There has to be a modeler who is familiar with the nuances of the techniques, the data being used and, perhaps most of all, the needs and practices of the organization. Creating a useful model is only partly a technical exercise.

Once descriptive analytics can paint a clearer picture, the next question is, “What can we do about it?”

### Predictive analysis

Because something is predictive, it naturally involves the future, but the future can be as near as the next instant, or as murky as five years from now. A predictive model everyone is familiar with is weather. Predictions for the next few days have become pretty reliable, thanks to both better models and better input data, but beyond a week or so, they aren’t so reliable. In business, a predictive model attempts to answer the question, “What will happen,” and, in some cases, “What are the consequences?” For example, a predictive model might alert purchasing that a supplier’s actions predict they will fail to fulfill a contractual obligation or that their creditworthiness is likely to degrade. Election polls are another well-known predictive model, where predictions about the outcome of an election are made based on very complicated models applied to a small sample of voters or exit poll surveys.

Predictive models typically are not dynamic, meaning, once they are assembled, they are put in place and do not change unless replaced. They can be run in the background to provide results for further analysis or they can be part of an on-line system where the inputs to the model can be quite dynamic. Response time is critical in these latter cases, but not so much in the former.

The implications for data management are two-fold with any analytics. Modelers work with existing data whose accuracy and reliability are important to understand and quantify, as they influence the model’s quality. Keep in mind that unlike many

applications in data warehousing and BI, where the quality of the data has to be very high, or even perfect for some applications, such as external reporting, predictive models are capable of spotting abnormalities in the data and dealing with them. The goal is not to report with accuracy, it is to understand the causes and relationships in the data in order to make predictions.

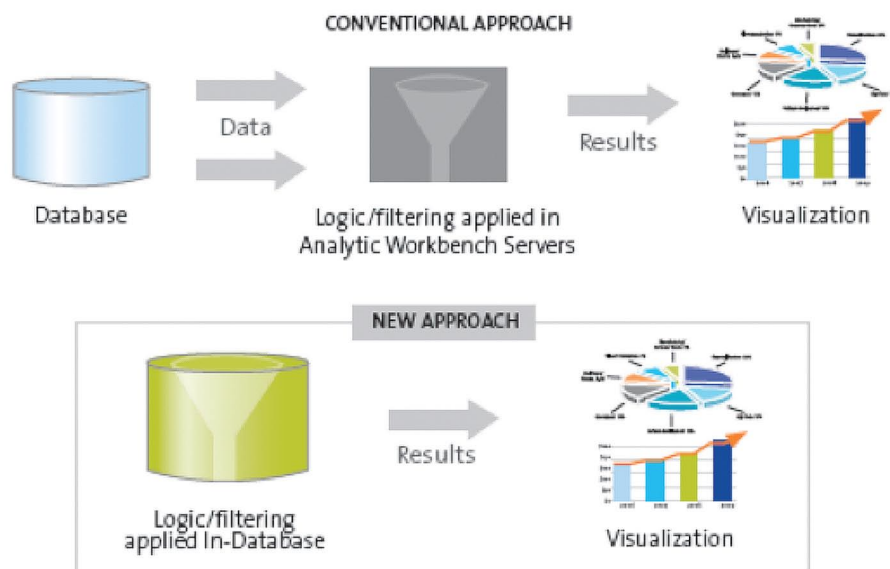
Data miners/predictive modelers often create subsets of data called training sets, to test their models against data with known outcomes. This is a very important step and one that requires great skill and care so as not to introduce bias into the models. There is even a concept called “overtraining” where the models become too sensitized to the training sets and lose their ability to predict well.

Most analytic modeling tools and analytic modelers don't get their data directly from operational systems or from data warehouses; instead, they spend a great deal of time assembling their data from multiple sources, both internal and external, to extract the data they need in a format that lets them work on it. This process actually wastes a great of the analyst's time, up to 80% by some estimates, because the data has to be assembled and moved to a platform where the algorithms can sift through it. The diagram below, at the top, depicts this process.

A better approach is to move the logic to the location of the data, thereby eliminating both the need to move the bulk unprocessed and unfiltered data from one platform to another, as well the more efficient process of actually preparing the data for analytical processing and also performing the analytics, in one place.

*This is an area that has finally been addressed by embedding statistical and other quantitative functions directly in a relational database. Sybase IQ incorporates the DB Lytix™ library from Fuzzy Logix which allows modelers to execute a rich set of routines directly within the database system, eliminating the need to create datasets.*

Some of the models used in predictive analytics have forbidding sounding names such as neural networks, decision trees, support vector machines, general regression, regression, clustering models, association rules and naïve Bayes classifiers, to name a few. Despite these unapproachable names, most of these are fairly intuitive and not terribly difficult to deploy with some training.



## Optimization

Optimization is a logical step beyond prediction. If the causal agents are understood, the goal is not to just predict, but to reorganize for the best possible outcome. Optimization models can range from the very simple to the truly weird. In the latter case, consider the revenue management practices of commercial airlines. No one can really say that the pricing of airline tickets makes any sense but, presumably, the models to combine maximizing revenue, crew and equipment scheduling and timing, among other things, are truly sublime. There are, in fact, a number of third-party organizations that specialize in optimization models for airlines. At a simpler level, an individual's portfolio management is an attempt at optimizing return versus risk or a university's admissions process to get the desired type and number of entering freshman. Granted, many of these models are more "rule of thumb" than truly optimization, but the rule usually represents some careful thinking about many variables over time.

Enterprise-level optimization models combine many descriptive and predictive models and go on to employ many other features of their own, even using probabilistic and stochastic methods like Monte Carlo Simulation or Bayesian models. This is clearly not for the faint-of-heart, but many organizations use third-party optimizations as part of packaged software applications that have been created and vetted over time.

## Decision Services

Although not really a part of advanced analytics, Decision Services is an approach to putting the results of predictive models into action. A decision service is invoked from an operational system, especially one employing predictive models. The decision service may be implemented as a rules engine, or it may be less formal, such as some procedural code in the application. A growing and very interesting driver of decisions instead of a rules engine is an ontology developed with RDF/OWL and accompanied by a "reasoner" that can understand the relationships and meanings in the ontology to derive, through deductive logic, new information. In any case, the decision service actually decides what to do and closes the loop by initiating certain actions in the operational system. Common examples include fraud detection (including what to do when the fraud is detected), credit approval, decisions to switch supply channels in real-time and recommendation engines (perhaps you would like this umbrella to go with those new boots?).

## Data

To derive insight from data, you must have data available first. The accuracy, completeness, and integration of the data you have determine the quality of insight you can derive. Building predictive analytic models from inaccurate data or mining inaccurate data for business rules could be fatal because the results build on inaccuracies in the data and product inaccurate predictions. You can get analytic insight from accurate data, even if it's limited in scope or completeness.

Clearly, however, the more complete the data, the more robust your analytic models are likely to be. Similarly, customer-based data provides more opportunities for interesting analytics than account-level data does.

## Sources

All stored data in any form represents some state or description that is potentially useful. Even data in operational systems used to document the process, not the result, can be useful. Data mining and predictive analytics are usually applied to data that is "rich" in attributes. For example, the quantity of five may not be very meaningful, but it is when it represents a measurement of something that is described by many attributes and/or it can be related to some wider context. For example, five may be the number of members of a household, but a database may contain hundreds of other bits of information about that household. Or, five may represent the power of an aftershock of an earthquake, one of thousands of aftershocks recorded.

One of the more common uses of analytics is trying to understand the behavior of people. At one level, looking at the history of an individual, say, their purchases and returns over time and comparing them to others, is a sort of direct form of data mining. A more indirect method is to evaluate a person's process behavior, such as their journey through your website. In this case, the data resource would not be the order-entry system but rather the clickstream log from which one would attempt to understand the person's actions and predict their behavior under various situations. It is rather likely that purchase behavior is in your data warehouse, but somewhat less like that your clickstream data is.

Data for analytics can and will come from many sources. The more of those sources that can be integrated into a data warehouse the better. Otherwise, it is best to find a way to federate the data, putting it in one place "virtually" by means of some sort of abstraction or semantic layer. There are some very mature tools for doing this.

## Use Cases

### Retail

A common question in retailing is how to turn customers into recurring customers. One retailer with hundreds of stores, relied heavily on discounts to draw in business, especially for seasonal offers, and had no way to tell whether offer selection was optimal. It had limited customer understanding and many internal opinions (but no insight) about shopper frequency, lifetime value, and demographics. With the addition of Web and phone channels, it also suffered from more poorly integrated data and channel-by-channel decisions.

The solution was to integrate transactional data with external demographic data to develop descriptive analytic models that matched all customers with those

covered in qualitative research projects. These models explained the demographics of different channels and product ranges.

Trials of different offers

supported developing predictive analytic models for marketing response and future revenue. Trade-off analysis between these models affected offer rules.

For instance, different offers were made to high responders who generate low revenue and low responders who generate high revenue. A continuous improvement infrastructure for the models supported decisions that were integrated across channels and used heavily segmented discount promotions.

The benefits derived from the program were:

- Store traffic increased 200 percent over the control group.
- Sales went up \$3.4 million and increased profit net of offers.
- Understanding of customers increased, including how few multiple purchasers existed and how many visits per purchase were required.

### Banking

A credit card issuer had an expanding customer base and was using a single model for predicting risk for all customers. The risk model was showing signs of increased losses and incorrect actions for customers, although it was rank-ordering customers (by risk) accurately.

Lift is a measure of model effectiveness. To the extent the model produces better results than by chance alone, the lift is

positive. For example, suppose 6.2 percent of customers receiving a retention offer are actually retained by using a subjective approach. If a model increases that return to 11.6 percent, the model's lift is 1.87—the ratio between 11.6 and 6.2. Lift is often used to compare predictive analytic models. In the case of the credit card issuer, the goal is to increase the number of retentions of “good” customers as well as to not make retention offers to less desirable ones.

Descriptive analytics divide customers into multiple segments, each with a distinct risk model. These risk models are more specific because they handle only a subset of all customers. The benefits from this approach are:

- Improved accuracy in risk modeling
- More than \$1 million in loss savings

### Telecommunications: Self-Repairing Network

In a telecommunications company, all network faults were reviewed manually, and an engineer was dispatched to make a repair. Meanwhile, an engineer in the control center reviewed the existing network and traffic and reassigned traffic around faulty equipment. Major customers with service-level agreements (SLAs) might be affected by a failure, but the company didn't know until the end-of-month analysis, which caused customer service problems.

#### Solution

Predictive models are able to more evenly anticipate outages, and those that are not predictable are more quickly resolved. A decision service handles network errors and alerts without requiring staff intervention. To respond to system failures, the service uses business models developed by advanced analytics to assign field service engineers based on region, product expertise and urgency. The system ensures accurate assignment of field service engineers and allows easy modification and deployment of statuses about new engineers or products. The service tracks and correlates system wide alarm information to determine uptime and downtime for equipment on the network and routes calls through equipment that's running. It can prioritize major customers. As new equipment is added or company experts learn more about equipment, experts can add new rules to the service. It can even balance service request responses based on what compensation is owed to major customers in the case of system failure.

From a cost perspective, the models are able to more accurately predict the need for maintenance, both preventative and clearing, and balance labor forces between employees, overtime, contractors and union requirements.

#### **Benefits**

- More effective assignment of engineers
- Faster response time to network faults and outages
- Cost savings for contractors and overtime
- Proactive management of major customers

#### **Insurance**

##### ***Mid-Tier U.S. Insurer: Auto Insurance Underwriting the Old Way***

At a U.S. property and casualty insurance company quoting and underwriting auto policies were manual processes. For new policies, a prospective customer needed to see an agent, answer questions and receive a quote. After customers had several quotes, they returned and worked with the agent to fill in information for a policy request. This request was transmitted to the insurance company and

queued up for an underwriter. When an underwriter reached that policy request, she consulted the policy manual, ordered additional reports, asked the agent for more information, and eventually assembled the data she needed to make a decision and underwrite the policy for a specific price. This price might not match the quote given the customer, however, which could mean several conversations with the agent to gather all the right information.

Almost all the underwriters' time was spent on this policy-by-policy decision making. Underwriters also spent time reviewing renewals going through the system; however, their workload meant that 80 percent of policies were renewed without review.

Predictive analytics were added to the application portfolio which insured that decisions met requirements for regulatory compliance (with federal and state regulations), knock-out rules—that reject applicants for specific reasons, and company policies, embedding risk models (scorecards that used information about prospects to predict their future claims risk) into this decision. The new models worked smoothly behind the scenes to enable consistent, real-time decisions.

The benefits are:

- The same number of underwriters can process 35% more applications.
- Agents and the company have a clear understanding of risk management policies.
- The company can profitably win the good business it wants and lose the bad business it doesn't want.

#### **Marketing Services**

Customer segmentation: A company was already using predictive analytics to make credit line and other account decisions. The service supported adaptive control, and the company had seen good results. Customer segmentation, however, was largely subjective and seemed to offer potential for improvement.

Descriptive analytics, particularly the classification and regression tree

(CART) approach, were used to build new customer segmentation rules from historical data. The trees had different performance objectives at each level. First, they segment on retention, then on profitability, and then on likelihood to respond. Large data volumes are processed rapidly to develop new models on a regular basis. They develop and profile new trees, compare differences offline, and ultimately transfer new rules (in the form of a decision tree) to production by using the adaptive control infrastructure.

The benefits are:

- Quick implementation, just three months from start to finish
- More specific segmentation where it adds the most value
- Segmentation that's closer to optimal for the company's goals
- Growth in revenue of 8 percent per account in six months

---

<sup>5</sup>Adaptive control is a technique to continuously test the usefulness of a predictive model by "challenging" the current model. In production, a small slice of the current transactions are fed through the "challenger" and if better results are gotten, the "champion" is replaced with the new model.

## Organization

There are a number of models for deploying analytical capability in organizations. Some organizations choose to build a centralized unit with very highly trained people who act as a sort of analytics development organization for the company. They more or less select the projects they pursue, usually with input from senior management. Many years ago, this arrangement was called the OR group for Operations Research, but the term has largely fallen out of use.

A variation on the above is a similar group, but one that performs like an internal consulting organization, helping other groups within the organization apply analytical techniques to their business units or functions. There are many positives about this arrangement, such as the potential for peer review within the group given their proximity. Those who use the services of this group need some assurance that the advice and/or deliverables they get represents the best choices since they often understand very little about the process. Peer review provides this assurance.

A more common arrangement currently is to place some DM/PA capability directly in the business organizations. In this arrangement, the quantitative staff is part of that group, such as marketing, risk management, logistics, etc. This clearly provides enterprise not only in the quantitative methods, but also a deeper understanding of that part of the business.

Many organizations outsource these capabilities, using consulting firms, contracting with universities and even the professional services of software vendors who provide statistical/predictive tools and even packaged software with embedded analytics.

## Resources/Skills

One of the often-cited problems with people who have advanced quantitative skills is that they are often perceived as “too brainy” or towards those who lack quantitative skills. People often cite this experience with quantitative experts and would prefer someone who uses language and concepts they understand and doesn’t get impatient when asked naïve questions or having to explain and justify their approach even if it is standard. One the other hand, there is a reasonable chance that this reputation is borne of the subject matter, math that most people fear the most. Quantitative experts, in practice, don’t act any differently than any other highly trained professionals who have to explain themselves to those who have little to no background in the subject and show no interest in acquiring it. These same complaints have been made at various times about IT people or doctors, for that matter.

Advanced analytics are routinely performed by people without a Ph D. Actuaries are a good example. Most actuaries can (and do) apply very complicated quantitative methods to their work, yet very few have a PhD. Obviously, some course work in probability and statistics is helpful, but many actuarial students do not take those courses in college. They learn them and pass the actuarial exams through self-study over a number of years. Hiring promising people out of college with the right background, and allowing them to grow into the role of a quantitative expert makes sense and alleviates the PhD shortage problem.

Clearly, very specialized people at the PhD level are still needed for new algorithm and model development and for the very sophisticated applications. Building a pricing model using multiple regression or Bayesian analysis is not; this can be done by a well-trained professional.

## Getting Started

There are many challenges to getting started with a program of advanced analytics. The first thing an organization should do is assess whether or not their culture is amenable to quantitative methods, and in what areas. There are organizations that are overly reliant on their gut feel and will realize value from an analytics initiative. In those cases, it's best to get started with embedded models and packaged analytics and to start slowly in areas that are not too visible. Another factor that is often overlooked is the fact that most Predictive Analytics applications need a strong data management repository as a backbone that feeds the building and scoring of analytical models. Increasingly, organizations are deploying columnar databases with in-database analytical capabilities that are specifically designed for advanced analytics.

Most companies, however, are ready to apply these tools, but don't know where to start. Here are some areas to examine:

- Business and IT collaboration
- Data readiness
- Analytic understanding
- Evaluation of technology platforms
- Willingness to change
- Management focus on operations

The most important thing, though, is to work at freeing analytics from its image of super-science that can't be understood by anyone except the anointed quant's. This takes time, of course, but once a successful application like the ones above is put into place, those opinions begin to dissipate. It just takes some time and some good messengers.

## Conclusion

Advanced analytics will be adopted by most organizations and attain the status of "must have." While the majority of people in organizations will not become quantitative experts and modelers, the affect of predictive models will be felt across the organization and beyond. They already are. It would be wise to take steps now, and a good first step is to begin evaluating technology solutions that will be suitable for the development and implementation of analytics. From a technology perspective, one clear requirement is an analytic engine embedded in your analytical database technology.

It will take a few years to get people hired and productive, and a little while longer for those without all the skills to get trained and mentored. In the meantime, use consultants where necessary, get some good advice about where advanced analytics will be most useful in the short term and gradually develop an analytic culture.