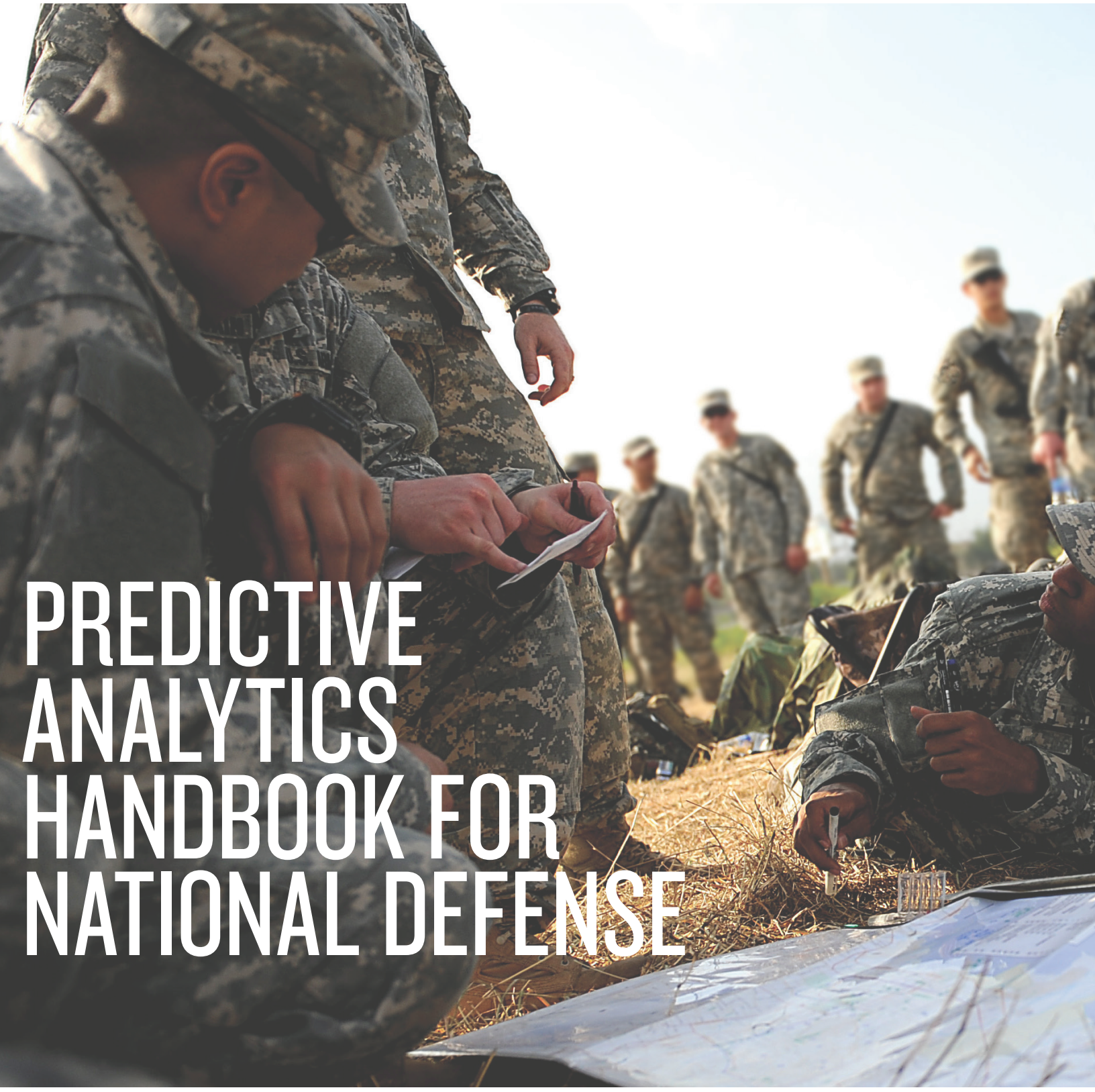
Booz | Allen | Hamilton



PREDICTIVE ANALYTICS HANDBOOK FOR NATIONAL DEFENSE

CONSULTING | ANALYTICS | DIGITAL SOLUTIONS | ENGINEERING | CYBER

Forward: From Hindsight to Foresight 1

The Predictive Analytics Story 3

Moving Beyond Descriptive and Lower-Level Analytics 3
The First Breakthrough: Higher-Level Diagnostic Analytics 4
Turning the Dials: Predictive Analytics 6
Beyond the Predictive 7
A New Level of Decision-Making 7

Why Predictive Analytics Are Achievable Now 8

Why Predictive Analytics are Technologically Feasible 8
Why Predictive Analytics are Practical 11
Why Predictive Analytics are Cost-Effective 14

Predictive Analytics for the Defense Community 17

Predictive Analytics for Readiness Planning 18
Predictive Analytics for Logistics Planning 20
Predictive Analytics for Workforce Management 22
Predictive Analytics for Military Intelligence 24

How Data Science Works: A Brief Overview 26

What Makes Data Science Different 26
The Impact of Data Science 28
What is Different Now 28
The Four Key Activities of Data Science 28
How Data Science Comes Together 31

Looking Ahead 32

FORWARD: FROM HINDSIGHT TO FORESIGHT

Defense organizations have long sought to take full advantage of one of their most valuable resources—the vast amount of data they collect day in and day out. They want to be able to use that data to make more insightful, forward-looking decisions about readiness, logistics, manpower, intelligence, and a host of other critical defense concerns. A new generation of advanced analytics—high-level diagnostic and predictive—can provide them that opportunity.

With these analytics, defense organizations can go beyond looking back at data to make projections. They can begin to drill deep into cause and effect, and determine the mathematical probability of future occurrences. This is the transition from hindsight to foresight.

With the conventional approaches commonly used by defense organizations, it is difficult to bring together all of the data for analysis. Typically, organizations look at data in isolated silos—such as training, personnel, equipment or supply chains—and then theorize about the impact those areas may have on one another. With advanced analytics, they can break down those silos to see the larger picture, with the full scope of interrelationships. Perhaps even more significantly, they can "turn the dials" to see what might happen to that picture if they take a specific action. How much can they cut back on training without increasing the failure rate of mission-critical aircraft parts? How will the failure of specific parts affect overall mission effectiveness?

Highly accurate predictive analytics are no longer aspirational—they are now within reach of every defense organization. Thanks to breakthroughs in data science, predictive analytics are:

1. Technologically feasible. New approaches, designed expressly for the age of big data, have overcome the obstacles that have long limited analytics. Defense organizations can now bring together and explore their data in profound new ways.
2. Practical. Defense organizations have already done much of the groundwork, making the bar of entry to predictive analytics relatively low. And, new tools are making it possible for these organizations to ramp up analytic capabilities quickly, and put them directly in the hands of defense analysts and commanders.
3. Highly cost effective. Substantially less time and effort are needed for data preparation and analysis, enabling defense organizations to increase their analytic output at lower costs. Organizations can save even

Highly accurate predictive analytics are no longer aspirational—they are now within reach of every defense organization.

more by using the advanced analytics to accurately target investments and cuts, and meet schedule and budget requirements.

Each of these factors is an outgrowth of rapid advances in data science. While it may be possible to conduct limited predictive analytics with current methods, they are typically too time-consuming and cumbersome to be of real value. With data science, what formerly took days, weeks or months to process can now be done in hours, minutes, seconds—or even microseconds. And we can now use all of the data—and all at once—not just select subsets.

Defense organizations do not need to break new ground to put predictive analytics into action. Other areas of government, particularly the intelligence community, have paved the way in using predictive analytics to guide decision-making. And in the private sector, companies like Google, Amazon and Facebook have helped make sophisticated predictive analytics a part of our everyday lives.

Booz Allen believes national defense organizations are poised to move into the realm of foresight. As a pioneer in predictive analytics and data science, and as a longtime partner of the defense community, we've created this handbook to help organizations make the transition.

THE PREDICTIVE ANALYTICS STORY

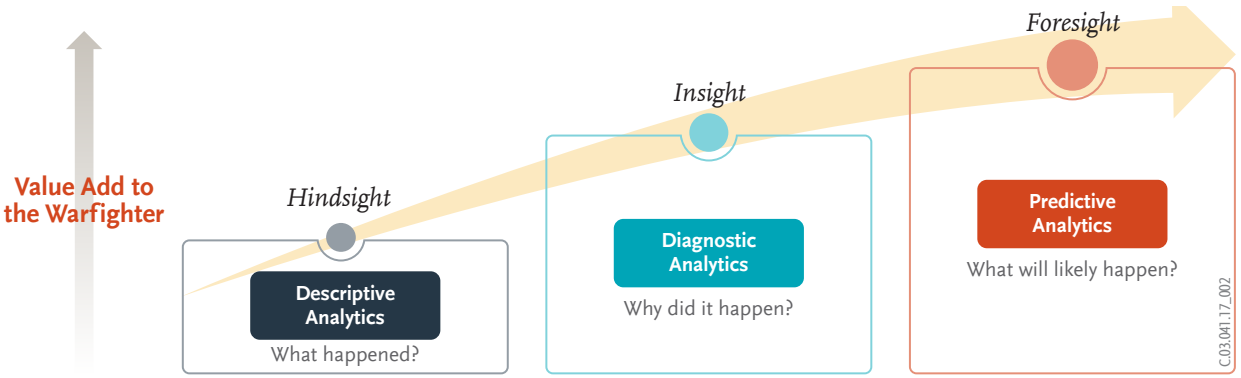
Predictive analytics represent a game-changing leap over current practices. Here is an essential guide to what they are, and what can they do that you're not doing now.

MOVING BEYOND DESCRIPTIVE AND LOWER-LEVEL DIAGNOSTIC ANALYTICS

The defense community has made major strides with two phases in the evolution of analytics—**descriptive** and lower-level **diagnostic**. Descriptive analytics are the bread-and-butter of decision-making in defense organizations today. They sort through and summarize raw data to make it understandable, through spreadsheets, charts, reports and other kinds of presentations. Descriptive analytics commonly help defense organizations determine the current situation—levels of manpower, training, or equipment, for example, or where adversaries are moving personnel, material or financial assets. These analytics are also used to examine past trends, with the aim of guiding future actions. Essentially, these analytics provide *hindsight*.

Descriptive analytics are a necessary building block for the next phase in the evolution of analytics, the **diagnostic**. While descriptive analytics can show what has happened in the past, or what is happening now, they don't explain why things happen. That's the role of diagnostic analytics, which look for relationships in the data that can provide clues to causes and effects. Diagnostic analytics work with the foundational descriptive analytics to provide the next level of understanding: *insight*. Intelligence analysts might use diagnostic analytics, for example, to determine whether money being moved by a terrorist group is intended to buy weapons.

While many defense organizations employ diagnostics, they are typically limited to lower-level aspects of the analytic.



Advances in data science are enabling defense organizations to transition to higher-level diagnostic analytics, and then to predictive analytics.

Most organizations rely on the older approaches to data and analytics, which were developed before the age of big data. These approaches make it difficult for defense organizations to integrate and analyze the vast amounts of data, from diverse sources, that are available today.

Much of the problem lies in the way data is currently stored and accessed. For example, analysts looking for the terrorist money-weapons connection would first have to hypothesize that such a connection exists. They then would typically build a relational database, with a number of datasets, to test the hypothesis. This tends to be a very time-consuming task—analysts can spend as much as 80 percent of their time formatting, or "normalizing," data so that it can be analyzed. If analysts don't find a connection, they have to toss out the custom-made data structure and start over—with a new hypothesis, new data, and a new round of data normalization. And so with only limited resources available, analysts don't have time to ask all the questions they want.

This process makes it difficult for analysts to get a complete picture. As a result, if they do find a connection, it may not be the most important one. For example, the terrorist money might indeed be intended for weapons, but it might represent only a small fraction of a much larger scheme. Most of the money might be coming from other financial sources, which the analysts didn't think to look for, and may never find.

The difficulty in bringing together large amounts of data has other implications. Defense analysts—whether in military intelligence, logistics, readiness, manpower or any other area—may have hundreds or even thousands of datasets to work with. They often have no choice but to cull that data, selecting the information they think is most important to the inquiry.

The danger is that the analysts may be trying to answer a question with only a small portion of the relevant information. More important data, that relates to other, unasked questions, may end up left on the cutting-room floor. And there's another problem. Because the analysts are focusing on a particular hypothesis, they are inevitably biasing the results. They may only be searching for what they already expect to find.

THE FIRST BREAKTHROUGH: HIGHER-LEVEL DIAGNOSTIC ANALYTICS

Advances in data science are overcoming these and other barriers, to continue the evolution of analytics. They are enabling defense organizations to transition to higher-level diagnostic analytics, and then to predictive analytics.

While descriptive and lower-level diagnostic analytics provide defense organizations with important information, they cannot easily show the full set of interrelationships in a given situation. Such analytics often cannot answer the key questions commanders are asking, such as, "What are the hidden causes and effects at play here? What is driving the outcomes I'm seeing?"

With high-level diagnostics, these questions can start to be answered. Breakthroughs in data science have made it possible for defense organizations to integrate and analyze their vast data stores—all at once, and without the need for extensive data preparation. High-level diagnostics can then map out complex connections and patterns in the data, and show the impact of various forces on others.

For example, high-level diagnostics might reveal a host of factors that in some way or another have led to unexpectedly high attrition in a particular unit over the course of basic training and the initial assignment. The diagnostic would show which factors had the most impact, which had a moderate impact, and which had the least impact.

Lower-level, hypothesis-driven diagnostics might find only a small number of the many factors that played a role in the attrition—and would have difficulty determining their relative influence. And, because these lower-level diagnostics don't provide the larger context, analysts may see a correlation, and mistake it for a cause. Higher-level diagnostics let the data speak for itself, creating a more full picture that can begin to reveal probable cause and effect. This is the move from hindsight—what has been happening—to true insight—what it really means.

One of the most important features of higher-level diagnostics is that they can show us anomalies in overall patterns, uncovering critical factors we hadn't thought to look for. Such diagnostic analytics, for example, might reveal connections linking two militant groups that had previously been thought to be operating independently. Higher-level diagnostics can probe deep into cause and effect. An analytic might find, for example, that an increase in failures of a tank part had less to do with terrain—the presumed culprit—than with a certain group of tank mechanics who all happened to have trained at the same base, during the same time period, under a particular instructor.

The search for cause-and-effect is not a solely data-driven exercise. It requires the insight of domain experts who have an operational understanding of readiness, logistics, etc., and of the related data. Their expertise is needed throughout the process—from building the context of the data into the analytics, to helping interpret the results. For example, an analytic may show an apparent cause-and-effect relationship in the data. The domain expert might agree with that finding and perhaps refine it—or point out reasons why it is likely misleading and shouldn't be considered. This is an iterative process that builds increasing fidelity into the analytic.

Breakthroughs in data science have made it possible for defense organizations to integrate and analyze their vast data stores—all at once, and without the need for extensive data preparation.

With predictive analytics, we can ask entirely new kinds of questions.

TURNING THE DIALS: PREDICTIVE ANALYTICS

Predictive analytics are built on the foundation of higher-level diagnostic analytics. With the predictive, we can turn the dials to how see how the picture we've created in the diagnostic changes if we take a particular action. Predictive analytics go far beyond the common practice of manually extrapolating from spreadsheets to make projections. These advanced analytics rely on computer models to create and think through any number of possible scenarios, and assign each one a likelihood of occurrence. The analytics do not actually predict the future. They merely provide the probability that events will unfold in a certain way, based on data about how events have unfolded in similar ways in the past. They represent the move to ***foresight***.

With predictive analytics, we can ask entirely new kinds of questions. If I reduce the number of civilian aircraft mechanics on my base, how will that likely affect not just maintenance, but unit and overall logistics and readiness? If I interdict the flow of money by a terrorist group in one country, in what other countries will the group likely try to compensate? What will happen if I change the maintenance cycle for certain ships?

With this type of "what-if analysis," we can see the ripple effect of possible decisions. For example, a diagnostic analytic might show how 15 elements of readiness work together, how 10 elements of logistics work together, and how each of those 25 elements influence each of the others. If I use predictive analytics to turn the dials of those elements—either singly or in combination—I can see the implications across the entire landscape.

The outputs of predictive analytics are typically expressed as percentages. In a given scenario, for example, there might be an 85 percent probability of Outcome A, a 10 percent probability of Outcome B, and a 5 percent probability of Outcome C. Such percentages are based entirely on data from events that have occurred in the past. There may be data from months or years ago—or data from just a few seconds ago that is now streaming in. When weather forecasters show a map of potential paths of a hurricane, they're using predictive models that combine the latest information with historical data about how previous hurricanes have behaved. The more data that is available—whether in predicting hurricanes or readiness or recruit retention—the more accurate the predictive model.

The outputs of predictive analytics are particularly helpful in assessing risk, such as when making a tactical decision, or in deciding whether to make cuts or investments. Predictive analytics evaluate the risk, and express it as a mathematical probability. This provides commanders with powerful information when they're balancing priorities and considering tradeoffs.

Predictive analytics don't always have to involve turning dials. Defense analysts can plug in the outcome they're looking for, and ask the analytic

to create a picture of how all the contributing factors would have to come together. For example, the director of a federal lab might want to know how to achieve the most scientific breakthroughs with the R&D projects that are planned. A predictive analytic, using data from the other federal labs, as well as from a host of labs in business and academia, might identify the ideal mix of expertise needed among the lab's scientists and other researchers—possibly a combination of disciplines that no one had considered.

BEYOND THE PREDICTIVE

As defense organizations become adept at predictive analytics, they can begin to move to the next phase, **prescriptive** analytics. These analytics rate each option of the predictive, taking into consideration the organization's goals, needs and limitations and other factors. In creating prescriptive analytics, data scientists, working closely with domain experts, build in information about the organization, from its mission to its day-to-day operations. The prescriptive analytics can then use this context to sort through the various predictions, and recommend the ones that will best meet the organization's goals.

A simple example of how predictive and prescriptive analytics work together is the "maps" application on a smartphone. Underlying the app is a computer model of streets and highways throughout the country, many with both their average and real-time traffic flows. When you enter in your current location and destination, the app gives you several possible routes, with the up-to-the-moment estimated travel times for each. That's the predictive analytic. The app also highlights the fastest route—perhaps in blue—which is essentially a recommendation of the best path to take. That's the prescriptive analytic.

A NEW LEVEL OF DECISION-MAKING

Neither the predictive nor prescriptive analytic takes away decision-making from commanders or other leaders. The analytics are simply tools that give the decision-makers more hard data to work with. In this case, the data might be in the form of the mathematical probability that an event will occur, or a recommended action based on certain parameters. Commanders still need to use their experience, knowledge and judgment to evaluate the analytic outputs.

Advanced analytics have the power to substantially increase the quality of decision-making. Currently, commanders may spend much of their time trying to understand the relationships between isolated silos of information, such as data on equipment, training and maintenance. With advanced analytics, those relationships are already clear—freeing up commanders to focus on higher-level decisions. In essence, the machines are doing what they do best, so that people can do what they do best.

The outputs of predictive analytics are particularly helpful in assessing risk, such as when making a tactical decision, or in deciding whether to make cuts or investments. Predictive analytics evaluate the risk, and express it as a mathematical probability.

WHY PREDICTIVE ANALYTICS ARE ACHIEVABLE NOW

New data science approaches, tools and technologies are making it possible for the defense community to harness the power of predictive analytics. Defense organizations are well positioned to make this transition, thanks to their strong foundation in data and analytics. There are three compelling reasons why organizations can start taking their analytics to the next level. Predictive analytics are:

Technologically feasible

Practical to implement and use throughout the organization

More cost-effective than current approaches

Why Predictive Analytics are Technologically Feasible

Data science is enabling defense organizations to overcome many of the constraints of conventional computing techniques, which were developed before the age of big data. One of the chief limitations of the conventional approach is that it is difficult for defense organizations to take full advantage of their vast amounts of data.

For example, before analysts can analyze data, they must first build rigid data structures, such as relational databases, and then format the information to fit them. This is typically a laborious task—analysts often spend the bulk of their time simply preparing the data. With each new line of inquiry, the data structures are torn down and rebuilt in a cumbersome, iterative process. In addition, only a limited amount of data can be placed in any one of the structures, forcing analysts to set aside potentially valuable data.

With these and other constraints, defense organizations, as a practical matter, are only able to take advantage of a relatively small portion of their data to make decisions about readiness, logistics, intelligence and other areas. However, new data science approaches are enabling organizations to use all of their data, and to harness for it decision-making with advanced analytics.

At the heart of these new approaches is the implementation of a **data abstraction layer**. One of its key breakthroughs is that instead of storing data in silos—the traditional method—it liberates the data, bringing together and integrating all of the data available to an organization, including from outside sources. Just as important, the entirety of the data is available, all at once, for any inquiry.

CONVENTIONAL APPROACHES	NEW DATA SCIENCE APPROACHES
Schema-on-write	Schema-on-read
Data stored in hard-to-connect silos	The entirety of the data is brought together
Difficult to use unstructured data	All types of data can be used easily
Query data by testing hypotheses	Let the data "speak for itself" to reveal patterns and connections
Data typically culled for inquiries	All of the data is available for every inquiry

This makes it possible to find previously hidden patterns and interrelationships in the larger data, which is a necessary foundation of predictive and other advanced analytics. Many of the approaches and technologies that underlie the data abstraction layer were pioneered by Booz Allen in conjunction with the intelligence community to help fight terrorism and other threats. The ability to find patterns and interrelationships has been a boon to the intelligence community, and can inform decision-making in every defense organization.

The data abstraction layer overcomes another major constraint of the conventional approach. We no longer need to query the data by framing and testing hypotheses. We can simply pose questions of the data as we go along, quickly and easily refining our queries as we gain greater insight.

What makes these advances possible is the data abstraction layer's transformative shift from "schema-on-write" to "schema-on-read." With **schema-on-write**, which underlies conventional computing methods, we must first decide how we want to organize the data for analysis—for example, what the row and column headings of a database will be. The choices we make not only circumscribe the data that can be examined, they also define the kinds of questions we can ask. Before we have even begun using the data, we have already narrowed our possibilities.

Once conventional structures have been built, the extract/transform/load (ETL) process is used to pull information from a data source and transform, or mold it to fit the strict parameters that have been set. In this way, the schema—or basic organization—is imposed upon the data as it is written into the structure. Everything is now locked into place. It is difficult to change the structure, the range of usable data, or the types of questions we can ask.

Schema-on-read operates on an entirely different model. Before being analyzed, each piece of data in a data source—such as a name, a photograph, a document or a news feed—is identified with metadata tags. These tags serve the same purpose as old-style card catalogues, which allow readers to find a book by searching the author, title, or subject. As with the card catalogues, tags enable us to find particular information from a number of different starting points—but with today's tagging abilities, we can characterize data in nearly limitless ways. The data abstraction layer "reads" these tags as it pulls data from the various data sources. The more tags, the more complex and rich the analytics can become. In addition, it does not matter how or even whether the data is formatted. Because all of the data connected by the data abstraction layer can easily be connected through the tags, the time-intensive frameworks of ETL are no longer necessary.

An advantage of the data abstraction layer is that it easily reads all forms of data, including unstructured, such as photos and text.

When we are ready to conduct an analysis, we use the tags to choose any or all of the data, and ask whatever questions we wish. Unlike with the conventional approach, we do not need to guess at which portion of the data to use—all of the data is available. Users can easily switch variables in and out to pursue entirely new lines of inquiry.

An advantage of the data abstraction layer is that it easily reads all forms of data, including unstructured, such as photos and text. Unstructured data is often left out of conventional analyses because it can be difficult to format—a problem, considering that such data makes up an increasingly large portion of the information collected by defense organizations today. With the data abstraction layer, the big data of defense organizations is fully available for predictive analytics.

MACHINE LEARNING

Another way data science is making advanced analytics technologically feasible is through machine learning, which helps conduct the search for patterns and interrelationships in the data. Machine learning analytics have the ability to develop a continuous and growing awareness of content and connections. For example, if the analytic "knows" the many variations of a name, such as common and alternative spellings, it can pull together all the data on a single individual who may use different spellings, and store that for later queries.

With conventional data structures, we typically have to know in advance what the commonalities in data might be—we might, for example, need a Social Security number to connect data about an individual. But the machine learning analytic finds commonalities on its own—through, say, addresses or family relationships—to form a much richer picture. Because the analytic stores all of its learning, it is able to instantly respond to queries. Machine learning analytics allow the data to "talk to us"—to reveal, on its own, the hidden patterns and interrelationships that can provide clues to cause and effect.

OPEN SOURCE AND OPEN ARCHITECTURE

Conventional approaches typically rely on proprietary software and architectures. These often limit system agility, and can lead to expensive vendor-lock. By contrast, the new approaches use both free open-source software (FOSS) and open architectures. Free of vendor-lock, defense organizations can build their systems with plug-and-play, modular capabilities. This provides far greater flexibility in provisioning, automating, orchestrating and implementing containerization, which greatly facilitates the development and deployment of predictive and other advanced analytics. At the same time, the open source and open architecture approach is substantially more cost-effective. One example of an open architecture is the data lake, which stores and manages large amounts of data.

Why Predictive Analytics Are Practical

Moving from hindsight to foresight with predictive analytics is a big jump—but it's easier for defense organizations than it might seem. There are two key reasons. First, most defense organizations have already laid the necessary groundwork, and are now ready to move to the next level. And second, advanced analytics—unlike older approaches—are not the exclusive domain of computer scientists or other experts in a small corner of an organization. One of the more remarkable breakthroughs of data science is that even people without specialized computer expertise can use sophisticated analytics on their own—to become more efficient, and gain both deep insight and foresight to aid decision-making. Factors such as these make the bar of entry to advanced analytics relatively low.

A STRONG FOUNDATION

The defense community has long been a leader in collecting data and using it to drive decision-making. And it has moved aggressively into the age of big data. Defense organizations are taking advantage of new methods of digitization, and are steadily expanding their data repositories. This rich and vast data supply is the fuel necessary for predictive and other advanced analytics.

Defense organizations are also laying the groundwork for advanced analytics by migrating to cloud computing, and employing new types of analytics. For example, intelligence analysts are using new analytic tools to update and consolidate their findings on people and events in near-real time—giving commanders greater situational awareness. By building on such foundational approaches, defense organizations can move into the realm of predictive and other high-level analytics.

DEMOCRATIZING THE DATA AND ANALYTICS

Another reason advanced analytics are practical is that they're more accessible, and to many more people in the organization, than are conventional analytics. For example, defense analysts and other subject-matter experts can use high-level diagnostic and predictive analytics without the need for computer scientists or computer engineers to serve as intermediaries. With conventional approaches, such experts are needed to design custom data-storage structures and queries. They essentially act as agents for the defense analyst, translating analysts' goals for the data into the language of the machine. Whenever there is a middleman in any field, things tend to get lost in the translation, and data analysis is no exception. Here, it leads to a disconnect between the people who need knowledge and insight (defense analysts and others) and the data itself. It also substantially slows the process.

With advanced analytics, the middleman syndrome disappears. Analysts and others subject-matter experts have direct access to the data, and with it the ability to explore the data, asking intuitive questions and looking for critical patterns and interrelationships.

Computer scientists and other experts are still needed to conduct important tasks, such as designing the underlying analytics. But such experts are now no longer needed to drive most of the inquiries into the data.

NEW TOOLS TO EXPLORE THE DATA

The advanced analytics themselves also tend to be highly user-friendly. For example, people without specialized computer expertise can ask questions in ordinary language, without needing to know SQL or other programming languages. And this can be done with the ease of an everyday Internet search.

These tools enable subject-matter experts to shape and focus their inquiries using a broad range of modules that can be switched in and out. Modules are essentially mini-analytics that work with the larger diagnostic and predictive analytics to slice and dice the data in various ways. The tools are designed to help users drill deep into the data, with features like autocomplete, which suggests potential questions to ask. Analysts and other subject-matter experts can customize the modules, using drop-down menus, for more tailored questions and answers. They can also create different versions of a module for specific types of projects. And users can save these customized modules for future use, as part of their permanent portfolios, as well as share them with others.

BUILDING THE VISUALIZATIONS INTO THE ANALYTICS

Organizations may be understandably concerned that even with the ease of visualization tools, working with all the data may be overwhelming, leading to information overload. The advanced analytics avoid this risk by incorporating the visualization from the outset. That is, the analytics not only conduct the inquiries, they help contextualize and focus the results.

This enables analysts to more easily make sense of the information, to frame better, more intuitive inquiries, and to gain deeper insights. Building the visualization into the analytics has another advantage—it provides the ability for quick and effective feedback between the analyst and the data, so that the findings can be continually refined for the ultimate decision-maker.

The visualization tools also make it possible for different organizations to tailor how they see the same data. For example, one organization might want

to examine training data through a readiness lens, while another might want a logistics-related focus. The advanced analytics easily accommodate both views of the data—and any number of others.

GRANULAR SECURITY IN THE ANALYTIC OUTPUTS

Another potential concern of defense organizations is what happens when the data that goes into an analytic is unclassified, but the result of the analytic—the insight and foresight—is in itself classified. This issue is resolved with the advances in technology used in data science, which provide security to analytic outputs at a granular level.

In the same way that data is tagged to help identify and locate it for analysis, it is also tagged for security. Each piece of data—such as a name, a photograph, an incident report or a Twitter feed—is tagged with its "visibility." This is essentially a security filter that governs who has access to the data and under what circumstances, and ensures all compliance regulations, standards and legal restrictions are applied. This multi-layered visibility is embedded using an Attribute-Based Access Control (ABAC) system.

Using the "visibility" tags of the data, the data abstraction layer embeds the logic and rules for combining the data through analytics. As with the data, the security of the analytic outputs is decided in advance through governance policies. When the analytics are performed, they implement the rules and logic that govern the security—for example, that when data from three particular unclassified sources are combined for analysis, the results of the analytic must be classified. In this way, advanced analytics fully support the defense organization's governance policies.

Why Predictive Analytics Are Cost Effective

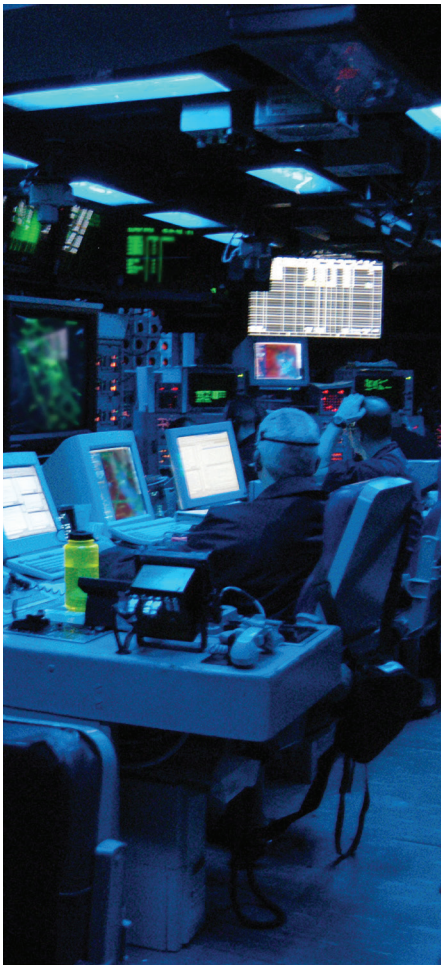
With defense organizations facing continuing budget constraints, predictive and other advanced analytics would be out of reach—just a pipe dream—if they could not pay for themselves. But the analytics can do more than that—they can bring substantial savings across the organization, both in analytic costs and in overall operations. The analytics aren't just cheaper to use. They can show defense organizations where they can safely make cuts, where they can invest for maximum return, and what specific actions they can take—on a day-to-day basis—to make their operations more efficient. The potential for savings and efficiencies with advanced analytics has now become so great that in the private sector, businesses that fail to adopt these analytics cannot hope to keep up with the competition.

REDUCED MANPOWER COSTS

Conventional analytic approaches commonly used by defense organizations tend to be highly labor intensive. At many organizations, analysts may spend as much as 80 percent of their time preparing the data, leaving just 20 percent for conducting actual analysis. The reason is that the conventional extract/transform/load (ETL) process requires that a specific data structure and analytic be built for each new line of inquiry. All information entered into the data structure must first be converted into a recognizable format, often a slow, painstaking task. For example, an analyst might be faced with merging several different data sources that each use different fields. The analyst must decide which fields to use and whether new ones need to be created. The more complex the query, the more data sources that typically must be homogenized.

Conventional methods also require that analysts spend a great deal of time selecting samples to be analyzed, posing hypotheses, and sifting through and refining results. That intense level of effort may be workable for small amounts of data, but becomes a heavy cost burden for organizations attempting to analyze big data.

By contrast, advanced analytic approaches turn over most of the work to the computer, particularly tasks that are repetitive and computationally intensive. This greatly reduces labor costs, and at the same time speeds up the work. Analysts no longer have to laboriously build and rebuild data structures—large amounts of data can be ingested and made available via the data abstraction layer with a minimum amount of preparation, and become instantly available for analysis. When we use advanced analytics to find interrelationships in the data, or make predictions, we are essentially asking the computer to take us as close as it can to finding the answers we want. It is then up to us, using our cognitive skills, to find meaning in those answers. By separating out what the computer can do—the analytics—and what only people can do—the actual analysis—the new approaches greatly ease the human workload.



REDUCED INFRASTRUCTURE COSTS

Expandability

With conventional approaches, organizations typically must expand their infrastructure as they add data stores, a burdensome expense. This is the case even with the cloud. The problem is that much of the space in current data frameworks is wasted. Imagine a spreadsheet combining two data sources, an original one with 100 fields and a new one with 50. To combine the two sources, we have to add 50 new 'columns' into the original spreadsheet. Rows from the original will hold no data for the new columns, and rows from the new source will hold no data from the original. The result will be a great deal of empty cells—and wasted storage space. With the data abstraction layer, no space is wasted, making it possible to store vast amounts of data in far less space than is found in even relatively small conventional data structures. As a result, organizations can cost-effectively scale their growing data, including from multiple outside data sources.

Reusability

In conventional methods, the analytic structures and other tools have to be continually built, torn down, and rebuilt as lines of inquiry change. With the advanced analytic approaches, tools are highly reusable for almost any number of inquiries. For example, a military intelligence unit might build a pre-analytic that transliterates a name like Karen into every possible spelling (e.g., Karin, Kerryn, Karyn, Caren). This would enable the computer to collect and analyze information about a particular person, even if that person's name is spelled differently in different sources of data. Although pre-analytical tools are commonly used in the conventional approach, they are typically part of the rigid structure that must be torn down and rebuilt as inquiries change. Generally, they cannot be reused—for example, each name to be transliterated would require an entirely new pre-analytic. With the new approach, the same pre-analytic can be used any number of times.

Distributed file system vs. SAN

A scalable framework for advanced analytics is often supported by a distributed file system, rather than by a conventional storage area network (SAN). With a SAN, data is taken out of storage for processing and then returned, traveling back and forth through a narrow fiber channel that substantially limits speed and capacity. With the distributed file system, however, the processing is conducted right at the point of storage—on thousands of nodes, all networked together in a cloud environment. It is considerably less expensive to add storage with a distributed file system than with a SAN. Instead of continually purchasing and configuring new storage systems, as with a SAN, more nodes can simply be added to the distributed file system as needed.

By reducing both manpower and infrastructure costs, predictive analytics enable defense organizations to address budgetary constraints while significantly expanding—rather than limiting—their analytic output and data-informed decision-making.

INCREASED OPERATIONAL EFFICIENCY AND ROI

Defense organization can realize greater cost savings by using predictive and other advanced analytics to effectively guide their budget decision-making. Predictive analytics can be used to identify potential reductions in personnel, equipment or maintenance that will have a minimum impact on readiness. For example, an organization may want to save money by deferring maintenance of a lightly-used aircraft. Would that be the right decision? A predictive analytic could look at evolving unit and overall mission requirements, and give a clear picture of how much the aircraft is likely to be flown—and optimize the type and frequency of maintenance that would actually be needed.

Predictive analytics can also provide cost savings by helping organizations target investments. For example, an organization might try to protect itself against cyberattacks by spending money to shore up every potential vulnerability, at great expense. But with predictive analytics, that might not be necessary. A predictive analytic could look at a wide range of factors—the methods and targets of potential cyber attackers, the most critical points in the organization's network, whether attackers are likely to know about those points, as well as the likely damage that a successful attack might cause. The analytic could then predict where the network—and the organization—is at greatest risk, and so where it should direct its cyber investments.

Still another way predictive analytics are cost-effective is by helping organizations

make the best use of resources. For example, it is often challenging for base commanders to determine what their civilian workforce should look like—how many people, of various skill sets and levels of experience, are needed for the required tasks. Commanders can't rely on industrially engineered standards—which call for far more personnel than the military can afford in these lean times, and are no longer useful. Often, commanders have to guess at who to hire, based on historical numbers that may not be tied to performance. Predictive analytics could look at the full scope of historical data on civilian workers at that base and many others, and consider skill level, aptitude, experience, performance, hourly rates, and host of other factors, to show the most cost-effective mix of civilians.

PREDICTIVE ANALYTICS FOR THE DEFENSE COMMUNITY

Higher-level diagnostic and predictive analytics can transform every area of defense. These analytics have the ability to map out the complex interrelationships between key factors—and then show what specific actions commanders can take to increase efficiency, cut costs, and support the warfighter.

Defense organizations have long sought to develop objective, data-driven systems, and have taken the approach of collecting ever greater amounts and different kinds of data. But the problem isn't a lack of data, or even lack of the right kind of data. For the most part, defense organizations are already collecting all the data they need to fuel predictive and other powerful new analytics. The challenge comes in bringing the data together as a whole, understanding how it is interrelated, and uncovering deeper insight and foresight.

Advanced analytics accomplish this. New approaches like the data abstraction layer make it possible for organizations to integrate and harness all their available data. And organizations are no longer limited to only structured data, the kind that might be found on spreadsheets, but now have full access to structured, unstructured and semi-structured data—such as notes, photographs and video feeds. Throughout the process of ingesting and accessing the data, domain experts who understand readiness, logistics, etc., work with data scientists to put the data into context.

Next, higher-level diagnostic analytics—using machine learning and other data science approaches—begin assembling vast webs of patterns and interrelationships in the data. Even without predictive analytics, such advanced analytics can provide commanders with actionable insight. They can identify the hidden factors that strongly influence efficiency and cost, and point to areas that may warrant more close attention. In addition, where we may have expected only one or two factors to be critical to a certain aspect of a problem, the analytics may reveal clusters of important factors.

With predictive analytics, commanders can then begin to turn the dials to see the likely impact of actions they might take. This section illustrates how predictive analytics can help commanders make decisions in four areas of defense—readiness planning, logistics planning, workforce management, and military intelligence. These are just examples—predictive analytics can provide the same type of valuable foresight for operational and planning decision-making across the entire defense community.

Throughout the process of ingesting and accessing the data, domain experts who understand readiness, logistics, etc., work with data scientists to put the data into context.

Predictive Analytics for Readiness Planning

Military commanders have a deep operational understanding—based on their knowledge and experience—of the factors that affect readiness, such as training, personnel and equipment. But no single person, or even a group of decision-makers, can fully consider the complex interrelationships of the hundreds, perhaps thousands of variables that may underlie a readiness decision. It's often not enough to know that a unit has a certain level of experience and training—is it the right experience and right training for the mission at hand? What cascading effects would changes to the supply chain have on spare-parts inventories, equipment levels, maintenance cycles—and on overall readiness? How does training on communications equipment affect the ability to direct firepower in different environments? What about the interdependencies between training, personnel, and equipment?

Those are the kinds of questions that high-level diagnostic and predictive analytics can begin to answer. With high-level diagnostics, analysts and decision-makers can quickly model and visualize the complex interrelation of information now isolated in data silos. For example, an analytic might show how 10 factors related to personnel, and 10 factors related to training, work in various combinations to increase or decrease readiness in a particular unit under varying conditions. With predictive analytics, we can turn the dials to see what would happen if we made changes to any one or more of those factors. In addition, such analytics can depict those scenarios in near-real time. This enables decision-makers to spend more time evaluating trade-offs, rather than compiling vast arrays of data into a coherent picture, to support rapid turn-around impact assessments.

UNDERSTANDING READINESS TRADEOFFS

This kind of foresight can help analysts and decision-makers tease out complex tradeoffs. For example, when a ship enters a shore repair facility, many sailors might be assigned to tasks not related to their core skill competencies. When the ship returns to seas months or years later, how much will the sailors' skills have atrophied, both individually and collectively? How is the uniqueness of the ship captured to accurately depict its real readiness? How are other contributing factors, such as unit cohesion, accounted for?

Similarly, when shore repair facilities support is outsourced to contracted personnel—instead of billeting active military personnel into these positions—active-duty personnel take shore duty rotations in billets outside their core competency. Although skills perish, an individual's time-in-rate continues to accrue. What are the best options for weighing the trade-offs? Predictive analytics can answer these kinds of questions by showing what an ideal readiness picture would look like, and how all the various contributing factors need to come together.

Predictive analytics can also help with contingency planning. Analytics might be used, for example, to predict the force structure demand for a particular geographic location over a two-year period—based on information about adversaries and other factors. They can then look at what needs to be done to make sure the military is ready. Analytics can also use post-mission data to provide insight into how well a unit performed, along with lessons learned.

TYING READINESS TO COST

Predictive analytics can show the impact of specific investments or cuts. For example, a certain investment in training could have a high likelihood of increasing readiness. If articulated in a traceable way, linked to readiness impact, this kind of fact-based assessment can help justify budgets and make the case for funding.

At the same time, predictive analytics can show decision-makers how they can save money without impacting readiness. A commander might use the analytic to ask a specific question, such as, "How will deferring maintenance on a particular type of aircraft affect readiness—not just in one unit, but across the fleet in the next two, five, or ten years, or beyond?" Or, the commander might use the analytic in a kind of reverse way—plugging in the desired outcome, and then asking how the various contributing factors would have to be brought together. In this case, the commander might ask, "What are all of the ways I can cut costs without decreasing readiness?"

Predictive Analytics for Logistics Planning

As in other areas of the defense community, it is difficult for logistics planners to make full use of the oceans of data around them. Part of the problem is that the data is widely scattered—it can be found everywhere from data warehouses, to the cloud, to the laptops of personnel who have unique data. But even if commanders were able to put their data all in one place, it wouldn't necessarily tell them how the disparate elements of logistics fit together to create a picture of the whole.

By taking advantage of new approaches of data science like the data abstraction layer, predictive analytics can be used to answer a broad range of logistics-planning questions. How will changes in manpower affect maintenance planning and execution? What kinds of skill sets and training are most critical to the maintenance of specific types of complex equipment and systems? What hidden patterns in usage and repair data can be used to improve the reliability of certain aircraft parts?

INTEGRATING AND ANALYZING ALL OF THE DATA

One example of how predictive analytics might be used in logistics is in planning for and executing maintenance availabilities at public and private shipyards. If an availability is to meet schedule and budget requirements, commanders need to know in advance what maintenance and modernization work needs to be done on the ship, how long it will take, how many artisans will be required in each trade, and how much it will cost. Maintenance data for equipment/systems from previous availabilities can provide valuable insight into these questions. However, the current analytic approaches used by most defense organizations are extremely labor-intensive, and typically provide commanders with only a limited ability to look back at past data to plan for new availabilities. They often have to rely on anecdotal information for guidance.

With predictive analytics, all of the relevant information for individual equipment/systems, such as documented maintenance, observed

reliability, crew training, supply posture, etc., can now be integrated and analyzed. For example, as part of the planning for a surface ship availability, the analytics might look at data for radar system maintenance and modernization for the entire class, as well as all of the casualty reports and other related data, over the past 20 years. The analytics might show that a certain type of repair had to be performed 75 percent of the time—and in addition, how long the work took, and how much it would cost in today's dollars.

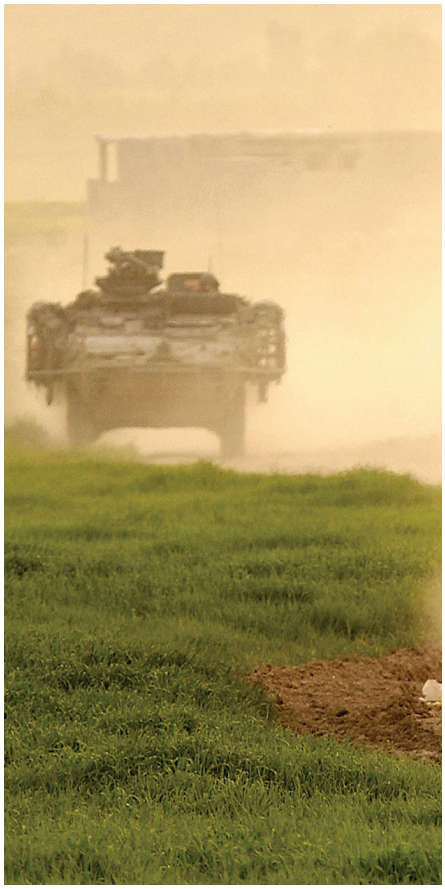
The analytics might take it a step further and find other potential causative factors, such as lack of training, obsolescence, manpower shortages and operational influences on those ships. They might find certain patterns in the 75 percent of the ships that needed the repair, and in the 25 percent that didn't. That kind of information alone could prove valuable—such patterns might suggest causes and effects in system reliability that might be investigated. The analytics could perform the same type of analysis on every type of system on every ship.

USING MATHEMATICAL PROBABILITIES IN PLANNING

By bringing together all of this information, the predictive analytics would give commanders relevant information they need in planning for a specific ship's maintenance availability. The analytics, for example, could provide the probability of every type of maintenance and repair that might be needed—and do the same for the potential modernization work. This would tell commanders—with mathematical precision—how long the upcoming availability is likely to take, how much maintenance and how much modernization are likely to be needed—and the cost of each. The analytics might also provide other logistics detail, such as how many people and what skill sets will likely be needed to perform the work. If commanders are concerned that the work might take too long—delaying the start of the ship's training cycle—they could ask the analytics to tell them what maintenance and modernization work could be deferred with minimal impact, and what the new timelines would be.

The same overall approach can be used throughout logistics planning. For example, advanced analytics could bring together comprehensive data on training, completed maintenance, casualty reports and the operating environment of a strike group to predict what the optimal supply chain would look like, as well as the details of optimal training, maintenance, etc.

These types of examples illustrate the kinds of profound cost savings that high-level diagnostics and predictive analytics can provide in logistics planning. Advanced analytics are not just cost-effective because they substantially reduce the resources needed to conduct data analysis. A far greater potential for cost savings lies in their ability to help organizations to become leaner and more efficient in supporting the warfighter.



Predictive Analytics for Workforce Management

Advanced analytics have the potential to transform how defense organizations carry out manpower planning and other areas of workforce management in all elements of the total force: active duty, Guard and Reserve units, government civilians, and even contractors. Currently, commanders and executives rely mostly on what they already know—what the key factors are in recruitment and retention, for example, or the best ways to prevent sexual assault. But, high-level diagnostic and predictive analytics can tell commanders what they don't know—the interrelationships and hidden causes and effects that can often have an even bigger impact.

IMPROVING RETENTION

One example of how this can work is in retention. The Army typically has to over-recruit by as much a third to account for attrition of personnel over the course of basic training and the first assignment. With current approaches, commanders generally have a limited ability to use data to understand the problem, so they just "live with it" by recruiting more personnel than they need. At best, they may be able to correlate retention with technical aptitude (or person/job "fit"), assessed through vocational and aptitude testing. Such information can be analyzed in relational databases by testing hypotheses—does a particular skill set have an impact on retention?

According to military workforce management experts, however, recruit success has far more to do with leadership, cultural fit and a host of other "soft" factors that are difficult to measure. Such factors can range from personality traits and family background to the culture of the unit and even the personality traits of the recruits' sergeants, which have been shown to play a significant role in retention. Because high-level diagnostic analytics can look at all of these factors—and all at the same time—they can reveal unexpected patterns in why some recruits stay, and others leave.

By analyzing these patterns and how they relate to information about an individual recruit, predictive analytics might provide the mathematical likelihood that the recruit will drop out. If commanders are able to make such an assessment even before the person

signs up, the success rate of new recruits—as measured by first- and second-year retention—could be improved dramatically, which would substantially reduce the need for over-recruiting. And commanders could use the insights gained through the advanced analytics to make changes that will improve retention rates, both in an individual unit and beyond.

ASSEMBLING THE OPTIMAL SUPPORT STRUCTURE

Another way predictive analytics can aid workforce management is in finding the right mix of military, civilian government, and contractor personnel necessary to support particular missions.

In a cybersecurity operation, for example, predictive analytics might start by understanding the nature of the threats against a particular defense organization. By bringing together a broad array of data, the analytics could create a picture of the potential attackers and their methods, where they're likely to attack, what vulnerabilities the defense organization has, and to what extent potential attackers may be able to exploit those vulnerabilities to mount a successful attack. At the same time, the analytics could look across the defense community at the mix of cybersecurity personnel, skills, abilities and experience that have proven to be the most effective in preventing and responding to attacks. By combining and analyzing all of this information, the predictive analytics would show what the optimal cybersecurity team would look like for the specific organization. The analytics could also factor in individual and other costs, to support requests for funding, as well to help commanders stay within budget constraints.

DEALING WITH COMPLEX WORKFORCE MANAGEMENT ISSUES

Predictive analytics can be particularly valuable in helping commanders address complex issues that are beyond the scope of traditional data analysis. For example, it can be difficult to identify military units that have a higher risk of sexual assaults. Advanced analytics could begin by looking at the organizational and environmental factors that were present in units across the military where sexual assaults occurred—and where they didn't. The factors might include the demographic makeup of units and their leaders, for instance—the mix of gender, age, diversity, experience, etc.—or the nature of the mission (e.g., did the assaults occur in support or front-line organizations?)

The analytics might find patterns that, when applied to specific units, could show which ones have a higher risk of assaults. Such analytics could also show which interventions are most likely to be successful, both in the individual units being analyzed, as well as across the military as a whole. This analytic approach could also be used to address other difficult workforce management issues, such as suicide prevention, discrimination complaints and violence in the workplace.

Predictive Analytics for Military Intelligence

Predictive analytics enable military intelligence analysts to rapidly sort through vast amounts of diverse data to find hidden patterns and connections, and then make that insight actionable. This is a significant advance over current analytic approaches. Such methods require analysts to choose a limited amount of data to work with, and then to pose and test hypotheses about what those connections might be—an often difficult and time-consuming task.

With advanced analytics, all of the data is securely brought together and made available for analysis. In place of the hypothesis-based approach, the analytics let the data speak for itself. The patterns and connections that emerge can then be used to guide action.

Predictive analytics, for example, have helped U.S. military convoys avoid roadside bombs and suicide bombers in Afghanistan. With older methods, intelligence analysts choosing the best convoy routes didn't have the time to analyze all the data. They were only able to pick the datasets they thought might be most valuable. Much of the data had to be set aside, particularly unstructured data, such as cellphone chatter and video images, that was difficult to format. Using the new analytic approaches, however, the analysts were able to bring all the data together—both structured and unstructured—to see the hidden patterns. This gave them the ability to predict which sections of a town—and even which roads and intersections—carried the greatest probability of an attack. Analysts then created "heat maps" that showed the danger zones, and enabled them to map out the best routes, turn-by turn.

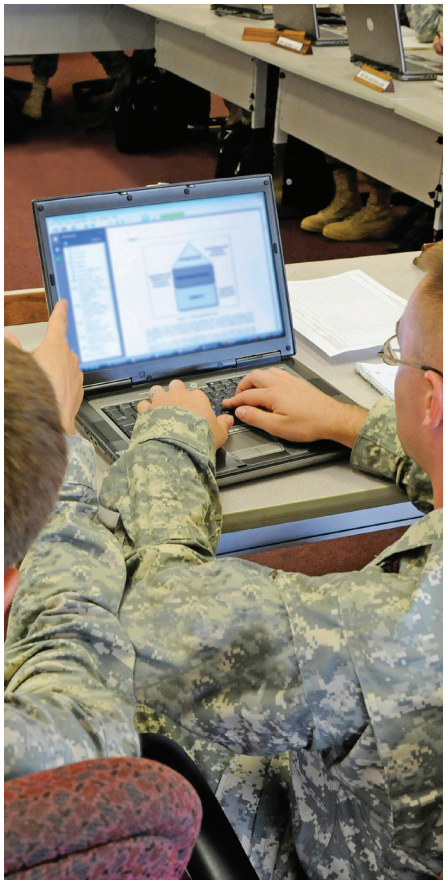
Predictive analytics have a broad range of applications in military intelligence. For example, they can be used to give analysts better insight—and foresight—into how terrorist groups transfer and spend money. With conventional approaches, analysts are often limited to

"snapshots" of the data—they may see a transfer of money, but not the larger picture of where it fits in. Predictive analytics are able to bring together all of the available information about the terrorist groups, as well as other networks and individuals they might associate with. The analytics can combine that with current and past financial and banking data, along with information on weapons suppliers, fraud, drug trafficking and the many other areas associated with terrorist funding.

Using machine learning and other tools, the analytics would then start mapping out webs of connections in the data, developing insight into how the terrorists groups have been moving money—where it has come from, where it has gone, and what it has been used to buy. Based on this information, the analytics might also be used to track current money movements—predicting where the money will likely end up, and how it will be spent. The analytics could also look ahead, and provide the probability, for example, that terrorists will move money from one bank to another during the next 10 days, with the intent to buy certain kinds of weapons. In addition, the analytics may show which types of interdictions or other actions are likely to be most effective in preventing or stopping a transaction.

Predictive analytics can also help the military intelligence community keep pace with a rapidly changing threat landscape. For example, if an adversary takes certain actions—such as military activity, cyber warfare or psychological operations, the analytics can be used to predict what additional intelligence resources the military might need. This might include, for example, a mix of new analysts, translators, regional experts, intelligence sources, or analytics tools. The same predictive approach might help answer many other intelligence questions, from whether to launch satellites to what kind of fighter jet will be needed in 10 years.

Military intelligence organizations may be concerned that integrating large amounts of data may make it less secure. But the advanced analytic techniques actually make the data far more secure. These techniques build in data visibility and control at a granular level—down to the image, sentence, and even word. This makes it possible for organizations to bring together data from every available source for predictive analytics.



HOW DATA SCIENCE WORKS: A BRIEF OVERVIEW

Data science is the art of turning data into action. This is accomplished through the creation of data products, which provide actionable information without exposing decision makers to the inner workings of the analytics.

Examples of data products in the defense community include answers to questions such as, What cuts can I make in training that will have little or no impact on readiness? What mix of disciplines, skill sets and experience do my civilian workers need to perform the necessary tasks? Where are the vulnerabilities in my cyber networks, and which ones put me at most risk?

WHAT MAKES DATA SCIENCE DIFFERENT

Data science supports and encourages shifting between deductive (hypothesis-based) and inductive (pattern-based) reasoning. This is a fundamental change from traditional analytic approaches. Inductive reasoning and exploratory data analysis provide a means to form or refine hypotheses and discover new analytic paths. In fact, to discover the significant insights that are the hallmark of data science, you must have both the tradecraft and the interplay between inductive and deductive reasoning. By actively combining the ability to reason deductively and inductively, data science creates an environment where models of reality no longer need to be static and empirically based. Instead, they are constantly tested, updated and improved until better models are found. These concepts are summarized in the figure below:

THE TYPES OF REASON...

DEDUCTIVE REASONING

- Commonly associated with "formal logic."
- Involves reasoning from known premises, or premises presumed to be true, to a certain conclusion.
- The conclusions reached are certain, inevitable, inescapable

INDUCTIVE REASONING

- Commonly known as "informal logic," or "everyday argument."
- Involves drawing uncertain inferences, based on probabilistic reasoning.
- The conclusions reached are probable, reasonable, plausible, believable.

AND THEIR ROLE IN DATA SCIENCE TRADECRAFT.

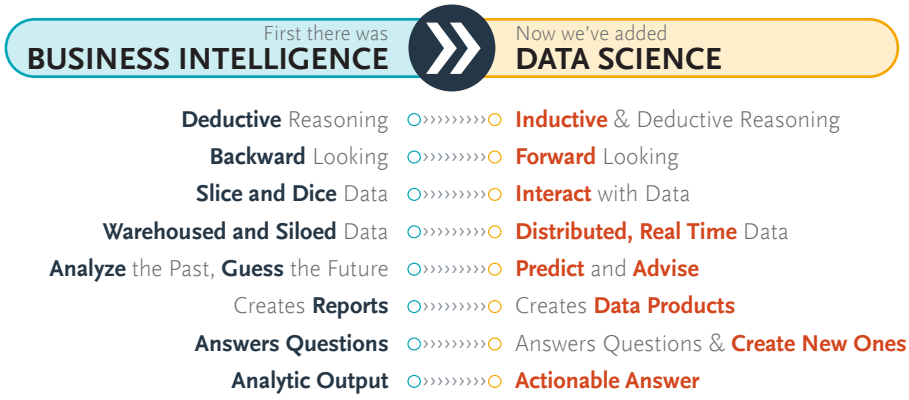
- Formulate hypotheses about relationships and underlying models
- Carry out experiments with the data to test hypotheses and models
- Exploratory data analysis to discover or refine hypotheses.
- Discover new relationships, insights, and analystic paths from the data.

The differences between data science and traditional analytic approaches do not end at seamless shifting between deductive and inductive reasoning. Data science offers a distinctly different perspective than capabilities such as business intelligence. Data science should not replace business intelligence functions within an organization, however. The two capabilities are additive and complementary, each offering a necessary view of business operations and the operating environment. The figure Business Intelligence and Data Science—A Comparison highlights the differences between the two capabilities.

Key contrasts include:

- **Discovery vs. Pre-canned Questions:** Data science actually works on discovering the question to ask as opposed to just asking it.
- **Power of Many vs. Ability of One:** An entire team providing a common forum for pulling together computer science, mathematics and domain expertise.
- **Prospective vs. Retrospective:** Focused on getting actionable information from data as opposed to historical reporting.

LOOKING BACKWARD & FORWARD



THE IMPACT OF DATA SCIENCE

The way organizations make decisions has been evolving for half a century. Before the introduction of business intelligence, the only options were gut instinct, loudest voice, and best argument. Although those methods still exist, most organizations now inform their decisions with real information through the application of simple statistics. However, we are outgrowing the ability of simple stats to keep pace with military decision-making demands. The rapid expansion of available data, and the tools to access and make use of the data at scale, are enabling fundamental changes to the way organizations make decisions.

Data science is required to keep pace with the increasingly data-rich environment. Much like the application of simple statistics, organizations that embrace data science will be rewarded while those that do not will risk falling behind. As more complex, disparate datasets become available, the chasm between these groups will only continue to widen.

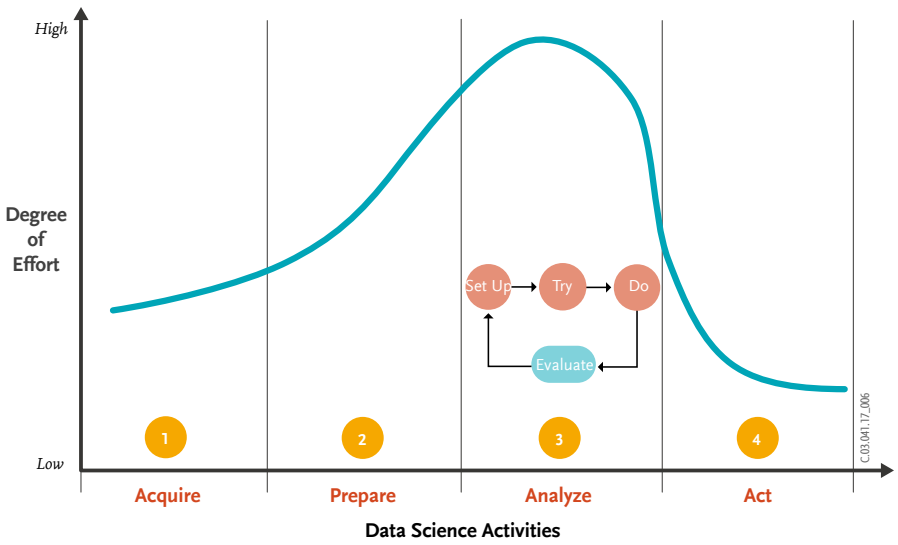
WHAT IS DIFFERENT NOW

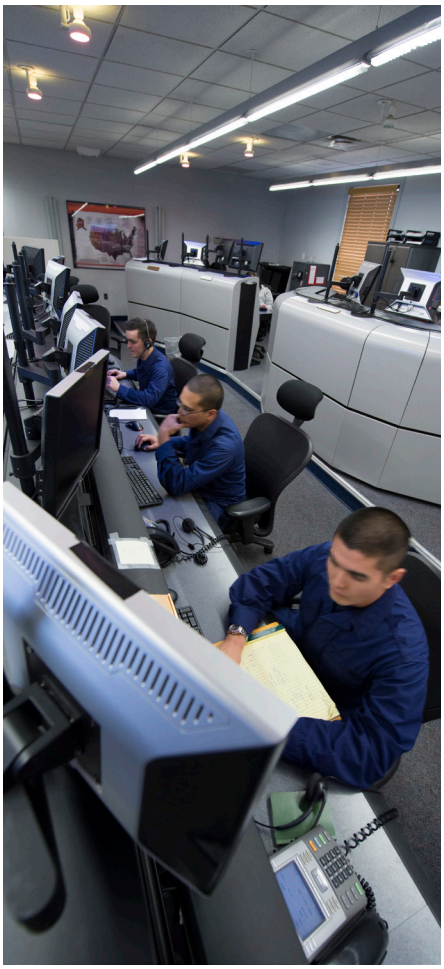
For 20 years, IT systems were built the same way. We separated the people who ran business and operations from the people who managed the infrastructure (and therefore saw data as simply another thing they had to manage). With the advent of new technologies and analytic techniques like the data abstraction layer, this artificial—and highly ineffective—separation of critical skills is no longer necessary. For the first time, organizations can directly connect decision-makers to the data. This simple step transforms data from being "something to be managed" into "something to be valued."

In the wake of the transformation, organizations face a stark choice: you can continue to build data silos and piece together disparate information or you can access your data with a data abstraction layer and distill answers. From the data science perspective, this is a false choice: The siloed approach is untenable when you consider the (a) the opportunity cost of not making maximum use of all available data to help an organization succeed, and (b) the resource and time costs of continuing down the same path with outdated processes.

THE FOUR KEY ACTIVITIES OF DATA SCIENCE

Data science is a complex field. It is difficult, intellectually taxing work, which requires the sophisticated integration of talent, tools and techniques. It can be understood through the four key data science activities show on the chart below:





Activity I: Acquire

All analysis starts with access to data, and for the data scientist this axiom holds true. But there are some significant differences—particularly with respect to the question of who stores, maintains and owns the data in an organization.

Traditionally, the rigid data silos artificially define the data to be acquired. Stated another way, the silos create a filter that lets in a very small amount of data and ignores the rest. These filtered processes give us an artificial view of the world based on the 'surviving data,' rather than one that shows full reality and meaning. Without a broad and expansive dataset, we can never immerse ourselves in the diversity of the data. We instead make decisions based on limited and constrained information.

Eliminating the need for silos gives us access to all the data at once—including data from multiple outside sources. It embraces the reality that diversity and complexity are good. This mindset creates a completely different way of thinking about data in an organization by giving it a new and differentiated role. Data represents a significant new mission-enhancement opportunity for organizations.

Keys to Acquire:

- Look inside first—what data do you have current access to that you are not using? Data you have left behind during the filtering process may be incredibly valuable.
- Remove the format constraints—stop limiting your data acquisition mindset to the realm of structured databases. Instead, think about unstructured and semi-structured data as viable sources.
- Figure out what's missing—ask yourself what data would make a big difference to your processes if you had access to it, then find it.
- Embrace diversity—try to engage and connect to publicly available sources of data that may have relevance to your domain area.

Activity 2: Prepare

Once you have the data, you need to prepare it for analysis. Organizations often make decisions based on inexact data. Data stovepipes mean that organizations may have blind spots. They are not able to see the whole picture and fail to look at their data and challenges holistically. The end result is that valuable information is withheld from decision-makers.

When data scientists are able to explore and analyze all the data, new opportunities arise for analysis and data-driven decision making. The insights gained from these new opportunities will significantly change the course of action and decisions within an organization. Gaining access to an organization's complete repository of data, however, requires preparation.

Experience has shown that the best tool for data scientists to prepare data for analysis is the approach known as the data lake. Instead of storing information in discrete data structures, the data lake consolidates an organization's complete repository of data in a single, large view. The entire body of information is available for every inquiry—and all at once. The data laken Layer eliminates the expensive and cumbersome data-preparation process—necessary with data silos—that can take up as much as 80 percent of analysts' time.

Activity 3: Analyze

Once the data is acquired and prepared, it is ready to be analyzed. This activity requires the greatest effort of all the activities in a data science endeavor. The data scientist actually builds the analytics that create value from data. Analytics in this context are iterative applications of specialized and scalable computational resources and tools used to provide relevant insights from exponentially growing data. This type of analysis enables real-time understanding of risks and opportunities by evaluating situational, operational and behavioral data.

With the totality of data fully accessible by a data abstraction layer, organizations can use analytics to find the kinds of connections and patterns that point to promising opportunities. This high-speed analytic connection is done with a data abstraction layer, as opposed to traditional style sampling methods that can only make use of a narrow slice of the data. Those methods require us to pull data out of different dataset—essentially bringing the data to the analytics. But with the data abstraction layer, we can bring the analytics directly to the data, a much faster and more efficient process.

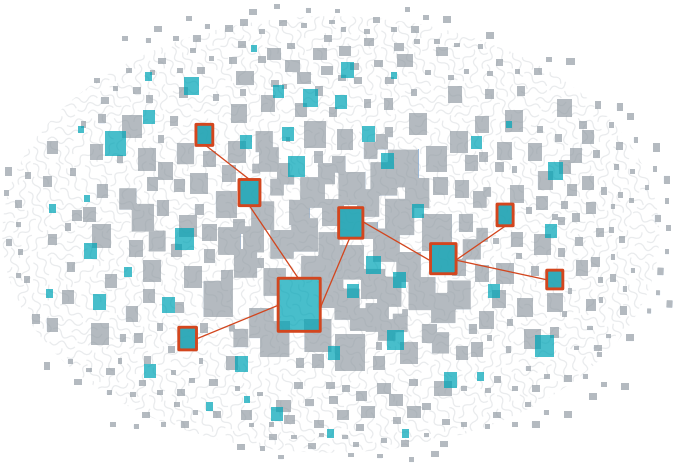
The figure (right) illustrates the concept of discovering new connections and patterns with the data abstraction layer.

Data scientists work across the spectrum of analytic goals—Describe, Discover, Predict and Advise. The maturity of an analytic capability determines the analytic goals encompassed. Many variables play key roles in determining the difficulty and suitability of each goal for an organization. Some of these variables are the size and budget of an organization and the type of data products needed by the decision-makers. In addition to consuming the greatest effort, the Analyze activity is by far the most complex. The tradecraft of data science is an art.

Activity 4: Act

The ability to make use of the analysis is critical. It is also very situational. Below are some key points to keep in mind when using analytics in decision-making:

1. The findings must make sense with relatively little up-front training or preparation on the part of the decision-maker.
2. The findings must make the most meaningful patterns, trends and exceptions easy to see and interpret.
3. Every effort must be made to encode quantitative data accurately so the decision maker can accurately interpret the data and easily compare them to one another.
4. The logic used to arrive at the finding must be clear and compelling as well as traceable back through the data.
5. The findings must be able to drive mission success.

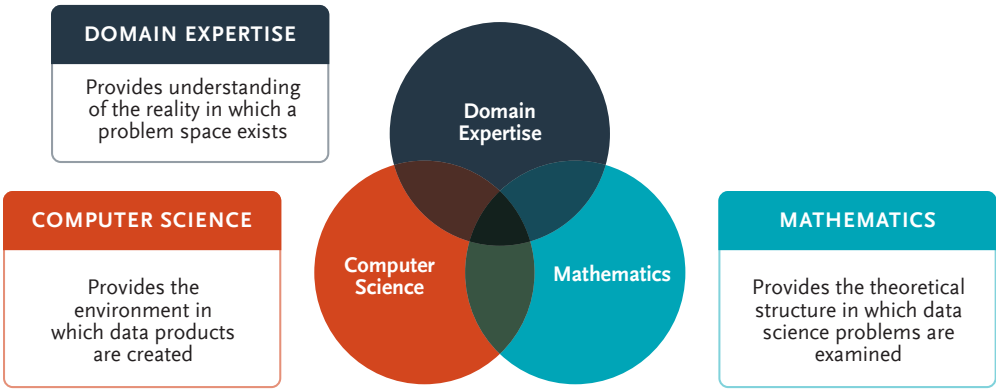


Analytic connection in the data abstraction layer

HOW DATA SCIENCE COMES TOGETHER

Data science requires proficiency in three foundational technical skills—computer science, mathematics and domain expertise (as shown in the figure below). Computers provide the environment in which data-driven hypotheses are tested, and as such computer science is necessary for data manipulation and processing. Mathematics provides the theoretical structure in which data science problems are examined. A rich background in statistics, geometry, linear algebra, and calculus are all important to understand the basis for many algorithms and tools. Finally, domain expertise contributes is needed to understand the context of the data, as well as what problems actually need to be solved, what kind of data exists in the domain, and how the problem space may be instrumented and measured. Individuals who excel in all three areas are rare. Typically, data science requires teams of people to bring together the necessary expertise.

The World of Data Science

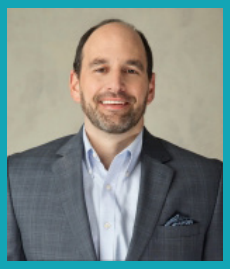


LOOKING AHEAD

The defense community is well positioned to move to predictive analytics. Defense organizations possess the necessary fuel—their massive data stores—and have made rapid progress in data-driven decision-making. By adopting recent advances in data science technologies and approaches, defense organizations can make a smooth, cost-effective transition from hindsight to foresight.

Booz Allen Hamilton worked with the intelligence community—the earliest government adopter of advanced analytics—to develop many of these data science breakthroughs, which have been proven in the fight against terrorism. We are now helping forward-looking organizations across government leverage predictive analytics that are technologically feasible, practical, cost-effective and secure. The time is right for the defense community to capitalize on this progress—and harness predictive analytics to support the warfighter, and drive mission success, in powerful new ways.

FOR MORE INFORMATION



Mark "Jake" Jacobsohn
Senior Vice President
jacobsohn_mark@bah.com
(703) 902-5290

Scott Jachimski
Principal
jachimski_scott@bah.com
(703) 559-6019

About Booz Allen

For more than 100 years, business, government, and military leaders have turned to Booz Allen Hamilton to solve their most complex problems. They trust us to bring together the right minds: those who devote themselves to the challenge at hand, who speak with relentless candor, and who act with courage and character. They expect original solutions where there are no roadmaps. They rely on us because they know that—together—we will find the answers and change the world. To learn more, visit BoozAllen.com.