Monday, April 12th

All times in Eastern Standard Time (EST).

Use of "A / B" in the schedule denotes parallel sessions; choose one session to attend.

8:45 AM-8:55 AM

Opening Remarks

Norton Schwartz

Institute for Defense Analyses

Abstract: Norton A. Schwartz serves as President of the Institute for Defense Analyses (IDA), a nonprofit corporation operating in the public interest. IDA manages three Federally Funded Research and Development Centers that answer the most challenging U.S. security and science policy questions with objective analysis leveraging extraordinary scientific, technical, and analytic expertise. At IDA, General Schwartz (U.S. Air Force, retired) directs the activities of more than 1,000 scientists and technologists employed by IDA. General Schwartz has a long and prestigious career of service and leadership that spans over 5 decades. He was most recently President and CEO of Business Executives for National Security (BENS). During his 6-year tenure at BENS, he was also a member of IDA's Board of Trustees. Prior to retiring from the U.S. Air Force, General Schwartz served as the 19th Chief of Staff of the U.S. Air Force from 2008 to 2012. He previously held senior joint positions as Director of the Joint Staff and as the Commander of the U.S. Transportation Command. He began his service as a pilot with the airlift evacuation out of Vietnam in 1975. General Schwartz is a U.S. Air Force Academy graduate and holds a master's degree in business administration from Central Michigan University. He is also an alumnus of the Armed Forces Staff College and the National War College. He is a member of the Council on Foreign Relations and a 1994 Fellow of Massachusetts Institute of Technology's Seminar XXI. General Schwartz has been married to Suzie since 1981.

8:55 AM-9:05 AM

Opening Remarks

Raymond O'Toole

DOT&E

Abstract: Dr. O'Toole is the Acting Director, Operational Test and Evaluation as of January 20, 2021. Dr. O'Toole was appointed as the Principal Deputy Director, Operational Test and Evaluation in February 2020. In this capacity he is the principal staff assistant for all functional areas assigned to the office. He participates in the formulation, development, advocacy, and oversight of policies of the Secretary of Defense and in the development and implementation of test and test resource programs. He supports the Director in the planning, conduct, evaluation and reporting of operational and live fire testing. He serves as the Appropriation Director and Comptroller for the Operational Test and Evaluation, Defense Appropriation and the principal advisor to the Director on all Planning, Programming, and Budgeting System matters. Dr. O'Toole is the former Deputy Director for Naval Warfare within DOT&E. He oversaw the operational and live-fire testing of ships and submarines and their associated sensors; combat and communications systems, and weapons. He was also responsible for overseeing the adequacy of the test infrastructure and resources to support operational and live-fire testing for all acquisition programs across the Defense Department. Dr. O'Toole was previously an employee of the Naval Sea Systems Command as the Deputy Group Director of Aircraft Carrier Design and Systems Engineering. Prior to that, he was the Director of Systems Engineering Division (Submarines and Undersea Systems) where he led a diverse team of engineers who supported all Submarine Program Managers. His other assignments include being a Ship Design Manager/Navy's Technical Authority for the USS VIRGINIA Class submarines during design and new construction and for Amphibious Ships, Auxiliary Ships, and Command & Control Ships during inservice operations. Dr. O'Toole has also held other positions within the Department of Defense such as Deputy Program Executive Officer (Maritime and Rotary Wing) at the United States Special Operations Acquisition Command, Staff to the Deputy Assistant Secretary of the Navy for Research, Development & Acquisition (Ship Programs), and Deputy Director of Regional Maintenance for COMPACFLT (N43). In addition, Dr. O'Toole has over 30 years of experience as a Naval Officer (Active and Reserve) retiring at the rank of CAPTAIN. His significant tours include 5 Commanding Officer tours. Dr. Raymond D. O'Toole, Jr. is a native of Long Island NY and a graduate of the State University of New York – Maritime College earning a Bachelor of Engineering in Marine Engineering. He also holds a Master of Engineering Degree in Systems Engineering from Virginia Polytechnic Institute and State University, a Master of Science Degree in National Resource Strategy from the Industrial College of the Armed Forces, and a Doctorate in Engineering in the field of Engineering Management from the George Washington University, where he is now a Professional Lecturer of Engineering Management and Systems Engineering. He has received the SECDEF Meritorious Civilian Service Award and the USN Meritorious and Superior Civilian Service Awards.

9:00 AM-9:45 AM

Keynote
Assessing Human-Autonomy Interaction in Driving-Assist Settings

Mary “Missy” Cummings

Duke University

Abstract: In order to determine how the perception, Autopilot, and driver monitoring systems of Tesla Model 3s interact with one another, and also to determine the scale of between- and within-car variability, a series of four on-road tests were conducted. Three sets of tests were conducted on a closed track and one was conducted on a public highway. Results show wide variability across and within three Tesla Model 3s, with excellent performance in some cases but also likely catastrophic performance in others. This presentation will not only highlight how such interactions can be tested, but also how results can inform requirements and designs of future autonomous systems.

9:45 AM-10:00 AM — *Break*

10:00 AM-11:30 AM

Session 1A
Theme: Modeling & Simulation
Tom Johnson, Session Chair

Advancements in Characterizing Warhead Fragmentation Events

John Haman

Institute for Defense Analyses

10:00 AM – 10:30 AM

Abstract: Fragmentation analysis is a critical piece of the live fire test and evaluation (LFT&E) of lethality and vulnerability aspects of warheads. But the traditional methods for data collection are expensive and laborious. New optical tracking technology is promising to increase the fidelity of fragmentation data, and decrease the time and costs associated with data collection. However, the new data will be complex, three dimensional ‘fragmentation clouds’, possibly with a time component as well. This raises questions about how testers can effectively summarize spatial data to draw conclusions for sponsors. In this briefing, we will discuss the Bayesian spatial models that are fast and effective for characterizing the patterns in fragmentation data, along with several exploratory data analysis techniques that help us make sense of the data. Our analytic goals are to – Produce simple statistics and visuals that help the live fire analyst compare and contrast warhead fragmentations; – Characterize important performance attributes or confirm design/spec compliance; and – Provide data methods that ensure higher fidelity data collection translates to higher fidelity modeling and simulation down the line. This talk is a version of the first-step feasibility study IDA is taking – hopefully much more to come as we continue to work on this important topic.

Modeling and Simulation in Support of the Decision Analysis Process

Michael Greco

U.S. Army CCDC Armaments Center

10:30 AM – 11:00 AM

Abstract: Informed enterprise and program decision making is central to DoD’s Digital Engineering’s purpose statement. Decision analysis serves as a key mechanism to link the voice of the sponsor/end user with the voice of the engineer and the voice of the budgetary analyst in order to enable a closed loop requirements writing approach that is informed by rigorous assessments of a broad range of system-level alternatives across a thorough set of stakeholder value criteria to include life-cycle costs, schedule, performance,

Session 1B
Theme: Integrated Testing
Joe Warfield, Session Chair

A Framework for Efficient Operational Testing through Bayesian Adaptive Design

Victoria Sieck

University of New Mexico / Air Force Institute of Technology

10:00 AM – 10:30 AM

Abstract: When developing a system, it is important to consider system performance from a user perspective. This can be done through operational testing—assessing the ability of representative users to satisfactorily accomplish tasks or missions with the system in operationally-representative environments. This process can be expensive and time-consuming, but is critical for evaluating a system. We show how an existing design of experiments (DOE) process for operational testing can be leveraged to construct a Bayesian adaptive design. This method, nested within the larger design created by the DOE process, allows interim analyses using predictive probabilities to stop testing early for success or futility. Furthermore, operational environments with varying probabilities of encountering are directly used in product evaluation. Representative simulations demonstrate how these interim analyses can be used in an operational test setting, and reductions in necessary test events are shown. The method allows for using either weakly informative priors when data from previous testing is not available, or for priors built using developmental testing data when it is available. The proposed method for creating priors using developmental testing data allows for more flexibility in which data can be incorporated into analysis than the current process does, and demonstrates that it is possible to get more precise parameter estimates. This method will allow future testing to be conducted in less time and at less expense, on average, without compromising the ability of the existing process to verify the system meets the user’s needs.

A Great Test Requires a Great Plan

Aaron Ramert

Scientific Test and Analytics Techniques Center of Excellence (STAT COE)

10:30 AM – 11:00 AM

Abstract: The Scientific Test and Analysis Techniques (STAT) process is designed to provide structure for a test team to progress from a requirement to decision quality information. The four phases of the STAT process are Plan, Design, Execute, and Analyze. Within the Test and

and long term viability. . The decision analytics framework employed by the U.S. Army's Combat Capabilities Development Command (CCDC) Armaments Center (AC) is underpinned by a state-of-the-art modeling and simulation framework called PRISM (Performance Related and Integrated Suite of Models) developed at CCDC-AC. PRISM was designed in a way to allow performance estimates of a weapon system to evolve as more information and higher fidelity representations of those systems become available. PRISM provides the most up to date performance estimates into the decision analysis framework so that decision makers have the best information available when making complex strategic decisions. This briefing will unpack PRISM and highlight the model design elements that make it the foundation of CCDC-AC's weapon system architecture and design decision making process.

Uncertainty Quantification and Sensitivity Analysis Methodology for AJEM

Craig Andres

U.S. Army CCDC Data & Analysis Center

11:00 AM – 11:30 AM

Abstract: The Advanced Joint Effectiveness Model (AJEM) is a joint forces model developed by the U.S. Army that is used in vulnerability and lethality (V/L) predictions for threat/target interactions. This complex model primarily generates a probability response for various components, scenarios, loss of capabilities, or summary conditions. Sensitivity analysis (SA) and uncertainty quantification (UQ), referred to jointly as SA/UQ, are disciplines that provide the working space for how model estimates changes with respect to changes in input variables. A comparative measure that will be used to characterize the effect of an input change on the predicted outcome was developed and is reviewed and illustrated in this presentation. This measure provides a practical context that stakeholders can better understand and utilize. We show graphical and tabular results using this measure.

Evaluation (T&E) community we tend to focus on the quantifiable metrics and the hard science of testing, which are the Design and the Analyze phases. At the STAT Center of Excellence (COE) we have emphasized an increased focus on the planning phase and in this presentation we focus on the elements necessary for a comprehensive planning session. In order to efficiently and effectively test a system it is vital that the test team understand the requirements, the System Under Test (SUT) to include any subsystems that will be tested, and the test facility. To accomplish this the right team members with the necessary knowledge must be in the room and prepared to present their information and have an educated discussion to arrive at a comprehensive agreement about the desired end stated of the test. Our recommendations for the initial planning meeting are based on a thorough study of the STAT process and on lessons learned from actual planning meetings.

Prior Formulation in a Bayesian Analysis of Biomechanical Data

Amanda French

Johns Hopkins University Applied Physics Laboratory

11:00 AM – 11:30 AM

Abstract: Biomechanical experiments investigating the failure modes of biological tissue require a significant investment of time and money due to the complexity of procuring, preparing, and testing tissue. Furthermore, the potentially destructive nature of these tests makes repeated testing infeasible. This leads to experiments with notably small sample sizes in light of the high variance common to biological material. When the goal is to estimate parameters for an analytic artifact such as an injury risk curve (IRC), which relates an input quantity to a probability of injury, small sample sizes result in undesirable uncertainty. One way to ameliorate this effect is through a Bayesian approach, incorporating expert opinion and previous experimental data into a prior distribution. This has the advantage of leveraging the information contained in expert opinion and related experimental data to obtain faster convergence to an appropriate parameter estimation with a desired certainty threshold. We explore several ways of implementing Bayesian methods in a biomechanical setting, including permutations on the use of expert knowledge and prior experimental data. Specifically, we begin with a set of experimental data from which we generate a reference IRC. We then elicit expert predictions of the 10th and 90th quantiles of injury, and use them to formulate both uniform and normal prior distributions. We also generate priors from qualitatively similar experimental data, both directly on the IRC parameters and on the injury quantiles, and explore the use of weighting schemes to assign more influence to better datasets. By adjusting the standard deviation and shifting the mean, we can create priors of variable quality. Using a subset of the experimental data in conjunction with our derived priors, we then re-fit the IRC and compare it to the reference curve. For all methods we will measure the certainty, speed of convergence, and accuracy relative to the reference IRC, with the aim of recommending a best practices approach for the application of Bayesian methods

		in this setting. Ultimately an optimized approach for handling small samples sizes with Bayesian methods has the potential to increase the information content of individual biomechanical experiments by integrating them into the context of expert knowledge and prior experimentation.
11:30 AM-12:00 PM	*Break*	
12:00 PM-1:00 PM	Roundtables *Pre-Registration Required*	
	Roundtable Discussion Theme: Special Topics **Opportunities and Challenges for Openly Publishing Statistics Research for National Defense** Lyndsay Shand *Sandia National Laboratories* 12:00 PM – 1:00 PM **Abstract:** Openly publishing on statistics for defense and national security poses certain challenges and is often not straightforward, but research in this area is important to share with the open community to advance the field. Since statistical research for national defense applications is rather niche, target journals and audiences are challenging to identify. Adding an additional hurdle, much of the data for practical implementation is sensitive and surrogate datasets must be relied upon for publication. Lastly, many statisticians in these areas do not face the same expectations to openly publish as their colleagues. This roundtable is an opportunity for statisticians in the defense and national security community to come together and discuss the importance and challenges for publishing within this space. Participants will be asked to share challenges and successes related to publishing research for national defense applications. A handout summarizing common challenges and tips for succeeding will be provided. Specific topics for discussion will include: What expectations exist for statisticians to publish in this community? Are these expectations reasonable? Are you encouraged and supported by your funding institution to openly publish? Are there opportunities for collaborative work across institutions that might further encourage publications?	Roundtable Discussion Theme: Special Topics **Overcoming Challenges and Applying Sequential Procedures to T&E** Rebecca Medlin *Institute for Defense Analyses* 12:00 PM – 1:00 PM **Abstract:** The majority of statistical analyses involves observing a fixed set of data and analyzing those data after the final observation has been collected to draw some inference about the population from which they came. Unlike these traditional methods, sequential analysis is concerned with situations for which the number, pattern, or composition of the data is not determined at the start of the investigation but instead depends upon the information acquired throughout the course of the investigation. Expanding the use of sequential analysis in DoD testing has the potential to save substantial test dollars and decrease test time. However, switching from traditional to sequential planning will likely induce unique challenges. The goal of this round table is to provide an open forum for topics related to sequential analyses. We aim to discuss potential challenges, identify potential ways to overcome them, and talk about successful stories of sequential analyses implementation and lessons learned. Specific questions for discussion will be provided to participants prior to the event.
1:00 PM-2:30 PM	Mini Tutorial 1A Theme: Modeling & Simulation *Sarah Burke, Session Chair* **Statistical Approaches to V&V and Adaptive Sampling in M&S – Part 1** Jim Wisnowski *Adsurgo LLC* 1:00 PM – 2:30 PM **Abstract:** Leadership has placed a high premium on analytically defensible results for M&S Verification and Validation. This mini-tutorial will provide a quick overview of relevant standard methods to establish equivalency in mean, variance, and distribution shape such as Two One-Sided Tests (TOST), K-S tests, Fisher's Exact, and Fisher's	Mini Tutorial 1B Theme: Modeling Human-System Interaction *Rebecca Medlin, Session Chair* **Introduction to Qualitative Methods – Part 1** Kristina Carter *Institute for Defense Analyses* 1:00 PM – 2:30 PM **Abstract:** Qualitative data, captured through freeform comment boxes, interviews, focus groups, and activity observation is heavily employed in testing and evaluation (T&E). The qualitative research approach can offer many benefits, but knowledge of how to implement methods, collect data, and analyze data according to rigorous qualitative research standards is not broadly understood

Combined Probability. The focus will be on more advanced methods such as the equality between model parameters in statistical emulators versus live tests (Hotelling T2, loglinear variance), equivalence of output curves (functional data analysis), and bootstrap methods. Additionally, we introduce a new method for near real-time adaptive sampling that places the next set of M&S runs at boundary regions of high gradient in the responses to more efficiently characterize complex surfaces such as those seen in autonomous systems.

Statistical Approaches to V&V and Adaptive Sampling in M&S – Part 2

Jim Simpson

JK Analytics

–

within the T&E community. This tutorial offers insight into the foundational concepts of method and practice that embody defensible approaches to qualitative research. We discuss where qualitative data comes from, how it can be captured, what kind of value it offers, and how to capitalize on that value through methods and best practices.

Introduction to Qualitative Methods – Part 2

Daniel Hellman

Institute for Defense Analyses

–

Introduction to Qualitative Methods – Part 3

Emily Fedele

Institute for Defense Analyses

–

2:30 PM-2:45 PM

Break

2:45 PM-4:15 PM

Session 2A
Theme: Special Topics
Ken Toro, Session Chair

Statistical Engineering in Practice

Angie Patterson

GE Aviation

2:45 PM – 3:00 PM

Abstract: Problems faced in defense and aerospace often go well beyond textbook problems presented in academic settings. Textbook problems are typically very well defined, and can be solved through application of one “correct” tool. Conversely, many defense and aerospace problems are ill-defined, at least initially, and require an overall strategy to attack, one involving multiple tools and often multiple disciplines. Statistical engineering is an approach recently developed for addressing large, complex, unstructured problems, particularly those for which data can be effectively utilized. This session will present a brief overview of statistical engineering, and how it can be applied to engineer solutions to complex problems. Following this introduction, two case studies of statistical engineering will be presented, to illustrate the concepts.

Statistical Engineering in Practice

Peter Parker

NASA Langley

3:00 PM – 3:25 PM

Statistical Engineering in Practice

Roger Hoerl

Union College

3:25 PM – 3:50 PM

Session 2B
Theme: Test & Evaluation Methods for Emerging Technology and Domains
Jason Schlup, Session Chair

Operational Cybersecurity Test and Evaluation of Non-IP and Wireless Networks

Peter Mancini

Institute for Defense Analyses

2:45 PM – 3:15 PM

Abstract: Nearly all land, air, and sea maneuver systems (e.g. vehicles, ships, aircraft, and missiles) are becoming more software-reliant and blending internal communication across both Internet Protocol (IP) and non-IP buses. IP communication is widely understood among the cybersecurity community, whereas expertise and available test tools for non-IP protocols such as Controller Area Network (CAN) and MIL-STD-1553 are not as commonplace. However, a core tenet of operational cybersecurity testing is to asses all potential pathways of information exchange present on the system, to include IP and non-IP. In this presentation, we will introduce a few non-IP protocols (e.g. CAN, MIL-STD-1553) and provide a live demonstration of how to attack a CAN network using malicious message injection. We will also discuss how potential cyber effects on non-IP busses can lead to catastrophic mission effects to the target system.

Cybersecurity Metrics and Quantification: Problems, Some Results, and Research Directions

Shouhuai Xu

University of Colorado Colorado Springs

3:15 PM – 3:45 PM

Abstract: Cybersecurity Metrics and Quantification is a fundamental but notoriously hard problem. It is one of the

Statistical Engineering in Practice

Alex Varbanov

Procter and Gamble

3:50 PM – 4:15 PM

pillars underlying the emerging Science of Cybersecurity. In this talk, I will describe a number of cybersecurity metrics quantification research problems that are encountered in evaluating the effectiveness of a range of cyber defense tools. I will review the research results we have obtained over the past years. I will also discuss future research directions, including the ones that are undertaken in my research group.

Collaborative Human AI Red Teaming

Galen Mullins

Johns Hopkins University Applied Physics Laboratory

3:45 PM – 4:15 PM

Abstract: The Collaborative Human AI Red Teaming (CHART) project is an effort to develop an AI Collaborator which can help human test engineers quickly develop test plans for AI systems. CHART was built around processes developed for cybersecurity red-teaming. Using a goal-focused approach based upon iteratively testing and attacking a system then updating the testers model to discover novel failure modes not discovered by traditional T&E processes. Red teaming is traditionally a time intensive process which requires subject matter expert to study the system they are testing for months in order to develop attack strategies. CHART will accelerate this process by guiding the user through the process of diagraming the AI system under test and drawing upon a pre-established body of knowledge to identify the most probably vulnerabilities. CHART was provided internal seedling funds during FY20 to perform a feasibility study of the technology. During this period the team developed a taxonomy of AI vulnerabilities and an ontology of AI irruptions. Irruptions being events (either caused by a malicious actor or unintended consequences) which trigger the vulnerability and lead to an undesirable result. Using this taxonomy we built a threat modeling tool that allows users to diagram their AI system and identifies all the possible irruptions which could occur. This initial demonstration was based around two scenarios. An smartphone-based ECG system for telemedicine and a UAV trained reinforcement learning to avoid mid-air collisions. In this talk we will first discuss how Red Teaming differs from adversarial machine learning and traditional testing and evaluation. Next, we will provide an overview of how industry is approaching the problem of AI Red Teaming and how our approach differs. Finally, we will discuss how we developed our taxonomy of AI vulnerabilities, how to apply goal-focused testing to AI systems, and our strategy for automatically generating test plans.

4:15 PM-4:30 PM	*Break*	
4:30 PM-5:30 PM	Session 3A Theme: Modeling & Simulation *Melissa Rivers, Session Chair*	Session 3B Theme: Modeling & Simulation *Han Yi, Session Chair*

Verification and Validation of Elastodynamic Simulation Software for Aerospace Research

Erik Frankforter
NASA Langley Research Center

4:30 PM – 5:00 PM

Abstract: Physics-based simulation of nondestructive evaluation (NDE) inspection can help to advance the inspectability and reliability of mechanical systems. However, NDE simulations applicable to non-idealized mechanical components often require large compute domains and long run times. This has prompted development of custom NDE simulation software tailored to high performance computing (HPC) hardware. Verification and validation (V&V) is an integral part of developing this software to ensure implementations are robust and applicable to inspection problems, producing tools and simulations suitable for computational NDE research. This presentation addresses factors common to V&V of several elastodynamic simulation codes applicable to ultrasonic NDE. Examples are drawn from in-house simulation software at NASA Langley Research Center, ranging from ensuring reliability in a 1D heterogeneous media wave equation solver to the V&V needs of 3D cluster-parallel elastodynamic software. Factors specific to a research environment are addressed, where individual simulation results can be as relevant as the software product itself. Distinct facets of V&V are discussed including testing to establish software reliability, employing systematic approaches for consistency with fundamental conservation laws, establishing the numerical stability of algorithms, and demonstrating concurrence with empirical data. This talk also addresses V&V practices for small groups of researchers. This includes establishing resources (e.g. time and personnel) for V&V during project planning to mitigate and control the risk of setbacks. Similarly, we identify ways for individual researchers to use V&V during simulation software development itself to both speed up the development process and reduce incurred technical debt.

Challenges in Verification and Validation of CFD for Industrial Aerospace Applications

Andrew Cary
Boeing Research and Technology

5:00 PM – 5:30 PM

Abstract: Verification and validation represent important steps for appropriate use of CFD codes and it is presently considered the user's responsibility to ensure that these steps are completed. Inconsistent definitions and use of these terms in aerospace complicate the effort. For industrial-use CFD codes, there are a number of challenges that can further confound these efforts including varying grid topology, non-linearities in the solution, challenges in isolating individual components, and difficulties in finding validation experiments. In this presentation, a number of these challenges will be reviewed with some specific examples that demonstrate why verification is much more involved and challenging than typically implied in numerical method courses, but remains an important exercise. Some of the challenges associated with validation will also be highlighted using a range of different cases, from canonical

Estimating Pure-Error from Near Replicates in Design of Experiments

Caleb King
SAS Institute

4:30 PM – 5:00 PM

Abstract: In design of experiments, setting exact replicates of factor settings enables estimation of pure-error; a model-independent estimate of experimental error useful in communicating inherent system noise and testing of model lack-of-fit. Often in practice, the factor levels for replicates are precisely measured rather than precisely set, resulting in near-replicates. This can result in inflated estimates of pure-error due to uncompensated set-point variation. In this article, we review previous strategies for estimating pure-error from near-replicates and propose a simple alternative. We derive key analytical properties and investigate them via simulation. Finally, we illustrate the new approach with an application.

Surrogate Models and Sampling Plans for Multi-fidelity Aerodynamic Performance Databases

Kevin Quinlan
Lawrence Livermore National Laboratory

5:00 PM – 5:30 PM

Abstract: Generating aerodynamic coefficients can be computationally expensive, especially for the viscous CFD solvers in which multiple complex models are iteratively solved. When filling large design spaces, utilizing only a high accuracy viscous CFD solver can be infeasible. We apply state-of-the-art methods for design and analysis of computer experiments to efficiently develop an emulator for high-fidelity simulations. First, we apply a cokriging model to leverage information from fast low-fidelity simulations to improve predictions with more expensive high-fidelity simulations. Combining space-filling designs with a Gaussian process model-based sequential sampling criterion allows us to efficiently generate sample points and limit the number of costly simulations needed to achieve the desired model accuracy. We demonstrate the effectiveness of these methods with an aerodynamic simulation study using a conic shape geometry. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Release Number: LLNL-ABS-818163

flow elements to complete aircraft models. Benchmarking is often used to develop confidence in CFD solutions for engineering purposes, but falls short of validation in the absence of being able to predict bounds on the simulation error. The key considerations in performing benchmarking and validation will be highlighted and some current shortcomings in practice will be presented, leading to recommendations for conducting validation exercises. CFD workshops have considerably improved in their application of these practices, but there continues to be need for additional steps.

Certification by Analysis: A 20-year Vision for Virtual Flight and Engine Testing

Timothy Mauery

Boeing

5:30 PM – 6:00 PM

Abstract: Analysis-based means of compliance for airplane and engine certification, commonly known as “Certification by Analysis” (CbA), provides a strong motivation for the development and maturation of current and future flight and engine modeling technology. The most obvious benefit of CbA is streamlined product certification testing programs at lower cost while maintaining equivalent levels of safety. The current state of technologies and processes for analysis is not sufficient to adequately address most aspects of CbA today, and concerted efforts to drastically improve analysis capability are required to fully bring the benefits of CbA to fruition. While the short-term cost and schedule benefits of reduced flight and engine testing are clearly visible, the fidelity of analysis capability required to realize CbA across a much larger percentage of product certification is not yet sufficient. Higher-fidelity analysis can help reduce the product development cycle and avoid costly and unpredictable performance and operability surprises that sometimes happen late in the development cycle. Perhaps the greatest long-term value afforded by CbA is the potential to accelerate the introduction of more aerodynamically and environmentally efficient products to market, benefitting not just manufacturers, but also airlines, passengers, and the environment. A far-reaching vision for CbA has been constructed to offer guidance in developing lofty yet realizable expectations regarding technology development and maturity through stakeholder involvement. This vision is composed of the following four elements: The ability to numerically simulate the integrated system performance and response of full-scale airplane and engine configurations in an accurate, robust, and computationally efficient manner. The development of quantified flight and engine modeling uncertainties to establish appropriate confidence in the use of numerical analysis for certification. The rigorous validation of flight and engine modeling capabilities against full-scale data from critical airplane and engine testing. The use of flight and engine modeling to enable Certification by Simulation. Key technical challenges include the ability to accurately predict airplane and engine performance for a single discipline, the robust and efficient integration of multiple disciplines, and the appropriate modeling of system-level assessment. Current modeling methods lack the

capability to adequately model conditions that exist at the edges of the operating envelope where the majority of certification testing generally takes place. Additionally, large-scale engine or airplane multidisciplinary integration has not matured to the level where it can be reliably used to efficiently model the intricate interactions that exist in current or future aerospace products. Logistical concerns center primarily on the future High Performance Computing capability needed to perform the large number of computationally intensive simulations needed for CbA. Complex, time-dependent, multidisciplinary analyses will require a computing capacity increase several orders of magnitude greater than is currently available. Developing methods to ensure credible simulation results is critically important for regulatory acceptance of CbA. Confidence in analysis methodology and solutions is examined so that application validation cases can be properly identified. Other means of measuring confidence such as uncertainty quantification and “validation-domain” approaches may increase the credibility and trust in the predictions. Certification by Analysis is a challenging long-term endeavor that will motivate many areas of simulation technology development, while driving the potential to decrease cost, improve safety, and improve airplane and engine efficiency. Requirements to satisfy certification regulations provide a measurable definition for the types of analytical capabilities required for success. There is general optimism that CbA is a goal that can be achieved, and that a significant amount of flight testing can be reduced in the next few decades.

Tuesday, April 13th

All times in Eastern Standard Time (EST).

Use of "A / B" in the schedule denotes parallel sessions; choose one session to attend.

8:45 AM-9:45 AM

Session 4A
Theme: Modeling & Simulation
Ying Shi, Session Chair

Physics-Informed Deep Learning for Modeling and Simulation under Uncertainty

Patrick Leser

NASA Langley Research Center

8:45 AM – 9:15 AM

Abstract: Certification by analysis (CBA) involves the supplementation of expensive physical testing with modeling and simulation. In high-risk fields such as defense and aerospace, it is critical that these models accurately represent the real world, and thus they must be verified, validated and provide measures of uncertainty. While machine learning (ML) algorithms such as deep neural networks have seen significant success in low-risk sectors, they are typically opaque, difficult to interpret and often fail to meet these stringent requirements. Recently, a Department of Energy (DOE) report was released on the concept of scientific machine learning (SML) [1] with the aim of generally improving confidence in ML and enabling broader use in the scientific and engineering communities. The report identified three critical attributes that ML algorithms should possess: domain-awareness, interpretability, and robustness. Recent advances in physics-informed neural networks (PINNs) are promising in that they can provide both domain awareness and a degree of interpretability [2, 3, 4] by using governing partial differential equations as constraints during training. In this way, PINNs output physically admissible, albeit deterministic, solutions. Another noteworthy deep learning algorithm is the generative adversarial network (GAN), which can learn probability distributions [5] and provide robustness through uncertainty quantification. A limited number of works have recently demonstrated success by combining these two methods into what is referred to as a physics-informed GAN, or PIGAN [6, 7]. The PIGAN has the ability to produce physically admissible, non-deterministic predictions as well as solve non-deterministic inverse problems, potentially meeting the goals of domain awareness, interpretability, and robustness. This talk will present an introduction to PIGANs as well as an example of current NASA research implementing these networks. REFERENCES [1] Nathan Baker, Frank Alexander, Timo Bremer, Aric Hagberg, Yannis Kevrekidis, Habib Najm, Manish Parashar, Abani Patra, James Sethian, Stefan Wild, Karen Willcox, and Steven Lee. Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence. Technical report, USDOE Office of Science (SC) Washington, DC (United States), 2019. [2] Maziar Raissi, Paris Perdikaris, and George Karniadakis. Physics-informed

Session 4B
Theme: Special Topics
Matthew Avery, Session Chair

Empirical Analysis of COVID-19 in U.S. States and Counties

Emily Heuring

Institute for Defense Analyses

8:45 AM – 9:15 AM

Abstract: The zoonotic emergence of the coronavirus SARS-CoV-2 at the beginning of 2020 and the subsequent global pandemic of COVID-19 has caused massive disruptions to economies and health care systems, particularly in the United States. Using the results of serology testing, we have developed true prevalence estimates for COVID-19 case counts in the U.S. over time, which allows for more clear estimates of infection and case fatality rates throughout the course of the pandemic. In order to elucidate policy, demographic, weather, and behavioral factors that contribute to or inhibit the spread of COVID-19, IDA compiled panel data sets of empirically derived, publicly available COVID-19 data and analyzed which factors were most highly correlated with increased and decreased spread within U.S. states and counties. These analyses lead to several recommendations for future pandemic response preparedness.

Spatio-Temporal Modeling of Pandemics

Nicholas Clark

West Point

9:15 AM – 9:45 AM

Abstract: The spread of COVID-19 across the United States provides an interesting case study in the modeling of spatio-temporal data. In this breakout session we will provide an overview of commonly used spatio-temporal models and demonstrate how Bayesian Inference can be performed using both exact and approximate inferential techniques. Using COVID data, we will demonstrate visualization techniques in R and introduce "off-the-shelf" spatio-temporal models. We will introduce participants to the Integrated Nested LaPlace approximation (INLA) methodology and show how results from this technique compare to using Markov Chain Monte Carlo (MCMC) techniques. Finally, we will demonstrate the short-falls in using "off-the-shelf" models and show how epidemiological motivated partial differential equations can be used to generate spatio-temporal models and discuss inferential issues when we move away from common models.

neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378:686-707, 2019. [3] Alexandre Tartakovsky, Carlos Ortiz Marrero, Paris Perdikaris, Guzel Tartakovsky, and David Barajas-Solano. Learning parameters and constitutive relationships with physics informed deep neural networks. arXiv preprint arXiv:1808.03398, 2018. [4] Julia Ling, Andrew Kurzawski, and Jeremy Templeton. Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. Journal of Fluid Mechanics, 807:155-166, 2016. [5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, pages 2672-2680, 2014. [6] Liu Yang, Dongkun Zhang, and George Karniadakis. Physics-informed generative adversarial networks for stochastic differential equations. arXiv preprint arXiv:1811.02033, 2018. [7] Yibo Yang and Paris Perdikaris. Adversarial uncertainty quantification in physics-informed neural networks. Journal of Computational Physics, 394:136-152, 2019.

Fast, Unbiased Uncertainty Propagation with Multi-model Monte Carlo

Geoffrey Bomarito

NASA Langley Research Center

9:15 AM – 9:45 AM

Abstract: With the rise of machine learning and artificial intelligence, there has been a huge surge in data-driven approaches to solve computational science and engineering problems. In the context of uncertainty propagation, machine learning is often employed for the construction of efficient surrogate models (i.e., response surfaces) to replace expensive, physics-based simulations. However, relying solely on surrogate models without any recourse to the original high-fidelity simulation will produce biased estimators and can yield unreliable or non-physical results. This talk discusses multi-model Monte Carlo methods that combine predictions from both fast, low-fidelity models with reliable, high-fidelity simulations to enable efficient and accurate uncertainty propagation. For instance, the low-fidelity models could arise from coarsened discretizations in space/time (e.g., Multilevel Monte Carlo – MLMC) or from general data-driven or reduced order models (e.g., Multifidelity Monte Carlo – MFMC; Approximate Control Variates – ACV). Given a fixed computational budget and a collection of models of varying cost/accuracy, the goal of these methods is to optimally allocate and combine samples across the models. The talk will also present a NASA-developed open-source Python library that acts as a general multi-model uncertainty propagation capability. The effectiveness of the discussed methods and Python library is demonstrated on a trajectory simulation application. Here, orders of magnitude computational speedup and accuracy are obtained for predicting the landing location of an umbrella heat shield under significant uncertainties in initial state, atmospheric conditions, etc.

10:00 AM-11:30 AM

Session 5A
Theme: Modeling & Simulation
Ken Toro, Session Chair

Assessing Next-Gen Spacesuit Reliability: A Probabilistic Analysis Case Study at NASA

James Warner

NASA Langley Research Center

10:00 AM – 10:30 AM

Abstract: Under the Artemis program, the Exploration Extravehicular Mobility Unit (xEMU) spacesuit will ensure the safety of NASA astronauts during the targeted 2024 return to the moon. Efforts are currently underway to finalize and certify the xEMU design. There is a delicate balance between producing a spacesuit that is robust enough to safely withstand potential fall events while still satisfying stringent mass and mobility requirements. The traditional approach of considering worst case-type loading and applying conservative factors of safety (FoS) to account for uncertainties in the analysis was unlikely to meet the narrow design margins. Thus, the xEMU design requirement was modified to include a probability of no impact failure (PnIF) threshold that must be verified through probabilistic analysis. As part of a broader one year effort to help integrate modern uncertainty quantification (UQ) methodology into engineering practice at NASA, the certification of the xEMU spacesuit was selected as the primary case study. The project, led by NASA Langley Research Center (LaRC) under the Engineering Research & Analysis (R&A) Program in 2020, aimed to develop an end-to-end UQ workflow for engineering problems and to help facilitate reliability-based design at NASA. The main components of the UQ workflow included 1) sensitivity analysis to identify the most influential model parameters, 2) model calibration to quantified model parameter uncertainties using experimental data, and 3) uncertainty propagation for producing probabilistic model predictions and estimating reliability. Particular emphasis was placed on overcoming the common practical barrier of prohibitive computational expense associated with probabilistic analysis by leveraging state-of-the-art UQ methods and high performance computing (HPC). In lieu of mature computational models and test data for the xEMU at the time of the R&A Program, the UQ workflow for estimating PnIF was demonstrated using existing models and data from the previous generation of spacesuits (the Z-2). However, the lessons learned and capabilities developed in the process of the R&A are directly transferable to the ongoing xEMU certification effort and are currently being integrated in 2021. This talk provides an overview of the goals of and findings under NASA's UQ R&A project, focusing on the spacesuit certification case study. The steps of the UQ workflow applied to the Z-2 spacesuit using the available finite element method (FEM) models and impact test data will be detailed. The ability to quantify uncertainty in the most influential subset of FEM model input parameters and then propagate that uncertainty to estimates of PnIF is demonstrated. Since the FEM model of the full Z-2 assembly took nearly 1 day to execute just once, the advanced UQ

Session 5B
Theme: Test & Evaluation Methods for Emerging Technology and Domains
Jeffery Smith, Session Chair

Intelligent Integration of Limited-Knowledge IoT Services in a Cross-Reality Environment

Mark Dennison

U.S. Army DEVCOM Army Research Laboratory

10:00 AM – 10:30 AM

Abstract: The recent emergence of affordable, high-quality augmented-, mixed-, and virtual-reality (AR, MR, VR), technologies presents an opportunity to dramatically change the way users consume and interact with information. It has been shown that these immersive systems can be leveraged to enhance comprehension and accelerate decision-making in situations where data can be linked to spatial information, such as maps or terrain models. Furthermore, when immersive technologies are networked together, they allow for decentralized collaboration and provide perspective-taking not possible with traditional displays. However, enabling this shared space requires novel techniques in intelligent information management and data exchange. In this experiment, we explored a framework for leveraging distributed AI/ML processing to enable clusters of low-power, limited-functionality devices to deliver complex capabilities in aggregate to users distributed across the country collaborating simultaneously in a shared virtual environment. We deployed a motion detecting camera and triggered detection events to send information using a distributed request/reply worker framework to a remotely located YOLO image classification cluster. This work demonstrates the capability for various IoT and IoBT systems to invoke functionality without a priori knowledge of the specific endpoint to use to execute that functionality but by submitting a request based on a desired capability concept (e.g. image classification) with requiring only: 1) the knowledge of the broker location, 2) valid public/private key pair required to authenticate with the broker, and 3) the capability concept UUID and knowledge of request/reply formats used by that concept.

Multi-Agent Adaptive Coordinated Autonomy in Contested Battlefields

John Rogers

U.S. Army DEVCOM Army Research Laboratory

10:30 AM – 11:00 AM

Abstract: Autonomous multi-robot systems have the potential to augment the future force with enhanced capability while reducing the risk to human personnel in multi-domain operations (MDO). Mobile robots can constitute nodes in a heterogeneous Internet of Battlefield Things (IoBT); they can offer additional capability in the form of mobility to effectively make observations useful for planning and executing military operations against adversaries. In this talk, I will present the result of a series of field experiments where robots are tasked to perform military-relevant missions in realistic environments, in addition to describing the integration of mobile robot assets in the Multi-Purpose

methods and HPC utilization required to make the probabilistic analysis tractable are discussed. Finally, the lessons learned from conducting the case study are provided along with planned ongoing/future work for the xEMU certification in 2021.

Entropy-Based Adaptive Design for Contour Finding and Estimating Reliability

Austin Cole

Virginia Tech

10:30 AM – 11:00 AM

Abstract: In reliability, methods used to estimate failure probability are often limited by the costs associated with model evaluations. Many of these methods, such as multi-fidelity importance sampling (MFIS), rely upon a cheap, surrogate model like a Gaussian process (GP) to quickly generate predictions. The quality of the GP fit, at least in the vicinity of the failure region(s), is instrumental in propping up such estimation strategies. We introduce an entropy-based GP adaptive design that, when paired with MFIS, provides more accurate failure probability estimates and with higher confidence. We show that our greedy data acquisition scheme better identifies multiple failure regions compared to existing contour-finding schemes. We then extend the method to batch selection. Illustrative examples are provided on benchmark data as well as an application to the impact damage simulator of a NASA spacesuit design.

Debunking Stress Rupture Theories Using Weibull Regression Plots

Anne Driscoll

Virginia Tech

11:00 AM – 11:30 AM

Abstract: As statisticians, we are always working on new ways to explain statistical methodologies to non-statisticians. It is in this realm that we never underestimate the value of graphics and patience! In this presentation, we present a case study that involves stress rupture data where a Weibull regression is needed to estimate the parameters. The context of the case study results from a multi-stage project supported by NASA's Engineering Safety Center (NESC) where the objective was to assess the safety of composite overwrapped pressure vessels (COPVs). The analytical team was tasked with devising a test plan to model stress rupture failure risk in carbon fiber strands that encase the COPVs with the goal of understanding the reliability of the strands at use conditions for the expected mission life. While analyzing the data, we found that the proper analysis contradicts accepted theories about the stress rupture phenomena. In this talk, we will introduce ways to graph the stress rupture data to better explain the proper analysis and also explore assumptions.

Sensing Array Distributed Proving Ground (MSA-DPG) for the purpose of augmenting IoBT systems.

Army's Open Experimentation Test Range for Internet of Battlefield Things: MSA-DPG

Jade Freeman

U.S. Army DEVCOM Army Research Laboratory

11:00 AM – 11:30 AM

Abstract: One key feature of future Multi-Domain Operations (MDO) is expected to be the ubiquity of devices providing information connected in an Internet of Battlefield Things (IoBT). To this end, U.S. Army aims to advance the underlying science of pervasive and heterogeneous IoBT sensing, networking, and actuation. In this effort, IoBT experimentation testbed is an integral part of the capability development, which evaluates and validates the scientific theories, algorithms, and technologies integrated with C2 systems under the military scenarios. Originally conceived for this purpose, Multi-Purpose Sensing Area Distributed Proving Ground (MSA-DPG) is an open-range test bed developed by the Army Research Laboratory (ARL). We discuss the vision and the development of MSA-DPG and its fundamental roles of MSA-DPG in research serving the communities of Military Sciences.

11:30 AM-12:00 PM	*Break*
12:00 PM-1:00 PM	Roundtables *Pre-Registration Required*
	Roundtable Discussion Theme: Special Topics
	Roundtable Discussion Theme: Special Topics

The Role of the Statistics Profession in the DoD's Current AI Initiative

Laura Freeman

Virginia Tech

12:00 PM – 1:00 PM

Abstract: In 2019, the DoD unveiled comprehensive strategies related to Artificial Intelligence, Digital Modernization, and Enterprise Data Analytics. Recognizing that data science and analytics are fundamental to these strategies, in October 2020 the DoD has issued a comprehensive Data Strategy for national security and defense. For over a hundred years, statistical sciences have played a pivotal role in our national defense, from quality assurance and reliability analysis of munitions fielded in WWII, to operational analyses defining battlefield force structure and tactics, to helping optimize the engineering design of complex products, to rigorous testing and evaluating of Warfighter systems. The American Statistical Association (ASA) in 2015 recognized in its statement on *The Role of Statistics in Data Science* that "statistics is foundational to data science… and its use in this emerging field empowers researchers to extract knowledge and obtain better results from Big Data and other analytics projects." It is clearly recognized that data as information is a key asset to the DoD. The challenge we face is how to transform existing talent to add value where it counts.

Identifying Challenges and Solutions to T&E of Non-IP Networks

Peter Mancini

Institute for Defense Analyses

12:00 PM – 1:00 PM

Abstract: Many systems within the Department of Defense (DoD) contain networks that use both Internet Protocol (IP) and non-IP forms of information exchange. While IP communication is widely understood among the cybersecurity community, expertise and available test tools for non-IP protocols such as Controller Area Network (CAN), MIL-STD-1553, and SCADA are not as commonplace. Over the past decade, the DoD has repeatedly identified gaps in data collection and analysis when assessing the cybersecurity of non-IP buses. This roundtable is intended to open a discussion among testers and evaluators on the existing measurement and analysis tools for non-IP buses used across the community and also propose solutions to recurring roadblocks experienced when performing operational testing on non-IP components. Specific topics of discussion will include: What tools do you or your supporting teams use during cybersecurity events to attack, scan, and monitor non-IP communications? What raw quantitative data do you collect that captures the adversarial activity and/or system response from cyber aggression to non-IP components? Please provide examples of test instrumentation and data collection methods. What data analysis tools do you use to draw conclusions from measured data? What types of non-IP buses, including components on those buses, have you personally been able to test? What components were you not able to test? Why were you not able to test them? Was it due to safety concerns, lack of permission, lack of available tools and expertise, or other? Had you been given authority to test those components, do you think it would have improved the quality of test and comprehensiveness of the assessment?

1:00 PM-2:30 PM

Mini Tutorial 2A
Theme: Modeling Human-System Interaction
Tom Donnelly, Session Chair

Introduction to Structural Equation Modeling: Implications for Human-System Interactions

Laura Castro-Schilo

SAS Institute

1:00 PM – 2:30 PM

Abstract: Structural Equation Modeling (SEM) is an analytical framework that offers unique opportunities for investigating human-system interactions. SEM is used heavily in the social and behavioral sciences, where emphasis is placed on (1) explanation rather than prediction, and (2) measuring variables that are not observed directly (e.g., perceived performance, satisfaction, quality, trust, etcetera). The framework facilitates modeling of survey data through confirmatory factor analysis and latent (i.e., unobserved) variable regression models. We provide a general introduction to SEM by describing what it is, the unique features it offers to analysts and researchers, and how it is easily implemented in JMP Pro 16.0. Attendees will

Mini Tutorial 2B
Theme: Test & Evaluation Methods for Emerging Technology and Domains
Lance Fiondella, Session Chair

Combinatorial Interaction Testing

Erin Lanus

Virginia Tech

1:00 PM – 2:30 PM

Abstract: This mini-tutorial provides an introduction to combinatorial interaction testing (CIT). The main idea behind CIT is to pseudo-exhaustively test software and hardware systems by covering combinations of components in order to detect faults. In 90 minutes, we provide an overview of this domain that includes the following topics: the role of CIT in software and hardware testing, how it complements and differs from design of experiments, considerations such as variable strength and constraints, the typical combinatorial arrays used for constructing test suites, and existing tools for test suite construction. Last, defense systems are increasingly relying on software with embedded machine learning (ML), yet ML poses unique challenges to applying

	learn how to perform path analysis and confirmatory factor analysis, assess model fit, compare alternative models, and interpret results provided in SEM. The presentation relies on a real-data example everyone can relate to. Finally, we shed light on a few published studies that have used SEM to unveil insights on human performance factors and the mechanisms by which performance is affected. The key goal of this presentation is to provide general exposure to a modeling tool that is likely new to most in the fields of defense and aerospace.	conventional software testing due to characteristics such as the large input space, effort required for white box testing, and emergent behaviors apparent only at integration or system levels. As a well-studied black box approach to testing integrated systems with a pseudo-exhaustive strategy for handling large input spaces, CIT provides a good foundation for testing ML. In closing, we present recent research adapting concepts of combinatorial coverage to test design for ML.
2:30 PM-2:45 PM	*Break*	
2:45 PM-4:00 PM	Panel **Finding the Human in the Loop: Evaluating Warfighters' Ability to Employ AI Capabilities** Dan Porter *Institute for Defense Analyses* **Abstract:** Although artificial intelligence may take over tasks traditionally performed by humans or power systems that act autonomously, humans will still interact with these systems in some way. The need to ensure these interactions are fluid and effective does not disappear—if anything, this need only grows with AI-enabled capabilities. These technologies introduce multiple new hazards for achieving high quality human-system integration. Testers will need to evaluate both traditional HSI issues as well as these novel concerns in order to establish the trustworthiness of a system for activity in the field, and we will need to develop new T&E methods in order to do this. In this session, we will hear how three national security organizations are preparing for these HSI challenges, followed by a broader panel discussion on which of these problems is most pressing and which is most promising for DoD research investments. **HSI | Trustworthy AI** Stoney Trent *Virginia Tech; The Bulls Run Group, LLC* **Abstract:** Recent successes and shortcomings of AI implementations have highlighted the importance of understanding how to design and interpret trustworthiness. AI Assurance is becoming a popular objective for some stakeholders, however, assurance and trustworthiness are context-sensitive concepts that rely not only on software performance and cybersecurity, but also on human-centered design. This talk summarizes Cognitive Engineering principles in the context of resilient AI engineering. It also introduces approaches for successful Human-Machine Teaming in high risk work domains. **Considerations for AI in Decision Making** Joe Lyons *711 Human Performance Wing at Wright-Patterson AFB* **Coming Soon** Jane Pinelis *Department of Defense Joint Artificial Intelligence Center (JAIC)* **Panelist** Chad Bieber *Johns Hopkins University Applied Physics Laboratory* **Panelist** Poornima Madhavan *MITRE* **Panelist** Rachel Haga *Institute for Defense Analyses*	
4:00 PM-4:15 PM	**Presentation of Army Wilks Memorial Award** Winner To Be Announced	

Wednesday, April 14th

All times in Eastern Standard Time (EST).

Use of "A / B" in the schedule denotes parallel sessions; choose one session to attend.

8:45 AM-9:45 AM

Session 6A
Theme: Test & Evaluation Methods for Emerging Technology and Domains
Curtis Miller, Session Chair

A Metrics-based Software Tool to Guide Test Activity Allocation

Jacob Aubertine

University of Massachusetts Dartmouth

8:45 AM – 9:15 AM

Abstract: Existing software reliability growth models are limited to parametric models that characterize the number of defects detected as a function of testing time or the number of vulnerabilities discovered with security testing. However, the amount and types of testing effort applied are rarely considered. This lack of detail regarding specific testing activities limits the application of software reliability growth models to general inferences such as the additional amount of testing required to achieve a desired failure intensity, mean time to failure, or reliability (period of failure free operation). This presentation provides an overview of an open source software reliability tool implementing covariate software reliability models [1] to aid DoD organizations and their contractors who desire to quantitatively measure and predict the reliability and security improvement of software. Unlike traditional software reliability growth models, the models implemented in the tool can accept multiple discrete time series corresponding to the amount of each type of test activity performed as well as dynamic metrics computed in each interval. When applied in the context of software failure or vulnerability discovery data, the parameters of each activity can be interpreted as the effectiveness of that activity to expose reliability defects or security vulnerabilities. Thus, these enhanced models provide the structure to assess existing and emerging techniques in an objective framework that promotes thorough testing and process improvement, motivating the collection of relevant metrics and precise measurements of the time spent performing various testing activities. References [1] Vidhyashree Nagaraju, Chathuri Jayasinghe, Lance Fiondella, Optimal test activity allocation for covariate software reliability and security models, Journal of Systems and Software, Volume 168, 2020, 110643.

An Adaptive Approach to Shock Train Detection

Greg Hunt

William & Mary

9:15 AM – 9:45 AM

Abstract: Development of new technology always incorporates model testing. This is certainly true for hypersonics, where flight tests are expensive and testing of component- and system-level models has significantly advanced the field. Unfortunately, model tests are often limited in scope, being only approximations of reality and

Session 6B
Theme: Modeling & Simulation
Kevin Singer, Session Chair

A DOE Case Study: Multidisciplinary Approach to Design an Army Gun Propulsion Charge

Melissa Jablonski

US Army CCDC Armaments Center

8:45 AM – 9:15 AM

Abstract: This session will focus on the novel application of a design of experiments approach to optimize a propulsion charge configuration for a U.S. Army artillery round. The interdisciplinary design effort included contributions from subject matter experts in statistics, propulsion charge design, computational physics, and experimentation. The process, which we will present in this session, consisted of an initial, low fidelity modeling and simulation study to reduce the parametric space by eliminating inactive variables and reducing the ranges of active variables for the final design. The final design used a multi-tiered approach that consolidated data from multiple sources including low fidelity modeling and simulation, high fidelity modeling and simulation and live test data from firings in a ballistic simulator. Specific challenges of the effort that will be addressed include: integrating data from multiple sources, a highly constrained design space, functional response data, multiple competing design objectives, and real-world test constraints. The result of the effort is a final, optimized propulsion charge design that will be fabricated for live gun firing.

A DOE Case Study: Multidisciplinary Approach to Design an Army Gun Propulsion Charge

Sarah Longo

US Army CCDC Armaments Center

9:15 AM – 9:45 AM

Abstract: This session will focus on the novel application of a design of experiments approach to optimize a propulsion charge configuration for a U.S. Army artillery round. The interdisciplinary design effort included contributions from subject matter experts in statistics, propulsion charge design, computational physics and experimentation. The process, which we will present in this session, consisted of an initial, low fidelity modeling and simulation study to reduce the parametric space by eliminating inactive variables and reducing the ranges of active variables for the final design. The final design used a multi-tiered approach that consolidated data from multiple sources including low fidelity modeling and simulation, high fidelity modeling and simulation and live test data from firings in a ballistic simulator. Specific challenges of the effort that will be addressed include: integrating data from multiple sources, a highly constrained design space, functional response data,

typically only partially covering the range of potential realistic conditions. In this talk, we focus on the problem of real-time detection of the shock train leading edge in high-speed air-breathing engines, such as dual-mode scramjets. Detecting and controlling the shock train leading edge is important to the performance and stability of such engines, and a problem that has seen significant model testing on the ground and some flight testing. Often, methods developed for shock train detection are specific to the model used. Thus, they may not generalize well when tested in another facility or in flight as they typically require a significant amount of prior characterization of the model and flow regime. A successful method for shock train detection needs to be robust to changes in features like isolator geometry, inlet and combustor states, flow regimes, and available sensors. Such data can be difficult or impossible to obtain if the isolator operating regime is large. To this end, we propose the an approach for real-time detection of the isolator shock train. Our approach uses real-time pressure measurements to adaptively estimate the shock train position in a data-driven manner. We show that the method works well across different isolator models, placement of pressure transducers, and flow regimes. We believe that a data-driven approach is the way forward for bridging the gap between testing and reality, saving development time and money.

multiple competing design objectives and real-world test constraints. The result of the effort is a final, optimized propulsion charge design that will be fabricated for live gun firing.

9:45 AM-10:00 AM

Break

10:00 AM-11:30 AM

Session 7A
Theme: Modeling Human-System Interaction
Rachel Haga, Session Chair

Characterizing Human-Machine Teaming Metrics for Test and Evaluation

Brian Vickers

Institute for Defense Analyses

10:00 AM – 10:30 AM

Abstract: As advanced technologies and capabilities are enabling machines to engage in tasks that only humans have done previously, new challenges have emerged for the rigorous testing and evaluation (T&E) of human-machine teaming (HMT) concepts. We differentiate the distinction between a HMT and a human using a tool, and new challenges are enumerated: Agents' mental models are opaque, machine-to-human communications need to be evaluated, and self-tasking and autonomy need to be evaluated. We argue that a focus on mission outcomes cannot fully characterize team performance due to the increased problem space evaluated and that the T&E community needs to develop and refine new metrics for agents of teams and teammate interactions. Our IDA HMT framework outlines major categories for HMT evaluation, emphasizing team metrics and parallelizing agent metrics across humans and machines. Major categories are tied to the literature and proposed as a starting point for additional T&E metric specification for robust evaluation.

Session 7B
Theme: Special Topics
Shelley Cazares, Session Chair

Dashboard for Equipment Failure Reports

Robert Cole Molloy

Johns Hopkins University Applied Physics Laboratory

10:00 AM – 10:30 AM

Abstract: Equipment Failure Reports (EFRs) describe equipment failures and the steps taken as a result of these failures. EFRs contain both structured and unstructured data. Currently, analysts manually read through EFRs to understand failure modes and make recommendations to reduce future failures. This is a tedious process where important trends and information can get lost. This motivated the creation of an interactive dashboard that extracts relevant information from the unstructured (i.e. free-form text) data and combines it with structured data like failure date, corrective action and part number. The dashboard is an RShiny application that utilizes numerous text mining and visualization packages, including tm, plotly, edgebundler, and topicmodels. It allows the end-user to filter to the EFRs that they care about and visualize meta-data, such as geographic region where the failure occurred, over time allowing previously unknown trends to be seen. The dashboard also applies topic modeling to the unstructured data to identify key themes. Analysts are now able to quickly identify

Automated Test Case Generation for Human-Machine Interaction

Matthew Bolton

University at Buffalo, the State University of New York

10:30 AM – 11:00 AM

Abstract: The growing complexity of interactive systems requires increasing amounts of effort to ensure reliability and usability. Testing is an effective approach for finding and correcting problems with implemented systems. However, testing is often regarded as the most intellectual-demanding, time-consuming, and expensive part of system development. Furthermore, it can be difficult (if not impossible) for testers to anticipate all of the conditions that need to be evaluated. This is especially true of human-machine systems. This is because the human operator (who is attempting to achieve his or her task goals) is an additional concurrent component of the system and one whose behavior is not strictly governed by the implementation of designed system elements. To address these issues, researchers have developed approaches for automatically generating test cases. Among these are formal methods: rigorous, mathematical languages, tools, and techniques for modeling, specifying, and verifying (proving properties about) systems. These support model-based approaches (almost exclusively used in computer engineering) for creating tests that are efficient and provide guarantees about their completeness (at least with respect to the model). In particular, model checking can be used for automated test case generation. In this, efficient and exhaustive algorithms search a system model to find traces (test cases) through that model that satisfy specified coverage criteria: descriptions of the conditions the tests should encounter during execution. This talk focuses on a formal automated test generation method developed in my lab for creating cases for human-system interaction. This approach makes use of task models. Task models are a standard human factors method for describing how humans normatively achieve goals when interacting with a system. When these models are given formal semantics, they can be paired with models of system behavior to account for human-system interaction. Formal, automated test case generation can then be performed for coverage criteria asserted over the system (for example, to cover the entire human interface) or human task (to ensure all human activities or actions are performed). Generated tasks, when manually executed with the system, can serve two purposes. First, testers can observe whether the human behavior in test always produces the system behavior from the test. This can help analysts validate the models and, if no problems are found, be sure that any desirable properties exhibited by the model hold in the actual system. Second, testers will be able to use their insights about system usability and performance to subjectively evaluate the system under all of conditions contained in the tests. Given the coverage guarantees provided by the process, this means that testers can be confident they have seen every system condition relevant to the coverage criteria. In this talk, I will describe this approach to automated test case generation and illustrate its utility with a simple example. I will then describe how this approach could be extended to frequent failure modes and look at time and region-based trends in these common equipment failures.

Metrics for Assessing Underwater Demonstrations for Detection and Classification of UXO

Jacob Bartel

Institute for Defense Analyses

10:30 AM – 11:00 AM

Abstract: Receiver Operating Characteristic curves (ROC curves) are often used to assess the performance of detection and classification systems. ROC curves can have unexpected subtleties that make them difficult to interpret. For example, the Strategic Environmental Research and Development Program and the Environmental Security Technology Certification Program (SERDP/ESTCP) is sponsoring the development of novel systems for the detection and classification of Unexploded Ordnance (UXO) in underwater environments. SERDP is also sponsoring underwater testbeds to demonstrate the performance of these novel systems. The Institute for Defense Analyses (IDA) is currently designing and implementing the scoring process for these underwater demonstrations that addresses the subtleties of ROC curve interpretation. This presentation will provide an overview of the main considerations for ROC curve parameter selection when scoring underwater demonstrations for UXO detection and classification.

Machine Learning Reveals that the Russian IRA's Twitter Topic Patterns Evolved over Time

Emily Parrish

Institute for Defense Analyses

11:00 AM – 11:30 AM

Abstract: Introduction: Information Operations (IO) are a key component of our adversaries' strategy to undermine U.S. military power without escalating to more traditional (and more easily identifiable) military strikes. Social media activity is one method of IO. In 2017 and 2018, Twitter suspended thousands of accounts likely belonging to the Kremlin-backed Internet Research Agency (IRA). Clemson University archived a large subset of these tweets (2.9M tweets posted by over 2800 IRA accounts), tagged each tweet with metadata (date, time, language, supposed geographical region, number of followers, etc.), and published this dataset on the polling aggregation website FiveThirtyEight. Methods: Machine Learning researchers at the Institute for Defense Analyses (IDA) downloaded Clemson's dataset from FiveThirtyEight and analyzed both the content of the IRA tweets and their accompanying metadata. Using unsupervised learning techniques (Latent Dirichlet Allocation), IDA researchers mapped out how the patterns in the IRA's tweet topics evolved over time. Results: Results showed that the IRA started tweeting in/before February 2012, but ramped up significantly in May/June 2015. Most tweets were in English, and most likely targeted the U.S. The IRA created new accounts after the first Twitter suspension in November 2017, with each new account quickly establishing an audience. Between at least January 2015 and October 2017, the IRA's English tweet topics evolved over time, becoming tighter, more specific, more negative, and more polarizing, with the final pattern emerging in late

account for different dimensions of human cognitive performance and emerging challenges in human-autonomy interaction.

Cognitive Work Analysis – From System Requirements to Validation and Verification

Matthew Miller

Jacobs/NASA Johnson Space Center

11:00 AM – 11:30 AM

Abstract: Human-system interaction is a critical yet often neglected aspect of the system development process. It is mostly commonly incorporated into system performance assessments late in the design process leaving little opportunity for any substantive changes to be made to ensure satisfactory system performance achieved. As a result, workarounds and compromises become a patchwork of “corrections” that end up in the final fielded system. But what if mission outcomes, the work context, and performance expectations can be articulated earlier in the process, thereby influencing the development process throughout? This presentation will discuss how a formative method from the field of cognitive systems engineering, cognitive work analysis, can be leveraged to derive design requirements compatible with traditional systems engineering processes. This method establishes not only requirements from which system designs can be constructed, but also how system performance expectations can be more acutely defined a priori to guide the validation and verification process. Cognitive work analysis methods will be described to highlight how ‘cognitive work’ and ‘information relationship’ requirements can be derived and will be showcased in a case-study application of building a decision support system for future human spaceflight operations. Specifically, a description of the testing campaign employed to verify and validate the fielded system will be provided. In summary, this presentation will cover how system requirements can be established early in the design phase, guide the development of design solutions, and subsequently be used to assess the operational performance of the solutions within the context of the work domain it is intended to support.

2015. Discussion: The United States government must expect that our adversaries’ social media activity will continue to evolve over time. Efficient processing pipelines are needed for semi-automated analyses of time-evolving social media activity.

11:30 AM-12:00 PM

Break

12:00 PM-1:00 PM

Roundtables *Pre-Registration Required*

Roundtable Discussion
Theme: Special Topics

Organizing and Sharing Data within the T&E Community

Matthew Avery

Institute for Defense Analyses

12:00 PM – 1:00 PM

Abstract: Effective data sharing requires alignment of personnel, systems, and policies. Data are costly and precious, and to get the most value out of the data we collect, it is important that we share and reuse it whenever possible and appropriate. Data are typically collected and organized with a single specific use or goal in mind, and after that goal

Roundtable Discussion
Theme: Special Topics

Test Design and Analysis for Modeling & Simulation Validation

Kelly Avery

Institute for Defense Analyses

12:00 PM – 1:00 PM

Abstract: System evaluations increasingly rely on modeling and simulation (M&S) to supplement live testing. It is thus crucial to thoroughly validate these M&S tools using rigorous data collection and analysis strategies. At this roundtable, we will identify and discuss some of the core challenges currently

	has been achieved (e.g., the report is published), the data are no longer viewed as important or useful. This process is self-fulfilling, as future analysts who might want to use these data will not be able to find them or will be unable to understand the data sufficiently due to lack of documentation and metadata. Implementing data standards and facilitating sharing are challenging in the national security environment. There are many data repositories within the DoD, but most of them are specific to certain organizations and are accessible by a limited number of people. Valid concerns about security make the process of sharing particular data sets challenging, and the opacity of data ownership often complicates the issue. This roundtable will facilitate discussion of these issues. Participants will have opportunities to share their experiences trying to share data and make use of data from previous testing. We hope to identify useful lessons learned and find ways to encourage data sharing within the community.	associated with implementing M&S validation for T&E. First, appropriate design of experiments (DOE) for M&S is not universally adopted across the T&E community. This arises in part due to limited knowledge of gold standard techniques from academic research (e.g., space filing designs; Gaussian Process emulators) as well as lack of expertise with the requisite software tools. Second, T&E poses unique demands in testing, such as extreme constraints in live testing conditions and reliance on binary outcomes. There is no consensus on how to incorporate these needs into the existing academic framework for M&S. Finally, some practical considerations lack clear solutions yet have direct consequences on design choice. In particular, we may discuss the following: (1) sample size determination when calculating power and confidence is not applicable, and (2) non-deterministic M&S output with high levels of noise, which may benefit from replication samples as in classical DOE.
1:00 PM-2:30 PM	Mini Tutorial 3A Theme: Test & Evaluation Methods for Emerging Technology and Domains *Tom Donnelly, Session Chair* **Pseudo-Exhaustive Testing – Part 1** Ryan Lekivetz *SAS Institute* 1:00 PM – 1:45 PM **Abstract:** Exhaustive testing is infeasible when testing complex engineered systems. Fortunately, a combinatorial testing approach can be almost as effective as exhaustive testing but at dramatically lower cost. The effectiveness of this approach is due to the underlying construct on which it is based, that is a mathematical construct known as a covering array. This tutorial is divided into two sections. Section 1 introduces covering arrays, introduces a few covering array metrics, and then shows how covering arrays are used in combinatorial testing methodologies. Section 2 focuses on practical applications of combinatorial testing, including a commercial aviation example, an example that focuses on a widely used machine learning library, plus other examples that illustrate how common testing challenges can be addressed. In the process of working through these examples, an easy-to-use tool for generating covering arrays will be demonstrated. **Pseudo-Exhaustive Testing – Part 2** Joseph Morgan *SAS Institute* 1:45 PM – 2:30 PM	Mini Tutorial 3B Theme: Test & Evaluation Methods for Emerging Technology and Domains *Doug Ray, Session Chair* **Introduction to Neural Networks for Deep Learning with Tensorflow** Roshan Patel *US Army CCDC Armaments Center* 1:00 PM – 2:30 PM **Abstract:** This mini-tutorial session discusses the practical application of neural networks from a lay person's perspective and will walk through a hands-on case study in which we build, train, and analyze a few neural network models using TensorFlow. The course will review the basics of neural networks and touch on more complex neural network architecture variants for deep learning applications. Deep learning techniques are becoming more prevalent throughout the development of autonomous and AI-enabled systems, and this session will provide students with the foundational intuition needed to understand these systems.
2:30 PM-2:45 PM	*Break*	
2:45 PM-4:00 PM	Panel **The Keys to Successful Collaborations during Test and Evaluation** Christine Anderson-Cook *Los Alamos National Lab* **Abstract:** The defense industry faces increasingly complex systems in test and evaluation (T&E) that require interdisciplinary teams to successfully plan testing. A critical aspect in test planning is a successful collaboration between T&E experts, subject	

	matter experts, program leadership, statisticians, and others. This panel, based on their own experiences as consulting statisticians, will discuss elements that lead to successful collaborations, barriers during collaboration, and recommendations to improve collaborations during T&E planning. **Panelist** Sarah Burke *STAT Center of Excellence* **Panelist** Willis Jensen *W.L. Gore & Associates* **Panelist** John Haman *Institute for Defense Analyses* **Panelist** Peter Parker *NASA Langley*
4:00 PM-4:10 PM	**Closing Remarks** William “Allen” Kilgore *NASA Langley Research Center* **Abstract:** Mr. William (Allen) Kilgore serves as Director, Research Directorate at NASA Langley Research Center. He previously served as Deputy Director of Aerosciences providing executive leadership and oversight for the Center’s Aerosciences fundamental and applied research and technology capabilities with the responsibility over Aeroscience experimental and computational research. After being appointed to the Senior Executive Service (SES) in 2013, Mr. Kilgore served as the Deputy Director, Facilities and Laboratory Operations in the Research Directorate. Prior to this position, Mr. Kilgore spent over twenty years in the operations of NASA Langley’s major aerospace research facilities including budget formulation and execution, maintenance, strategic investments, workforce planning and development, facility advocacy, and integration of facilities’ schedules. During his time at Langley, he has worked in nearly all of the major wind tunnels with a primary focus on process controls, operations and testing techniques supporting aerosciences research. For several years, Mr. Kilgore led the National Transonic Facility, the world’s largest cryogenic wind tunnel. Mr. Kilgore has been at NASA Langley Research Center since 1989, starting as a graduate student. Mr. Kilgore earned a B.S. and M.S. in Mechanical Engineering with concentration in dynamics and controls from Old Dominion University in 1984 and 1989, respectively. He is the recipient of NASA’s Exceptional Engineering Achievement Medal in 2008 and Exceptional Service Medal in 2012.