Booz | Allen | Hamilton

DATA SCIENCE PLAYBOOK

Version 2.1

CONSULTING | ANALYTICS | DIGITAL SOLUTIONS | ENGINEERING | CYBER

Table of Contents

1 | INTRODUCTION

We live in a data revolution. Data is growing at an exponential rate. Each time we check the weather on our phones, drive using a navigation app, pay with a credit card, update our social media, or catch up on a Netflix show, we generate data that can be used in meaningful ways. This is just data generated from our personal lives. Imagine how much more data organizations collect that monitor facilities usage, air quality, security, technology, company vehicles, and budgets – and that doesn't include data collected on customers. This vast quantity of data has the power to provide immense value and it is the key to staying competitive in any market or industry.

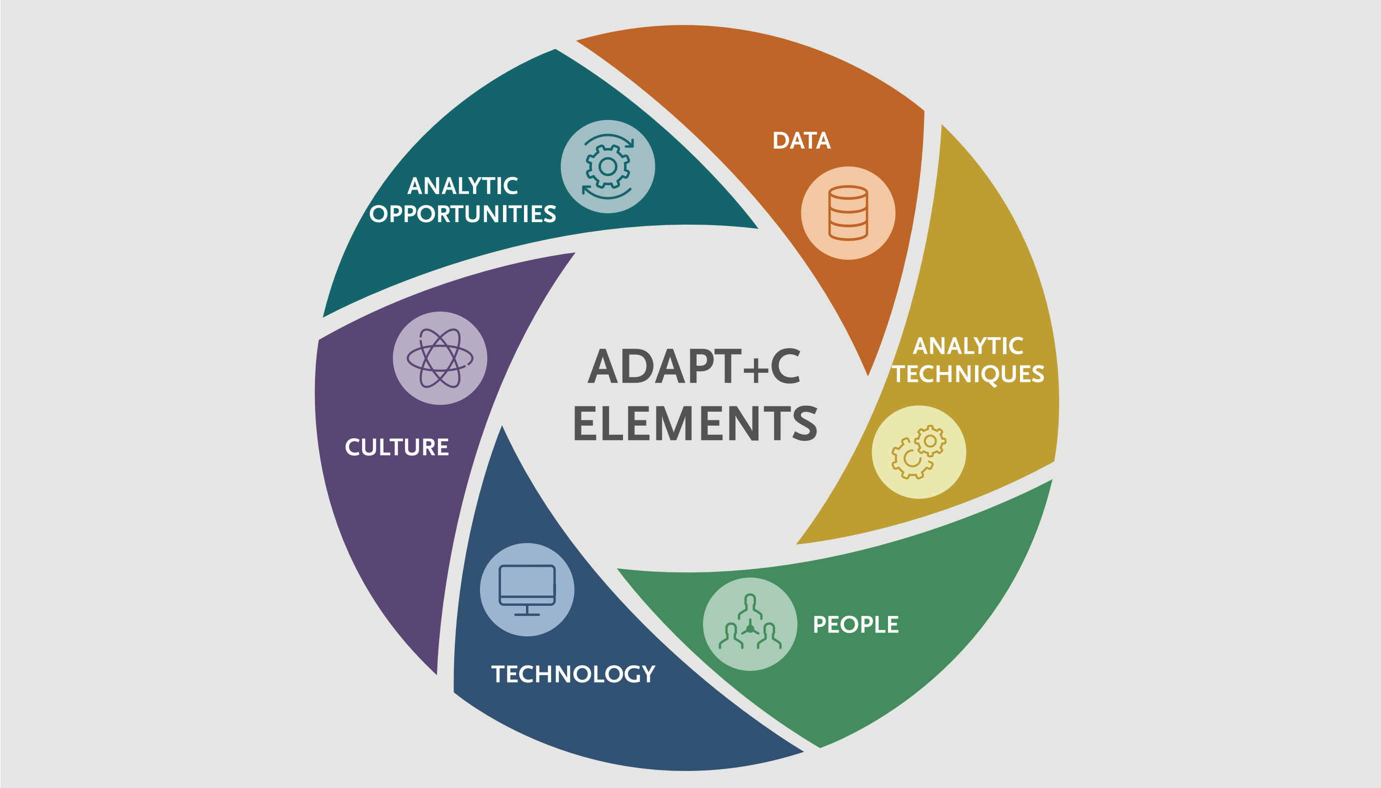
The time is now – organizations must transform themselves through data science. Data science is the art of turning data into insights to inform decisions. Effective data science can bring higher productivity, greater returns, and overall improved performance to organizations. Many organizations have been collecting data for a long time, which means they are now well-positioned to use data science to drive mission and business success. A focus on data science helps organizations understand past events, predict future trends, and prescribe optimal courses of action. Ultimately this will lead to improved operations, differentiated customer value propositions, and competitive advantage.

So where does one start down the path to effective data science? The solution isn't easy. Some groups learn the hard way. They make large investments in technology or hire data science teams that promise the world, but can't deliver. Others make smaller mistakes, but still must course correct along the way. All come to the same conclusion – there is no one-size-fits-all approach. But, we have the benefit of learning from the innovators and early adopters who have already navigated the journey to data science success.

Figure 1: Benefits of Data Science

BENEFITS OF DATA SCIENCE	
17-49%	...increase in productivity when organizations increase data usability by 10%
11-42%	... return on assets (ROA) when organizations increase data access by 10%
241%	... increase in ROI when organizations use big data to improve competitiveness
1000%	... increase in ROI when deploying analytics across most of the organization, aligning daily operations with senior management's goals, and incorporating big data
5-6%	... performance improvement for organizations making data-driven decisions

Figure 2: ADAPT+C Elements



After working with a large number of public and private sector organizations, Booz Allen learned that data science depends on six critical elements:

Analytic Opportunities: Considers new and existing use cases to apply data science to improve organization mission and operations

Data: Considers opportunities to use new and existing data sets and better manage and govern data in support of data science projects

Analytic Techniques: Considers the analytic tradecraft and techniques to be applied to generate insights from data

People: Considers the set of human capital programs required to develop a talented and capable team of data science practitioners

Technology: Considers the optimal ways to use existing and new technologies including applications, data platforms, and infrastructure to perform data science projects

Culture: Considers the set of mechanisms that communicate, share, and reinforce the value of data science across an organization to change the behavior of the staff

Effective data science organizations build, grow, and reinforce their data science capabilities by prioritizing these six "**ADAPT+C**" elements over time. This Data Science Playbook dives into each of the six elements, sharing Booz Allen's lessons learned from successful organizations and describing relevant tools, techniques, and resources to make the most out of data science within your organization.

Effectively developing or growing data science within your organization is no small feat. It may seem overwhelming at times, but remember, it all boils down to just six elements, ADAPT+C. Regardless of where it began, any organization that has successfully transformed knew where it wanted to go. Set a vision and go after it!

2 | ANALYTIC OPPORTUNITIES

+ *Who in your organization is thinking about how data science can help it operate better, smarter, faster, or more efficiently?*
+ *How do employees notify leadership of ideas for operational improvements?*
+ *How are proposals for new analytic solutions reviewed for investment?*
+ *Is someone researching new capabilities that can drive better decision-making?*
+ *Who evaluates the effectiveness of ongoing analytics projects?*
+ *How do you capture lessons-learned from your ongoing analyses?*

Chapter Preview:

Data science is only as valuable to an organization as the questions that it can help answer. The answers to these questions may result in operational efficiencies, better market sensing, higher quality service to the customer, or nothing at all, but an organization sees no return if it makes no investment. So, how should you cultivate and implement opportunities for data science?

Fundamentally, it begins with the right business questions and a strategic process to turn those questions into realized business outcomes through analytics. The opportunities can emerge from business pressures, market pressures, or just organizational curiosity. When they emerge, an organization can scope and transform business questions into analytics projects through the same rigor as any other project.

Booz Allen helps clients identify business questions and develop a process for turning ideas into solutions. The most successful organizations institute formalized ways to manage each phase of a data science project: soliciting and reviewing new ideas, managing their portfolio, executing projects, and implementing solutions. These formal mechanisms include clearly defined roles and responsibilities for idea creation, idea filtering, and solution management. Booz Allen ensures organizations harness the combined originality of their people with the strategic decision-making of their leaders to use data science for a competitive advantage.

OPPORTUNITY DEVELOPMENT: HOW YOU FIND WHAT YOU NEED

Developing the right business questions for data science is the key to success. Often, the best ideas for data science projects sit deep within the organization. Employees who know the business often know what questions to ask and what insights the data may reveal. An effective organization brings these ideas to the foreground and prioritizes those ideas to best meet strategic objectives.

Figure 3: How an Idea Becomes a Project

Ideas are conceptual notions from employee of how analytics can be applied to a business challenge

Opportunities represent a set of ideas that have been further explored and reviewed by a team of experts. They represent actual opportunities to apply analytics to solve a particular business issue

Projects are vetted analytics opportunities that have been selected for further development and application of data science

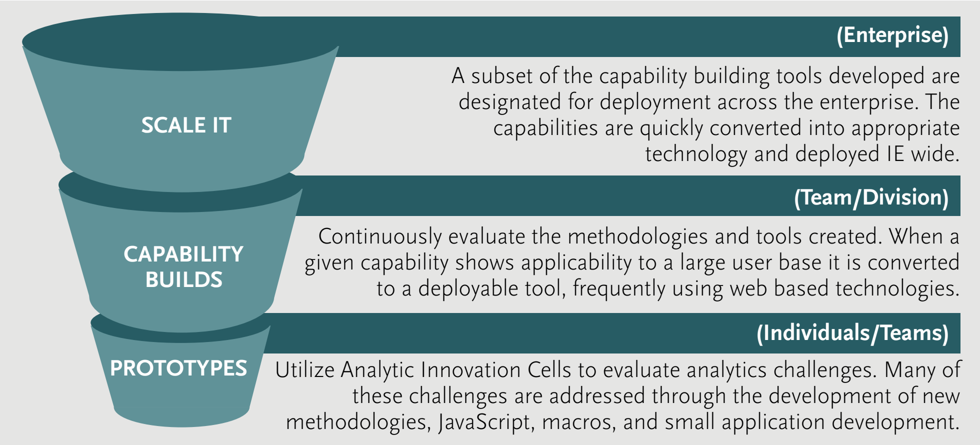
Case Study: Optimizing Operations for Immigration and Customs Enforcement (ICE) AIR Operations (IAO)

IAO regularly repatriates removed persons back to their home countries via owned and chartered aircraft. Due to the ever-changing removal volumes, destination countries, and departure cities throughout the country, most flight scheduling was managed manually, which was inefficient and led to operational challenges. Booz Allen built a model and visualization tool using ArcGIS to apply dynamic multi-period, multi-vehicle, multi-hub, and holding-cost concepts from traditional vehicle routing problems to determine an optimal flight plan. To effectively implement and integrate the analytic solution, the project included a simple training course and user guides. These tools aided the usage and adoption of the tool within ICE. With these resources, IAO analysts were able to use the tool to take advantage of underutilized charter planes and better automate the scheduling and routing processes.

Opportunity Identification

There are direct and indirect ways to identify analytic opportunities. One direct way is to create a centralized group who scans organizational data to find opportunities. This method relies on the group's ability to discover business questions in the data. Booz Allen offers a proven method for a centralized data science group called the **Analytic Innovation Cell (AIC)**. An AIC can be particularly effective for organizations lacking upfront investment in data science capability. These organizations may not have the advantage of robust data or sophisticated tools for analysis. In this type of organization, the AIC can both identify the opportunities and create the solutions. An AIC is particularly valuable to an organization that wants to demonstrate the immediate value of data science. This is because the AIC process uses the data and desktop tools available to develop quick wins. Further, the AIC approach uses "lowest common denominator" technology, so that its rapid prototypes are immediately implementable to the end user.

Figure 4: Analytic Innovation Cells



A subset of the capability building tools developed are designated for deployment across the enterprise. The capabilities are quickly converted into appropriate technology and deployed IE wide.

Continuously evaluate the methodologies and tools created. When a given capability shows applicability to a large user base it is converted to a deployable tool, frequently using web based technologies.

Utilize Analytic Innovation Cells to evaluate analytics challenges. Many of these challenges are addressed through the development of new methodologies, JavaScript, macros, and small application development.

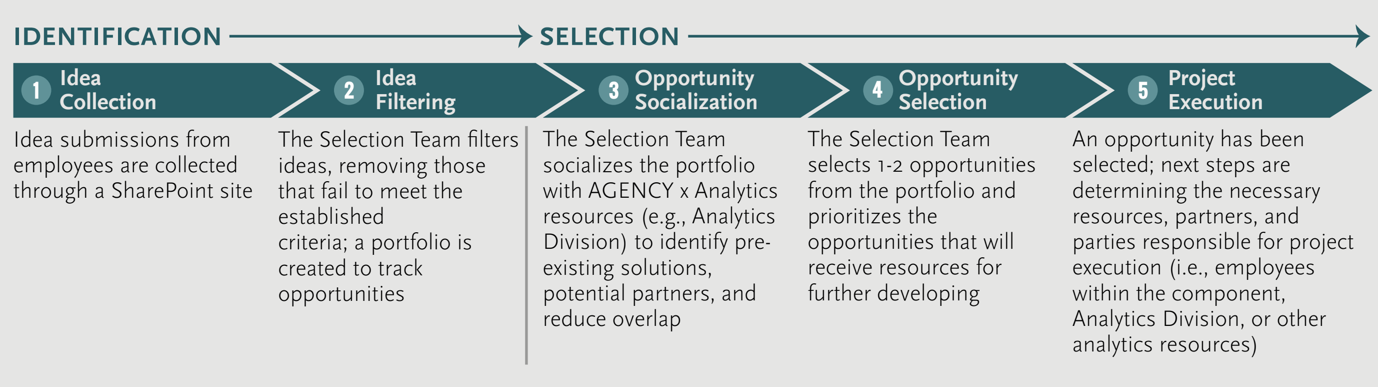
An indirect way to gather analytic opportunities can be through open forums and digital crowdsourcing tools. Through these mechanisms, all employees (Data Scientists or not) can submit ideas or suggestions to improve the organization. This method provides a steady stream of new ideas for organizations to consider and fosters staff engagement around data science. This approach works effectively in organizations where there is a pre-existing group or a network of analysts that can triage the ideas collected from an organization. Within Booz Allen, the Garage serves as a crowdsourcing tool to help connect problems to the vast network of problem-solvers employed by the firm. Employees can post challenges (business questions), and other employees can submit ideas (creative solutions for how they might be solved). Ideas are then voted on, experts select the best ideas, and the most viable ones receive investment for execution.

Opportunity Selection

Much like cutting-edge technology firms have internal incubators and venture programs, leading organizations have a formal process to invest in the best opportunities to use data science to drive the business. After developing a pipeline of ideas, either directly or indirectly, organizations should evaluate those opportunities. The criteria to prioritize and select the opportunities should include factors such as: cost, likelihood of success, and relevance and impact to the organization's mission and near-term goals. Organizations often say "we want to use our data" without any clear assessment of what questions to ask and whether those questions help the organization meet its strategic goals. A methodical review based on formal selection criteria is the most effective way to select analytics projects.

Booz Allen uses an **Analytics Opportunities Toolkit** to provide clients with the tools to determine which analytic opportunities are best suited to their individual needs. Our five-step Toolkit includes advice, best practices, templates, and detailed guidance on how to identify and select the most valuable analytic opportunities for an organization's unique needs.

Figure 5: Analytic Opportunities Process



1 Idea Collection: Idea submissions from employees are collected through a SharePoint site

2 Idea Filtering: The Selection Team filters ideas, removing those that fail to meet the established criteria; a portfolio is created to track opportunities

3 Opportunity Socialization: The Selection Team socializes the portfolio with AGENCY x Analytics resources (e.g., Analytics Division) to identify pre-existing solutions, potential partners, and reduce overlap

4 Opportunity Selection: The Selection Team selects 1-2 opportunities from the portfolio and prioritizes the opportunities that will receive resources for further developing

5 Project Execution: An opportunity has been selected; next steps are determining the necessary resources, partners, and parties responsible for project execution (i.e., employees within the component, Analytics Division, or other analytics resources)

PROJECT MANAGEMENT: HOW YOU ENSURE SUCCESS

As analytic opportunities emerge, they need to be treated with the same project rigor as any other investment in an organization. Organizations must structure project management processes against analytics objectives, success criteria, realistic timelines, and proper resources.

Project Planning & Oversight

As an organization begins down the path of data science, it runs the significant risk of overcommitting and underdelivering. This will create unnecessary concern about data science rather than generate the rush that comes from early victories. Even for organizations with significant project management experience, managing data science projects can be a new challenge. Do not skip the critical project planning steps of defining the scope, cost, and timeline for each data science project.

Alternatively, many data scientists have begun using a Development Operations (DevOps) approach to their projects, working in sprints and relying on Agile methodologies. This process provides project oversight that encourages data discovery and exploration because failure happens quickly and data science teams can quickly course correct when they identify a roadblock. While either the traditional or Agile approach can work for data science, proper project management is necessary to execute projects successfully.

Project Portfolio Review

In addition to the management of single projects, organizations must regularly review the enterprise-wide portfolio of data science projects. An analytics portfolio review maintains the alignment of all analytical projects against the organization's analytical objectives. Moreover, it limits the "one off" projects and helps to break down potential data silos across the organization. An effective portfolio management review should consider three key factors:

- **Project allocation**, which determines the long-term mix of projects
- **Diversification**, which limits exposure to risk
- **Rebalancing**, which regularly revisits the portfolio of projects to ensure they continue to meet the goals determined up front

Effective portfolio management may also lead to a reduction in overall data science investment, as it helps to identify the existing "pockets of greatness" within the organization. Pockets of greatness are small groups, teams, or divisions that have strong capabilities in certain areas that may be able to provide greater resources to the enterprise at large upon discovery. For example, a portfolio review within an organization might identify a pocket of highly specialized skills in survey design, analysis, and statistical modeling that can be deployed throughout the organization, avoiding duplicative work and supporting future endeavors.

Case Study: Identifying Fraud for Government Client

Over the course of five years, Booz Allen assisted a government client to stop a large volume of fraudulent payments. The firm delivered strategy, data science, and business process re-engineering services to combat fraud associated with stolen personally identifiable information (PII). Booz Allen developed, tested, and supported the implementation of a variety of data-driven detection algorithms including decision trees, neural networks, regression models, and customized clustering mechanisms. The team continued to apply these anomaly detection techniques to scan for new and evolving fraud schemes by leveraging a combination of internal and external open source data.

Project Resource Management

To complete an analytics project on time, it must have the right resources. Properly resourcing projects—in terms of time, money, and people—is a requirement for any organization building a data science capability. Many organizations overlook the importance of an ongoing resource review. Data science is not a one-time effort, but instead, a long-term investment in organizational decision-making. The best analytics solutions are inherently iterative, and the best portfolio will include some projects that fail. Analytics project failure shows that the organization is taking risks.

As a result, projects need continuous review to determine if they are reducing workload, providing valuable insights, and are staffed and resourced at the right level. Some analytics projects require heavy resourcing and personnel at the outset, but less personnel once operational. Other projects ramp-up over time, requiring only a few analysts to get off the ground, but need increased support as the projects scale up. Data science organizations constantly evaluate their projects' resourcing needs to ensure success and good stewardship of valuable resources. An organization's existing project management and portfolio management tools will easily meet the needs of data science project and portfolio management. Booz Allen uses the proven PMI Project Management Approach and our own **Portfolio Management Approach** to catalogue, categorize, and prioritize analytics projects to ensure they are scheduled and resourced in alignment with overall organizational goals.

SOLUTION EXECUTION: HOW YOU IMPLEMENT & IMPROVE

Data-driven organizations recognize that the solutions that support decision-making are just as dynamic and ever changing as data itself. Identifying what success looks like at the end of an analytics project can be an important shared activity for the data scientists and end users to norm around. But don't define success by the outcome of the quantitative analysis, define success by the organizational changes that result with the quantitative analysis. Prototyping and pilot testing both the solutions, as well as the decision-making, can provide important information to the data science team for future projects and provide valuable information to improve the existing solution.

Solution Implementation & Integration

Many organizations think of analytics solutions as dashboards and on-going reports— tools that just need to be updated with new data. However, effective analytic solutions are living tools that support decision-making. As such, users must understand the intended value and use of each analytics solution. Even if the end user originally identified the analytic opportunity and managed its execution, those efforts (and expenses) will have been wasted if the organization does not integrate the analytics solution into its regular operations. For analytics teams that are customer service providers to their broader organizations, Booz Allen recommends developing a Project Scoping Agreement or Service Agreement, where the end user agrees to implement the solution upon receipt. Otherwise, how can the analytics team prove its value back to leadership? Many of the principles that support the deployment or adoption of new technologies apply to analytics solutions as well (i.e., employee training, clear communications, and an implementation plan). Organizations need to formalize the execution of analytics solutions to capture their potential value.

Solution Evaluation

Data science organizations focus on the evaluation of analytics as much as the development of analytics. As living tools that support decision-making, effective analytics require constant review, revision, and even re-imagining. The program evaluation process brings two benefits: 1) that analytics never become stale and always maintain the latest operational thinking of the organization, and 2) that analytics become institutionalized and live beyond just a single advocate or stakeholder.

Strong program evaluation requires agreed-upon project goals so that once implementation begins, periodic reviews can check the status against stated benchmarks. Additionally, these benchmarks help justify long-term investments in analytics programs.

The Booz Allen Difference

Booz Allen data scientists develop some of the most sophisticated and cutting-edge analytics tools available to public and private sector organizations. We also know that those tools are for naught if they aren't used for decision-making. Identifying, managing, and integrating analytics is critical to building a data science capability. But, each organization is unique in that there is no one-size-fits-all plan for exploiting data science opportunities. For this reason, we use customizable frameworks and modifiable approaches to meet our clients' needs.

Organizations do not need to take a "0 to 60" approach to data science. It can seem overwhelming, but with Booz Allen's help, clients can be assured they are using their analytics resources most efficiently to drive enhanced business and operational decision-making.

CHALLENGE	BOOZ ALLEN SOLUTION	DESCRIPTION
Don't know how to apply analytics in your organization	**Analytic Innovation Cell**	SWAT team of data scientists and developers removed from the daily grind of reporting and requests for information, focused entirely on exploratory analysis of existing data sets and rapid prototyping to build pilot analytics solutions within a matter of days and put it in the hands of the end user
Don't know how to choose which analytics opportunities to pursue	**Analytics Opportunities Toolkit**	Collection of best practices and step-by-step advice to help identify and select projects that will yield the maximum analytics value
Need to prioritize and maximize resource allocation among multiple analytics projects	**Analytic Portfolio Management Approach**	Strategic approach to catalogue, categorize, and prioritize the analytics portfolio to ensure projects are scheduled and resourced in alignment with overall organizational goals

3 | DATA

- *Where do you begin when thinking about your organization’s data?*
- *Does your organization have a vision for how to use its data?*
- *Who is thinking about how to maximize the potential of your data?*
- *Is your data regularly used to drive both strategic and operational decisions?*
- *Do your employees all treat data the same way?*
- *Does your organization have data definitions and categories for all collected data?*

Despite the explosion of big data in recent years (more data will be created in 2017 than in the previous 5,000 years combined!), less than 0.5% of data is being used to help guide decision-making[1]. But, the data world is highly dynamic. New sources emerge daily, “rogue databases” are found weekly, and organizations retire old systems monthly. These shifts may lead many organizations to ask themselves, “Why bother?”

Booz Allen advises organizations that they may miss out on critical opportunities when they don’t plan, collect, and manage their data effectively. Getting their “data house” in order through a strategic and structured approach can help drive efficiencies, increase transparency, provide richer insights, inform strategic decisions, improve performance, ensure regulatory compliance, and minimize risks. The key to data management success is the formalization of a strategy and the underlying processes to support the evolving nature of data.

The common challenges in data management may seem insurmountable, but they can have a large impact on an organization’s ability to succeed through analytics:

Figure 6: Data Challenges and Impacts

CHALLENGE	IMPACT TO ORGANIZATION
Inconsistent adoption	Inconsistencies in use of governance standards limit their effectiveness due to continued data quality and sharing issues.
Multiple IT systems	Interoperability conflicts between data systems limit sharing and may result in “data leaks” as staff develop workarounds.
Ineffective governing bodies	Lack of authority to enforce and deploy governance standards leads to a patchwork system and limits the member’s credibility.
Regulation and/or policy issues	Insufficient controls on data generation, use, and sharing per regulatory and policy requirements likely leads to non-compliance and costly consequences.
Too much data	Increasing volume and diversity of data to manage can overwhelm organizations, causing them to miss opportunities to use that data.
Siloed organizations	Unwillingness to collaborate and share data leads to duplicative and inconsistent efforts.

Booz Allen develops approaches that address these challenges by supporting organizational agility in this dynamic data environment. These approaches build on three

Chapter Preview:

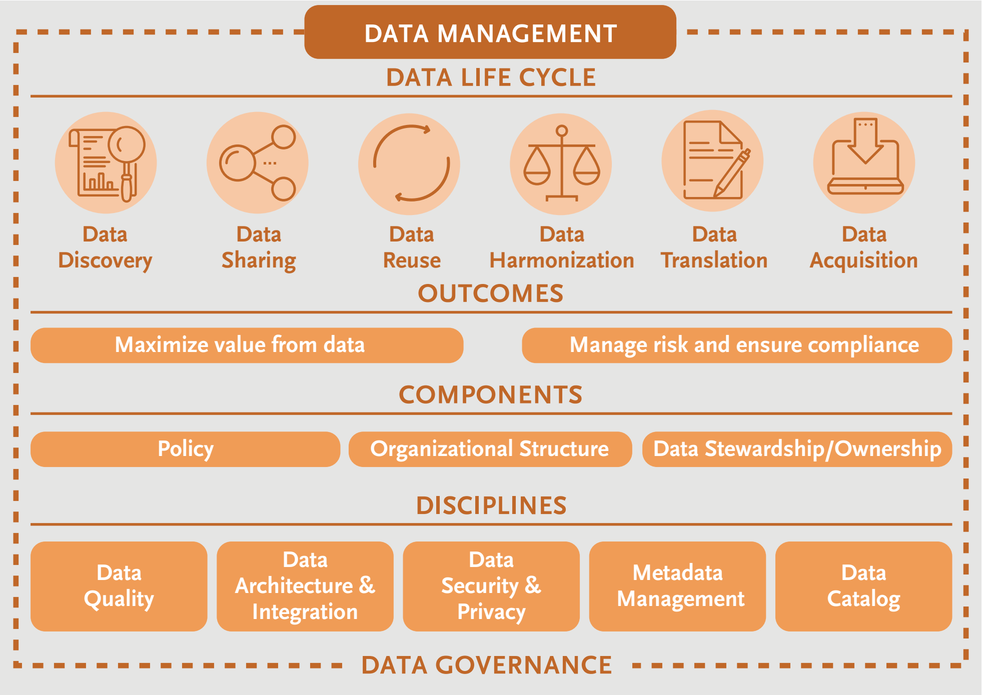
[1] *https://www.cio.com/article/3217020/data-management/your-data-accessibility-problem-is-costing-you-65-million.html*

key components for effective data management: data strategy, data operations, and data compliance. Together, these components increase the value of data throughout its lifecycle.

DATA STRATEGY: HOW YOU MANAGE YOUR DATA

Data is a capital asset, and just as an organization would undertake a capital improvement to a factory or building, an organization should undertake the strategic management of its data. Below is one approach to effective data management. A data strategy is a data management approach tailored to the goals of the organization and helps set the course for how data will provide value to an organization. A successful data strategy determines how the organization will invest to continue to extract that value. Like any other strategy, it should be focused on outcomes, supported by strong components, and continually revisited through continuous disciplines at each phase of the lifecycle.

Figure 7: Data Management Approach



Data Outcomes & Vision

Every successful organization, division, or program starts with a vision and required outcomes. The same holds true for your data. Simply put, an organization needs to lay out a vision for the value it wants to extract from its data. The data vision is important not only for an organization's leaders to make executive business decisions, but also for stakeholders across the organization to clearly understand the goals and help drive the organization towards the same future state. This is not a technology challenge, although it is often managed by technology leaders. For a successful data vision, business owners must be true owners in the process – their vision is arguably the most important.

The data vision should include goals and key indicators for what the organization seeks to gain from data. Goals should be short objective statements that describe specific achievements the organization would like to meet for each data phase. An example of a goal might be transforming data sharing processes by reducing limitations to data so analysts can uncover new opportunities for analytic insights. It is also important to set measures for each goal; those measures should be quantifiable and relevant to convey progress toward the specific goal, as well as the overall status of the data strategy. Specifically, organizations should consider establishing both outcome and progress oriented measures. Outcome measures relate to an organization's ability to achieve its end goals (e.g., the percent of users with access to data, decreasing lag time between systems, and increasing data access and transparency to inform marketing decisions), while progress measures will assess interim progress toward achieving data strategy goals (e.g., hitting major milestones such as the socialization of the strategy itself with key executives).

Designing a data vision is hard. This is true whether an organization is just "dipping its toe" into managing its data for the first time or has extensive experience with data management. Booz Allen uses design-thinking principles to conduct **Visioning Workshops** with our clients that enable leaders to create an ambitious, yet attainable, data vision for their organization.

Data Governance

With an effective vision, data governance provides the oversight bodies, processes, and procedures to facilitate successful execution of the data strategy. Many organizations attempt to accomplish data governance by simply standing up a new executive board or writing a few policy memos. These are almost never sufficient and may hinder real change efforts because leaders feel they've "checked the box" on data governance.

Booz Allen uses industry best practices such as the Data Management Association (DAMA) International's Data Management Body of Knowledge (DMBOK) as a foundation, along with our client delivery experience, to help clients implement effective data governance.

Designing and implementing a data governance program requires careful planning as well as the right mix of people, tools, and technologies. Some example first steps might include:

- Set up a data governance team that adheres to best practices and use cases
- Develop a data security plan and data privacy policies in accordance with Federal Trade Commission (FTC) policies
- Define roles and responsibilities for each role on the team, keep leadership impartial and objective, and ensure that there is a plan for staff transitions (e.g., if staff separate from organization, or are unable to continue in role)
- Strive for a global outlook–move towards global data governance processes that can function across systems
- Determine the right metadata to collect
- Define the expected timeline for your data governance investments to deliver results
- Move towards adopting industry standards (i.e. United Nations (UN) / International Standards Organization (ISO) data standards)
- Follow Data Architecture Principles

With a more holistic view of data governance, an organization can not only create a structure for governance, but drive efficiency, insight, and value from data.

Data Requirements Lifecycle

Just as systems have a lifecycle cost and value, so too does data. Certain data is most valuable at the moment of collection and loses its value over time. Other data lacks operational value in the moment but has significant value when examining trends over time. Most organizations have both types of data. While the data may not be constant and the specific requirements of the data may shift, the processes for maintaining these requirements must remain constant. The data requirements lifecycle represents the processes of identifying, analyzing, and verifying the business and operational requirements for data, including how the data components are collected and tracked. For example, when updating a system to include an additional data field, the organization should capture and evaluate what data the end user needs and why, how that impacts the individual entering the data, and how the data should be stored. This ensures that the data can be of value while also considering its cost.

Data vision, data governance, and data requirements lifecycle are each an essential part of an effective data strategy. Booz Allen works collaboratively with organizations to consider all three within our comprehensive **Data Strategy Framework**. This framework walks clients through the entire data strategy development process, helping them make enterprise-level decisions and yield data-driven outcomes. It starts with an organization's data vision and then works across six distinct phases of the data lifecycle, while keeping the strategy aligned with on-the ground operational realities. For each lifecycle phase, organizations define clear data goals, identify key people and technology, and consider the cultural implications. Finally, data governance and data management underpin the framework and define the functions that ensure oversight, control, and execution of these data activities.

Case Study: Supporting the Data Act at Department of Treasury

The DATA Act of 2014 required the development of data standards to improve the quality of federal spending data. Booz Allen led the way by helping the Department of the Treasury develop the DATA Act Information Model Schema (DAIMS), a Data Broker for ingesting and validating new spending datasets, and giving an overhaul of USASpending.gov to publish the new spending datasets, which have financial information and procurement information connected and published for the first time. Using our Digital Solutions and Data Science capabilities, Booz Allen is helping to consolidate, standardize, and publish over $3.8 trillion dollars in annual federal spending data to give Americans a better understanding of how taxpayer dollars are used.

Case Study: Integrating Siloed Data at Department of Homeland Security (DHS) Undersecretary for Management (USM)

The Undersecretary for Management (USM) needed to improve the efficiency of reporting and analysis on Department of Homeland Security (DHS) management data across the entire department. USM's seven Lines of Business (LOBs) separately collected data from all DHS components so that the data was siloed within each LOB. However, cross-LOB data was required to make department-wide decisions and respond to congressional requests. Booz Allen investigated the types of data used throughout these siloed data sources and built a series of workflows integrating all relevant data within a single warehouse. This allowed USM Analyst users to view and analyze the information across the entire enterprise, generating new cross-LOB insights and informing better business decisions for DHS.

DATA OPERATIONS: HOW YOU COLLECT AND STORE DATA

Data comes from all over the place and must be stored. Data collection is the internal creation of data, whereas data sourcing involves importing data from partners or purchasing it from vendors. In both situations, an organization must formalize the processes used to gather the data before storing it in a centralized repository.

Data Collection & Data Sourcing

Many organizations find centralizing data collection and sourcing to be a significant challenge. They may rely on manual processes for data intake (e.g., paper forms) or house data across many silos. Different entities throughout the organization may store data differently or use inconsistent protocols. Many organizations even lack full awareness of the data that is available to them or that they already own. To address these challenges, Booz Allen walks our clients through an **Organizational Data Assessment** that helps them understand their data operations and leads to enterprise-wide standardized data collection and sourcing procedures (e.g., Enterprise Data Catalogue).

Many organizations may also seek to supplement their own data with external data sources. These data sources may come through contracts from data service providers, or through service level agreements (SLAs) from federal organizations such as census, weather, or other economic data, or through social media streams and other publicly available sources. These data sources provide a way for organizations to augment and infuse more power into their own data, providing more complete and nuanced insights.

Data Storage

Having access to data is one thing, but storing it properly (which includes formatting and matching the data) to enable its use in analytics is an entirely separate endeavor. Each organization should assess its unique data storage needs depending on the volume, structure, and access requirements of its data. Many organizations find that their existing rigid data infrastructure lacks the flexibility to support their analytical objectives. Data storage should be designed to meet the organization's need, from traditional relational storage, to document, key/value, and/or columnar stores. Data is never going to be stored in a single way, instead organizations should consider the optimal way for the data in question; but, organizations are increasingly moving towards ”one-stop-shops" for their data, also known as "data lakes." As described in the technology section, data lakes help organizations aggregate their data so that more data is usable for analytics. In addition, data lakes allow organizations to scale drastically better than traditional data storage solutions, providing increased computing power to users.

DATA COMPLIANCE: HOW YOU ENSURE DATA INTEGRITY

Just as data is a capital asset that requires investment and resources, it is also an asset that requires protection and security because of its inherent vulnerability. Because data is generated, used, and often violated by humans, it is inherently vulnerable and imperfect. Fortunately, strong processes enabled by technology can help overcome and limit that vulnerability, increasing data quality, privacy, and security while not inhibiting data access.

Data Quality

Data is rarely perfect. Even for organizations with above average data quality, the increasing volume of data, processes, and reporting makes maintenance increasingly difficult. But it also makes it increasingly important. Poor data quality can result in poor decisions, failure on key mission metrics, mistrust in an organization's abilities, and time and resources spent on correction and rework. On the other hand, effective data quality allows an organization to confidently make decisions, report accurately to stakeholders, operate more efficiently, and avoid extraneous costs.

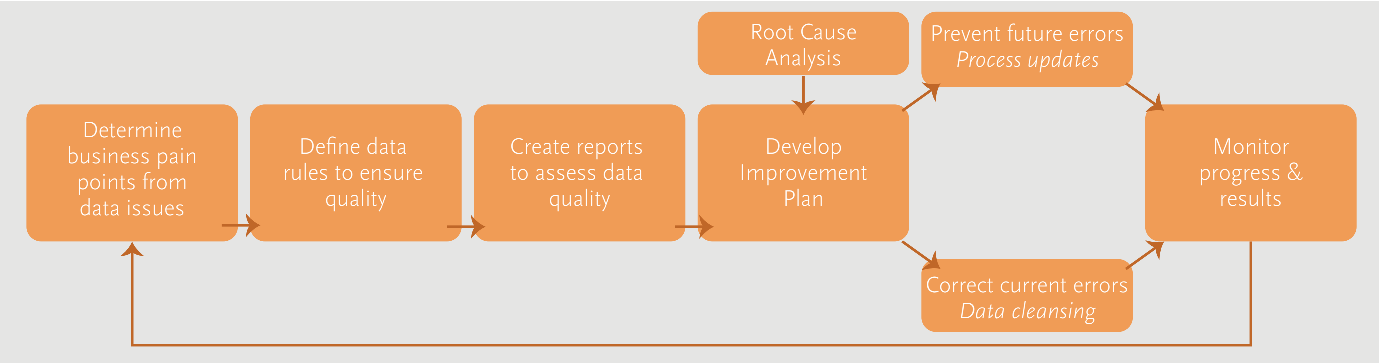
Imagine having all your data perfectly organized and synced in one place in such a way that everyone in your organization has perfect information equity. With strong processes for continuous review, audit, and improvement of data quality, this is an attainable dream.

Good data quality relies on six data quality principles that can help your team get the most out of your data. They are:

- **Completeness** – Are critical data fields (names, IDs, locations, etc.) missing few or no values?
- **Uniqueness** – When measured against other data sets, is there only one entry of its kind?
- **Timeliness** – How much of an impact does date and time have on the data? This could be previous sales, product launches, or any information that is relied on over a period of time to be accurate.
- **Validity** – Does the data conform to the respective standards set for it?
- **Accuracy** – How well does the data reflect the real-world person or thing that is identified by it?
- **Consistency** – How well does the data align with a preconceived pattern?

Booz Allen's **Data Quality Approach** helps organizations ensure they meet these data quality principles in an automated manner. Booz Allen helps our clients employ a holistic approach to achieve these high-quality data standards using a combination of well-defined procedures and advanced technology.

Figure 8: Data Quality Approach


Determine business pain points from data issues
Define data rules to ensure quality
Create reports to assess data quality
Root Cause Analysis
Develop Improvement Plan
Prevent future errors
Process updates
Correct current errors
Data cleansing
Monitor progress & results

Data Security, Privacy, and Access

Many organizations today rely on techniques for securing data that were created before the rise of big data and the cloud. These traditional approaches stored data in siloed servers with varying levels of security restrictions. In turn, data protection focused on the individual silos. Proper data management allows organizations to make connections between the silos of data without creating security risks. Modern technologies make it simple to grant access to containerized portions of the data and log who is accessing the data and how they are using it. For example, the White House's Digital Government Strategy called for federal agencies to make better use of digital technologies while maintaining data security.

Strong data security and proper data access don't simply happen on their own. They only exist in organizations with a strong culture of accountability and commitment to data integrity. The most effective organizations have formalized standards and procedures for data access that are regularly reviewed and consistently enforced. This sounds easier than it is. Sending out a few management directives is not sufficient to make institutional changes.

At Booz Allen, we've helped our clients overcome these challenges to build lasting change within their organizations, even when it comes to seemingly mundane business protocols such as data access. For example, user-centered design principles, when applied to data access, can open doors previously closed to many data scientists. Without changing a single data access policy, but instead accelerating and enabling the permission process through a customer service desk for data access, organizations can effectively provide more data to analytics practitioners. For security, an effective way to ensure organizations follow security protocols is to bake them into the technology (e.g., user authentication, monitoring). Organizations with strict security protocols may even choose to use cell level permissions for their data access.

The Booz Allen Difference

Booz Allen's approach to data strategy and management prioritizes the activities that facilitate organizational change in line with the data strategy. The strategic and structured approach helps drive efficiencies, increase transparency, provide richer insights, inform strategic decisions, improve performance, ensure regulatory compliance, and minimize risks. Through our experience helping clients navigate data strategy and management challenges, along with our own 600+ data scientists, Booz Allen brings new ways to adapt to the constantly changing data environment.

CHALLENGE	BOOZ ALLEN SOLUTION	DESCRIPTION
Not sure where to start in creating a data vision	**Visioning Workshops**	Workshops that use design thinking principles to help leaders create a data vision for their organization
Need guidance in fulfilling your data science vision	**Data Strategy Framework**	Framework that helps organizations develop a data strategy that will enable them to make organizational data-driven decisions and pursue their data vision
Unsure of what data you have and what you need	**Organizational Data Assessment**	Approach that helps clients understand their data operations and leads to enterprise-wide standardized data collection and sourcing procedures (e.g., Enterprise Data Catalogue)
Unable to keep up with maintaining the quality of growing volume of data	**Data Quality Approach**	Process and tools that assess the quality of an organization's data, establish corrective actions, and automatically evaluate data quality on an ongoing basis

4 | ANALYTIC TECHNIQUES

- *Are you able to quickly and easily access the specific pieces of information your organization needs at any given moment?*
- *Do your analyses typically include historical data and predictions for the future?*
- *Does your organization know which analytic techniques to use to generate actionable insights from your data?*
- *Are data results presented in a way that makes the conclusions obvious, or do they require explanation?*
- *How confident are you in the statistical accuracy of the analyses on which your organization is making business decisions?*

Successful data science starts with the right analytic techniques to drive business decisions. This does not mean the organization needs everyone to be proficient in the latest statistical modeling techniques and algorithms. Rather, organizations need to be thinking about how analytics can help solve key business problems. Data-driven organizations connect the dots to find the appropriate analytics techniques needed to meet the specific challenges they are facing.

In order to effectively connect the dots, analysts go through a two-part process. They must first prepare the data and then perform their tradecraft. Analytic preparation includes how an organization ingests and organizes structured data and unstructured data – essentially, the inputs. Analytic tradecraft is all about the analysis, visualization, reporting, and validation – essentially, the outputs.

Chapter Preview:

ANALYTIC PREPARATION: HOW YOU ORGANIZE YOUR DATA

To deliver valuable analytics solutions, analysts must prepare the raw data. Raw data may be ingested from a variety of different sources: large and small databases, documents, unstructured text, and even streaming data or log files. Because of the complexity of these source types, numerous techniques exist solely to organize data from its many sources into a useable and consistent format. These techniques vary based on the type, volume, and accessibility of an organization's data, but they all provide one basic result: clean, organized data that is ready for analysis.

Effective analytic preparation involves choosing the right data organization techniques. These techniques focus on: capturing all available data, cleaning ingested data, storing reliable data, and exposing prepared data to analysts. The goal of organizing data is to anticipate potential analytic exercises and reduce the time required between acquisition and action.

Capture & Clean Raw Data

Data that is already structured into rows, columns, or cells (i.e., Excel spreadsheets) or is semi-structured (i.e., PDFs or legacy database systems) is much easier for organizations to manipulate, but that does not mean the data is ready for deep analysis. Regardless of the initial state of the data, the first step in data organization is to find a single consistent location where the versions of the data that will be used for analytics can reside. For historical reported data, it may be appropriate to create a brand-new data repository and forward all known data in a one time transfer. For some production data, it might be reasonable to capture analytic data directly in production environments in real time or make a copy in a downstream test environment. One of the key considerations at this step is to ensure that data from multiple sources can be brought together in the analytics environment. Open source tools such as NIFI or Sqoop can help set up reliable efficient extract, transform, load (ETL) processes that ensure data is always captured in the analytics environment in the exact same way.

A multitude of techniques can organize the data more efficiently to support pattern recognition, detect trends, and identify key insights but often these techniques assume a base level of data quality to succeed. The second key component of analytic preparation is data cleaning. Data cleaning might include ensuring all data meets prescribed formats, for example no hyphens in ingested social security numbers, or is realistic, for example transaction start dates always precede transaction end dates. While some data cleaning can be accomplished with relatively simple queries, more complicated interpolation techniques may be necessary to compensate for sequences of missing data. The early application of these analytic techniques can significantly reduce the time spent wrangling the data, accelerate the actual analysis, and improve the outcomes of eventual modeling.

Store & Expose Organized Data

Data storage involves multiple physical and logistic decisions related to encryption and security but also involves developing robust indexing schemes so that future algorithmic operations can be completed quickly when executed on demand. Data storage should also begin to consider what additional resources will be needed further in the process to hold analytic results or temporary datasets generated while analytic projects are underway. Open source tools such as Hive, Yarn, Elasticsearch, and Hbase have functions that support ingestion and processing/indexing data. Specific data storage modes might also be more appropriate depending on the type of analytics, for example Neo4j provides an efficient underlying graph database for certain types of relational queries.

Stored data, thousands of gigabytes of streaming web traffic or hundreds of quarterly reports held in flat files, are only as useful as access to the underlying data allows. For small projects or simple analyses, it may make sense to allow team members to directly query the underlying data stores, design their own schema, and perform their own extracts. In more complex projects, or projects where analytic results need to be rerun hourly or near real time, it may be more reliable to keep a federated copy of the data that analysts can access via an API or application services layer. In an environment where some data is considered sensitive, exposing data might be accomplished by creating sanitized data products from the main data repository and providing them to analysts through a regular distribution process.

ANALYTIC TRADECRAFT: HOW YOU TRANSFORM DATA INTO INSIGHTS

Formally, tradecraft refers to the analytic, visualization, and reporting techniques that transform organized data into the insights required for decision-making. Informally, organizations that successfully perform analytic tradecraft are thoughtful and nuanced in how they analyze, visualize, and report information.

Data Analysis

Once the data is organized, it is time for data analysis. Often analysts try to categorize techniques by the specific methods they use, for example regression analysis, clustering analysis, optimization, or simulation. From a planning and design perspective, it can be more reliable to begin to break up analytic techniques based on their desired outcome and the maturity of the analytics being performed *(see next page)*.

Case Study: Employing TUNE at a Life Sciences Company

A Life Sciences Company was attempting to integrate data about suppliers that was trapped in data silos across the organization. They were wasting time and money by integrating multiple data types and models across incomplete and inconsistent data with different entity field types. Booz Allen employed its innovative technique called TUNE, which provides entity resolution across multiple data sets at a very large scale. TUNE uses machine learning algorithms to aggregate and de-duplicate data. The resulting Master Data Set streamlined and reduced supplier data from 900,000 entities down to a consolidated list of 300,000 unique suppliers. The organization applied the same TUNE capability to its Customer Relationship Management systems to develop one view of the customer. These de-duplication techniques often generate immediate cost savings by reducing repetitive communications and contact.

Figure 9: Four Types of Analytics

TYPE	DESCRIPTION
Descriptive: What Happened?	Descriptive analytics summarize what has already occurred. For example: After reviewing last year's sales records, descriptive analysis could reveal which of ABC Retail Store's products was the most popular or which days of the week were the busiest.
Diagnostic: Why Did It Happen?	Diagnostic analytics go one step further and often involve a regression to compare multiple variables against each other. For example: ABC Retail Store could compare weather data with its sales performance to determine that sales went up when it was sunny.
Predictive: What Will Happen?	Predictive analytics take the next step by identifying trends, clusters, and exceptions with the data to make a prediction about future outcomes. For example: ABC Retail Store analyzes the sunshine-sales connection to determine if there is any correlation that would suggest sales will similarly be higher on sunny days next year as well.
Prescriptive: What Should Be Done?	Prescriptive analytics are a relatively new field that analyze different possible scenarios or courses of action to determine which one has the most optimal outcome. For example: ABC Retail Store runs a scenario-testing analysis to determine that increasing store staff on sunny days would likely increase profits due to shorter customer wait times and higher customer satisfaction.

Selecting the right analytic technique isn't necessarily about selecting what is the most mature, though, it is about selecting what is the appropriate technique for the question you are attempting to answer. Booz Allen's **Field Guide to Data Science** provides a *Guide to Analytic Selection* which provides recommendations for some of the types of analytic techniques to consider. For example, when considering the right techniques for classification of data into known groups, the *Guide to Analytic Selection* points the reader to several different options:

- If you have known dependent relationships between variables, start with:
  - Bayesian network
- If you are unsure of feature importance, start with:
  - Neural nets
  - Random forests
  - Deep learning
- If you require a highly transparent model, start with:
  - Decision trees
- If you have <20 data dimensions, start with:
  - K-nearest neighbors
- If you have a large dataset with an unknown classification signal, start with:
  - Naive bayes
- If you want to estimate an unobservable state based on observable variables, start with:
  - Hidden markov model
- If you don't know where else to begin, start with:
  - Support vector machines (SVM)
  - Random forests

In addition to helping a reader select the right analytic technique, the Field Guide also includes a comprehensive *Detailed Table of Analytics* which describes all techniques, gives

Case Study: Natural Language Processing (NLP) for a Large Healthcare Organization

A large healthcare organization had to manually read and analyze thousands of unstructured text items daily to identify trends and rising issues reported by more than 12,000 call center representatives about their interactions with callers. Booz Allen identified key topics in text data by using supervised machine learning algorithms including neural networks, hierarchical clustering, Latent Dirichlet Allocation, and k-nearest neighbor to interpret and categorize text files. This capability replaced the manual processes by instantly clustering findings by relevant topics, including the type of issue and concerns about program policies and procedures. Booz Allen's open source techniques enable clients to sort through millions of lines of text in a matter of minutes, organizing the data into logical groups and clusters for analysis.

tips from the pros, and references for further reading. The detailed table is included in the appendix of this Playbook.

Booz Allen's Field Guide to Data Science is more than just about selecting analytic techniques. It defines the field of data science and why it matters to organizations. The Field Guide contains proven methodologies, personal tips and tricks, and real-life case studies. Senior leaders will walk away with a deeper understanding of the concepts at the heart of data science, practitioners will add to their toolboxes, and beginners will find insights to help them start on their data science journey.

Each organization is unique, requiring its own data environment, architecture, and business or data needs. Consequently, organizations face the difficult task of choosing which analytic techniques to apply to answer their business question. It's never as simple as building a "predictive model," as this may include several techniques. For example, imagine an organization that is building a model to predict if certain transactions will succeed or fail. The output of this model is likely a probability between 0-100. Even after the model is built and tuned, business decisions need to be made about how to handle the results. Should every transaction with a risk greater than 75% be stopped? Should every transaction with a total volume above $50,000 and a risk greater than 50% be stopped? Should the five riskiest transactions a day be pulled aside for manual review before they are allowed to proceed? Figuring out how to integrate the results of a successfully deployed analytic technique can often be more challenging than the statistical portion of the process.

To help make the nuanced decision of which analytic techniques to select, organizations should seek to break projects into smaller pieces and assess how the results will ultimately be used. This will also enable better staffing and resources decisions and help to identify the appropriate technologies required.

Data Visualization

Pictures and images will be remembered more than words – it's what scientists call the "picture superiority effect." Visualizations enable people to remember information for longer, access it more easily, and remember it in more detail than if they read about it in text. Essentially, we identify patterns through color or spatial arrangement better than looking at numbers alone. Data scientists recognize the value of visualizations by incorporating best practices for visualizations into most standard reporting and technology platforms and offering training programs on data visualization.

Booz Allen produced the in-depth and easy-to-use **Data Visualization Guide** to help organizations understand data visualization and design fundamentals. These guides use the latest psychological persuasion research and serve as step-by-step tools for teams with limited data science or visualization experience to determine what types of visuals (bar graph, line graph, pie chart, etc.) are most effective to support their analytics objectives. An organization's decision about data visualization depends on the types of data as well as the key takeaways it wants to emphasize for the audience. Obviously high-level guidance is useful to everyone, but more detailed data visualization techniques may provide increased support for specific types of organizations. For example, an organization that produces navigational information may want a data visualization guide entirely on geographic visualizations (e.g., choropleth, heat and dot/symbol map); whereas a comptroller's office documenting cashflows may prefer a visualization guide that focuses entirely on waterfall charts.

Case Study: Creating Dashboards & Visualizations at Department of Homeland Security (DHS) Office of Immigration Statistics (OIS)

Booz Allen built a DHS Enforcement Lifecycle Tracker for the DHS Office of Immigration Statistics (OIS). The user-friendly dashboard provides a status update on enforcement and benefit outcomes for alien apprehension cohorts in different time periods after apprehension. This visualization tool enables users to select different alien demographic information and review changes in apprehension outcomes over time. It is also used to identify bottlenecks and gaps in the immigration life cycle.

Reporting & Distribution

Once an organization conducts its data analysis and creates the supporting visualizations, the information must be shared with the right audience through reporting and distribution. Leading data science organizations realize that responding to the never-ending flow of requests for ad-hoc reports, creating daily briefs, and attending meetings can pull data scientists away from the work of data science. Reporting and distribution shouldn't fall into this category, it should instead be part of the natural outflow of rigorous analysis. Organizations that do this well protect their data scientists from the "daily grind" of management information reports, and keep them focused on sharing actionable insights with leaders.

For further efficiency, Booz Allen helps organizations take advantage of technology-enabled solutions (e.g., automated data collection, customized web-enabled reports tailored to the end user) and analytic storytelling to influence the audience. Analytic rigor and strong visualizations should be self-sufficient, but they don't always tell the whole story. Organizations need effective reporting to share results with leaders, as this can play a significant role in the delivery and interpretation of results.

Analytic Integrity

The accurate application of analytic techniques is essential to building trust in data science. Leading academic institutions include peer review as a fundamental part of their process. Similarly, data scientists recognize the value of ensuring analytic rigor prior to making large decisions. As such, prior to briefing out the results of any analytics solution, organizations should incorporate routine, standard checks on its analyses to ensure their accuracy and statistical validity. Organizations must regularly review and maintain existing models to ensure the assumptions the model was built upon are still valid and to identify and incorporate newly-available data. Analytic techniques used to validate the continued accuracy of analyses include: regular methodology testing, comparing results against other statistical approaches, and conducting code reviews on a routine basis. Some organizations even develop an internal review board to provide oversight over analytical projects.

Case Study: Integrating Costs and Schedules at the National Aeronautics and Space Administration (NASA)

The National Aeronautics and Space Administration (NASA) needed to estimate joint costs and schedules between several large-scale projects to streamline and estimate the integrated program launch date. Booz Allen harnessed Polaris (which was developed by the firm at NASA) that integrated detailed cost, schedule, and risk analysis from multiple sources. The team also implemented additional techniques for forecasting, situational testing, and visualizations that illuminated and consolidated budget and schedule areas. NASA is now able to take a holistic view of its program costs at all levels, forecast costs going into the future, and test situational scenarios that may arise and their effect on the budget, schedule, and launch dates.

The Booz Allen Difference

Booz Allen works hand-in-hand with a variety of organizations to strategically identify, develop, and implement the most effective analytic techniques. The *Guide to Analytic Selection* in the Field Guide to Data Science (pp. 71-82) provides an adaptable tool that supports organizations in the selection and use of analytical techniques. Organizations successfully integrate their data into decision-making and extract the most valuable insight when they carefully think about their analytic needs and the most effective analytic techniques.

When helping to drive an organization toward its analytic goals, Booz Allen takes a holistic approach that begins with preparing data to maximize its potential value before implementing the analytic tradecraft to analyze it. Therefore, Booz Allen works with organizations to first identify the analytic techniques needed to ingest, integrate, and organize their data. Once their data is properly arranged, Booz Allen helps organizations identify and implement the analytic techniques to analyze, visualize, and report the key insights from the data.

In light of emerging techniques and capabilities, organizations must constantly rethink their data infrastructure and evaluate new analytic approaches to stay at the forefront of data science. Booz Allen continually strives to be at the forefront of new and existing analytic techniques. With experience in applying a broad scope of analytic techniques across a wide set of operational spaces, Booz Allen successfully provides the most impactful solutions for the analytic challenge.

CHALLENGE	BOOZ ALLEN SOLUTION	DESCRIPTION
Unsure of how long it will take to derive action from your data	**Data Science Growth Chart**	Visualization tool that helps organizations determine the time it will take between acquiring their data and executing data-driven actions given the type of data it is and how it is structured
Need help integrating incomplete/inconsistent data from different sources	**TUNE**	Data science tool that allows for large scale data alignment from multiple sources and uses machine learning algorithms to aggregate and de-duplicate data
Don't know how to analyze and draw insight from text data	**Natural Language Processing (NLP)**	Machine learning tool that processes unstructured and semi-structured text documents to perform data analysis
Need to monitor large quantities of different types of data for anomalies	**Multivariate Analysis of Stealth Quantities (MASQ)**	Algorithm that mines data for known and unknown signatures to detect fraudulent data transactions
Need to analyze unstructured data from images and videos	**Computer Visioning**	Combination of analytic techniques including machine learning that helps identify patterns, trends, or insights from unstructured data such as photos or video recordings
Don't know how to display insight from data in a way that is easy to understand	**Data Visualization Guide**	Collection of best practices for selecting and creating data visualizations that clearly depict the key insights derived from data analysis
Employees don't have analytical skills to use data insights for decision-making	**Sailfish Explore**	Analytics platform that enables analytics-driven decision-making for all members of an organization regardless of their analytics skills level
Unsure of how your organization's data science problems could be solved	**Field Guide to Data Science**	Booklet that details Booz Allen's unique data science capabilities and how these capabilities have been used to solve clients' data science problems (available here)

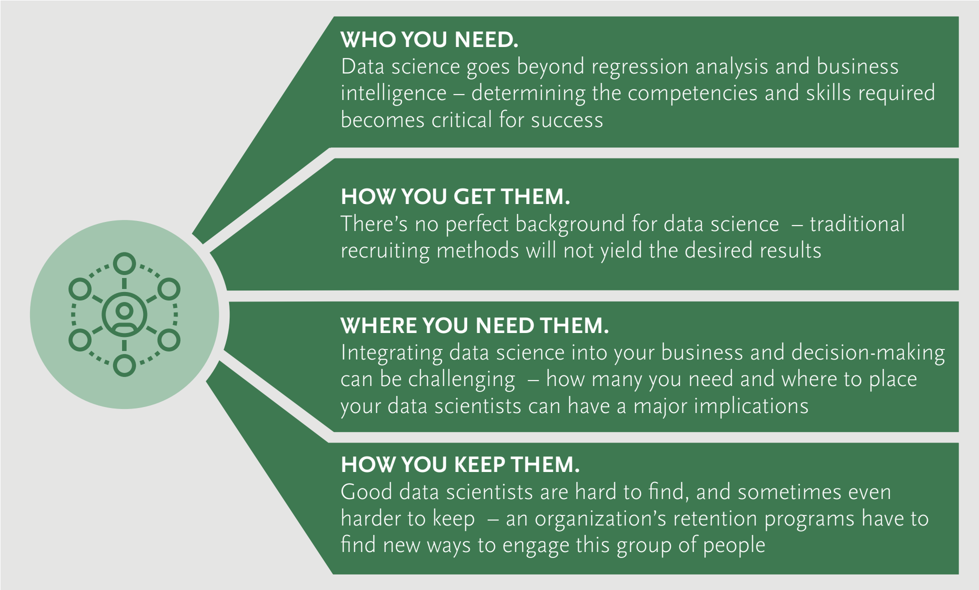
5 | PEOPLE

- *Do you know who the data scientists are in your organization?*
- *How does your organization find, recruit, on-board, develop, and retain an analytics workforce?*
- *Do you know where to place data scientists so that they are best positioned to help the organization answer its most critical questions?*
- *Is your organization continually reviewing its analytics goals to understand its talent needs?*

The demand for data scientists and analytics practitioners is at an all-time high. In 2012, Harvard Business Review called Data Scientist, "the sexiest job of the 21st century[1]" and a recent report[2] predicted that the number of positions for data and analytics talent in the US will increase by 365,000 (+15%) by 2020. However, the demand for data scientists outweighs supply and conventional talent management fails to hit the mark. So, how can organizations win the war for data science talent?

All organizations, large and small, have human capital processes to hire and develop staff, but the best organizations understand how to adjust these processes to get the specialized skills needed for data science. Within the human capital lifecycle, there are four areas where the processes must be modified to address these specialized talent needs. Those areas are: Defining Talent, Recruiting Talent, Placing Talent, and Developing Talent.

Figure 10: Talent Management Model



WHO YOU NEED.
Data science goes beyond regression analysis and business intelligence – determining the competencies and skills required becomes critical for success

HOW YOU GET THEM.
There's no perfect background for data science – traditional recruiting methods will not yield the desired results

WHERE YOU NEED THEM.
Integrating data science into your business and decision-making can be challenging – how many you need and where to place your data scientists can have a major implications

HOW YOU KEEP THEM.
Good data scientists are hard to find, and sometimes even harder to keep – an organization's retention programs have to find new ways to engage this group of people

Chapter Preview:

[1] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

[2] http://burning-glass.com/wp-content/uploads/The_Quant_Crunch.pdf

Case Study: Growing Data Science Talent at a Commercial Airline

A commercial airline developed a data science capability to remain competitive in an environment in an industry with a rapid proliferation of data and the need for more insights. Booz Allen helped the airline remain competitive by attracting and retaining the right data science talent. For this effort, Booz Allen developed a Data Science Skills Model that provides the airline with the guidance and tools to identify, attract, and retain new data scientists, while developing existing employees. The Data Science Skills Model is broken down into three primary components: Data Science Competencies to outline the desired competencies for identifying data scientists; Data Science Job Descriptions that provide resources for understanding the role of a data scientist and how to acquire talent; and Engagement Tools for on-going learning and development, which foster a desirable experience to promote growth and retention. As a result, the airline was able to foster a culture which supported data scientists and their ongoing development, and better hire data scientists.

TALENT DEFINITION: WHO YOU NEED

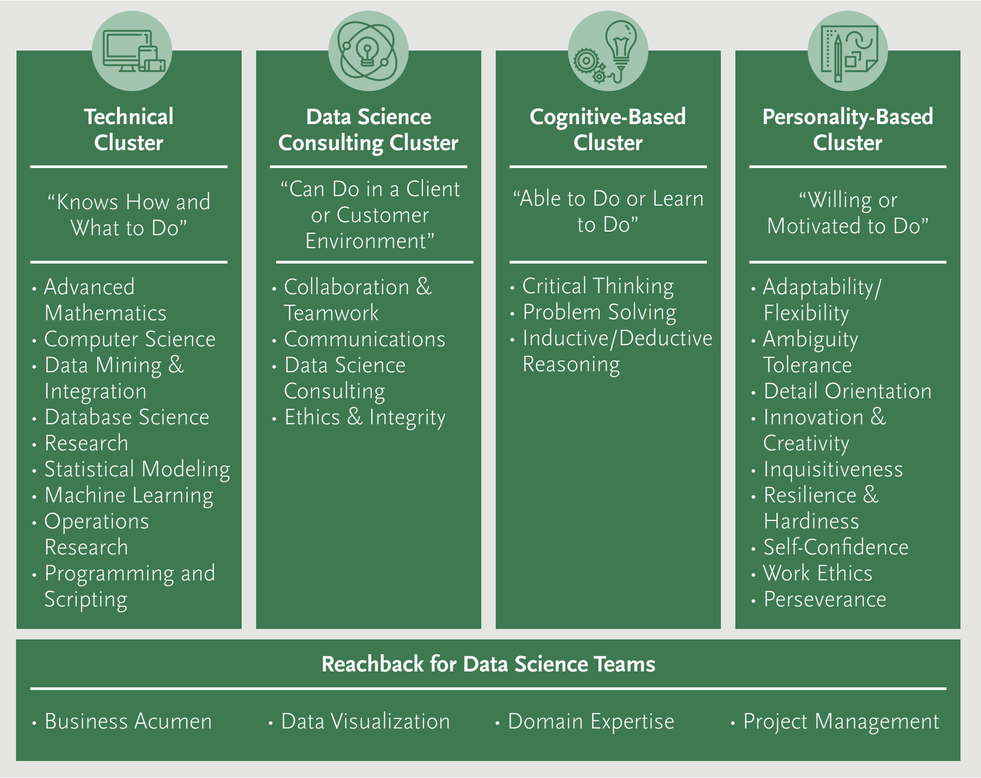
It takes a unique mix of skills to generate analytic insights. Most simplistically, an individual should be able to blend three key disciplines – mathematics, computer science, and domain expertise. But, data science isn't an individual athletic event; it's a team sport that requires a combination of competencies and skills. Data science isn't just about the technical skills, it's about the intellectual curiosity, creativity, and determination that work in harmony with technical skills to drive to a solution.

Therefore, if an organization uses its traditional approach to defining talent needs, it may miss capturing opportunities for rare talent. At Booz Allen alone, some of our leading data scientists have advanced degrees in subjects as broad as Biophysics, Computer Graphics, English, Forest Management, History, Operations Research, and Wildlife & Fisheries Management.

Job Analysis & Position Definition

Effective data science requires a combination of skills. Booz Allen helps build data science teams using two different approaches. In the first approach, organizations hire specified talent. The group defines the work required and conducts a job analysis. A job analysis reveals the particular job duties and requirements as well as the relative importance of these for a given job. The job analysis then translates to a set of required knowledge, skills, and abilities (KSAs), which are documented and described in very specific position descriptions to create a "to be" organizational chart. Many organizations build their talent in this manner, especially the federal government which has rigorous policies around staffing, classification, and hiring. In this approach, each position is specific and allows for the hiring of different "shades" of data science (e.g., expertise in machine learning, artificial intelligence, data mining, etc.).

Figure 11: Data Science Talent Competency Framework



Technical Cluster	Data Science Consulting Cluster	Cognitive-Based Cluster	Personality-Based Cluster
"Knows How and What to Do"	"Can Do in a Client or Customer Environment"	"Able to Do or Learn to Do"	"Willing or Motivated to Do"
• Advanced Mathematics • Computer Science • Data Mining & Integration • Database Science • Research • Statistical Modeling • Machine Learning • Operations Research • Programming and Scripting	• Collaboration & Teamwork • Communications • Data Science Consulting • Ethics & Integrity	• Critical Thinking • Problem Solving • Inductive/Deductive Reasoning	• Adaptability/ Flexibility • Ambiguity Tolerance • Detail Orientation • Innovation & Creativity • Inquisitiveness • Resilience & Hardiness • Self-Confidence • Work Ethics • Perseverance

Reachback for Data Science Teams

• Business Acumen • Data Visualization • Domain Expertise • Project Management

The second approach focuses on building a strong cadre of data scientists with generalized skills and augmenting the data science team with highly specialized skills as needed. Each approach has strengths and weaknesses. In the first option, organizations ensure they have all the skills to do any type of analysis, but specialized individuals may grow uninterested in work outside their specialty. In the second example, organizations have a more flexible data science team, but risk lacking expertise in nuanced areas of data science.

To learn how to build a data science team, Booz Allen first surveyed its own data scientists and developed the **Data Science Competency Framework**. The framework predefines the knowledge, skills, and abilities associated with a successful data science team. Organizations use this framework to help think through the blend of technical skills, personality, and problem-solving skills, rather than trying to find the elusive "unicorn" who meets all the needs.

TALENT RECRUITMENT: HOW YOU GET THEM

Knowing that demand is high and supply is low, Booz Allen also helps organizations figure out where and how to recruit the best data scientists. Hiring doesn't begin when an applicant applies for a job, and recruiting doesn't just mean setting up a table at your alma mater's career fair. To compete in today's global talent economy, especially for high demand data scientists, organizations must develop a recognizable and attractive employer brand targeted to high value sources of top-talent. The brand is then supported by streamlined and efficient hiring processes.

Attracting & Sourcing Talent

Recruitment today does not look like it did twenty, ten, or even five years ago. There are two primary reasons why. First, 70% of talent are "passive job-seekers": they are likely currently employed, but open to talking with a recruiter or within their personal network. Conversely, only 30% of the job-seeking population actively pursues new jobs[3]. Second, 79% of job seekers use social media as part of their job search, allowing them to compare job listings and read reviews from current employees[4]. Given the significant demand for quality data science talent, this affords top candidates the opportunity to patiently explore their options. Either of these changes alone significantly impacts the recruiting practices of an organization. When combined, organizations seeking data science talent must react swiftly to attract and engage employees in the competitive marketplace.

Booz Allen helps organizations rethink their brand identity and social media presence through the lens of an employer. Organizations commonly devote significant time and resources to cultivating an effective and valuable customer-facing brand. This same messaging and identity must be incorporated into all employer marketing materials as well. An organization seeking the top data science talent must ask itself how it can incorporate its public image into its employer brand and culture to create an attractive workplace that attracts top talent.

Once an organization has the right employer brand, it must decide where to showcase that employer brand. Organizations that have been hiring talent with strong technical skills for many years can use their existing HR data to identify their best sources of talent. Organizations that are seeking to build a data science team should research current

Case Study: Measuring & Enhancing Employer Brand for the National Aeronautics and Space Administration (NASA)

For NASA's Office of Human Capital Management (OCHM), Booz Allen developed an assessment to measure the strength of the agency's employer brand in attracting 'top talent'. Using social sentiment analysis, Booz Allen explored the difference in positive sentiment toward NASA as an employer versus NASA's external brand. Data from the initial assessment drove a comprehensive transformation project to enhance NASA's employer brand and to streamline recruiting operations.

To strengthen NASA's employer brand, Booz Allen leveraged insights to develop a data-driven Employer Value Proposition (EVP). As part of NASA's digital recruiting efforts, Booz Allen launched social media campaigns on LinkedIn, Twitter, Instagram, and Snapchat to highlight what makes employees proud to work at NASA, influencing LinkedIn's first TV commercial which featured NASA. These efforts resulted in the exponential growth (17x) of engagement with NASA's employer brand compared to the previous year. The EVP is also driving the creation of a new careers website as well as targeted LinkedIn landing pages for NASA's core hiring personas, where talent can find personalized content about working at NASA.

[3] *https://business.linkedin.com/content/dam/business/talent-solutions/global/en_us/c/pdfs/global-talent-trends-report.pdf*

[4] *http://www.careerarc.com/blog/2016/01/13-recruiting-stats-hr-pro-must-know-2016/*

academic data science programs, making use of existing university relationships, or build new relationships with companies such as General Assembly, Galvanize, Tech Elevator, or other bootcamp programs.

Talent sourcing can seem easy enough, but some pitfalls do exist. Leading organizations know how to avoid these pitfalls and have identified the metrics that best predict whether a university, school, or other educational program will yield the "right" talent for their data science program. Essentially, it's about not falling for "false metrics." For example, the employee tenure from particular sources may be a false indicator, because it may predict that an employee has no value outside of his/her current organization. Instead, employee performance data is a likely better predictor of the value of a source of talent. For example, if all talent hired from University X is averaging 4.82/5 on their annual performance reviews with an average tenure of 1.8 yrs., that may be an indicator that University X produces strong talent. While tenure may seem disappointing given the current costs of recruiting, the value that the employee provides in the 1.8 yrs. may be worth the costs. It may also trigger a further discussion within the organization about the employee value proposition and whether it is in alignment with competitors.

Booz Allen cultivated one of the industry's largest data science teams by leveraging some of these techniques. Our established partnerships with over 20 universities have created strong pipelines of technical talent into our organization. In addition, Booz Allen's efforts to support data science for social good have in turn helped our own employer brand. The launch of the Data Science Bowl in 2014, the world's premier data science competition for good, is particularly appealing to the next generation of workers who, research has shown, have greater interest in being connected to purpose-focused work.

Hiring Talent

Attracting and finding the right talent is the first half of the battle. The second half is ensuring the talent that you've found is the right fit. This is the goal of hiring. Hiring data scientists can be challenging because an organization needs to balance the right technical skills with cultural fit and attitude. It is important to note that in many organizations, data science becomes an internal customer service role. Therefore, data scientists must have the right cultural alignment for the organization. So how can the hiring process be designed to measure skills and cultural fit?

Booz Allen knows that successful data science teams conduct more than one interview and they advise interviewers to probe for different skills and competencies. For example, the first interviewer may be asked to assess a candidate's technical skills and may use questions such as:

- Provide an example of when you conducted data pre-processing.
- Describe a project you worked on in which you developed strategic insights from large data sets.
- Tell me about a time when you programmed a custom algorithm and how it impacted the project.

A second interviewer instead may probe a candidate's communications skills and ability to work within a team environment. For that interview, questions might include:

- Explain regression analysis as if you are talking to your grandmother.
- Describe a multi-disciplinary team that you have worked on – how did your skill sets come together to get the job done? How could you have worked more effectively to leverage each other's unique skills?

The examples above dig into two key personality areas, but organizations with strong cultures may have more than two areas that they value. The National Business Research Institute study shows that a bad hire can have significant costs to an organization – between $25,000 and $300,000[5]. Asking a candidate to partake in four interviews may seem like overkill, but it seems trivial compared to the cost of a "bad hire." Organizations that hire the best data science talent ensure they spend the time to use the best hiring practices.

[5] *https://www.nbrii.com/blog/the-cost-of-a-bad-hire-infographic/*

TALENT PLACEMENT: WHERE YOU NEED THEM

Leading data science organizations plan for their people long before recruiting staff. Talent Placement focuses on planning for, managing, and optimally placing analytics practitioners within the organization. Therefore, an effective data science team is well-balanced and well-positioned to drive analytics throughout the organization.

Workforce Planning

Organizations can save time and resources by regularly evaluating and addressing their analytics talent needs. Data-driven organizations clearly define and forecast their workforce needs (specifically for analytics); otherwise, they risk hiring people with unnecessary, redundant, or inadequate skills. Therefore, effectively planning for and managing a data science team is an ongoing process.

To plan for an analytics workforce, organizations should rely on the expertise of their workforce planners in combination with their current data scientists. Booz Allen partners with analytics organizations to review the portfolio of analytics projects "on deck" for the team. Organizations can develop a rough estimate of the size of the team required by combining a portfolio analysis with a metric for the average time to build a model. For example, if the average project takes 3 FTEs over 16 weeks, assuming normal data wrangling and full access to data and technology, and there are 8 projects to complete in a given year, then the size of the team would be approximately 6 data scientists (not including project management and leadership support).

Workforce Management

In the past, it was easier for organizations to ensure that they had the right number of employees with the right skills in the right positions. For a discipline like data science that crosses so many career fields or occupational series, it is no longer a simple task. Instead, it becomes critical to define the analytics workforce so that talent management programs can support the right staff. Besides, how can analysis be done on the analytics workforce without data?

Organizations need to identify their data science talent in a profession that lacks clear definitions. There is no occupational series for a data scientist in the federal government, and many companies have individuals operating as data scientists without the formal title. So, how do you identify the workforce in a meaningful way so that you can track critical trends (e.g., turnover, tenure, etc.) and plan for/create tailored career development programs based on the number of staff with training needs? In addition, if your organization has the goal to become more data-driven, how can you begin to identify those staff with the personality or basic skills that lend themselves to data science?

Building upon our Data Science Competency Framework, Booz Allen partnered with Hogan Assessments, a leader in personality assessments, to develop a **Hogan Assessment for Data Science Potential**. This scorecard measures potential across 13 different personality and cognitive competencies that result in strong data science performance, such as collaboration and teamwork, data science consulting, and perseverance. Candidates take two of Hogan's online surveys and receive a customized scorecard that indicates a Low, Moderate, or High data science potential marking on each of the 13 competencies with an overall potential score out of 100.

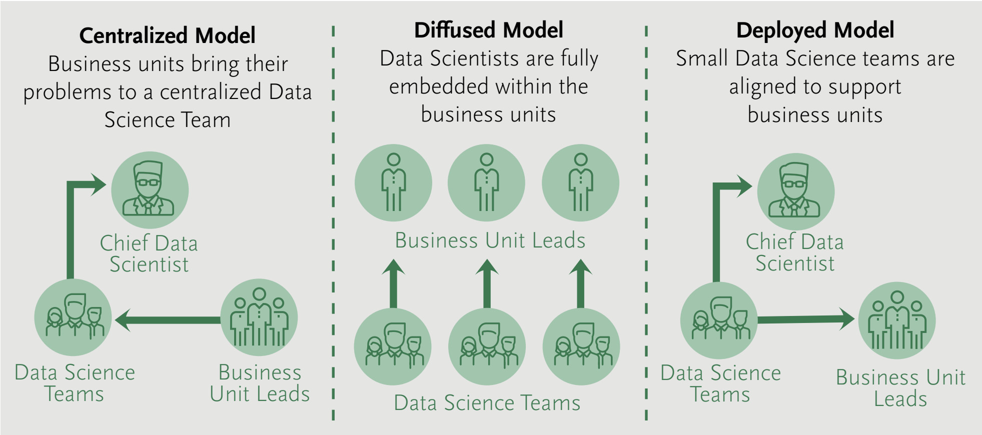
For organizations interested in measuring the technical proficiency of their current workforce, Booz Allen's Data Science Competency Framework may also be used to perform a **Data Science Technical Competency Assessment**. The approach provides four levels of proficiency – Basic, Foundational, Full Performance, and Expert – for each of the nine technical competencies and Data Visualization, a reach back competency. The assessment is administered through a survey which takes between 15-20 minutes, and can be designed to be a self-assessment completed by the individual, or may be completed by the manager. The survey relies on behavioral indicators which describe what it means to have proficiency at each of the four levels. Additional questions about the use of data analytics in the course of employee job, years of experience, and related training courses may be incorporated as needed.

Workforce Design

There is no right way or wrong way to position your data science team. Leading organizations think regularly about whether the current organizational structure is driving or inhibiting data-driven decision-making and they consider how to modify their existing design to support their strategic objectives.

There are infinite organizational charts for where to place your data science team, but there are three guiding constructs shown below. Each has positives and negatives that should be considered as part of any organization's design. For example, a centralized model may have greater efficiency with limited resources, but may require data scientists to “sell analytics” to the business units. In a diffused model, data scientists have a deepened understanding of the business, and there is less peer collaboration and rigor placed on consistent analytic standards. In the deployed model, project diversity will be great and practitioners build functional and domain knowledge, but there is always the risk of competing priorities when staff “serve two masters.”

Figure 12: Data Science Organizational Structures



One recent trend across industry and government alike is the rise of leaders with titles such as Chief Data Officer (CDO), Chief Analytics Officer (CAO), Chief Data Scientist (CDS), or Director of Data and Analytics. The emerging and evolving role of these Data and Analytics Catalysts addresses the need to better collect, manage, and exploit data assets and apply analytics to enable richer insights, regulatory compliance, and transparency. Reporting structures vary but typically a CDO or similar role reports to the organizational head, Chief Information Officer/Chief Technology Officer, or other business unit. Catalysts typically perform the following functions: data strategist, data evangelist, information/data governance, data steward/management, data standards/quality, analytics, and technologist. Booz Allen has broken down the challenges and ways to overcome them for four common types of Catalysts in the **Data and Analytics Catalyst Playbook**.

Also, Booz Allen conducted the first-of-its-kind **Data and Analytics Catalyst Trends Study**. The study finds that leaders are driving a culture change, not simply a technological shift. Achieving a culture change requires well-defined but adaptive governance processes and enabling technologies, but these solutions alone are not sufficient to bring the cultural revolution. We found that regardless of industry group, Catalysts achieve culture change through four strategies: demonstrate value, inspire innovation, establish partnerships, and nurture strong talent.

TALENT DEVELOPMENT: HOW YOU KEEP THEM

As the best analytics professionals will always be in high demand, organizations must continually provide opportunities for growth and development within their data science community. Effective talent management strategies, professional learning opportunities, consistent and constructive performance management, and motivational systems like rewards or growth opportunities help organizations retain and harness the maximum value from the data science talent.

Talent Management Strategy

Organizations require a dedicated talent management strategy focused on building and growing their data science practitioner base. The strategy identifies how existing human resource programs can be modified or adapted to address the unique needs of the analytics workforce. Booz Allen implemented a best practice in data science talent management by creating an HR leader focused entirely on the needs of the analytics workforce. This leader operates almost like an HR Business Partner pulling from the functional areas of HR to deliver services to the needs of a business unit, except the HR Partner focuses on an organizational discipline. Based on the data and the organization’s challenges, the HR leader should identify the two or three objectives to focus on (e.g., building a data science leadership pipeline, developing a learning roadmap).

Career Development

Organizations must combine traditional learning and development programs with new programs to engage and motivate high-demand data scientists. Experiential learning, rotational programs, and project-based opportunities are the best types of career development. Rather than simply sitting in a traditional classroom-based learning program, employees can build professional relationships while working and improving their skills – experiential training benefits everyone.

Leading organizations recognize that classroom-based training accounts for only 20% of learning, while the rest comes from experiential opportunities. That doesn't mean that traditional programs should be overlooked. New hires recently out of academic programs are familiar with classroom environments and may trend toward those types of opportunities. Therefore, organizations must offer a range of career development programs that span all types of learning modes and provide staff with the best opportunities for growth. Booz Allen recently launched a data science training program called the Data Science 5K Challenge to train an additional five thousand data scientists through General Assembly. Individuals that didn't qualify for General Assembly's program were offered online training through Coursera to prepare for General Assembly's second cohort.

Performance Management

Just as data science is both an art and a science, performance management follows the same philosophy. Good performance management systems should have the rigor of science, industrial and organizational psychology as well as the art of emotional intelligence and leadership. Although performance management shouldn't be vastly different across different disciplines, leading organizations know how to frame assessment around the user. Instead of tying goals to communication, tie the goal to problem solving and shift the outcome away from quantitative analysis. No matter the individual, employees need to feel like they are getting personalized attention and support. Employees also like to know that they are part of a team with a bigger mission and overarching vision. Understanding how wrangling data fits into that broader mission can often be overlooked by technical leaders, but shouldn't be.

Retention Strategies

The very same characteristics that make someone a good analyst – intellectual curiosity, a desire for experimentation, and a drive to always do and learn more – may also push them to look for new, more challenging opportunities both within their current organization and potentially beyond. Retention strategies are mostly driven by career development, but it doesn't end there. Organizations with high retention maximize all available "levers". For example, if their employees want better work/life balance, they offer more telework options (in the public sector) and unlimited time off (in the private sector).

Overwhelming research demonstrates that employees regularly leave due to dissatisfaction with their managers, not their jobs. Therefore, organizations must ensure that data scientists receive the right leadership and management. Within the chapter on Culture, Booz Allen provides information on the importance of having strong leadership to build, develop, and reinforce your data science team. Open door policies sometimes just mean that the door is open, but may not facilitate open communication. Instead, managers should engage with staff, ask to see a demonstration of an analytical model they have built, or ask for an explanation of their philosophy for writing code – open rapport is critical.

The Booz Allen Difference

Using our industry-leading insights, expertise, and proven techniques, Booz Allen helps organizations take the right steps to capitalize on the potential of data science by identifying their analytics needs and attracting, hiring, and retaining talent that can deliver. Booz Allen ensures that human capital development happens in a systematic, scientifically-proven way to maximize results. Our deep human capital experience, coupled with our unique understanding of the data science workforce, can help organizations win the war for high-quality data science talent.

CHALLENGE	BOOZ ALLEN SOLUTION	DESCRIPTION
Need help building and managing a data science workforce	**Data Science Talent Management Model**	Suite of integrated tools that helps organizations strategically manage data science employees across the talent life cycle
Don't know what competencies make a good data scientist	**Data Science Competency Framework**	Solution for helping organizations define the knowledge, skills, and abilities a data science team needs to be successful
Identifying the set of people required to round out a data science team	**Data Science Profiles**	Set of profiles which describe the roles, responsibilities, requirements, and likely occupational series (for federal government) for a data science team
Not sure where to find data science talent	**University Relationships**	Resource for recruiting top data science talent
Need to identify who has the potential for data science	**Hogan Assessment for Data Science Potential**	Individual survey to measure the personality characteristics aligned to success in data science
Need to assess proficiency in data science competencies	**Data Science Technical Competency Assessment**	Survey to assess an individual's technical skill using behavioral indicators – can be self-assessed or taken by manager
Don't know where to place your data science talent	**Data Science Organizational Structures**	Models that help organizations determine the optimal organizational strategy for incorporating analytics personnel into the workforce
Need to better understand the role of Data of Analytics Catalysts	**Data and Analytics Catalyst Playbook**	Provides a holistic framework to help Catalysts guide their organizations through the data and analytics journey
Need to understand how Data and Analytics Catalysts can be successful	**Data and Analytics Catalyst Trends Study**	First-of-its-kind study on Data and Analytics Catalysts that describes the strategies Catalysts should employ to be successful in their role

6 | TECHNOLOGY

- + *Does your organization ingest and store data in a way that enables it to be easily managed and analyzed?*
- + *Does your organization have tools that allow users of varying skill levels to interact with data?*
- + *Do you put as much emphasis on the front-end user interface and visualizations as the back-end technology?*
- + *How does your organization think about, plan for, and integrate new and emerging technologies that may make certain analyses more efficient?*

Technology enables data science techniques with effective tools to ingest, curate, and analyze data. There is no single perfect technology for data science – each technology has benefits and weaknesses and is designed for specific purposes. Organizations evaluate technologies and develop a suite of tools that support its strategy and analytic operations. An organization's needs, resources, and current analytic maturity all influence technology selection.

An organization must place equal attention on each stage of technology adoption to ensure it successfully incorporates the right IT. This begins with a technology assessment to determine needs and possible procurements. Then, you evaluate and build the tools and services necessary to maintain data science technologies. Next, organizations ensure access to the technology for the right people across the enterprise. With a skilled team implementing, using, and evaluating the outputs of the technologies, an organization positions itself to extract the maximum value and insights from data.

PLANNING & DEVELOPMENT: WHAT YOU NEED AND HOW TO PROCURE IT

It isn't enough to identify a need for new analytics technology. Effective organizations determine the best way to provide these technologies to their data scientists. Most technology organizations have processes for technology acquisition, but the nuances might be slightly different for analytics technology. Data science organizations recognize those differences and address them by modifying and adapting their acquisition strategy to meet their data science needs.

Planning and Needs Assessment & Analytics Technology Strategy

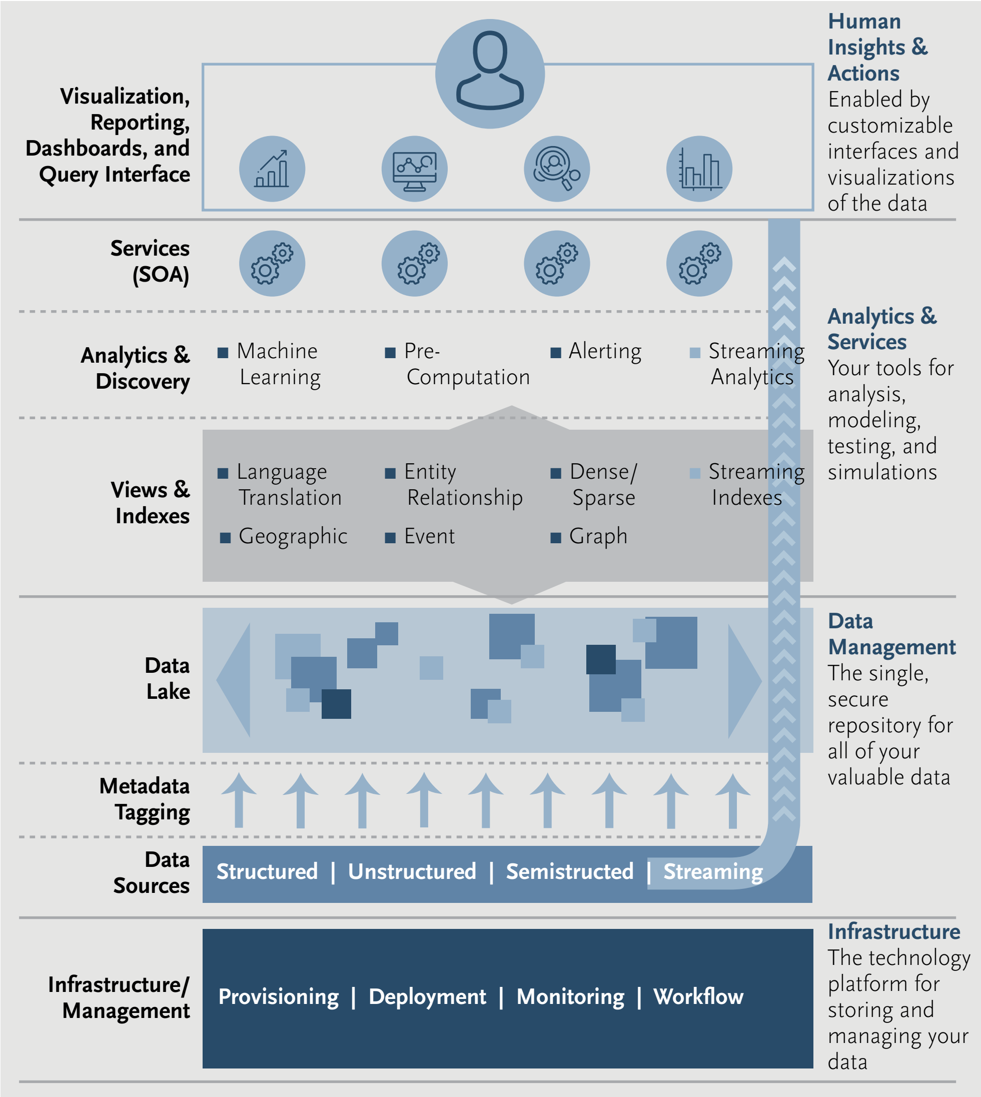
You must begin by evaluating and articulating technology needs. Organizations should conduct a comprehensive, internal assessment of the current state of their data science technology, anticipated or desired future state needs, and the gap between the two states. In the white space between current state and future state, organizations will find data science technologies that specifically meet their needs. From this gap analysis, an organization can develop a well-defined strategy (with refined mission, goals, actions, roles, responsibilities, and budget considerations) to develop, implement, and maintain the technologies that will drive the organization toward the desired future state.

Data science requires different strategic considerations for technology. One such consideration is that technologies must accommodate the different types of data that an organization ingests, such as APIs, cloud-based data systems, and machine-generated outputs. Another consideration is how an organization's technology will integrate with other data science capabilities. Third, organizations must consider how much of their data science needs can be addressed through open source technology.

Chapter Preview:

Booz Allen, in collaboration with its U.S. Government partners, developed a **Cloud Analytics Reference Architecture (CARA)** to address the nuanced, interconnected questions of technology management. CARA intelligently integrates big data and cloud computing technology along with advanced analytic capabilities and data science. Further, CARA automatically tags each piece of data with security metadata as it enters a central data lake. Organizations can use a variety of commercial off-the-shelf (COTS), government off-the-shelf (GOTS), and open source tools to organize the data. CARA has proven to be an effective way for organizations to consolidate their disparate data sources into one single source of trusted enterprise-wide data.

Figure 13: Cloud Analytics Reference Architecture



Visualization, Reporting, Dashboards, and Query Interface
Services (SOA)
Analytics & Discovery
■ Machine Learning
■ Pre-Computation
■ Alerting
■ Streaming Analytics
Views & Indexes
■ Language Translation
■ Entity Relationship
■ Dense/ Sparse
■ Streaming Indexes
■ Geographic
■ Event
■ Graph
Data Lake
Metadata Tagging
Data Sources
Structured | Unstructured | Semistructed | Streaming
Infrastructure/ Management
Provisioning | Deployment | Monitoring | Workflow
Human Insights & Actions
Enabled by customizable interfaces and visualizations of the data
Analytics & Services
Your tools for analysis, modeling, testing, and simulations
Data Management
The single, secure repository for all of your valuable data
Infrastructure
The technology platform for storing and managing your data

Development, Purchasing, & Maintenance of Technology

As part of a thorough self-assessment, effective organizations evaluate not only their technology needs and desires, but also whether it currently has the right skills and processes to develop, use, and maintain data science technologies. This is important for an organization to understand early because Office of the Chief Information Officer (OCIO) processes will determine whether an organization can build the technology in-house or must purchase it externally. In both cases, once the organization builds or procures the right tools and technology, it must also have the processes and resources in place to efficiently test, vet, implement, and maintain technologies. Going forward, organizations sustain formalized processes to continually improve the data science tools, technologies, and infrastructure.

Booz Allen believes in incrementally building technology through an Agile methodology. As the organization's data science infrastructure matures, it can incrementally add more advanced tools and technologies. This framework is founded on the premise that an organization is most successful with solid technology building blocks, rather than jumping straight to the most advanced technology. In other words, an organization should focus its attention first on developing the more basic infrastructure and capabilities needed to meet business goals. This is another reason why it is so important for an organization to have an accurate understanding of its current level of analytic maturity and a clear strategy toward the future state.

One challenge facing governmental organizations is the inability to get qualified and cleared data scientists. Booz Allen recommends developing an organization's data science technology suite through low to high development. Through Agile development, pieces of a solution can be built without access to classified data systems. These can be done on the low (non-cleared) side. Then, these technologies can be moved onto the high-side at the right time and evolved over time to meet shifting needs.

INFRASTRUCTURE: ENABLING DATA SCIENCE TECHNOLOGIES

Data science is constantly evolving, the methodologies and technologies are advancing rapidly, and it is critical to ensure that the infrastructure can keep pace. While it is easy to focus on the cutting-edge software and trendy tools, an organization designs its resources and services to support data science. Infrastructure provides the foundation for all other technologies and software.

Technical Architecture & Requirements

Data science technology is only effective when the technical requirements and architectural framework necessary to ingest, model, store, deliver, and integrate data into downstream systems are all in place. Careful data engineering and effective data ingestion up front are critical to the entire analytical process that follows. If an organization cleanly ingests, organizes, and indexes data, its ensuing efforts to retrieve and analyze the data downstream are also much easier and more accurate. Therefore, the processes and architecture for data ingestion must integrate seamlessly with the downstream technologies used for storing and analyzing the data. Once the data are curated and prepared for analysis downstream, organizations must determine the mechanisms they will put in place to secure and manage access to that data. Application-based interfaces (API), when established in a secure and scalable environment, enable organizations to enforce a single point of user access to the data stores, which ensures both data accessibility as well as data security.

Booz Allen developed the **Open Data Platform (ODP)** to facilitate the process of ingesting, integrating, and accessing large amounts of distributed and disparate data to support a wide range of data science and analytic requirements. The ODP solution is an enterprise-scale, data management platform that integrates best-of-breed open source technologies derived from Booz Allen's extensive experience supporting the Department of Defense (DoD), Intelligence Community, Federal Government, and commercial clients. The open architecture approach empowers organizations to select and integrate the most effective combination of tools from the open source community to meet their specific data science needs.

ANALYTIC TOOLS AND SERVICES: ACCESSING AND EXPLOITING DATA

At the core of data science technology are the tools and capabilities users need to access and analyze data. These are the technologies themselves that users will leverage to interact with their data and extract valuable business insights. Generally speaking, effective data science technologies are easily accessible to users, adaptable to specific organizational needs, and adept at interacting with particular types of data and infrastructure. However, the adage "no technology is created equal" still stands, and organizations must refer to their own data science self-assessments and strategies to identify the data science tools best suited to their analytical and business needs.

Sophistication & Accessibility

Several platforms deliver the data storage, management, and access that organizations need today. For example, Amazon Web Services (AWS) is a secure, cloud services platform offering computing power, database storage, content delivery, and other functionality for

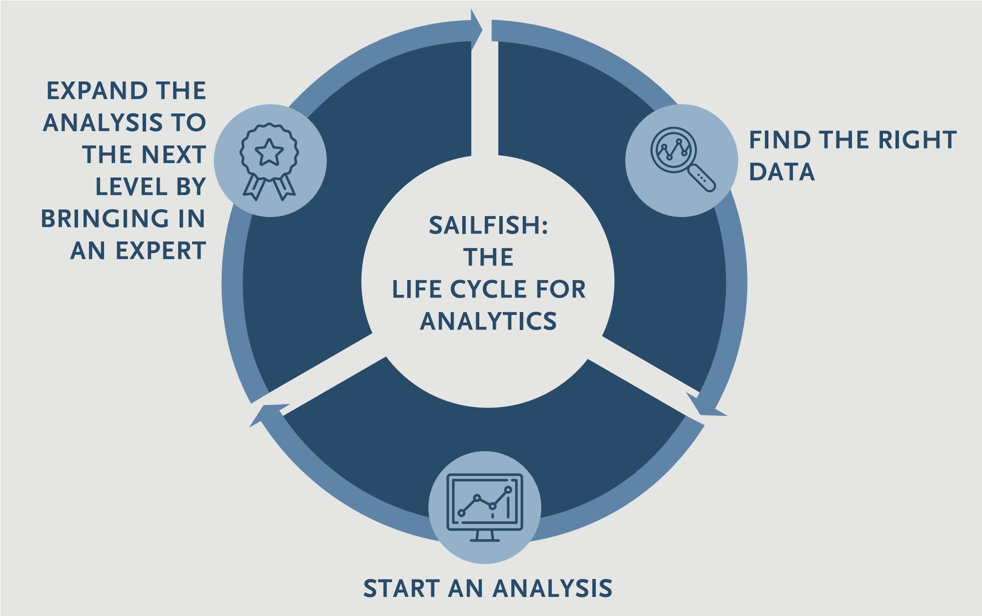
Case Study: Harnessing a Data Lake for Merck

Merck, a pharmaceuticals company, was experiencing high variability in vaccine yields during manufacturing and production. Booz Allen provided advanced aggregation techniques that aligned 16 disparate data sources around common metadata dimensions, forming a data lake and fully realizing the power of the company's vast data collection practices. As a result, Booz Allen could focus more attention on analysis, including the application of sophisticated pattern detection analytics to identify patterns associated with high and low yield. They then developed models to associate process controls with these patterns to ultimately determine the best ways to control yield. Due to these models and Booz Allen's customized investigative visualizations, Merck could explore the phases in vaccine manufacturing most closely tied to yield, potentially saving them millions of dollars per year in lost revenue.

over a million active users. For organizations in both the public and private sectors, the AWS secure cloud platform provides unprecedented levels of flexibility, scalability, and reliability. Microsoft Azure is another similar cloud-based platform that enables users to store and leverage data.

Similarly, the **Analytics Workbench (WB)** is a Booz Allen developed, cloud-based environment that integrates tools, data, and people to drive insights. The WB creates a platform that brings together industry-leading analytics tools, strategic partnerships within the industry, hundreds of curated data sets, and highly scalable infrastructure to store big data. Through the WB, Booz Allen places all of these resources in the hands of cross-domain teams. The Analytics WB itself is comprised of three main components: an analytics tools suite, cloud-based data repositories, and a custom portfolio. Each solution created on WB is uniquely adaptable to every user because WB allows users to integrate their current tool sets, access past work products, import data sets from any source, and use their own people. The visualization, processing, and storage functionalities of the WB enable users to better present, explore, collect, and access data.

Figure 14: Sailfish Life Cycle


EXPAND THE ANALYSIS TO THE NEXT LEVEL BY BRINGING IN AN EXPERT
SAILFISH: THE LIFE CYCLE FOR ANALYTICS
FIND THE RIGHT DATA
START AN ANALYSIS

Booz Allen offers additional data science tools and technologies that are flexible to the particular needs of an organization. One of these tools, **Sailfish**, provides a comprehensive suite of data science tools and resources that are customizable to the particular data sets and analytic needs of organizations. Sailfish is composed of three applications that enable an organization to grow a data science capability. Sailfish Exchange allows users to curate data sets across the organization, regardless of location and file type. Sailfish Explore empowers users to analyze big data without needing to code in programming languages. Together with the Answers on Demand service that Sailfish also offers, Sailfish enables users to curate, analyze, and leverage their data to grow their data science capability.

HUMAN INSIGHTS AND ACTIONS: DISTILLING AND VISUALIZING INSIGHTS

After ingesting, curating, and analyzing data, organizations can gain further value and insights from their data by visualizing the data and analytics graphically. Data visualizations are an important tool for translating complex analytics into informative products that explain the transformation of data, reveal additional insights and value, and inform decision-making. Therefore, it is critical that organizations not only emphasize analytical technologies, but also value tools for data visualization.

Visualization Tools & User Interfaces

The technologies for data science visualization must align with the type of data or analysis and have an accessible user interface that empowers users to effectively leverage the tool. Often, visualizations provide further clarity, insight, and value to the underlying data, tools, and analytics.

One such visualization tool, **Quick Look Assessments (QLAs)**, provides a graphical evaluation of an organization's analytical maturity in certain areas. These insights help leaders transform and modernize their organizations. Booz Allen has conducted QLAs in over 30 organizations across the areas of: cloud transition, data and analytics, DevOps, cybersecurity, and mobile. In each area, Booz Allen developed and implemented assessments to evaluate an organization's maturity, the results of which are displayed in a visual analysis, scorecard, framework, or similar visualization.

Epidemico is another Booz Allen tool for monitoring various topics in mainstream and social media designed for high-fidelity, actionable insights. **Epidemico** cloud-based software offers custom portfolios by domain, including biovigilance (i.e., disease outbreaks, food safety), medvigilance (i.e., drug safety), toxicovigilance (i.e., illicit drugs), and more (i.e., brand perception, cybersecurity, event monitoring). Data are acquired from APIs, client databases, and direct aggregation. The data are securely stored and analyzed by domain machine learning tools with an extensive library of natural language processing and manually by health experts. Descriptive and inferential statistics are applied and results are delivered through a dashboard, summarized in reports, or pushed via API for further analysis. Epidemico offers processing as-a-service, data collection and processing, and subject matter expertise.

The Booz Allen Difference

Booz Allen recognizes the importance of data science infrastructure throughout the data lifecycle – from data engineering and ingestion to data storage and analysis. Consequently, Booz Allen developed a comprehensive suite of data science technologies that organizations can use to better understand their data, build more resistant and secure infrastructure, enhance analytical sophistication, and ensure that users of all skill levels can extract key insights from the data through easy-to-use interfaces. Booz Allen does not offer one-size-fits-all solutions. Instead, Booz Allen's technologies are completely customizable in scale, scope, and utility, offering the maximum resource efficiency to clients by enabling them to implement only what their unique data situation requires. The Open Data Platform allows organizations to integrate data from many disparate sources. Cloud Analytics Reference Architecture empowers organizations to conduct unique analyses on the data by standardizing data tags to create a single source of trusted enterprise-wide data. Sailfish is a big data analytics platform that democratizes data science and enables anyone within an organization to make data-driven decisions. With these and many other technologies, Booz Allen helps organizations execute their technology strategy, maximize the value of their data, enhance their analytic capabilities, and build a leaner, more successful data science operation.

CHALLENGE	BOOZ ALLEN SOLUTION	DESCRIPTION
Need to consolidate disparate data sources into one single source	**Cloud Analytics Reference Architecture**	Technology that tags each piece of data entering a data lake with security metadata, creating a single source of trusted enterprise-wide data
Need help facilitating movement of data across the enterprise	**Open Data Platform (ODP)**	Free and Open Source Software that is built to accelerate an organization's ability to ingest, integrate, analyze, and interact with data of all source types by facilitating the secure movement of data across the organization
Don't know how to share data across the organization	**Sailfish Exchange**	Social platform for managing and engaging with your data, making it easy to collect, organize, and catalog data sets from any source or file type
Employees don't have analytical skills to use data insights for decision-making	**Sailfish Explore**	Analytics platform that enables analytics-driven decision-making for all members of an organization regardless of their analytics skills level
Need to monitor large quantities of different types of data for anomalies	**Multivariate Analysis of Stealth Quantities (MASQ)**	Algorithm that mines data for known and unknown signatures to detect fraudulent data transactions
Need help selecting and using the right technologies	**Analytics Workbench**	Cloud-based collaborative platform that integrates tools, data, and people through an analytics tools suite, data repositories, and a custom portfolio
Need to manage, automate, and control your cloud environment	**Project Jellyfish**	Community-driven, open source self-service portal that delivers IT services, automation, and hybrid cloud management in a secure and policy-driven manner
Want to provide organizational leaders with a visual of the organization's analytics maturity levels	**Quick Look Assessments (QLAs)**	Visualization tool that provides a graphical evaluation of an organization's analytical maturity in a variety of areas, including cloud transition, data and analytics, DevOps, cybersecurity, and mobile

7 | CULTURE

- *Has your organization's leadership made it clear how data science plays a role in furthering your organization's mission and strategy?*
- *Do your organization's policies and procedures promote employees' active engagement with data science?*
- *Is it an organizational norm for all employees to use evidence-based decision-making?*
- *Does everyone across your organization, including non-analytics staff, agree upon and understand the value of data science?*

"Culture is King." "Culture eats strategy for breakfast." The phrasing changes, but the bottom line is the same – nothing new happens in an organization without either aligning to the culture or changing the culture. Becoming a data-driven organization is no different. Analytics success depends on the full organization accepting data science as a new norm and actively seeking opportunities to use data science across the organization.

Creating a data science culture depends on three key factors: activating leaders, engaging employees, and building organizational mechanisms. Leading organizations market and brand their analytics, communicate them through every possible channel, and find leaders to drive the effort. On the other side of the equation, organizational enablers create the conditions for data science to flourish through known and communicated mechanisms. Analytics must be democratized to sustain data science over the long-term – with each member of an organization taking collective ownership for its success.

Long-term investments in technology platforms and hiring data science talent will yield no return without the right culture to support these investments. Booz Allen's experience integrating data science into organizations demonstrates that culture is often the hardest feature to change, and the easiest to overlook. However, starting the cultural shift toward data science can be as simple as: agreeing that the organization isn't perfect, acknowledging it could get better, and recognizing that maybe data science can help.

Chapter Preview:

ENGAGEMENT: HOW YOUR STAFF EMBRACES ANALYTICS

Engagement is about creating a dialogue between the organization and employees on the topic of data science. Early on, the dialogue should focus on establishing a baseline level of awareness and building a shared vocabulary for data and analytics. As the organization matures it can focus more on building knowledge bases, sharing available resources, and celebrating early successes. A best practice organization isn't even actively thinking about how it communicates, markets, brands or recognizes analytics because it is so enmeshed in the fabric of the organization. Every activity and decision calls for analytics and all strategic decisions are pilot tested to measure the potential effectiveness before a major release. Engagement is critical early on, but may become second nature as an organization improves its data science practice.

Communications

Messaging analytics is an important first step in building adoption and change within an organization, as it ensures that you are communicating the right things to the right people. Organizations need to emphasize the value of data science in supporting the success of the business. An organization's key messages should: 1) reveal what analytics can do and how, 2) why this is important, 3) how this fits with the strategic vision, and 4) speak to all audiences.

Because data science is a new and growing field, there is likely information inequity amongst employees. Therefore, communications should provide basic education on the topic of analytics to a broad audience. But for advanced practitioners, communications and outreach may include technical tips and tricks on natural language processing or Python. For leaders, there may be discomfort with the growing expectation that they must make decisions relying on sophisticated analytic techniques they don't understand. For these leaders, effective communications may require more than just stating facts. One of the most impactful methods of sharing a message is through applicable examples. Demonstrating how a piloted technique brought time savings or process efficiencies to a part of the organization is one way.

Booz Allen helps organizations create communications strategies which may include analytics-focused newsletters and portals on internal websites where anecdotes can be shared, ultimately showing how valuable analytics can be. Continued communication shows the workforce that the organization takes data science seriously and is committed to its enduring success. This not only instills confidence in the data scientists, but encourages the rest of the organization to embrace the change.

Marketing & Branding

Advertisers and retailers know the value of marketing and branding – when done correctly, it can subtly influence change without the end user's awareness. Placement of a cereal box at exactly the right height in a supermarket, with exactly the right color and design, may encourage you to try the cereal without your conscious thought. Organizations that want to create the same desire for analytics should rely on the similar techniques. Marketing and branding should be a critical part of the culture of analytics. Many organizations have "brand identities" that come with an associated set of templates, color palettes, and associated collateral – but true data science organizations extend those brand identities into data science. These leading organizations go the extra step and offer customized excel templates pre-loaded with the right color schemes, automated reports using standardized formats, and guidance available to all employees to maintain a visual identity around the organization's analytics.

Leadership

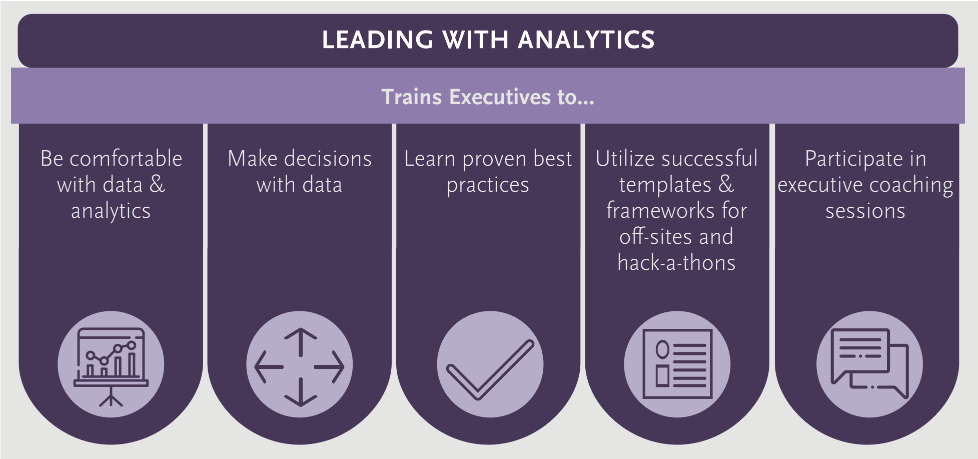
Building or enhancing a successful data science capability requires visible and committed leaders at the top of the organization. Leaders emphasize the importance of integrating analytics into everyday operations. However, there are very few leaders who have fully embraced this uncharted role. And, as each organization is unique, a one-size-fits-all approach doesn't work. Therefore, organizations must charge ahead and build the path along the way, leveraging the lessons learned from innovators and early adopters in the field.

Booz Allen develops leaders by emphasizing both data science advocacy (how to support data science) and data science literacy (an understanding of data science tool and techniques) with our comprehensive, multi-modal leadership development program, **Leading with Analytics**. Leaders are provided with: 1) a crash course in data science, 2) a peer-driven speaker series, and 3) executive coaching sessions designed to coach them through decision-making. Armed with this type of development program, leaders can actively engage in analytics discourse and identify increased opportunities to apply analytics within their organizations. This will result in higher adoption of analytics and more sustained cultural change.

Case Study: Improving Visualizations at a Large Benefits Agency

The Analytics Hub at a large Benefits agency is responsible for building an Advanced Analytics Community of Interest, Workforce Development (in Advanced Analytics), and Improving the Capability of the Agency (in Advanced Analytics). The Hub partnered with Booz Allen to deliver on these three services and develop a Data Visualization Guide to market Advanced Analytics, build interest for the community, and improve the skill base simultaneously. Booz Allen's guide lays out the steps and considerations the agency must take when creating visualizations, including: consider the audience and purpose; define what needs to be shown; choose an appropriate type of visualization; and select the right tool(s) to create the product. Together with a visualization effectiveness evaluation form, the agency can use this guide to better translate complex analytics into valuable data visualizations that inform evidence-based decision-making.

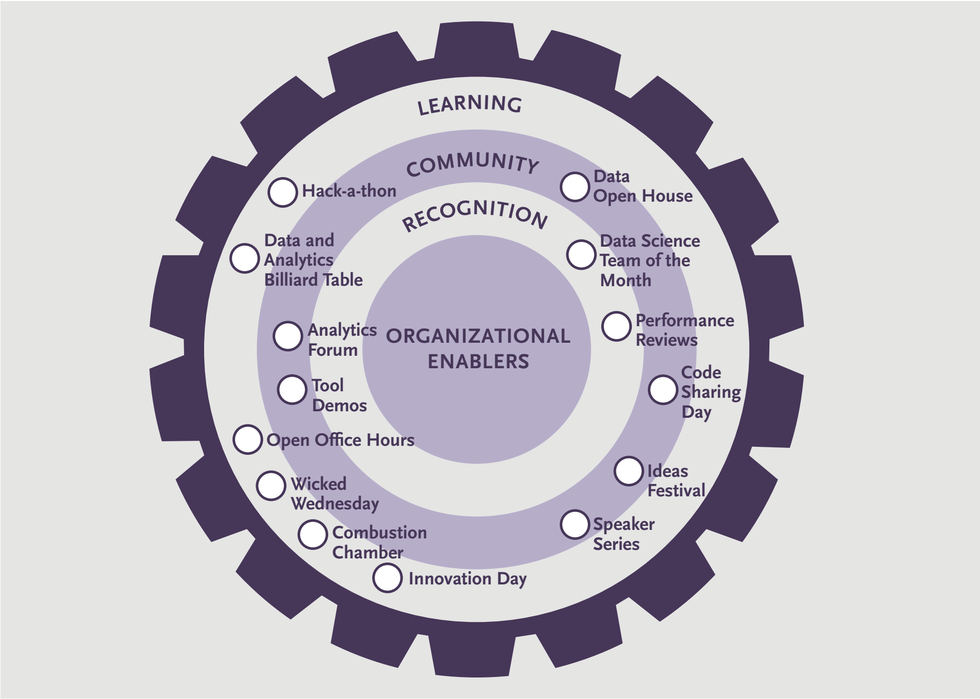
Figure 15: Leading with Analytics


LEADING WITH ANALYTICS
Trains Executives to...
Be comfortable with data & analytics
Make decisions with data
Learn proven best practices
Utilize successful templates & frameworks for off-sites and hack-a-thons
Participate in executive coaching sessions

ORGANIZATIONAL ENABLERS: HOW YOU FOSTER ENDURING CHANGE

In order to reinforce and sustain data science, organizations must establish analytics programs that support data science staff. Booz Allen uses proven change management approaches to establish an analytics capability and cement a culture of data science. These approaches help organizations understand stakeholder needs and equip stakeholders with the activities required to prepare the organization for coming changes. It gives an organization the tools to: train employees on effectively adopting cultural change, measure the impact of change, and communicate the successes resulting from change. Data science is the wave of the future, and organizations will only succeed if they reinforce the value of data science. Through knowledge management, awards programs, and other organizational enablers, organizations must create a data science ecosystem for analytics to thrive.

Figure 16: The Analytics Ecosystem



Case Study: Applying Organizational Change Management at Department of Homeland Security (DHS) Immigration and Customs Enforcement (ICE)

Booz Allen supported the long-term business transformation planning for the stand-up of the Pacific Enforcement Response Center (PERC) at Immigration and Customs Enforcement (ICE). Booz Allen aligned multiple work streams, developed a full continuity of operations (COOP) plan, and created a phased project plan. Additionally, to support these workload and business process changes, the Booz Allen team provided advanced change management and business planning support. This included developing an Alternatives Analysis to guide executive-level decision-making on the most efficient and effective approach to staffing and an updated workforce/workload model (based on historical and current data, interviews, and on-site observation) to ensure that the PERC has the appropriate number of personnel to handle its current-state and future-state analytics and processing responsibilities. By using data to drive the decision-making and formalizing new policies and procedures, Booz Allen helped ICE to more efficiently distribute and execute the organization's analytic workload and successfully guide the organization through the changes.

Recognition

A visible way to highlight the value of data science in an organization is to recognize the impact of data science – and the teams that helped bring about that change. Further, analytics should be accessible to each and every staff member in order to realize its full value. The individuals that identify the opportunities, raise them to leadership and manage the project through to completion often have as much value to the realization of the solution as the analysts themselves.

Recognition can come in many forms, all of which carry significance. An acknowledgment does not always have to come as monetary compensation, which may be problematic for federal organizations. Here are a couple of ways to provide non-monetary recognition:

- Identify a "Data Science Team of the Month," a distinction recognized across all communication platforms and highlighted by leadership in monthly meetings
- Add recognition of data science efforts in performance reviews to contribute to the proposed cultural shift
- Highlight successes and appreciation during team meetings or organizational town halls to show the rest of the staff that data science work is valued
- Highlight the role of data science in the "every day" work of the organization by bringing out the data science role in organizational outcomes

Policies & Procedures

To formalize analytics into its culture, an organization must also assess and potentially re-evaluate its policies and procedures. This may mean institutionalizing data-driven decision-making in policies throughout the organization. For example, before deciding to open a new retail location, a company's internal rules may require market research and customer analysis to justify the expense. It should not allow a regional manager to open a new store just because he/she "has a hunch." However, there are other types of decisions that are done with analysis, but not the type of robust analysis that may be available with today's tools and technologies, simply because "that's the way it's always been done."

Organizations must cast off these ties to the past and accept that they can do better, and then attempt to do so. Federal organizations are often associated with this type of thinking, but the data required for achieving government missions is often less publicly available. However, this presents a new opportunity for federal organizations to create crowd-sourced challenges internally to help "solve unsolvable problems with analytics." If solved, the analytic process can be incorporated into policies and decision-making practices. If unsolved, then nothing has changed – leaders can continue to make decisions as they always have.

Although formalizing policies can be a key pillar of building a data science organization, it can also be one of the most difficult to accomplish. Policy changes, especially in federal organizations, go through formal approval processes, which can be time-consuming. In addition, once implemented, policies must be diligently disseminated and monitored to ensure stakeholders are held accountable for adhering to the guidance. While this may seem overwhelming, a policy change can be the most effective way to ensure that data science endures within an organization.

Community

Organizations built upon a strong discipline of data science also maintain strong communities of practice/interest. These communities provide networking opportunities to grow deep technical skills and increase staff knowledge base, while facilitating direct data sharing across departmental lines. Community events such as data science exchange seminars or open houses offer practitioners the opportunity to self-identify and meet others with similar interests. Although events such as this are often the first to be cut in times of fiscal austerity, communities of practice/interest provide efficiencies for the organization, as practitioners share resources, code, and informally learn from each other. Additionally, these events increase staff engagement, which increases retention of high demand talent.

Learning

Organizations can further build data science culture by offering casual community-based learning opportunities. Since data science is a complex and emerging field, an organization can go a long way in cultivating knowledge and skill through learning opportunities. These programs can range from video blogs on "What is Data Science", to more abstract discussions led by organizational luminaries on whether data science is a single discipline or is multi-disciplinary. They can be generated by the organization or purchased from outside vendors developing an ever-growing set of content.

On SharePoint, data science teams can offer reusable tools and methodologies to their data science community. These tools and methodologies can include: a code repository, data sets cleansed and prepared for exploratory analysis, reusable models for repeatable challenges (e.g., resource allocation), or even documentation of analytic methodologies.

Booz Allen has conducted organizational reviews of analytics resources, and designed potential knowledge management sites that best aggregate and align documents and artifacts to support the practitioner base. As one of the largest employers of data scientists in the world, Booz Allen has focused resources on creating experiential learning programs for its own staff. As part of its data science journey, Booz Allen sought ways for leaders to share organizational knowledge through informal (non-classroom) settings. Some examples of these activities are:

Figure 17 Learning Activities

LEARNING ACTIVITY	DESCRIPTION
Wicked Wednesdays	On a Wednesday, a "wicked problem" is posted, and staff come together to work side-by-side away from their clients and solve the challenge presented
Hack-a-thon	Working in teams to solve a data science challenge and present your solution to a panel of data science leaders
Innovation Day	Large scale open-house events to showcase different data science capabilities, tools, and techniques in combination with industry partners
Combustion Chamber	Modeled on the TV show Shark Tank, Combustion Chamber competitors pitch their ideas in front of a group of Booz Allen's senior leaders, who decide if the firm should invest in the team's idea as a future client solution
Data and Analytics Billiard Table	An immense data set was gathered together and placed atop a billiards table in a small office. All employees were offered the chance to explore this analytics sandbox to expand their skills, try new ideas, develop visualizations, integrate mobile technology, and whatever else felt inspiring
Open Office Hours	Open time available for staff to drop in, meet leaders, and learn more about data science from a leadership perspective
Kaggle	Online challenge events to inspire staff, identify key talent, and build our presence in the global community

Case Study: Building a Community of Interest at a Large Benefits Agency

In support of a large benefits agency's Analytics Hub, Booz Allen helped stand up an Advanced Analytics Community of Interest (COI). As part of this effort, the team developed branding, communications products (posters, flyers, table tents, newsletters, etc.), and events. The kick-off event for the COI had over 400 attendees from all components of the agency. During this event, the acting commissioner and other executives reminded staff of the strategic value and importance of advanced analytics to driving the agency's mission. Each component was provided a booth, where they were able to highlight their work in the area of advanced analytics, and could recognize others' contributions to the field. The event created strong relationships between practitioners, identified additional opportunities for the application of advanced analytics, and reinforced the impact that advanced analytics could have when applied in a focused and deliberate manner.

The Booz Allen Difference

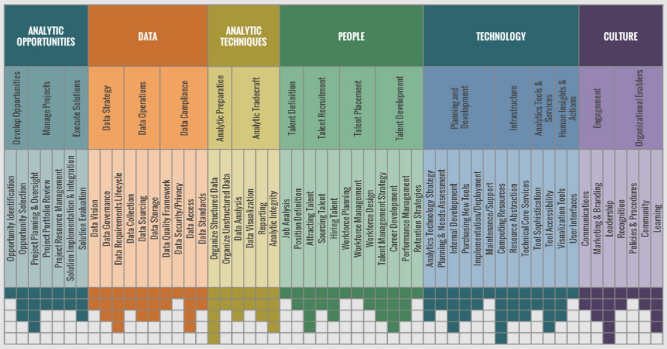


Figure 18: ADAPT+C Capability Maturity Model

Culture is the final element of the ADAPT+C model, and perhaps the most important. Organizations that dedicate resources to building a data science culture will realize success far earlier than those that do not. Booz Allen's leading practice in change management, combined with its depth and breadth of data science experience, positions us to help organizations overcome the cultural barrier and create organizational conditions where analytics will thrive.

Data science has already changed our world. It continues to do so in new and more exciting ways each and every day. With Booz Allen, organizations can bring those same exciting opportunities inside their own doors and begin to explore what is truly possible. But first, organizations must create a culture where data science can thrive.

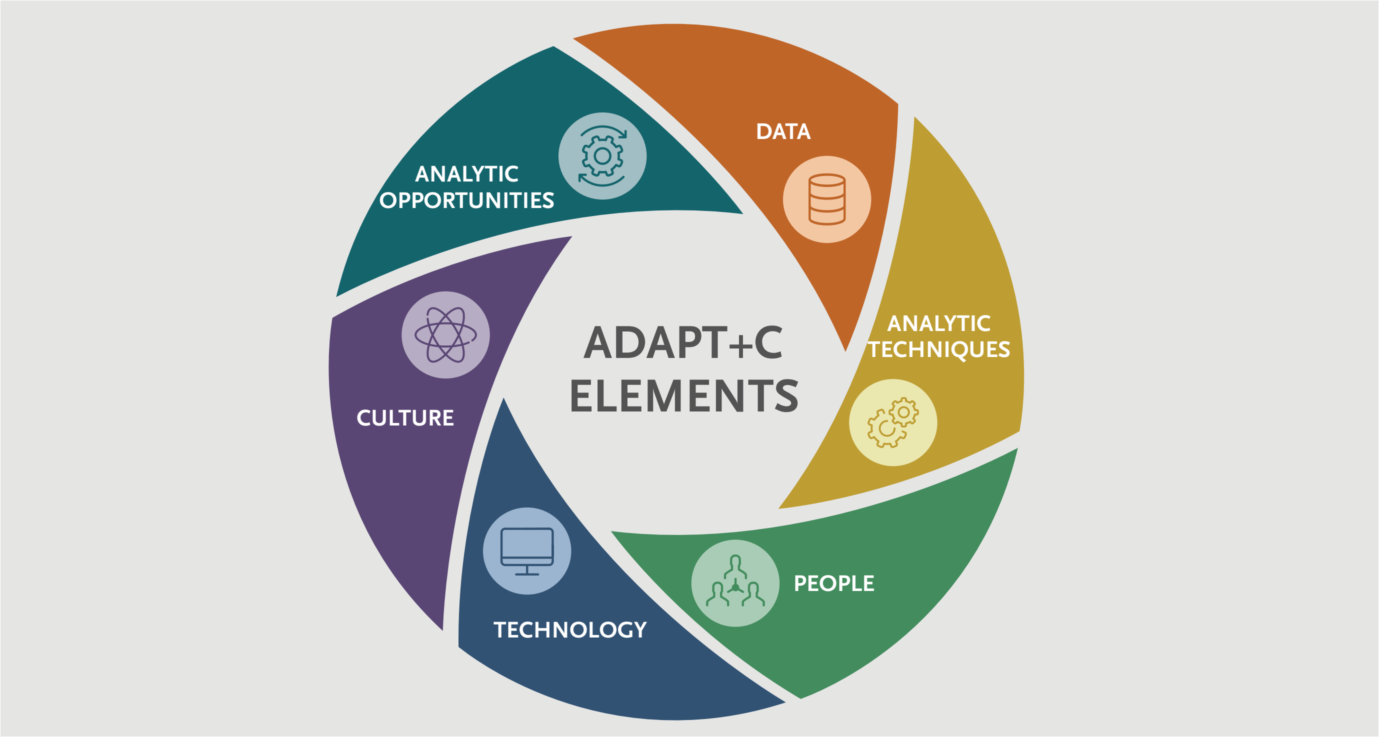
CHALLENGE	BOOZ ALLEN SOLUTION	DESCRIPTION
Don't know how to display insight from data in a way that is easy to understand	**Data Visualization Guide**	Collection of best practices for selecting and creating data visualizations that clearly depict the key insights derived from data analysis
Leaders in your organization aren't committed to using data and analytics	**Leading with Analytics**	Leadership development training program specifically designed to train executives to be more comfortable with data and analytics
Need help managing the change that occurs with establishing a data-driven culture	**Data Science Change Management Approach**	Strategic approach for standing up an analytics capability and establishing a data-driven culture and helping leaders identify and understand their organizations' stakeholders' needs and concerns
Not sure what your organization's strengths and weaknesses are	**ADAPT+C Maturity Model Assessment**	Assessment tool that helps organizations understand their maturity levels in each of the six elements that are critical to building an analytics capability
Want to develop a data science community throughout organization	**Data Science Bowl, D.C. Data Coders, Girls Who Code, Women in Data Science**	Organizations and events held or sponsored by Booz Allen to provide data science opportunities for employees, expand the firm's data science community beyond the internal staff, and use data science expertise for social good in the local community and beyond

8 | CONCLUSION

Data science may seem like the next trendy management fad designed to convince organizations to purchase software platforms and hire consulting organizations, but when applied correctly, it can transform businesses and mission attainment. Thinking, acting, and operating quantitatively – and using analytics as part of your everyday decision-making – is a major change. By breaking down the change into six simple elements, any organization can transform. It doesn't mean it won't take hard work and there won't be roadblocks ahead, but a comprehensive approach that keeps the elements in harmony will ensure that organizations evolve toward data science successfully.

It is no mistake that **ADAPT+C** was designed to represent a wheel, with each of the elements equal in size to the others. Successful data science organizations know that an analytics technology platform with self-service analytics is irrelevant without practitioners actively using it, and that a team of data scientists can't rely on data lacking quality controls. An organization must ensure that all capability elements are aligned and prioritized to successfully institutionalize data science. Technology might come before people, but not at the expense of people. People may come before data, but people can't do anything without data. This is where the hard work comes in: to effectively build, grow, and reinforce data science, an organization must take the time to understand what capabilities it has, what it needs, and the path it must chart to get there.

@BoozAllenDataScience



9 | APPENDIX

Figure 19: Detailed Table of Analytics from Booz Allen's Field Guide to Data Science

TECHNIQUE	DESCRIPTION	TIPS FROM THE PROS	REFERENCES WE LOVE TO READ
Active Learning	Intelligent sample selection to improve performance of model. Samples are selected to provide the greatest information.	Can be paired with a human in-the-loop to help capture domain knowledge.	Burr, Settles B. "Active Learning: Synthesis Lectures on Artificial Intelligence and Machine Learning." Morgan & Claypool, 2012. Print.
Agent Based Simulation	Simulates the actions and interactions of autonomous agents.	In many systems, complex behavior results from surprisingly simple rules. Keep the logic of your agents simple and gradually build in sophistication.	Macal, Charles, and Michael North. "Agent-based Modeling and Simulation." Winter Simulation Conference. Austin, TX. 2009. Conference Presentation.
ANOVA	Hypothesis testing for differences between more than two groups.	Check model assumptions before utilizing, and watch out for Family Wise error when running multiple tests.	Bhattacharyya, Gouri K., and Richard A. Johnson. Statistical Concepts and Models. Wiley, 1977. Print.
Association Rule Mining (Apriori)	Data mining technique to identify the common co-occurances of items.	Utilize when you have a need to understand potential relationships between elements.	Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast Algorithms for Mining Association Rules." Proc. Of 20th Intl. Conf. on VLDB. 1994. Conference Presentation.
Bayesian Network	Models conditional probabilities amongst elements, visualized as a Directed Acyclic Graph.	Calculate by hand before using larger models to ensure understanding.	Russel, Stuart, and Peter Norvig. "Artificial Intelligence: A Modern Approach." Prentice Hall, 2009 Print.
Collaborative Filtering	Also known as 'Recommendation,' suggest or eliminate items from a set by comparing a history of actions against items performed by users. Finds similar items based on who used them or similar users based on the items they use.	Use Singular Value Decomposition based Recommendation for cases where there are latent factors in your domain, e.g., genres in movies.	Owen, Sean, Robin Anil, Ted Dunning, and Ellen Friedman. Mahout in Action. New Jersey: Manning, 2012. Print.
Coordinate Transformation	Provides a different perspective on data.	Changing the coordinate system for data, for example, using polar or cylindrical coordinates, may more readily highlight key structure in the data. A key step in coordinate transformations is to appreciate multidimensionality and to systematically analyze subspaces of the data.	Abbott, Edwin A., Flatland: A Romance of Many Dimensions. United Kingdom: Seely & Co., 1884. Print.
Deep Learning	Method that learns features that leads to higher concept learning .Usually very deep neural network architectures.	Utilize a GPU to efficiently train complex models.	Bengio, Yoshua, and Yann LeCun. "Scaling Learning Algorithms towards AI." Large- Scale Kernel Machines. New York: MIT Press, 2007. Print.
Design of Experiments	Applies controlled experiments to quantify effects on system outputcaused by changes to inputs.	Fractional factorial designs can significantly reduce the number of different types of experiments you must conduct.	Montgomery, Douglas. Design and Analysis of Experiments. New Jersey: John Wiley & Sons, 2012. Print.

Figure 19: Detailed Table of Analytics from Booz Allen's Field Guide to Data Science (Continued)

TECHNIQUE	DESCRIPTION	TIPS FROM THE PROS	REFERENCES WE LOVE TO READ
Differential Equations	Used to express relationships between functions and their derivatives, for example, change over time.	Differential equations can be used to formalize models and make predictions. The equations themselves can be solved numerically and tested with different initial conditions to study system trajectories.	Zill, Dennis, Warren Wright, and Michael Cullen. Differential Equations with Boundary-Value Problems. Connecticut: Cengage Learning, 2012. Print.
Discrete Event Simulation	Simulates a discrete sequence of events where each event occurs at a particular instant in time. The model updates its state only at points in time when events occur.	Discrete event simulation is useful when analyzing event based processes such as production lines and service centers to determine how system level behavior changes as different process parameters change. Optimization can integrate with simulation to gain efficiencies in a process.	Burrus, C. Sidney, Ramesh A. Gopinath, Haitao Guo, Jan E. Odegard and Ivan W. Selesnick. Introduction to Wavelets and Wavelet Transforms: A Primer. New Jersey: Prentice Hall, 1998. Print.
Discrete Wavelet Transform	Transforms time series data into frequency domain preserving locality information.	Offers very good time and frequency localization. The advantage over Fourier transforms is that it preserves both frequency and locality.	Burrus, C.Sidney, Ramesh A. Gopinath, Haitao Guo, Jan E. Odegard, and Ivan W. Selesnick. Introduction to Wavelets and Wavelet Transforms: A Primer. New Jersey: Prentice Hall, 1998. Print.
Ensemble Learning	Learning multiple models and combining output to achieve better performance.	Be careful not to overfit data by having too many model parameters and overtraining.	Dietterich, Thomas G. "Ensemble Methods in Machine Learning." Lecture Notes in Computer Science. Springer, 2000. Print.
Expert Systems	Systems that use symbolic logic to reason about facts. Emulates human reasoning.	Useful to have a human readable explanation ofmwhy a system came to a conclusion.	Shortliffe, Edward H., and Bruce G. Buchanan. "A Model ofmInexact Reasoning in Medicine." Mathematical Biosciences. Elsevier B.V., 1975. Print.
Exponential Smoothing	Used to remove artifacts expected from collection error or outliers.	In comparison to a using moving average where past observations are weighted equally, exponential smoothing assigns exponentially decreasing weights over time.	Chatfield, Chris, Anne B. Koehler, J. Keith Ord, and Ralph D. Snyder. "A New Look at Models for Exponential Smoothing." Journal of the Royal Statistical Society: Series D (The Statistician). Royal Statistical Society, 2001. Print.
Factor Analysis	Describes variability among correlated variables with the goal of lowering the number of unobserved variables, namely, the factors.	If you suspect there are inmeasurable influences on your data, then you may want to try factor analysis.	Child, Dennis. The Essentials of Factor Analysis. United Kingdom: Cassell Educational, 1990. Print.
Fast Fourier Transform	Transforms time series from time to frequency domain efficiently. Can also be used for image improvement by spatial transforms.	Filtering a time varying signal can be done more effectively in the frequency domain. Also, noise can often be identified in such signals by observing power at aberrant frequencies.	Mitra, Partha P., and Hemant Bokil. Observed Brain Dynamics. United Kingdom: Oxford University Press, 2008. Print.
Format Conversion	Creates a standard representation of data regardless of source format. For example, extracting raw UTF-8 encoded text from binary file formats such as Microsoft Word or PDFs.	There are a number of open source software packages that support format conversion and can interpret a wide variety of formats. One notable package is Apache Tikia.	Ingersoll, Grant S., Thomas S. Morton, and Andrew L. Farris. Taming Text: How to Find, Organize, and Manipulate It. New Jersey: Manning, 2013. Print.

Figure 19: Detailed Table of Analytics from Booz Allen's Field Guide to Data Science (Continued)

TECHNIQUE	DESCRIPTION	TIPS FROM THE PROS	REFERENCES WE LOVE TO READ
Fuzzy Logic	Logical reasoning that allows for degrees of truth for a statement.	Utilize when categories are not clearly defined. Concepts such as "warm", "cold", and "hot" can mean different things at different temperatures and domains.	Zadeh L.A., "Fuzzy Sets." Information and Control. California: University of California, Berkeley, 1965. Print.
Gaussian Filtering	Acts to remove noise or blur data.	Can be used to remove speckle noise from images.	Parker, James R. Algorithms for Image Processing and Computer Vision. New Jersey: John Wiley & Sons, 2010. Print.
Generalized Linear Models	Expands ordinary linear regression to allow for error distribution that is not normal.	Use if the observed error in your system does not follow the normal distribution.	MacCullagh, P., and John A. Nelder. Generalized Linear Models. Florida: CRC Press, 1989. Print.
Genetic Algorithms	Evolves candidate models over generations by evolutionary inspired operators of mutation and crossover of parameters.	Increasing the generation size adds diversity in considering parameter combinations, but requires more objective function evaluation. Calculating individuals within a generation is strongly parallelizable. Representation of candidate solutions can impact performance.	De Jong, Kenneth A. Evolutionary Computation - A Unified Approach. Massachusetts: MIT Press, 2002. Print.
Grid Search	Systematic search across discrete parameter values for parameter exploration problems.	A grid across the parameters is used to visualize the parameter landscape and assess whether multiple minima are present.	Kolda, Tamara G., Robert M. Lewis, and Virginia Torczon. "Optimization by Direct Search: New Perspectives on Some Classical and Modern Methods." SIAM Review. Society for Industrial and Applied Mathematics, 2003. Print.
Hidden Markov Models	Models sequential data by determining the discrete latent variables, but the observables may be continuous or discrete.	One of the most powerful properties of Hidden Markov Models is their ability to exhibit some degree of invariance to local warping (compression and stretching) of the time axis. However, a significant weakness of the Hidden Markov Model is the way in which it represents the distribution of times for which the system remains in a given state.	Bishop, Christopher M. Pattern Recognition and Machine Learning. New York: Springer, 2006. Print.
Hierarchical Clustering	Connectivity based clustering approach that sequentially builds bigger (agglomerative) or smaller (divisive) clusters in the data.	Provides views of clusters at multiple resolutions of closeness. Algorithms begin to slow for larger datasets due to most implementations exhibiting O(N3) or O(N2) complexity.	Rui Xu, and Don Wunsch. Clustering. New Jersey: Wiley- IEEE Press, 2008. Print.
K-means and X-means Clustering	Centroid based clustering algorithms, where with K means the number of clusters is set and X means the number of clusters is unknown.	Centroid based clustering algorithms, where with K means the number of clusters is set and X means the number of clusters is unknown.	Rui Xu, and Don Wunsch. Clustering. New Jersey: Wiley- IEEE Press, 2008. Print.

Figure 19: Detailed Table of Analytics from Booz Allen's Field Guide to Data Science (Continued)

TECHNIQUE	DESCRIPTION	TIPS FROM THE PROS	REFERENCES WE LOVE TO READ
Linear, Non-linear, and Integer Programming	Set of techniques for minimizing or maximizing a function over a constrained set of input parameters.	Start with linear programs because algorithms for integer and non-linear variables can take much longer to run.	Winston, Wayne L. Operations Research: Applications and Algorithms. Connecticut: Cengage Learning, 2003. Print.
Markov Chain Monte Carlo (MCMC)	A method of sampling typically used in Bayesian models to estimate the joint distribution of parameters given the data.	Problems that are intractable using analytic approaches can become tractable using MCMC, when even considering high-dimensional problems. The tractability is a result of using statistics on the underlying distributions of interest, namely, sampling with Monte Carlo and considering the stochastic sequential process of Markov Chains.	Andrieu, Christophe, Nando de Freitas, Amaud Doucet, and Michael I. Jordan. "An Introduction to MCMC for Machine Learning." Machine Learning. Kluwer Academic Publishers, 2003. Print.
Monte Carlo Methods	Set of computational techniques to generate random numbers.	Particularly useful for numerical integration, solutions of differential equations, computing Bayesian posteriors, and high dimensional multivariate sampling.	Fishman, George S. Monte Carlo: Concepts, Algorithms, and Applications. New York: Springer, 2003. Print.
Naïve Bayes	Predicts classes following Bayes Theorem that states the probability of an outcome given a set of features is based on the probability of features given an outcome.	Assumes that all variables are independent, so it can have issues learning in the context of highly interdependent variables. The model can be learned on a single pass of data using simple counts and therefore is useful in determining whether exploitable patterns exist in large datasets with minimal development time.	Ingersoll, Grant S., Thomas S. Morton, and Andrew L. Farris. Taming Text: How to Find, Organize, and Manipulate It. New Jersey: Manning, 2013. Print.
Neural Networks	Learns salient features in data by adjusting weights between nodes through a learning rule.	Training a neural network takes substantially longer than evaluating new data with an already trained network. Sparser network connectivity can help to segment the input space and improve performance on classification tasks.	Haykin, Simon O. Neural Networks and Learning Machines. New Jersey: Prentice Hall, 2008. Print.
Outlier Removal	Method for identifying and removing noise or artifacts from data.	Be cautious when removing outliers. Sometimes the most interesting behavior of a system is at times when there are aberrant data points.	Maimon, Oded, and Lior Rockach. Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers. The Netherlands: Kluwer Academic Publishers, 2005. Print.
Principal Components Analysis	Enables dimensionality reduction by identifying highly correlated dimensions.	Many large datasets contain correlations between dimensions; therefore part of the dataset is redundant. When analyzing the resulting principal components, rank order them by variance as this is the highest information view of your data. Use scree plots to infer the optimal number of components.	Wallisch, Pascal, Michael E. Lusignan, Marc D. Benayoun, Tanya I. Baker, Adam Seth Dickey, and Nicholas G. Hatsopoulos. Matlab for Neuroscientists. New Jersey: Prentice Hall, 2009. Print.

Figure 19: Detailed Table of Analytics from Booz Allen's Field Guide to Data Science (Continued)

TECHNIQUE	DESCRIPTION	TIPS FROM THE PROS	REFERENCES WE LOVE TO READ
Random Search	Randomly adjust parameters to find a better solution than currently found.	Use as a benchmark for how well a search algorithm is performing. Be careful to use a good random number generator and new seed.	Bergstra J. and Bengio Y. Random Search for Hyper- Parameter Optimization, Journal of Machine Learning Research 13, 2012.
Regression with Shrinkage (Lasso)	A method of variable selection and prediction combined into a possibly biased linear model.	There are different methods to select the lambda parameter. A typical choice is cross validation with MSE as the metric.	Tibshirani, Robert. "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society. Series B (Methodological). Toronto: Royal Statistical Society, 1996. Print.
Sensitivity Analysis	Involves testing individual parameters in an analytic or model and observing the magnitude of the effect.	Insensitive model parameters during an optimization are candidates for being set to constants. This reduces the dimensionality of optimization problems and provides an opportunity for speed up.	Saltelli, A., Marco Ratto, Terry Andres, Francesca Campolongo, Jessica Cariboni, Debora Gatelli, Michaela Saisana, and Stefano Tarantola. Global Sensitivity Analysis: the Primer. New Jersey: John Wiley & Sons, 2008. Print.
Simulated Annealing	Named after a controlled cooling process in metallurgy, and by analogy using a changing temperature or annealing schedule to vary algorithmic convergence.	The standard annealing function allows for initial wide exploration of the parameter space followed by a narrower search. Depending on the search priority the annealing function can be modified to allow for longer explorative search at a high temperature.	Bertsimas, Dimitris, and John Tsitsiklis. "Simulated Annealing." Statistical Science. 1993. Print.
Stepwise Regression	A method of variable selection and prediction. Akaike's information criterion AIC is used as the metric for selection. The resulting predictive model is based uponordinary least squares, or a general linear model with parameter estimation via maximum likelihood.	Caution must be used when considering Stepwise Regression, as over fitting often occurs. To mitigate over fitting try to limit the number of free variables used.	Hocking, R.R. "The Analysis and Selection of Variables in Linear Regression." Biometrics. 1976. Print.
Stochastic Gradient Descent	General-purpose optimization for learning of neural networks, support vector machines, and logistic regression models.	Applied in cases where the objective function is not completely differentiable when using sub-gradients.	Witten, Ian H., Eibe Frank, and Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. Massachusetts: Morgan Kaufmann, 2011. Print.
Support Vector Machines	Projection of feature vectors using a kernel function into a space where classes are more separable.	Try multiple kernels and use k-fold cross validation to validate the choice of the best one.	Hsu, Chih-Wei, Chih-Chung Chang, and Chih-Jen Lin. "A Practical Guide to Support Vector Classification." National Taiwan University Press, 2003. Print.

Figure 19: Detailed Table of Analytics from Booz Allen's Field Guide to Data Science (Continued)

TECHNIQUE	DESCRIPTION	TIPS FROM THE PROS	REFERENCES WE LOVE TO READ
Term Frequency Inverse Document Frequency	A statistic that measures the relative importance of a term from a corpus.	Typically used in text mining. Assuming a corpus of news articles, a term that is very frequent such as "the" will likely appear many times in many documents, having a low value. A term that is infrequent such as a person's last name that appears in a single article will have a higher TD IDF score.	Ingersoll, Grant S., Thomas S. Morton, and Andrew L. Farris. Taming Text: How to Find, Organize, and Manipulate It. New Jersey: Manning, 2013. Print.
Topic Modeling (Latent Dirichlet Allocation)	Identifies latent topics in text by examining word co-occurrence.	Employ part-of-speech tagging to eliminate words other than nouns and verbs. Use raw term counts instead of TF/IDF weighted terms.	Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent Dirichlet Allocation." Journal of Machine Learning Research. 2003. Print.
Tree Based Methods	Models structured as graph trees where branches indicate decisions.	Can be used to systematize a process or act as a classifier.	James, G., D. Witten, T. Hastie, and R. Tibshirani. "Tree Based Methods." An Introduction to Statistical Learning. New York: Springer, 2013. Print.
T-Test	Hypothesis test used to test for differences between two groups.	Make sure you meet the tests assumptions and watch out for Family Wise error when running multiple tests.	Bhattacharyya, Gouri K., andRichard A. Johnson. Statistical Concepts and Models. Wiley, 1977. Print.
Wrapper Methods	Feature set reduction method that utilizes performance of a set of features on a model, as a measure of the feature set's performance. Can help identify combinations of features in models that achieve high performance.	Utilize k-fold cross validation to control over fitting.	John, George H., Ron Kohavi, and Karl Pfleger. "Irrelevant Features and the Subset Selection Problem." Proceedings of ICML-94, 11th International Converence on Machine Learning. New Brunswick, New Jersey. 1994. Conference Presentation.

To learn more about how we can help you grow data science within your organization, visit boozallen.com/datascience.

Ezmeralda Khalil Sager
Vice President
sager_ezmeralda@bah.com

Kirk Borne
Principal Data Scientist and Executive Advisor
borne_kirk@bah.com

Shelly Davis
Principal/Director
davis_shelly@bah.com

Ray Hensberger
Distinguished Technologist
hensberger_ray@bah.com

Patrick McCreesh
Principal/Director
mccreesh_patrick@bah.com

Stephanie Beben
Chief Scientist
beben_stephanie@bah.com

Doug Gartner
Chief Technologist
gartner_douglas@bah.com

Graham Gilmer
Chief Technologist
gilmer_graham@bah.com

Jesus Jackson
Chief Scientist
jackson_jesus@bah.com

Eric Syphard
Chief Scientist
syphard_eric@bah.com

Alex Decknick
Lead Associate
decknick_alexandra@bah.com

Kate Helfet
Lead Associate
helfet_katherine@bah.com

Mark Kokoska
Senior Lead Scientist
kokoska_mark@bah.com

Susan Michener Johnston
Lead Associate
johnston_susan@bah.com

Aoibheann Thinnes
Associate
thinnes_aoibheann@bah.com

Sarah Toigo
Associate
toigo_sarah@bah.com

About Booz Allen

Booz Allen Hamilton has been at the forefront of strategy and technology for more than 100 years. Today, the firm provides management and technology consulting and engineering services to leading *Fortune* 500 corporations, governments, and not-for-profits across the globe. Booz Allen partners with public and private sector clients to solve their most difficult challenges through a combination of consulting, analytics, mission operations, technology, systems delivery, cybersecurity, engineering, and innovation expertise.

With international headquarters in McLean, Virginia, the firm employs more than 22,600 people globally and had revenue of $5.41 billion for the 12 months ended March 31, 2016. To learn more, visit BoozAllen.com. (NYSE: BAH)