Achieving Mission Impact with Data Science

John Piorkowski

ABSTRACT

Data science emerged as a popular technical field by leveraging the advances in data storage, computing, and machine learning. Practical applications of data science are far-reaching and include marketing, fraud detection, logistics, crime prediction, social engagement, sports team management, and health care. Recognizing this profound impact, the Johns Hopkins University Applied Physics Laboratory (APL) Asymmetric Operations Sector (AOS) created the Data Science Initiative (DSI) to apply data science to national security challenges and health care. The DSI accelerated APL data science contributions to national security and health care by creating new research initiatives and establishing deep technical competencies that shaped and directed novel solutions across the AOS mission space.

INTRODUCTION

In 2016, each APL sector established next-generation initiatives with internal investment to position the Laboratory to create innovations that were yet envisioned by our sponsors. Accordingly, the Asymmetric Operations Sector (AOS) created the Data Science Initiative (DSI) to address technologies for solving the enduring problem of exponential data growth across the national security and medical communities with insufficient analysts. (The term *analyst* broadly refers to technology users of interest to AOS. They include cyber operators, intelligence analysts, clinicians, and special operators.) Overall, the DSI led to the sector establishing a new competency and several technologies that have resulted in innovative contributions across mission areas.

WHAT IS DATA SCIENCE?

The Beginnings of Data Science

Data collection and analysis have been around long before the advent of the computer. A notable example is the work of Matthew Fontaine Maury, who was known as the Scientist of the Seas. Maury was a pioneer in the field of ocean navigation during the mid-1800s.[1] He joined the Navy at the age of 19, but a stagecoach accident forced him to give up traveling the seas and to take an assignment at the Navy with the Depot of Charts and Instruments. The Depot of Charts and Instruments would later become the US Naval Observatory. By studying meteorology, collecting data from ship's logs, and creating charts, Maury revolutionized our understanding of oceanography and marine naviga-

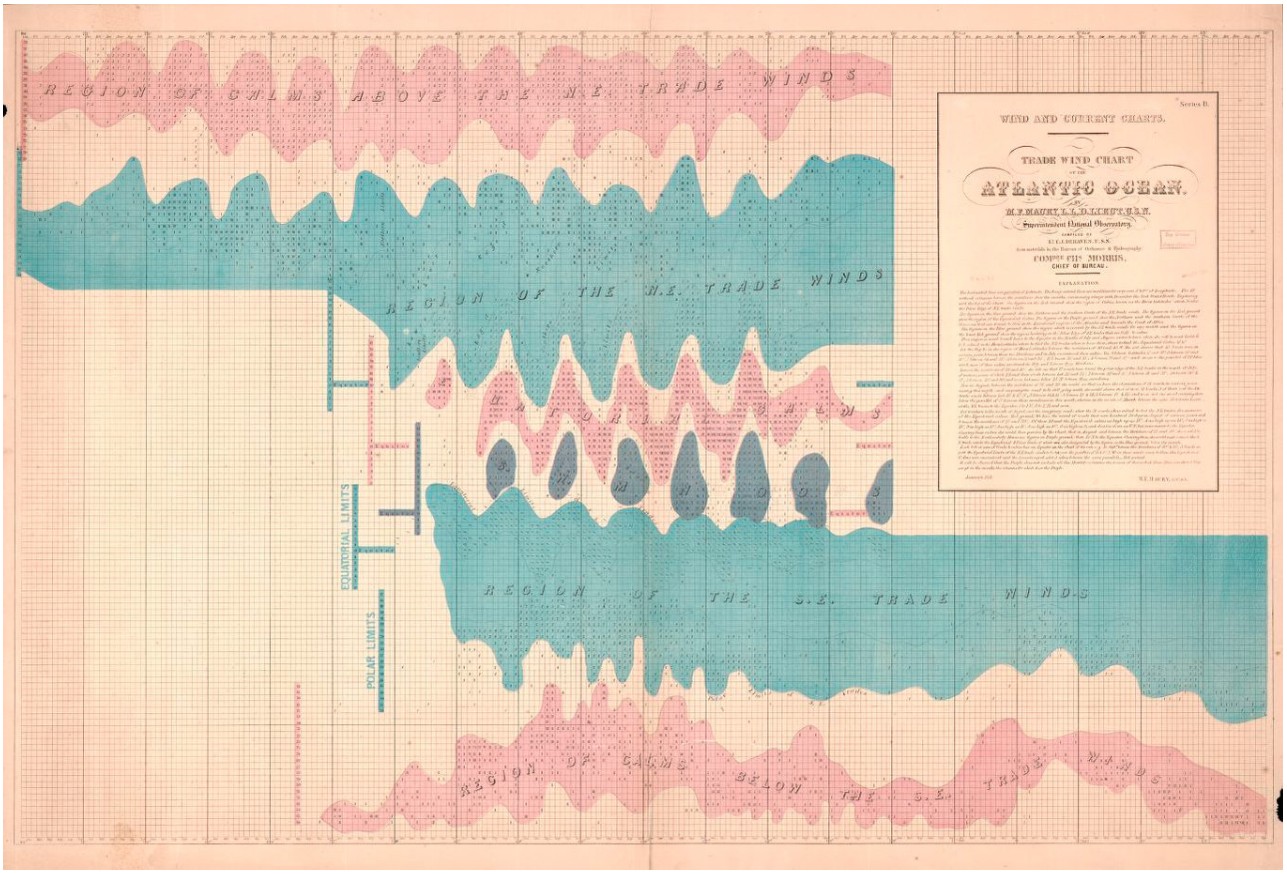
REGION OF CALMS ABOVE THE N.E. TRADE WINDS
REGION OF THE N.E. TRADE WINDS
EQUATORIAL CALMS
MONSOONS
EQUATORIAL LIMITS
POLAR LIMITS
REGION OF THE S.E. TRADE WINDS
REGION OF CALMS BELOW THE S.E. TRADE WINDS
Series B.
WIND AND CURRENT CHARTS.
TRADE WIND CHART
OF THE
ATLANTIC OCEAN.
BY
M. F. MAURY, L.L.D. LIEUT. U.S.N.
Superintendent National Observatory.
COMPILED BY
Lt. E. J. DEHAVEN, U.S.N.
from materials in the Bureau of Ordnance & Hydrography
COMDRE CHS MORRIS,
CHIEF OF BUREAU.
EXPLANATION

Figure 1. *Trade Wind Chart of the Atlantic Ocean* by Matthew Fontaine Maury, 1851. Maury was a pioneer in the field of ocean navigation during the mid-1800s. This map, one of many he created, assisted ship captains with their cross-Atlantic journeys. (From Geography and Map Division, Library of Congress.)

tion. Figure 1 illustrates his 1851 *Trade Wind Chart of the Atlantic Ocean*, which assisted ship captains at the time with their cross-Atlantic journeys.

A modern history of data science enabled by computing is often credited to a 1962 paper by John Tukey titled "The Future of Data Analysis."[2] In this paper, he describes procedures for analyzing data, interpreting results, and planning for the gathering of data, as well as the statistics that apply to these procedures. Tukey's prophecy of data analysis motivated a shift from theoretical statistics and advocated for applied statistics to become data analytics. Tukey's paper has been reviewed more recently and still stands as a foundation for modern data science.[3]

In 1974, Peter Naur published the "Concise Survey of Computer Methods" and repeatedly used the term *data science*, defining it as "the science of dealing with data, once they have been established, while the relation of the data to what they represent is delegated to other fields and sciences."[4] Despite the repeated use of the term in Naur's publication, many credit William Cleveland with coining the term *data science* with his publication in 2001. In his paper,[5] he advocates for a substantial change to the field of statistics. To reinforce a significant change, he advocated for a new field called data science. He asserted that data science should include the following:

- Multidisciplinary investigations
- Models and methods for data
- Computational systems
- Pedagogy for education
- Evaluation of tools
- Theoretical foundations

Cleveland's paper is cited as the seminal data paper; however, the field did not gain popularity until the explosion of internet connectivity, the low cost of data storage, and the big data era. The term *big data* refers to large and complex data that cannot be addressed with traditional relational database tool sets.[6]

Data Science Today

Leveraging the advances in data storage, computing, and machine learning, data science emerged as a popular field. Practical applications of data science are

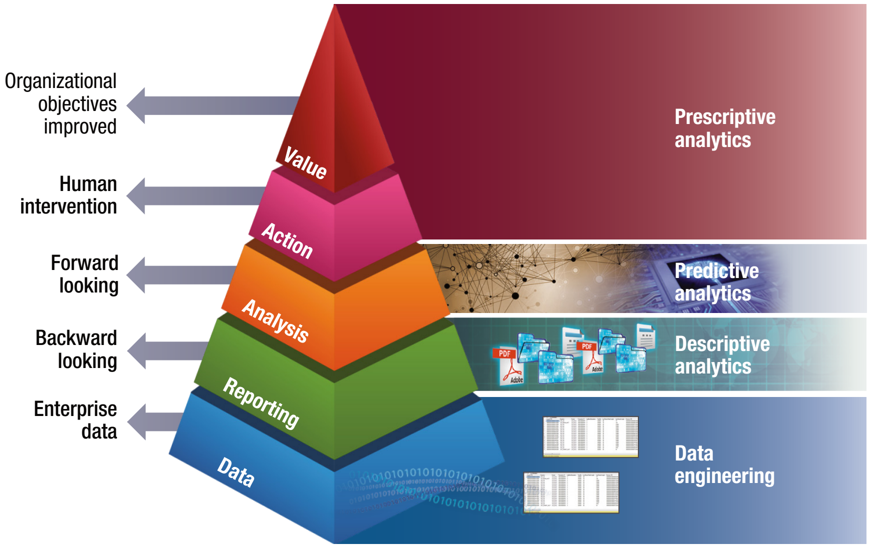

Figure 2. Data science maturity model. Derived from Anderson's analytics value chain,[8] the model shows the progression from backward-looking to forward-looking data. Moving from descriptive analytics (understanding the characteristics of the data) to predictive analytics (using data about the past to make predictions about the future) to prescriptive analytics (accounting for human intervention by providing for decision options) enables analysts to make decisions and achieve objectives using the data.

far-reaching and include marketing, fraud detection, logistics, crime prediction, social engagement, sports team management, and health care.[7] Any data science application can be considered along the data science maturity ladder illustrated in Figure 2. The maturity model is adapted from the analytics value chain presented by Anderson.[8] The initial (and often the most resource-intensive) step to mature a data-driven approach is the data engineering. Data engineering can also be described as data wrangling; curation; or extract, transform, load (ETL). As previously mentioned, the reduced cost of memory has led to the creation of enormous amounts of data. However, the data are often contained in disparate systems and are not well suited for modern data science algorithms. In the discussion of machine learning, the data must be engineered to allow feature representations. Neil Lawrence offers a framework to assess data readiness for analytics, shown in Figure 3.[9] He uses three levels of data readiness. The lowest level (Class C) describes the challenges with data engineering and wrangling. As Lawrence explains, many organizations claim they have data, but the data have not been made available for analytic use. He refers to this type of data as "hearsay" data. Class B data require an understanding of the faithfulness and representation of the data. Finally, Class A data are data in context. An analyst understands whether Class A data can answer organizational questions.

Once data are made available in a data warehouse or data lake, they can be reported. Many organizations create reports using spreadsheets or text documents. This approach looks backward, reflecting what has happened in the past. The promise of data science is to move beyond backward-looking reporting to forward-looking analysis. In the field of data science, analytics are typically described as descriptive, predictive, and prescriptive. Descriptive analytics involve understanding the characteristics of the data. For numerical data, descriptive analytics would include statistical measures such as means, standard deviations, modes, and medians. Other analytics may include histograms. Descriptive analytics help to discover anomalous and missing data examples. Descriptive analytics are backward looking as well.

Moving from backward-looking data to forward-looking data can be achieved with predictive analytics. Predictive analytics use data about the past to make predictions about the future. Supervised machine learning provides an analytical tool for predictive analytics. Supervised machine learning uses training data to create a machine learning model. The machine learning model can then be used to make predictions about new data sets.

Prescriptive analytics, the third type of data science analytics, address the human intervention by providing for decision options. Moving beyond predictive

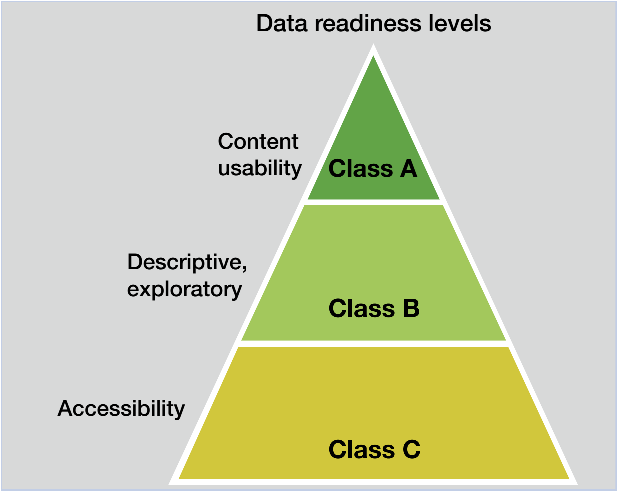

Figure 3. Data readiness levels defined by Neil Lawrence.[9] Class C data face challenges with data engineering and wrangling. Class B data require an understanding of their faithfulness and representation. Class A data are data in context. An analyst understands whether Class A data can answer organizational questions.

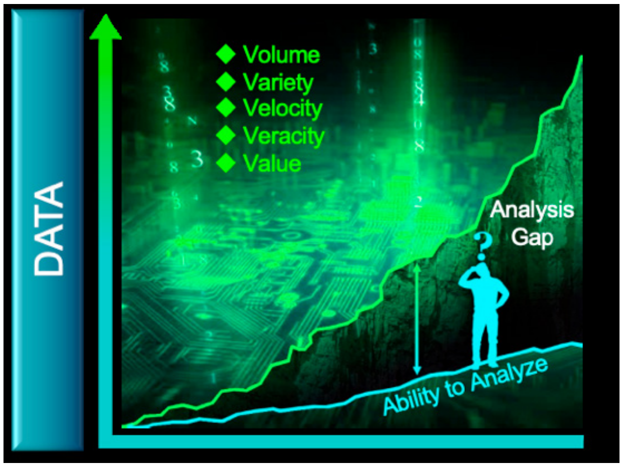

Figure 4. The analysis gap. Significant internet connectivity and the low cost of computer storage enabled an exponential increase in the amount of data APL sponsor communities were able to collect and store. However, their analyst populations were not growing at the same rate, creating an analysis gap.

analytics, which describe a future state, prescriptive analytics offer courses of action to bring value to an organizational objective. Reinforcement learning is a machine learning approach that provides a foundation for prescriptive analytics.

THE AOS DATA SCIENCE INITIATIVE

In 2016, AOS staff members observed that their sponsor communities, including cyber analysts, intelligence analysts, and customs and border patrol agents, were being inundated with data. As these communities benefited from significant internet connectivity and the low cost of computer storage, the amount of data they were able to collect and store was growing at exponential rates. However, their analyst populations were not growing at exponential rates, creating the analysis gap shown in Figure 4. This gap motivated AOS to stand up the DSI to use machine computation to close the analysis gap. Internal research initiatives addressed a broad set of data science and artificial intelligence (AI) approaches. Examples include applying transfer learning techniques in deep learning to address the training machine learning classifiers with small amounts of labeled data that represent national security targets. For instance, we explored predictive models to improve targeting with global signals from intelligence data, as well as new algorithms to address scalability issues when applying unsupervised graph analytics to practical real-world applications.

Not only did the DSI yield innovative research, but it also allowed the sector to significantly increase its competencies and to make contributions across multiple mission domains. The next sections describe contributions that leveraged research and competencies across mission domains including cyber; intelligence, surveillance, and reconnaissance (ISR); trade fraud detection; and health care.

Data Science in Cyber

The recognized gap in cyber situational awareness for enterprise networks motivated the sector to apply data science to cyber operations and led to the creation of APL's LIVE Lab (Live data, Integration, Validation, and Experimentation Lab), which focused on creating an analytics platform using internal APL network data (Figure 5). In this innovative facility, researchers mapped a duplicate of APL's internal network on banks of huge television screens. They then tested different cyber monitoring and security technologies to see how they affected information flow as an intruder moved around systems. The creation of LIVE Lab's analytic platform, which originally focused on internal network

Figure 5. APL's LIVE Lab. Analysts and engineers work in this unique facility that helps researchers develop solutions to continually evolving cyber threats by detecting and monitoring intrusion attempts on APL's network in real time. To assess new cyber defense techniques for the government, APL uses LIVE Lab to mirror its real-time network as a testing ground.

situational awareness, caught the attention of sponsors in the national security community and led to several long-term sponsor-funded efforts pursuing data science for cyber operations. One notable effort is the Department of Homeland Security (DHS) Advanced Cyber Analytics Environment. APL explored numerous analytic techniques to increase machine processing to assist operators in defending government networks; techniques that were investigated include unsupervised, supervised, and semi-supervised machine learning. One successful result included the identification of Domain Name Service (DNS) tunneling attacks. The DNS is the network protocol that maps human-readable domain names to Internet Protocol (IP) addresses that machines can use to route packets across the internet. DNS packets flow continuously between network-connected devices and comprise massive amounts of data. Because it is known to contain vulnerabilities, the DNS protocol is a common attack vector for malicious actors. Buzcak et al.[10] describe a data science approach that was able to detect over 99% of malicious DNS tunnels.

To achieve these results, the team first created accurate data sets to train the DNS detection algorithms, leveraging the LIVE Lab analytic platform using the APL network to curate high-quality data. They then undertook a penetration testing collection effort. With this training data, the team applied traditional machine learning techniques such as random forest. Before applying machine learning techniques, an import step of feature engineering must be completed. Significant research of prior DNS tunneling detection algorithms and deep knowledge of DNS protocols led to a set of relevant features. Example features include:

- DNS query type
- DNS packet length
- Number of distinct substrings in the DNS “QNAME” field
- DNS response packet length
- DNS query string length

Through thoughtful feature engineering, the team readily applied traditional random forest machine learning techniques, resulting in detection of over 99% of malicious DNS traffic.

In addition, an APL team explored a broad set of analytical approaches for cyber operations, including supervised learning approaches (such as pattern mining and deep neural networks), unsupervised techniques (such as K-means), and semi-supervised techniques.[11–13] A set of DHS-sponsored exploratory experiments led to APL guiding the implementation of analytic platforms and analytics approaches in DHS cyber operations.

Data Science in ISR

APL’s research in ISR challenges leveraged groundbreaking academic work in deep neural networks (DNNs) by Krizhevsky, Sutskever, and Hinton.[14] One significant effort was the development of the ImageNet data set[15] that commenced in 2009; it took 2.5 years to label 3.2 million images. The data set served as the foundation for the ImageNet Large Scale Visual Recognition Challenge that started in 2010. In 2012, Krizhevsky, Sutskever, and Hinton won the competition by a significant margin, and this result inspired the current AI boom.[14] Their technique used neural networks that had existed for decades. However, by applying neural networks with a large labeled data set and leveraging modern graphics processing units (GPUs), they advanced the field with their work. GPUs provided extraordinary computation power that allowed for the design of neural networks with significant layers (i.e., DNNs).

APL explored this research and its application to national security problems, especially in the area of ISR for special operations. The ImageNet data set was built by labeling pictures posted on the internet. However, the ISR mission involves a mix of different sensors to include synthetic aperture radar (SAR), full-motion video (FMV), and radio frequency (RF) sensors. The challenge faced in ISR was the lack of labeled data at the scale of ImageNet corpus. For example, only hundreds of labeled images existed. APL applied the use of transfer learning to create a universal feature extractor that allowed DNNs to be trained with small amounts of training data. The innovative concept of the universal feature extractor is described by Rodriguez et al.[16] Its advantages are twofold. First it enables a design to train classifiers with small amounts of labeled data (i.e., sparse data). The second benefit is that training time for new objects can be significantly reduced. The DSI further matured the universal feature extractor and parallelized the algorithm to address object detection in full-motion video.

APL aptly applied the universal feature extractor to targets of military interest to the Intelligence Community and Department of Defense. The impressive accuracy achieved with this approach led to several projects where APL contributed machine learning solutions to these communities.

Data Science in Illicit Trade Discovery

Given the Lab’s trusted role with US Customs and Border Protection, the agency provided APL access to a large data set of 4 years of trade data. Specifically, the data set included entry summary reports for over 4 years, which included over 200 million shipping records. To discover illicit shipments, the agency was using an automatic target recognition capability that contained rule-based systems. Rule-based systems fail when nefarious actors are caught and then change their tactics, techniques, and procedures, essentially working around the rule-based systems. Seeking a new analytical approach for discovering illicit trade, the agency provided APL trade records so that the APL team could explore novel analytics. To discover new illicit patterns, the APL team applied unsupervised machine learning using probabilistic graph analytics. Graphs provide a flexible data structure that facilitates fusion of disparate data sets.

We leveraged SOCRATES, a graph analytics capability[17] to analyze large-scale graphs. The SOCRATES platform was developed as part of internal investment and direct-funded projects. The successful implementation of graph analytics revolves around several key considerations: rapid data ingestion and retrieval, scalable storage, and parallel processing. SOCRATES provides a graph analytics platform that is particularly focused on facilitating analysis of large-scale data sets. As part of the platform, SOCRATES provides a rich set of analytics. The clique tree analytic[18] provided critical insights into discovering anomalies in trade data. The clique tree technique addressed probabilistic graph modeling of the trade records, which involved categorical type data in a highly dimensional graph space. Applying probabilistic techniques, including the APL clique tree analytic, to the large-scale high-dimensional data set produced new insights from the trade data by discovering anomalous patterns. The exploratory work that APL performed convinced Customs and Border Protection that advanced analytics possessed merit beyond their traditional rule-based systems. Subsequently, APL served in a trusted agent role to assist the agency in working to introduce these advanced analytics into their operational systems.

Data Science in Precision Medicine

In 2017, APL partnered with Johns Hopkins Medicine to bring data science capabilities to precision medicine. The first step was to create an architecture for an analytic platform. APL provided thought leadership by leveraging the analytic platform being created in the cyber operations efforts previously mentioned. The architecture led to the creation of the Precision Medicine Analytics Platform (PMAP). As described by Alan Ravitz in his article in this issue and illustrated in Figure 6, "PMAP handles the 'dirty work' of creating pipelines to access disparate, high-velocity, high-volume data (i.e., big data)." It aggregates these disparate data into a single Data Commons to facilitate access, obviating the existence of multiple researchers independently creating different tools to access the same data and storing them separately. The Data Commons affords a single repository that combines the transactional data of the electronic health record with other sources of data while also providing a single point of storage from which secure study-specific projections of data can be provisioned to researchers with institutional approval to access them."

The creation of PMAP enabled researchers to pursue new data science solutions. One example includes applying natural language processing techniques to the problem of mining medical records. Building on PMAP, Chee, Joice, and Johnson[19] created a pipeline to discover key information for prostate cancer in electronic

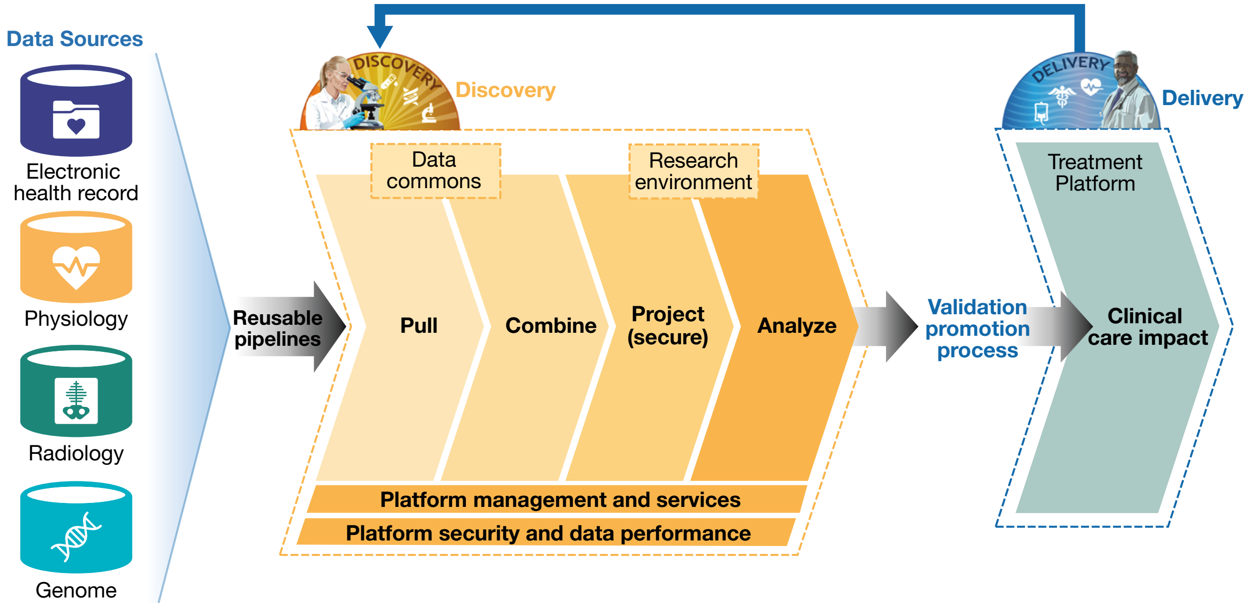

Figure 6. The PMAP platform. PMAP pulls data from multiple sources and aggregates them into the Data Commons. Approved researchers can then access needed data in a secure Research Environment where they can also access a suite of tools and capabilities built for other studies.

medical records. Furthermore, this same group[20,21] extracted information such as Gleason scores to identify patient populations for research. The information can also be used in clinical settings to identify information that is potentially ambiguous or difficult to extract and requires clarification. The pipeline was built on the PMAP foundation.

SUMMARY

Data science emerged in 2010 and continues as a vibrant technology area that is pervasive across many industries, including national security. AOS leaned strongly into this field by creating the DSI. This 2-year initiative had profound impact on the work AOS staff members do, enabling contributions in cyber operations, international trade, ISR, and health care. Additional projects benefited from the DSI, and data science has become a core competency for AOS. The DSI and the contributions it propelled demonstrate the benefits of focused investment in a strategic area to achieve impact for national security missions. When looking toward the future of the data science, we see the field maturing and becoming more commonplace across missions and organizations. With that said, with its systems engineering perspective, APL can provide thought leadership in advancing the science of data science.

REFERENCES

[1] T. St. Onge, “Scientist of the seas: The legacy of Matthew Fontaine Maury,” Library of Congress blog, Jul. 25, 2018, https://blogs.loc.gov/maps/2018/07/scientist-of-the-seas-the-legacy-of-matthew-fontaine-maury/.

[2] J. W. Tukey, “The future of data analysis,” *Ann. Math. Statist.*, vol. 33, no. 1, pp. 1–67, 1962, https://doi.org/10.1214/aoms/1177704711.

[3] C. Mallows, “Tukey’s paper after 40 years,” *Technometrics*, vol. 48, no. 3, pp. 319–325, 2006, https://doi.org/10.1198/004017006000000219.

[4] P. Naur, *Concise Survey of Computer Methods*, Lund, Sweden: Studentlitteratur, 1974.

[5] W. Cleveland, “Data science: An action plan for expanding the technical areas of the field of statistics,” *Int. Statistic. Rev.*, vol. 69, no. 1, pp. 21–26, 2001, https://doi.org/10.1111/j.1751-5823.2001.tb00477.x.

[6] C. Andrus, J. Cook, and S. Sood, “Chapter 01: A history of data science,” in *Data Science: An Introduction* (Wikibook), https://en.m.wikibooks.org/wiki/Data_Science:_An_Introduction/A_History_of_Data_Science.

[7] M. Rice, “17 data science applications & examples,” Built In, Jul. 23, 2019 (updated Mar. 25, 2020), https://builtin.com/data-science/data-science-applications-examples.

[8] C. Anderson, *Creating a Data-Driven Organization*, 1st ed., Sebastopol, CA: O’Reilly Media, Inc., 2015.

[9] N. D. Lawrence, “Data readiness levels,” 2017, https://arxiv.org/abs/1705.02245.

[10] A. L. Buczak, P. A. Hanke, G. J. Cancro, M. K. Toma, L. A. Watkins, and J. S. Chavis, “Detection of tunnels in PCAP data by random forests,” in *Proc. 11th Annu. Cyber and Inf. Sec. Res. Conf. (CISRC ’16)*, 2016, pp. 1–6, https://doi.org/10.1145/2897795.2897804.

[11] A. L. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” in *Proc. IEEE Commun. Surv. Tut.*, vol. 18, no. 2, pp. 1153–1176, 2016, https://doi.org/10.1109/COMST.2015.2494502.

[12] L. Watkins, S. Beck, J. Zook, A. Buczak, J. Chavis, et al., “Using semi-supervised machine learning to address the big data problem in DNS networks,” in *Proc. 2017 IEEE 7th Annu. Comput. Commun. Workshop and Conf., CCWC 2017*, 2017, pp. 1–6, https://doi.org/10.1109/CCWC.2017.7868376.

[13] L. Watkins, J. Chavis, A. L. Buczak, D. S. Berman, S. W. Yen, and L. T. Duong, “Using sequential pattern mining for common event format (CEF) cyber data,” in *Proc. 12th Annu. Cyber and Inf. Secur. Res. Conf. (CISRC ‘17)*, 2017, pp. 1–4, https://doi.org/10.1145/3064814.3064822.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. 25th Int. Conf. Neural Inf. Process. Syst. (NIPS 2012)*, pp. 1097–1105, 2012.

[15] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. 2009 IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2009, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.

[16] P. A. Rodriguez, N. Drenkow, D. DeMenthon, Z. Koterba, K. Kauffman, et al., “Selection of universal features for image classification,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2014, pp. 355–362, https://doi.org/10.1109/WACV.2014.6836078.

[17] C. Savkli, R. Carr, M. Chapman, B. Chee, and D. Minch, “Socrates,” in *Proc. IEEE High Perform. Extreme Comput. Conf. (HPEC)*, 2014, pp. 1–6, https://doi.org/10.1109/HPEC.2014.7040993.

[18] C. Savkli, J. R. Carr, P. Graff, and L. Kennell, “Bayesian learning of clique tree structure,” in *Proc. Int. Conf. Data Mining (DMIN)*, 2016, https://arxiv.org/abs/1708.07025.

[19] B. Chee, G. Joice, and M. Johnson, “A novel natural language processing pipeline automates unstructured data extraction within medical reports,” presented at the National Comprehensive Cancer Network 23rd Annual Conf., Orlando, FL, Mar. 2018.

[20] B. Chee, G. Joice, and M. Johnson, “Natural language processing allows for accurate and automated extraction of data from prostate biopsy pathology reports,” presented at the National Comprehensive Cancer Network 23rd Annual Conf., Orlando, FL, Mar. 2018.

[21] G. Joice, B. Chee, N. Gupta, and M. Johnson, “MP76-18 natural language processing allows for accurate and automated extraction of data from prostate biopsy pathology reports,” *J. Urol.*, vol. 199, no. 4S, pp. e1025–e1026, 2018, https://doi.org/10.1016/j.juro.2018.02.2586.

John Piorkowski, Asymmetric Operations Sector, Johns Hopkins University Applied Physics Laboratory, Laurel, MD

John Piorkowski is the chief artificial intelligence architect and Applied Information Sciences Branch head within APL’s Asymmetric Operations Sector. Dr. Piorkowski received a BS in electrical engineering from Pennsylvania State University (Penn State), an MS in electrical engineering from Johns Hopkins University, and a PhD in information systems from the University of Maryland, Baltimore County. In his current roles, he provides technical oversight and technical staff management for national security and health care efforts. Dr. Piorkowski also serves as the chair for the Artificial Intelligence program and co-chair for the Data Science program in the Whiting School of Engineering at Johns Hopkins University. As chair, he provides guidance on strategic direction, including strategies for excellence in curriculum design and faculty quality. As an adjunct faculty member, he teaches courses in social media analytics and artificial intelligence. His email address is john.piorkowski@jhuapl.edu.