research

BEST PRACTICES REPORT

Q3 2017

Advanced Analytics: Moving Toward AI, Machine Learning, and Natural Language Processing

By Fern Halper

tdwi
Transforming Data With Intelligence™

Research Sponsors

SAS

ThoughtSpot, Inc.

Vertica

research

BEST PRACTICES REPORT

Q3 2017

Advanced Analytics: Moving Toward AI, Machine Learning, and Natural Language Processing

By Fern Halper

Product and company names mentioned herein may be trademarks and/or registered trademarks of their respective companies. Inclusion of a vendor, product, or service in TDWI research does not constitute an endorsement by TDWI or its management. Sponsorship of a publication should not be construed as an endorsement of the sponsor organization or validation of its claims.

This report is based on independent research and represents TDWI's findings; reader experience may differ. The information contained in this report was obtained from sources believed to be reliable at the time of publication. Features and specifications can and do change frequently; readers are encouraged to visit vendor websites for updated information. TDWI shall not be liable for any omissions or errors in the information in this report.

Table of Contents

About the Author

FERN HALPER, Ph.D., is vice president and senior director of TDWI Research for advanced analytics. She is well-known in the analytics community, having been published hundreds of times on data mining and information technology over the past 20 years. Halper is also coauthor of several Dummies books on cloud computing and big data. She focuses on advanced analytics, including predictive analytics, text and social media analysis, machine learning, AI, cognitive computing, and big data analytics approaches. She has been a partner at industry analyst firm Hurwitz & Associates and a lead data analyst for Bell Labs. Her Ph.D. is from Texas A&M University. You can reach her by email (fhalper@tdwi.org), on Twitter (twitter.com/fhalper), and on LinkedIn (linkedin.com/in/fbhalper).

About TDWI Research

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on data management and analytics issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of data management and analytics solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

About the TDWI Best Practices Reports Series

This series is designed to educate technical and business professionals about new business intelligence (BI) technologies, concepts, or approaches that address a significant problem or issue. Research for the reports is conducted via interviews with industry experts and leading-edge user companies and is supplemented by surveys of BI professionals. To support the program, TDWI seeks vendors that collectively wish to evangelize a new approach to solving BI problems or an emerging technology discipline. By banding together, sponsors can validate a new market niche and educate organizations about alternative solutions to critical BI issues. To suggest a topic that meets these requirements, please contact TDWI senior research directors Fern Halper (fhalper@tdwi.org), Philip Russom (prussom@tdwi.org), and David Stodder (dstodder@tdwi.org).

Acknowledgments

TDWI would like to thank many people who contributed to this report. First, we appreciate the many users who responded to our survey, especially those who agreed to our requests for phone interviews. Second, our report sponsors, who diligently reviewed outlines, survey questions, and report drafts. Finally, we would like to recognize TDWI's production team: James Powell, James Haley, Peter Considine, Lindsay Stares, and Michael Boyda.

Sponsors

SAS, ThoughtSpot, Inc., and Vertica sponsored the research and writing of this report.

Research Methodology and Demographics

Report purpose. There is a lot of confusion in the market regarding machine learning (ML), natural language processing (NLP), and artificial intelligence (AI). This report educates readers about these technologies. The report also examines how organizations are using the technologies and organizational and technology best practices for getting started and gaining value from them.

Terminology. Respondents were provided with specific definitions for AI, machine learning, and NLP. These are detailed in the definitions section below.

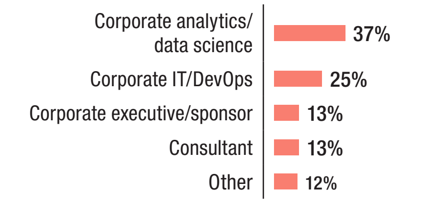
Survey methodology. In June 2017, TDWI sent an invitation via email to the business intelligence (BI) and data professionals in our database, asking them to complete an online survey. The invitation was also posted online and in publications from TDWI and other firms. The survey collected responses from 324 respondents who were either not implementing these technologies (15%), planning to implement them (42%), or already implementing them (43%). A total of 267 respondents completed all questions. All responses are valuable and so are included in this report's data sample. This explains why the number of respondents varies per question.

Research methods. In addition to the survey, TDWI Research conducted telephone interviews with technical users, business sponsors, and analytics experts. TDWI also received briefings from vendors that offer products and services related to these technologies.

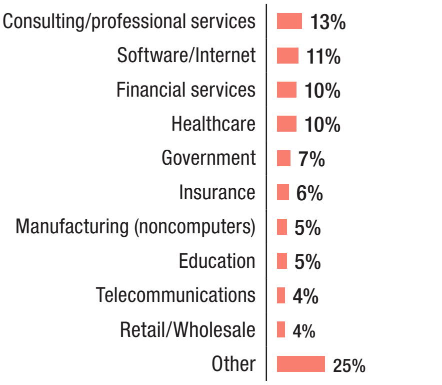
Survey demographics. The majority of survey respondents are in corporate analytics (37%), followed by those in IT (25%), executives and sponsors (13%), and consultants (13%).

The consulting (13%), software/Internet (11%), financial services (10%), and healthcare (10%) industries dominate the respondent population, followed by government (7%), insurance (6%), manufacturing (5%), and other industries. Most survey respondents reside in the U.S. (57%) or Europe (12%). Respondents come from enterprises of all sizes.

Position

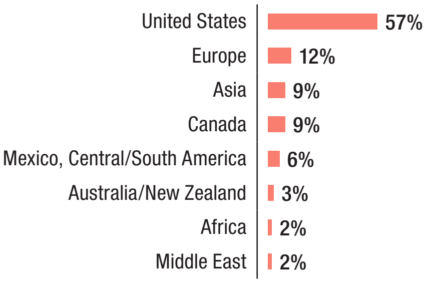


Industry

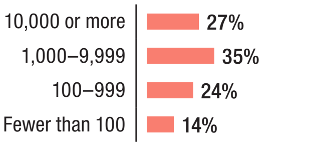


("Other" consists of multiple industries, each represented by less than 4% of respondents.)

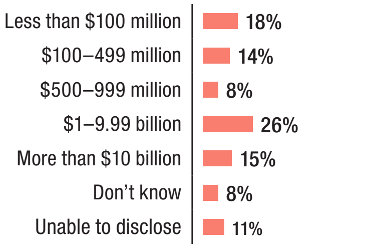
Geography



Number of Employees



Company Size by Revenue



Based on 267 respondents who completed every question in the survey.

Executive Summary

New advancements in compute power along with some new algorithmic developments are making machine learning, NLP, and AI attractive to many companies.

There is a lot of excitement in the market about artificial intelligence (AI), machine learning (ML), and natural language processing (NLP). Although many of these technologies have been available for decades, new advancements in compute power along with new algorithmic developments are making these technologies more attractive to early adopter companies. These organizations are embracing advanced analytics technologies for a number of reasons including improving operational efficiencies, better understanding behaviors, and gaining competitive advantage.

We have found that organizations are making use of these technologies in numerous ways. Some are applying machine learning for traditional use cases such as fraud and risk analysis or analyzing customer behavior. Others are using machine learning for preventive maintenance. Still others are building interactive chatbots and B2B applications that provide intelligence using a natural language interface. Deep learning is being employed to classify images and diagnose diseases. The use cases are wide and growing.

Data scientists are leading the way in terms of model building using various technology approaches. They are making use of open source analytics technologies such as R and Python as an important part of the advanced analytics efforts. Commercial analytics products are also being deployed by many, and some use open source in conjunction with commercial platforms. Organizations are also continuing to build out their data environments for analytics, with many beginning to utilize multiplatform data architectures.

Another important trend is that more AI technology approaches are targeting users beyond data scientists (e.g., a broad range of business users and "citizen" data scientists). Analytics applications more often include built-in AI/ML algorithms that are targeted to make it easier for business analysts and users to find insights. These include natural-language-based search interfaces, automated suggestions, and automated model building.

Early adopter experience provides clues as to best practices for those getting started with these technologies to gain advantage more quickly. For instance, early adopters are building centers of excellence (CoEs) and are hiring data scientists and analytics leaders. They are focused on data quality for analytics, operationalizing their analytics, and providing training opportunities.

Overall, one thing is clear—organizations that are utilizing these technologies now are gaining value. In fact, early adopters are much more likely to be satisfied with their analytics deployments than those that are just getting started with more advanced analytics or those that have no plans. Organizations are also seeing value as they move through the analytics success cycle.

This TDWI Best Practices Report examines organizations' experiences with and plans for machine learning, NLP, and AI, including technology plans as well as organizational strategies. It also looks at various advanced analytics challenges and how organizations are overcoming them. It examines the importance of new open source models and automated intelligence. Finally, it offers recommendations and best practices for successfully implementing more advanced analytics such as machine learning and AI in the organization.

An Introduction to AI, Machine Learning, and Natural Language Processing

What Goes Around Comes Around

Organizations are at an inflection point when it comes to analytics. Many are already deploying technologies such as predictive analytics, geospatial analytics, and text analytics. Some even have a unified plan for using these technologies together because they understand that greater insight and the ability to take action require it. These organizations realize that in order to be competitive, they need to be predictive and proactive.

There is also a smaller group of leading-edge companies that are pushing the envelope by deploying "newer" technologies such as machine learning, natural language processing, and artificial intelligence either to build models or put under the hood of their analytics platforms. Although early adopters are using these technologies now, other organizations are just starting to explore them. TDWI research indicates that these organizations are interested in learning more about these emerging technologies and finding out if these solutions are right for them, either in the near term or in the near future, so they can be prepared.

Many of these technologies are not particularly new (see Figure 1), but the combination of readily available compute power, platforms, and an organizational imperative to analyze and utilize data is creating an analytics renaissance. Many organizations now realize that to be competitive they need to be proactive in their analytics strategy. This includes utilizing and deploying advanced technologies such as machine learning and AI as part of an ever-expanding analytics ecosystem.

TDWI research consistently finds organizations that deploy advanced analytics technologies are more likely to measure a top- or bottom-line impact.

What Goes Around Comes Around

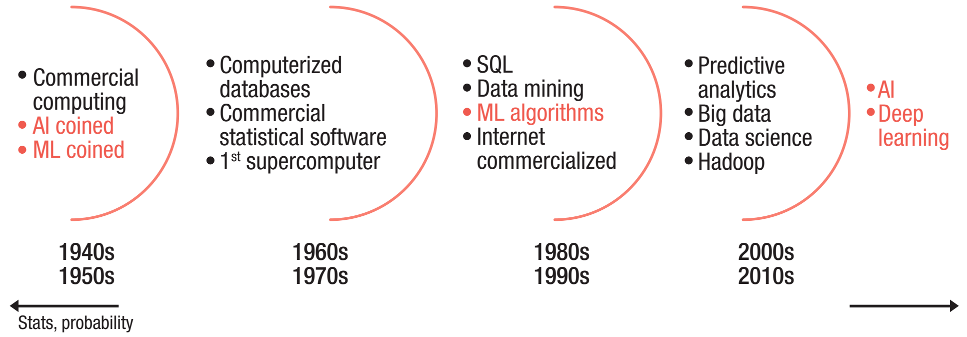


Figure 1. *Machine learning and AI are not new technologies. Both terms were coined in the 1950s.*

TDWI research consistently finds organizations that deploy more advanced analytics technologies are more likely to measure value. This makes sense and in some ways is a virtuous circle. As an organization adds more data for analytics, it gets better results; those better results drive success and the company then starts to build on its success, perhaps bringing in new analytics techniques. As management sees it is successful, the organization continues to build on its achievements. This is what TDWI refers to as the *analytics success cycle*. The idea of this cycle is important and we will return to it at various points in the report.

Technologies Introduction

Machine learning utilizes mathematical and computational science approaches.

Machine learning, natural language processing, and artificial intelligence have been part of the advanced analytics lexicon for years. In fact, the terms *machine learning* and *artificial intelligence* were coined in the 1950s. Today, however, there is a lot of market confusion about what they mean. Some people use the terms interchangeably, but they are not the same thing, regardless of the overlap between some of the techniques used.

- **Machine learning (ML).** This method of data analysis originated in the field of computational science. In machine learning, systems learn from data to identify patterns with minimal human intervention. The computer learns from examples, typically using either supervised or unsupervised approaches. In supervised learning, the system is given a target (also known as an output or label) of interest. The system is trained on these outcomes using various attributes (also called features). In unsupervised learning, there are no outcomes specified.

  Some people think of machine learning as being completely different from predictive analytics. There is a tendency to associate utilizing statistical approaches, such as regression, with predictive analytics. Machine learning, however, is used in predictive analytics and offers a different approach. Popular use cases include churn and fraud analysis, which are also popular for statistical approaches.

  Part of the market hype around machine learning includes its subset, deep learning, which uses algorithms to learn functions that can classify complex patterns such as images. Some deep learning algorithms, such as artificial neural networks, have input nodes and a number of hidden layers that act as a kind of "black box" to model one or more output nodes. Whereas early neural networks could not recognize complex patterns, some algorithmic advances in the last decade or so, together with the availability of vast compute power, have made deep learning feasible as well as more accurate than some thought possible. This has spurred greater interest in deep learning for use in audio-, image-, and text-classification problems. Much of the hype around AI comes from deep learning.

- **Natural language processing (NLP).** This involves analyzing, understanding, and generating responses ultimately to enable interfacing with systems using human rather than computer languages. For text, NLP often uses semantics to parse sentences for entities (people, places, things), concepts (words and phrases that indicate a particular idea), themes (groups of co-occurring concepts), or sentiments (positive, negative, neutral). Today, NLP is used in text and social media analytics tools to analyze issues and opinions.

  A popular use case for NLP is analyzing tweets or reviewing sites for feedback on products. For example, a marketing department for an electronics company might launch a campaign for its new reasonably priced portable chargers. Based on sales, it might think that it is doing well. However, in reality, the customer might not like the product and may take to social media to complain about it (perhaps devices using this product do not hold a charge for long). If the company can analyze those tweets and reviews using NLP technologies, it will be able to understand what people are talking about, their sentiment (positive, negative, neutral), and even how emotional they are about it (from the words used in the tweets).

  Although analyzing text for marketing is extremely important, another use of natural language processing is to generate languages to enable interfacing with systems using human language. This is found in interactive applications, such as chatbots, or other customer experience applications, such as routing a customer to a certain agent based on status and what was said.

Also related here is search-driven analytics where users employ a natural language experience to search and analyze their data to find insights. Some of these search engines learn about users from their analytics history and then provide them with search suggestions based on what might be most relevant to them.

NLP, together with machine learning, is also used in other applications such as text summarization and classification.

- **Artificial Intelligence (AI).** The idea that machines could act "intelligently" has been around since the ancient Greeks, but there has been no real consensus about what *artificial intelligence* actually means. Back in the 1950s, when John McCarthy of Dartmouth College coined the term, he described it as, "Making a machine behave in ways that would be called intelligent if a human were so behaving."[1]

  There has been debate on the subject ever since, with some definitions including the ability of a computational device to learn, reason, and interact. Other narrower, task-oriented definitions include cognitive tasks that humans do easily, such as image recognition or natural language understanding. Some researchers talk about general versus applied AI. Others discuss weak or narrow AI to perform specific tasks versus strong or true AI, where machine intelligence is equivalent to human intelligence.

  AI can mean different things to different people. The confusion is compounded by the fact that some software vendors want to label everything as AI while some researchers take a purer view of the definition. For the purpose of the survey used in this report, the following definition was used: AI is the theory and practice of building computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision making, and translation between languages. Of course, the bar continues to move. A few years ago, a recommendation engine might have been viewed as AI. Now, witness the emergence of self-driving cars, personalized bots, and computers that can detect cancers with the same reliability as a physician.

Machine learning and NLP are both subfields of AI and will be discussed in detail in this report. Other subfields of AI include robotics, scene recognition, and automatic programming.

EXPERT OPINION THE LANDSCAPE OF ARTIFICIAL INTELLIGENCE FROM A RESEARCHER'S PERSPECTIVE

According to Arend Hintze, assistant professor for integrative biology and computer science and engineering at Michigan State University, "The field of AI involves many things, starting from developing 'smart' algorithms (A-star is probably the best algorithmic example), and includes classification, clustering, optimization, and machine learning. As a CS (computer science) discipline, the focus is typically on the engineering side of solutions and applications, and not so much on basic research as we know it from other experimental sciences, even though engineers apply the scientific method. In other words, in CS, it is not so important to find out how something works in nature but that the code or algorithm developed does what it is supposed to do.

"At the beginning of CS, researchers were convinced they could bring about speaking and thinking machines rather quickly, but in the 1980s, we saw a depression of funding in AI research. As it turns out, AI is much more complicated than assumed by early researchers. This also led to a more pragmatic application of AI tools, and we now find that mainstream AI research is mostly concerned with classifier systems that are now very powerful due to the advances in deep learning and improved computational

[1] John McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," August 31, 1955. *AI Magazine* 27, no. 4 (2006): 12.

> infrastructure. Business applications of this kind of technology are numerous, and I think we will see a lot of this technology become mainstream (self-driving cars are just one example)."
>
> As a researcher, Hintze questions whether the field is making progress on the original idea of creating general AI machines that are indistinguishable from humans. He believes that "we have neuro-evolution as a research topic within the field of AI, but only very few researchers such as myself work in this field. In neuro-evolution, we try to understand how nature brought about natural intelligence and try to recapitulate that process in a virtual environment, evolving AI. Although I am convinced that this is the only viable path to humanlike intelligent machines, I don't think this approach will replace conventional AI research, which just tries to 'computationalize' specific tasks and might not even seek to understand intelligence."

Newer Use Cases Made Possible by These Technologies

Later in this report, you will see that many of the use cases for machine learning, NLP, and AI are use cases that are already familiar in the market, such as churn, fraud, customer behavior analysis, recommendation engines, predicting probabilities of events, text analytics, and social media analytics. However, some newer use cases are worth mentioning as well.

Operations is often a popular area for machine learning.

Preventive maintenance in asset management. Many respondents in this survey mentioned predictive or preventive maintenance as a use case for machine learning. Some of these examples make use of sensor data from the Internet of Things (IoT).[2] For instance, a fleet operator might use sensors to collect data from their various trucks. Such data might include the temperature or number of vibrations per second of a particular part or parts. This data can then be analyzed using machine learning to determine what precipitates a part failure or when undue wear and tear is occurring. The system "learns" the patterns that constitute the need for repair. That information might then be encoded into a set of rules or a model and finally used to score new data from trucks in order to improve fleet maintenance and operational efficiency.

Chatbots in customer experience. Chatbots have been in the market for a number of years, but the newer ones have a better understanding of language and are more interactive. Some organizations use chatbots to answer routine questions in help desks. Some use bots to help in routing help desk questions. Here, based on who you are (e.g., whether you have status with the company) and what you asked for (using NLP for text analysis), you will be routed to the right call-center person to answer your specific questions. Other companies use it for personalized shopping that involves understanding what you and people similar to you bought, in addition to what you are searching for. These use cases require smart NLP-based search as well as machine learning.

Text summarization. Text summarization is the automatic process of shortening a document to generate a summary of the major points of the original text document. There are multiple use cases for this, from distilling down the major takeaways from text and monitoring new documents for research needs, to summarizing what people are emailing to your company. Multiple approaches from NLP and machine learning/deep learning are being used and tested for text summarization. Most experts would say these algorithms still need work but the field is moving forward.

Classifying image and sound. Using deep learning, a system can be trained to recognize images and sounds. The systems learn from examples that are labeled in order to become accurate in classifying new images or sounds. For instance, a computer can be trained to identify certain sounds that indicate that a motor is failing. This kind of application is being used in cars and aviation. Though we think about auto tagging images in social networks, this kind of technology

[2] For more information on IoT, see *TDWI IoT Readiness Guide* (2016), online at tdwi.org/iot-assessment.

is also being used to classify photos in business for online auto sales or for identifying other products. A photo of an object to be sold in an online auction can be automatically labeled, for example. Image recognition is being used in medicine to classify mammograms as potentially cancerous and in genomics to understand disease markers. Vendors such as Google and others provide image recognition APIs for building these kinds of applications.

Smart cars. Who hasn't heard of self-driving cars? These cars use multiple internal and external sensors (cameras, LIDAR, sound, GPS) to do things such as detect pedestrians and other objects on the road by learning how to recognize them.

Smart analytics applications. Most people are familiar with interactive search applications such as Siri or Alexa. However, more frequently, other kinds of applications are using some sort of natural language interface. This includes BI applications where users can ask questions, either by voice or by text, in a natural language fashion and get answers back from the application. Additionally, machine learning is being used inside of data management and BI applications to help with everything from data integration to data preparation to the actual analytics analysis. Some applications provide users with insights that they may not have thought to look for on their own. Several of the sponsors for this report are providing this kind of functionality.

B2B applications. In line with the example above, natural-language-based apps are increasingly becoming part of B2B applications. For instance, some companies embed these apps into portals so partners who might not be analytically well-informed can easily get at insights. Others might build a conversational application.

Other use cases. These include translation, drug discovery, cognitive tutors, and cybersecurity.

USER STORY DETECTING EMOTIONS USING VIDEO AND ADVANCED ANALYTICS

Will people like a new movie or TV show? Do their facial expressions mean they are lying? These are the kinds of questions that SilverLogic Labs, an emotion recognition company, helps answer. According to Michael Lisin, the company's VP of data science, a combination of open source, commercial, and proprietary algorithms are used to read emotional responses. "We identify facial features and then track those features. For instance, we know you're smiling when we see that the end of the mouth is moving, an eye is squinted, and cheekbone is up." In all, the company can track seven emotions: anger, fear, disgust, happiness, sadness, contempt, and surprise.

"When experts watch focus groups and decide if a TV show or movie is going to be successful, that is gut feel. Gut feel can be right or wrong; there is no science involved. Additionally, focus groups are small. An expert may be able to pick up on a single person's emotions but not audience emotions in combination or at scale."

Although the emerging business of "marketing emotions" has played a key role in marketing for a decade, SilverLogic has taken it to the next level. This is not Big Brother watching. Content creators target specific demographics by obtaining viewers' consent to be filmed while watching prerelease television and movie content. Digitized images are then passed through SilverLogic's emotion recognition engine, which utilizes a combination of deep learning and machine learning to quantify emotions at the frame level. According to Lisin, it takes numerous audiences and participants to train the system to accurately quantify emotions. The models are continuously updated and retrained to improve prediction performance. In all, the company is dealing with 10,000 users and millions of facial expressions. SilverLogic uses a hybrid data management approach; some data is in the cloud, some on premises.

> Lisin recommends that those getting started with learning complex computer technologies should be "ready to enjoy a roller coaster of learning and challenges." The skills required to be successful include software development, data manipulation (data wrangling), math and statistics, and substantive expertise.[3] In addition, to give your work focus, it is important to start your data science discoveries with a question (such as the one at the beginning of this user story) that your business needs to answer.

The Current State of Machine Learning, NLP, and AI

To understand the current state of machine learning, NLP, and AI, we asked respondents if they use these technologies and how they use them. If they are not using them currently but plan to do so, we asked when they planned to do this. For those who are not using the technologies and have no plans to do so, we asked, "Why not?"

Survey results indicate that 42% of respondents are planning to use these technologies.

These technologies, particularly machine learning and NLP, are making their way into organizations. Forty-two percent of respondents to this survey are planning to use the technologies but are not doing so now. When this group is called out specifically in this report, it will be referred to as the "investigating group." Slightly more than 43% claim to already be using at least one of these technologies; this group will be referred to as the "active group." The remainder of the respondents (about 15%) are not using any of these technologies now or don't know if they are using them (all not shown).

Of the 43% that claim to use at least one of the technologies, about three-quarters use machine learning. This jibes with what we see in other research, where typically about 30%–40% of respondents are making use of some sort of predictive analytics technology. Likewise, about 50% of respondents are using NLP technologies, or about 22% of total respondents. TDWI has not captured data for AI-related technologies in the past. In this survey, about 12%–15% of respondents claimed to be using AI technologies (all not shown).

Why Not Use the Technologies?

As stated above, a small group of respondents (about 15%) is not making use of the technologies, which may be due to the nature of the survey topic. Respondents tend to gravitate toward familiar topics, and though we stated in the invitation for the survey that we want input from everyone, this may explain the relatively low percentage found here.

For those who do not use machine learning, NLP, or AI, the top three reasons include no support, no budget, and no skills. Nonusers believe in the value of these advanced analytics but cannot seem to get buy-in to the process. Oftentimes, this is because they lack an analytics culture. In fact, TDWI frequently sees cultural issues playing heavily into analytics challenges. These issues can stop any analytics program in its tracks. Only a handful of respondents said they currently had no need to utilize these technologies.

No support, no budget, and no skills are the top three reasons organizations don't use advanced analytics technologies.

Importantly, when asked about satisfaction with analytics deployments, 63% of respondents in the nonuser group stated they are not satisfied with the analytics in place at their organization today. Thirty-five percent said it is OK for now. These responses seem to suggest that these businesses would like to do something more, but organizational issues are preventing them from doing so, especially because the vast majority of this group (approximately 80%; not shown) believe there is value in these technologies. About half hope that there might be an analytics culture in place over the next two to three years to make this happen; the other half don't know if that will ever happen (all not shown).

[3] See data scientist Drew Conway's Venn diagram at http://bit.ly/2h2y3Ge.

Budget and Deployment Plans

There is no doubt that the market is excited about machine learning, NLP, and AI. These are the buzzwords du jour. Additionally, as described later in this report, these technologies can provide significant value to organizations fortunate enough to use them.

For those planning to deploy these technologies, close to 60% believe they will have the funding in the next 18 months to do so (Figure 2). Close to 10% believe they will have it in the next six months. If users can stick to their plans, that is a market on the move.

Note the use of the word *can* here. Organizations tend to believe they will get funding faster than it happens in reality—budget is often an obstacle for any kind of analytics project. It takes time to convince sponsors of the value of a technology and put a plan in place. It takes time to build trust. For instance, TDWI has been tracking the use of predictive analytics in organizations for years now. If users were able to stick to their plans, more than 75% of organizations would be making use of the technology now. In reality, we see about a third (as stated above) making use of it. In fact, even in this study, close to a quarter of respondents felt that lack of business sponsorship might be a barrier to implementing these technologies (not shown).

If users were able to stick to their plans, more than 75% of organizations would be making use of predictive analytics now.

When do you expect to have budget to implement any of these technologies?

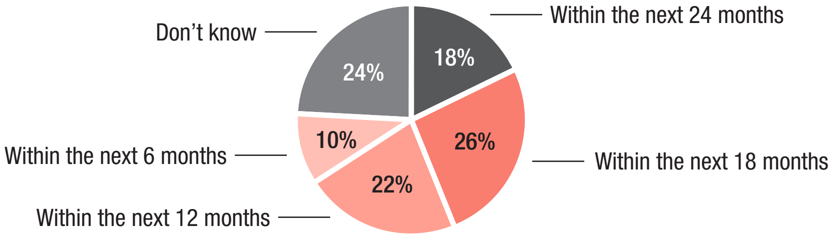


Figure 2. *Based on 122 respondents in the investigating group.*

A key is converting the ideas and plans to reality, which will require careful preparation as well as organizational and technology change. New skills will be required along with potential changes in leadership. Some executives are excited about the technology and have green-lighted their teams, but others need to be sold on its value. The data presented in this research will be helpful in making the case. This report also provides success strategies from respondents already engaged in deployments for those planning to utilize these technologies.

Big Plans Ahead for Some Organizations

Planned use cases fall into a number of categories including customer focused, operations focused, and unfocused.

As with those not using machine learning, NLP, or AI, many in the investigating group are not satisfied with what they have in place today. In fact, 67% of the investigating group are not satisfied with their company's current analytics program and are planning for these newer technologies (not shown). Many of the use cases under discussion are traditional, while others are newer. Technology plans appear to involve the use of open source technologies in conjunction with newer platform options. Of course, whether those plans succeed for all of these respondents remains to be seen.

Business Problems Being Addressed

In an open-ended question, we asked those investigating these advanced technologies about their current plans for the technologies. Their responses fell naturally into a number of categories:

- **Customer focused:** There were many marketing-related applications mentioned in the survey including better understanding of market segments, customers, and the products customers want; building recommendation engines; predicting churn; and personalizing marketing to better engage customers. Many respondents also spoke about improving the customer journey.
- **Person focused:** On the healthcare front, respondents mentioned use cases such as predicting clinical support needs, patient-missed appointments, and patient risk management. Some use cases were research focused, including cancer and pathology research, with unstructured data such as images. Similarly, on the student front, use cases included predicting retention.
- **Operations focused:** There were numerous survey responses focused on being able to catch issues before they occur, proactively identify machine problems, and become more efficient with plant maintenance, supply chain scheduling, and so on.
- **Fraud and risk focused:** There were also numerous responses about identifying fraud in different kinds of data, such as claims data. Risk-focused use cases often involved insurance pricing.
- **Vague and not focused:** Many responses were vague in nature, either because respondents did not want to disclose what they were planning or because they simply did not yet know. Leaders at some organizations think they want to do more advanced analytics but do not necessarily know what questions they need to answer. That can make it harder to get funding. Although it is fine to explore data and experiment with it, it is important to know the questions you are interested in answering and what is important to the business. In fact, we will see later in this report that determining these parameters is an important organizational success strategy.

If these use cases sound like established ones for more advanced analytics, it is because many are. What is different will be some of the technologies used. Additionally, though marketing was often the top area for early adopters in predictive analytics, operations is important here, too. In fact, we find that operations and marketing are two of the most popular areas for the active group (see Figure 5, page 17).

Open Source vs. Commercial Software

Open source R scores second on the list of tooling used for machine learning, NLP, or AI.

Open source provides a low-cost source community for innovation utilized by many data scientists and analytics application developers. The open source model is a collaborative development model where code is made available for free, and the copyright holder has the rights to study, change, or distribute the code. Open source has become quite popular, especially for big data and data science.

We asked those planning to use these technologies to rate a series of technologies for advanced analytics on a scale from 1 to 5, where 1 was "not at all important" and 5 was "very important." Figure 3 illustrates the percent of respondents who rated these technologies as important or very important.

Please rate the following technologies in terms of importance in meeting your future goals in ML, NLP, and/or AI on a scale from 1 to 5, where 5 is very important.

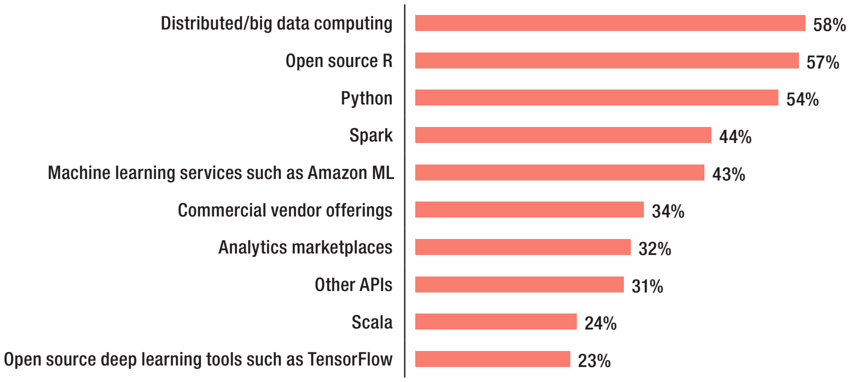


Figure 3. *Based on 119 respondents in the investigating group. Percent ranked as "important" or "very important."*

- **R scores near the top.** R is a language and environment for statistical analysis that is part of the GNU free software/open source project. It includes data handling and storage facilities, a large set of tools for data analysis (including machine learning and NLP), tools for graphical analysis, and a programming environment. Developed in the 1990s and similar to the earlier S language, R is widely used as a statistical environment by universities and organizations. There is huge community support for R. In this survey, 26% ranked R as very important (not shown).

  R includes algorithms for machine learning as well as NLP. Recently, vendors have begun to support R either in a data-science workbench environment or as part of full-life-cycle products. If the vendor has a drag-and-drop visual interface, it will often let a user connect to a model developed in R that way, as long as R is already installed on the machine.

- **Python is popular.** Python is an interpreted, interactive, object-oriented scripting language now available through the Python Foundation. It was created in the 1990s to be an easy-to-read language and has a library for analytics (as does R). Developers appreciate its flexibility and simplicity. Hadoop is written in Java, but MapReduce applications can be written in Python. Python supports other Hadoop ecosystem projects such as Spark, Storm, Hive, and HBase. Python is often used in developing Web applications.

For Python's analytics libraries, NumPy provides foundational libraries for manipulating data. SciPy builds on this with some algorithms, including optimization. Scikit-learn builds on these further to include machine learning and data mining tasks such as clustering, regression, and classification. NLTK is a set of libraries designed for natural language processing. As with R, many vendors provide interfaces to models developed in Python.

- **Commercial tools score lower than expected.** Although over 30% of the investigating group ranked commercial tools to be important or very important, less than half ranked them as very important to machine learning, NLP, and AI efforts. This highlights the popularity and hype around open source tools as well as the vast community of support. As one respondent stated, "It is easier to find people who know open source rather than commercial tools."

  Interestingly, the active group typically uses a combination of approaches, including commercial software and open source tools, in more advanced deployments. For instance, some organizations will use open source to build models but then put the models into production using commercial tools—often embedding these models in the commercial environments. This is discussed more fully in the next section.

USER STORY USING OPEN SOURCE MACHINE LEARNING TO BUILD INSURANCE RISK MODELS

"In insurance, it is all about how you price risks," said Sai Giridharan, director of enterprise analytics at property and casualty insurance provider Heritage Insurance. "If you write policies, you don't expect losses on all policies. Historically, this risk analysis was done at an aggregate level. We are going to look at a granular level, using boosting algorithms and random forests across hundreds of attributes and using historical data going back 10 to 15 years. We'll look at all of the variable interactions that contribute to loss and look at the probabilities of loss costs associated with different customers." Giridharan plans to operationalize the model, running customers and potential customers through the algorithm to predict the loss cost for each to determine whether the risk is adequately priced or not.

The model itself will be built in R. Giridharan's strategy is to refine the model in R and then get buy-in into the models. "Most likely we will use a commercial platform to put the model into production. The limitation with R is creating a runtime to integrate with the production systems on a real-time basis." He added that "the building of the model is going to be an iterative process, and by the time we think we will be ready for deployment, it is most likely going to be toward the end of next year."

The company is careful about the models it puts into production. Data scientists and technologists will be building the models, but actuaries will validate it. This is one control that will be put in place. Giridharan also believes strongly in monitoring models for degradation. "No model is perfect forever. As models start to deviate, we'll have to get back on board to refine them." Meanwhile, the models are monitored through dashboards.

Platform Plans

The plans for machine learning, NLP, and AI make use of new kinds of platforms. We asked the investigating group what kind of platforms they are already using for analytics and what they plan to deploy. Their answers fell into broad buckets, which are illustrated in Figure 4.

Of those investigating advanced analytics technologies, 48% plan to use a cloud data warehouse in the next two years.

The on-premises data warehouse rules, but the cloud is coming on strong. As TDWI has seen in other research, the vast majority (79%) of respondents in this survey are already using a data warehouse on premises to support their analytics efforts. Those investigating the technology plan to use their data warehouse for advanced analytics plans; however, the majority of this group either know they may have to extend their platform or are already using what TDWI terms a *multiplatform data environment* that includes the cloud, Hadoop, and other platforms.

For instance, it is interesting that while 23% of the investigating group are using a cloud data warehouse today, approximately 48% more plan to use one in the next two years—and this number is higher for those who believe they have the budget for it in the next 18 months (not shown). This makes sense because the cloud is scalable and flexible for big data projects, and many machine learning and AI projects fall into the big data category. More often, data is generated in the cloud and it makes sense to analyze it there. In addition to the cloud data warehouse, 29% plan to use Hadoop in the cloud and 28% plan to use data appliances in the cloud in the next two years. We do not expect all of the respondents to take action on their plans, but cloud usage will no doubt grow.

Interest is growing in data lakes. As Philip Russom describes in the recent *TDWI Best Practices Report: Data Lakes*,[4] a data lake is a collection of often diverse data that can scale to tens or hundreds of petabytes in size. These lakes are also becoming part of a modern multiplatform environment. Although the warehouse is used for reporting and some other kinds of BI, these other platforms are more often used for advanced analytics. In this survey, while 23% are using data lakes currently, another 47% plan to make use of data lakes in the next few years. As Figure 4 illustrates, this is one of the top areas for growth in terms of platforms supporting these advanced analytics technologies. Data scientists frequently see the data lake as a place to experiment with data.

Newer technologies are making an impact. As shown in Figure 4, a group of technologies is moving into the early mainstream stage of adoption (at least among our respondents) that includes data appliances, columnar databases, Hadoop, NoSQL DBMS, and cloud data warehouses. All of these are being used by about 20%–25% of respondents, with additional plans for growth (if users can stick to their plans). This is significant and reinforces the idea that it is not just one kind of data platform that will be used for analytics; rather, analytics will be carried out across a multiplatform data environment where the right tool is used for the right job.

Emerging technologies stand to grow. In addition, a group of technologies that includes graph databases and streaming platforms is emerging and attracting the interest of organizations. For instance, streaming platforms allow organizations to deal with real-time data for analytics such as IoT, cybersecurity, and recommendations by making use of machine learning to detect patterns and anomalies and act on them in real time. Organizations are also becoming more interested in graph databases, which use graph theory to relate data in the store. These databases are useful for a multitude of business problems ranging from fraud to IoT.

[4] For more information, the complete *TDWI Best Practices Report: Data Lakes* is online at tdwi.org/bpreports.

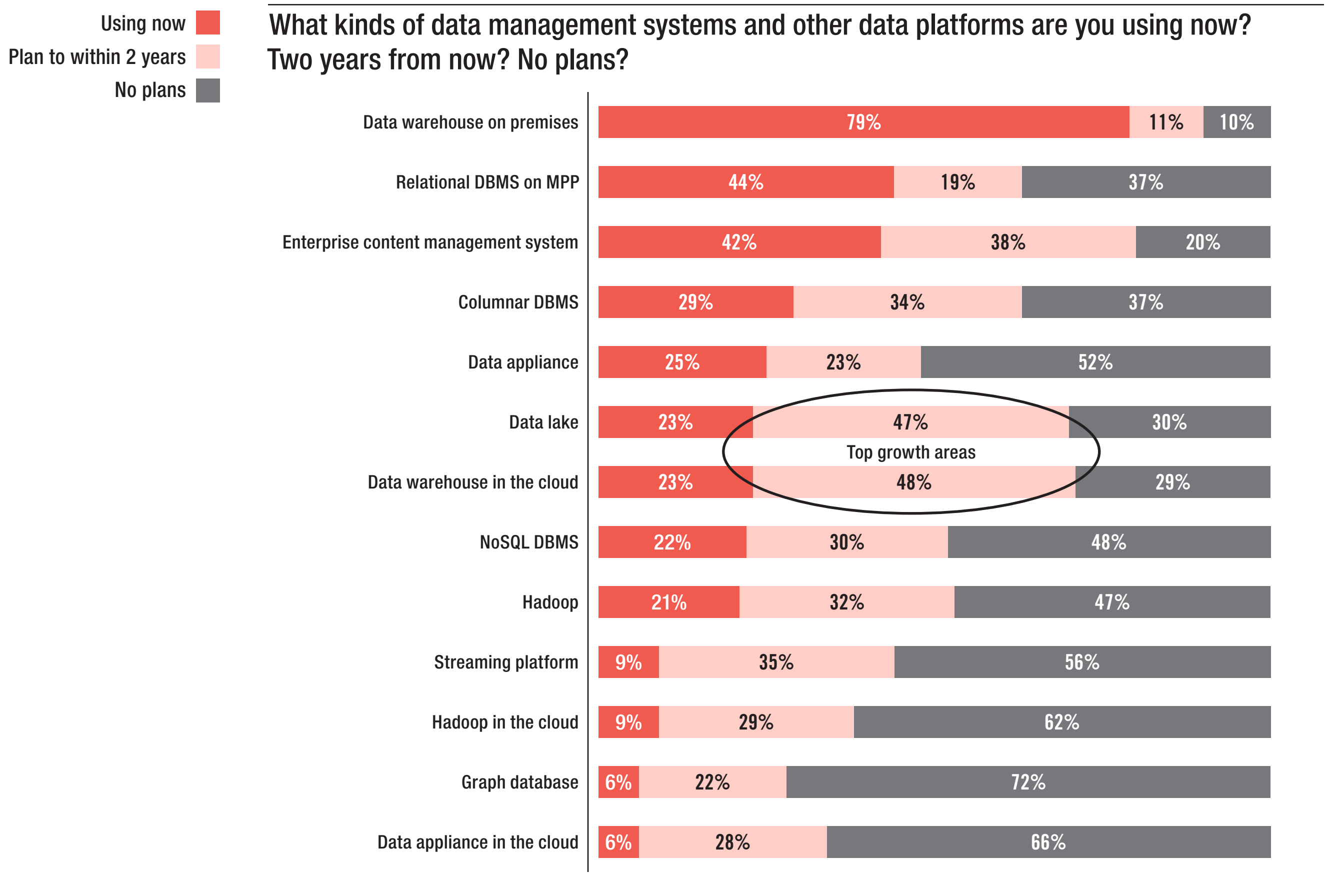


Figure 4. *Based on 123 respondents in the investigating group.*

Early Adopter Experience

The investigating group is beginning to develop some solid plans for machine learning, NLP, and AI. However, exploring the responses from the active group provides insights that can help those investigating the technologies learn more. These advanced analytics technologies are being used by early adopters across the organization for a range of use cases. Operations is the leader in the active group, with 58% of respondents using the technologies in operations. Fifty-three percent are using the technologies in marketing, which is often an early adopter of more advanced analytics. Some of the use cases are illustrated in Figure 5.

Area	Percent using technologies and case examples
Operations	58% • Predictive maintenance • Equipment optimization • Process optimization, including supply chain • Predicting fraud, waste, abuse
Marketing	53% • Social media and sentiment analysis • Churn analysis and other customer behaviors • Proactive intervention
IT	48% • Equipment failure prediction • Security threat detection
Customer service	42% • Helpbots, chatbots • Product feedback analysis based on text • Post-purchase interaction analysis
Sales	31% • Sales prediction • Product mix prediction
Finance	27% • Credit scoring • Financial transaction categorization
E-commerce	18% • Chatbots to suggest products on websites • Classifying products by images • Automated categorization of URLs
HR	15% • Employee churn • Internal HR bots • Analyzing internal communications

Figure 5. *Percent based on 130 respondents in the active group.*

USER STORY PINNACLE INSURANCE EXPLORES MACHINE LEARNING AND NLP TO DRIVE EFFICIENCIES

"We recently started to do more in terms of machine learning, NLP, and AI," said David Sauther, BI director at Pinnacle Insurance, a workers' compensation company in Colorado. To do more required hiring several data scientists at the beginning of the year. They are building NLP pipelines in Python for email and voicemail so that the company can better understand why people are contacting them.

The company is also exploring other use cases. "We already have predictive models that do claim routing," said Sauther. "We are looking at areas where we can automate tedious work, such as understanding what a business does from the information, including images, on its website."

Sauther said his organization is lucky because company executives are excited about increasing efficiencies using some of these advanced technologies. Nevertheless, he advises organizations to start small. "We are realizing little victories that help us understand how the technologies work. We manage expectations by using small sprint processes that can add value and get work done, as well as keep people from becoming dispirited by unrealistic goals. My organization sends out emails every few weeks to executives describing what they have done and to give them something to look at."

Open Source Products Used

The majority of active group respondents use both commercial and open source platforms.

Whereas the majority of investigating group respondents believe they will be using open source technologies to accomplish their plans for machine learning, NLP, or AI, those in the active group use commercial tools in addition to a strong use of open source tools. In fact, the vast majority of active group respondents (more than 90%, not shown) use *both* commercial and open source tools/products. Many use open source *within* a commercial product.

Active group respondents are big believers in open source, too. Some users like the fact that open source is freely available. One respondent, for instance, mentioned that his group found an open source project developed by someone at a university that fit their problem perfectly. They contacted the author and were able to make use of the code. This saved them time and money. Others cited the fact that people like to use what they are trained in and do not necessarily want to learn new tools. Many newly minted data scientists learn open source in universities.

On the open source front, R and Python again lead the way (see Figure 6). However, other open source tools are also important to the active group, including:

- **Apache Spark libraries.** Spark is an open source big data processing framework that is part of the Apache project. The framework provides processing capabilities for multiple kinds of big data (text, graph, streaming). Spark also offers analytics libraries, including a machine learning library. It theoretically is able to process data faster than MapReduce because it processes data in-memory, while MapReduce persists back to the disk after a map or reduce action. This is helpful for iterative analysis. Forty-one percent of the active group are making use of Spark for their machine learning, NLP, and AI projects.
- **Apache OpenNLP.** OpenNLP is a machine learning toolkit for natural language processing. It supports NLP tasks such as tokenization (a subset of parsing, where parts of sentences are broken up into pieces for further processes), sentence segmentation (dividing text into sentences), part-of-speech tagging (nouns, verbs, etc.), named entity extraction (e.g., persons, organizations), parsing, language detection, and others. Twenty-three percent of the active group are making use of these technologies for natural language processing.

- **Scala libraries.** Scala is a concise general programming language created in the early 2000s to address some of the criticisms of Java. It includes libraries such as Breeze for machine learning and Epic for text parsing. Recently, it has been gaining more of a following. Among the active users, 18% are already using Scala.
- **Weka and Knime.** Active group respondents are making use of Weka and Knime (about 7% for each). Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning algorithms developed in the 1990s at the University of Waikato in New Zealand and licensed under the GNU General Public License. It is written in Java, contains data preprocessing tools, and includes classification, regression, and clustering algorithms. It also has a GUI-based interface option. Knime, introduced in 2004 at the University of Konstanz, is a commercial open source analytics platform with over 1,000 modules that also includes data preparation features. It is written in Java and based on Eclipse.

Vendors understand the popularity of open source. As mentioned above, some vendors provide commercialized open source solutions as part of their predictive analytics offering—generally in a data-science workbench environment. Others have opened their proprietary GUI-based products to open source. For instance, many vendors provide users with the capability of calling a model built in an open source environment into their tools; a few let users call commercial models from open source tools. Some vendors even provide nodes on their visual GUI interfaces for open source environments such as R or Python. All that is required is that the users load these open source environments onto their machines.[5]

Of course, some data scientists and statisticians don't like to code and prefer the GUIs that come with many commercial packages. Others were trained on commercial packages and want to continue to use them. Still others appreciate the fact that some commercial products are actually platforms that can help with the entire analytics life cycle. The choice of tooling depends on a number of factors including comfort level, skill, features, and affordability. For instance, organizations with numerous business analysts and business users might look for commercial tools that support automated model building.

Which of the following open source tools does your organization currently use for ML, NLP, or AI? Please select all that apply.

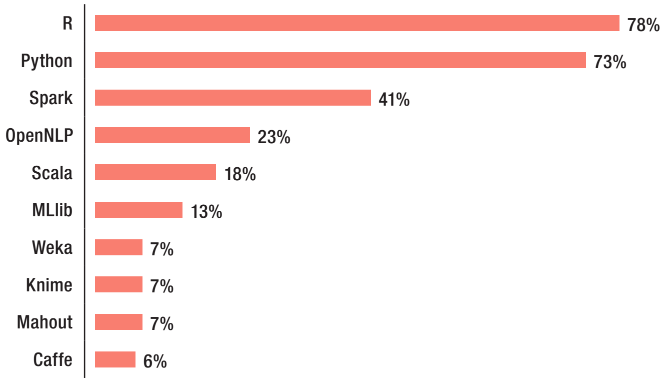


Figure 6. *Based on 130 respondents in the active group. Multiple responses permitted.*

[5] For more information on vendors that support open source models, see *TDWI Navigator Report: Predictive Analytics*, online at tdwi.org/navigator.

Popular Machine Learning Algorithms

Whether you are using commercial or open source products, there are numerous techniques that can be applied for machine learning. We asked those in the active group to identify the kinds of techniques they are currently using (Figure 7).

Decision trees and regression lead the way. Decision trees and linear regression (a statistical approach) were the top two responses for those in the active group. Both methods are fairly straightforward and easy to understand. Linear regression tries to model the relationship between variables by fitting a line to the observed data. This simple model is widely used in statistics. Seventy-nine percent of the active group are using this method.

In the active group, 80% use decision trees.

Decision trees are also popular, especially among nonstatisticians, with 80% of active group respondents using them. A decision tree is a supervised learning approach that uses a branching or a treelike approach to model specific target variables of interest. For instance, an outcome might be leave or stay, buy or not buy. A user would typically provide a set of training data with known outcomes to the tool. The decision tree then builds a model that can be interpreted as a set of probabilities. People like decision trees because they are easy to understand and are not a black box. Both decision trees and regression were also the most common techniques used for predictive analytics in a 2014 survey.[6]

Clustering is also popular. As indicated in Figure 7, clustering was also cited as a popular machine learning approach. Clustering is an unsupervised technique where groupings are based on similarities and the target variable, such as a segment, is not known. In this survey, 64% are using this technique.

Neural networks and Naïve Bayes classifiers are gaining steam. While decision trees, regression, and clustering have been mainstream technologies for many years, it is interesting that neural networks and Naïve Bayes classifiers are gaining momentum. In previous TDWI studies, less than 20% of respondents were using these techniques. Here, greater than 50% are using neural networks.[7] This may be due to the audience, or it may be because those organizations that are using more advanced techniques tend to branch out with the techniques used. This is part of the success cycle.

What machine learning techniques are you using? Please select all that apply.

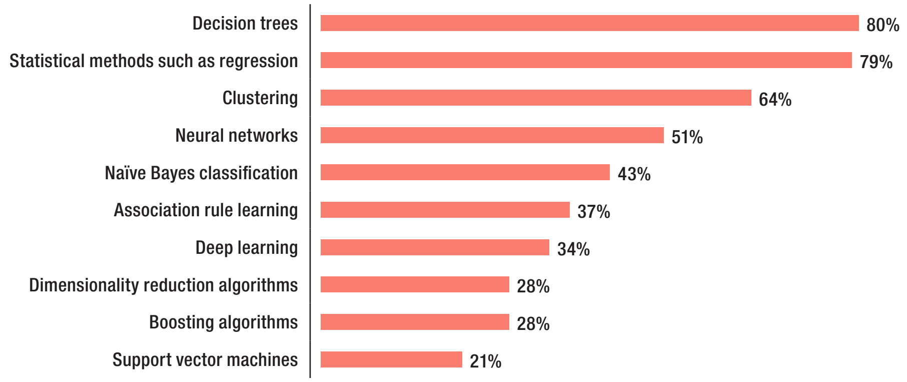


Figure 7. *Based on 90 respondents in the active group. Multiple responses permitted.*

[6] See *TDWI Best Practices Report: Predictive Analytics for Business Advantage*, online at tdwi.org/bpreports.

[7] Ibid.

Neural networks were briefly described earlier in the report. Naïve Bayes classifiers are a family of algorithms based on Bayes' theorem, which looks at the probability of an instance being part of a certain class, where every feature being classified is independent of the value of any other feature. These classifiers are often fast, which is one of their strengths.

Data Platforms Used

As is the case with the investigating group, the active group also realizes that the data warehouse alone will probably not be able to support machine learning, NLP, or AI efforts. About a third state that the data warehouse is working for now, but they will have to extend their data platforms (Figure 8). Forty-one percent say they use a multiplatform data approach now. For instance, some organizations use Hadoop in the cloud while their data warehouse is on premises. Others use platforms from different vendors for their data warehouse, Hadoop, and analytics needs. Only a small group (8%) does not use a data warehouse at all.

Does your data warehouse support your ML, NLP, or AI efforts?

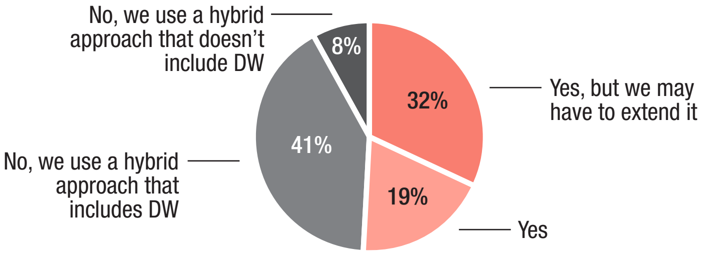


Figure 8. *Based on 113 respondents in the active group.*

A multiplatform data environment makes sense for analytics such as machine learning or deep learning, which often involve large amounts of disparate data. A data warehouse is usually used when the kinds of questions to be asked, as well as the output, are known. In the case of machine learning, NLP, or AI, it is often necessary to experiment with the data or refine models further and further, which can take time and system resources.

Who Is Behind the Curtain?

Data scientists have statisticians and business analysts beat 2-to-1 in terms of who is building models today.

Who is building out the analytics in the multiplatform environments that make use of open source and commercial products? Previous TDWI research indicates a movement by vendors to make their software easy enough to use so that business analysts (aka “citizen” data scientists) can build a predictive model. In fact, we have seen the expectation by respondents in earlier surveys that business analysts, along with statisticians and data scientists, will be the primary builders of predictive analytics models.[8]

In this study, however, data scientists have statisticians and business analysts beat 2-to-1 in terms of who is building models. We asked respondents in the active group who in the organization was building machine learning models. Figure 9 shows that the top answer is data scientists (92%), followed by developers (40%) and statisticians (40%).

A typical organizational model is to hire a few data scientists to become part of the analytics team. Some organizations want to train business analysts from within, while others believe it depends on the business analyst. As one respondent noted, “Some people are not up for it.” In other organizations, the staff on hand is already overwhelmed with what they have on their plates. However, some business analysts do want to build models and may use a variety of tools, including coding and automated model building (see next section), as part of a data science team or in a center of excellence (CoE).

Who is using ML to build models? Please select all that apply.

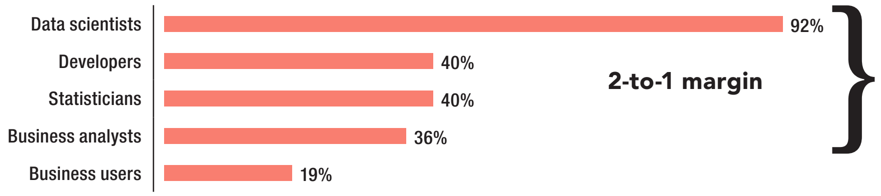


Figure 9. *Based on 130 respondents in the active group.*

Then, too, there is also a move to build applications that embed machine learning and more advanced analytics. This involves development skills that data scientists and developers are more likely to possess. In this survey, 44% (not shown) of the active group were building machine-learning-based applications such as image recognition applications, supply chain optimization applications, or customer-alert-based applications.

A Closer Look: The Impact of AI

As AI begins to ramp up, inevitably there are important questions about privacy, job loss, and what AI means to the economy and humanity’s future. Yes, there are many opportunities for AI in terms of automation, improving operational efficiencies, understanding behaviors, and diagnosis. There is no doubt, however, that AI will displace jobs. A 2013 University of Oxford study, for example, ranked jobs at risk based on the probability of occupations being fully automated.[9] It concluded that workers in transportation and logistics as well as administrative support were at significant risk.

For instance, self-driving cars will displace taxi and limo drivers as well as truck drivers. Text summarization can replace the legal assistants who help lawyers put together briefs.

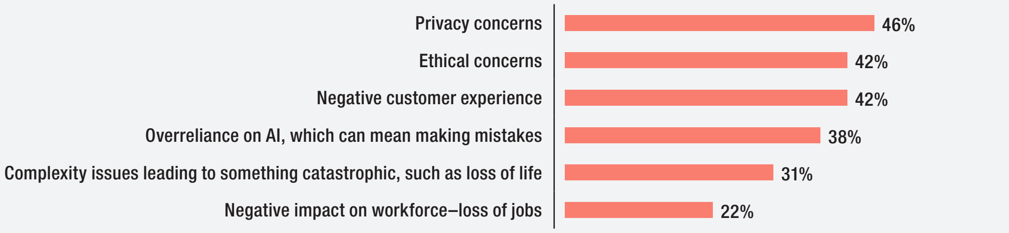
[8] See *TDWI Best Practices Report: Next Generation Analytics for Business Advantage*, online at tdwi.org/bpreports.

[9] See Carl Benedikt Frey and Michael A. Osborne, “The Future of Employment: How Susceptible Are Jobs to Computerisation?,” September 17, 2013. http://bit.ly/2f5fq7J

Automation in manufacturing will displace factory workers. The list is long. Other studies have reached a different conclusion under the assumption that jobs will not be fully automated or that new jobs will grow up around those industries where jobs have been lost due to automation, albeit requiring different skill sets. The debate still rages.

We asked respondents to rate some perceived risks associated with AI on a scale from 1 to 5, where 1 was not at all significant and 5 was very significant. Their responses are illustrated in the figure below.

Please rate the following risks associated with AI on a scale from 1 to 5, where 5 is very significant risk.



Based on 252 responses from the not using group, investigating group, and active group. Responses included "significant" and "very significant."

Big Brother is watching. Privacy concerns ranked highest on the list, with 46% of respondents ranking this either significant or very significant. Privacy issues take many forms—respondents, for example, might be concerned about outside entities gaining access to their private information and using it in unscrupulous ways. This could result in receiving unwanted offers, or an insurance company might determine a premium based on telematics data captured without consent. Other examples include invasive surveillance technology and speeding tickets issued on the results of automated sensor technology.

Ethical concerns. These also ranked high on the list, with 42% of respondents citing them as a risk. Ethical issues in AI run the gamut from unemployment (only 22% of respondents cited this as a concern) to AI making unintended decisions to the rights of AI systems. Interestingly, 38% of respondents felt there was a significant or very significant risk of AI making mistakes because of overreliance on it to perform tasks, such as image recognition or text recognition, where it might not do well in important situations (e.g., diagnosing disease). Slightly less (31%) were concerned about a complex AI encountering a situation it hasn't seen with bad results, such as loss of life.

Although AI is not yet to the point of a doomsday scenario, it is important to address these concerns sooner rather than later. To that end, a number of organizations have already formed to promote safe and responsible AI, including OpenAI, the Machine Intelligence Research Institute (MIRI), the Future of Humanity Institute, and the AI Fund.

Automated Insights for Business Users

An important market trend is the introduction of tools that provide automated/intelligent insights. Our survey results highlight the fact that data scientists are continually at the forefront of using machine learning, AI, NLP, and other advanced technologies to build models and applications. Nevertheless, Figure 9 also illustrates that business analysts and business users are building machine learning models. Some of these business analysts are making use of tools that include advanced analytics under the hood to generate results in a more automated fashion. The goal is to make it easier and faster for nontechnical users to uncover insights. Vendors are providing these tools to help to address some of the issues around lack of people skills cited earlier.

Figure 10 illustrates the percent using these tools in both the active and investigating groups. These tools include:

- **Natural-language-based search.** As mentioned previously, these tools use a natural language experience to search and analyze data to find insights. The search engine learns the user profile, user and group search history, and data characteristics, and then ranks search suggestions that are most relevant to the user. Some of these tools have audio interfaces as well. Less than 5% of respondents are using these tools today, a reflection of their relatively recent introduction to the market.
- **Automated model building.** As previously mentioned, because data scientists and statisticians are often in short supply, many vendors are offering tools that help business analysts and even business users construct predictive models. In some of the tools, all the user needs to do is supply the target or outcome variables of interest, along with the attributes believed to be predictive. The software picks the best model. Some tools even generate derived attributes, such as popular ratios, to use as model input. A number of these tools provide details about the statistics and math used; some do not. In this survey, about 13% of respondents are using automated model building tools, though the number jumps to over 20% (not shown) for the active group.
- **Automated visualization and analytics.** Tools that suggest visualizations to users are an example of automation. Suggestions and guides are included in a number of visualization tools on the market. Over 25% of respondents are currently making use of these tools. Newer tools use machine learning and other techniques to automatically analyze data and provide analyses on what is trending, leading and lagging indicators, outliers, etc. In other words, these tools can provide users with insights to questions they may not have thought to ask. Some describe or generate a narrative about the data/analytics so the user does not have to know how to interpret the data.

Does your organization make use of any intelligent data management or analytics tools?

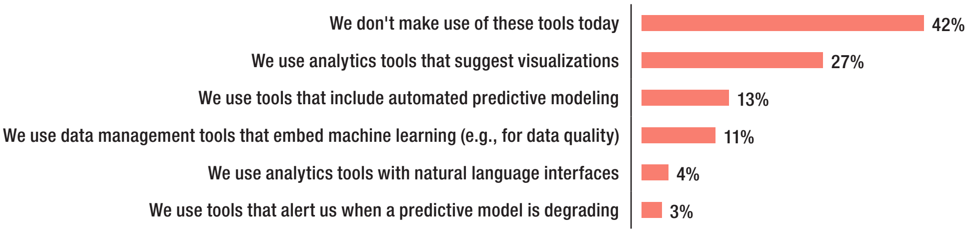


Figure 10. *Based on 222 respondents from both the investigating group and the active group.*

These tools can be a big productivity booster for the organizations that use them. Two caveats here are worth mentioning. First, it is important that these tools be transparent. In other words, they need to provide the user with information about what is going on behind the scenes (e.g., models used, statistical measures, etc.). In this way, business analysts and other users have information about the analyses, to determine if it makes sense. Second, users should have some understanding of exactly what *is* happening behind the scenes. This means getting some training in at least the basics of machine learning techniques. This will help in interpretation and in defending the analysis.

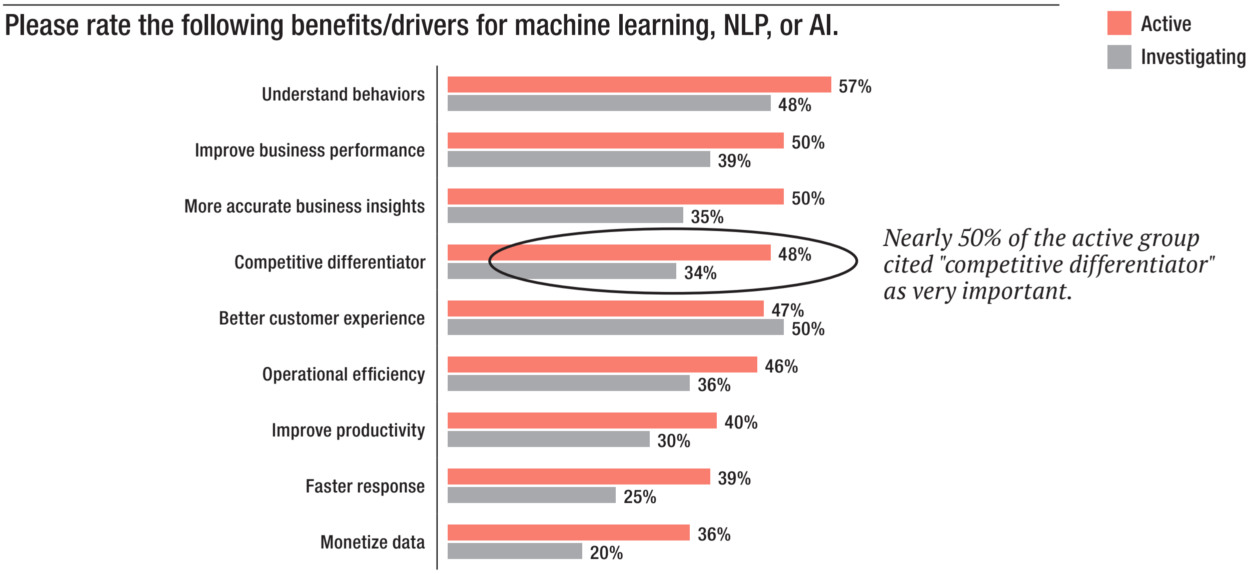
Some organizations put controls in place if business analysts are building models that will be put into production—and this is a good practice. For example, a business analyst who builds a model will need to get sign-off from a data scientist or someone else with model building skills to determine whether or not the model is cleared to be put into production. In this way, the organization can scale the number of people building models, while minimizing potential problems with the analytics down the road, when real money is at stake. Of course, organizations also need to monitor in-production model output because models tend to degrade over time.

Benefits, Challenges, Best Practices, and Value

If you haven't already picked up on this fact, there is a lot of enthusiasm around the benefits of these advanced analytics technologies. Regardless of the group, when respondents were asked whether they think AI opportunities outweigh risks with AI, the overwhelming answer was "yes." Clearly, organizations are bullish on the benefits that AI provides.

Benefits of the Technology

When we asked respondents about the benefits of machine learning, NLP, or AI (Figure 11), they cited the following as very important:


Please rate the following benefits/drivers for machine learning, NLP, or AI.

Nearly 50% of the active group cited "competitive differentiator" as very important.

Figure 11. *Based on 124 respondents from the investigating group and 130 respondents from the active group. Ratings of "very important" shown.*

The biggest benefit that respondents cited was understanding customer behavior.

Customer-focused benefits lead the way. Across both the investigating and active groups, the biggest benefit that respondents cited was understanding behavior. As we have seen throughout this report, machine learning, NLP, and AI can help in a variety of ways to understand behaviors using a variety of data types (57% of active, 48% of investigating). Most of these behaviors are likely people related (customers, patients, students, etc.). Additionally, better customer experience, which is related to customer behavior, was also very important to these respondents (47% active group, 50% investigating group).

Active group more definitive. Interestingly, with the exception of the customer experience choice, active group respondents were more likely to select "very important" for the benefits listed in this question than investigating group respondents. Those in the active group have experience with these technologies and are using them for multiple use cases. They have experienced the benefits and realize how important they are. For instance, active group respondents stated that improved business performance and more accurate business insights were very important benefits (50% for each). They are more definitive in their responses because they are experiencing the benefits.

Close to 50% of the active group cited competitive differentiator as a very important benefit.

Advanced analytics as a competitive differentiator. Competitive differentiation and monetizing analytics are also very important for the active group. Close to 50% of active group respondents cited competitive differentiator as an important benefit, whereas among the investigating group, this was one of the lower-rated drivers. Likewise, almost twice as many of those in the active group cited monetizing data as an important differentiator compared to those in the investigating group. The early adopters understand the value of these technologies in helping them to compete and grow revenue. This should be a wake-up call to others.

Overcoming Challenges

Though the benefits and the value of advanced technologies should help sell projects, there are still numerous challenges to getting started. Those in the investigating group cite (all not shown) finding skills (45%), executive support (32%), and cost (30%) as top barriers to implementing machine learning, NLP, and AI. Most likely, a large percent of the active group faced similar challenges when getting started. In an open-ended question, we asked those in the active group how they overcame some of these challenges. Their responses include:

- **Building proof of concepts and taking small steps.** Some respondents spoke about "winning the battle rather than the war." Others look to show business value with "small, quick wins."
- **Change management, communicating, and building trust.** Some respondents cited that change management approaches are required along with patience and persistence. One respondent in the manufacturing space said, "We spend a significant amount of time educating business leaders in the benefits and basic methods in machine learning to get their support in trying new approaches and gaining trust." Others make a big effort to get stakeholders involved and educate them, which often involves many people in the loop.
- **Executive support.** Some organizations are lucky enough to have an executive go to bat for them and "run interference." This can help to move projects along more quickly.

- **Build skills in numerous ways.** Skills are obviously key to success. Online training to help build skills is a popular method used by some early adopters. Others hold onsite training taught either by an internal group or by a third party. Some send their employees to conferences and other external events. Still others supplement this with lunch-and-learns or internal Kaggle-based competitions. As mentioned above, even if a company is automating the generation of insights, it is still important for people to understand, even at a basic level, the techniques behind the tools.

In addition to overcoming organizational hurdles, those in the active group cited a series of technical issues that others should keep in mind. These challenges primarily involve technical data issues (no figure is shown for results in this section):

- **Noisy and dirty data is a problem.** We asked respondents about the top three challenges their organization faced in implementing machine learning, NLP, and/or AI. More than half (52%) of respondents cited noisy and dirty data as a problem. Contrary to popular belief, data scientists do like clean data. Big, disparate data types can be noisy; consider social media data, for instance. Even structured data can be dirty. Business analysts and data scientists still spend a vast majority of their time cleaning up data and transforming it. Some analysts and data scientists are OK with that because it brings them closer to their data and helps them gain a better understanding of it.

  Respondents also mentioned that integrating data was a challenge. Fifty percent cited this as a top challenge, which makes sense. As data becomes disparate and comes from more sources, it can become harder to integrate. Other issues include trust (37%) and data access (37%). Oftentimes, even if an organization is already performing statistical analysis such as regression, there is still a trust curve that needs to be scaled to get people comfortable with tools such as machine learning. People often don't trust what they don't understand, and many people don't understand advanced analytics.

- **NLP challenges are centered around building taxonomies and training systems.** Many of those in the active group utilizing NLP (a small group of about 60 respondents) are using it for text analytics (81%) and social media analytics (46%). Some are using it for building chatbots to interact with customers (40%), smart assistants (23%), or classifying documents (30%). The biggest problems they tend to face are around building taxonomies and training systems.

  A taxonomy is a method for organizing information into hierarchical relationships. This is important in NLP, especially in dealing with specific vocabularies in certain industries. For instance, some organizations need to create specific taxonomies about products and services or about certain kinds of diseases. Some vendors will provide baseline taxonomies out of the box, but do not expect that they will work out of the box.

  Machine learning is used in text processing to identify certain parts of speech or sentiments or entities. Tagged documents are used for training. Respondents cited that training these models can take more time than they would like.

Best Practices for Advanced Analytics

Understanding business objectives was the top organizational strategy for success.

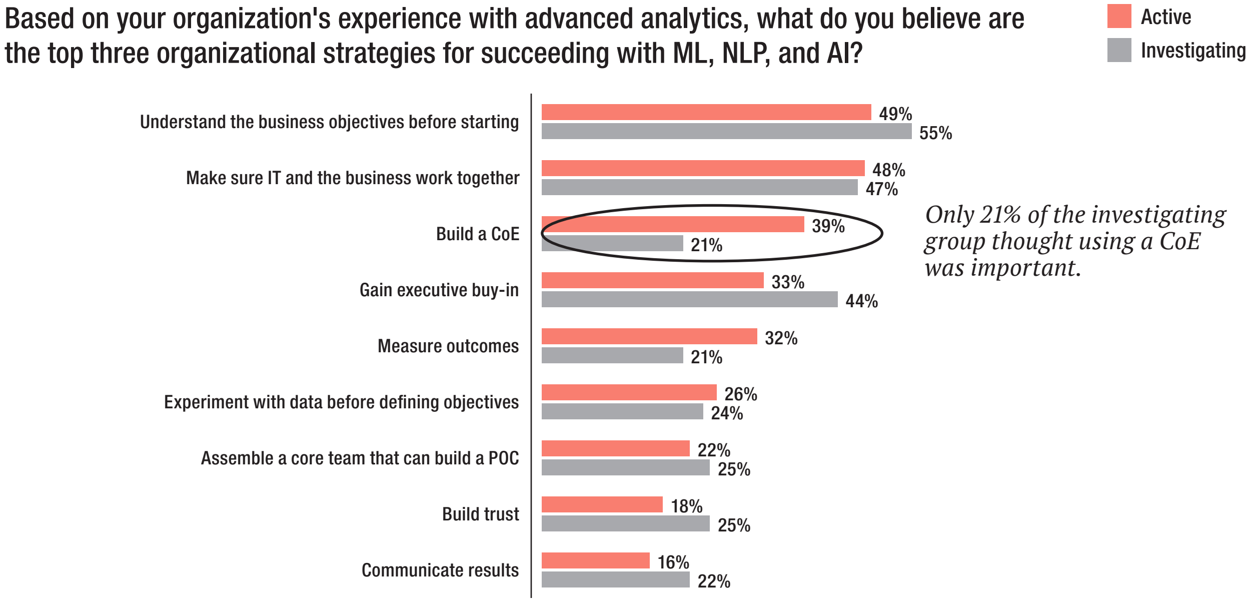
We have already provided some strategies organizations use to get their more advanced analytics projects off the ground. In the survey for this report, we also asked a series of questions about organizational strategies, talent building strategies, technology strategies, and leadership strategies to try to get a handle on what some of the best practices are for those organizations that have succeeded with machine learning, NLP, and AI. As part of the analysis, we compared the active group and the investigating group across a number of responses to look for differences. Additionally, we examined a small group of respondents (less than 50) who had measured a positive top- or bottom-line impact to see how their responses compared. Some interesting findings emerged by examining the data this way.

Organizational best practices. Understanding business objectives (Figure 12) was the top organizational strategy for success across all groups, followed closely by IT and business working together. These best practices are well-known in analytics circles. For instance, in previous TDWI research, we have seen that poor definition of project objectives is one of the chief obstacles to reducing time to value.[10] Though business analysts and data scientists like to experiment with data and it is important to do so, it is still important to have some project objectives in mind. Otherwise, there is a greater chance the project will fail. Likewise, it is important to get business stakeholders involved early on so they can "buy in" to the project. Advanced analytics efforts conceived and executed exclusively within one department without collaboration across business and IT do not tend to have the impact needed to survive, let alone thrive.

Thirty-nine percent of the active group felt the CoE to be an important success strategy.

Both the investigating and active group cited the same top two organizational best practices. However, there were some interesting differences between the investigating group and the active group in terms of organizational strategy—probably a result of analytics maturity. One of the most noticeable was the use of a center of excellence (CoE). Thirty-nine percent of the active group found this to be an important success strategy, which put the CoE in the top three organizational strategies for advanced analytics. Only 21% percent of the investigating group thought this was important—a statistically significant difference. Many of the investigating group likely do not yet have a CoE, but TDWI research indicates that a using one can provide value in numerous ways. A CoE typically consists of a cross-functional group that provides leadership in advanced analytics. In addition to building and deploying analytics, CoE teams are often responsible for training and disseminating best practices.

[10] See *TDWI Best Practices Report: Accelerating the Path to Value with BI and Analytics*, online at tdwi.org/bpreports.



Based on your organization's experience with advanced analytics, what do you believe are the top three organizational strategies for succeeding with ML, NLP, and AI?

Only 21% of the investigating group thought using a CoE was important.

Figure 12. *Based on 116 respondents from the active group and 114 respondents from the investigating group.*

Technology best practices. The active group ranked operationalizing analytics as the top technology strategy (61%, Figure 13), while the investigating group ranked putting an architecture together at the top. Operationalizing analytics is key to analytics success. As Thomas Edison said, "The value of an idea lies in the using of it."[11] It is one thing to create a great model, but the value comes when it is put into action as part of a business process, into a system, or into an application. In fact, for those who measured value, 69% ranked this as the top technology strategy.

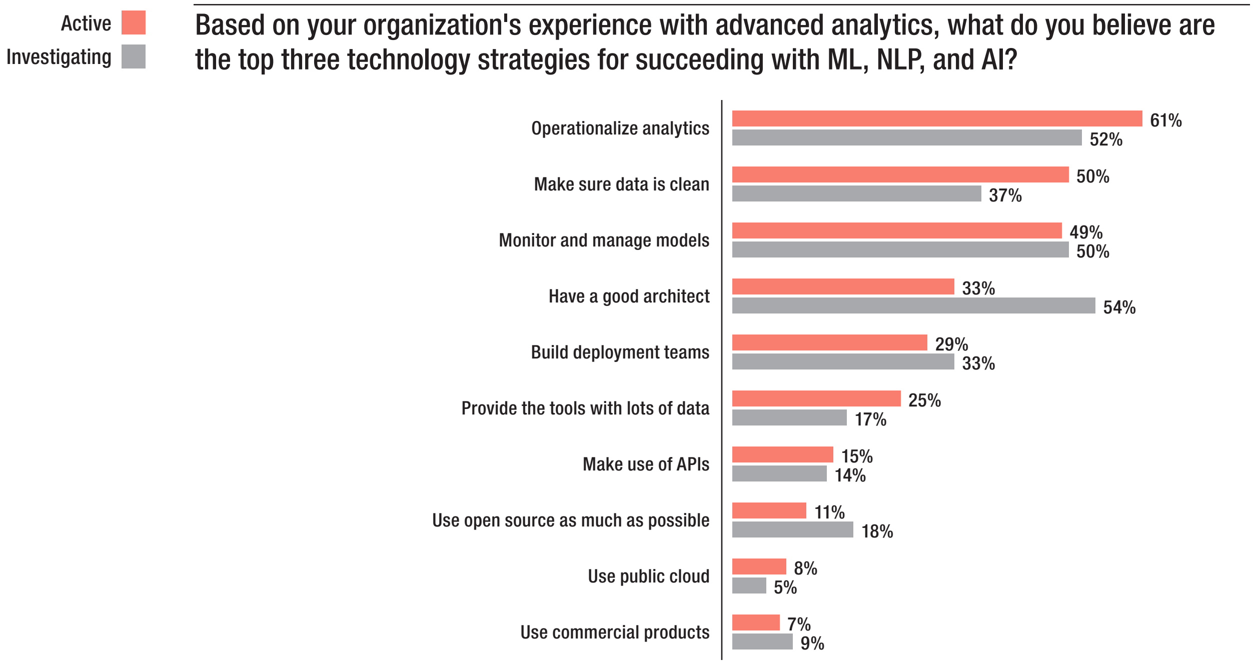
Operationalizing analytics is a top success strategy.

Another top technology strategy is clean data. Fifty percent of the active group rated this as a top priority; 59% of those who measured value ranked it as important. Thirty-seven percent of the investigating group ranked this in the top. Those who analyze data know how important data quality is—the adage "garbage in, garbage out" applies here. Some newer data preparation tools use machine learning techniques to cleanse data; however, these are still evolving and address mostly structured data. Many of the "new" kinds of data, such as text data or sensor data, will also need to be processed. Some organizations are hiring data engineers to do this type of work, and data scientists also get involved in it.

Another top technology strategy is clean data.

Another technology best practice that ranked high in all groups is managing and monitoring models. The reality is that models will degrade over time as external conditions change. It is important to monitor model performance. Some vendors provide tools for helping users do this. A handful of source data-science workbenches are starting to provide this kind of functionality, as well.

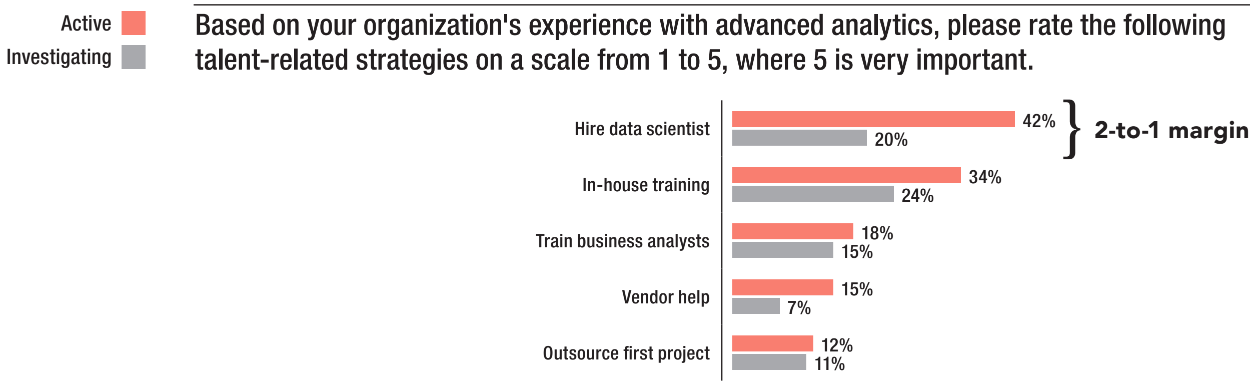
[11] See http://bit.ly/2joV76H.



Based on your organization's experience with advanced analytics, what do you believe are the top three technology strategies for succeeding with ML, NLP, and AI?

Figure 13. *Based on 116 respondents from the active group and 114 respondents from the investigating group.*

Talent strategies. Based on their experience, we asked respondents to rate talent-related strategies for machine learning, NLP, and AI on a scale from 1 to 5, where 5 was very important. These included hiring data scientists, training business analysts to become data scientists, and outsourcing (Figure 14). At 42%, hiring a data scientist was a top strategy for the active group versus 20% for the investigating group—a significant difference. This is in line with the fact that the active group was more likely to use data scientists to build models than the investigating group by a margin of 2-to-1. The investigating group may not yet need data scientists, but it is definitely a strategy to consider. Some organizations hire a few data scientists as part of the team. As one respondent stated, “It is important to *really* understand the techniques.” Some organizations may want to simply grow talent from within and train from within, though this may not be a suitable success strategy.



Based on your organization's experience with advanced analytics, please rate the following talent-related strategies on a scale from 1 to 5, where 5 is very important.

Figure 14. *Based on 116 respondents from the active group and 114 respondents from the investigating group. Ratings of "very important" shown.*

Leadership strategies. Organizations utilizing machine learning, NLP, and AI were more likely to have someone in charge of analytics other than IT (no figure is shown for these results). Top titles include VP of analytics (28%) and chief data scientist (15%). Those in the investigating group were more likely to have someone in IT in charge of their analytics program (36% investigating group versus 20% active group). This also makes intuitive sense because once organizations become more sophisticated in analytics, they typically want someone with a specific skill set to lead the efforts.

Value and Satisfaction with Analytics

Not surprisingly, those who have implemented these technologies are more likely to be satisfied with their analytics program than those who have not. Twenty-one percent of the investigating group are satisfied with their programs versus 52% of active group—a significant difference (not shown). Active group respondents believe that they “can do things [they] could not before, in a way that will ultimately drive value,” and that they “have gained more insights.”

This suggests that with more analytics maturity comes more satisfaction and value. In this study (all not shown), 82% of the active group had either measured a positive impact, felt confident that there was an impact, or believed advanced analytics was helping their organization. Six percent said it was having no impact. Twelve percent said it was not relevant at this time.

This all makes sense and is part of the success cycle mentioned earlier. As organizations see success with their analytics program, they start to do more. As they do more and as they have more experience, they tend to see positive results. This success builds on itself and is perhaps one reason why those that are more advanced analytically tend to be more satisfied and measure value.

Vendor Solutions

The firms that sponsored this report are among the leaders in BI, analytics, and data management. To get a sense of where the industry as a whole is headed, this section takes a look at the portfolio of these vendors. (Note: The vendors and products mentioned here are representative, and the list is not intended to be comprehensive.)

SAS

SAS has focused on solving business problems using analytics since 1976. Within the SAS platform, there are engines designed to address a wide diversity of analytics use cases and end-user needs. SAS 9 is being used in virtually every industry and across 149 countries. SAS Viya is the newest engine to the SAS Platform, scaling analytics workloads to cloud architectures and new types of big data.

With SAS Viya, SAS is evolving into a self-learning platform that makes use of the company's AI and cognitive capabilities, including machine learning and natural language processing. For example, SAS will utilize its strength in NLP to provide natural language interaction within its analytics software. Users interact with the software using natural language. Restful services enable developers to embed conversational services inside of applications. Additionally, certain products in the platform will be automated with intelligent pipelines the company calls guided analytics. This will include automatically performing analysis that might be relevant, such as churn.

ThoughtSpot, Inc.

Founded in 2012, ThoughtSpot's AI-Driven analytics platform enables business users to use search to analyze their data using a natural language experience or get insights pushed to them with a click. Under the hood are several piece parts that leverage AI to make this happen. This includes the ThoughtSpot Relational Search engine that provides a guided search experience as well as SpotIQ, the AI engine that generates automated insights by running dozens of advanced analytics algorithms on billions of data points. In addition, ThoughtSpot leverages an in-memory relational data cache to handle large amounts of data from different data sources. The data cache is distributed and can grow as data volumes increase.

Artificial Intelligence is built into the platform. With Relational Search, the system learns, using a machine learning algorithm, to rank search suggestions as a user types in a query. The algorithm takes into account user profile, user and group search history, and metadata and data characteristics to generate the most relevant search suggestions to the user. SpotIQ uses AI to enable "one-click" automated insights by automatically asking thousands of questions about billions of data points and bringing back dozens of insights in seconds, each with its own narrative generated in natural language. SpotIQ works hand in hand with Relational Search and uses usage-based ranking so that insights are personalized for each user based on the patterns detected from their search history, profile, and data characteristics. With supervised learning, users can also tune the built-in algorithms as they "like" or "unlike" an insight. SpotIQ also provides transparency to the queries being generated and the algorithms that have been performed. In this way, the human remains in the loop to ensure relevant, trusted, and accurate machine-discovered insights.

Vertica

Conceived and developed 10 years ago, Vertica addresses the need for organizations to derive insights from very large volumes of data. To this end, the Vertica platform provides a big data columnar store, architected for MPP. Vertica is built to run on commodity hardware, in the data center, across cloud platforms, and natively on Hadoop nodes.

Recently, the company has begun to offer its own machine learning algorithms to run inside its platform and leverage the MPP cluster to execute. Models built in Vertica can be deployed there, as well as for real-time scoring. Data can be analyzed in Vertica or, for external data to enrich a data set, can be pulled internal to the database in a flex table or used outside of the database in an external table. In addition to providing algorithms, the company also offers solutions for the Internet of Things (IoT), fraud and risk management, network management, and customer behavior analysis. The platform is open and although models built in open source languages such as R and Python can run on Vertica, the company recommends using Vertica algorithms for the best performance.

Recommendations

In closing, we summarize the report by listing the top 10 best practices for machine learning, NLP, and AI along with a few comments about why each is important. Think of the best practices as recommendations that can guide your organization into successful implementations of big data and data science.

Know your business problem. It is important to start with a *real* business problem and clear objectives when embarking on a more advanced analytics project. You should still experiment with data—exploration is part of analysis. However, understanding the business problem at hand is more likely to lead to success than looking for patterns in a sea of data.

Start somewhere. It is important to start somewhere. Change is happening fast and organizations cannot afford to be complacent. Machine learning and other advanced analytics tools have a learning curve—the more you practice, the better you get. Organizations need to start getting on the learning curve.

Consider the pros and cons of open source. Open source is rapidly becoming a go-to software for machine learning, NLP, and AI. Some of this has to do with the large community around open source as well as the fact that it is an inexpensive way to get started. Many good algorithms have been built using open source frameworks. Organizations should consider open source for their analytics needs, if they have the skills.

That said, open source analytics tools typically involve a learning curve and do not always provide all of the consistency and stability of a commercial platform. Commercial tool vendors have worked to make their tools easy to use with nice GUIs and software that addresses the analytics life cycle with nonglamorous features such as data preparation, model management, and monitoring—capabilities that are really needed to do advanced analytics at scale. If your company has the budget, commercial tools can make users more productive, faster. Remember, as well, that many organizations use commercial and open source tools together. A data scientist might build some models using open source and then run them on a commercial platform.

Hire some data scientists. There are excellent tools on the market that can help everyone become more productive. However, the reality is that if your organization really wants to do sophisticated analytics, it is probably going to have to hire at least a few data scientists. In addition to helping execute the work and potentially build apps, these people can also provide guidance to those business analysts trying to come up to speed.

Build a center of excellence. As described above, a CoE can be a great way to ensure the infrastructure and analytics you implement are coherent. CoEs can help your organization disseminate information, provide training, or maintain governance.

Hire an analytics guru to be in charge. Often the best person to be responsible for an analytics effort is someone who really understands analytics. Their titles may vary—some organizations hire CAOs and others hire VPs of analytics—but the point is that this person will lead the charge, evangelize the concepts, and provide guidance to the team.

Think about the architecture, including the cloud. The data warehouse is not going anywhere anytime soon. However, machine learning, NLP, and AI may necessitate moving beyond the data warehouse to platforms that can support multistructured data and iterative analytics. These multiplatform data architectures might include the data warehouse, Hadoop, and other platforms both on premises and in a public cloud. Users should investigate whether their in-place platform provider has plans to support these types of analytics, cloud environments, and some of the newer platform environments.

Pay attention to data quality. The issue of data quality is also front and center for many organizations. The notion put forth by some "experts" that quality issues would "wash out" with big data is not true. We have seen in this report that data quality is a very important issue—especially if attributes become obfuscated by some easy-to-use tools that hide data from the user. That can produce patterns that do not make sense. This means that IT needs to work with data scientists and business analytics to ensure data quality is in order.

Operationalize your analytics. Analytics provides the greatest amount of value when someone can take action on the results. That action can take many forms. Think about embedding advanced analytics into a system or a process or a mobile application. Think about using BI or analytics tools where advanced analytics such as machine learning is part of the process. This may help you become more productive.

Think about the success cycle. We have seen that success begets success. If you are successful with one project, do not stop there. Build on that success to become more sophisticated as an organization.

sas.com/analytics

SAS offers software for the entire analytics life cycle, including enterprise platform solutions for data management, discovery, and the deployment of analytics. As a global leader in advanced analytics, machine learning, and data mining, the company offers a comprehensive suite of software for structured, unstructured, and streaming data analysis for all users in an organization, ranging from the expert data scientist and business analyst, to the end consumer of the analytical results. The results are easy to deploy and based on automated creation of all required assets for operationalizing and embedding analytics, including model governance and management, model documentation, and model monitoring. SAS supports the process of embedding advanced analytics for automating operational decision actions. Business and IT can collaborate in an environment to design, build, and execute decision flows that combine business rules and advanced analytics for the automation of operational business decisions. With scenario testing and direct integration with operational data, governed by integrated workflow, organizations can reduce deployment times in execution environments.

Learn more at: sas.com/action

THOUGHTSPOT

thoughtspot.com

ThoughtSpot's AI-Driven analytics platform puts the power of a thousand analysts in every businessperson's hands. With ThoughtSpot, you can use search to easily analyze your data or automatically get trusted insights pushed to you with a single click. ThoughtSpot connects with any on-premises, cloud, big data, or desktop data source and deploys 85% faster than legacy technologies. BI and analytics teams have used ThoughtSpot to cut reporting backlogs by more than 90% and make more than 3 million decisions—and counting. ThoughtSpot's customers include Amway, Bed Bath and Beyond, Capital One, Celebrity Cruises, Chevron Federal Credit Union, DeBeers, Insurethebox, and Scotiabank. ThoughtSpot was cofounded in 2012 by its CEO, Ajeet Singh, and six other technical cofounders from Google, Microsoft, Amazon, and Oracle. The company is based in Palo Alto, California, and is currently expanding operations in North America, Europe, and Asia-Pacific. ThoughtSpot's mission is to enable analytics at "human scale" and put search-driven analytics in the hands of 20 million users by 2020. For more information, please visit thoughtspot.com.

VERTICA

vertica.com

The Vertica Analytics Platform is designed for use in data warehouses and other big data workloads where speed, scalability, simplicity, and openness are crucial to the success of analytics. Vertica relies on a tested, reliable distributed architecture and columnar compression to deliver blazingly fast speed. A simplified license and the capability to deploy anywhere delivers on the promise of big data analytics like no other solution.

Vertica provides you with the broadest range of deployment models so you have complete choice as your analytical needs evolve:

- **Vertica Enterprise:** The core "shared nothing," distributed analytical database designed to work on clusters of cost-effective, off-the-shelf servers in your data center with unparalleled performance and extreme scale.
- **Vertica in the Clouds:** Optimized and pre-configured to run on AWS, Microsoft Azure, and VMware clouds, Vertica is also available as a BYOL (bring your own license) model to enable you to transition your data analytical workloads to the cloud, on premises, and back seamlessly.
- **Vertica for SQL on Hadoop:** Run the industry's most comprehensive Vertica SQL analytics engine directly on your Hadoop cluster and tap into advanced SQL on Hadoop capabilities.

research

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on business intelligence, data warehousing, and analytics issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of business intelligence, data warehousing, and analytics solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

tdwi

Transforming Data
With Intelligence™

555 S. Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org