Developing a Scalable Data Analytics Pipeline for DoD AI/ML Applications

Anastacia MacAllister
General Atomics
San Diego, CA
Anastacia.Macallister@ga-asi.com

Patrick Rupp
Lockheed Martin Corporation
Fort Worth, TX
patrick.h.rupp@lmco.com

Daniel Javorsek, Louis Dube
Air Force Operational Test and Evaluation Center
Nellis Air Force Base, Nevada
daniel.javorsek@us.af.mil, louis.dube@us.af.mil

ABSTRACT

Artificial Intelligence (AI) and Machine Learning (ML) are the new must have technologies for both commercial industry and the Department of Defense (DoD). The budget and number of projects related to AI/ML in the DoD is continually increasing, with a 50% jump in 2022 alone. In addition, the National Security Commission on AI recommends doubling research funding for AI/ML until it reaches $32 billion in 2026. With this funding drive, the stakes for this technology development are high. Failing to properly develop AI/ML capability can hamper readiness for decades and creates the risk of falling even further behind near peer competitors in the space.

To ensure the technology delivers, government and industry developers need to learn from previous projects in the fast-paced space to develop sound AI/ML implementation strategies. Unfortunately, more attention is paid to flashy new AI/ML algorithms or computing enhancements rather than sound deployment fundamentals. Ultimately, algorithms and hardware are only tools and are a small part of the whole process. Unglamorous topics like data and process governance are arguably more important to ensuring success. Data/process governance encompasses topics like data collection, data cleaning, data formatting, data storage, feature extraction, and configuration management. These pieces create the underlying infrastructure or plumbing required to build robust and tactically relevant AI/ML in the real world rather than the lab. This paper starts by describing what data governance is and how it can be applied to phases in the data lifecycle found in AI/ML applications. From there, a series of examples and case studies are provided for each of these phases describing the successful or unsuccessful results of data governance structures. The paper ends with a discussion of lessons learned and how they can be applied to future AI/ML projects in the DoD to ensure successful AI/ML deployment.

ABOUT THE AUTHORS

Anastacia MacAllister, Ph.D., is an AI/ML Solutions Architect for General Atomics Aeronautical Systems. Her work focuses on prototyping novel machine learning algorithms, developing machine learning algorithms using sparse, heterogenous or imbalanced data sets, and exploratory data analytics. Throughout her career she has provided key contributions in a number of interdisciplinary areas such as prognostic health management, human performance augmentation, advanced sensing, and artificial intelligence aids for future warfare strategies. This work has received several awards and commendations from internal technical review bodies. Dr. MacAllister has published over two dozen peer reviewed technical papers, conference proceedings, and journal articles across an array of emerging technology concepts. Dr. MacAllister holds a bachelor's degree from Iowa State University (ISU) in Mechanical Engineering. She also holds a Master's and PhD from ISU in Mechanical Engineering and Human-Computer Interaction.

Patrick Rupp is a Data Engineer at Lockheed Martin's Skunk Works®. Mr. Rupp specializes in developing data pipelines particularly for use in analytics problems ranging in size from small prototypes to enterprise solutions. Mr. Rupp is currently developing data pipelines for various AI/ML programs and has previously supported Lockheed's work in Defense Advanced Research Projects Agency's (DARPA) Systems of Systems Integration Technology and Experimentation (SoSITE) program and Air Combat Evolution (ACE). Mr. Rupp received his Masters in Data Science from Johns Hopkins University.

Daniel Javorsek, Ph.D., is the Commander of Detachment 6, Air Force Operational Test and Evaluation Center, Nellis Air Force Base, Nev., and Director, F-35 U.S. Operational Test Team. AFOTEC's Detachment 6 plans, conducts, and reports on realistic, objective, and impartial operational test and evaluation of fighter aircraft. The detachment evaluates the operational effectiveness, suitability, and mission capability of the A-10, F-16, F-15C/E/EX, F-22, F-35, and Next Generation Air Dominance (NGAD) aircraft, and reports results in support of major acquisition program milestone decisions and combatant command fielding decisions.

Louis Dube is the F-35 Division Chief of Detachment 6, Air Force Operational Test and Evaluation Center, Nellis Air Force Base, Nev., and Chief Engineer, F-35 U.S. Operational Test Team. The F-35 Division of AFOTEC's Detachment 6 plans, conducts, and reports on realistic, objective, and impartial operational test and evaluation of the F-35 multi-role fighter aircraft. As the Technical Expert for the F-35 U.S. Operational Test Team, Mr. Dube provides technical advisorship and oversight of on tri-service operational test and evaluation designed to accurately represent modern joint doctrine concepts. Mr. Dube earned his Masters in Aerospace Engineering from University of Nevada, Las Vegas.

Developing a Scalable Data Analytics Pipeline for DoD AI/ML Applications

Anastacia MacAllister
General Atomics
San Diego, CA
Anastacia.Macallister@ga-asi.com

Patrick Rupp
Lockheed Martin Corporation
Fort Worth, TX
patrick.h.rupp@lmco.com

Daniel Javorsek, Louis Dube
Air Force Operational Test and Evaluation Center
Nellis Air Force Base, Nevada
daniel.javorsek@us.af.mil, louis.dube@us.af.mil

INTRODUCTION

Industry and the Department of Defense (DoD) have invested billions of dollars into algorithms and hardware associated with building Artificial Intelligence (AI) and Machine learning (ML) systems (Keller, 2021; Harper, 2021). This interest is driven by AI/ML's potential to revolutionize the battlespace, touching on everything from sensor scheduling (Woodward, 2016) and loyal wingmen maneuvers (Pope, et al., 2021) to battle management (Schneider, et al., 2021; Matsumoto, et al., 2021; MacAllister, et al., 2021) and wargaming (Avila, et al., 2021). However, projects that aim to apply this potentially revolutionary technology have a mixed track record of success in both industry and the DoD (Siegel, 2022; Visnjic, 2022; Sheppard, 2020). With this high level of investment and the need to hold off competitors such as China in the AI/ML arms race, the stakes for these projects are high. Looking at past lessons learned and establishing best practices from these is important to ensure the US does not waste valuable time in the AI/ML capability development race.

One such critical lesson is looking at the AI/ML development pipeline as a holistic system where the pieces work together to make development, testing, and deployment successful. Figure 1 shows an overview of the pieces commonly found in a holistic AI/ML deployment pipeline. Currently, more attention is directed to flashy new algorithms or new computing hardware which represents only a small fraction of the whole AI/ML development process. In reality, the algorithms and hardware are a small piece of the entire system (Sculley, et al., 2015). The early data collection and conditioning phases shown in Figure 1 are often the most critical piece for successful AI/ML model development and deployment. However, sound data governance practices during these early stages are often lacking.

Too often in unsuccessful AI/ML development projects, the team starts by asking what model can be built to solve a problem, overlooking the fidelity and suitability of data collected. This resulting haphazard data collection and storage hamstrings ML models because they do not have quality descriptive data, a necessity for many contemporary ML systems today. As a result, the models perform poorly, and the potential of the technology is not realized, or the reputation of its potential is tarnished. This sets back adoption and ensures the DoD falls further behind in the race to operationalize AI/ML.

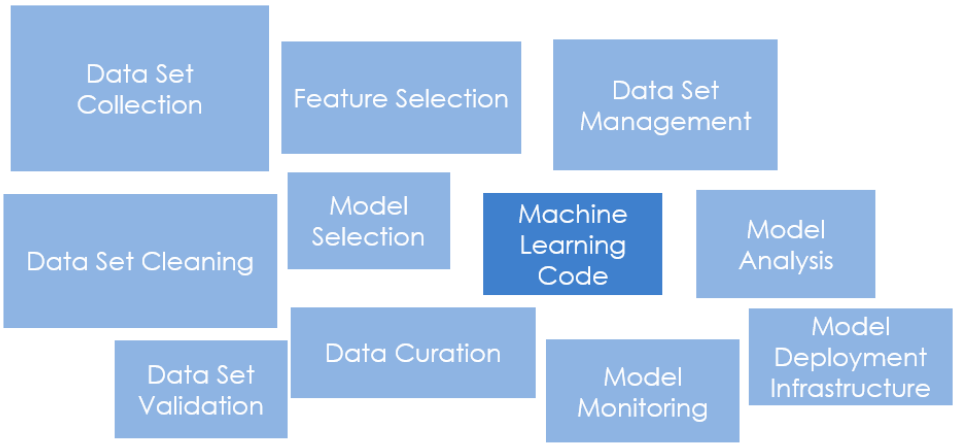

Figure 1. Machine Learning Code is a Small Fraction of System Development and Deployment

To avoid these pitfalls a sound data strategy, often referred to as data governance, is required, before model creation begins. Data governance practices often cover topics like data collection, data cleaning, data formatting, data storage, feature extraction, and configuration management. These

pieces create the underlying infrastructure, or plumbing, required to build robust and tactically relevant AI/ML in the real world rather than the lab. Unfortunately, many organizations do not develop good governance practices surrounding these steps, reducing the quality of data. Thankfully, there exist strategies that can be employed to help manage the data and ensure its suitability for model development tasks, but the DoD needs to start thinking about how they can be implemented in the unique existing DoD culture and computing frameworks. By leveraging lessons learned from past projects and best practices found in literature, the community can start a discussion about how data should be governed in the DoD for operationally relevant, large-scale AI/ML deployment.

This paper begins by discussing the concept of data governance and how that maps to the six common phases of the data life cycle for AI/ML capability development. It then then moves into a case study describing how these data life cycle phases were executed on the F-35 program and how the adherence or nonadherence to sound governance practices impacted data suitability for AI/ML. Lastly, the paper discusses how these shortcomings have impacted the deployability of AI/ML in the DoD and what policies should be implemented to help ensure quality data for future use cases. Ultimately, this work helps shape the discussion around data and process policy for the unique needs of the DoD, helping maximize the effectiveness of AI/ML for the warfighter.

BACKGROUND

While the amount of data in the DoD has increased exponentially in the last two decades, the quality of data for AI/ML systems has not. Due to a lack of emphasis on sound data practices, the DoD is not fully ready to take advantage of AI/ML. To make future autonomous warfighting concepts a reality the DoD needs to develop sound policies, standards, and management practices around data. This section begins by describing the concept of data governance and how key themes from the discipline can facilitate successful AI/ML deployment.

Data Governance – Establishing Processes, Procedures, and Implementation Strategies

The concept of data governance focuses on establishing frameworks of rules, responsibilities, and decisions required within an organization to ensure data are useful for accomplishing desired tasks (Alhassan, Sammon, & Daly, 2016; Khatri & Brown, 2010). Data governance has become increasingly more important throughout the early 21st Century. One factor of this growth in importance stems from the usage of large quantities of data for AI/ML applications. One article from the International Journal of Information Management shows the number of publications on data governance has substantially increased from just two per year in 2004-2006 to over fifteen on average from 2013-2019 (Abraham, Schneider, & vom Brocke, 2019). Specifically in the DoD, the Chief Information Officer (CIO) released a paper describing the DoD's net-centric data strategy which outlines the vision for managing data by "empowering users through access to data and faster availability of data as a result of posting before processing" (Stenbit, 2003). The CIO's report mentioned implementing data governance to meet the needs of the net-centric data strategy demonstrates an increased awareness of data governance concepts. However, while the importance of these concepts are being increasingly recognized, data governance is still an emerging field, requiring careful case study and analysis of the application before attempting to stand up a governance structure.

While there is no universal standard for the design and implementation of a data governance framework, many authors generally see it as a way to define rules, policies, and responsibilities around data (Alhassan, Sammon, & Daly, 2016). One conceptional framework defines three major components impacted by the governance mechanism: data scope, organizational scope, and domain scope (Abraham, Schneider, & vom Brocke, 2019). Additionally, another framework centered around the problem of big data offers these components of its data governance framework: communication and data management, identification of organizational structure, stakeholder selection, data storage, big data scope determination, measurement and monitoring quality, optimization and computation, and policies and standards initiation (Al-Badi, Tarhini, & Islam Khan, 2018).

While all of these different frameworks offer a unique perspective on data governance, there are several underlying themes and commonalities that particularly align with the needs for maximizing the value of data within the DoD for AI/ML. Common themes include: 1) identification of stakeholders roles and responsibilities to ensure data are suited to needs; 2) organizational structures for data creation, data storage, and data access; 3) establishing a clear understanding of long term data storage and refreshing requirements.

The first common theme centers around the identification of stakeholder's roles and responsibilities to ensure data are suited to their needs. Without understanding who the stakeholders are, the end goal of using the data cannot be properly grasped. This can result in poor collection and documentation outcomes because key pieces of information could be omitted. For example, in the DoD an emphasis is being placed on complex joint all domain warfare programs, and in order to use data as a strategic asset for these programs, stakeholders must be chosen to align the use of data to their programmatic strategy. However, stakeholders must also believe data can provide value to the organization, otherwise stakeholder support maybe lacking, leading initiatives to fall short.

The second common theme has to do with organizational structure or how different components within the process work with one another. Often in the DoD the organizational structure is defined by the command structure. To achieve data governance success within the DoD, the organization must implement a chain of command utilizing a structure that ensures successful creation of data, storage of data and access to data at different responsibility levels. This can be challenging due to communication and classification barriers between different people throughout the process. Since the DoD handles information at varying levels of classification, ensuring data accessibility and security while designing AI/ML algorithms will be a crucial aspect of a data governance strategy. Data accessibility must be properly balanced between ease-of-access for cross domain and cross functional efficiency while protecting need to know information. The organization must also emphasize data quality to ensure that the information at all levels is complete, usable, and trustable. Additionally, the second point regarding roles and responsibilities adds substance to the organizational structure by defining the roles needed within the organization. These roles need to have clearly defined responsibilities with the appropriate decision capabilities to ensure the implementation of the data strategy.

Finally, the third aspect of data governance themes identified is establishing a clear understanding of long-term data storage and refreshing requirements. Counterintuitively, data often has a shelf life where it can become stale and depends heavily on its intended use. For example, data from pre-pandemic supply chains was of limited utility in a post COVID world. As a result, that data might need to be either put in long term storage or deleted to free up space for new data collects. To ensure that the data on hand is representative of the situation trying to be modeled, governance strategies surrounding longitudinal data storage and deletion need to be developed.

Ultimately, data governance is a topic that touches on many cross-cutting organizational roles. Governance structures will impact everything from operational testing squadrons to the end user. As a result, the DoD needs to think interdisciplinary when designing and building the infrastructure for deploying AI/ML systems.

Data Governance Applied to the AI/ML Data Life Cycle

Currently, the majority of DoD AI/ML solutions are reliant on data driven models. As such, sound data governance practices are an extremely important part of successfully developing this emerging technology for 21st century warfare. The data governance concepts of operational organization, stakeholder identification, and long-term data use strategies detailed above cross into many parts of the data lifecycle when developing AI/ML models. Figure 2 shows the data lifecycle. It depicts the stages that data will exist in throughout its life. In order to ensure data governance is sufficiently implemented, one must consider the roles & responsibilities, data accessibility & security, data quality, stakeholders and strategy with relation to every phase within this life cycle. The entirety of data's scope must be governed to ensure risk mitigation while maximizing the potential value of the data for the organization and most importantly the

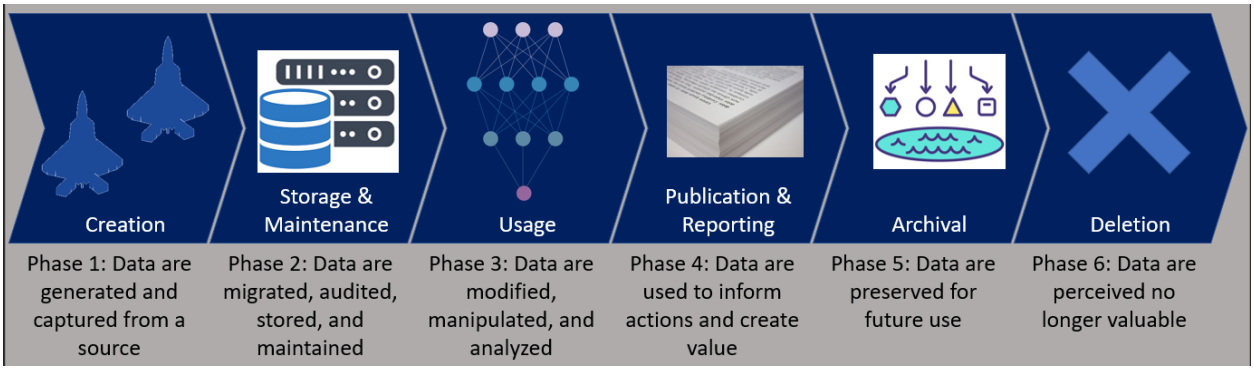

Figure 2. Six Phases of the Data Life Cycle

warfighter. Success or lack of success developing and implementing such a strategy can have a direct impact on the effectiveness of promising new AI/ML methods.

To recapitulate, robust data governance processes ensure that data collected for AI/ML in the DoD will be useful for building desired algorithms. Unfortunately, the DoD is not always applying best practices when it comes to data governance throughout data lifecycles. The sections below describe data collection for different lifecycle phases on the F-35 program and describe if strategies implemented were successful or lacked the proper data governance processes for meaningful AI/ML development.

F-35 CASE STUDY AND ANALYSIS

The F-35 flight test enterprise provides a compelling case study for the importance of data governance and its core concepts. The F-35 Integrated Test Team (ITT) serves as a tri-service umbrella for the developmental and operational test efforts performed in pursuit of verification, validation, and evaluation objectives. While the F-35 ITT serves a myriad of functions, its most important product is the flight test data it generates in pursuit of these objectives. Flight test data are collected in six geographical locations, encompassing over forty aircraft in multiple software and hardware configurations, all equipped with varying levels of instrumentation. Additionally, several simulation and laboratory environments funnel data to and from the flight test enterprise to verify and troubleshoot anomalies identified during test operations. This data collection apparatus has recently been implemented to increase collection across other non-test flying units. Ultimately, all of this information is leveraged by engineers and analysts to identify anomalies, evaluate system performance, characterize new capabilities in both controlled and uncontrolled environments, and develop and demonstrate new tactics and best practices. Meanwhile, the greater F-35 enterprise has its eyes set on AI/ML concepts that could further refine the lethality and survivability of the weapon system, but significant data governance challenges must be solved before this vision can be realized.

The sections below describe the data lifecycle for the F-35 and include the relative success of the employed data strategies. The overview of the data lifecycle for the F-35 case study is shown below in Figure 3. Green outlined phases indicate successfully implemented data governance best practices and red indicates where improvement is needed. In what follows, the first phase covers successful data creation methods employed by the developmental and operational test communities. The second phase describes data storage and access including how that system has become fragmented. The third phase covers how the data are being used and how lack of contextual information hampers meaningful analysis. Finally, the fourth phase covers how data are used to generate reports and the final two phases which cover the long-term storage and deletion, still need to be developed.

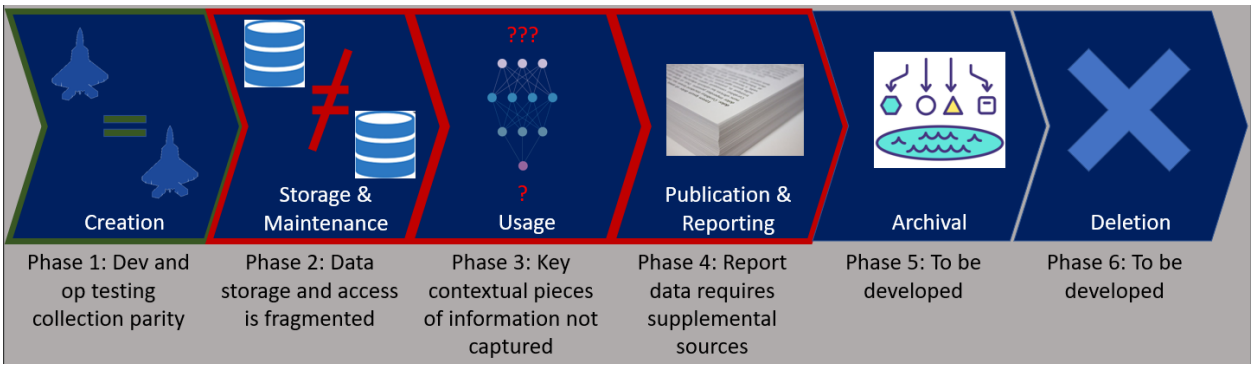

Figure 3. F-35 Case Study - Data Management Best Practices Results by Phase

F-35 PHASE 1: Flight Test Data Creation

The first step in the process is flight data creation. Every flight test effort is conducted in pursuit of some test objective, and some type of data collection occurs on each sortie. In the case of the F-35, that test data could originate from a myriad of sources: telemetered or internally captured own-ship network traffic, recording of on-board displays, off-board digital and analog data sources, and qualitative data such as pre- and post-mission surveys. Of particular interest is the capture of own-ship network traffic; Lockheed Martin and the U.S. government independently developed

instrumentation solutions to enable low- to mid-level own-ship data collection in response to familiar yet dissimilar test, and therefore data collection, requirements. While developmental testing required high fidelity data collection and near-complete bus data capture, operational testing accepted a lower fidelity solution to maximize proliferation of unobtrusive data recording across its sizable fleet of aircraft. The high-level differences between the two collection strategies are shown in Figure 4. This difference in data needs by the operational and developmental functions within the DoD have led to a fragmentation of data fidelity, however, the DoD employed sound data governance practices across both areas. This resulted in a strong set of data collection practices that allows for robust data collection for AI/ML algorithms depending on the application.

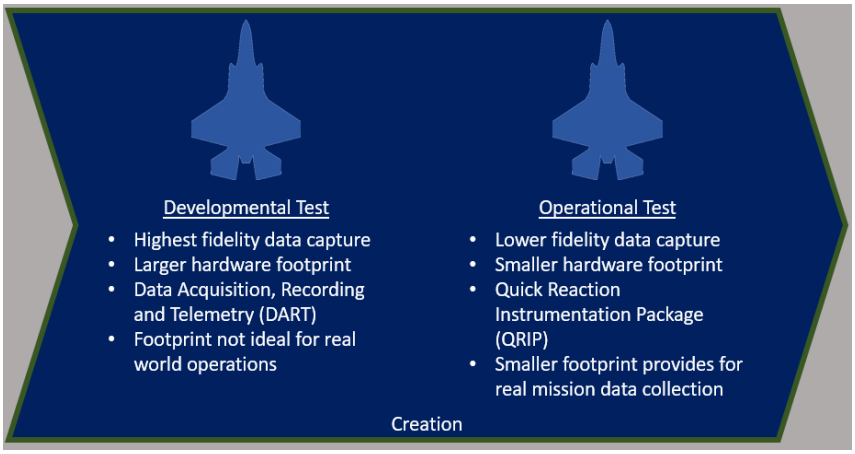

Figure 4. Developmental vs Operational Test Data Collection

The developmental testing community often collects data at a higher fidelity because of their traditional data analysis needs. For the F-35 program they used the Data Acquisition, Recording and Telemetry (DART) instrumentation. The DART was a podded solution carried in one of the weapon bays, therefore taking up a weapon station. The DART provided nearly full bus capture for several hours at a time, the ability to record multiple video feeds simultaneously, and advanced telemetry capabilities. This means it can collect high fidelity detailed data well suited to the needs of various AI/ML models. Despite the DART being the preferred instrumentation solution for development purposes, it was less favored in the operational test community because its internal carriage and intrusive modification requirements made the aircraft less operationally representative.

The operational testing community, however, often collects more coarse grained data due to their interest in overall mission outcomes rather than detailed performance metrics. To better meet the demands of the growing operational test fleet, the U.S. government developed the Quick Reaction Instrumentation Package (QRIP). The QRIP provided less recording capacity and bandwidth than DART in exchange for a smaller footprint in the weapons bay, freeing up a weapon station. Additionally, since QRIP was developed several years after DART had entered service the development benefited from advances in storage and digital input/output technology to produce a smaller overall package that could still meet some of the requirements of the developmental test community. Additionally, QRIP is a key enabler in the plan to instrument non-test aircraft, with the objective to collect large volumes of data to augment the limited sample sizes produced by the test enterprise.

One of the key successes emerging from the development of QRIP after DART was the continuation and consistency of the requirements from which these data collection methods originated, and the method of collection they share. In order for the data collected on QRIP aircraft to be compatible with data collected on DART-equipped aircraft, the collection methods had to be similar. Both instrumentation solutions offer flexibility to their associated engineering teams in setting up parameter lists to be recorded. They use similar formatting that make synchronization of recording efforts relatively seamless, and the output formats follow the same industry standard Chapter 10 format. On multiple occasions, the F-35 ITT has leveraged mixed instrumentation configurations for multi-ship testing, interleaving developmental and operational test aircraft seamlessly in pursuit of complex test objectives, collecting data sets that would otherwise be unattainable.

The very successful data creation efforts of DART and QRIP show that with proper data governance the growing F-35 database represents a remarkable opportunity for the application of AI/ML techniques. Additionally, when replicated on other aircraft (F-22, B-2, B-21, NGAD, etc) or even on weapons, ships, and satellites the truly astounding potential would exist to influence battle management and battlespace shaping. This data creation success enables AI/ML applications all the way from the sensor to the battle management level. In the end, leveraging a common set of variables and data collection procedures for the developmental and operational test community highlights that DoD

can successfully apply robust data collection strategies, allowing them to leverage data collected to build AI/ML models.

F-35 Phase 2: Data Storage & Maintenance

While the F-35 Phase 1 is considered a glowing success, Phases 2 and 3 are the exact opposite and represent a great example of an opportunity for better adoption of data governance. The on-going success of the DART and QRIP instrumentation solutions is in stark contrast to the data accessibility challenges currently plaguing the F-35 test enterprise which represent a direct symptom of a highly fragmented data sharing infrastructure. For example, at Edwards Air Force Base, data collected on developmental aircraft is placed on one classified network to which many, but not all, stakeholders have access. Meanwhile, data collected on operational test aircraft, which are parked only a hundred yards away from their developmental test counterparts, are uploaded to a completely separate network, to which a different subset of stakeholders have access. Data transfers from one network to another occur by exception instead of being the norm in an extremely manual process. And still, data collected on other operational test aircraft just three hundred miles away, at Nellis Air Force Base, are uploaded to yet another, completely separate classified network. While the barriers between developmental and operational test are being knocked down through various initiatives adopted by the F-35 ITT, the data infrastructure that should promote accessibility of F-35 data to all stakeholders instead presents significant logistics and usage challenges for the entire test enterprise. These challenges greatly complicate even rudimentary data analytics and make the adoption of data hungry AI/ML techniques virtually impossible.

Equally problematic as network fragmentation is the movement of data across the enterprise. As the number of collection sources increases, so does the need to move large volumes of data from one location to another in a timely fashion. Advanced instrumentation solutions such as DART and QRIP can produce hundreds of terabytes of data each week. The majority of the test enterprise relies on movement of data to high-performance workstations wherever analysis is occurring. Alternatively, to address bandwidth concerns on classified transport networks, the F-35 Joint Program Office (JPO) partially adopted and implemented an edge computing network called the F-35 Knowledge Management (KM), which is one component of a broader hardware and software strategy being developed by the Test Resource Management Center (TRMC). A properly provisioned edge computing solution would enable analysis to be performed at the location of data collection via a remote connection. Unfortunately, deployment of the KM across the F-35 enterprise has been marred by funding shortfalls, contracting difficulties, and challenges associated with migrating legacy analysis capabilities to a new environment. Although the arrival of KM was expected to unify the F-35 test enterprise, poor execution, lack of follow-through in the deployment of the network, and inadequate funding actually accomplished the opposite and instead further fragmented the F-35 data infrastructure. This highlights an important but often overlooked challenge of AI/ML implementation since failure to accurately perform Phase 2 of the data lifecycle significantly limits the potential to capitalize on what the collected data can do.

In the flight test community, subject-matter expertise tends to be centralized, which is in stark contrast to the highly federated nature of the F-35 test enterprise. To maximally leverage the subject-matter expertise resident with each organization that makes up the F-35 ITT and further enable the weapon system's developer to improve it, the flight test data must also become centralized. This is the only way to ensure maximum accessibility across the stakeholders and enable future applications of AI/ML techniques. While AI/ML concepts are becoming increasingly widespread across the DoD, the flight test world is working through the adoption of data science, the increased focus on ensemble statistics generated from large volumes of data, and AI/ML in pursuit of flight test data analysis and optimization of weapon system performance. It is only a matter of time until AI/ML work of high significance will be performed on F-35 flight test data, but that endeavor will fail unless a complete data set is readily accessible to all stakeholders.

F-35 Phase 3: Usage & Contextualization

Assuming the F-35 test enterprise can ensure centralization and accessibility of its flight test data, there is one important factor that must be addressed in both the creation and storage phases of this data's life cycle: contextualization. Because flight test data are collected in pursuit of test objectives, the context and manner in which the data are collected is necessary to compile a complete and useful dataset. Metadata or secondary data sources--pieces of information related to but not recorded by the primary instrumentation sources--can often unlock the meaningfulness of data recorded by an F-35 during a test. Additionally, as aircraft and weapon systems further

integrate with each other, the completeness of perceived truth becomes entirely dependent on the enterprise's ability to amalgamate all pertinent data sets in a singular environment.

To illustrate this concept, consider a single F-35 which fuses multiple on-board sensors' outputs to present a single air and ground picture to the pilot on the tactical situation display, as shown in Figure 5. Then, let us define the concept of truth completeness as it applies to the information displayed to the pilot. In this case, a perfect presentation of all air and ground assets detectable by the F-35's sensors to the pilot would be devoid of Type I and Type II errors. To evaluate the truth completeness presented by the F-35 to the pilot, an analyst would need a complete understanding of all airborne and ground assets within sensor range during a given test. This data are actually available to testers, but its collection and its accessibility have not been standardized. Instead, this type of data are collected on a case-by-case basis and the analysis process associated with making truth completeness evaluations is complex and time-consuming. Incidentally, this type of analysis and many others are rich in opportunity for AI/ML algorithms, but the flight test enterprise must ensure complete alignment of data collection objectives, and this data collection extend beyond the confines of data collected on the test article itself.

Figure 5. Usage and Contextualization Challenges

Furthermore, metadata and secondary data sources must be afforded the same level of accessibility as primary data sources. In some cases that might mean overcoming classification hurdles, but at a more fundamental level the storage and dissemination methods must be flexible enough to accept and organize dissimilar or new data sets in a timely manner. Recalling TRMC's KM initiative, a key software enabler to these concepts is the Cloud Hybrid Edge-to-Enterprise Evaluation & Test Analysis Suite (CHEETAS). CHEETAS is a TRMC-developed tool that facilitates a variety of functions tied to edge-computing solutions such as querying of local data sets, and cross-site querying of simple or complex data sets. Additionally, CHEE TAS directly addresses the subject of metadata and secondary data sets, as it enables the creation of localized metadata schemas that support an enterprise's needs, along with the capability to ingest secondary data sources, perform necessary transformations automatically, and make those secondary data sets available alongside the primary data sources. The implementation of software solutions such as CHEETAS, coupled with centralization of data sets, are key enablers to big data analytics and the deployment of AI/ML in the test world.

F-35 Phase 4: Publication & Reporting

The readiness of a weapon system such as the F-35 can be attributed to a variety of factors. For example, software stability can greatly affect the effectiveness and availability of the aircraft. The concept of software stability can be decomposed into a set of definitions which can then be applied to adjudicate data collected during flight test to provide a quantifiable evaluation of software stability, from one software version to the next. Assuming that collection methods are compatible from one source to another, the primary concern then becomes the consistency in which the usage and publication standards are applied. The F-35 software stability metrics have been tracked weekly for over seven years, and while minor changes have been made to the adjudication or usage standards, this body of publications is a useful litmus test on the F-35 software's health that has stood the test of time.

Historically, the entirety of software stability data was derived from the developmental test aircraft flight hours and the assumption was that the 7 to 10 aircraft in its fleet were representative of the operational fleet. However, as the software development timelines were shortened to achieve an increased tempo in Continuous Capability Development and Delivery (C2D2), the developmental test fleet generated fewer flight hours with each software than previous, therefore increasing the uncertainty in the software stability metrics from developmental test aircraft alone. Coincidentally, these accelerated software development timelines drove early involvement from operational test organizations to assist in developmental efforts. With over triple the aircraft flying much more often, the operational

test fleet generated flight hours at a much greater pace than the developmental test. Given these constraints, the F-35 enterprise requested that the operational test fleet augment the software stability data pool with its own data set.

The life cycle of F-35 software stability data, as well as the combined developmental and operational test publication and reporting of those results, are great examples of homogeneous data governance across dissimilar organization in pursuit of a common goal. By clearly identifying creation standards and usage protocols (i.e. adjudication parameters), the operational test data set was merged with the developmental test one, augmenting the statistical significance of the inferences published by the various stakeholders. These more robust usage and publication standards secured the potential for more advanced analytics later down the road.

F-35 Phase 5 and 6: Archival, Long-Term Storage, and Deletion

As acquisition programs like the F-35 take action to shorten development timelines, data governance infrastructures that support continuous evaluation concepts will be of increasing importance. Instead of the serial nature of developmental and operational test, an integrated test approach in which weapon system performance, effectiveness, survivability and suitability is demonstrated on a continuous basis must be pursued. To be successful, the continuous evaluation concept necessitates a consistent data collection protocols, prolonged data storage for comparison purposes, and the standardization of usage and publication standards. The software stability metrics generated on the F-35, and the adaptability of its adjudication process, are examples of this paradigm shift. In this example, it takes little imagination to plot a course for integration of AI/ML in a nearly immediate sense, provided appropriate data governance is in place.

Ultimately, the collected data on old versions of the software will become out of synch with the fielded aircraft, inherently limiting its utility. In a resource constrained environment, judgement decisions will need to be made on archival and deletion. As part of the governance process it is important to identify the risk acceptance authority associated with data deletion since even old data may have potential applications that go beyond the originator's vision.

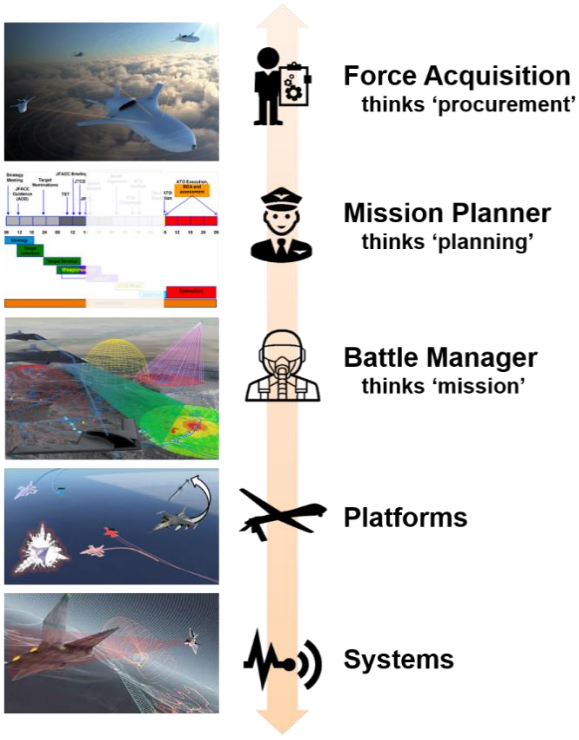

Figure 6. AI/ML applications in DoD exist from the sensor to the battle shaping levels.

DISCUSSION AND PATH FORWARD

There have been several recent examples of successful isolated applications of AI/ML and data analytics for DoD (see Figure 6) and in each case they relied on one of DoD's most precious, but undervalued, resources—Data. From predictive maintenance and logistics (Jimenez, Schwartz, Vingerhoeds, Grabot, & Salaun, 2020) to air combat (Pope, et al., 2021) and battle management (Schneider, et al., 2021; Matsumoto, et al., 2021) the potential is clearly recognized by department leadership. In fact, it is now impossible to imagine a Pentagon briefing without AI/ML pixie dust sprinkled throughout the presentation. However, we must shift the focus from the current fixation on the algorithms alone and address the pressing need for sound governance and infrastructure management to scale and deploy this promising technology.

Development of sound data governance strategies is becoming even more important with the joint and allied nature of the F-35. This is due to the need for the aircraft to integrate into a diverse battlespace to share information, a fact that should help catalyze the development of a scalable data analytics pipelines (Rupp, et al., 2021). This pipeline and its associated data governance will apply to everything from modeling and simulation as is currently being developed in the Joint Simulation Environment (JSE)

(Menke, 2019), and live, virtual, constructive execution such as the Live Mission Operations Center (LMOC) (Gao, Zhang, Zhou, & Lu, 2021). Although our F-35 case study focused on flight test data, the lessons learned regarding the need for data accessibility, contextualization, and standards is extensible far beyond any single platform.

One example of unintended future usage of data collected during F-35 flight test is the Defense Advanced Research Projects Agency (DARPA) Air Combat Evolution (ACE) program. The ACE program was designed to build trustable, scalable, human-level autonomy for air combat. As a result, the program built trust in AI/ML algorithms at the single platform level with the intent to maintain custody of that trust as the scenario complexity was increased to campaign scales (Javorsek, 2019). The program went to great length to design and build a data analytics framework (Albarado, et al., 2022) based on very rich and diverse datasets including over a decade of Large Force Exercises (LFE) such as Operation Red Flag (Berger, 2005). These were from both Nevada and Alaska and implemented AlphaMosaic AI/ML agents from several different contractors. Although the Red Flag data included exquisite position and shot information the datasets lacked the contextualization associated with both the objectives of the individual LFEs and the engagement outcomes. This is particularly important for reinforcement learning in which reward shaping and feedback is a critical aspect of the training phase of their development. The lack of good contextualization in the data severely limited its use for the ACE program, demonstrating that not following good data governance practices (such as collecting and storing auxiliary contextualizing information during tests) can have far reaching consequences. Ultimately, the lack of good data governance practices will limit the DoD's ability to use data collected for developing critical future AI/ML warfighting technology. As a result, leaders in the DoD and industry must start thinking about developing good governance strategies throughout the data lifecycle phases.

CONCLUSION AND FUTURE WORK

In the Department of Defense, new acquisitions and upgrades that get funded ultimately come down to those that provide the largest increase in combat capability. For our dated twentieth century mindset, and its accompanying acquisitions system, it is easy to show how a new aircraft, hypersonic weapon, or enhanced sensor increases combat capability. It is also then relatively straight forward to procure funding for big promises made in glossy brochures and PowerPoint because these types of items are fun to talk about at conferences and artist renditions make great additions to catchy media headlines. Unfortunately, in the evolving 21st century battlefield the true combat capability will be borne from investments in data infrastructure and governance, as opposed to the more contemporary model fixated on widgets.

As shown by the sections above, data governance is an evolving and important concept that needs to be applied to the data lifecycle for the DoD to increase the probability of success for AI/ML capability development. The examples provided show how the F-35 program has a mixed history of success following sound data governance strategies for different phases in the data lifecycle. This ultimately hampers not only current AI/ML capability development, but future warfighting concepts from subsystem resource managers and autonomous agents for maneuver to battle management command and control. In the end, until combat capability is redefined and both DoD and industry begin thinking more seriously about data governance, progress will be limited. Moving forward, it is up to both industry and DoD to devote resources to solving governance challenges and to applying the lessons to help ensure successful development of AI/ML capabilities for the warfighter.

ACKNOWLEDGEMENTS

The opinions, findings, conclusions, or recommendations expressed in this article are those of the authors and do not reflect the policy or position of the U.S. Air Force, the U.S. Department of Defense, or the U.S. Government.

REFERENCES

Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, 424-438.

Al-Badi, A., Tarhini, A., & Islam Khan, A. (2018). Exploring Big Data Governance Frameworks. *The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks* (pp. 271-277). Muscat: Elsevier Ltd.

Albarado, K., Conduti, L., Aloisio, D., Robinson, S., Drown, D., & Javorsek, D. (2022). AlphaMosaic: An Artificially Intelligent Battle Management Architecture. *Journal of Aerospace Information Systems, 19*(3), 1-11.

Alhassan, I., Sammon, D., & Daly, M. (2016). Data governance activities: an analysis of the literature. *Journal of Decision Systems*, 64-75.

Avila, A. M., Fonoberova, M., Hespanha, J., Mezic, I., Clymer, D., Goldstein, J., . . . Javorsek, D. (2021). Game Balancing using Koopman-based Learning. *2021 American Control Conference*, (pp. 710-717).

Berger, A. (2005, Summer). Beyond Blue Four: The Past and Future Transformation of Red Flag. *Air & Space Power Journal, 19*(2), 43-54.

Gao, Y., Zhang, Y., Zhou, X., & Lu, H. (2021). Overview of Simulation Architectures Supporting Live Virtual Constructive (LVC) Integrated Training. *2021 6th International Conference on Control, Robotics, and Cybernetics* (pp. 333-338). IEEE.

Harper, J. (2021, Feb 10). *Budget*. Retrieved from National Defense: https://www.nationaldefensemagazine.org/articles/2021/2/10/federal-ai-spending-to-top-$6-billion

Javorsek, D. (2019). Air Combat Evolution (ACE). *DARPA-BAA-HR001119S0051*, (pp. 1-24).

Jimenez, J., Schwartz, S., Vingerhoeds, R., Grabot, B., & Salaun, M. (2020). Towards multi-model approaches to predictive maintenance: A systematic literature survey on diagnostics and prognostics. *Journal of Manufacturing Systems 56*, 539-557.

Keller, J. (2021, June 4). *Computers*. Retrieved from Military + Aerospace: https://www.militaryaerospace.com/computers/article/14204595/artificial-intelligence-ai-dod-budget-machine-learning

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, 148-152.

MacAllister, A., Rupp, P., Hellstern, G., Garrison, J., Javorsek, D., & Chu, P. (2021). Using Machine Learning for Battle Management Analysis. *Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, (pp. #-#).

Matsumoto, S., Barreto, A., Costa, P. C., Benyo, B., Atighetchi, M., & Javorsek, D. (2021). Dynamic Explanation of Bayesian Networks with Abductive Bayes Factor Qualitative Propagation and Entropy-Based Qualitative Explanation. *2021 IEEE 24th International Conference on Information Fusion (FUSION)*, 1-9.

Menke, T. (2019). Joint Simulation Environment for United States Air Force Test Support. *WFAFB: NATO S&T Organization*.

Pope, A. P., Ide, J. S., Diaz, H., Rosenbluth, D., Ritholtz, L., Twedt, J. C., . . . Javorsek, D. (2021). Hierarchical Reinforcement Learning for Air-to-Air Combat. *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*, (pp. 275-284).

Rupp, P., MacAllister, A., Garrison, J., Hellstern, G., Javorsek, D., & Chu, P. (2021). Developing a Scalable Data Analytics Pipeline. *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, (pp. #-#).

Schneider, M. K., Barbulescu, L., Batlle-Rafferty, L., Cook, M., Kapler, T., Loppie, M., . . . Javorsek, D. (2021). Context-sensitive, distributed, multi-domain adaptive option generation. *Proc. SPIR 11746, Artificial Intelligence and Machine Learning for Multi-Domain Operations Acpplications III*, (p. 1174608).

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., . . . Dennison, D. (2015). Hidden Technical Debt in Machine Learning Systems. *NIPS*.

Sheppard, L. (2020). *Accelerating the Defense Department's AI Adoption.* Arlington: Council on Foreign Relations.

Siegel, E. (2022, Jaunary 17). Retrieved from KD Nuggets: https://www.kdnuggets.com/2022/01/models-rarely-deployed-industrywide-failure-machine-learning-leadership.html

Stenbit, J. P. (2003). *DoD Net-Centric Data Strategy*. Washington, DC: Department of Defense. Retrieved from https://dodcio.defense.gov/Portals/0/Documents/Net-Centric-Data-Strategy-2003-05-092.pdf

Visnjic, A. (2022, May 16). Retrieved from Venture Beat: https://venturebeat.com/2022/05/16/for-ai-model-success-utilize-mlops-and-get-the-data-right/

Woodward, T. (2016). Converged Collaborative Elements for RF Task Operations (CONCERTO). *DARPA-BAA-16-28*, 1-66.