Presented as:
A. Aved, E. Blasch, F. Darema, M. Rangaswamy, "Workshop Report on Dynamic Data Systems," *Presented at AAAI Spring Symposium: Workshop on Deep Models and Artificial Intelligence for Defense Applications: Potentials, Theories, Practices, Tools and Risks*, 2020.

Workshop Report *on* *Dynamic Data Systems*

August 6-7, 2018

Workshop Organizers:

Dr. Alex Aved, Research Scientist, AFRL, Information Technology Directorate
Dr. Erik Blasch, Program Officer, AFRL, Air Force Office of Scientific Research
Dr. Murali Rangaswamy, Senior Advisor, AFRL, Sensors Directorate

Workshop Chair:
Dr. Frederica Darema, Director, AFRL, Air Force Office of Scientific Research & Research Director, Chief Data Office

September 24, 2018

Elements of Data: Visible, Accessible, Understandable, Linked, Trusted (VAULT)

Workshop on Dynamic Data Systems

1. Introduction

1.1 Future Directions and Drivers

Future complex systems (systems-of-systems) supporting dynamic Air Force mission needs, require adaptive data utilization, relevant resource exploitation and management, to redirect and amass the right resources, and make effective decisions at any given time with decision speed. New Science and Technology (S&T) directions provide innovation through basic research and development (R&D) to applied R&D technology transitions and demonstrate that they can create new capabilities, towards such Air Force mission critical needs.

To assess the impact of such new S&T directions, the AF Chief Data Office (SAF/CO) sponsored a workshop that brought-together relevant advanced research and technology development communities to inform on S&T directions and data environments,to support the emerging Intelligence, Surveillance, and Reconnaissance (ISR) and Multi-Domain-Command & Control (MDC2) capability needs of the AF, and more broadly of the DoD. In particular, the scope of the S&T directions that the workshop addressed, included not only data-analytics but also systems-analytics, through emerging dynamically integrated modeling and systems instrumentation methods, exploitation of multiple types of archival and real-time data, and the underlying systems infrastructure computing and data services enabling operational support capabilities in the context of the following key areas of interest to the Air Force: MDC2 and ISR, as well as Predictive Maintenance (PM) and Clouds (IoT), the latter two being key in supporting the broader MDC2 needs.

MDC2 consists of three elements: extremely high-quality situational awareness, rapid decision making, and the ability to direct forces across domains and missions with continuous feedback (General Goldfein, Air Force Magazine, January 2018). MDC2 collects data from a global network of sensors and models rapidly fused into actionable intelligence and shared across all Air Force components.

The two-day workshop included over 60 attendees drawn from academia, industry, and several federal agencies, 20 briefings, and four break-out groups discussions, that sought to identify and prioritize fundamental needs for a data strategy, based on the new S&T directions in the above referenced and other application areas.

1.2 General Background about the Chief Data Office (SAF/CO)

Data is a key organizational asset, but is not optimally managed. Thus, the SAF/CO's role is to help with the realization of value from data across the Air Force enterprise. The SAF/CO is responsible for establishing an enterprise-wide data strategic vision and CONOPs encompassing an information strategy, governance, control, and effective exploitation. The SAF/CO's role will combine accountability and responsibility for information governance,

SAF/CO Vision

The Air Force is a data-driven organization – one that purposely collects, creates, shares, and acts upon trusted data in and across all mission areas and domains – empowering Airmen and the machines they rely upon with timely access to the right data to accelerate effective and decisive action at the speed and scale of operations.

data quality, access and life cycle management, along with the exploitation of data assets to support real-time decision-making, create value and reduce costs. A SAF/CO key tenet for broad data attributes is that they are VAULT[1] compliant.

1.3 Data Architecture Thrust Categories

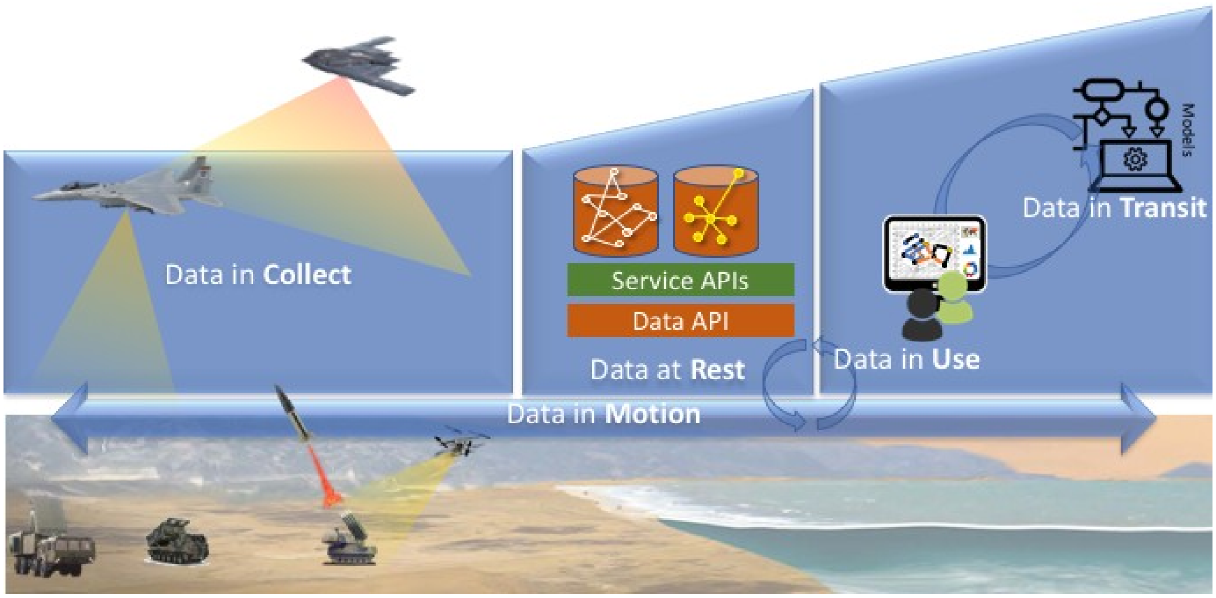


Figure 1: Use-based Data Taxonomy

We group the key areas of data architecture support needs and recommendations along the following thrusts[2]:

- **Data at Rest**: Develop a *Data as a Service (DaaS)* architecture that incorporates contextual information, metadata, and registration;
- **Data in Collect**: Provide *structure (e.g., translations)* between data for integration, analysis, and storage;
- **Data in Transit**: Leverage the *power of modeling* from which data is analyzed for information and delivered as knowledge;
- **Data in Use**: Afford data-based *needs collections* (recommendations) based on dynamic mission priorities, and balanced between need to know and need to share;
- **Data in Motion**: Utilize *feedback control loops* to dynamic ally adapt to changing priorities, timescales, and mission collects.

Figure 1 highlights that for real-time analytics, data preparation requires methods to begin the process for effective and efficient data processing.

[1]VAULT – Visible, Accessible, Understandable, Linked, Trusted; AF/CDO Strategy Document

[2]Blasch, E., et al., "Methods of AI for Multi-modal Sensing and Action for Complex Situations"; AI Magazine (2019)

Based on the requirements of the driving application areas addressed in the workshop and the new capabilities needed, there is a need to address these in the context of the symbiotic relation of systems-data and systems-models approaches. These new directions addressed in the workshop are elaborated in the context of the referenced applications and the data taxonomy above. The top five (R1 – R5) technical recommendations that emerged from the workshop are summarized below, and sub-recommendations for each category are provided in Section 7.1 of this report. In addition, policy and acquisition-driven recommendation sare proposed in Section 7.2 of the report.

- R1: Models interacting with data in a feed-back control loop (dynamically integrated data & models)
- R2: Extend the Multi-Domain MDM (Master Data Management) to include real-time multidimensional[3] and dynamic data, and data from models
- R3: Support enterprise-level federated service-based architecture
- R4: Proactive sharing of data based on need to know with intelligent granularity
- R5: Data for performance analysis and benchmarking

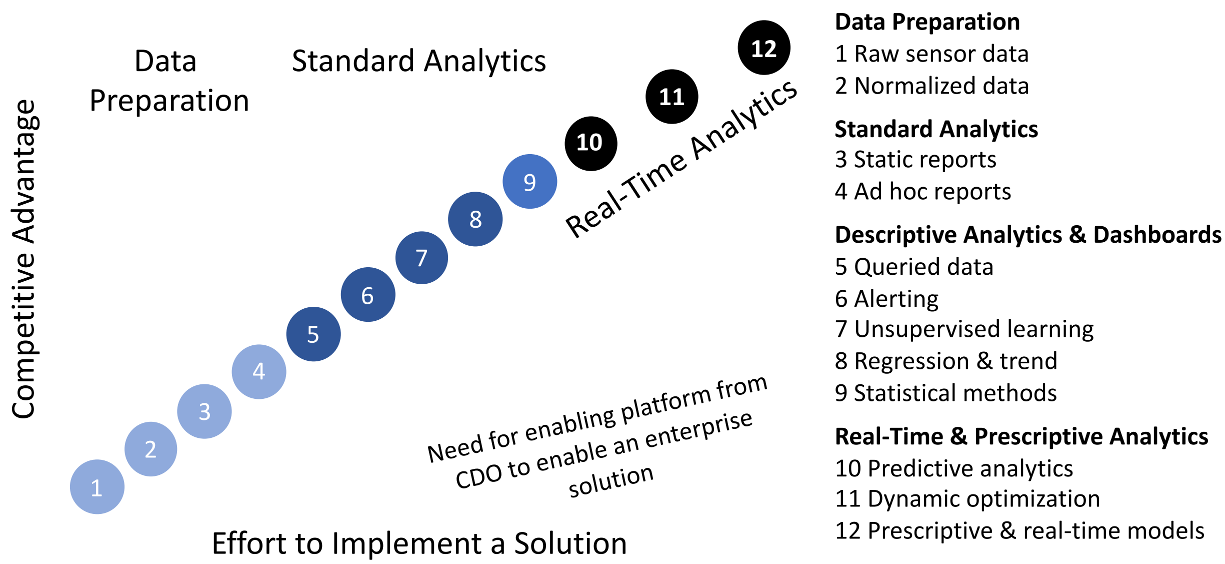


Figure2. Data needs to support real-time analytics

2. Data & Models

There has been increasing emphasis in data and "data analytics", where recent methods of machine learning/artificial intelligence (ML/AI) have demonstrated some success for static data analysis. However, there is a wealth of other ***comprehensive sets of modeling methods*** (also known as representations of systems, and their subsystems and sub-components), that have been developed from other data regression techniques and first principles modeling approaches. These (numeric and non-numeric)modeling approaches include:statistical methods (ML is in that class; other methods include SVM, MCMC, etc.[4]);

[3] Massive volumes of high-dimensional (multiple diverse sources and multiple diverse consumers, multiple timescales and formats; collections of heterogeneous sensors and controllers; humans also included)
[4] SVM-Support Vector Machine; MCMC – Markov-Chain Monte Carlo

"physics-based" models (e.g., differential equations solvers: FEM, CFD, DFT, FFT[5]; Monte Carlo - Boltzman statistics; Simulated Annealing, Molecular Dynamics etc.); agent-based (e.g., data flow methods); and graph-based techniques (e.g., probabilistic methods).

In particular, new directions whereby instrumentation data of a system are dynamically integrated with executing models of that system, in a dynamic feed-back control loop (e.g., those of Dynamic Data Driven Applications Systems (DDDAS)[6]), are demonstrating new capabilities in design and analysis (understanding) of systems. The data-model coordination enhances theability to represent, analyze, and predict behavior and optimize performance of systems with respect to operation, evolution, and interoperability aspects. Such methods can improve the accuracy of models, make the models faster, create decision support capabilities with the accuracy of full-scale models, guide the process of collecting data in targeted ways, coordinate and correlate instrumentation across multiple sets of measurements (sensors), as well as apply coordinated (model-driven) control across multiple actuators (controllers).

These methods[7] afford real time analysis from current data collection coordinated with multi-dimensional models and their data to predict unknowns, capture sparse data collections[8], and determine and quantify the uncertainty of the result.

Supporting such capabilities and their concomitant environments (architectural/infrastructure frameworks) requires new methods (some augmenting traditional ones) for managing the associated data:

1. ***Data collected from instrumentation (e.g., IoT)***, and analyzed in real time provide enhanced analysis capabilities and more effective response over archival data analysis alone, and require supporting the access of real time data with concomitant support for tagging/indexing (attribute/assignment)[9]. The continuity of static data ML/AI analysis rests in the model produced. Hence, a ***dynamic approach*** ensures that the model is updated, used for prediction and decision support, and cognizantly and effectively incorporates heterogeneous modalities of data.
2. ***Management of data*** goes beyond the static ways that entail archiving the collected data, and then "VAULT" the data off-line. Supporting real-time multi-domain models for responsive decisions requires supporting real-time multi-domain and multi-level models for dynamic

[5] FEM – Finite Element methods; CFD-Computational Fluid Dynamics; DFT – Density Functional Theory; FFT – Fast Fourier Transform, …

[6] DDDAS/InfoSymbiotics – DDDAS (Dynamic Data Driven Applications Systems), is a paradigm whereby "the computation and instrumentation aspects of an application system are dynamically integrated in a feed-back control loop, in the sense that instrumentation data can be dynamically incorporated in to the executing model of the application, and in reverse the executing model can control the instrumentation" (Darema 2000). Such approaches have been shown that can enable more accurate and faster modeling and analysis of the characteristics and behaviors of a system and can exploit data in intelligent ways to convert them to new capabilities, including decision support systems with the accuracy of full scale modeling, efficient data collection, management, and data mining.

[7] E.g. enhance standard methods such as reduced order modeling – interpolate and extrapolate not through a static surrogate for example, but through the real-time on-line (sensor) data - enhanced reduced order modeling (enhanced ROM).

[8] Also detect and compensate for faulty/failing sensors; e.g. Pitot tube failure – Prof. Carlos A Varela: Dynamic Data Driven Avionics Systems. Streaming Technology Requirements, Application and Middleware (STREAM2015), Indianapolis, IN, October 2015. Other applications include Next-generation spacesuits (Prof. Leia Stirling, MIT), and medical applications (e.g., artificial pancreas; Prof. Wayne Bequette, RPI).

[9] Ref: Nehme-Rundesteiner-Bertino, Tagging Stream Data for Rich Real-Time Services; VLDB2009, August 24-28, Lyon, France.

data analysis, model updates, and knowledge dissemination. The SAF/CO VAULT strategy requires the use of common (e.g., standard) data translations from data dictionaries, glossaries, ontologies, tagging, querying, and metadata; off-line methods and services are not sufficient for the dynamic, real-time data environments of concern here.

3. ***Data as a Service (DaaS)*** ensures that the right information is accessible, such as having the required security markings, metadata, structure, and data context. A services registry is paramount within the Data Architecture – Enterprise Information Mode to support a systems of systems approach that is extensible (models invoking other models dynamically, as encountered in the dynamic data driven environments of interest here), adaptive (Recommender System – Targeted-Peer-Cognitive), and flexible to the operational mission.

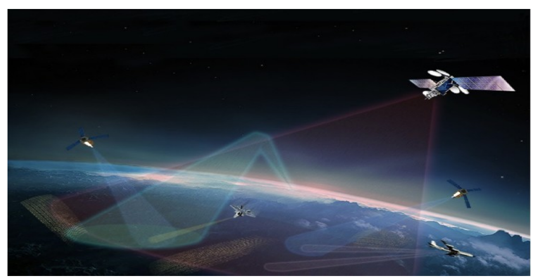
To assess the SAF/CO Data Architecture and Information Model support, the workshop discussions highlighted considerations driven by key Air Force dynamic adaptive missions and capabilities areas:MDC2 (Multi-Domain Command and Control), ISR (Multi-Domain Command and Control), PM (Predictive Maintenance), and Clouds/IoT (Internet of Things).

3. Driving Applications

3.1 MDC2

The Multi-Domain Command and Control (MDC2) mission simultaneously manages assets in space, air, and cyber (SAC) domains. MDC2 is both a data consumer and data producer. The MDC2 as a consumer uses data to coordinate platform operations, with primary objectives on aligning platforms and sensors to provide better situational awareness and help the Air Force make faster decisions, with continual feedback (Air Force Magazine, Jan 2018). The MDC2 as a producer of data provides updates to the mission theater on the battlespace environment.

Multi-Domain Command and Control



- **Objective:** provide better *situational awareness* and help the Air Force make faster decisions, with *continuous feedback*
- **Areas:** complex *adaptive systems*, complex effects analysis, and machine intelligence
- **Challenge:** simultaneously *manage* command-and-control of the air, space, and cyber domains
- **Focus**: machine intelligence is required to enable the human user to master this *complexity at speeds* greater than the adversary

Air Force Magazine, Jan 2018

To orchestrate the SAC domains, it requires elements of Dynamic Data Driven Applications systems (DDDAS) to accomplish complex adaptive response. Complex analysis requires the infrastructure, methodologies, and techniques for data-information-knowledge (DIK) to enable the user to master complexity at speeds greater than the adversary. Enabling such a strategic-operational-tactical capability requires virtualization of SAC platform locations and context data, accessible data content, linked multi-sensor data fusion, usable situation data awareness, and trusted networks for secure and resilient data transmission. A foundation of data support through enterprise services will support multi-domain data alignment, and compatibility and refactor analysis across domains.

3.2 ISR

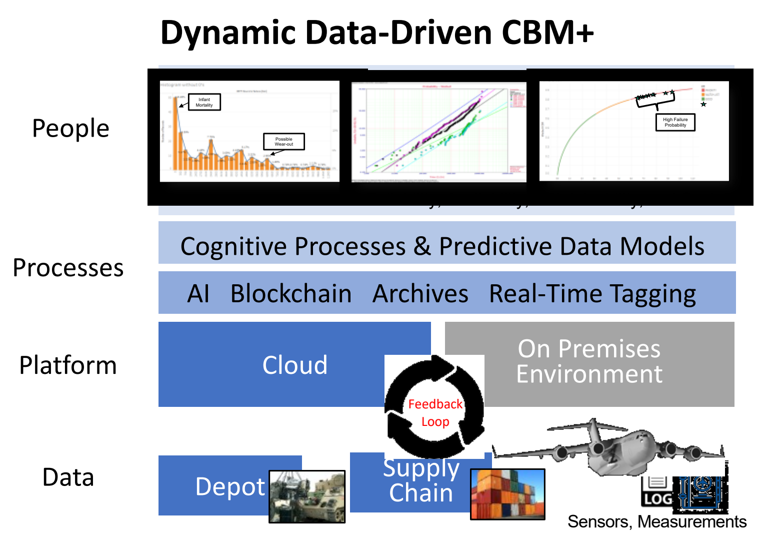
ISR requirements for Anti-Access/Area-Denial (A2AD) environments call for the coordinated collection and processing of data from several platforms. This involves a multitude of sensors operating in various modes, and performing multiple functions to provide relevant, timely, accurate, precise, and actionable data to the warfighter in a cost-effective manner.

Due to the contested and congested nature of A2AD environments, the old method of "loiter, collect at will" and "process at leisure" does not meet the real time requirements of collection, processing, and dissemination. Consequently, a paradigm change is required, which calls for "closed-loop" experimentation, involving concurrent data collection, processing, and analysis of the data. The dynamic nature of the environment and targets of interest necessitate a closed loop interaction between models, data, and processing algorithms - all of these forming a key tenet underlying DDDAS. Critical issues in this context include data collection, modeling and simulation, data exchange between human and machine, data storage, retrieval, transmission, and security for effective information integration. Together these enable an integrated real time visualization of the operational environment.

Emerging ISR Data Needs

- "In-the-loop" experimentation
- Concurrent collection and processing
- Computational challenge-curse of dimensionality
- Storage, transmission, and data retrieval-real time database cueing
- Secure communication links
- Massive Physics
- Massive Cores
- Massive Data "in-the-loop" collection and processing
- Massive Architectures
- Massive Integration

3.3 SHM-PM (Structural Health Monitoring and Predictive Maintenance)



The introduction of advanced analytics into asset maintenance and planning can achieve reduced costs and improved asset lifecycle management with optimized replacement and maintenance planning. Such advanced methods allow to incorporate real-time data (e.g., atmospheric turbulence and weather), which are aspects affecting sensor performance, with the ensuing needs for mitigating and correcting for those effects[10]. Sensor measurements and inspection reports, together with historical and maintenance records (e.g., design and construction records, historical weather models and incident reports, etc.) would extend the life-time of AF assets. Cost, lifecycle and

[10] Hoffman/Kerekes/Vodacek, work on: adaptive Fabry-Perot interferometer (Kerekes – JWA15, 2011); UAV jitter correction and continuous imaging and tracking (Hoffman-Vodacek – Proceedings SPIE 9407, 2015)

logistics savings are achievable by optimizing part or asset maintenance vs. replacement decisions, as well as part inventory and depot management decisions and processes.

Three key enablers support the realization of SHM into Air Force and DoD processes[11]: (1) refining planning processes to incorporate predictive insights and analytical planning, (2) identifying and extracting data for incorporation into real-time and offline analytical models and tools, and (3) ensuring the appropriate people-staffing and training are in place to incorporate SHM-PM into daily and strategic planning processes. Using such predictive methods, efficient processes are established. For example, by determining the expected end of the lifetime of a component, the part can be carried on-board, or stocked (ahead) at the next destination to reduce the down-time for the repair; or as suitable and feasible, additive manufacturing can be exploited, to produce the part, either on-board or at the next destination.

3.4 Clouds/IoT

As collection, processing, and exploitation capabilities advance, the primary challenge in the systems-of-systems environments of relevance in future AF operations (e.g., MDC2, ISR, PM), becomes the lack of integration among disparate and stove-piped data sources. For example, in MDC2 environments, data from all domains is interdependent and connected. Those who recognize this otherwise hidden relationship gain a MDC2 advantage; therefore, data integration must be at the forefront of any MDC2 strategy. Successful integration is characterized by rapid data discovery and access to authoritative multi-domain data by users across the enterprise. Compiling and integrating enterprise wide critical data into a comprehensive repository requires data to be stored in an open and platform agnostic manner where data schemas are established at the time of analysis, not during storage[12]. Identification of all data sources and an authoritative data handling policy is also critical to a master data management and integration plan. A menu of collection capabilities, their respective outputs, and associated authorities are needed to establish the sources for an enterprise wide data repository. To facilitate openness and accessibility, data handling, security, and policy administrative concerns (e.g., coalition arrangements) should be transferred from dispersed collection and processing entities to a centralized data management entity (e.g., center that controls data policies). Furthermore, an enterprise-wide data management strategy requires planning resourcing and infrastructure support at the enterprise level. Current development is focuses on data pipelines that are designed with a program level scope in mind, lacking consideration for enterprise-wide needs. Centralized funding would increase resource sharing and incentivize common data standards. Finally, identification of specific users and use cases would facilitate identification of data gaps, analytic workflows, and prioritization for enterprise wide efforts.

The needs of the driving applications discussed above, are decomposed further into the specific open themes for data needs, discussed in Section 5.

[11] Example of applying DDDAS-methods for Condition Based Maintenance Plus (CBM+) to achieve objectives articulated by the Office of the Secretary of Defense (OSD) on predictive maintenance.

[12] Possible implementations to at least tag/index/attribute/metadata/etc. for some of the data (in real-time) at the time of the analysis, and process other/remainder of data off-line. DDDAS helps here because can determine which data one needs to develop their schemas at time of the analysis (aka in real-time).

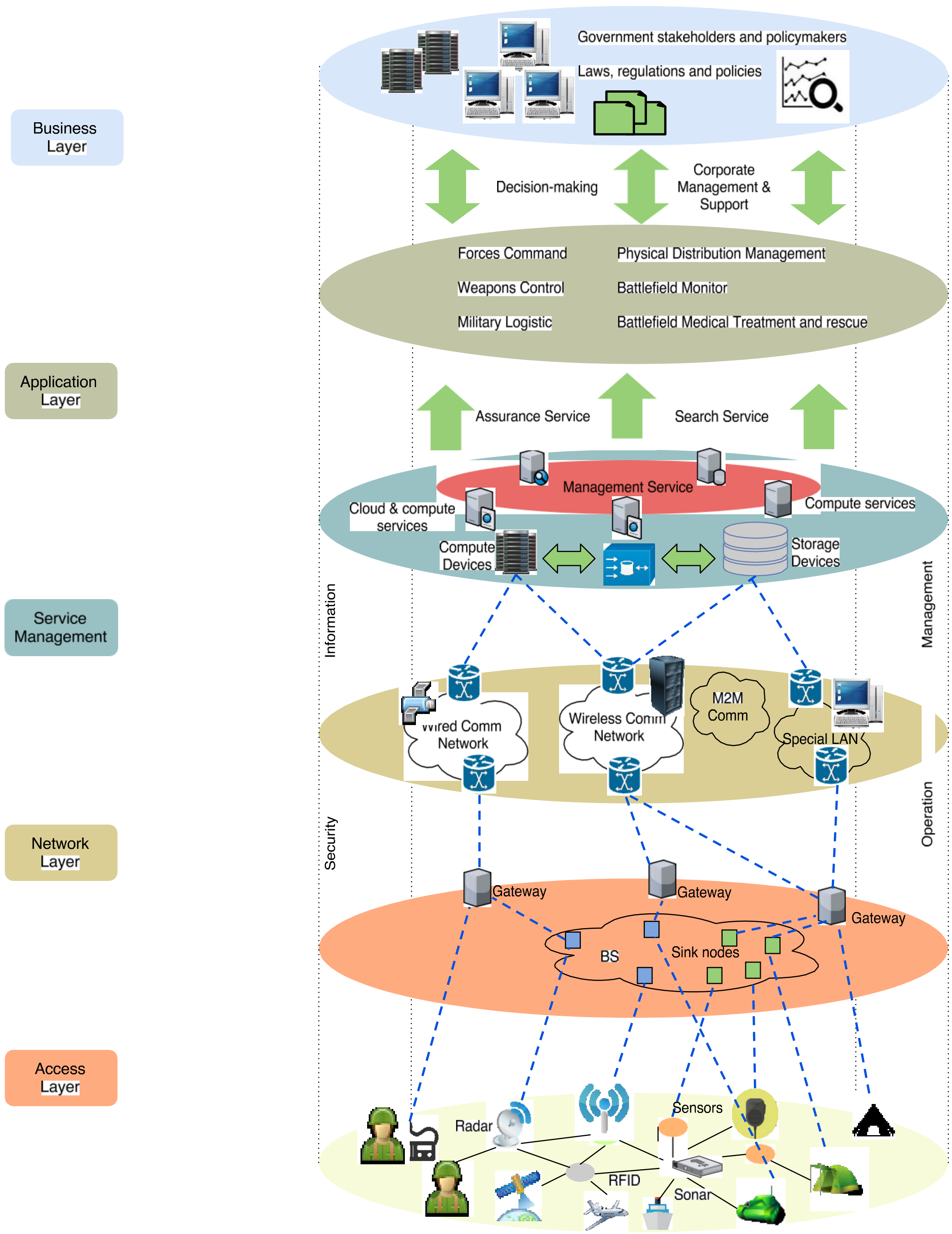


Figure 3. Example of military architecture with six layers

4. VAULT Requirements for DDDAS-based Application Examples

The table below represents some of the discussion topics generated by the working groups as to the needs for different applications in support of the VAULT strategy.

4.1.1 MDC2: Micro-UAV, space, air, and ground sensing (SAG)

SAG coordination extends beyond S/A (space and air) needs to support a multi-domain efficient design Space/Air/Ground; these are necessary capabilities in MDC2 operations.

AF sensor collection coordination (Layered Sensing Concept) includes aircraft (adaptive radar carrying, and UAVs), in coordination with space, cyber, and ground assets, to increase battlespace awareness. Various data collection and modeling (e.g., sensors, weather terrain, and threats) need coordination across massive data repositories that require ***Data Optimization*** access.

DDDAS based methods have shown the ability to synchronize and adapt sensor coordination to prosecute dynamic moving targets such as time-critical targeting based on threat awareness. Additional size, weight, and power (SWAP) data processor designs for micro-UAVs enhance performance. Future testing, scaling to larger systems, would be enabled with more on-board DDDAS processing.

The emerging VAULT concerns (Accessible, Understandable, Linked) must align with all scales from the large to small, on many domains. Accessibility of data transmission and on-board use requires strategies to combine the software (algorithms) with the hardware (chips). Many compute-intensive operations can be done on the platform for real-time analysis, which has to be coordinated with other platforms in different domains. Several research project contribute to such capabilities – e.g., micro-UAVs and very low power computing[13],[14].

4.1.2 MDC2: Structural Health Monitoring

Preventive Maintenance (PM) and Structural Health Monitoring (SHM) is important for MDC2 capabilities, as knowing the health and maneuverability of a UAV platform would support the air domain. At the same time, applications, such as a router in the sky, enable global reach, and optimality and effectiveness in logistics (in addition to cost savings).

The data intensive AF application exploits DDDAS methods for the Dynamic Response for Aircraft (e.g., the Low Cost Attritable Aircraft Technology (LCAAT) Concept). Multi-fidelity models (e.g., detailed FEM (wing), simplified physics (airframe), and reduced models (control)) integrate with real-time data collections using both DDDAS-based (which advance the modeling capabilities over other standard methods such as ***Data Assimilation***[15])**.** DDDAS-based methods enable advanced capabilities such as maneuverability with the accuracy of full-scale models[16],[17]. The transmission of data supports Predictive Health

[13] Karaman S., et al, Exploitation by Informed Exploration between Isolated Operatives for Information-Theoretic Data Harvesting, IEEE CDC 2015, Osaka, Japan

[14] Wang, Z., E. Blasch, et al., "A Low-Cost Near Real Time Two-UAS Based UWB Emitter Monitoring System," *IEEE AESS Magazine*, 30(11): 4-11, Nov. 2015.; and Mohseni, K., et al., "Symplectic model reduction of Hamiltonian systems," *SIAM Journal on Scientific Computing*, 38(1), A1-A27, 2016.

[15] DDDAS is a more powerful concept than Data Assimilation; e.g. integrates data into the phase space of a model, not only to correct a field vector (in positions with uncertainty exceeding some bounds), and includes the data collection control aspect.

[16] Willcox, K., et al, Dynamic Data-Driven Reduced-Order Models CACME, 2015; and Darema, F., New *S&T Directions for New AF Capabilities and CDO Strategy Drivers*

Maintenance (PHM), aircraft routing, and survivability. The approach leverages onboard sensing technologies, aero-elasticity model updates, UAV performance curves and lifecycle costs. Flight tests demonstrate recovery from structure failure of composite UAVs structures. The SAF/CO VAULT aspects (Visible, Accessible, Understandable) are essential in the AF's use of UAV technology enabling non-collocated airmen to monitor and assess a situation, while simultaneously ensuring survivability and mission extensions.

4.2.1 ISR: Space Situation Awareness

SSA for ISR capabilities includes Processing, Exploitation, and Dissemination (PED), which drive S&T requirements in the space community. The small debris especially, presents challenges to many platforms. Data that is useful might not be visible; e.g., small objects less than an RSO (resident space object), and events effecting communications. Radiation and solar flares present a need for real time prediction analysis, to make satellite changes of time and over many time intervals. Hence, predictive estimation is critical to SSA.

Future ISR systems will rely on the sensor exploitation challenged by availability, atmospherics, and coordination requiring advanced methods in (Nonlinear Data Estimation). Sparse views, data rendering in different frames, and translations include data interaction with models forcing ***Data Transformation***. Data transformation includes key aspects of data preparation (see Figure 1). New, DDDAS-based capabilities include the use of physics models (e.g., the Global theoretical ionospheric model- GTIM[18] - that includes atmospheric, optical, and control methods; and adaptive radar work[19]; the latter has transitioned to industry and to OSD). The transmission of data subject to atmospheric effects, requires on the fly processing. With advances in super-resolution of optics data, uncertainty quantification, and nonlinear estimation, RSO tracking is possible. The VAULT aspects (Visible, Accessible, Understandable, Linked) are critical in SSA applications. SSA has many issues with sparsity of views (looks), low resolution, and data transformation. Much of the data needs to be preprocessed to be visible, accessible, and understandable; and linking with other historical and physics data is required.

4.2.2 ISR: Multi-INT Situational Awareness

Battlespace Awareness includes correlating and associating data (including through Graphical Fusion). Multi-INT involves modalities with spatial/temporal/frequency aspects; includes association of Physics-based and Human-based information Fusion (PHIF) by

---

[17] Bazilevs, Y., et al, "Toward a computational steering framework for large-scale composite structures based Bazilevs, Y., et al, "Isogeometric Fatigue Damage Prediction in Large-Scale Composite Structures Driven by Dynamic Sensor Data", Journal of Applied Mechanics 82 (2015)

[18] Bernstein, D. S., et al, ``Correction of the Photoelectron Heating Efficiency Within the Global Ionosphere-Thermosphere Model Using Retrospective Cost Model Refinement,'' J. Atmospheric Solar-Terrestrial Physics, Vol. 124, pp. 30--38, 2015, DOI 10.1016/j.jastp.2015.01.004; and 2. A. A. Ali, A. Goel, A. J. Ridley, and D. S. Bernstein, ``Retrospective-Cost-Based Adaptive Input and State Estimation for the Ionosphere–Thermosphere,'' J. Aerospace Information Systems, vol. 12, 767-783, 2015.

[19] Rangaswami, M.: A basic research effort (using DDDAS-based methods created capabilities for joint adaptive processing on transmit-and-receive to close the loop on the radar from receive to transmit. The capabilities developed allow for real-time instantiation of the threat environment, enabling informed command&control decision making. From the technology point of view, in addition to the DDDAS based algorithms, such capabilities require real-time data management on-the-fly. The capability has important advantages for electronic protection and countermeasures.

gathering intelligence data, such as multimedia; synthesizing them requires many probabilistic combinations.

Not all data collected are needed for a particular analysis objective (e.g. to satisfy mission objectives), though many of these data may be useful for a future analysis (for example historical reference, prior behaviors which may be important, etc.). Getting the real-time dynamic information from analytics is required. Numerous human in/on/out of the loop-experiments demonstrated that many issues are still required for mission success. Access and linking are underappreciated in multi-INT fusion solutions.

DDDAS success includes case examples which involve: human (text analytics) and physics (image analytics) models, along with 3D environment-related models (terrain, roads, time of day, railroad) for ***Data Fusion.***[20] Text analytics (entity, relationship, and event) uses social/contextual models to augment the decision at speeds faster than the adversary. Additional insights come from data architectures (e.g., Docker for data processing) to expedite the federated query. The emerging VAULT concerns (Visible, Accessible, Linked) include data fusion (correlation and association); while at the same time, the analyst requires visualization of the situation (e.g., user defined operating picture – UDOP)[21]. There is need to enable access to the data for the operator to minimize processing and facilitate exploitation and dissemination. Packaging the metadata in folders through structured object modeling (SOM) enhanced distributed and real-time analysis.

4.3.1 Cloud/IoT: Cyber awareness

Enterprise security aligns with trusted information. Assessing trustworthiness may impact performance, as corrupted nodes and files need to be tested at multiple stages (and sent on multiple channels) thus increasing the data communicated and resting processes. Nevertheless, to use the data their reliability must be ensured.

Cyber operations result in data flow from computers to networked platforms. Autonomic Monitoring, Analysis and Protection (AMAP) detects the value of data using machine learning (ML) to model normalcy and abnormal/malicious files detection. Cyber awareness includes data transmission, content, and storage. Trust analysis is relevant to ***Data Protection*** that comes from dynamic data analysis and model assessment. DDDAS data traffic, attack, data flow, malicious file detection models monitoring the safety and security of information. The real-time approach reflects behaviors (network, activity) that change as fast as the adversary. Adherence to the VAULT aspects (Understandable, Linked, Trusted) is a must to address the concerns that malicious attacks on networks will continue to increase, even if for only disruptive intent. Usable data across the network can support emerging and proactive migration for destructive failures. Since a network detection is pattern analysis, a library of patterns has to be linked and aggregated for cyber awareness. Further issues on the many areas of trust (network, communication, machine, and human) are intertwined as data on the move is subject to corruption and loss.

4.3.2 Cloud/IoT: Networked Resources Awareness

Analysis and decision support for networked resources is important in many AF operations and their related systems, including energy/power delivery systems, and more broadly infrastructures relating to AF operations; or consumers and production line in manufacturing

[20] Blasch, E., et al., *High-Level Information Fusion Management and Systems Design*, Artech House, Norwood, MA, 2012.

[21] Blasch, E., "Enhanced Air Operations Using JView for an Air-Ground Fused Situation Awareness UDOP," *AIAA/IEEE Digital Avionics Systems Conference*, Oct. 2013.

operations; or optimized urban traffic and transportation, and other civilian infrastructures such as energy/electric and water distribution grids. Network analysis of energy/power includes controlled (local) and uncontrolled (global) issues where weather, use, and behaviors need to accounted for. The ability of a Supervisory Control and Data Acquisition (SCADA) failure has to be resolved, assessed, and mitigated. Energy is required for data storage, movement, and analysis; so, it is fundamental that it operates as expected. In computational- and data-grids, there is heterogeneity and adaptivity at multiple levels.

The computational grids of relevance to the environments of interest here require the support of seamless integration of resources ranging from the high-end/mid-range computational platforms to the real-time data-acquisition and control. The approaches support the dynamic integration of instrumentation data into executing (application systems) models, and in reverse the models to guide in a coordinated way the data acquisition (e.g., adaptively coordinated measurements from heterogeneous collections of multiple levels of sensors), and also apply coordinated (application model-driven) control of collections of controllers. In such environments the computational complexity of the application can change (based on dynamic data inputs) and/or at execution time with other models (describing other modalities of a system) can be invoked (based on dynamic data inputs). These changing computational needs require at run-time adaptively acquiring the (cloud-based) computational resources needed or determining what parts of the computation may be performed at the sensor (and/or controller) levels. To support this adaptivity and delivery of service with optimized performance (at the computational platforms and data systems layer – hardware and systems-software), requires data (information) on performance characteristics of the computational systems layer, performance data on the applications models and/or their algorithms, and recommender systems, employed to ensure optimized performance and just-in-time delivery of services (for decision support with the speed of action).

The coordination of multiple heterogeneous resources(e.g., IoT) under multiple and dynamically changing demands, such as those manifested in the system networks referenced above, require adaptive, real-time management and response. A representative example is that of electric power-grids, and ensuring optimized operation under multiple energy sources (power-plants and renewable, the latter being of variable output) and multiple classes of consumers, with varying levels of priorities (e.g. a hospital or a military base). Optimized energy availability is subject to network vulnerability assessment of collections of power resources and power consumers[22]. Data interaction with models requires***Data Scaling***, such that systems of systems operate with changing conditions: network load, seasonal changes, radiation amounts, and SCADA attacks. DDDAS solutions utilize physics, human, social, and economic data factors to demonstrate facilities management for low/high balancing effects. New data domains exist in cyber, social, sensor, and energy power analytics.

[22]Celik, N.: Data and Computation-aware Modeling for Resilient Power Grids create decision support with the accuracy of full-scale models for dynamically changing electricity microgrids, under dynamically changing resource availability (power-plants and renewable – solar-wind, resources- of varying output) and multiple consumers (each with multiple levels of priorities) and the need to satisfy optimally their needs, predict potential failures onset, and optimize delivery of services, satisfying critical priorities in the case of a failure in the power grid (including microgrids). The capabilities enabled are real-time inferencing in a large-scale system under limited computational resources and excessive number of parameters and massive data loads involved, need to address massive, highly complex, and rapid moving data are necessary to make useful discoveries and achieve timely control over these networks. Dynamic integration of measurements, computation, and control; algorithms invoking and utilizing other algorithms of different scales and different fidelities, and their concomitant data, of multiple levels of fidelity; assessing on the fly the data integrity (detect and avoid or mitigate accidental or intentional corruption of data).

The VAULT requirements (Visible, Linked, Trustworthy) include large scale and increased demand on energy, and many stakeholders are involved crossing the civilian-military domains. The visibility challenge comes from many stakeholders viewing the operations differently (e.g., different organizations having differing objectives and strategic goals).Also, changing world conditions and constraints require the adaptivity needed (e.g. a recent challenge is the wide-spread fires that encroach on electric power and energy infrastructure systems). Therefore, multiple complex and intertwining challenges need to be considered and supported.

5. Emerging Themes in Data Requirements

1) Unknown data (Data in Collect) ingestion from raw, structured and unstructured opportunistic data collections, or modeling and simulation (M&S) data

Overarching points: Robust decisions derive from static or dynamic structured data collections[23]. However, unstructured data requires curation for future analysis. There is a wealth of data collected from external sources, opportunistic collections, and refactored[24] repositories. These data sources require innovative methods for storage, access and retrieval for exploratory analysis. Future developments require backward compatible approaches that leverage historical data, while future collections would have prescriptive methods for analysis. Likewise, data collected in one domain may in the future be useful for other domains, and be used in environments not originally intended. The multitude of data includes directed collections such as platform coordination for overlap collections; as well as indirect collections in one domain providing opportunistic data to support current and emerging concerns for other domains.

In the case of MDC2:An example is space weather data that informs cyber networks on potential radiation and network outages that reveal natural effects versus coordinated attacks. While rare events are possible, modeling and simulation can determine in a (DDDAS-based) complementary way with targeted datato improve the accuracy of the model,[25]and predict the onset of potential (adverse) situations for more informed decisions. The linking of SAC data provides a global understanding such that knowledge (gained from intelligence information gathering over authoritative data sources) can aid commanders in both domain assets under control, as well as provide useful information to other domain commanders.

In the case of ISR: An example in this context is training data for characterizing signal dependent interference scenarios. Classical space-time adaptive radar processing deals with the problem of covariance estimation from replicas of training data sharing the same characteristics. However, owing to the transmit degrees of freedom invoked in fully adaptive radar, each realization of training data corresponds to a different covariance matrix. Consequently, the covariance matrix estimate from this class of training data becomes singular, precluding implementation of the adaptive processor. Therefore, a new approach is required for this problem-physics based modeling and simulation approaches, which

[23] E.g., essential for example in the case of *contested environments with information/data sparsity*

[24] *Refactored* – as used in “refactor data” and “refactor database schema”: A database refactoring is a simple change to a database schema that improves its design while retaining both its behavioral and informational semantics.

[25] Blasch, E., S. Ravela, A. Aved (eds.) *Handbook on Dynamic Data Driven Applications Systems*, Springer, 2018.

opens up the issues of generation, accessibility, and availability of sufficient quantity and quality of training data involving more general principles of channel estimation. In addition, with respect to requirements on data tagging, typically the Air Force employs many people to analyze and report on streaming video. The time-consuming operations have been known as weak in operational speed and agility. As images stream by, an analyst calls out information in the video while another analyst manually records the information in a fixed business tool. The data use is antiquated and limiting. Future data strategies would need to leverage computer automation to collect, index, and store data that can be retired for dynamic product generation. The dynamic product generation supports the ability to access the data based on queries of interest, reducing timelines to manually manipulate data.

In the case of PM: An example is the supporting S&T for Predictive Maintenance, which requires access to real-time or near-real-time data sources, brokering agreements between data owners and other users of the data. Driven by AFMC/AFLCMC Predictive Maintenance needs, several examples of data sources that would provide particular value to the PM effort include those used for maintenance (e.g., IMDS, G081, REMIS, etc), sortie, supply, aircraft telemetry and on-board sensor data. Whenever possible the data sources should be accessed from the authoritative source, consolidated into centralized repositories/interfaces and easily accessible. This recommendation supports the Visible, Accessible, Understandable and Linked components of the VAULT strategy.

In the case of Cloud/IoT: A specific example of IoT is to track shipments of products to a soldier, and in general, to provide situational awareness and visibility into supply and demand trends in an area. This includes, fundamental questions pertaining to on hand supplies, estimates pertaining to order arrival times and “just in time arrival”. Problems include lack of accountability in shipment arrivals, mis-alignment between shipment priority and the criticality of supplies (e.g., food, fuel) and warehouse stockpiles. The current prevailing technology relied on barcodes, which required line of sight to read, and lacked accuracy and manual data entry. Prevailing IoT technologies to overcome these challenges are RFID (both active and passive) technologies and the corresponding automation. Those IoT technologies provide increased accuracy and real-time visibility into supply chain logistics. In addition to the tags themselves and readers, communication structures (wired and wireless), as well as sensor-agnostic IoT infrastructures allowing a multitude of devices to sense and communicate within and across networks as assets pass through. Additional sensors (e.g. electro-optical, hyperspectral, etc.) can also be leveraged to assess item health. This use case demonstrates SAF/CO tenants Visible, Accessible, Understandable, and Linked.

2) Background data (Data at Rest) access for efficient context analysis to support change detection, anomalies, and causal analysis

Overarching points: Decisions-to-Data (knowledge from information and data) requires an understandable data representation that includes a common set of methods for semantics, ontologies, and dictionaries to support curation. The background data provides context of normalcy to support change detection, anomaly characterization, and hypothesized causal analysis. Data fusion includes correlation and association. To support correlation, data registration in time, space, and frequency enables a more complete situation understanding. Supporting data association requires accessible data from different domains and classification levels; however, not all data needs to be sharable, and for performance efficiencies considerations may (or should) be delivered based on a need for immediate use (or need to know basis). Service Level Agreements (SLA) provide a more orchestrated granularity of data authorization; mission-based service requests (e.g., ISR as a service) with token-based approvals; and query-based detection for discovering versus control for threat analysis.

In the case of MDC2: Data at rest includes the many collections at different domains, which exists in a database. Over time, as specific domains are exploited (e.g., an airspace); knowledge is developed over time. The data can be used to build models, provide situational understanding, and predictions of information needs. An example is the common airspace routes available in a complex civil-military operation which generates airspace restrictions that can be used to develop effective routing procedures to get the best perspective to observe the area of regard.

In the case of ISR: The data challenge can be broken up into signal level data coming from a sensor suite and mission level meta-data. The processing involved in converting sensor level data into meaningful information from a mission perspective involves significant computational cost. Data association from multiple sensor types in order to extract meaningful information and create meta-data poses an important challenge in terms of either complexity of models, complexity of rules or complexity of algorithms. Additionally, creation of metadata can influence future collection of data from a sensor suite. This calls for two-way feedback between signal level and mission level data. SLA need to be in place to permit data sharing among parties with the appropriate need to know.

In the case of PM: Supporting predictive maintenance requires establishing criteria for defining which kinds of items in the AF enterprise should be serial marked and tracked. Notably, there are many items in the AF enterprise that are not serially-tracked (e.g., LRUs[26]), which makes it difficult to track usage, complicating the task of executing predictive analytics. These require approaches to ensure the appropriate level of mandatory serialization[27], as well as recommended mechanisms for performing tracking of serialized parts such as the adoption of universally unique identifiers (UUID) or radio frequency identifiers (RFID), where appropriate. A sample use-case is the difficulty of knowing which parts are located behind-LO[28] panels – failure to adequately serialize, track and log the usage of these parts are can result in expensive maintenance actions to inspect and verify. We believe this recommendation supports the Accessible and Understandable components of the VAULT strategy.

In the case of Cloud/IoT: IoT encompasses a multitude of technologies; to include hardware sensors (e.g., optical, pressure, temperature) and compute units (e.g., Raspberry PI, Nvidia Jetson to name a few), as well as communication protocols and network devices (e.g., network sinks or communication gateways). The devices connect to a network and communicate with each other and software, a gateway facilitates device communication with consumers on a different network, and various configurations of compute facilitate sensor data processing and tagging. Real-time models can serve to apply machine intelligence to

[26] Line-replaceable units (LRU's) are modular components designed to be quickly replaced at an operating location
[27] Serialization and serialized parts here refer to the designation and tracking of routable (e.g. replaceable) components via a unique number
[28] Low observable (LO) refers to the stealth-enabling protective coatings applied to aircraft panels

the tagging process, and intelligently filter data to optimize spectrum contention and optimize power consumption and compute resources. Military application of offline and archival IoT data includes base security protection and situational awareness; for example, detecting trends correlated with time of data in terms of personnel movement, or long-term mobility patterns for infrastructure modernization and capacity planning (e.g. roadway expansions).

3) Repository of Models (Data in Transit): from instrumentation and algorithms (numeric, non-numeric) coupled with massive volumes of high-dimensional data from heterogeneous sources for mission awareness

Overarching points: Architectures require multiple layers of feedback to support multi-domains, distributed users from data instrumentation, algorithms, and models. Physics-based modeling is needed, for effective coordination, orchestration, communication in a control feedback loop support decision speed. High fidelity physics (e.g., radiometry solver) models for scene generation/rendering enable real time data to be presented for situation awareness. The data production from (MDC2) systems reside in a model and data repository with an interface to support man and machine interfaces, collaborations, and course of action simulations.

In the case of MDC2: Data in transit requires effective and efficient big data processing methods such as Hadoop, parallelization, and container-based approaches. An example includes transmission of data being collected (as a data producer) from one-domain and transmitted to another domain (data consumer). Transmitting all the information is not helpful when communication bandwidth is limited. Hence, strategies of providing the useful and trusted information from one domain requires strategic decisions based on the need to share versus the need to know for continuous coordination of data across the MDC2 platforms.

In the case of ISR: The dynamic nature of the environment and targets of interest necessitate a closed loop interaction between models, data, and processing algorithms. This leads to feedback at multiple levels in order to match up models to collected data and in turn allow collected data to update models and guide future collections. In this context sequential and block updates to the models and processing algorithms are needed in order to carry out the data to model match using prescribed metrics. Matching the models to the data is replete with challenges of real-time data acquisition, processing, and updates. A key issue in this context is that of acceptable latency limits.

In the case of PM: As with the other applications, the data being transmitted might not be all the raw data, but an aggregate – such as distribution of PM data and estimates. The large volume and velocity of PM data requires methods that afford a approach for normal data as well as being able to highlight anomalies. For example, warnings and alerts that logistical parts needs would cause a stop in mission operations.

In the case of Cloud/IoT: A use case is situational awareness of fuel storage capacities and its transportation in alignment with vehicle mobility patterns and stockpiles. In order to monitor vehicles and depot fuel resources, sensing systems consisting of physical measurement devices (which sense the environment and transform the signal into a digital representation) are directly connected to programmable logic controllers (PLCs), which monitor and combine sensor readings for further communication. Frequently PLCs are connected with wired serial communication (e.g. RS-232). Relying on physical wires to facilitate PLC communication with computers and databases are the norm. However, these wires make the system rigid and can limit how devices communicate and potential system upgrades and modernization efforts. Wireless IoT (e.g., Wi-Fi, 3G, sidelight) communication technologies can enable these sensors and local processing capabilities to be dynamically reconfigured in dynamic ad hoc configurations to enable prioritized sensing and communication. Wireless IoT can also enable the ability to leverage local hardware and software upgrades dynamically to reduce "long-haul" communication bottlenecks as the velocity and variety of devices become pervasive. Such a physical sensing infrastructure facilitates real-time capacity monitoring and trend analysis and correlation with impromptu missions, as well as alignment with strategic long-term military asset modernization initiatives and reduction unnecessary logistic delays, costs and asset return on investment.

4) Dynamic Analysis (Data in Motion): based on mission needs, multiple time scales with feedback control support mission efficiency

Overarching points: The driving challenge areas in this report (MDC2,ISR, PM, Clouds/IOT) require coordinated feedback agility (cyber, tech, apps) for real-time response autonomy, contemporary breakthroughs, and data refactoring. The MDC2 architecture supports a cloud analysis versus an edge (domain process) to bring together real time data at the edge, and M&S in the cloud for exploratory and predictive analytics. The data that includes spatial/time attribute stamping for synchronization of mission operations would enable real-time platform control and routing. The data transmission coordinated in the network layer can optimize over bandwidth availability, delays, and quality of service (QOS) performance. As a service-based architecture will support the multi-service data aggregation coordination for command-based field of regard (FoR[29]). The data recording lifecycle sustainment and maintenance based on program cost, schedule, and performance enables the best data available.

In the case of MDC2:An example is space weather (effects of solar flares) on monitoring on the ground for cyber-attacks, which could also be observed under conditions based on a blizzard, heat, or flood. These then challenge aircraft telemetry data rendering of EW information to determine cyber-attack. The dynamic data ensures effective operations of the system. While not all data in transmit is needed for dynamic operations, the motion information of platforms requires dynamic data that impacts immediate platform control.

In the case of ISR: Massive data architectures are required for ensuring the availability of massive volumes of high dimensional data from heterogeneous sources (sensors and human sources for example) to relevant decision makers with the need to know. Of greater consequence is the ability to transform massive volumes of high dimensional data to reduced dimension data and disseminate relevant data to parties with the need to know. This, in turn requires, scalable, flexible, and modular architectures to allow for efficient data dissemination in real time. The ability to deal with creation and dissemination of metadata adds to the architecture complexity. This requires a delineation of on-board processing and

[29]FoR – the area captured/mapped by a moving sensor – also referred to as AoR - Area of Regard

off-board data processing with supportable, cost effective architectures enabling the dissemination of relevant information in a timely manner.

In the case of PM: Predictive maintenance of dynamic data is a result of the speed at which data is available and usable. A data in motion situation includes the challenges associated with real-time operations and transmission of current asset capabilities to the enterprise. If anin-flight failure occurs, then the dynamic data could be sent to the enterprise from which maintenance and analysis could be expedited instead of waiting for static data analysis to reveal the possible course of remedy.

In the case of Cloud/IoT, Cloud based solutions for data in motion have well been publicized – such as that of IBM. Many strategies have been proposed where Cloud/IoT are the software methods to realize the data in motion. The AF could adopt some of these methods to enhance data processing to increase the speed at which decisions are made. (See P.C. Zikopoulos, et al., Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw Hill, 2012)

5) User interaction (Data in Use): for near real-time exploitation, analysis, and clarify data relevance.

Overarching points: Architectures can be a data push access (producer) as well as a data pull usable (consumer) that requires a federated brokerage to enable decision speed balancing the need to share with the need to know. The need to know arbitrates the sharable data to afford query based "access" to (formal) data, information, or/and knowledge. Getting the right data requires meta-tagging annotation of data to support archival replay, access to usable data, and the ability to aggregate and predict future data collects based on data needs uncertainty. The query-based services include agent-based contextual inquires to learn the user work domain procedures and tolerances for recommender systems. Additional insights come from cross entity user-ID access management (balancing the need to share vs. need to know) for distributed user collaboration.

In the case of MDC2: MDC2 requires data views sent to data providers/collections for real-time control. A user agent serves as a proxy and partner wingman to jointly learn, supporting bi-directional mentorship, group learning, and user-machine collaborate. The data of importance comes from federated Trust as validated by relevancy based on use. Data meta-tagging based on user and machine analytics for authoritative data reference provide added value from source data.

In the case of ISR: In-the-loop collection and processing calls for significant human/human, human/machine, machine/human, and machine/machine interaction. The coordination requires reliable interfaces to permit the information exchange. A federated brokerage is required in this instance. Quite often the need to know and need to share impose conflicting demands on service which then calls for a reliable control that efficiently balances these demands.

In the case of PM: Supporting Predictive Maintenance requires the creation of a development environment supported within the AF Data Architecture to enable the capability of hosting analytics tools. This would provide a dedicated area for developers to build, test and stage their analytical tools before releasing them publicly for general population usage. Ideally, this development environment would also offer collaborative mechanisms where developers can share existing analytical approaches or code.

In the case of Cloud/IoT: IoT enables the real-time sharing of the environment (e.g. sensed data) for situational awareness and coordinating informed decision-making; enabling both machine-to-human and machine-to-machine scenarios. To achieve this, convergence in hardware, networking, data management, spectrum management and analytics, as well as various forms of cloud computing, are required. An example application of this is disaster recovery; where human life and military readiness can be critically dependent upon a timely and coordinated response. First responders can have their deployments optimized according to the criticality of the situations of personnel requiring assistance. Their routes can be coordinated and optimized for emergent conditions (infrastructure, changing weather conditions, etc). Later, infrastructure recovery can also be enhanced (for example, consider the "black start" problem of restarting a power grid; it will be critically important as a factor of this to determine which part of the grid was rendered offline due to a cyber-attack, which part may still have malware present and become active when the restart is attempted).

6. Data Management, Data Architecture and EIM Support

Given the examples discussed above (derived from the break-out group discussions), a common theme was the Enterprise Information Management (EIM) that supports the data processing. This Section provides some of the possible approaches considered to address the needs (data dictionaries, glossaries, ontologies, tagging/indexing/attributes, other metadata, querying, security markings, structure of data, and context of the data, services registry, workflow-based tagging, and Data Architecture & Enterprise Information Model. These capabilities need to be supported for the dynamic and adaptive environments of interest here (e.g., MDC2,ISR, PM, Cloud/IoT), and in the context of DDDAS-based multi-modal/multi-fidelity models interacting with multisource multi-modal/multi-fidelity data). The capabilities support the requirements of both real-time as well as archival data used in conjunction (dynamically integrated) with models (ala DDDAS), as well as workflows of models invoking other models, based on dynamic data inputs[30]. The sample approaches provided here may be viewed as test cases for SAF/CO to consider, with respect to Data Architecture support, and data and computational resources sharing.

These approaches include: adopt a *data-centric middleware* approach to provide the required support for real-time, archival, and DDDAS application related data. By adopting this approach, data analytics applications do not need to deal with data types, real-time or archival, message structure or network delays. All these issues are handled by the middleware in a transparent manner. The middleware will address the issues related to message passing, data filtering, integrity check, QoS and security requirements for each data type and/or source. Furthermore, since the middleware is responsible for all these

[30]Other aspects include: need for the data to be tagged with accurate time (and location for mobile assets). Where possible, ensure real-time or near-real-time access to data sources – data needs to be accessible, consolidated. And, need a dedicated area for developers to build, test and stage their tools prior to pushing to production for general population usage.

issues, applications are connected with the data objects without need to be changed or adopted. Also, new data types can be added in a transparent manner[31].

In addition, to satisfy the requirements and data architecture support with respect to dictionaries, glossaries, ontologies, tagging, querying, security markings, meta data, structure of the data, and context of the data, a possible approach is to adopt the Data Virtualization Service (DVS). The DVS performs many of the traditional data transformation functions such as Extract-Transform-Load (ETL), data replication, data federation, Enterprise Service Bus (ESB), etc. In addition to these functions, the DVS leverages emerging virtualization technology to provide real time data integration at lower cost with more speed and agility. By separating the underlying hardware dependencies from the logical capabilities of the data sources, we can provide the automatic linkage and adaptation of the data analytics environments as required by the DDDAS paradigm. This means the DVS can easily integrate and support all the metadata structures, and technologies, such as dictionaries, glossaries, ontologies, etc.

Adopting service-oriented architecture (SOA), allows to publish and discover all the computational models and tools needed to perform sophisticated data analytics that can be executed as a data analytics workflow. SoA supports the needs of services registry, workflow-based tagging, and Data Architecture & Enterprise Information Model. A number of DDDAS approaches use Apache Storm[32], with an open-source distributed computation system to implement the data analytics workflows[33]. DVS provides the capabilities to share these data sources among several VDAPs and provide the data in the required format and when it is required; this may solve one of the major challenges that SAF/CO is aimed to address, which is to provide the data in the required format, when it is needed by the application[34].

With respect to data sharing, the Data Lakes model provides flexibility and ease of access to the data, from a common space and common data models, as by definition that data in Data Lakes are “VAULT”-ed. However, in practice, for ease of implementation, security aspects, performance considerations, and the fact that data are collected and reside in different organizations, the model of “data-puddles” has emerged and needs to be supported. In addition, often data are too large to move (and it will be the computation that will be sent to where the data resides – this is akin to the old “query-shipping” methods that have been supported in from early-on in DBMS systems). Moreover, for real-time data (and for

[31] The implementation of the data centric middleware will be based on the OMG Distributed Data Service (DDS) standard that provides the following features: 1) DDS moves message construction, communication, data filtering and validation to the middleware software; 2) DDS utilizes real-time publisher-subscriber protocol to enable peer-to-peer communication; 3) DDS support automatic discovery of newly joined devices and their data structures; and 4) DDS can define different QoS or security for each data type, source, etc. DDS middleware is a standard tool and its real-time publisher-subscriber protocol (RTPS) ensures interoperability, real-time performance and automatic discovery. It is widely used in several mission critical applications including Radar and Ship Management System, Air Traffic Control Centers in some European countries, Automation and Scada Systems, Duke Energy for Open Field Message Bus, etc.

[32] Hariri et al.; DDDAS-Based Resilient Cyber Battle Management Services (D-RCBMS), 2015

[33] For big data storage and analytics, one may use Hadoop, and Map/Reduce programming model when it is appropriate.

[34] Other methods include NIST SMART DATA FLOW system [Ref. Martial Michel, Systems Plus Inc, and NIST: Network Transfer Control Data, An Application of the NIST DATA FLOW system (part of the NIST SMART SPACE project).

DDDAS-based environments) the (single) Data-Lake model also poses performance related challenges, e.g. tagging/indexing and metadata constructed in real-time. Therefore, the “Data-Federation” model needs to be supported by the Data Architecture and the EIM. Likewise, for Cloud Computing, the related (computational) model is the Federated Community Clouds[35].

In the case of networked systems (e.g., power grids), the “Data Architecture” needs to support: a) Data synchronization between the distributed sensors and decision networks; b) Determination of how and where the missing or unrecognized data types will be reported/addressed (e.g., at the low-end of the computation or the high-end); c) Automation of the security markings so that they could pace up with the real-time capabilities; and d) Standardization of the data architecture so that it is easily transferable over multiple domains/applications are the concerns related to our works. If addressed, these could have significant impact on our capabilities of (near-) real time analysis and prediction in advanced simulation systems.

Medical imaging is an area that has modeling and algorithmic methods requirements and approaches that as are akin with ISR needs. In that area, deep-learning (DL) methods are used to detect, classify/label and segment imagery at different levels of resolution. DL involves coupled analysis at multiple scales - a given image will have between 50K to several million objects (cells) with the cells forming a variety of structures at different scales. Examples of efforts also potentially relevant to the Air Force includeusing deep learning to identify regions in gigapixel images infiltrated with immune cells, and where each image can have roughly 1M cells and the level of resolution of the immune cell identification process gives us something like 250K little patches per image. These immune infiltrated areas are then aggregated with clusters characterized. Another deep-learning application is the use of these methods to analyze deep learning derived segmentations (easily extended to classifications). All these efforts require appropriate data representations, controlled vocabularies[36]. There is also industry interest and have already developed and are marketing the instrument which employs augmented reality and support APIs that allow in the loop image analysis.

7.Recommendations – Technical and Policy& Acquisition Related

The recommendations provided here derive from workshop participant presentations and working-groups discussions.

7.1 Technical Recommendations

R1: Models interacting with data in a feedback control loop (dynamically integrated data & models)

- Establish In-the-loop Experimentation with real, simulated, and predicted data
- Design M&S-enabled Decision Support to provide real-time context awareness
- Invoke multilevel, multiscale, and multimodal, dynamically invoked based on dynamic data inputs to determine effective EIM data delivery and use
- Provide on-demand asset and sensor tasking “via the cloud” so as to test data methods for real-time operations
- Establish metrics to determine the key performance parameters of future EIM systems

[35] Ref Craig Lee; and Robert Bohn – NIST
[36]These efforts are supported by NIH (NCI) grant, led by Prof. Saltz; SUNY Stonybrook.

- Develop uncertainty representation and propagation for dynamic data and dynamically invoked models to monitor data needs and delivery
- Consider semantic data modeling and representation for inter-operability VAULT-Enables DDDAS system software for high-end streaming data for multisensor fusion support including at "the edge".

R2: Extend the Multi-Domain MDM (Master Data Management) to include real-time multidimensional[37] and dynamic data, and data from models

- Provide support data from models, and data describing models.
- Utilize data refactoring
- Leverage labeled data as a part of existing analytic workflows
- Develop automated methods to create metadata
- Create catalogue of sensors/sensor owners
- Utilize on-demand data – accessible, query-able
- Determine data interoperability / standardization – what is different for real-time/dynamic data?
- Identify what is needed to ensure –correctness, provenance, authority, trust worthiness of real-time/dynamic data AND do so on the fly
- Support correlated data tagging and metadata creation for efficient context analysis to support data-anomalies detection and causal analysis. E.g. DDDAS methods allow, detection of onset of data-anomalies and associated cause.
- Establish data exchange for human/machine machine/machine, human/machine, human/human (including model data)

R3: Support enterprise-level federated service-based architecture

- Provide federated semantic/numeric/configuration data modeling and representation for inter-operability VAULT (for clouds and at the edge)
- Accommodate real-time sensor data over the systems enterprise
- Develop flexible architecture to support correlated data tagging and metadata creation for efficient context analysis to support data-anomalies detection and causal analysis; e.g., DDDAS methods allow the detection of onset of data-anomalies and associated cause.

R4: Proactive sharing of data based on need to know and with intelligent granularity

- Establish risk management and governance of the data (control, access, distribution)
- Support Recommender System capabilities
- Support repositories of characteristics and parametrization of models with respect to suitability in terms of data models used, and performance of underlying computational, communication, and storage (memory hierarchies) infrastructures
- Support of dynamic partitioning and mapping of model components and algorithms to the underlying resources ranging from the high-end to the real-time computing architectures
- R5: Data for Performance analysis and benchmarking
- Determine measurable attributes to drive success

[37] Massive volumes of high-dimensional (multiple diverse sources and multiple diverse consumers, multiple timescales and formats; collections of heterogeneous sensors and controllers; humans also included)

- "Manage Complexity at Decision Speed"

R5: Data for Performance analysis and benchmarking
- Determine measurable attributes to support performance objectives
- "Manage Complexity at Decision Speed"

7.2 Policy- & Acquisition-related Recommendations

a. The AF SAF/CO should outline a set of minimum recommended and required data rights that future acquisition programs should obtain (e.g., a data dictionary to make sense of the rest of the data). Additionally, contract language should maintain this as an enduring requirement, so that future changes/upgrades to existing systems do not fail to produce updated documentation for changes to existing or newly available data sources. We believe this recommendation supports the Visible, Accessible and Understandable components of the VAULT strategy.

b. The AF SAF/CO should advocate for a cultural shift within the Air Force, changing the mentality from "need to know" to "need to share". Data owners should be required (by AFI) to register any data they create in an appropriate data registry (or otherwise provide written justification why the data will not be registered). Common restrictions on the use of data should also be identified at the time of registration so that it can be properly protected (such as data governed by HIPPA, the Privacy Act, or similar). We believe this recommendation supports the Accessible, Understandable and Linked components of the VAULT strategy.

c. Additional Policy related Recommendations:
   - Data owners may wish to purge data that has outlived its usefulness to the owner, however that data may still contain value for other users. Data owners should be required (by AFI) to maintain a disposition plan for all digital information (e.g., via records management processes). The AF SAF/CO office should establish a data warehouse and maintain a "first right of refusal" for this data, with the option to take ownership of this data from the data owner at disposition, or otherwise authorize the data owner to delete the data as requested.
   - System components/parts (e.g., LRUs, referenced earlier) are not always serially-tracked, which makes it difficult to track and leverage predictive analytics against. The AF SAF/CO should investigate approaches to ensure the appropriate level and mechanisms for tracking parts (e.g., by UUID or using RFID to help automate the process. A sample use-case is the difficulty of knowing which parts are located behind-LO panels).
   - The AF SAF/CO should establish minimum required data requirements for all programs of record (whether or not the AF chooses to purchase this data from the contractor), such that there may be certain types of data that should exist in the event it is required in the future.
   - The AF SAF/CO should outline recommended and required data rights that future acquisition programs must obtain (e.g., a data dictionary to make sense of the rest of the data).
   - Advocate for tools and licensing for enterprise license agreements for analytical tools for end users; SPSS etc.

- Understand data generated under specific rights should be respected by users not originally intended

8. Summary

This workshop addressed data management requirements that need to be supported by the SAF/CO, in terms of VAULT compliance, as well as by the SAF/CO supported Data Architecture and Enterprise Information Models frameworks.

In particular the workshop addressed the needs that are derived from dealing not only with static data, but also of real-time data related to AF relevant systems and environments, and where these data are exploited and acquired through approaches integrating them with models of a symbiotic feed-back control loop. The modeling methods discussed here go beyond the Machine Learning methods which have captured recently the attention for "data analytics", as the ultimate objective is not "data analytics", by "system analytics" (that is the applications systems of interest to AF environments).

Future systems (and systems-of-systems)and data-and-models environments, in which data dynamically integrated with (systems') models were common in the scope of the technical directions addressed in the workshop. The workshop discussions led to consensus that such environments require on-the-fly data management (e.g., indexing, tagging, and other metadata creation); require support of models workflows (where models may invoke other models – other modalities and behaviors of an application system or of systems-of-systems), and data management, not only to deal with data of the applications systems, but also of the heterogeneous and resource adaptive computational, communication, and data platforms, to ensure optimized utilization of resources at all levels. A Data as a Service (DaaS) approach ensures timely and informed decision making,to accomplish the AF superiority tenets of "*extremely high-quality situational awareness, rapid decision making, and the ability to direct forces across domains and missions with continuous feedback*".

This report features material from presentations and discussions from the SAF/CO-sponsored workshop involving leaders in government, academia and industry. The VAULT set of key attributes, remains a valid focus, for numerous examples and use cases, anda basis for an end-to-end data architecture linking sensing to modeling and end-users.

A key VAULT scenario is the real-time tagging of data based upon computation over its content. The DDDAS paradigm (which incorporatesdata with models to guide computation and model execution, andguidescorresponding data flows in a real-time feedback-control-loop) generates such needs and also provides a methodology to manage such needs. Driving application areas for such environments and in context of the DDDAS methodologyexamined in the workshop include MDC2, ISR, SHM-PM and Cloud/IoT[38].

The report considers emerging themes in data and analytical requirements, providing example narratives of scenarios that a VAULT-based platform can enable (via data management, tagging, security, i.e., the VAULT principles), to support not only data-analytics but systems-analytics.The capabilities sought are delineated in the 5 key technical recommendations (as well as about 40 sub-recommendations) and two broad policy and acquisition recommendations for the removal of acquisition policy roadblocks. The Report

[38]Additional applications are discussed in New S&T Directions for New AF Capabilities and CDO Strategy Drivers.

also identifies a number of possible "use-cases" (identified in Section 4) to be considered by SAF/CO to drive and refine further the prioritization and implementation of the recommendations in this report.

References

Ali, A. A., *et al*., Retrospective-Cost-Based Adaptive Input and State Estimation for the Ionosphere–Thermosphere, *J. Aerospace Information Systems* 12:767-783, (2015).

Aved, A., *et al*., Dynamic data Driven Application Systems, Website, (2014).

Aved, A., *et al*., Regularized difference criterion for computing discriminants for dimensionality reduction, IEEE Transactions on Aerospace and Electronic Systems 53 (5), 2372-2384, (2017).

Bazilevs, Y., et al, Toward a computational steering framework for large-scale composite structures based, *Procedia Computer Science 9*: 1149 – 1158, (2012)

Bazilevs, Y., et al., Isogeometric Fatigue Damage Prediction in Large-Scale Composite Structures Driven by Dynamic Sensor Data, Journal of Applied Mechanics 82, 091008-1:091008-12 (2015)

Bernstein, D. S., *et al*., ``Correction of the Photoelectron Heating Efficiency Within the Global Ionosphere-Thermosphere Model Using Retrospective Cost Model Refinement,'' *J. Atmospheric Solar-Terrestrial Physics*, 124: 30--38, (2015)

Blasch, E., *et al*., *High-Level Information Fusion Management and Systems Design*, Artech, Norwood, MA, (2012).

Blasch, E., Enhanced Air Operations Using JView for an Air-Ground Fused Situation Awareness UDOP, *AIAA/IEEE Digital Avionics Systems Conference*, (2013).

Blasch, E., S. Ravela, A. Aved (eds.) *Handbook on Dynamic Data Driven Applications Systems*, Springer, (2018).

Blasch, E., *et al*., Methods of AI for Multimodal Sensing and Action for Complex Situations, AI Mag. 40(4):50-65 (2019)

Celik, N., *et al*., DDDAMS-based dispatch control in power networks, *Procedia Computer Sci*. 18: 1899 – 1908 (2013)

Darema, F., Grid computing and beyond: The context of dynamic data driven applications systems, IEEE (2005)

Du, L., *et al*., GARP-face: Balancing privacy protection and utility preservation in face de-identification, IJCB, (2014).

Guerci, J. R., et al., A new MIMO clutter model for cognitive radar, *IEEE Radar Conference (RadarConf)*, (2016).

Hariri, S., *et al*., DDDAS-Based Resilient Cyber Battle Management Services (D-RCBMS), HiPCW, (2015).

Insaurralde, C.C., *et al.* Ontologies in Aeronautics, Advances in Aeronautical Informatics: Technologies Towards Flight 4.0 (2018)

Imai, S., *et al*., Airplane Flight Safety Using Error-Tolerant Data Stream Processing, *IEEE Aerospace and Electronics Systems Magazine*, 32(4):4-17 (2017).

Karaman S., *et al*., Exploitation by Informed Exploration between Isolated Operatives for Information-Theoretic Data Harvesting, IEEE CDC (2015)

Krishnamurthy, V., et al., How to calibrate your adversary's capabilities? Inverse filtering for counter-autonomous systems," *IEEE Transactions on Signal Processing*, 67:6511–6525, (2019).

Mohseni, K., et al., Symplectic model reduction of Hamiltonian systems, *SIAM J. on Sci. Comp.* 38(1), A1-27, (2016).

Liu, S., IR2VI: Enhanced Night Environmental Perception by Unsupervised Thermal Image Translation, CVPRW, (2018)

Salerno, J. J., et al., Evaluating algorithmic techniques in supporting situation awareness, *Proc. SPIE 5813*, 96-104 (2005).

Shen, D., et al., A game theoretic data fusion aided path planning approach for cooperative UAV ISR, Aerospace, (2008).

Snidaro, L., et al. (eds.), *Context-Enhanced Information Fusion*: *Boosting Real-World Performance with Domain Knowledge*, Springer, 2016.

Uzkent, B., et al., Real-time Vehicle Tracking in Aerial Video using Hyperspectral Features, IEEE CVPRW, (2016).

Varela, C. A, Dynamic Data Driven Avionics Systems. *Streaming Tech. Requirements, App. and Middleware*. (2015).

Wang, Z., *et al*., A Low-Cost Near Real Time Two-UAS Based UWB Emitter Monitoring System, *IEEE AESS Magazine*, 30(11): 4-11, Nov. (2015)

Wei, M., *et al*. Multi-pursuer multi-evader pursuit-evasion games with jamming confrontation, *Journal of Aerospace Computing, Information, and Communication*, 4 (3), 693-706 (2007).

Willcox, K., *et al*, Dynamic Data-Driven Reduced-Order Models. *Comp. Methods in App. Mechanics and Eng*., (2015).

Wu, R., *et al*., A Container-based Elastic Cloud Architecture for Pseudo Real-time Exploitation of Wide Area Motion Imagery (WAMI) Stream, *The J. of Signal Processing Systems*, 88(2):219-231 (2017)

Xu, R., *et al*., Blendcac: A smart contract enabled decentralized capability-based access control mechanism for the iot, *Computers* 7 (3), 39. (2018).

Xu, R. et al., Blendcac: A blockchain-enabled decentralized capability-based access control for iots, IEEE iThings (2018).

Zheng, Y., *et al*., *Multispectral Image Fusion and Colorization*, SPIE Press, (2018)

Zikopoulos, P., *et al*., *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw, (2012).