*Three-phase optimal design of sensitivity experiments* (Keynote)

C.F. Jeff Wu
Georgia Institute of Technology

In sensitivity testing the test specimens are subjected to a variety of stress levels to generate response or non-response. These data are used to estimate the critical stimulus (or threshold) level of the experimental object. Because of its versatile applications, several sensitivity testing procedures have been proposed and used in practice. There remains the outstanding question of finding an efficient procedure, especially when the sample size is small and the interest lies in the extreme percentiles. In the paper, we propose a novel three-phase procedure, dubbed 3pod, which can be described as a trilogy of “search-estimate-approximate”. A core novel idea is to choose the stress levels to quickly achieve an overlapping data pattern which ensures the estimability of the underlying parameters. Simulation comparisons show that 3pod outperforms existing procedures over a range of scenarios in terms of efficiency and robustness. (Joint work with Yubin Tian.)

*Lessons Learned during the Test and Evaluation of Reliability Growth*
(Special Session on Reliability Growth)

Robert N. Tamburello and Michael J. Cushing
United States Army Evaluation Center

The US Army implemented policy to improve reliability in December 2007 which required major US Army programs to have a comprehensive reliability growth strategy with a reliability growth planning curve in the Test and Evaluation Master Plan (TEMP). The US Department of Defense followed with a reliability policy in 2008 and an update in March 2011. As well, the US Army updated its reliability policy in June 2011, which requires the reliability growth planning curve to also be included in the Systems Engineering Plan and the Engineering Manufacturing and Development contract in addition to the TEMP. The US Army reliability policy directs each program manager to perform design-for-reliability activities prior to acquisition Milestone B in order to mitigate the risk of failing to demonstrate reliability requirements. These policies have driven a marked rise in the application of reliability growth principles to programs undergoing test and evaluation. Over time, the number of reliability growth programs has increased from just a handful to about twenty currently in testing. Additionally, roughly twenty more reliability growth programs are in the planning stage. Early test results are in for many of the programs now undergoing reliability growth testing. In this presentation, we discuss the general trends observed in the reliability growth programs, including the reasons why many programs are struggling to achieve their goals. As well, we propose improvements to the engineering and acquisition processes to simultaneously manage government and developer risks and improve the likelihood that US Army programs will achieve their reliability requirements.

*A Dynamic Competing Risk Model and Its Implications*
(Special Session on Reliability Growth)

Owen S. Martin
RSA Group

In the problem of modeling the growth or change in reliability of a system that undergoes debugging, we have found a class of new avenues to explore in creating models. This class, the Poisson mixture of M failure times, has appeared in the literature before but actually begets paradox with respect to infallibility. We explore this potential paradox and propose a few methods to resolving it.

*Statistical Inference for a parametric model for repairable systems under multiple failure modes*
(Special Session on Reliability Growth)

Ananda Sen
University of Michigan

The focus of this talk is on failure history of a repairable system for which the relevant data comprise successive event times for a recurrent phenomenon along with an event count indicator. I shall report an investigation for analyzing failures from repairable systems that are subject to multiple failure modes. The context discussed is the parametric framework of the Army Material Systems Analysis Activity (AMSAA) model that has found considerable attention in modeling the failure process of defense systems. I will highlight some interesting and non-standard inference results that ensue. Findings from extensive simulation will be reported. An extension to the case where the specific cause of failure may be missing will be indicated. I shall discuss the application of the methodology on data on recurrent failures of a vertical boring machine that documents the specific component responsible for failure.

*Measuring R50 in Aircraft Radar Blip Scan Data: Lessons Learned in Increasing Analysis Rigor*

Todd Remund and Jeff Weisz
Edwards AFB, CA

A series of tests are run to portray windmill effects on aircraft radar systems. Three questions are posed for analysis of the data. In reference to radar detection distance of $R_{50}$, or range at which there is 50% detection probability, the questions are:

1. For three increasing distances from windmills where two head to head aircraft are in a direct approach vector to the windmills, is there a difference among test conditions in radar detection distance, $R_{50}$?
2. For four head to head encounters of two aircraft in a horizontally shifted approach vector to the windmills, is there a difference among test conditions in $R_{50}$?
3. Is there a difference between radar moving target rejection settings 55 and 110 in $R_{50}$?

Old methods of measuring this distance proved inadequate due to lack of rigor in analysis between $R_{50}$ values from different groups. While it is possible to achieve an estimate of the $R_{50}$ value from old methods, it is difficult to infer statistical significance in the estimate as well as bounding it with a confidence interval. It is yet more difficult to achieve a confidence interval on the difference between two $R_{50}$ values from different test conditions. Using new applications of logistic regression and bootstrapping techniques, statistical rigor can be added to analysis of blip scan data. This paper will present the analysis of radar data where the $R_{50}$ values are not only estimated but compared across different test conditions. Direct analysis and confidence intervals are computed to address the question of difference in the previously stated analysis instead of comparing confidence intervals from each $R_{50}$ at the different test conditions with one another. This method is known to be inefficient and is discussed in the paper. This application will open the door of consideration for more rigorous and efficient statistical analysis techniques in flight test.

*Confidence Intervals for Binary Responses – R50 and the Logistic Model*

Arnon Hurwitz
Edwards AFB, CA

Logistic regression is a non-linear method for modeling a binary response variable. For example, y = {success, failure} for blip-scan radar detections. Such responses cannot be modeled using regular linear regression. In our work, many applications of logistic regression present themselves. Models allowing independent slopes and independent intercepts are considered for comparing multiple groups of measures. The question that we consider here is the construction of a confidence interval about the difference in the Range 50 (R50) values for two such curves with each value (R1, R0) arising from the separate curve. R50 represents the range at which radar achieves 50% detection probability. This problem is akin to the problem of prediction of the LD50 ('lethal dose 50%') value in medical science. We approach the problem analytically and check our results using bootstrap simulation. A feature is the use of 'inverse prediction' or calibration methods.

*Detection of Changes in Resilience and Level of Coordination in Terrorist Groups*

Vasanthan Raghavan, Aram Galstyan, and Alexander G. Tartakovsky
University of Southern California

We are interested in developing models for the activity profile of a terrorist group, detecting sudden spurts and downfalls in this profile, and in attributing spurts to either an increased resilience in the organizational dynamics or an increased level of coordination between different sub-groups in the group or both of these aspects. For this, a $d$-state hidden Markov model (HMM) that captures the latent states underlying the dynamics of the group is proposed. The simplest setting of $d = 2$ corresponds to the case where the dynamics are coarsely quantized as *Active* and *Inactive*, respectively. It is shown that a state estimation strategy based on the Viterbi algorithm detects sudden spurts in the activity profile and tracks them over a period of time very well. Aided by a majorization theory-based framework, spurts are then classified into those corresponding to increased resilience or increased level of coordination. Case studies with real terrorism data are provided to illustrate the performance of the proposed methodologies.

*A Bayesian Approach to Evaluation of Land Warfare Systems*

Lee S. Dewald, Sr., Virginia Military Institute
Robert Holcomb, Sam Parry, and Alyson Wilson, Institute for Defense Analyses

For some time, there has been a general dissatisfaction with the statistical analyses used in operational testing of land warfare systems. In this paper we review the current paradigm; critique perceived deficiencies with the current methodology; and demonstrate how a Bayesian approach could be applied to the planning, conduct, and analysis of land warfare systems under test by utilizing data from previous field tests and simulations to assist in designing the next field test to take advantage of prior information. We first apply this approach to the analysis of Blue Survivors and Red Killed as univariate measures of effectiveness (MOE) in the operational testing of digitized and non-digitized forces from AWE testing of a mechanized force and the subsequent DCX exercises. We conclude by illustrating a Bayesian approach using hierarchical modeling to the analysis of the Loss-Exchange Ratio (LER) from AWE testing, DCX exercises, and simulation runs of JCATS.

*Statistical Models for Combining Information: Stryker Reliability Case Study*

Rebecca Dickinson, Virginia Tech
Laura Freeman and Alyson Wilson, Institute for Defense Analyses

Reliability is an essential element of system suitability in the Department of Defense. It often takes a prominent role in both the design and analysis of operational tests. However, in the current era of reducing budgets and increasing reliability requirements, it is challenging to verify reliability requirements during an initial operational test (IOT). This talk illustrates the benefits of using parametric statistical models to combine information across multiple test events. Both Bayesian and Non-Bayesian inference techniques are employed and contrasted to illustrate different statistical methods for combining information. We apply these methods to both developmental and operational test data for Stryker Armored Vehicle. We illustrate how the conclusions of the reliability analysis would have differed if these parametric models had been used at the conclusion of the IOT.

*F-16 radar performance analysis- do Bayesian hierarchical models provide better insight into system characterization than standard techniques?*

James Brownlow
Edwards AFB, CA

Radar system analysis addresses a variety of system performance parameters. Target detection range is one of the most common measures of performance. Typically, detection range is specified as a probability (ref 1) R.50, the range at which the probability of detecting a target is 0.5. Methods for estimating R.50, based on blip-scan data, include averaging range-to-target bin data, logistic regression generalized linear models (GLMs), and Bayesian hierarchical models. This paper presents the background and analysis advantages of using Bayesian hierarchical logistic- and arc sine regression models. Arcsin models in particular may offer the advantage of lower overall root mean square errors. This paper presents an analysis of F-16 R.50 radar detection using (unclassified) flight test data and a variety of generalized linear models to estimate R.50. That is, prob(Detection=1|range, test run number) is modeled using logit, cloglog, probit, and cauchit generalized linear models. The same data are analyzed with Bayesian hierarchical models, using both logit and arcsine links. Model comparisons are made using information likelihood (AIC) methods. Shortcomings of the GLM models include possible over dispersion and the inability to identify (in some rigorous sense) "atypical" flight test runs. The two Bayesian hierarchical models (logit, and arcsin) are shown to provide additional insight into R.50 estimation.

*DaViTo: An Open-Source Mapping, Graphics, and Analysis Tool*

Sam Buttrey, Lyn R. Whitaker, and Arnold Buss, Naval Postgraduate School
James Henry, US Army

Analysts in contemporary military operational environments accumulate vast amounts of data. These include data on, for example, incidents like IEDs; vehicle tracking; surveys; local economics; and demographics. Much of this will include a spatio-temporal component. Visualizing, exploring, and understanding this data reveals insights about our own forces' behavior as well as patterns within the environment. The Data Visualization Tool (DaViTo) merges several open-source software packages to create a single data exploration tool capable of uncovering these insights and patterns. The tool provides a suite of data exploration methods (current focused on point processes) that can be applied to the data by military analysts, who will be experts in operations rather than statistics. DaViTo's power derives from the capabilities of the open-source software upon which it has been built: OpenMap, JFreeChart, and R. Statisticians can provide additional functionality through R and its packages without Java programming. DaViTo is government-owned software, built solely with open-source tools, which may be run from a CD even in secure computing environments.

*Africa Knowledge, Data Source, and Analytic Effort Exploration*

Thomas Deveans
US Army TRADOC Analysis Center

The TRADOC Analysis Center (TRAC), Naval Postgraduate School, and other Department of Defense organizations are conducting large data capture and analysis efforts on areas all around the world. As efforts in the US Central Command Area of Responsibility (AOR) - particularly in Iraq and Afghanistan - draw down, many senior decision makers expect future efforts will be focused on the US African Command (AFRICOM) AOR. The Africa Knowledge, Data Source, and Analytic Effort Exploration will build an assessment framework focused on the AFRICOM AOR to identify the ideal data to be gathered and measured in a COIN environment. Then, by actually gathering the data points from multiple sources, gaps in the available data will be identified in comparison to the ideal. This effort will concurrently develop the necessary software within the Data Visualization Tool (DaViTo) - an open source, government owned exploratory data analysis tool - to allow the end user to construct an assessment framework using a customized weighting scheme and the ability to display results. Finally, this project will develop an irregular warfare scenario methodology and a small proof-of-principle use case in Nigeria by conducting factor analysis of survey data. It will use generalized linear models, both linear and multinomial logistic regression, to predict future issue stance scores, and the population's observed attitudes and behaviors, which will directly support TRAC's Irregular Warfare Tactical Wargame (IW TWG).

*Regression trees for censored data with applications to subgroup identification*

Wei-Yin Loh
University of Wisconsin

Previous regression tree algorithms for censored response data are mostly based on the CART (Breiman et al. 1984) approach, which searches for splits that maximally reduce some node impurity function. As a result, the algorithms are biased toward selecting variables that afford more splits. Besides, since the best split must be found for every variable, the algorithms are limited in practice to fitting piecewise constant models. An alternative algorithm based on the GUIDE (Loh 2002, 2009) approach is proposed for fitting piecewise multiple-linear proportional hazards models. It uses on an old trick that uses Poisson regression to fit proportional hazards models. An importance ordering of the variables is obtained as a by-product. By utilizing the treatment variable as a linear predictor in the model at each node, the method can also be used to identify subgroups of the data that exhibit differential treatment effects. Evaluations based on real and simulated data show that the proposed method is as good or better than other tree methods and random survival forest, in terms of accuracy and computational speed in identifying the important variables.

*Accelerated Test Methods for Reliability Prediction*

David H. Collins, Aparna V. Huzurbazar, and Brian P. Weaver, Los Alamos National Laboratory
Jason K. Freels and Richard L. Warr, Air Force Institute of Technology

Perusal of quality and reliability engineering literature indicates some confusion over the meaning of accelerated life testing (ALT), highly accelerated life testing (HALT), highly accelerated stress screening (HASS), and highly accelerated stress auditing (HASA). In addition, there is a significant conflict between testing as part of an iterative process of finding and removing defects, and testing as a means of estimating or predicting product reliability. We review the basics of these testing methods and describe how they relate to statistical methods for estimation and prediction of reliability and reliability growth. We also outline some synergies to help reconcile statistical and engineering approaches to accelerated testing, resulting in better product quality at lower cost.

*On the "No-Internal-Zeros Conjecture" for system signatures in reliability*

Frank Samaniego
University of California-Davis

Consider a coherent system in n components having independent, identically distributed (i.i.d.) lifetimes. The signature of the system is an n-dimensional vector $\underline{s} = (s_1, \ldots, s_n)$ representing the probability distribution of the index of the ordered component failure which causes the system to fail. A brief review of the theory and applications of system signatures is given. Since its introduction in Samaniego (1985), signature theory has played an important role in studying the comparative performance of engineered systems and communication networks. It has long been conjectured that a signature vector of a coherent system can have no internal zeros. But the main argument for the conjecture has been that no-one had ever seen such a signature. Recently, we found a paper in an Operations Research journal that proves a rather general theorem that implies that the conjecture is true. The proof is elegant but quite complex. We present a proof of the NIZ property of signatures from basic principles, that is, using only known properties of the cut sets of a coherent system. The new proof is quite revealing and the result has immediate applications. The work is joint with Professor Jorge Navarro.

*Operational Risk Analysis: New analysis methods for numerical and non-numerical survey responses*

Michele Wolfe
US Army TRADOC Analysis Center (TRAC)

Determining operational risk in support of acquisition decisions is a complex problem that cannot be adequately addressed with the Army's basic risk assessment doctrine. Current operations research delves into non-numerical areas where concrete data do not exist. Categories such as "consequence" or "severity" are not limited to a single answer; instead, these factors are characterized by ranges of responses. To better investigate these complex issues, the US Army Training and Doctrine Command (TRADOC) Analysis Center (TRAC) developed a new approach to operational risk analysis. Using the triangular distribution as a means to evaluate the ranges of responses, this new methodology provides a solid mathematical foundation to conduct sensitivity analysis and to inform leadership. Because of its mathematical foundation, gaps and the mitigation of gaps may be assessed with greater analytical rigor, providing numerical information as to how well capability gaps are mitigated. This presentation highlights the new risk analysis method, and how to collect and process the data. The presentation also discusses the implementation of a gap assessment tool and an example within a current analysis of alternatives study conducted by TRAC.

*Speed vs. Accuracy, the Shooter’s Dilemma*

Benjamin Thirey and Christopher E. Marks
United States Military Academy

Effectively engaging a target with a direct-fire weapon requires both speed and accuracy. However, there is a tradeoff: shooting too quickly results in inaccuracy, while attempting to achieve near-perfect accuracy results in prolonged engagement times. In this paper, we present a method of quantifying the trade-off between speed and accuracy in the context of competition shooting. Using shot placement standard deviation as a measure of a shooter's accuracy, we present a time-dependent model quantifying the increase in accuracy for longer shot times. We fit our model to data collected from two college-level competition shooters using maximum likelihood parameter estimation, and apply our results in order to predict a shooter's optimal engagement time, set marksmanship training objectives, and make assessments and comparisons of shooters' abilities.

*Measurement System Analysis and Applications to Improvements in Body Army Testing*

Sandra Meiselwitz and Barbara J. Gillich
US Army Aberdeen Test Center

Measurement System Analysis (MSA) is a method that examines the variability of a measurement process to determine the ability of the process to precisely measure an item. At the US Army Aberdeen Test Center (ATC), many different types of items are tested, including body armor items. This presentation will show how ATC has used MSA to examine the precision of different measurement processes related to body armor, including helmet thickness measurements and qualifying operators for the hard body armor plate dimension measurement process. Using MSA has allowed ATC to document reductions in process variability as continuous improvements are made to a measurement system, show customers that measurement procedures are capable of measuring to the desired level, and demonstrate that operators are properly trained prior to participating in customer projects.

*Comparative Testing*

Jacob J. Warren
Marine Corps Operational Test and Evaluation Activity

This presentation will describe a test and evaluation strategy applicable to systems whose requirements are stated via a comparison to another system. An example of a requirement stated as such is: "System X must be able to acquire targets on the move greater than or equal to the legacy." The presentation will compare and contrast standard hypothesis tests (which utilize critical and test values) with: p-values, and confidence intervals for testing a set of hypotheses. Equivalence, superiority, and non-inferiority comparative testing procedures common to clinical trials will be presented, and a non-inferiority testing example will be presented providing a link to how these methodologies can be used in an Operational Test and Evaluation. An R script will be presented to amplify the non-inferiority testing example.

*How to complete the incomplete; a statistical antidote for incomplete data sets*

Ofer Harel
University of Connecticut

Missing data is a common complication in medical, sociological and public health research. Most of the time researchers tend to ignore this obvious difficulty. In this presentation, I will introduce the missing data problem, introduce several different solutions and argue that considerable thought is needed before analyzing incomplete data.

*Demystifying Kriging Models with Massive Data*
(Special Session on Simulation Experiments and Efficient Design)

Peter Z.G. Qian
University of Wisconsin-Madison

The Kriging model is the prevailing choice in metamodeling. Fitting a Kriging model with massive data is not only a challenge but also a mystery. On one hand, the nominal accuracy of a Kriging model is supposed to increase with the number of data points. On the other hand, fitting such a model to a large number of points is known to have numerical singularity. To reconcile this contradiction, I will present a sequential method to simultaneously achieve numerical stability and theoretical accuracy in large-scale Kriging models. This method forms nested space-filling subsets of the data, builds kernel models for different subsets and then combines all submodels together to obtain an accurate surrogate model. We introduce a mathematical decomposition of the overall error of a Kriging model into nominal and numeric portions. Theoretical bounds on the numeric and nominal error are developed to show that substantial gains in overall accuracy can be attained with this sequential method. Examples are given to illustrate the effectiveness of the developed method. If time permits, I will discuss extensions of this method to incorporate derivative information and multiple responses.

*Recent Breakthroughs in Large-Scale Simulation Studies*
(Special Session on Simulation Experiments and Efficient Design)

Susan M. Sanchez, Thomas W. Lucas, Paul J. Sanchez, and Alejandro Hernandez
Naval Postgraduate School

Designed experiments are powerful methodologies for gaining insights into the behavior of complex systems. In recent years, several new designs have been created to address the large number of factors and complex response surfaces that often arise in simulation studies. These include procedures capable of handling discrete-valued or qualitative factors. We describe some of these advances, and summarize the results of using large-scale designed experiments in several recent defense simulation studies. Application areas include the development of more effective training scenarios for peace support operations; logistics life cycle management; and model-based ship design. The new designs provide structured methods for performing trade-off analyses in situations where other techniques have fallen short. Our approaches are scalable and compare well to existing methods that construct custom designs for smaller experiments. These breakthroughs were originally developed for exploring large stochastic simulation models, but may also be of interest for exploring computer models with limited numbers of factors.

*Sequential Importance Sampling for Rare Event Estimation with Computer Experiments*
(Special Session on Simulation Experiments and Efficient Design)

Brian Williams
Los Alamos National Laboratory

Importance sampling often drastically improves the variance of percentile and quantile estimators of rare events. We propose a sequential strategy for iterative refinement of importance distributions for sampling uncertain inputs to a computer model to estimate quantiles of model output or the probability that the model output exceeds a fixed or random threshold. A framework is introduced for updating a model surrogate to maximize its predictive capability for rare event estimation with sequential importance sampling. Examples of the proposed methodology involving materials strength and nuclear reactor applications will be presented.

*Designing Observational Studies to Avoid Being Misled*

Thomas E. Love
Case Western Reserve University School of Medicine

This session provides practical advice for researchers and scientists reviewing and designing observational studies, focusing on strategies for dealing with selection bias. The key issue is how we may be appropriately circumspect in evaluating the evidence from observational studies, especially when randomized experiments or clinical trials are unavailable. The presentation focuses on the use of propensity scores, and discusses this approach in light of several case studies, taken from recent reports on clinical medicine and public health.

*Propensity Score Analysis: Concepts, Issues, and Examples*
(Special Session on Design and Analysis of Observational Studies)

Wei Pan
Duke University

Researchers often use observational data to estimate treatment effects because randomized controlled trials or experimental designs are not always feasible. The use of observational data threatens the validity of causal inference by introducing selection bias into the treatment assignment. To deal with this challenge, Rosenbaum and Rubin (1983) proposed propensity score analysis (PSA), which balances the distributions of observed covariates between treatment conditions (i.e., treatment vs. control), as a means to reduce selection bias. Over the past three decades, PSA has become increasingly popular for making causal inferences based on observational studies. To further promote the use of PSA, this presentation will introduce basic concepts of PSA and discuss current issues in PSA. Empirical examples will also be used to illustrate how to select and implement appropriate PSA methods for observational studies.

*Designing observational studies to be more like randomized experiments*
(Special Session on Design and Analysis of Observational Studies)

Mike Baiocchi
Stanford University

Randomized controlled trials (RCTs) are the gold standard for estimating the impact of an intervention (e.g., improvements in outcomes due to a new training regime). There are several reasons RCTs are highly regarded, chief among these are the ability to randomize units of observations into the types of treatment. Through randomization we are able to control both the observed covariates as well as the unobserved covariates (even the infamous "unknown unknowns"). This produces accurate, reliable estimates of the treatment effect. But how can we harness randomization in an observational setting? This talk will cover a very clever technique known as "instrumental variables," which allows for RCT-like studies in some observational settings. The technique will be motivated using a case study which estimated the relative efficacy of intensity of care on quality outcomes.

*Continuous Updating Procedure for Flow Field Forecasting*

Kyle Caudle, South Dakota School of Mines & Technology
Michael Frey, Bucknell University

Flow field forecasting is a statistical procedure for time series forecasting that is based on the premise that if the time series is currently following a dynamic similar to one observed in its past record, then the process is likely to advance forward in a manner that is similar to how it behaved in the past. Flow field forecasting has three steps: a penalized spline regression (PSR), a Gaussian process regression (GPR) built upon the PSR, and a forecasting mechanism that uses the GPR to interpolate future process changes step-by-step. Often data presents as a flow requiring the need to continuously, efficiently and autonomously update forecasts. Updating the PSR in a strict sense involves a complete reevaluation of its global smoothing parameter by analyzing all data, not just recent data. Through experimentation we show that updating the complete PSR is often not necessary. By using a local smoothing parameter, we show the PSR estimate is better than had a single global smoothing parameter been used. Once the PSR is updated, if we assume that hyperparameters of the GPR have not been drastically altered by the new data point, updating the GPR can also be done in a straightforward fashion that is computationally efficient.

*Computing POD for Structural Health Monitoring Systems*

Christine M. Schubert Kabban, Mark M. Derriso, and Brandon Greenwell
Air Force Institute of Technology

Embracing the objectives of condition based maintenance, structural health monitoring (SHM) systems are sought to improve maintenance performance with respect to materiel availability and reliability as well as reduce mean down time through instantaneous or continual monitoring of structural integrity. SHM can be defined as automated methods for determining adverse changes in the integrity of mechanical systems. The ultimate goal of SHM is to provide an automated and real-time assessment of a structure's ability to serve its intended purpose. SHM assessments consist of a diagnosis of the current state of a structure and a prognosis about the capability of the structure to perform its function in the future. The diagnosis should include the detection, location, and assessment of damage. The prognosis might be that the structure is as good as new; safe to operate for only a certain number of flight hours; or, that immediate repair is required. The detection, location, assessment, and prognosis of damage are considered the four levels of SHM. A level four SHM system potentially provide the greatest benefit in that not only may the system be designed for damage detection, but may also provide estimates of the extent of structural damage as well as continuous monitoring of such damage, leading to better estimation (e.g., crack growth) and time to structural failure. Even with this added capability, however, damage detection (i.e., level one SHM systems) remains a focus for SHM systems. In particular, validation of SHM systems is much debated as the concept of probability of detection (POD) of structural damage remains an open question. Although understood as the key validity measure for detection, the computation of POD is less understood for SHM systems. Specifically for aircraft, thresholds based on the crack size that is detected 90% of the time with 95% statistical confidence is used to establish inspection intervals occurring at ½ the time between this and the crack size associated with structural failure. Statistical methods have been mandated that appropriately address the computation of POD and determination of this crack size for current inspection methods. This methodology may also translate over to SHM systems where data is collected in a manner analogous to that of current inspection routines, that is, at pre-established inspection intervals. For many SHM systems, however, data is collected continuously and therefore constitutes a different structure, a structure that is not compatible with current statistical methods of POD and critical crack length computations. The purpose of this paper is to review statistical methods for POD and critical crack length computations, demonstrating the usefulness of current methodology and to suggest advanced statistical models for comparable computations using the continuous data collection of SHM systems. Examples using recently collected data will be used to demonstrate these concepts.

*Recognition of Microcalcifications in Mammograms: Feature Extraction*

Ben Riggan and Wesley Snyder, North Carolina State University
Robert Basilico, Rita Freimanis, and Joseph Mammarappallil, Wake Forest School of Medicine

A careful analysis has been performed of the process of identifying Microcalcifications in mammograms. Of particular interest is the difference between two approaches: (1) developing a data base of small (e.g. 9x9) images and classifying with a Support Vector Machine and (2) using a much lower dimensionality data set based on extracted features. The feature extraction philosophy requires techniques used for preprocessing, segmentation, feature selection, and these are addressed. In particular, features used for classification are discussed in terms of their ability to adequately represent microcalcificaitons. By virtue of the features' physical and statistical interpretations, they can be well adapted for classifying microcalcifications.

## *Geographical Trends in Cancer Mortality: Using Spatial Smoothers and Methods for Adjustment*

Karen Kafadar
Indiana University

Mapping health-related data can lead to important insights and generate hypotheses about causes and potential effects. Usually such data are adjusted for age and gender so that inferences are not influenced by these obvious factors. Similarly, data should be (but usually are not) adjusted for known risk factors that have been identified. One method of adjustment is suggested here, and insights from adjusted data are enhanced by smoothing the data in two dimensions (longitude and latitude). The process of adjustment and smoothing is illustrated on cancer mortality data from three different sites: lung, prostate (among nonwhite males), and melanoma (among white males and females). Issues of linear versus nonlinear smoothers and display of trends and their uncertainties are discussed.

*Lookahead or Lookback: A Hybrid Architecture in Approximate Dynamic Programming*

Haitao Li and Keith Womer
University of Missouri - St. Louis

The complementary strengths of lookup table and rollout policy in approximate dynamic programming have motivated the idea of combining the two. On one hand, the lookup table directly associates the cost-to-go function with a state-action pair, which offers an accurate and reliable function approximation. On the other hand, for a high-dimensional MDP, it is often not possible to visit every state. For those state-action pairs that have not been visited/evaluated by the lookup table, the rollout approach can be employed to evaluate them. This hybrid approximation architecture has been implemented for stochastic project scheduling problems and shows promising results.

*Made-to-Order Production Scheduling using Bayesian Updating*

Keith Womer, J. Camm, C. Osterman, R. Radhakrishnan, and Haitao Li
University of Missouri - St. Louis

Production planning for made-to-order (MTO) manufacturing optimizes the production rate in each period before a promised delivery date. As products in MTO are often unique and customized, the associated learning curve slope and other production parameters cannot be accurately estimated before production starts. In this paper, a dynamic and adaptive solution approach is developed to estimate the effects of learning and production rate in conjunction with optimization. This approach offers a closed-loop solution through dynamic programming. Monthly production data are used to update the joint probability distributions of production parameters via Bayesian methods. Our approach is illustrated using historical data from the US Army's Black Hawk Helicopter Program. Managerial insights are obtained and discussed.

*Uncertainty Quantification using Soft Information and Exponential Epi-Splines*

Johannes O. Royset, Naval Postgraduate School
Roger J-B Wets, University of California

We address uncertainty quantification in complex systems by a flexible, nonparametric framework for estimation of density and performance functions. The framework systematically incorporates hard information derived from physics-based sensors, field test data, and computer simulations as well as soft information from human sources and experiences. The framework is based on epi-splines for consistent approximation of infinite-dimensional optimization problems arising in the estimation process. Preliminary numerical results indicate that the framework is highly promising.

*Developing Joint Model of Repeated Measure and Survival Outcome*

Jason Zhou and Harry Ji, Digital system, Inc.
Yuanzhang Li, US Army Medical Research and Materiel Command

In order to assess the causal effect of treatment s in longitudinal, observational data, it is need to address repeated measure and survival outcome with incomplete data. Statistical Model Development in Epidemiological Research can be used to develop joint model of repeated measure and survival outcome for epidemiology statistical analysis that enables users to conduct those longitudinal data. SMDER is built on top of BASE SAS with another user friendly interface. In this presentation, we will express that SMDER allow the researchers to develop joint model of repeated measure and survival outcome to conduct a survival analysis using all SAS advanced procedures to analyze time to event--survival outcome.

*DOE and Tolerance Design for Antenna Design*

William C. Thomas
Raytheon

A sequence of 5 DOE's was performed on an antenna design. The single experiment I will discuss was a 3^8-2 fractional factorial design. This means running 729 different combinations from a total of 6561. This number of experiments is possible because they are simulations. A statistical tolerance analysis was performed to insure performance and affordability. I will discuss the methodology and the interesting results that provided the Engineers an affordable solution.

*Applications of Clustered Designs*

Selden Crary
NewallStreet, Inc.

Careful searches for exact, optimal designs for computer experiments, under the minimum-Gaussian-process-IMSE objective and assuming fixed Gaussian-covariance parameters, led to the discovery, c. Y2000, of frequent proximal points (clusters) in these designs. These “clustered designs” specify that both a function evaluation and one or more specific directional derivative should be taken at the location of the twin points. The existence of these optimal designs runs counter to the prevailing view that designs for computer experiments should be space filling in the usual sense (maximin distance between nearest neighbors, minimax Voronoi-cell volume, uniform, etc.). We show examples of these designs, their symmetry properties, and the possibility of a universality class for the resulting phase diagram. We provide an example of the application of clustered designs to sequential generation of metamodels, as well as the application of clustered-design theory to the numerical inversion of highly ill-conditioned covariance matrices.

*Verifying System Performance Using Designed Simulation Experiments*

Terril N. Hurst, Allan T. Mense, and Colin Pouchet
Raytheon Missile Systems

Recent guidance has been issued from the Department of Defense Operational Test and Evaluation Directorate (DOD OT&E) on the creation of Test and Evaluation Master Plans (TEMPs). The Guide includes instructions and several examples clarifying the use of Design of Experiments (DOE) during Developmental Test, Live-Fire Evaluation, and Operational Evaluation. The Guide also repeatedly mentions the use of Modeling and Simulation (M&S), which has become an increasingly important tool for verifying that system performance requirements are met. However, the Guide is silent on how conventional DOE should be used or adapted when simulation experiments are employed to satisfy OT&E objectives. M&S-based DOE is an area undergoing rapid change, particularly as more practitioners seek to employ it to develop and verify software-intensive systems.

During the past few years, Raytheon Missile Systems has been learning how to apply DOE to M&S. Four categories of objectives have been identified for the design and analysis of simulation experiments (DASE). A 7-step protocol has been developed, taught, and practiced for planning, executing and briefing results. Performance Verification, labeled as a "Category 1" DASE objective, differs substantially from the other DASE categories, including Design Exploration (Category 2). The latter commonly uses conventional experimental designs and response surface methods. Category 1 DASE is used in "scoring" performance, and involves sampling over the full extent and dimensionality (factor space) of the system's operational envelope. Scoring thus requires descriptive, rather than predictive response models. The difference between the two model types dictates different types of experimental design, with space-filling designs being commonly used for building descriptive models expressed as moments or quantiles of a scoring bin's population. The use of hypothesis tests (confidence and power) and confidence intervals also differs between Category 1 and Category 2 DASE. This paper uses several examples to illustrate lessons learned in applying DOE and M&S for performance verification and is intended to engage the DOD Acquisition community in active dialog on how to more fully integrate DOE and M&S within the system performance verification process.

*Extracting communities from networks*

Ji Zhu
University of Michigan

Analysis of networks and in particular discovering communities within networks has been a focus of recent work in several fields, with applications ranging from citation and friendship networks to food webs and gene regulatory networks. Most of the existing community detection methods focus on partitioning the network into cohesive communities, with the expectation of many links between the members of the same community and few links between different communities. However, many real-world networks contain, in addition to communities, a number of sparsely connected nodes that are best classified as "background". To address this problem, we propose a new criterion for community extraction, which aims to separate tightly linked communities from a sparsely connected background, extracting one community at a time. The new criterion is shown to perform well in simulation studies and on several real networks. We also establish asymptotic consistency of the proposed method under the block model assumption. This is joint work with Yunpeng Zhao and Elizaveta Levina.

*Threshold Boolean Reformulation of Probabilistic Constraints with Multi-Row Random Technology Matrix*

Miguel Lejeune, George Washington University
Alexander Kogan, Rutgers University

We construct a partially defined Boolean function (pdBf) representing the satisfiability of a joint probabilistic constraint with random technology matrix. We extend the pdBf as a threshold Boolean tight minorant to derive a series of integer reformulations equivalent to the stochastic problem. Computational experiments will be presented. We develop a modeling and exact solution method for stochastic programming problems including a joint probabilistic constraints with multi-row random technology matrix. We binarize the probability distribution of the random variables and extract a threshold partially defined Boolean function (pdBf) representing the probabilistic constraint. We then construct a tight threshold Boolean minorant for the pdBf. The separating structure of the tight threshold Boolean minorant defines sufficient conditions for the satisfaction of the probabilistic constraint and takes the form of a system of linear constraints. We use the separating structure to derive three new deterministic formulations equivalent to the stochastic problem. A crucial feature of the derived integer formulations is that the number of integer variables does not depend on the number of realizations that the vector of random variables can take. Computational results will be presented.

*An integrated model for locating and dispatching ambulances*

Laura A. McLay and Sardar Ansari
Virginia Commonwealth University

There is a rich literature that develops covering location problem models that seek to locate servers (ambulances) at a set of stations to maximize the coverage level, the fraction of demand that the servers can reach in a prespecified amount of time. These models focus solely on how to locate ambulances and they make assumptions about how the ambulances would be dispatched to customers. In this paper, we lift the dispatching assumptions to develop a novel model that simultaneously locates ambulances at stations and creates dispatching rules using a mixed integer programming (MIP). Together, these decisions can be used to identify ambulance response districts. Contiguity constraints are added to the model to maintain contiguous response districts. The modeling approach applies a Hypercube approximation to a queuing system to identify the MIP model input parameters. An iterative procedure updates the parameters and balances the workload among the servers.

*Characterizing Probability Distributions Using Econometric Concentrations*

Nozer D. Singpurwalla
George Washington University

Characterizing probability distributions is an old and honourable pass time of applied probabilists, if not probability theorists. Concentration measures like the Kolmogorov-Smirnov, the Anderson-Darling, and the Gini Index, are a part of the toolkit of statisticians, econometricians, and those involved with the diagnostic testing of diseases and threats. In this talk I make the argument that the statistician's toolkit can supplement the probabilist's pass time. To do so, I introduce a new measure of concentration based on income inequalities and demonstrate its use for the stochastic chracterization of ageing. My work is motivated by an idea of Moshe Shaked and his coworkers, who introduced the notion of "excess wealth" as a property of distribution functions, and my interest in developing meaningful measures of concentration for diagnostic testing. Aspects of this work are to appear in the Annals of Operations Research.

*Divisive normalization in neural circuits drives context-dependent choice*

Kenway Louie and Paul W. Glimcher
New York University

A fundamental assumption of rational theories of choice is that decisions are free of context; how a chooser decides between any two options should not depend on the presence of other options, a property known as independence of irrelevant alternatives. In contrast to theory, context profoundly affects empirical choice behavior in species ranging from insects to birds to humans. Various psychological explanations have been proposed for specific context effects, but these proposed mechanisms do not incorporate information about neural processing, and no general model exists for context-dependence. Here, we describe a stochastic choice model based on normalized value coding, characterize novel context- dependent behavior in simulation, and test the model predictions in animal and human choosers. Observations in sensory pathways suggest that signal normalization is a canonical feature of neural computation. Recent electrophysiological studies have extended these insights beyond the sensory domain, demonstrating an analogous algorithm for the value signals that directly guide behavior. This normalized value coding, in which firing rates depend on intra- and extra-response field targets, introduces an inherent contextual dependency in the decision process, but its implications for behavior are unknown. In a simple computational model incorporating normalized value coding, we find that choice efficiency between high-valued options decrements as a function of the value of irrelevant alternatives. This effect occurs across a broad range of tested conditions, and is dependent on both the magnitude of divisive scaling and level of stochastic variability. Notably, the normalization model predicts context-dependent effects from varying both distracter value and number, and generalizes to well-known set size phenomena. In behavioral experiments, both monkey and human choosers reproduce the context-dependence predicted by normalized value coding. In both paradigms, we found shallower conditional choice curves and increasingly inefficient choice behavior with larger distracter values. In the human data, we also observed a reversal of context-dependence at very high distracter values, a biphasic effect also predicted in model simulations. These findings suggest that the neural mechanism of value coding critically influences stochastic choice behavior and provides a generalizable quantitative framework for examining context effects in decision-making

*The role of incentive schemes in probability assessment under uncertainty*

Adam S. Goodie, Prashant Doshi, and Daniel G. Hall
University of Georgia

A series of empirical studies investigated probability assessment under conditions of high uncertainty that characterize much combat operation space. This study utilized a simulated UAV theater with realistic scenery of the Bagram AFB in Afghanistan. The subject is tasked with guiding her UAV I from the initial sector to the target sector without being spotted by an opponent UAV J. J follows a trajectory that loops around continuously and independently of I. Studies utilized human participants from the University of Georgia's Psychology Research Pool, as well as UGA Army and Air Force ROTC cadets. Participants experienced a training phase of 5 trials, followed by s a test phase of 15 trials, updating subjective probability of reaching the target sector at each of up to 10 decision points within each trial. In Study 1, no systematic inflationary or deflationary bias in the uncertainty expressions could be inferred. In Study 2 probability assessments were tested under proper scoring rules, non-proper rules or no scoring rules. Results showed significant improvement under a "zero-one" scoring rule that was simple but non-proper. In Study 3, an attempt to replicate this finding in a within-subjects design was not successful. In Study 4, a motivational intervention following the 5 training trials was accompanied by learning in all groups during the course of the study, but no significant differences were observed between scoring rule groups.

*Neurally-informed graph-based transductive models for rapid decision making*

Paul Sajda
Columbia University

Our visual systems are amazingly complex rapid decision making machines. Using our brain's visual system we can recognize objects at a glance, under varying pose, illumination, and scale, and are able to rapidly learn and recognize new configurations of objects and exploit relevant context even in highly cluttered scenes. However our brains are subject to fatigue and have difficulty finding patterns in high-dimensional feature spaces that are often useful representations for multimedia data. In this talk I will describe our work in developing a synergistic integration of human visual processing and computer vision via a novel brain computer interface (BCI). Our approach, which we term cortically-coupled computer vision (C3Vision), uses non-invasively measured neural signatures from the electroencephalogram (EEG) that are indicative of user intent, interest and high-level, subjective and rapid reactions to visual and multimedia data. These neural signatures are ingested into a graph-based representation of an image search space, and transductive inference is used to propagate these labels and identify subsets of images in the graph which are likely to be of interest to the individual. I will describe several system designs for C3Vision and current applications that are being developed for government and commercial applications.

*Modeling Real and Artificial Neurons for Robot Control*

Holly Yanco and Abraham Shultz
University of Massachusetts Lowell

Multi-Electrode Arrays (MEAs) are a useful tool for inspecting the activity of living, interacting brain cells in a manner that is not intrusive to the cells. However, the limited lifespan of the cells in vitro and the impossibility of exactly duplicating any particular network of living cells are problems for research with MEAs. This talk describes a simulation method for modeling the layout and networking of the cells in a culture, and for simulating the activity of that network. We are investigating the predictive value of such simulated networks for some aspects of MEA research to allow easier "rough drafts" of experiments, which may allow researchers to determine if the effort of culturing a living neural network is warranted or can help determine possible parameters for the in vitro cultures. We will describe our use of this simulation and in vitro cultures to control a robot arm to point at an object.

*Consensus, Control and Minority Opinion* (Withdrawn)

Bruce J. West
US Army Research Office

A few years ago I helped initiate a program to understand networked decision making and which reestablishes connections that had been severed by the boundaries separating the physical and human sciences. A particularly significant outcome of this research has been the statistical Decision Making Model (DMM) for complex networks that undergoes a phase transition at a critical value of the coupling parameter. The DMM has provided insight into the information exchange between complex networks and why inflexible minorities can dominate the behavior of complex social networks; possible mechanisms to explain the Arab Spring and the Occupy Wall Street movement. DMM dynamics self-organize linked elements on a simple two-dimensional lattice through a phase transition to generate a scale-free interdependent topological structure with an inverse power-law exponent of 1.0 and an inverse power-law distribution of consensus times with index 1.5. The DMM dynamics are sensitive to internally committed minorities and to internal minorities that are externally controlled. These minorities are unresponsive to the host network and are shown to either stabilize or induce social crisis.

## *Achieving the First Reliability Benchmark of a Reliability Program Plan* (Withdrawn)

Nathan Herbert
US Army Materiel Systems Analysis Activity

This paper presents reliability testing best practices and how to gain accurate assessments of the system's reliability through reliability growth modeling. Planning curves are developed in advance of the system-level reliability growth program. The design for reliability effort should eliminate debilitating failure drivers and help get the system reliability on the first benchmark of the planning curve. In order to demonstrate the reliability of the first benchmark, the testing must be operationally relevant and balanced with respect to the main stressors of the system. Such testing can produce data that are more reflective of the system's reliability under tactical use conditions. Utilizing this balanced data set and assigned fix effectiveness factors of the corrective actions, a reliability projection may be compared to the second benchmark of the planning curve in order to monitor progress of the system reliability growth. This paper illustrates the testing procedures and modeling techniques used by an actual Department of Defense acquisition program that achieved their first benchmark on the planning curve.

## *Reliability Growth Projection Applicable to Multiple Units Under Test* (Withdrawn)

Paul Ellner and Martin Wayne
US Army Materiel Systems Analysis Activity

This paper presents a variant of the currently used estimation procedure for the parameters of the AMSAA Maturity Projection Model (AMPM). The AMPM model projects the reliability impact of corrective actions (termed fixes) to problem failure modes discovered in test. The original version of AMPM utilized the problem failure mode cumulative first occurrence times to estimate the model parameters. This allowed the model to be applied to corrective action strategies that employed all delayed fixes (implemented at the end of a test period), fixes implemented during the test period, or a mixture of both delayed fixes and fixes occurring sometime after the first occurrence of the failure mode and before the end of the test period. The paper presents an estimation procedure based on the first occurrence and all subsequent occurrences of failure due to each potential problem failure mode. This allows for the use of individual failure mode fix effectiveness factor assessments to be utilized, i.e., the expected reduction due to a fix in the occurrence rate of failures associated with a failure mode. The presented estimation procedure can also provide more accurate assessments of a number of projection metrics such as the initial problem mode rate of occurrence, the rate of occurrence of new problem failure modes, and the growth potential mean-time-between failure (the value approached as the rate of occurrence of new problem failure modes approaches zero). The presented estimation procedure is especially suited to assessing the projected reliability and associated metrics for the often encountered case where multiple units of a system are undergoing reliability growth testing over test periods of different time durations that may be executed over different calendar periods. The presented estimation procedure can also be applied to the situation where due to the test schedule, corrective action facility and equipment availability, and other practical reasons, the fix to a failure mode observed on at least one of the units under test is physically implemented on the test units at different times.

*Discrete Approximations of Controlled Stochastic Systems with Memory* (Withdrawn)

Mou-Hsiung (Harry) Chang
US Army Research Office

We consider some computational issues for an optimal control problem in which the controlled state equation is described by a general class of stochastic functional differential equations with bounded memory. In particular, a finite difference method for numerically computing the viscosity and a solution of the infinite dimensional Hamilton-Jacobi-Bellman equation for the optimal control problem will be discussed.

*A Robotic Trenching Capability Experiment* (Withdrawn)

Barry Bodt
US Army Research Laboratory

The Robotics Collaborative Technology Alliance administered by the US Army Research Laboratory has focus areas in Cognitive Architectures, Human/Robot Interaction, Perception/Sensing, and Dexterous Manipulation and Unique Mobility. Periodically the research program conducts Integrated Research Assessments (IRA) to evaluate component technologies integrated on a robotic platform. In IRA2, a robotic arm with claw was mounted on a small Dragon Runner robotic platform. The goal was to assess performance of the autonomous trenching function over varied terrain composition, topography, obstacle clutter, robot control, and arm motion. A straightforward experimental design is discussed in the context of an interesting technology, with encouraging findings reported for an autonomous trenching capability.