Statistical and Data Literacy in the Era of Big Data (Keynote)

Kirk Borne Booz Allen Hamilton

This talk will address one of the fundamental building blocks of data science know-how and analytics capability that is often over-looked. That core competency is statistical and data literacy. There is so much focus on the abundant uses of data resources and analytics tools in conjunction with statistical and data science algorithms that we often overlook some of the unavoidable abuses and misuses that creep in. Concepts that contribute to data and statistical literacies will be discussed. As data scientists, we may take an understanding of such basic concepts as a given, and consequently we would assume that our stakeholders and decision-makers will grasp our jargon and appreciate the wonders of our coolest algorithms. The requisite data and statistical literacies could exist in some cases, but that is not a safe starting assumption when consulting, advising, training, and communicating data science benefits and results to a broader audience. Various data and statistical literacies will be illustrated through examples, including some classic cases and some reallife failures in the media.

Estimating Network Degree Distributions from Sampled Networks: An Inverse Problem

Eric Kolaczyk Boston University

Networks are a popular tool for representing elements in a system and their interconnectedness. Many observed networks can be viewed as only samples of some true underlying network. In that case, a question of fundamental interest is how to estimate characteristics of the underlying network using only information in the sampled network. One of the most fundamental descriptors of a network is its degree distribution. We study the problem of how to estimate the degree distribution of a true underlying network from its sampled network, under various common network sampling designs. We show that it can be formulated as a linear inverse problem that is, in many cases, ill-posed. Accordingly, we offer a penalized least-squares approach to solving this problem, with the option of additional constraints. The resulting estimator is a linear combination of singular vectors of a matrix, relating the expectation of our sampled degree distribution to the true underlying degree distribution, which is defined entirely in terms of the sampling plan. We discuss several aspects of this problem that are atypical for linear inverse problems, due to the unique nature of network sampling, and offer solutions accordingly. We present the results of a simulation study, characterizing the performance of our proposed method, and we illustrate its use in the context of monitoring large-scale social media networks.

Cyberlanguage: Enhancing cybersecurity through statistical and natural language processing techniques

J.A. Marr, DZYNE Technologies

B.D. Richardson, D.B. Krisiloff, B.J. Radford, and C.M. Morris, Sotera Defense Solutions

Cybersecurity is a rapidly growing field for which big data technologies and cutting-edge statistical techniques promise to deliver enhanced situational awareness for defensive applications. Existing software solutions in this domain rely largely on heuristic-based approaches to process the high volume of logs produced by network sensors. Unfortunately, simple analytics and signature-matching depend on known attack patterns and provide little assurance against advanced adversaries. In this paper, we treat the cybersecurity problem as one of anomaly detection and investigate unsupervised learning techniques from natural language processing and related fields to identify network intrusions. Additionally, we discuss the challenge of model validation for unsupervised cybersecurity applications.

Bayesian spatial model selection for detection and identification in chemical plumes based on hyperspectral imagery data

Abel Rodriguez University of California, Santa Cruz

The use of hyperspectral imagery in the remote sensing has proven to be important for a wide variety of defense applications and beyond. For example, hyperspectral images can be used to detect gas plumes that are invisible to the human eye, and to identify their chemical structure. A hyperspectral image is a massive cube of data consisting of thousands of pixels each with ~100 observations over a range of frequencies in the electromagnetic spectrum. In this talk we discuss novel Bayesian models for detection and identification of chemicals in hyperspectral images. The model combines ideas from Gaussian processes regression, conditionally autoregressive processes and g-priors for model selection to generate models that can account for the non-linear relationships between sensor's inputs and outputs and for the expected structure of gas plumes. We evaluate the models using a variety of datasets and show that they outperform other state-of-the-art methods, in some cases substantially.

Data Farming: Reaping Insights from Large-Scale Simulation Experiments

Susan Sanchez

Simulation models are integral to modern scientific research, national defense, industry and manufacturing, and public policy debates. These models tend to be extremely complicated, often with large numbers of factors and many sources of uncertainty, but recent breakthroughs help analysts deal with this complexity. Data farming is a descriptive metaphor that captures the notion of generating data purposefully in order to maximize the information "yield" from simulation models. Large-scale designed experiments let us grow the simulation output efficiently and effectively. We can explore massive input spaces, uncover interesting features of complex simulation response surfaces, and explicitly identify cause-and-effect relationships. Data farming has been used in the defense community for over a decade, and has resulted in quantum leaps in the breadth, depth, and timeliness of the insights yielded by simulation models. In this talk, I will give an overview of the principles of data farming, describe some recent applications in defense and homeland security, and present a portfolio of designs suitable for efficiently obtaining insights from high-dimensional simulation models. I will finish with some thoughts about opportunities and challenges for further improving the state of the art, and transforming the state of the practice, in this domain.

Adversarial Risk Analysis (90-minute tutorial)

David Banks Duke University

Adversarial Risk Analysis (ARA) is a Bayesian approach to strategic decision-making. One builds a model of one's opponents, expressing subjective uncertainty about the solution concept each opponent uses, as well as their utilities, probabilities, and capabilities. Within that framework, the decision-maker makes the choice that maximizes expected utility. ARA allows the opponent to seek a Nash equilibrium solution, or a mirroring equilibrium, or to use level-k thinking, or prospect theory, and so forth, and it allows the decision-maker to relax the common-knowledge assumption that arises in classical game theory. The methodology applies to corporate competition and counterterrorism. The main ideas are illustrated in the context of auctions, the Borel game La Relance, and a toy counterterrorism example.

Bayesian Analysis (90-minute tutorial)

Kassie Fronczyk Institute for Defense Analyses

In an era of reduced budgets and limited testing, verifying that requirements have been met in a single test period can be challenging, particularly using traditional analysis methods that ignore all available information. The Bayesian paradigm is tailor made for these situations, allowing for the combination of multiple sources of data and resulting in more robust inference and uncertainty quantification. Consequently, Bayesian analyses are becoming increasingly popular in T&E. This tutorial briefly introduces the basic concepts of Bayesian Statistics, with implementation details illustrated in R through two case studies: reliability for the Core Mission functional area of the Littoral Combat Ship (LCS) and performance curves for a chemical detector in the Common Analytical Laboratory System (CALS) with different agents and matrices.

Combinatorial Methods in Software Testing (90-minute tutorial)

Rick Kuhn NIST

Combinatorial methods have attracted attention as a means of providing strong assurance at reduced cost. Combinatorial testing takes advantage of the interaction rule, which is based on analysis of thousands of software failures. The rule states that most failures are induced by single factor faults or by the joint combinatorial effect (interaction) of two factors, with progressively fewer failures induced by interactions between three or more factors. Therefore if all faults in a system can be induced by a combination of t or fewer parameters, then testing all t-way combinations of parameter values is pseudo-exhaustive and provides a high rate of fault detection. The talk explains background, method, and tools available for combinatorial testing, with examples from several case studies. Methods for combinatorial coverage measurement are also introduced, with measures of input space values combinations and the relationship of these measures with traditional structural coverage of code.

Resampling Methods (90-minute tutorial)

Dave Ruth US Naval Academy

This tutorial presents widely used resampling methods to include: bootstrapping, permutation tests, and cross-validation. Underlying theories will be presented briefly, but the primary focus will be on applications. Examples will be demonstrated in R; participants are encouraged to bring their own portable computers to follow along using datasets provided by the instructor.

An Introduction to Survey Research (90-minute tutorial)

Heather Wojton Institute for Defense Analyses

Researchers and policy makers throughout the Department of Defense encounter surveys in their daily work. Surveys help us understand, for example, how military personnel interact with weapons systems and if implementing certain programs or policies improves the well-being of military personnel and their families. This 90-minute course serves as an introduction to survey research. Specifically, we explore the role of psychology in survey construction and response, common test designs in survey research, and the implications of these designs for data analysis. There are no specific prerequisites for this course. However, attendees should bring paper and pencil to fully engage with the course material.

The DASE Axioms: Designing Simulation Experiments for Verifying Performance of Software-Intensive Systems (Special Session on Computer Experiments)

T. Hurst, I. Goodrich, C. Leigh, C. Pouchet, M. Tolman, M. Wynn, and J. Zink Raytheon Missile Systems

The paper describes several innovations and lessons learned through applying and expanding design-of-experiments (DOE) principles to conduct simulation experiments for verifying requirements compliance of software-intensive, closed-loop systems.

Design for Large-scale Statistical Computation and Distributed Computer Experiments (Special Session on Computer Experiments)

Peter Qian University of Wisconsin-Madison

Big Data appear in a growing number of areas like marketing, physics, biology, engineering, and the Internet. For example, for every hour, more than one million transaction data are stored in WalMart database and a HPC based computer model can produce results of millions of runs. While large volume of data offers more statistical power, it also brings computational challenges.

We first introduce an experimental design algorithm, called orthogonalizing EM (OEM), intended for various least squares problems. The main idea of the procedure is to orthogonalize a design matrix by adding new rows and then solve the original problem by embedding the augmented design in a missing data framework. We demonstrate that OEM is highly efficient for large-scale least squares problems.

We then present a reformulation and generalization of OEM that leads to a reduction in computational complexity for least squares and penalized least squares problems. The reformulation, named the GOEM (Generalized Orthogonalizing EM) algorithm, is further extended to a wider class of models including generalized linear models and Cox's proportional hazards model. Synthetic and real data examples are included to illustrate its efficiency compared with standard techniques.

Finally, we will discuss several new classes of space-filling designs inspired by Samurai Sudoku for conducting distributed computer experiments. A growing trend in science and engineering is to distribute runs of a large computer experiment across different groups, machines or locations. Due to the complexity of the hardware and the simulation code, some batches in such an experiment may malfunction or fail to converge. By ensuring that the analysis can be done at the batch level and the experiment level, these new designs provide a robust solution to this problem. We will also talk about applications of these designs in solving optimization under uncertainty problems.

Activity-Based Intelligence and Pattern-of-Life Analysis (Special Session on Quantitative Methods in Intelligence)

Pat Biltgen Vencore

Activity-based intelligence (ABI) is a new methodology for intelligence analysis that uses spatially enabled "big data" to enhance national security. ABI methods – developed for counterterrorism – are being applied to a broad spectrum of intelligence issues. Increasingly, analysts are being challenged to think statistically and integrate multiple types of data to understand complex human behaviors over space and time. This presentation will demonstrate how statistical tools are used to visually explore data across space and time. We demonstrate how the linked data paradigm allow an intelligence analyst to seamlessly move between spatial and statistical views, dynamically filtering data to understand patterns and trends. An intuitive graphical interface lets analysts ask questions to discover previously unknown entities (people, vehicles), correlate activities across data sets, and develop interactive "spatial stories" that describe complex activities in the human dimension. Although these methods are used by government intelligence analysts, the presenter will demonstrate public domain examples from commercial marketing and the Internet of Things to illustrate basic principles and stimulate discussion about how these methods can be cross-fertilized with techniques in other disciplines.

Data Science for Threat Finance Intelligence (Special Session on Quantitative Methods in Intelligence)

I. Kloo

Center for Army Analysis

This presentation demonstrates how data science methods can be used to identify critical vulnerabilities in threat networks. Specifically, we describe a scientific method for processing bulk financial transaction records to develop meaningful (graph) representations of threat finance networks. Analyzing these network representations facilitates the identification of critical vulnerabilities within them, such as financial institutions supporting multiple criminal/terrorist networks.

Leveraging Data Across Operational Test Unit Variants to Assess Reliability

S. Brady, P. Ellner, and M. Wayne Army Materiel Systems Analysis Activity (AMSAA)

The Army conducts Operational Testing (OT) to determine the effectiveness and suitability of a system. To adequately assess whether a system's reliability requirement has been demonstrated with statistical confidence using classical methods, OT typically requires multiple test units to be operated for relatively long durations. For programs with high reliability requirements, costly test units, or various resource and schedule constraints, a sufficient amount of testing is often infeasible. In cases where OT provides limited test data (i.e., short test duration or no/few failures), there can be large uncertainty associated with the reliability estimates for those particular units. As a result, the Army assumes an increased risk of fielding systems that may perform below expectations which negatively impacts mission success and increases operating and support costs.

To address this issue AMSAA developed methodology for providing improved reliability assessments under constrained test and evaluation environments. The methodology utilizes a test data driven Bayesian approach to appropriately leverage data across test unit variants to assess reliability. The paper presents the methodology, demonstrates its benefits using sample test data, and outlines potential follow-on initiatives.

Test Design Adequacy for Logistic Regression

Kendal Ferguson and Scott Hunter Dugway Proving Ground

Assessment of test design adequacy, typically in terms of statistical power, is an area of continuing research. Often, power is assessed with respect to one or more assumed logarithms of odds ratios. Most decision makers would benefit from a clearer portrayal of the adequacy of a proposed test design. This paper presents a SAS macro that estimates several measures of test design adequacy for a basic logistic regression model with three coefficients. Measures include statistical power for finding the marginal and/or joint significance of the coefficients and the width of user-specified confidence intervals. The macro uses simulations to assess test design adequacy and produces graphical output for improved communication of analysis results. Results are based on data randomly generated from a "true" model that is generated by simple-to-understand, user-specified assumptions about the relationship between the probability of success and the independent variables used in the analysis. Further flexibility is added by incorporating one or more random components, which change the function during each simulation, to account for uncertainty regarding the "true" model. An example of this macro, with respect to a chemical detector, is presented.

Fast Patchwork Bootstrap for Quantifying Estimation Uncertainties in Sparse Random Network

Y. Gel, V. Lyubchich, and L. Ramirez

University of Texas at Dallas, University of Maryland Center for Environmental Sciences

We propose a new method of nonparametric bootstrap to quantify estimation uncertainties in large and possibly sparse random networks. The method is tailored for inference on functions of network degree distribution, under the assumption that both network degree distribution and network order are unknown. Moreover, the network can be only partially observable. The key idea is based on adaptation of the blocking argument, developed for bootstrapping of time series and re-tiling of spatial data, to random networks. That is, our idea behind the bootstrap path is intuitive: as the classical bootstrap was originally suggested for independent and identically distributed data and then adapted to time series and spatial processes, we borrow the blocking argument developed for resampling of space and time dependent processes and adjust it to networks. To diminish bias in estimating a degree distribution, we employ a new sampling procedure of Labelled Snowball Sampling with Multiple Inclusion (LSMI). We also develop a new computationally efficient and data-driven cross-validation algorithm for selecting an optimal block size. The significance of our approach is the following. First, while there exists a vast literature on graph sampling for estimating network properties, very little is known on how to reliably evaluate associated errors of estimation (outside of extensive, information costly and typically impractical simple random sampling). Second, computing statistics on the entire graph data may be either infeasible due to computational costs or because the entire graph data may be unavailable, e.g., due to privacy issues. Third, the problem with estimation uncertainties is further aggravated when network analysis is performed from a single network or from a series of time evolving dependent graphs rather than a series of independent graphs with the same probabilistic structure. For instance, while evaluating networks of potential money laundering schemes or terrorist communication, we cannot assume that graphs observed at consecutive time points are independent. Hence, the classical optimality properties for maximum likelihood and other conventional parametric methods are no longer valid. We validate the new bootstrap procedure by extensive simulations and show that the new bootstrap method outperforms the available competing approaches by providing more reliable quantification of uncertainties for functions of a network degree distribution. We discuss future extension of our methodology to bootstrap-based anomaly detection in complex networks and anomaly warning in data stream analysis/processing algorithms.

Many Transformations, 400 New Distributions

A. Glen, S. Butler, and N. Mankovich Colorado College

Using "A Probability Programming Language," the authors present extensive research that automates the probabilistic transformation process to produce over 400 new families of distributions. A review of the process is given, as well as some examples of the unique distributions that result. The effort in cataloging and presenting this large amount of statistical information is also presented.

Cyclic Stochastic Approximation for Multiagent Stochastic Optimization

K. Hernandez, Johns Hopkins University J. Spall, JHU Applied Physics Laboratory

In this work we are concerned with the problem of optimization of multiple agents that are to be used in some process of interest. The agents may represent vehicles, people, machines, etc. Our aim is to collectively optimize the relevant properties (e.g., location, velocity, etc.) of the multiple agents in order to achieve some optimal state of the overall system. An example application might be one where the agents are undersea vehicles and we wish to optimally govern their motion in order to maximize the probability of detecting a target vehicle. A challenging factor in real-world problems is the uncertainty associated with agent properties and the uncertainty in the overall state of the system. The uncertainty implies that stochastic optimization techniques are necessary in order to optimize agent properties. We discuss how cyclic stochastic approximation can be used to perform the necessary stochastic optimization in this multiagent setting. In addition, we discuss theoretical results on cyclic stochastic approximation and comment on possible lines of future research.

An Assessment of Energy Efficient Planning

C. Lennon, M. Childers, M. Harper, C. Ordonez, N. Gupta, J. Pace, R. Kopinsky, A. Sharma, E. Collins, and J. Clark US Army Research Laboratory and Florida A&M University-Florida State University

The Army Research Laboratory's Robotics Collaborative Technology Alliance (RCTA) is a program intended to change robots from tools that soldiers use into teammates with which soldiers can work. One desired ability of such a teammate is the ability to operate in an energy efficient manner on a variety of surfaces. Researchers from the RCTA have developed planning algorithms which incorporate knowledge of the vehicle's steering and control system into path planning. This algorithm learns appropriate parameter values by conducting a brief set of trial maneuvers, and is intended to enable the robot to operate in manner which is both more energy efficient, and in which the risk of collision is reduced. We present the results of an assessment of this technology, conducted by comparing the RCTA planning algorithm to a traditional minimum distance planning algorithm.

Software Failure and Reliability Assessment Tool (SFRAT): An Open Source Application for the Practitioner and Research Community

V. Nagaraju, T. Wandji, and L. Fiondella University of Massachusetts, Dartmouth and Naval Air Systems Command, Patuxent River

Reliability has been an ongoing challenge for the Department of Defense. More recently, software reliability has become a significant concern as the majority of systems acquired are software intensive. It can require twenty-four months or more to stabilize the failure rate of software below a target threshold, extending completion time and delivery to the warfighter. Such delays also come with additional cost. Methods to proactively assess software reliability throughout acquisition are needed to mitigate these schedule and cost risks. This talk describes Software Failure and Reliability Assessment Tool (SFRAT), which is an open source application designed for the practitioner and research community. SFRAT is suitable for practitioners because it implements software reliability research into an application that promotes an intuitive workflow. The tool has been programmed in R and provides functionality through a Shiny graphical user interface, greatly reducing the need for knowledge of the underlying statistical techniques. This interface will help contractors quantitatively assess the software they develop as part of their data collection and reporting process. However, SFRAT is not just a modernized version of previous tools such as SMERFS and CASRE. It is both open source and extensible. The open source nature of the tool means that the source code is free and publicly accessible. This serves multiple purposes. Access to the source code simplifies the information assurance process, facilitating approval for use in defense and security settings. The code can also be modified by individuals and incorporated into the automated testing procedures of their organizations, enabling more frequent assessment for internal purposes as well as reporting. The extensible nature of the tool allows researchers to add additional models and measures of goodness of fit. This will enable researchers to reach users who can assess the ability of alternative models to characterize failure data. The tool architecture can be further extended to encompass reliability assessment in additional stages of the software lifecycle.

Group-wise and Regional Terrorism Trends via Clustering in the Model Space

V. Raghavan Qualcomm Flarion Technologies

The focus of this work is on a comparative analysis of terrorism trends across a broad geographical swathe over the 1970-2010 period. While a number of such studies exist in the literature, the focus of this work is from a model learning perspective. In particular, this work studies changes in terrorism trends by their impact on the parameters in a recently introduced hidden Markov modeling (HMM) approach for capturing the activity of a terrorist group. The terrorist data considered in this work corresponds to 28209, 19166, 6802, 17727 and 14701 attacks in i) Latin and South America, ii) West Asia, North Africa and Central Asia, iii) Southeast/East Asia and Australasia, iv) South Asia, and v) Western Europe, respectively — all from the Global Terrorism Database (GTD) maintained by the UMD START Center. After addressing trends in standard metrics such as number of attacks, known fatalities and injuries, number of "massive events," type and target of attacks, etc., this work branches off into group-wise and regional trends via clustering in the model space. It is shown that such clustering can help in identifying both macroscopic commonalities across groups as well as in microscopic behavior such as attack/target types.

DoD Applications of Sensitivity Testing and DOE for Binary Response

D. Ray, C. Drake, and P. Roediger US Army ARDEC, UTRS, Inc.

This presentation will provide a brief overview of sensitivity testing, detail the US Army ARDEC's development and implementation of the '3pod' sensitivity test algorithm in R statistical computing language, and emphasize a variety of adaptations of sensitivity testing for binary response data, including adaptive, nonadaptive, and hybrid approaches applied to several products and system of importance to the US Army and Joint Services, including Insensitive Energetics, Ballistic testing of protective armor, testing of munition fuzes and Microelectromechanical Systems (MEMS) components, and safety testing of high-pressure test ammunition.

Uncertainty Quantification of Armament Engineering Models and Simulations

D. Ray and M. Jablonski US Army ARDEC

The US Army ARDEC has recently established an initiative to integrate statistical and probabilistic techniques into engineering modeling and simulation (M&S) analytics typically used early in the design lifecycle to guide prototype development. DOE-driven Uncertainty Quantification techniques enable engineering design teams to study the impact of variations in design parameters, and identify opportunities to make technologies more robust, reliable, and resilient earlier in the product's lifecycle. Several recent armament engineering case studies – each with unique considerations and challenges – will be discussed.

A new sensitivity testing procedure when there are two stress variables

C.F.J. Wu

Georgia Institute of Technology

Sequential sensitivity testing occurs commonly in testing of military hardware or ammunition and in toxicity studies. I will first review a recent work called "3pod design" (Wu and Tian, JSPI, 2014). It consists of three phases which can be described as a trilogy of search-estimate-approximate. Then I will address a rarely studied problem with two stress variables. It occurs quite commonly but most work in the literature is on one stress variable. Some new ideas will be outlined, part of which is inspired by the modular nature of the 3pod design mentioned above, especially the concept of "separation" and "overlapping" pattern. An illustration of the new procedure will be presented.