June 2020

Perspective
EXPERT INSIGHTS ON A TIMELY POLICY ISSUE

AARON C. DAVENPORT, MICHELLE D. ZIEGLER, ABBIE TINGSTAD, KATHERINE ANANIA, DANIEL ISH, NIDHI KALRA, SCOTT SAVITZ, RACHEL LIANG, MELISSA BAUMAN

Decoding Data Science

The U.S. Coast Guard's Evolving Needs and Their Implications

Like many large organizations, the Coast Guard has vast amounts of data that it could use to identify, predict, and solve pressing challenges. Data science could be valuable to the Coast Guard in a variety of domains, such as forecasting the resources needed for future trends in search-and-rescue (SAR) missions, further automating aids to navigation, or automating fishery observations. In personnel areas, data science could help improve billet assignments, determine where to focus recruiting efforts, and boost employee retention.

The Coast Guard has an opportunity to plot the path to determine service-specific uses, identify the strategy and driving mechanisms, and begin laying out a plan. This Perspective outlines the role that data science can play in decisionmaking processes and provides a selected set of key questions and sensitivities for the Coast Guard to consider in developing its future usage of data science.

Data science uses elements from many disparate fields and includes such methodologies and concepts as artificial intelligence (AI), big data, and machine learning (ML). *Data science* refers to the processes through which data are collected,

HSOAC
HOMELAND SECURITY OPERATIONAL ANALYSIS CENTER
An FFRDC operated by the RAND Corporation under contract with DHS

manipulated, analyzed, and understood. Data science takes huge, complex, and dynamic arrays of data and distills them into patterns and trends, enabling organizations like the Coast Guard to glean valuable insights into personnel, processes, and procedures.

To harness the power of data science, an organization needs the right data, the right people, and the right culture willing to understand the new information and insights and factor them into decisionmaking processes. Employing data science is not without costs or risks, so it is critical that an organization develops its plans and milestones to align to its strategies, priorities, and goals. Small-scale, proof-of-concept work using a variety of data science techniques and currently available data are underway in various areas of the Coast Guard, the U.S. Department of Defense, other government agencies, and the private sector. The broad employment of data science applications across an organization is evolving and maturing but is not so new as to require navigating uncharted waters.

Data Science Explained in Brief

Data science is an umbrella term covering an expansive field of approaches to and techniques for collecting, manipulating, analyzing, and understanding data. Something akin to this science has existed since humans developed the need for data and the ability to track them for various applications. Both the presence of big data and the ability to manipulate them relies on computers and other technological advances (e.g., the internet) that they have helped enable. Whereas *big data* once referred to data sets that were too large to be housed on a single laptop, it now includes large and complex data sets, often with unstructured data and mixed media, such as video, images, numbers, and raw text. Although some of the mathematics and basic mechanisms behind data science have existed for decades, this field has recently gained increasing public, private, and government attention. The proliferation of high-density networks (social and otherwise) have generated dramatically more data, increasing the availability and proliferation of big data. Computing power has increased to the point at which complex algorithms to analyze and infer meaning from data can be more widely employed. The final piece is the use of the insights and information derived from the data analytics to support decisionmaking.

Project Evergreen V

The topic of this Perspective is the result of a series of Project Evergreen strategic foresight subject-matter expert workshops (called Pinecones) that were conducted by the U.S. Coast Guard Office of Emerging Policy (DCO-X), from September 2019 to March 2020.

Project Evergreen employs scenario-based planning to identify strategic needs for incoming service chiefs, with the goal of supporting executive leaders in their roles as the Coast Guard's strategic decisionmakers.

For specific themes and implications related to the individual workshops and applications to data science, see the appendix.

Within data science are two principal categories:

- *data management:* defining, collecting, storing, and presenting measurements of relevant characteristics. This can be thought of as capturing and cultivating data.
- *data analysis or analytics:* systematically using qualitative or quantitative methods (or a combination of these) to describe and evaluate patterns within data ("Data Analysis," undated).

Both categories require the development and deployment of policy, practices, and infrastructure for using data science and data analysis and analytics to support decisionmaking and for addressing the challenges that their use might present (e.g., protecting privacy and preventing hacking).

One example of the data value chain was developed by IBM. Figure 1 illustrates the nine levers that IBM identified as having significant influence on an organization's ability to gain value from data and analytics.

These nine levers fall into three categories, as described in Table 1 (Balboni et al., 2013).

Taking the time to work through mechanisms to integrate data analysis across the organization, ways to support the development and implementation of strategy, what technology will be required, and how data science

FIGURE 1

Nine Levers That Influence the Ability to Derive Value from Data and Analytics

SOURCE: Balboni et al., 2013, p. 4.

TABLE 1
Categories of the Nine Levers That Affect an Organization's Ability to Derive Value from Data

Mechanism Category	Function	Example
Enabling	Create the foundation for data analytics and big data use within organizations.[a] By aligning the data strategy with the enterprise strategy, leadership establishes the organizational direction for analytics.	Metric or platform. For example, to promote the strategic goals of equality and diversity, the Coast Guard might want to analyze racial or gender bias in its performance review system; an enabling mechanism would be modification of the existing database of performance review data to allow statistical analysis of differences in evaluations.
Driving	Push an organization from analytic discovery to value creation.	Culture, data, trust, or strong governance and security so that data will be valued as a decisionmaking tool. An example of a driving mechanism for the Coast Guard would be developing a billet assignment optimization model that ensures that billets are filled with the correct skills and informs service members which assignments are most likely to support their desired career trajectories.
Amplifying	Generate the energy and capabilities necessary to translate the results of data analysis into actions that positively affect the organization.	Expertise or funding. An example of a Coast Guard amplifying mechanism would be decisionmakers applying data analytic investments across all missions and requirements.

[a] Balboni et al., 2013.

can help the Coast Guard execute its missions will provide senior leadership with tools to integrate data analytics into its culture and processes.

Data Science Terminology

Some terms, such as *AI*, *ML*, and *big data*, are often used interchangeably because clear definitions were lacking as data science techniques quickly evolved and became more mainstream. Table 2 summarizes some of the key terms associated with data science, although some definitions lack consensus and the relationship between terms is not always clearly delineated. Each term is a field of study in and of itself, with entire careers focused on expertise in just one area. However, in this Perspective, we generally discuss data science as a whole to avoid prescribing a particular path or solution, given that it is still unknown which analytic tools and aspects of data science the Coast Guard will employ. "A Deeper Dive: Types and Examples of Data Analysis," later in this Perspective, ventures into a few of these niche areas to highlight examples, possibilities, and challenges. Figure 2 illustrates the relationships between some of these concepts.

Big data and ML have delivered impressive results in a wide variety of use cases, including financial models that help investors make better decisions, image analysis that monitors the effects of climate change, and natural-language processing that turns spoken words into

TABLE 2
Definitions for Some Key Data Science–Related Terms

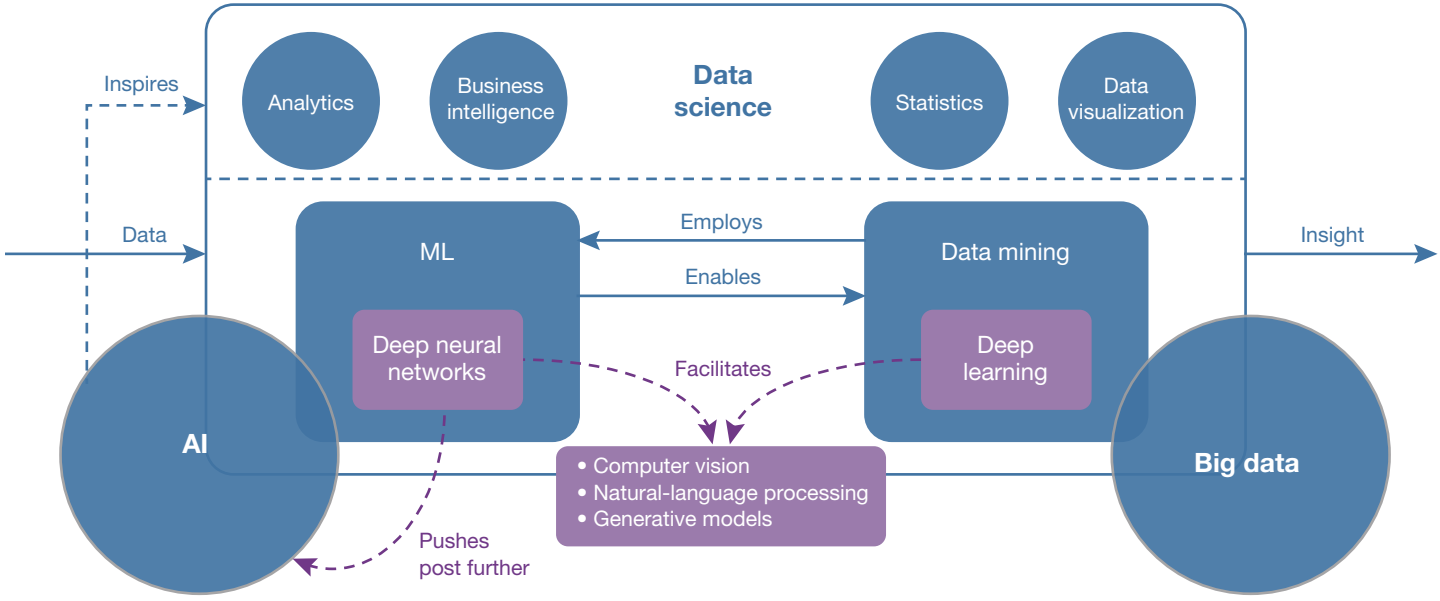
Term	Definition
Data science (Provost and Fawcett, 2013)	Umbrella term for a field of study and practice centered on a broad set of approaches and techniques that enable information and knowledge to be derived from data sets
Data analysis ("Data Analysis," undated)	Act of systematically employing qualitative or quantitative approaches (or a combination of these) to describe and evaluate patterns within data
Data conditioning, curation, and normalization (Boyer et al., 2015)	Steps taken to "clean" data sets so that data analytic processes (such as mining) can be more easily utilized
Data mining (Provost and Fawcett, 2013)	Using technology to extract information and knowledge from data
Big data (Gandomi and Haider, 2015)	Largely unstructured data existing in such formats as text, audio, imagery, and video that can be collectively characterized by such attributes as *volume*, *variety*, *velocity*, *veracity*, *variability* (and *complexity*), and *value*
AI ("artificial intelligence," 2020)	The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decisionmaking, and translation between languages
ML ("machine learning," 2020)	The capacity of a computer to learn from experience (i.e., to modify its processing on the basis of newly acquired information)

text. These techniques require a significant amount of data (and the ability to collect, store, process, and prepare those data for analysis), as well as human input to define the problem and integrate the results into a decisionmaking framework. People who are skilled in coding and data analysis are also needed to develop, train, and provide assurance that analyses and automated processes are working correctly. As a result, these techniques should be thought of as tools that perform a narrow set of tasks to inform decisionmakers or execute predefined priorities, not as a substitute for human input into the decisionmaking process.

Data Science and the U.S. Coast Guard

The field of data science is broad and provides many opportunities—and challenges—for the Coast Guard to streamline manual systems, track patterns across the organization, and better plan for the future by using data-driven analysis. Maximizing the benefits of data science will require three main steps, described in the rest of this section.

FIGURE 2
Relationships Between Key Data Science Concepts



SOURCE: Adapted from Mayo, 2016.

Decide How the Coast Guard Should Use Data Science

Data science systems are most useful when they have been thoughtfully designed to support an organization's needs and long-term goals. Data science is not a silver bullet, but it can be a powerful tool for supporting decisionmakers. Before investing heavily, the Coast Guard should define its long-term vision and goals for using data science in the context of its responsibilities and long-term strategies. Some considerations that the Coast Guard should assess when designing a data science strategy include the desired level and requirements for data gathering and storage, the trade-offs that come with using increasingly complex data-driven analytic tools, the development of in-house data science managers, and the ways in which data science can complement and support policy- and decisionmaking.

To integrate data science into long-term plans and daily operations, the Coast Guard needs to gain an appreciation for how data science can and should be best used within the organization and its potential implications

on future operations. Important considerations in Coast Guard plan development include

- *strategy:* establishing a strategic agenda for data science and analytic development, aligned to Coast Guard strategy and goals. This agenda should be incorporated into the funding process by prioritizing projects that speak to those goals.
- *technology:* making decisions related to information technology (IT) infrastructure and expertise based on the service's future needs. A pool of individuals with strong analytical talent or knowledge of day-to-day operations (or both) should be consulted to provide valuable, organization-specific insights while designing Coast Guard–specific systems.
- *organizational culture:* creating an organizational focus on ensuring that the data gathered and analyzed provide specific, timely answers that aid decisionmaking. Leaders should use analytics in decisionmaking to promote a fact-based culture.

Determine the Data Needed and How to Manage Them

The Coast Guard has several needs to consider in expanding its use of data science for any purpose.

Managing Data

Data management is the baseline requirement for all data science systems. To implement any sort of data science system, data are needed. This may seem obvious, but developing useful and effective systematic processes for collecting, storing, protecting, and cleaning data can be complex and resource-intensive. A new data-gathering initiative often requires time for training across multiple groups to ensure that consistent and high-quality data are gathered.

The challenges are not insurmountable; indeed, the capacity to gather data is likely critically required for the future. Developing these systems requires leadership vision and enabling mechanisms to set the tone for investment in capturing data and doing so in ways that align with the Coast Guard's long-term strategies. Understanding how the data will be used is critical to ensuring the right balance of factors to get "quality data," as illustrated in Figure 3. It is important to note that data-management tasks are not discrete stovepipes; instead, there are feedback loops between them and within the policy and decisionmaking structures. It is critical to approach data management not as a stagnant or permanent state but as an iterative development process based on new information and dynamic input from leadership on priorities and goals.

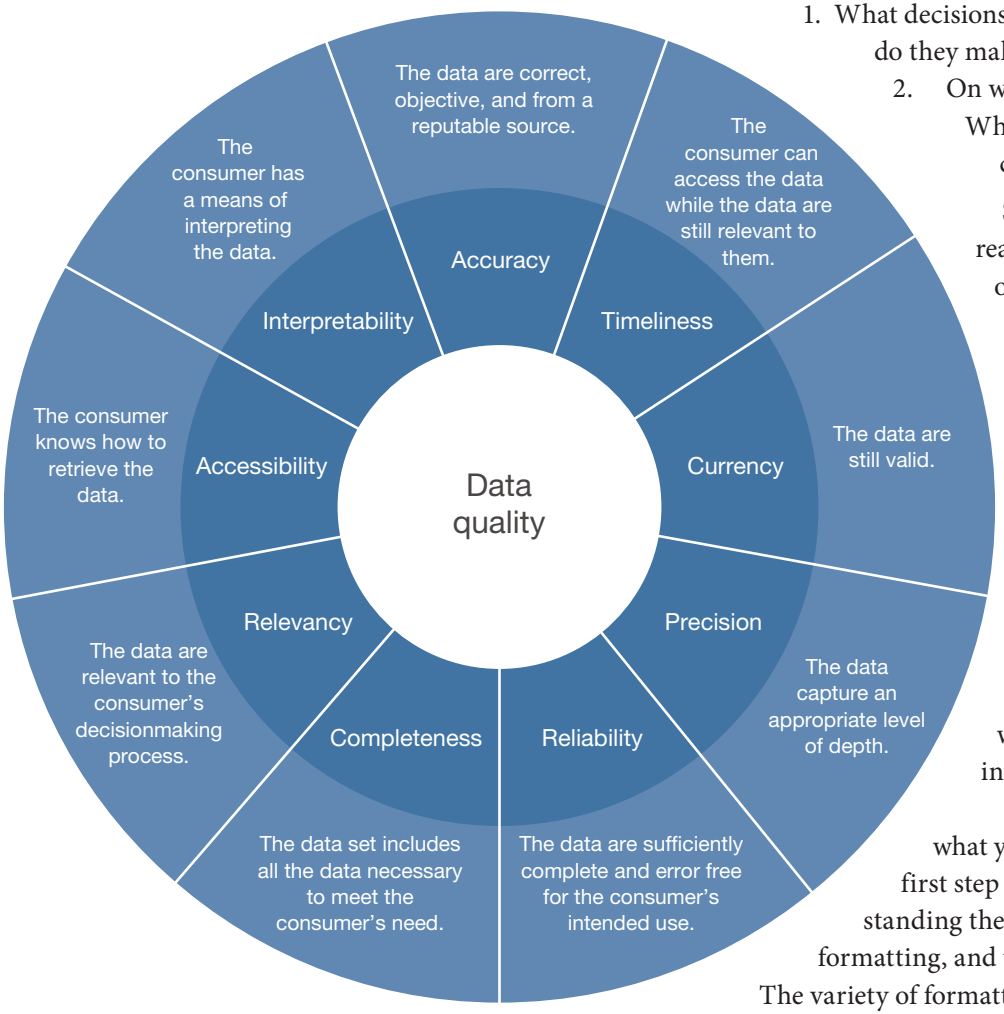
Setting up a system for data gathering should be considered a significant up-front investment in both time and resources. But if planned for and aligned to strategic plans and resources, the data-gathering step could no longer be a challenge once effective systems and processes are in place, and the benefits of analysis should build over time to support decisionmaking and planning.

Defining and Collecting Data

Defining which set of attributes and metrics to record is critical for data to be useful for decisionmakers. The task begins with an inventory of the decisionmaking process as

FIGURE 3

Factors in Collecting Quality Data



SOURCE: Wang and Strong, 1996.

it stands. For the purposes of designing a data-collection procedure, one needs to answer two sets of questions:

1. What decisions are made? Who makes them? When do they make them?
2. On what criteria are the decisions based? What attributes are germane to these criteria?

Some of these attributes may be readily defined based on simple facts or easily observable outcomes. For example, in weather forecasting, the temperature over time at a specific location is clearly defined even before the forecaster sits down to design a data collection. In the context of personnel management, such factors as the timing or length of an assignment, the length of a service member's tenure and in what capacity, and whether any accidents occurred while someone was on duty likely fall into this category.

With the adage "You can't improve what you can't measure" in mind, a crucial first step would be identifying and understanding the types of data currently available, the formatting, and the quality within the data systems. The variety of formatting and storage methods of information from evaluations, skill assessments, and scores will likely make for a complex big data set.

The set of attributes and metrics to collect defines the scope of the questions that can be addressed with the data set. At the same time, practical and legal restrictions might limit that scope. Collection of personnel information is especially likely to be subject to legal restrictions, and the design of the data-collection process will likely need to be undertaken in concert with legal counsel.

Storing and Presenting Data

Considerable investment is necessary to safely store and agilely present the data that have been collected. Though many of the most dramatic uses of data come from analysis, simply having the data available to relevant stakeholders in a digestible form when needed can provide vital situational awareness for decisionmaking.

Protecting the data and the authorities to access those data can present challenges for integration and use. Classified, law enforcement sensitive, and personally identifiable information are just three types of access control that need to be identified and planned for in the management and analysis process. For example, personnel data can be particularly sensitive and represent an especially attractive target for malicious actors as a result. The 2015 cyberintrusion into the Office of Personnel Management (Zengerle and Cassella, 2015) is a dramatic demonstration of the value of U.S. government personnel data to well-resourced, sophisticated malicious actors. Regardless of how the data are stored, it will take effort, resources, policies, and practices to protect against intrusion and unauthorized access.

Designing the process and infrastructure that present the data to decisionmakers begins with answering the first set of questions above. Successful utilization of the data will require investments in infrastructure to deliver the data, training to ensure that the data are understood, and policies to prevent abuse of the data.

Develop Data Analysis Systems with Several Considerations in Mind

Although popular articles promote the seemingly boundless benefits of data science (especially ML and AI), there are many issues to take into account while developing a data science–based strategy for analysis.

In-House Talent

A consideration for developing a robust data analysis system is having the data science talent in-house. A cohort of individuals who understand and can facilitate machine-aided decisionmaking is needed to ensure that the algorithms are working correctly and the results are correctly interpreted. The algorithms' design can be outsourced, but internal capabilities in coding and math can be useful in designing the actual tools and applications that end users need. While each federal agency stands up an office of the chief data officer and determines the office's roles and responsibilities, this timely concern can be factored into that planning process.

Transparency and Explainability

More-complex ML techniques can be risky in several ways that could compete or conflict with some ideals, such as transparency. ML algorithms are developed by the machine itself and can be a "black box," obscuring from the user a full understanding of what algorithms are

being used and why. The answers derived often cannot be clearly traced to the algorithmic decisionmaking process, which reduces transparency. Because ML relies on existing data for predictive actions, it may unearth and propagate unconscious bias originating in the provided data sets. This poses significant risks when applied to certain actions, such as hiring and promotions (Zengerle and Cassella, 2015; DeBrusk, 2018). Risk increases with anything linked to networked computers; the trade-off between having a computer-assisted broad view of data also means that decisions guided by machines are vulnerable to cyberintrusions, such as denial of service or insertion of deceptive information. The issues are not insurmountable but require forethought, study, and planning.

A related issue in data analysis is how well the result of a particular algorithm can be explained to a decisionmaker. In applications that do not directly interact with humans, such as predictive load balancing in computer network routing, explainability might not be very important to the user. However, explainability is important when using data analysis algorithms to support decisionmakers, particularly when the decisions directly affect other people. If decisionmakers cannot offer any rationale for the algorithm-driven parts of the decision beyond broad confidence in the algorithm or an explanation of how the algorithm works in general, it may impede buy-in to the decision itself and the decisionmaking structure more broadly.

For decision tasks in general, it is exceedingly unlikely that a computer can automatically analyze every consideration going into the decision. It will eventually fall to a person to integrate information that may span qualitative information that cannot be digested by a computer, computer-optimized solutions, and computer-generated predictions in order to arrive at a final decision. To weigh these different types of information against one another, a decisionmaker must be able to understand what information the computer used and how it arrived at its prediction or solution, as well as the inherent computation limitations.

Culture

Decisions aided by machines are still subject to the policies that ultimately shape the decision landscape. Although data-driven decisionmaking can help an organization better adhere to these policies—or highlight their positive and negative impacts—the algorithms themselves cannot mitigate misaligned policies. Having a culture that values the results and insights suggested by data-driven decisionmaking and is prepared to consider them is key to reaping the benefits of data science. Often, policy change lags behind technological change, and those lagging policies will need to be identified early to pave the way for future development. Thus, it is important to cultivate a culture in which people are willing to consider technology-driven change and evolve policies accordingly.

Parallel Processes

To implement the expanded use of data science, the Coast Guard will first need to consider the long-term strategic goals for how it wants to use data and identify the gaps and hurdles that will be budget and policy drivers, especially with regard to its overarching data-collection and management structure. Long-term changes are not made in a vacuum, so even incremental advancements and changes must support current needs and workflows in the

interim while any prototype or new applications designed to improve workflow or decisionmaking are being developed and implemented. Understanding this parallel development process will be beneficial for smoothing the proverbial bumps in the road that come with significant process and system changes. Additionally, the Coast Guard will need to review policies to inform the scope of change possible using data science techniques. The Coast Guard should consider how to adapt such functions as acquisition, protection, and sustainment to meet needs in the digital environment as it increases its leverage of data science.

A Deeper Dive: Types and Examples of Data Analysis

A variety of analytical techniques fall under the "data analytics" heading. The methodology used to perform data analysis depends on the question being asked, the types and amounts of data available, computing and network capacity, and the desired level of accuracy. The list of techniques and examples in this section is not exhaustive but provides a sense of the breadth of the field and some idea of the process of deploying these techniques.

Supervised ML is a technique likely to be deployed to support day-to-day personnel decisionmaking. The term refers to a set of techniques that searches for a good approximation of the statistical relationship between inputs and outputs. Supervised ML is well adapted to certain tasks, such as predicting the future price of an asset based on historical price data, identifying the presence or absence of elements in an image, or predicting the next word in a sentence based on the preceding words. Abstractly, the steps to perform this technique are

1. *Select a model.* A model is a tunable representation of the relationship between inputs and outputs. When a model is selected, some parameters can be adjusted to the choice that best represents the data. Generally speaking, selecting a model is based on some partial knowledge about likely relationships in the data.
2. *Train.* A subset of the data, called training data, is used to tune the parameters of the model. The exact details of this process vary from model to model and can be adjusted to the particular case at hand, but, broadly speaking, the model is tuned with training data so that it produces an output as close as possible to the true (known) output.
3. *Validate.* The remaining data are then used to test the accuracy of the machine-learned map from inputs to outputs. This step is necessary because the model is likely to be more accurate on the training data than on the rest of the data.
4. *Deploy.* The model is used to predict likely outputs on real-world inputs.
5. *Monitor and update.* The model is updated as additional data are gathered.

In the real world, work does not proceed uniformly along these steps. Steps 3 and 4 frequently produce evidence that requires returning to step 1 to reengineer the model or returning to step 2 with more training data. Similarly, the simple descriptions of steps 1 and 2 elide the engineering effort or level of domain-specific expertise that may be necessary to select an appropriate model and

identify the amounts of data that might be necessary to train it. In general, the more complex the inputs, outputs, and their relationship, the more data and time will be necessary to successfully deploy a working model.

The classes of models available vary significantly. The framework outlined here focuses primarily on tunability, but other models may have more functionality (e.g., statistical measures of a prediction's uncertainty or tests for overall model quality). Additional functionality can offer decisionmakers more information, but the need for these functions should be weighed against predictive power.

Use Cases for Machine Learning

The following use cases attempt to illustrate some predictive ML applications and capabilities. Although this set of examples is not exhaustive, it gives a sense of the degree to which solutions based on supervised ML must be engineered to the specific problem they are attempting to address. Both the amount of data required and the design details of the state-of-the-art models vary from application to application.

Example 1: Image Classification

In some use cases, a large volume of digital image data needs to be analyzed. For example, the maker of an autonomous vehicle may need an analysis to help the vehicle's optical sensors automatically recognize a stop sign in its video feed, so that the vehicle comes to a halt. Or a smartphone company might want to use the front-facing camera to automatically unlock the phone when the user's face is detected. In these cases, the typical approach is to use supervised ML to automatically build a function that assigns the presence or absence of the target feature in the image.

As another example, image classification was one of the tasks in the ImageNet challenge (Russakovsky et al., 2015), a competition for image-labeling algorithms that ran from 2010 through 2017. The task required teams to build a model that predicts which of 1,000 possible objects are present in an image. To do this, the teams were presented with a training set of 1.2 million hand-labeled images. The winner of the competition achieved an accuracy rate of 97.7 percent (ImageNet, 2017).

As detailed by the challenge organizers (Russakovsky et al., 2015), assembling such a massive data set of manually labeled images was a significant feat in itself. Such a large data set was necessary because the task required of the final algorithm was so complex. A modern smartphone camera produces images bigger than 10 megapixels. This means that on the order of 33,554,432 variables characterize a full-color digital image. Combined with the subtlety and general nature of the task, the sheer size of the space of possible images means that a significant amount of data is required to reach these levels of accuracy. These considerations also drive the typical choice of model type and neural networks, which can automatically learn important predictors in large, complicated inputs. Even within this class of model, however, some designs of neural networks are more successful than others at making use of image data in particular. Careful engineering and computer vision–specific expertise are required to build a state-of-the-art model.

A possible Coast Guard application of this type of modeling is the 2017 "N+1 fish, N+2 fish" challenge (DrivenData, undated), offered by DrivenData and hosted

by the Gulf of Maine Research Institute and the Nature Conservancy to automate fishery observation data. For this challenge, video cameras were placed on board participating commercial fishing vessels, and algorithms were developed to automatically analyze the video images for the sequence of fish, fish species, counts, and measurements (DrivenData, undated).

Example 2: Time-Series Prediction

Time-series prediction is a commonly used supervised ML model that utilizes past data to predict future performance. Weather, stock market performance, and even traffic patterns are based on time-series prediction models. Most people use one or more time-series prediction models in their daily lives, and the concept of employing historical data to predict future outcomes is fairly straightforward.

Time-series data typically consist of only a handful of quantities tracked over time that have a relatively simple relationship with the output for which a predictor is desired. Consequently, one can frequently turn to older statistical models that are easier to understand and explain. These models also typically have better-developed tools for quantifying the uncertainty in their predictions. Similarly, the amount of data needed for good predictions is usually more modest. This is not universally true, however: Some data may have a complicated or obscure enough relationship between input and output to demand a more modern, opaque model or a larger amount of data.

A Coast Guard application of time-series prediction modeling is being developed by the Homeland Security Operational Analysis Center (HSOAC) to estimate future resourcing needs for SAR cases based on past SAR person-hour and equipment requirements. This and other data science efforts currently being pursued by the Coast Guard are indicative of interest and creativity in determining its usefulness for the service.

Example 3: Natural-Language Processing

Natural-language processing, defined broadly as the task of teaching computers to identify the meaning of speech or text, is a large and vibrant field. This is perhaps unsurprising, given the obvious utility of a computer that can interact productively with the way humans natively represent information. Natural-language processing has already produced a variety of compelling products, including speech-to-text, translation, and next-word prediction algorithms.

Text is a challenging input for ML algorithms, so a great deal of data is typically required. Although labeled data sets might not be large enough to accomplish the desired task, a wealth of unlabeled data is available as text-based content on the internet and in books. Consequently, many state-of-the-art methods for natural-language tasks use a methodology known as *transfer learning*. In transfer learning, a model is first trained on a task for which ample data are available, and then pieces of the trained model are used as building blocks for a model to be trained on a related task for which much less data are available.

A Coast Guard application of natural-language processing models could be developing a tool that automatically and systemically combs social media channels to confirm that a person reported missing is, in fact, missing and not currently posting on social media.

Although supervised ML is a natural fit for day-to-day decisionmaking, related techniques, such as unsupervised ML or reinforcement learning, could conceivably play a role in longer–time horizon analyses. Briefly, *unsupervised*

ML refers to a set of techniques designed to find compact representations of the data (or, equivalently, patterns in the data) rather than predicting likely values of a particular output. Unsupervised ML techniques can be useful for exploring a data set for unanticipated phenomena. *Reinforcement learning* refers to techniques that enable a virtual agent to optimally control its environment by learning only from data in its past interactions. Reinforcement learning is perhaps most famous for its use in designing AI that can outperform humans at competitive games, such as DeepMind's AlphaGo for the game of Go (Silver and Hassabis, 2017).

The preceding excursion into supervised ML and three examples expounding on methods within this field provide some insights and ideas on how this technique has been and could be applied.

Conclusion

Data science has the potential to provide valuable insight and capability to the U.S. Coast Guard, if the systems are carefully designed and managed. Applications can help decisionmakers make data-informed choices and improve the overall effectiveness of the service. Data science is also resource-intensive, requiring a significant number of inputs, including proper data-management systems and appropriately trained staffing to ensure that models are designed, explained, and applied appropriately. Data science should be used as a tool to support decisionmakers rather than considered a "catch-all" solution to managing the Coast Guard. That tool, being akin to a Swiss Army knife of options, has a wide variety of topic and mission areas of potential applications in the Coast Guard, from human resource (HR) management (HRM) to fishery management and enforcement processes.

Some potential next steps to evaluate the utility of data science applications for the Coast Guard include

- *Perform benchmarking.* Researching best practices and success stories of other governmental and private organizations that promoted data cultures can reveal concrete steps that the Coast Guard can take to strengthen its data management and analysis.
- *Support leaders looking for ways to make data-informed decisions.* Coast Guard decisionmakers can use advanced data analysis to promote a culture of data-driven choices. Data-based knowledge can be considered a resource that supports the organization as a whole.
- *Make the invisible visible.* Data analysis can answer questions and provide insight into patterns that individuals cannot detect without computer-supported data analysis. Predictive tools could help project future needs and implications of decisions and policies, all of which could serve to better inform leadership decisionmaking processes.

Understanding and planning for the investments required in such areas as IT infrastructure and talent needs and how each area are affected by long-term strategic plans and goals is critical to starting down the most efficient and effective path to harnessing the potential of data science. As William Cameron said, "Not everything that counts can be counted, and not everything that can be counted counts" (Cameron, 1963, p. 13). *The most important factor in implementing data science in an organization is clearly defining the organization's strategy and goals for using and promoting data science.*

Appendix A. Evergreen V Pinecone Results

To support executive service leaders in their roles as the Coast Guard's key strategic decisionmakers, Evergreen V is conducting several foresight engagements, called Pinecones, on key topics over a two-year period. After each Pinecone, the Evergreen team produces reports and other focal material to identify service implications and strategic choices for the 27th Commandant of the Coast Guard's leadership team. The first Pinecone was in September 2019, with the topic "Workforce 2040." The second Pinecone was part of the Maritime Risk Symposium, held in November 2019. The topic of the second Pinecone was "Future Risks to the Marine Transportation System." In both workshops, themes related to data science technology emerged as areas that warranted further analysis; this appendix is a discussion of the implications and applications of big data in these two topic areas.

Workforce 2040

Like most employers, the U.S. Coast Guard will face challenges recruiting qualified candidates because of emerging technologies and the gig economy's influence on employment trends. Participants in the Coast Guard's Workforce 2040 workshop determined several ways in which technology could also help the Coast Guard recruit and maintain its future workforce if the technology is used efficiently and systems are designed thoughtfully.

Participants explored future recruiting challenges and discussed ways to mitigate employment trends that adversely affect the Coast Guard, including better information-sharing practices and applying new technology in workforce management decisions. Data science techniques—from ML and predictive analytics to data conditioning and basic task automation—were recognized as having great potential impact on the Coast Guard and its personnel management systems. During the workshop, participants identified the importance of making investments today to establish and enable a data culture. Key points include (Office of Emerging Policy, 2019)

- making the invisible visible by using data to gain insight
- supporting leaders looking for ways to leverage data analytics and ML to make data-informed decisions
- creating and empowering a data workforce to gain a competitive advantage.

Applying Data Science to Coast Guard Human Resource Management

For several decades, data science has been effectively used in personnel and talent management (see, e.g., Fitz-enz, 1984), suggesting that the Coast Guard could expand its use of broadly defined data science techniques for HRM functions without pushing accepted boundaries or navigating uncharted waters. The growth of data availability and the advancements in data science technology have transformed business strategies and enabled organizations to bring together data from a variety of disparate sources. They can now go beyond standard administrative insights about their employees, applying ML techniques to foster the development of necessary skills and talents organizationwide (Chui et al., 2018). The Coast Guard can

TABLE A.1

Some Data Science Techniques and Applications That the Coast Guard Could Use for Human Resource Management

		Data Analysis	
Example HRM Activity	**Data Management**	**Less Complex**	**More Complex**
Developing strategic plans	Training future scenarios with data	Projecting force structure based on assumptions	Predicting future shortfalls in key personnel
Creating and filling positions	Compiling billet type, filled status, and location	Tracking billets and flagging unfilled billets	Reporting fleetwide trends in unfilled billets over time and connecting to future personnel planning
Recruiting candidates	Digitizing data collected by recruiters	Mapping the most-fruitful recruiting locations	Developing a recruitment portfolio
Managing compensation and benefits	Tracking organizational compensation and benefit costs	Projecting future compensation and benefit costs given force structure planning	Predicting future compensation and benefit costs based on external and internal market factors
Managing payroll and time	Recording hours worked and associated payroll requirements	Analyzing trends in overtime hours	Optimizing staffing size based on overtime projections
Enabling training and development	Documenting workflows	Analyzing feedback from training programs	Tracking training types to operational and career (promotion, tenure in service) outcomes
Managing talent, performance, and succession	Leveraging more-detailed experience identifiers in electronic personnel files	Determining how frequently personnel use special experience in future assignments	Optimizing personnel matches based on experience
Engaging employees	Developing and administering surveys	Analyzing survey trends and takeaways	Triggering adaptive plans
Retaining specific employees	Tabulating reasons for personnel departures	Associating departure reasons and timing with career fields	Calculating retention incentives needed to retain personnel

SOURCES: Labelle and Dyer, 1992; UK Civil Service Human Resources, 2018; U.S. Coast Guard, 2019.

investigate data science applications to improve decision-making with respect to such issues as

- determining workforce requirements through work measurement (see, e.g., Chen, 2019)
- recruiting, including how and from where to attract talent
- making assignments that better optimize individual skills, experience, and preferences with operational needs
- improving retention, such as predicting the effects of new policies or highlighting cohorts that might be at risk for separation, health, or other issues.

Many data science techniques and applications could be leveraged (further) for Coast Guard personnel management. Table A.1 outlines some example applications. The first column gives examples of HRM activities compiled and adapted from academic literature and discussions at the 2019 Workforce 2040 workshop. The next three columns summarize types of data science–related applications at the levels of *data management* and *less-complex and more-complex data analysis*. At its most fundamental structure, higher levels of data analysis must be built on a foundation of data management (see, e.g., Frické, 2009). We use these levels to demonstrate the varying levels of complexity to which data science can be applied.

Broadly, *data management*–related applications enable data to be collected and prepared for analysis. Fundamentally, these applications include data collection and conditioning techniques that prepare data for analysis. The applications categorized under *less-complex data analysis* prepare data for consideration by a decisionmaker or automate (entirely or parts of) basic HR tasks.[A] These amount to basic data analysis and coding. Finally, *more-complex data analysis* applications support decisionmaking by teaming with or replacing humans in the most-complex cognitive processes. These applications leverage advanced big data analytics and AI.

As discussed in the main body of this document, the Coast Guard must make some large-scale decisions about its plan for and use of data science before these applications can be realized. Nevertheless, there are significant opportunities for the Coast Guard to apply data science techniques to its HRM systems.

Autonomous Systems and the Maritime Transportation System

The Coast Guard faces key opportunities and challenges with the advance of autonomous systems in the maritime domain. Several recurring themes emerged from the 2019 Maritime Risk Symposium workshop—with implications for the future of the Marine Transportation System (MTS)—that are directly related to autonomous systems (Savitz, Davenport, and Ziegler, 2020):

- *differential paces of technological adoption:* Advances in autonomous systems that operate in the maritime domain can put less advanced response and regulatory agencies at a disadvantage in the timely completion of planning, regulatory, legal, and operational response to this technology.
- *workforce competency:* A major concern is that personnel must be capable of handling both advanced

[A] This can include such tasks as filling out forms, transferring data, and disseminating information to constituents.

and legacy technologies, in addition to partially autonomous systems. This affects the ability to recruit and retain personnel with those technological skill sets that are in increasingly high demand.
- *uncertainty about capacity demands throughout the MTS:* It is unknown how the industry will adapt and whether government agencies will be able to keep pace with the speed of technology as cargo ships, work boats, and port facilities become more automated.
- *governance challenges for governmental agencies:* As nations aim to address technological, economic, and environmental changes, they need to do so without imposing conflicting policies that either hinder MTS activities or pose new unintended and unanticipated risks. Developing new marine industry regulations and standards requires a deeper technical and operational understanding of the implications of an MTS that operates with greater autonomy.
- *increased Arctic activity:* As the Arctic becomes more accessible, increased maritime activity there could present daunting challenges and opportunities for the Coast Guard, even as land infrastructure becomes harder to build and maintain because of climate effects, such as thawing permafrost and coastal erosion. Commercial and tourism use of Arctic shipping routes could increase the need for rescue capabilities, and autonomous system technology might provide both a safer and more persistent capability in harsh, remote environments.
- *Coast Guard operations doctrine:* The operations doctrine will need revisions. Techniques, tactics, and procedures will need to be adapted to a different MTS operating environment, as will internal policy implications for responding to autonomous system failure and resulting collisions, allisions, groundings, spills and other safety-of-navigation incidents, security events, and human-caused and natural disasters.

How an Autonomous Maritime Transportation System Might Use Data Science

The broad field of data science has many applications for autonomous maritime transportation systems (AMTSs), such as the vessel shown in Figure A.1. An individual system can use AI and ML in many functions and subsystems:

- *perception:* A key task of many AMTSs is to understand the environment around the system and to recognize objects and activities (e.g., recognizing humans in a SAR context or recognizing threatening activity in port, waterway, and coastal security applications). ML is central to developing and improving object and activity recognition functions of the AMTS.
- *sensor fusion:* A single AMTS can have multiple sensors to gather data about and make sense of the environment. These can include multiple cameras, sonar, the Global Positioning System (GPS), inertial navigation systems, and depth sensors. There are also many proprioceptive sensors to measure the internal state of the system. These systems can produce enormous amounts of data per second that must be fused to produce a coherent state of

the system's external and internal world. Thus, sensor fusion requires big data and AI applications simultaneously.

- *simultaneous localization and mapping (SLAM):* Many AMTSs may operate in areas that are unknown or for which high-fidelity maps are unavailable—for example, in underwater applications or remote areas. In these environments, AMTSs may perform SLAM, in which the system simultaneously creates a map of the world while identifying its location within that mapped world. SLAM is a classic application of AI and ML.
- *path planning:* AMTSs by definition are mobile, so they usually require some sort of path planning (i.e., the ability to develop a plan for how to navigate from point A to point B). Like SLAM, path planning

FIGURE A.1

An Autonomous Vessel Prototype: *Sea Hunter*, U.S. Navy, Office of Naval Research

SOURCE: U.S. Coast Guard Innovation Program, 2017, p. 4.

is a fundamental AI application that builds on perception and sensor fusion to navigate around obstacles in the environment, reach intended destinations, and respond to changes.

- *task prioritization and allocation:* Many uses of AMTSs require performing tasks (e.g., retrieving objects A, B, and C and delivering them to locations X, Y, and Z). This higher-order functioning of organizing and prioritizing tasks is another classic AI/ML application. The system identifies the tasks it must complete, calculates the efficiency or value of performing tasks in a particular order or with particular resources, and chooses or prioritizes tasks that maximize some goal typically defined by the operator.

Data science is also important for the Coast Guard to manage future fleets of AMTSs in at least two ways:

- *operations management:* A set of AMTSs must be managed to ensure that tasks and activities are completed, resources are allocated, and the AMTSs are maintained. As with many kinds of assets, data analytics and AI are important for managing AMTS fleet operations.
- *fleet learning:* The observations and activities of a set of AMTSs may be greater than the sum of the observations of a single system. For example, imagery simultaneously gathered and integrated from several systems might be needed to effectively identify, monitor, and assess maritime environmental damage and recovery. Big data applications and ML may be critical for integrating data from across the fleet and identifying patterns in those data (e.g., assessing the spread and containment of oil after a spill from the operations of multiple AMTSs).

When interacting with an external market that, like the MTS, might embrace autonomous systems at a rapid pace, the Coast Guard should preemptively examine its related goals, capabilities, and responsibilities well in advance of any broad-scale adaptation and proliferation.

References

"artificial intelligence, n.," *OED Online*, Oxford University Press, March 2020. As of April 20, 2020: http://www.oed.com/view/Entry/271625

Balboni, Fred, Glenn Finch, Cathy Rodenbeck Reese, and Rebecca Shockley, *Analytics: A Blueprint for Value—Executive Summary*, IBM Institute for Business Value, October 2013. As of April 27, 2020: https://www.ibm.com/downloads/cas/4WBWGBJL

Boyer, Sebastien, Ben U. Gelman, Benjamin Schreck, and Kalyan Veeramachaneni, "Data Science Foundry for MOOCs," paper presented at the 2015 Institute of Electrical and Electronics Engineers International Conference on Data Science and Advanced Analytics, Paris, 2015, pp. 1–10.

Cameron, William Bruce, *Informal Sociology: A Casual Introduction to Sociological Thinking*, New York: Random House, 1963.

Chen, Te-Ping, "Three Hours of Work a Day? You're Not Fooling Anyone," *Wall Street Journal*, July 19, 2019.

Chui, Michael, James Manyika, Mehdi Miremadi, Nicolaus Henke, Rita Chung, Pieter Nel, and Sankalp Malhotra, *Notes from the AI Frontier: Applications and Value of Deep Learning*, discussion paper, McKinsey Global Institute, April 2018. As of April 20, 2020: https://www.mckinsey.com/featured-insights/artificial-intelligence/notes-from-the-ai-frontier-applications-and-value-of-deep-learning

"Data Analysis," *Responsible Conduct of Research*, Faculty Development and Instructional Design Center, Northern Illinois University, undated. As of November 5, 2019: https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/datopic.html

DeBrusk, Chris, "The Risk of Machine-Learning Bias (and How to Prevent It)," *MIT Sloan Management Review*, March 26, 2018.

DrivenData, "N+1 Fish, N+2 Fish: About the Project," undated. As of April 20, 2020: https://www.drivendata.org/competitions/48/identify-fish-challenge/page/91/

Fitz-enz, Jac, *How to Measure Human Resources Management*, New York: McGraw-Hill, 1984.

Frické, Martin, "The Knowledge Pyramid: A Critique of the DIKW Hierarchy," *Journal of Information Science*, Vol. 35, No. 2, 2009, pp. 131–142.

Gandomi, Amir, and Murtaza Haider, "Beyond the Hype: Big Data Concepts, Methods, and Analytics," *International Journal of Information Management*, Vol. 35, No. 2, April 2015, pp. 137–144.

ImageNet, "Large Scale Visual Recognition Challenge 2017 (ILSVRC2017)," c. 2017. As of October 2019: http://image-net.org/challenges/LSVRC/2017/results

Labelle, Christiane M., and Lee Dyer, *A Role-Based Taxonomy of Human Resource Organizations*, Ithaca, N.Y.: Cornell University, School of Industrial and Labor Relations, Center for Advanced Human Resource Studies, Working Paper 92-35, July 1992. As of April 20, 2020: https://digitalcommons.ilr.cornell.edu/cahrswp/322/

"machine learning, n.," *OED Online*, Oxford University Press, March 2020. As of April 20, 2020: http://www.oed.com/view/Entry/111850

Mayo, Matthew, "Deep Learning Key Terms, Explained," *KDnuggets*, October 2016. As of April 27, 2020: https://www.kdnuggets.com/2016/10/deep-learning-key-terms-explained.html

Office of Emerging Policy, U.S. Coast Guard, *Workforce 2040: Executive Summary*, October 2019.

Provost, Foster, and Tom Fawcett, "Data Science and Its Relationship to Big Data and Data-Driven Decision Making," *Big Data*, Vol. 1, No. 1, March 2013, pp. 51–59.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang,, Andrej Karpathy, Aditja Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, Vol. 115, No. 3, December 2015, pp. 211–252.

Savitz, Scott, Aaron C. Davenport, and Michelle D. Ziegler, *The Marine Transportation System, Autonomous Technology, and Implications for the U.S. Coast Guard*, Homeland Security Operational Analysis Center operated by the RAND Corporation, PE-359-DHS, 2020. As of May 4, 2020: https://www.rand.org/pubs/perspectives/PE359.html

Silver, David, and Demis Hassabis, "AlphaGo Zero: Starting from Scratch," DeepMind blog post, October 18, 2017. As of April 20, 2020: https://deepmind.com/blog/article/alphago-zero-starting-scratch

UK Civil Service Human Resources, *Global HR Design Principles and Process Taxonomy*, briefing, March 2018. As of April 20, 2020: https://www.gov.uk/government/publications/global-hr-design

U.S. Coast Guard, "Workforce 2040," Evergreen workshop discussions, September 2019.

U.S. Coast Guard Innovation Program, *Autonomous Systems Challenge*, July 2017. As of April 20, 2020: https://www.uscg.mil/Portals/0/Strategy/Autonomous%20Systems%20Challenge%20Report%202017.pdf

Wang, Richard Y., and Diane M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, Vol. 12, No. 4, Spring 1996, pp. 5–33.

Zengerle, Patricia, and Megan Cassella, "Millions More Americans Hit by Government Personnel Data Hack," Reuters, July 9, 2015. As of April 20, 2020: https://www.reuters.com/article/us-cybersecurity-usa/millions-more-americans-hit-by-government-personnel-data-hack-idUSKCN0PJ2M420150709

About the Authors

Aaron C. Davenport is a senior policy researcher at the RAND Corporation. His research focuses include border and maritime security, emergency preparedness and response, occupational health and safety, and national security strategy. He is a graduate of the U.S. Coast Guard Academy and a retired Coast Guard senior officer with security assistance, search-and-rescue, and law-enforcement experience. He has an M.S. in environmental sciences, with a certificate in industrial hygiene and a minor in hazardous materials.

Michelle D. Ziegler is a technical analyst at the RAND Corporation. Her research focuses include U.S. Army logistics, disaster recovery, U.S. Coast Guard capability and capacity analysis, and cooperation and domain awareness in the Arctic. She has an M.S. in astronomy.

Abbie Tingstad is a senior physical scientist and associate director of the Engineering and Applied Sciences Department at the RAND Corporation. Her research focuses on issues related to strategy and planning in defense and homeland security, and for the environment. Much of her work explores the intersections between organizations, processes, technologies, and people. She has a Ph.D. in geography.

Katherine Anania is a technical analyst at RAND and has multidisciplinary experience with the U.S. Department of Defense, the U.S. Department of Homeland Security, and in the environmental field. She has an M.E.S.M. in environmental science and management and an M.A. in economics.

Daniel Ish is an associate physical scientist at RAND, focusing on technical and quantitative analyses. He has done work involving cybersecurity, logistics, supply chain risk management, supervised machine learning and bibliometric analyses of scientific abstracts. He has a Ph.D. in physics.

Nidhi Kalra is a senior information scientist at the RAND Corporation. Her research focuses on autonomous vehicle policy, climate change adaptation, and tools and methods that help people and organizations make better decisions amid deep uncertainty. She spearheads RAND's autonomous vehicle policy work. She has a Ph.D. in robotics.

Scott Savitz is a senior engineer at the RAND Corporation. Much of his research focuses on how to improve the effectiveness and resilience of operational forces, as well as the impact of reallocating resources among those forces. He has a Ph.D. in chemical engineering.

Rachel Liang is the Nuclear Section chief for the Cybersecurity and Infrastructure Security Agency. She was a Department of Homeland Security fellow with the RAND Homeland Security Research Division in 2019–2020. She has a master's degree in weapons of mass destruction security policy from George Washington University's Elliott School of International Affairs.

Melissa Bauman is a communications analyst who helps researchers make their complex findings accessible to a sophisticated audience of lawmakers, journalists, and practitioners. She has a B.A. from the University of Kansas William Allen White School of Journalism and Mass Communications.

About This Perspective

This Perspective documents support by the Homeland Security Operational Analysis Center (HSOAC) to the U.S. Coast Guard's Evergreen project. Founded in 1996, Evergreen is the Coast Guard's strategic foresight initiative, which has historically run in four-year cycles and uses scenario-based planning to identify strategic needs for the incoming service chief. In 2019, Evergreen was restructured to best support executive leaders in their roles as the Coast Guard's decision engines. The project objective is to help posture the Coast Guard to better bridge the gap between future challenges and near-term plans, which typically focus on the urgent needs of the present. HSOAC analysts reviewed Evergreen activities, examined Coast Guard strategy-making and planning processes, adapted an approach for developing scenarios, and narrated a set of exemplar global planning scenarios. The individual Perspectives that resulted from this project reflect themes and specific subjects that have emerged from a series of workshops that were conducted with subject-matter experts and were identified as areas of particular interest for senior leadership strategic-planning activities and emerging policy development.

This research was sponsored by the U.S. Coast Guard Office of Emerging Policy and conducted within the Strategy, Policy, and Operations Program of the HSOAC federally funded research and development center (FFRDC).

The RAND Corporation operates HSOAC under contract to the U.S. Department of Homeland Security (DHS). RAND is a research organization that develops solutions to public policy challenges to help make communities throughout the world safer and more secure, healthier and more prosperous. RAND is nonprofit, nonpartisan, and committed to the public interest. For more information, visit www.rand.org/hsrd/hsoac.

HSOAC
HOMELAND SECURITY
OPERATIONAL ANALYSIS CENTER

An FFRDC operated by the
RAND Corporation under
contract with DHS

The Homeland Security Act of 2002 (Section 305 of Public Law 107-296, as codified at 6 U.S.C. § 185) authorizes the Secretary of Homeland Security, acting through the Under Secretary for Science and Technology, to establish one or more FFRDCs to provide independent analysis of homeland security issues. The RAND Corporation operates HSOAC as an FFRDC for DHS under contract HSHQDC-16-D-00007.

The HSOAC FFRDC provides the government with independent and objective analyses and advice in core areas important to the department in support of policy development, decisionmaking, alternative approaches, and new ideas on issues of significance. The HSOAC FFRDC also works with and supports other federal, state, local, tribal, and public- and private-sector organizations that make up the homeland security enterprise. The HSOAC FFRDC's research is undertaken by mutual consent with DHS and is organized as a set of discrete tasks.

The information presented in this Perspective does not necessarily reflect official DHS opinion or policy.

For more information on this publication, visit www.rand.org/t/PEA150-1.