Conference Program
About the CoDA Banner

The background image shows the circuitry of MANIAC, the Mathematical Analyzer, Numerator, Integrator, and Computer, an early computer designed and built at Los Alamos in the 1950s by a team led by Nicholas Metropolis.


In 1942 Nick Metropolis was working with Edward Teller on the reactor project at the University of Chicago when J. Robert Oppenheimer invited the young physicist to continue his collaboration with Teller, but at Los Alamos. There Metropolis joined the Manhattan Project as a member of the Theoretical Division, having been encouraged by Teller to move from experimental to theoretical physics. His first assignment was to develop equations of state for materials at high temperatures, pressures, and densities.

Over the years Metropolis turned increasingly to mathematics and computer design, and by 1948 he was leader of a Los Alamos team that designed and built the MANIAC, one of the first electronic digital computers. Many of the country’s foremost scientists were eager to try their experiments on the wonderful new machine and came to the Laboratory to work with Metropolis. A few years later, together with Teller, John von Neumann, Stanislaw Ulam, and Robert Richtmyer, Metropolis developed techniques and algorithms for using the Monte Carlo method (so named by Metropolis) on the new computers.

The Monte Carlo method is an application of the laws of probability and statistics to the natural sciences. The essence of the method is to use various distributions of random numbers, each distribution reflecting a particular process in a sequence of processes such as the diffusion of neutrons in various materials, to calculate samples that approximate the real diffusion history. Statistical sampling had been known for some time, but without computers the process of making the calculations was so laborious that the method was seldom used unless the need was compelling. The computer made the approach extremely useful for many physics problems.

Metropolis was also involved in the development of an importance-sampling scheme, called the Metropolis algorithm, that improves the effectiveness of the Monte Carlo method.

Paul Stein (left) and Nicholas Metropolis test MANIAC’s skill at chess. Image courtesy of Los Alamos National Laboratory. Back cover: Output from MANIAC.
Overview

Wednesday, February 29, 2012
- 8:45-12:15 Featured and Invited Talks
- 1:30-4:45 p.m. Featured and Invited Talks
- 5:15-7:30 p.m. Mixer and Poster Session

Thursday, March 1, 2012
- 8:45-12:15 Featured and Invited Talks
- 1:30-4:45 p.m. Featured and Invited Talks
- 6-8 p.m. Dinner Banquet (La Fonda, 100 E. San Francisco St.)

Friday, March 2, 2012
- 8:45-12:15 Invited Talks
- 1:45-5 p.m. Invited Talks

Organizing Committee
- Kary Myers, Los Alamos National Laboratory, Chair
- Earl Lawrence, Los Alamos National Laboratory, Co-Chair
- David Banks, Duke University
- Derek Bingham, Simon Fraser University
- Vera Bulaevskaya, Lawrence Livermore National Laboratory
- Paul Kidwell, Lawrence Livermore National Laboratory
- Max Morris, Iowa State University
- George Ostrouchov, Oak Ridge National Laboratory
- Shane Reese, Brigham Young University
- Dave Robinson, Sandia National Laboratories
- Mark Tardiff, Pacific Northwest National Laboratory
- Sandy Thompson, Pacific Northwest National Laboratory and DOE Headquarters

Sponsors
CoDA would not be possible without the financial support of the following sponsors:
- Los Alamos National Laboratory Information Science and Technology Center
- Los Alamos National Laboratory Science Program Office, Office of Science
- Los Alamos National Laboratory Statistical Sciences Group
- American Statistical Association Section on Physical and Engineering Sciences
- American Statistical Association Section on Defense and National Security
- SAS Institute Inc., JMP Division
- American Society for Quality, Statistics Division
Wednesday, February 29, 2012

8:45-9 a.m. | Welcome

9-10 a.m. | Dave Higdon, Los Alamos National Laboratory
Uncertainty Quantification: Combining Large Scale Computational Models with Physical Data for Inference

10-10:30 a.m. | Break

10:30 a.m.-12:15 p.m. | Statistical Issues in Climate and Energy Modeling
Organizer: Vera Bulaevskaya, Lawrence Livermore National Laboratory

- Gardar Johannesson, Lawrence Livermore National Laboratory
  Uncertainty Quantification and Predictive Computer Models: LLNL’s UQ Effort and its Application to Climate Models
- Ryan Elmore, National Renewable Energy Laboratory
  A Large-Scale Sensitivity Analysis of the Biomass Scenario Model’s Input Variables
  Addressing Model Selection Uncertainty in Bioremediation Experiments through Bayesian Model Aggregation

12:15-1:30 p.m. | Lunch (on your own)

1:30-2:30 p.m. | Katrin Heitmann, Argonne National Laboratory
Exploring the Dark Universe: Statistical and Data Challenges

2:30-3 p.m. | Break

3-4:45 p.m. | Large Scale Statistical Computing
Organizer: George Ostrouchov, Oak Ridge National Laboratory

- Tamara G. Kolda, Sandia National Laboratories in California
  The BTER Graph Generator: A Scalable Model with Measurable Community Structure
- Christopher Paciorek, University of California Berkeley
  Parallel Statistical Extreme Value Analysis of Climate
- George Ostrouchov, Oak Ridge National Laboratory
  Data Parallel Statistical Computing and R

5:15-7:30 p.m. | Mixer and Poster Session
Abstract
Combining physical measurements with computational models is key to many investigations involving validation and uncertainty quantification (UQ). This talk surveys some of the many approaches taken for validation and UQ, suggesting classifications of different types of problems with common features (e.g. data size, amount of empiricism in the model, availability of data, extent of extrapolation required, etc.). Similar problems may be amenable to similar UQ approaches, many of which are outlined in this talk. This talk concludes by considering potential new research directions for methodological development in UQ.

Biography
Dave Higdon was the Group Leader for the Statistical Sciences Group at Los Alamos National Laboratory until 2010. He is an internationally recognized expert in Bayesian statistical modeling of environmental and physical systems. He has also led numerous programmatic efforts at LANL in the quantification of margins and uncertainties and uncertainty quantification. His recent research has focused on simulation-aided inference in which physical observations are combined with computer simulation models for prediction and inference.
Uncertainty Quantification and Predictive Computer Models: LLNL’s UQ Effort and Its Application to Climate Models | 10:30-11:05 a.m.
Gardar Johannesson, Lawrence Livermore National Laboratory

Predictions from computer models can be off due to numerous factors, including: uncertainty in input parameters and data, approximated physics, numerical discretization, etc. Projections from global climate models are good examples. Here we report on ongoing research effort at the Lawrence Livermore National Laboratory to advance the science of uncertainty quantification (UQ) of simulation codes and the application of UQ to a climate models. The effort is mainly focused on ensemble-based UQ methods, where a large ensemble of simulations is created by perturbing uncertain inputs. We present some results on an ongoing UQ ensemble-study of the various model components of the Community Earth System Model (CESM), including results of sensitivity analysis and statistical calibration of uncertain input parameters.

A Large-Scale Sensitivity Analysis of the Biomass Scenario Model’s Input Variables | 11:05-11:40 a.m.
Ryan Elmore, National Renewable Energy Laboratory

The Biomass Scenario Model (BSM) is a systems dynamic simulation model of the domestic biofuels supply chain. Results from BSM simulations may provide insight into the entire biofuel system behavior and policy/incentive effectiveness and/or limitations. In this talk, we will discuss our experiences with the first large-scale sensitivity analysis (SA) involving approximately 2500+ input variables of the BSM (version 3). Our talk will focus on the computational challenges involved in our study, the experimental design, and the preliminary results. Finally, we will briefly discuss our future plans regarding the SA and uncertainty quantification efforts related to the BSM. This is joint work with Brian Bush, David Hsu, Yolanda Lin, Monte Lunacek.

Addressing Model Selection Uncertainty in Bioremediation Experiments through Bayesian Model Aggregation | 11:40 a.m.-12:15 p.m.
Luke Gosink, Pacific Northwest National Laboratory

Strategies for applying technologies to complex problems such as carbon sequestration, chemical waste bioremediation, or subsurface contaminant transport are often based on the predictive results obtained from numerical simulation. A fundamental challenge to accurately assessing the accuracy of the modeled results arises from the fact that the underlying conceptual models that define the processes and outcomes are rarely known with a high degree of confidence. In this context, model selection uncertainty – i.e. the uncertainty associated with correctly selecting a set of domain-specific interpretations, processes, and mathematical systems - is one of the leading sources for error in predictive modeling.

This talk presents a Bayesian approach for constructing an aggregate prediction from an ensemble of initial models, and examines the two-fold benefit of this aggregate to help ameliorate the uncertainties associated with model selection. We first present the predictive performance trends for aggregates in comparison to those of the ensemble’s constituents. Next, we examine the aggregate’s diagnostic capabilities as a tool to help identify which constituents in the ensemble are more or less accurate than others. We propose that this diagnostic information provides a critical feedback loop for modelers to help refine and tune their assumptions for the problem. Our experiments and analysis are based on field data from an ongoing bioremediation experiment.
Exploring the Dark Universe: Statistical and Data Challenges
1:30-2:30 p.m.

Abstract
Cosmology -- the study of the origin, evolution, and constituents of the Universe -- is now entering one of its most scientifically exciting phases. Two decades of surveying the sky have culminated in the celebrated “Cosmological Standard Model”. Yet, two of its key pillars, dark matter and dark energy -- together accounting for $95\%$ of the mass-energy of the Universe -- remain mysterious. Deep fundamental questions demand answers: What is dark matter made of? Why is the Universe's expansion rate accelerating? Should general relativity be modified? What is the nature of primordial fluctuations? What is the exact geometry of the Universe? To address these burning questions, survey capabilities are being spectacularly improved. Next-generation observatories will open new routes to understand the true nature of the “Dark Universe”. These observations will pose tremendous challenges on many fronts -- from the sheer size of the data that will be collected (more than a hundred Petabytes) to its modeling and interpretation. The interpretation of the data requires sophisticated simulations on the world's largest supercomputers. The cost of these simulations, the uncertainties in our modeling abilities, and the fact that we have only one Universe that we can observe opposed to carrying out controlled experiments, all come together to create a major test for statistical methods of data analysis.

In this talk I will give a very brief introduction to the Dark Universe and outline the challenges ahead. To combat these challenges, close cross-disciplinary collaborations between physicists, statisticians, and computer scientists will be crucial. I will discuss two examples of successful collaborative work and propose new tasks where cosmologists urgently need help from the data and statistics community.

Biography
Katrin Heitmann obtained her PhD in physics at the University of Dortmund, Germany, in 2000. After finishing a three-year postdoctoral fellowship at Los Alamos National Laboratory, Katrin was hired as technical staff member in the Space Science and Application Group at LANL. In 2011 she moved to Argonne National Laboratory where she holds a joint appointment in High Energy Physics and Mathematics and Computer Science Divisions. She is also a Senior Member of the Kavli Institute for Cosmological Physics at the University of Chicago. Katrin's research interests in cosmology span a large range of subjects, starting with the very early Universe and the inflationary phase, to the formation of large-scale structures. Her research on precision studies of the “Dark Universe” makes essential use of large-scale supercomputing and advanced statistical methods.
The BTER Graph Generator: A Scalable Model with Measurable Community Structure | 3-3:35 p.m.
Tamara G. Kolda, Sandia National Laboratories

Community structure plays a significant role in the analysis of social networks and similar graphs, yet this structure is little understood and not well captured by most generative graph models. We propose the Block Two-Level Erdős-Rényi (BTER) model, and demonstrate that it accurately captures the observable properties of many real-world social networks. Our model is based on the idea that a community is a subgraph that is internally highly connected and has no deeper substructure. We use tools of combinatorics to show that any such community must contain a dense Erdős-Rényi (ER) subgraph and moreover that any graph with a heavy-tailed degree distribution and community structure must contain a scale free collection of dense ER subgraphs. These theoretical observations corroborate well with empirical evidence and are the basis of our proposed model. The BTER generator can produce any desired degree distribution and maintain a high clustering coefficient while at the same time being a fully scalable model. We propose that BTER should replace the Stochastic Kronecker Graph (SKG) generator for the Graph 500 benchmark. This is joint work with C. Seshadhri and Ali Pinar.

Parallel Statistical Extreme Value Analysis of Climate | 3:35-4:10 p.m.
Christopher Paciorek, University of California, Berkeley

Extreme weather events are a key aspect of climate, but are difficult to study because of sparse data. Statistical extreme value analysis provides a rigorous methodology to analyze extremes, either through analysis of maxima within time blocks (Generalized Extreme Value: GEV) or exceedances over a high threshold (Peaks over Threshold: POT). However, current statistical methods for spatial extremes focus on joint models for multiple locations that are computationally infeasible for large numbers of locations. Even analysis at individual locations will be slow for high-resolution model output such as in the CMIP5 archive.

We focus on a statistical approach that allows for embarrassingly parallel analysis of individual locations. The basic approach is to fit individual GEV and POT models at individual locations, allowing for simple trends in time. To borrow strength spatially, we smooth in space using a local likelihood approach at each location. For uncertainty assessment, we use a bootstrap approach that preserves spatial structure. We analyze US station data and global climate model output for temperature and precipitation using these techniques.

Parallel processing is required to perform statistical EV analysis at thousands of locations, in particular when doing local likelihood and bootstrapping. Within the context of a DOE-funded, Berkeley Lab-headed effort to develop parallel software tools for climate analysis, we are integrating R with the parallel visualization software VisIt to allow climate scientists to perform parallel EV analysis. VisIt is handling parallel input/output, preprocessing of maxes and exceedances, and initialization of multiple serial R jobs. R is then used to fit EV models at each location. This is joint work with Michael Wehner and Prabhat (Lawrence Berkeley National Laboratory) and Dave Pugmire (ORNL).

Data Parallel Statistical Computing and R | 4:10-4:45 p.m.
George Ostrouchov, Oak Ridge National Laboratory

This overview talk will begin with some motivation for data parallel statistical computing and describe its current state within the R statistical computing environment. Parallel computing capability comes in roughly three flavors: message-passing (distributed), shared-memory (multicore), and coprocessor (GPU). All three have some representation in R with much overlap between packages. Truly massive data sets require end-to-end parallel treatment. This means parallel file systems, parallel data readers, parallel statistical analysis, and parallel connections between analysis steps and graphics. Many of these pieces are still missing in R but some can be made available by connecting with other software. The Department of Energy funds a few projects that are making an impact on R’s parallel capability and on applications that utilize this capability.
Thursday, March 1, 2012

8:45-9 a.m. | Announcement of the Statistics in Defense and National Security Student Poster Award Winner

9-10 a.m. | Charlie Nakhleh, Sandia National Laboratories
Addressing Uncertainties in High Energy Density Physics: Experiment and Simulation

10-10:30 a.m. | Break

10:30 a.m.-12:15 p.m. | Applications of Statistics in Homeland Security
Organizer: Kary Myers, Los Alamos National Laboratory

- Shane Reese, Brigham Young University
  Adaptive Testing of Chemical and Biological Detectors
- Kristin P. Lennox, Lawrence Livermore National Laboratory
  Validation of HADES MicroCT Radiograph Prediction
- Christopher Oehmen, Pacific Northwest National Laboratory
  Data Analysis for Homeland Security Applications: Bio-Inspired Approaches to Optimization and Signature Discovery

12:15-1:30 p.m. | Lunch (on your own)

1:30-2:30 p.m. | Deb Agarwal, Lawrence Berkeley National Laboratory
Changing Your Perspective from Serving the Data to Enabling Data Users

2:30-3 p.m. | Break

3-4:45 p.m. | Statistical Network Analysis
Organizer: Dave Robinson, Sandia National Laboratory

- Jaideep Ray, Sandia National Laboratories
  Estimating a Thinning Ratio for a Markov Chain of Graphs
- Scott Vander Wiel, Los Alamos National Laboratory
  Uncertainty Quantification for Networks with Power Distribution Applications
- Dan Nordman, Iowa State University
  An Illustration of Bayesian Nonparametric Models for Community Detection

6-8 p.m. | Banquet, La Terraza at la Fonda, 100 E. San Francisco St.

Speaker: Graham Walsh, Energetic Materials Research and Testing Center, New Mexico Tech
Blowing Stuff Up for Science: Explosives Experiments with the MythBusters
Abstract
The field of high-energy-density physics (HEDP) is concerned with the behavior of systems at pressures greater than one million atmospheres, or 1 Mbar. Because pressure is proportional to both density and temperature, HEDP systems can span a wide regime of physical conditions, from moderate density, but very hot plasmas, to relatively cold, but quite dense Fermi-degenerate systems. HEDP systems range from naturally occurring systems like planetary and stellar interiors to laboratory systems like Inertial Confinement Fusion (ICF) capsules. Obtaining a detailed understanding of HEDP systems faces challenging experimental and theoretical problems. On the experimental side, HEDP systems are created and studied in mega-Joule (MJ) class facilities, such as the National Ignition Facility (NIF) at Lawrence Livermore National Laboratory (LLNL) and the Z Pulsed Power Facility at Sandia National Laboratories. Such experiments typically last several billionths of a second, and often achieve temperatures in excess of 10 million Kelvin. It is a tremendous technological challenge to obtain detailed experimental information about such short-lived systems under such extreme conditions. On the theoretical side, HEDP systems are characterized by a very large number of degrees of freedom interacting in a highly nonlinear manner. Analytical theory, while important in developing a qualitative picture of HEDP systems, is usually limited in its ability to describe HEDP systems. Numerical simulation has played a critical role in designing, predicting, and understanding HEDP systems from the very beginnings of the field, and will continue to do so for the foreseeable future.

This presentation will give an overview of the field of HEDP, focusing on the types of experimental data characteristic of HEDP experiments, and on the overall challenge of bringing simulation codes into intimate and fruitful contact with reality. Using examples drawn from our experience at Z, I will try to point out along the way the many possible ways in which scientists, statisticians, and data analysts could collaborate to their mutual benefit and the benefit of their respective fields.

Biography
Charles W. Nakhleh has been managing the Inertial Confinement Fusion (ICF) Target Design Department in the Pulsed Power Sciences Center at Sandia National Laboratories since April 2010. He supervises theoretical design and analysis efforts for magnetically-driven ICF targets for the Z pulsed-power facility and indirect-drive experiments for the National Ignition Campaign (NIC). He joined Sandia National Laboratories in December 2007, focusing on the physics and design of ICF and radiation-effects targets. From 2005 to 2007, he was the Group Leader (acting) and Deputy Group Leader for the Thermonuclear Applications Group (X-2) of the Los Alamos National Laboratory. He spent nearly a decade before that as a staff member in X-2, serving as the Project Leader for the Quantification of Margins and Uncertainties (QMU) Tools and Methods project, on which he collaborated closely with the Statistical Sciences Group, and worked on a wide variety of weapons-physics and design issues. He is a graduate of the Theoretical Institute of Thermonuclear and Nuclear Studies (TITANS) program at Los Alamos. He was a member of study teams that received Department of Energy Awards of Excellence in 1999, 2000, 2005, 2007, and 2010. He has served on numerous advisory panels, including the NNSA’s Predictive Science Panel, a National Academy of Sciences (NAS) panel on validation and verification, the Los Alamos Director’s advisory panel on weapons certification, as a consultant to the 2009 JASON study on warhead Life Extension Programs, as an adviser to the Undersecretary of Energy on the NIC, and as an adviser to the NNSA on a variety of weapons physics issues. His research interests span a wide range of ICF radiation effects, and other applications of high-energy-density physics, and applications of Bayesian inference techniques. He received his PhD in Physics from Cornell University in 1996.
Adaptive Testing of Chemical and Biological Detectors | 10:30-11:05 a.m.
Shane Reese, Brigham Young University

Adaptive approaches to experimental protocol provide substantial benefit over standard protocols, especially as the resources involved in the study are expensive or otherwise limited in their availability. The use of adaptive experimental design approaches has not been used in testing emergent chemical and biological detection technologies. We present a Bayesian adaptive design approach to allocation of a test budget for chemical and biological detectors. We illustrate the approach with two examples: one from chemical threat detection and one from biological threat detection. We employ a probit based Gaussian process model as the primary analysis tool for modeling probability of detection. A comparison to standard, fixed experimental design protocols demonstrates that the adaptive approach yields substantial savings in test assets and test time, with statistical power comparable to the fixed design. We propose some opportunities for future research.

Validation of HADES MicroCT Radiograph Prediction | 11:05-11:40 a.m.
Kristin P. Lennox, Lawrence Livermore National Laboratory

A key area in modern homeland security research is the detection of contraband in cargo and luggage. X-ray imaging allows operators to screen large numbers of containers for suspicious substances quickly, given that the detection signature of the material is known. However, determining large numbers of these signatures in a cost-effective manner is difficult.

HADES is a radiographic simulation code that can be used to simulate the output of a micro-computed tomography (micro-CT) system in a fraction of the time required for experimental measurements, and without requiring the production or handling of dangerous compounds. In order for HADES to be used to supplement experimental micro-CT results, the quality of HADES predictions must be quantitatively evaluated. Potential sources of systematic errors, such as uncertainty in sample composition and experimental geometry, must be addressed in addition to characterizing the repeatability of the micro-CT experiments themselves. This talk provides a case study of some of the challenges inherent in the validation of computer models with real data, and of how such validation can lead to a better understanding of, and ultimately improvements to, the simulation system.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. This is joint work with Haiyin Chen, Maurice B. Aufderheide III, W. Travis White III, G. Patrick Roberson, Jerel Smith and Lee G. Glascoe.

Data Analysis for Homeland Security Applications: Bio-Inspired Approaches to Optimization and Signature Discovery | 11:40 a.m.-12:15 p.m.
Christopher Oehmen, Pacific Northwest National Laboratory

Similar to many scientific disciplines, security applications are undergoing a revolution in abundance and complexity of data. Often this presents a significant barrier to maximizing the value of the data that is being collected. Biology affords many models for dealing with incomplete, complex, distributed data. Some examples include techniques used for genome analysis, coordinated effort of social insects for problem solving, and bio-inspired data analysis for real-time defensive awareness in cyber systems. Genome analysis is a way to find patterns in letter sequences that is robust to changes in order and the introduction of mismatches or nonsense. Social insects collect local data over a distributed domain and acting on this data (without centralized control) are capable of coordinating to produce astoundingly robust and complex emergent behaviors. Similarly, biosystems are capable of defending themselves and continuing operations in the presence of constant attack and partial compromise—without having complete a priori knowledge of how attacks will occur. This resulting resilience can serve as a model for cybersystems in general.

In this presentation I will present some recent work at PNNL in all of these areas and show how bio-inspired approaches are poised to enhance traditional data analysis applications in national security and provide asymmetric advantage to defenders.
Abstract
Large-scale synthesis and modeling studies are often conducted by science teams that are geographically distributed and on datasets that are heterogeneous and global in scale. A broad array of collaboration and data analytics tools are now available that could potentially support these science teams. However, building tools that scientists actually find useful is hard. Also, moving scientists from an informal collaboration structure to one mediated by technology often exposes inconsistencies in the understanding of the rules of engagement between collaborators. Collaboration portals such as fluxdata.org support the community of scientists contributing and analyzing the global FLUXNET carbon-flux synthesis dataset and the National Soil Carbon Network. Portals under development include the ASCEM Data System. Through long-term partnership with scientists, we have the opportunity to develop data management systems/portals that not only support collaboration and synthesis efforts but also support the data over its lifecycle by enabling data curation, versioning, and metadata collection.

Biography
Deb Agarwal is a Senior Staff Scientist at the Lawrence Berkeley Laboratory and head of the Advanced Computing for Science Department. She is leading several teams developing cyber infrastructure to support scientific research. Her current projects are developing a data server infrastructure to enhance data browsing and analysis capabilities for eco-science and new computational modeling environments for environmental management and carbon capture at power plants. Dr. Agarwal holds a Ph.D. in electrical and computer engineering from University of California, Santa Barbara.
Estimating a Thinning Ratio for a Markov Chain of Graphs | 3-3:35 p.m.
Jaideep Ray, Sandia National Laboratories

Markov chain methods, in conjunction with a graph generation model or procedure, can be used to construct a sequence of graphs where a given graph metric e.g., degree distribution, is preserved. However, successive graphs in the sequence are highly correlated, and the Markov chains have to be run long enough to sample the distribution sufficiently. Being able to save a thinned version of the chain, where every kth realization is preserved, has obvious attractions in terms of convenience of use. However the thinning factor k must be so chosen that the thinned chain honors the distribution of graph properties e.g. clustering coefficients etc, that existed in the original chain.

We present a method to identify a suitable k. It is demonstrated on a Markov chain that generates a sequence of graphs, with labeled nodes, with a prescribed joint degree-distribution. The method is applied to the binary “time-series” indicating the existence/non-existence of an edge between a given pair of nodes. This sequence is recursively thinned, by factors of two, till it resembles independent draws from a biased coin. The resultant thinning factor is k. Note that this does not necessarily mean that the thinned chain leads to an independent set of graphs; rather, the chain can be thinned by k, and saved, without destroying the distribution of graph metrics in the original chain. We test this method empirically on a collection of graphs, where the Markov chain was previously demonstrated to converge i.e., the edge autocorrelation is seen to decay with increasing lag. We find that k calculated by our method is conservative i.e., the distribution of graph metrics obtained by our thinning ratio is similar to those obtained by thinning the original chain by factors which may be 4 – 12 times larger.

Uncertainty Quantification for Networks with Power Distribution Applications | 3:35-4:10 p.m.
Scott Vander Wiel, Los Alamos National Laboratory

Network-delivered commodities are important in national and global security, as well as everyday life. Unfortunately, the networks themselves are complex, vulnerable, unpredictable, and only partially observable. Despite lacking complete knowledge, analysts must make decisions about network operations, resiliency, and criticality. Currently these decisions are based on point estimates of the complete network based on expert judgment or basic state estimation techniques. We discuss the challenge of uncertainty quantification (UQ) for networks with partial observation of the topology, attributes, and flows. Specifically, we are developing methodology for inference about unobserved parts of a network, such as links, nodes, attributes, and commodity flows, using science-based simulation tools in conjunction with the observed network components. Examples from power distribution networks include missing link estimation, optimal dispatch with state uncertainty, criticality assessment and assignment of resources following a disaster when the network state is poorly characterized. This is joint work with Russell Bent and Earl Lawrence.

An Illustration of Bayesian Nonparametric Models for Community Detection | 4:10-4:45 p.m.
Dan Nordman, Iowa State University

In this talk, we propose a series of statistical models for community detection, through a case study analysis of a well-known network data set (consisting of football games played among various NCAA conferences in one season). We start with an initial, simple model and proposed subsequent models (increasing in sophistication) to correct deficiencies in previous models. Under these models, nodes (teams) may or may not fall into community structures, and the probabilities of edges forming between nodes within a community or between nodes in different communities can potentially vary with the community or pairs of communities in the network. Estimation in these models is performed through Markov Chain Monte Carlo-based Bayes analysis, and community detection is carried out through a decision theory approach based on posterior samples. The Bayes approach does not require the number of communities (or nodes not belonging to any communities) to be specified in advance, and these features are estimated from the data. For the motivating data set, we also discuss assessing the goodness-of-fit between proposed models and the data based on model checks with deviance information criterion or posterior predictive distributions of graph features.
6-8 p.m. Banquet, La Terraza at La Fonda, 100 E. San Francisco St.

Graham Walsh, New Mexico Institute of Mining and Technology
Blowing Stuff Up for Science:
Explosives Experiments with the MythBusters

Abstract
Multiple experiments (though not in the statistical sense) were performed by the Energetic Materials Research and Testing Center for the popular television show MythBusters. In the course of these shenanigans, EMRTC has made diamonds, bifurcated a car, pancaked a car, shot an RPG with a bullet, blown up a car, performed a 300 mph car crash and thrown a perfectly good (well, at least a functional) limousine off of a 200-foot cliff. In so doing, we have set many MythBuster milestones, the largest explosion, only bifurcation of a car, fastest car crash and only speechless moment for Adam in MythBusters history. This presentation will detail some of the science and engineering done behind the scenes for each stunt, and will, in the ultimate example of preaching to the choir, point out some of the shortcomings in declaring a myth to be busted or not busted based on a single data point.

Biography
Graham Walsh works at the Energetic Materials Research and Testing Center (EMRTC), an explosive research division of the New Mexico Institute of Mining and Technology (NMT). He has worked in the explosives field for 12 years, doing research projects on explosive pitting as a forensic indicator, microstructural investigations of shocked metals, reactive metal systems, pyrotechnics, high temperature explosives, shaped charges and explosively formed projectiles, etc. Graham is also certified by the Department of Transportation to test and evaluate explosive materials and articles and recommend hazard classifications. Dr Walsh produces all of EMRTC’s improvised explosives (yes, at the time this bio was written he had all his fingers), and has been active in improvised explosive device (IED) and homemade explosive (HME) training. Graham is an adjunct professor in the Department of Mechanical Engineering at NMT, where he teaches various classes on explosive effects and theory. All of this is a fancy way to say that he lights things on fire for a living. Lately, when not marooned in meetings, Graham has made several valuable contributions to academia and the scientific community by lending his explosive experience to such future Nobel laureates as television’s MythBusters, BBC’s Richard Hammond and other television productions. Graham used Duncan’s multiple range test about a decade ago, and occasionally says things like “standard deviation” when he wants to sound smart in front of a class. He is in no way a statistician.
Friday, March 2, 2012

8:45-10:30 a.m. | Reliability for Complex Systems
Organizer: Shane Reese, Brigham Young University

- J. Hathaway, Pacific Northwest National Laboratory
  The Transition to LED Lighting: Modeling Lumen Degradation as a Measure of Reliability and Quality
- Alyson Wilson, IDA Science and Technology Policy Institute
  Bayesian Methods for Estimating System Reliability Using Heterogeneous Multilevel Information
- Christine Anderson-Cook, Los Alamos National Laboratory
  Strategic Future Data Collection for Complex Systems Based on Multiple Reliability Objectives

10:30-10:45 a.m. | Break

10:45 a.m.-12:30 p.m. | Highlights of Department of Energy Collaborations with Universities
Organizer: Derek Bingham, Simon Fraser University

- James Paul Holloway, University of Michigan
  Methodologies for Assessing Uncertainties in Radiative Shock Prediction
- Matt Taddy, University of Chicago
  Optimal Design of Text Mining Experiments
- Tan Bui, University of Texas at Austin
  Scalable Methodologies for Large-Scale Statistical Inverse Problems with Uncertain Data

12:30-1:45 p.m. | Lunch (on your own)

1:45-3:30 p.m. | Video Analysis and Satellite Imagery
Organizer: Paul Kidwell, Lawrence Livermore National Laboratory

- Angela Mielke, Los Alamos National Laboratory
  Genie, Etc.
- David Messinger, Rochester Institute of Technology
  Graph Theory Approaches to Spectral Image Analysis
- Paul Kidwell, Lawrence Livermore National Laboratory
  A Context Based Approach to Modeling Vehicle Trajectories in Wide-Area Motion Imagery

3:30-3:45 p.m. | Break

3:45-4:55 p.m. | Computational Strategies for Visualization
Organizer: Earl Lawrence, Los Alamos National Laboratory

- Jon Woodring, Los Alamos National Laboratory
  Extending In-Situ Analysis through Feature and Event Detectors
- Ryan Hafen, Pacific Northwest National Laboratory
  Exploratory Data Analysis and Statistical Model Building with Large and Complex Data

4:55-5 p.m. | Closing Remarks
The Transition to LED Lighting: Modeling Lumen Degradation as a Measure of Reliability and Quality | 8:45-9:20 a.m.

J. Hathaway, Pacific Northwest National Laboratory

Light emitting diode (LED) technology has evolved significantly since the late 1960s and has become a viable option for lighting buildings and outdoor areas. Development over the last several years has introduced LED lamps that can replace the traditional 60-watt incandescent lamp. The Building Technologies Program within the U.S. Department of Energy (DOE) has been a key leader in pushing LED lamp research and development at DOE laboratories and in industry.

On behalf of the DOE, Pacific Northwest National Laboratory (PNNL) has managed the Bright Tomorrow Lighting Prize (L Prize) competition and participated on industry standards committees developing the testing procedures for solid-state lighting. As part of the L Prize, PNNL developed and implemented a long-term photometric test to evaluate the lumen maintenance and color sustainability of LED replacement lamps (field studies and stress testing were performed by other PNNL partners).

In this presentation we will provide background on LEDs, the development of current lighting standards and test methods, and the L Prize competition. Details of the statistical applications implemented to handle the unique characteristics involved in defining a methodology to test LED lamp reliability will be presented. Specifically, we will highlight the statistics that went into newly developed procedures for testing and estimating the lifetime of LED lamps, which can have very long lives (i.e., 3+ years of continuous use), making full life testing impractical. In addition, we will discuss the challenges of applying statistical methods within the constraints of industry standards and/or legislated mandates. This is joint work with T. Pulsipher, K. Anderson, R. Hafen, J. McCullough, E. Richman, G. Williams, and M. Ledbetter.


Alyson Wilson, IDA Science and Technology Policy Institute

We present a Bayesian approach for assessing the reliability of multicomponent systems. A novel feature of this model is the natural manner in which degradation, lifetime, binary, or elicited data collected over time at the component, subsystem, or system level are simultaneously integrated into the assessment. The model allows pooling of information between similar components, the incorporation of expert opinion, and straightforward handling of censored data. The methodology is motivated (and illustrated) with an example. Joint work with Jiqiang Guo.

Strategic Future Data Collection for Complex Systems Based on Multiple Reliability Objectives | 9:55-10:30 a.m.

Christine Anderson-Cook, Los Alamos National Laboratory

The talk presents an example of multiple criterion optimization in the context of sequential experimentation. When estimating the reliability of a complex system comprised of many components, there are often different types of data which might be collected at the system, sub-system and component levels. We consider an example where the goal is to better estimate the reliability of a system and two of its major sub-systems. After initial analysis has been performed, there is an opportunity to collect more data. The goal is to select new data which maximally improves the quality of the estimation of the system and sub-system reliabilities. The talk presents a process with accompanying graphical tools based on the Pareto front approach to multiple criterion optimization, which allows some possible collections of new data to be eliminated as clearly inferior. From the remaining allocations, the best set of new data can be identified based on the relative importance of the different criteria, and the anticipated improvement in the quality of prediction quantified. The methodology is widely applicable to different problem scenarios where several competing goals are considered simultaneously. This is joint work with Lu Lu and Jessica Chapman.
Methodologies for Assessing Uncertainties in Radiative Shock Prediction | 10:45-11:20 a.m.
James Paul Holloway, University of Michigan

The CRASH center is exploring methods for assessing and understanding the uncertainties in predictions of radiative shocks. In our system physics is dominated by the interaction among the radiative shock, an ablation-driven wall shock, and a material interface. This physics is relevant to astrophysics and fundamental high-energy-density physics research. Through several series of 1D and 2D computer runs and experimental campaigns we have collected data on various quantities of interest including shock location, wall shock location, and measures of the radial distribution of dense Xe. The different simulations have some physics parameters in common and some physics parameters that exist only in 1D or 2D. We have developed methodologies that can jointly calibrate these different fidelity of simulators. This requires two discrepancies: one the discrepancy between the 1D and 2D simulations, and one the discrepancy between the 2D simulation and field measurements. The discrepancy between simulation models depends on those physics parameters that are unique to the 2D model and the common physics parameters, but not on those that are unique to the 1D model. We have also explored the process of extrapolating models into new regions of input space where there are simulation runs but no field measurements. A major benefit of the work has actually been to stress both the CRASH code and the field experiments in ways that have led to better understanding of uncertainties in measurements and improvements in the simulation capability.

This research was supported by the DOE NNSA/ASC under the Predictive Science Academic Alliance Program by grant number DEFC52-08NA28616.

Optimal Design of Text Mining Experiments | 11:20-11:55 a.m.
Matt Taddy, University of Chicago

Sequential design of experiments is an interest area for many at DOE labs. Previous work with Sandia Labs looked at this problem in the context of robust search optimization, wherein we were tasked with augmenting a local pattern search algorithm with search locations that had a high probability of improvement. In particular, parallel execution required optimal sets of multiple new search points rather than a single 'best' new location. This talk will re-visit these ideas, and look to problems in text-sentiment analysis as a new application area. Here, the goal is to predict variables that motivated language use (e.g., the author's political beliefs in a news article). Typically, huge amounts of text are available, but obtaining sentiment-scored text samples for model training is very expensive. Hence we discuss methods for choosing optimal sub-samples from the available conversation. The technology will be illustrated in scoring of various sentiment indices for text data from the streaming twitter feed.

Scalable Methodologies for Large-Scale Statistical Inverse Problems with Uncertain Data | 11:55 a.m.-12:30 p.m.
Tan Bui, University of Texas at Austin

We address inverse problems with uncertain data in the framework of Bayesian inference. That is, given observational data and their uncertainty, the governing forward model (typically in the form of partial differential equations (PDEs)) and its uncertainty, and a prior probability density function (pdf) describing uncertainty in model parameters, infer the (distribution of) parameters that is consistent with the data, model, and prior. The solution using the Bayesian framework is the posterior probability distribution over parameter space. Often, the parameters represent continuous fields such as heterogeneous material properties, initial conditions, boundary conditions, and shape. In this case, the Bayesian solution is extremely challenging, since the (suitably-discretized) posterior is very high dimensional, and evaluating it at any point in parameter space requires a solution of the forward PDE model. Sampling the posterior by what is often the method of choice, MCMC, quickly becomes prohibitive, since even in modest dimensions, the number of samples required can be in the thousands or millions.

In this talk, we discuss scalable methods using high order derivative information (gradients and Hessians of the negative log posterior) and infinite dimensional structure of the underlying physical problem to address the challenges of large-scale Bayesian inversion. Numerical examples are given for shape inverse electromagnetic scattering and seismic inversion.
Video Analysis and Satellite Imagery

Organizer: Paul Kidwell, Lawrence Livermore National Laboratory
1:45-3:30 p.m.

Genie, Etc. | 1:45-2:20 p.m.
Angela Mielke, Los Alamos National Laboratory

This discussion presents a survey of work done by the Machine Learning team in the Intelligence and Space Research (ISR) Division at Los Alamos National Laboratory. This survey begins with a review of Genie, one of the team’s first and most visible projects. Analysts interact with Genie by painting over a remotely sensed image to identify features of interest, from roads and crops to municipal golf courses. Genie ‘learns’ algorithms that are able to find these features automatically. Post-Genie projects include polygonization of pixels, segmentation of imagery, rectilinear structure identification to aid broad searches for man-made activities, 3-D structure from multiple views, anomalous change detection and sparse modeling for video analysis. A brief discussion highlights the applicability of these techniques to broader systems employing video surveillance and distributed sensing systems for high level knowledge extraction.

Graph Theory Approaches to Spectral Image Analysis | 2:20-2:55 p.m.
David Messinger, Rochester Institute of Technology

Automated extraction of quantitative information from remotely sensed multi- & hyperspectral imagery requires one to have a mathematical model for the "background" and "foreground" signatures in the image. Given these data models, one can then make decisions per-pixel as to the likelihood of the presence of a signature of interest or perform other tasks such as spectral clustering. Traditional processing schemes rely heavily on algorithms that use simple first and second order statistics or linear subspace models of the data to make these decisions and have been successful in several applications such as sub-pixel target detection and anomaly detection. However, the new generation of sensors has a significant improvement in spatial and spectral coverage and resolution: the DigitalGlobe Worldview-2 sensor has spatial resolution of ~2m and has 8 spectral bands in the visible - near infrared part of the spectrum, while commercially available hyperspectral imagery has similar spatial resolution with ~200 contiguous, narrow spectral channels in the visible - short wave infrared. At these spatial and spectral resolutions, the sensors image the surface of the Earth at ever-greater levels of detail. It is simple to show that assumptions of multivariate normality, or that the data are well-defined by linear subspaces, are not well-met by the current generation of sensors. This presentation will describe new algorithms that leverage the mathematics of graph theory for applications such as spectral clustering, anomaly detection, data dimensionality estimation, and change detection. Specific applications require different characteristics of the graph, and various methods of graph construction will be presented. Additionally, several methods will be presented to extract quantitative information from the graph itself. Results will be shown for space-based multispectral imagery as well as airborne hyperspectral imagery.

A Context Based Approach to Modeling Vehicle Trajectories in Wide-Area Motion Imagery | 2:55-3:30 p.m.
Paul Kidwell, Lawrence Livermore National Laboratory

Wide area overhead imagery has proliferated in recent years necessitating the development of automated procedures for analyzing this data. Of particular interest are multi-object tracking algorithms which can persistently track vehicles from origin to destination for near real time or forensic analysis. In this work we propose a novel approach to modeling vehicle trajectories leveraging previously under utilized information.

More precisely, the output of a tracking algorithm is used to build context via a set of models describing the scene. While an ideal tracking algorithm would simultaneously model every individual track, in a practical sense this accomplished through shared models. An optimal set of vehicle tracks is arrived at by considering the generative model and using a set of data association moves as the basis for a greedy optimization. Experiments demonstrate that the track length is increased, while id switches and false alarms are decreased.
Computational Strategies for Visualization

Organizer: Earl Lawrence, Los Alamos National Laboratory
3:45-4:55 p.m.

Extending In Situ Analysis through Feature and Event Detectors | 3:45-4:20 p.m.
Jon Woodring, Los Alamos National Laboratory

The massive amounts of data that are generated by large-scale simulations and supercomputers have exceeded our technological capacity to store it and our cognitive capacity to understand it. A solution to manage the data deluge is in situ visualization and analysis, which stores smaller processed data by integrating analysis and visualization with simulations. Since most large-scale simulations are batch processes without a human-in-the-loop, in situ analysis can miss discoveries achieved through interactive data exploration. To try to fill this discovery role, we extend in situ analysis by adding feature and event detection algorithms to trigger visualization and analysis data pipelines. Triggering data processing pipelines for large-scale simulations is conceptually similar to high-energy physics analysis codes searching for events in accelerator experiments. We will discuss our previous work in feature and detection algorithms to support event-driven in situ analysis, various use-case scenarios, and our on-going in situ analysis with DOE simulation codes.

Exploratory Data Analysis and Statistical Model Building with Large and Complex Data | 4:20-4:55 p.m.
Ryan Hafen, Pacific Northwest National Laboratory

The marked growth of data size and complexity presents compelling challenges for the statistics community that threaten to severely impede the throughput of critical analysis efforts. To address these challenges, statisticians need scalable, practical tools that will efficiently extend statistical analysis to large data. The goal is to be able to apply exploratory data analysis (EDA) and statistical model building techniques to large data in a straightforward way. In EDA, we look for trends and patterns in the data as well as deviations from those expected, relying heavily on the use of visualization, in support of ultimately building statistical models that characterize the relationships and variability in the data. Simply resorting to summary statistics or preconceived, unvalidated models in this process is certain to lead to missed information or incorrect results. The statistician needs to be able to look at large amounts of data in great detail, combining the computational power of the machine with the reasoning power of the human in an iterative process of model development and refinement, regardless of the size of the data.

This talk presents emerging tools for scalable statistical analysis. The approach centers around the idea of "divide and recombine" (D&R), where analysis routines are applied in parallel to subsets of the data and recombined in a meaningful way. D&R is achieved through a map-reduce framework, linking the R statistical software environment to the Hadoop distributed computing system. R allows for rapid prototyping of models and methods in a language familiar to statisticians, and Hadoop takes care of the data management, fault-tolerance, and scalability. There are several considerations in this approach such as how to optimally subset the data and how to visualize the data throughout the analysis process. We will examine aspects of this approach and how it is being applied to solve problems related to the power grid and nuclear forensics.
1. **Optimal Linear Empirical Bayes Estimator of Finite Population Mean**  
   Mohanad F. Al-khasawneh, Qatar University

2. **The Commute Time Distance Transformation for Spectral Image Data and Application to Anomaly Detection**  
   Jamie Albano, Rochester Institute of Technology

3. **Small is Green Energy: A Biostatistician Compares (his) Rooftop PV in New Mexico to the Yankee Rowe Nuclear Power Station**  
   Hubert A. Allen, Jr., Hubert Allen and Associates

4. **Adaptive Correlation Filter for Video Action Detection and Classification**  
   Robert T. Arn, Colorado State University

5. **Bayesian Nonparametric Methods for Solids Analysis of Long-wave Hyperspectral Image Data**  
   Candace Berrett, Brigham Young University

6. **Scalable Bayesian Nonparametric Methods for Community Detection**  
   Jonathan Berry, Sandia National Laboratories

7. **Uncertainty Quantification for Carbon Capture Simulation**  
   Sham Bhat, Los Alamos National Laboratory

8. **Practical Data Analysis in Two Renewable Energy Applications**  
   David Biagioni, University of Colorado

9. **Neutron Multiplicity Counting in Non Proliferation Applications**  
   Tom Burr, Los Alamos National Laboratory

10. **Fire Modeling of a Glovebox Fire**  
    Leigh J. Cash, Los Alamos National Laboratory

11. **Random Graphs with Latent Spatial Structure**  
    Emily Casleton, Iowa State University

12. **Applying Model-Based Clustering for Analysis of Community Atmospheric Model Output**  
    Wei-Chen Chen, Oak Ridge National Laboratory

13. **Algorithms and Applications of Sparse Support Vector Machines**  
    Sofya Chepushtanova, Colorado State University

14. **An Information-Theoretic and Algebraic Approach to Financial Reconciliation**  
    Peter Chew, Galisteo Consulting Group, Inc.

15. **Multistate Stochastic Process Modeling of Nuclear Power Plant Reliability**  
    David Collins, Los Alamos National Laboratory

16. **MapReduce-Enabled Model Reduction for Large Scale Simulation Data**  
    Paul Constantine, Stanford University
17. **Are Megaquakes Clustered?**  
   Eric G. Daub, Los Alamos National Laboratory

18. **A Predictive Model for Geographic Statistical Data**  
   Jorge Diaz-Castro, University of Puerto Rico

19. **Image Registration for Change Detection Using Aerial and Satellite Imagery**  
   Matt Fair, Lakshman Prasad, Los Alamos National Laboratory

20. **Scalable Methods for Large-Scale Statistical Inverse Problems, with Applications to Subsurface Flow**  
    Pearl Flath, The University of Texas at Austin

21. **Functional Data Analysis in Cosmic Microwave Background Radiation Modeling**  
    Devin Francom, Brigham Young University

22. **Methods for Characterizing and Comparing Shock Wave Curves**  
    Michael Fugate, Curtis Storlie, Los Alamos National Laboratory

23. **Statistical Electromagnetics for Radio Frequency Remote Sensing**  
    John Galbraith, Los Alamos National Laboratory

24. **Feature Extraction Metrics for Quantitative Comparison Between Simulations and Experimental Results**  
    Michael Grosskopf, University of Michigan

25. **Monte Carlo Analysis for Repair/Restoration in Earthquake Scenarios**  
    Christina Hamada, University of New Mexico Los Alamos/Los Alamos National Laboratory

26. **Estimating a Proportion from Repeated Sampling of a Growing Population**  
    Michael Hamada, Los Alamos National Laboratory

27. **A Statistical Model for the Study of U-Nb Aging**  
    Geralyn M. Hemphill, Los Alamos National Laboratory

28. **Statistical Comparison of Toolmarks in Forensics**  
    Amy Hoeksema, Iowa State University

29. **Efficient Estimators for Sequential and Resolution-Limited Inverse Problems**  
    Darren Homrighausen, Carnegie Mellon University

30. **Sensitivity of the Community Land Model (CLM4.0) to Key Modeling Parameters and Modeling of Key Physical Processes with Focus on the Arctic Environment**  
    Elena Kalinina, Sandia National Laboratories

31. **Evaluating the Likelihood of a Fire-Following-An-Earthquake Scenario**  
    Elizabeth J. Kelly, Los Alamos National Laboratory

32. **Uncertainty Quantification of 3D Geophysical Models**  
    Carène Larmat, Monica Maceira, Los Alamos National Laboratory

33. **Exploring Large Scale Scientific Data Using Information Theory**  
    Teng-Yok Lee, The Ohio State University
34. **Applications of Wavelet Analysis to Solution Monitoring for Nuclear Safeguards**  
Claire Longo, University of New Mexico

35. **A Data Analysis Strategy for Prediction of High Dimensional Chaotic Systems**  
Michael LuValle, OFS Labs

36. **Arctic Land Cover Classification using Multispectral Imagery with Adaptive Sparse Representations**  
Daniela I. Moody, Los Alamos National Laboratory/University of Maryland College Park

37. **Comparing Mitigations Under Uncertainty in Simulated Influenza Outbreak**  
Leslie M. Moore, Los Alamos National Laboratory

38. **Scan Statistics for the Online Detection of Locally Anomalous Subgraphs**  
Josh Neil, Los Alamos National Laboratory

39. **Computational and Statistical Challenges in Cosmology**  
Adrian Pope, Argonne National Laboratory

40. **Inverse Methods for Dimensionality Reduction**  
Juan Ramirez, Jr., University of Colorado at Boulder

41. **Assessing Network Partitioning**  
David Robinson, Sandia National Laboratories

42. **Indago: A Novel Search and Analysis Tool for Electronically Stored or Transmitted Content**  
Jorge H. Román, Los Alamos National Laboratory

43. **Rhipe: R and Hadoop Integrated Programming Environment**  
Jeremiah Rounds, Purdue University

44. **Analysis of Biothreat Detection Technologies**  
Ryan Roundy, Brigham Young University

45. **Scalable Analysis of Network Measurements with Hadoop and Pig**  
Taghrid Samak, Lawrence Berkeley National Laboratory

46. **Topologies of the Conditional Ancestral Trees and Full Likelihood-based Inference in the General Coalescent Tree Framework**  
Ori Sargsyan, Los Alamos National Laboratory

47. **Estimating the Central Direction of Random Rotations in SO(3)**  
Bryan Stanfill, Iowa State University

Jon Stearley, Sandia National Laboratories

49. **Scientist Attrition at LANL: A Look over Time, Using Survival Analysis Methods**  
Blair O. Stephenson, Los Alamos National Laboratory

50. **Visualization at the Mesoscale**  
Richard Strelitz, Los Alamos National Laboratory
51. Spatio-Spectral Anomalous Change Detection in Hyperspectral Imagery  
James Theiler, Los Alamos National Laboratory

52. Uncertainty Analysis for GIS Estimates of CO2 Storage of the Oriskany Sandstone  
Andrew C. Thomas, Carnegie Mellon University

53. Analysis of Amplified Fragment Length Polymorphisms for Detection of Infectious Biological Agents  
Lawrence Ticknor, Los Alamos National Laboratory

54. Automated Analysis of Large Numbers of Diffraction Datasets Using the Rietveld Method  
Sven C. Vogel, Los Alamos National Laboratory

55. Derivation of QMU-like formalisms from Maximum Entropy Arguments  
Timothy Wallstrom, Los Alamos National Laboratory

Brian Weaver, Los Alamos National Laboratory

57. Neuromorphic Algorithms for Classifying and Locating Objects in a Fixed Camera Video Feed  
Dylan Paiton, Los Alamos National Laboratory

58. Sparse Principal Subspace Estimation  
Vincent Vu, Carnegie Mellon University

59. In-situ Exploration and Sampling of Computer Simulation Models  
Joanne Wendelberger, Los Alamos National Laboratory

60. Data Challenges for Scientific Data Analysis and Synthesis: An Example  
Arthur Wiedmer, UC Berkeley

61. Automatic Detection of PV System Configuration  
Matthew Williams, Shawn Kerrigan, Locus Energy

62. Using the Ensemble Kalman Filter for Computer Model Calibration  
Matt Pratola, Los Alamos National Laboratory

63. cranvas: A New Package for Interactive Statistical Graphics in R  
Yihui Xie, Iowa State University

64. Rank Approach in Assessment of Health Conditions among DOE Security Workers Using Data from DOE Illness and Injury Surveillance Program  
Joey Zhou, Office of Health, Safety and Security, Department of Energy

65. Negative Log-Gamma Modeling of Series System Reliability with Trend in Presence of Few Highly Reliable Components  
Roger Zoh, Iowa State University
1. Optimal Linear Empirical Bayes Estimator of Finite Population Mean  
Mohanad F. Al-khasawneh, Qatar University

The Bayes sequential estimation problem is to find an optimal sequential procedure which includes an optimal stopping rule as well as Bayes estimate. When the prior distribution is not completely specified, empirical Bayes procedures have to be used. In this article we consider the Bayes and empirical Bayes problem of estimating the population mean of a finite population when sample data are available from other similar finite populations. We investigate a general class of linear estimators and we obtain the optimal linear Bayes estimator of the finite population mean under a squared error loss. The optimal linear Bayes estimator and the sample size are obtained as a function of the parameters of the prior distribution. Then the unknown prior parameters are replaced by their estimates derived based on the historical data as well as the current data to get the linear empirical Bayes estimator. These estimates are updated at every stage of the sampling and a stopping time is utilized to stop sampling. A Monte Carlo simulation study is then carried out to investigate the performance of the proposed Empirical Bayes Estimator as compared to the usual Estimators.

2. The Commute Time Distance Transformation for Spectral Image Data and Application to Anomaly Detection  
Jamie Albano, Rochester Institute of Technology

Imaging spectrometer data can be represented as a three dimensional structure which encompasses both spatially and spectrally sampled data of a given scene. Analysis of this type of imagery often begins by applying a transformation that generates an alternative representation of the spectral data with the intention of exposing hidden features not discernable in the original space. The transformation introduced here is based on a Markov-chain model of a random walk on a graph and its utility in exposing anomalies in the spectral data is analyzed. To quantify the random walk, a measure known as the average commute time distance is used which is defined as the average length a random walker takes, when starting at one node, to transition to another and return to the starting node. The Commute Time Distance (CTD) transformation embeds the nodes of a graph in a Euclidean space such that their separation is equal to the square root of this random walk measure. This distance metric has the important characteristic of increasing when the number of paths between two nodes decreases and the lengths of those paths increase. When applied to spectral imagery, the first step is to construct a similarity graph on the data in the spectral domain. Therefore, the spectral vectors are represented as nodes of the graph while the edges connecting them depend upon the type of similarity graph applied. For our application, the similarity graph used is a combination of a mutual k-nearest neighbor graph and a minimum spanning tree graph. The prior graph is used to connect pixels of similar density while the latter graph ensures graph connectivity so that the random walker has a nonzero probability of transitioning to any node in the graph. This poster will present the similarity graph, the mathematical derivation of the CTD transformation and its application to anomaly detection in spectral imagery.

3. Small is Green Energy: A Biostatistician Compares (his) Rooftop PV in New Mexico to the Yankee Rowe Nuclear Power Station  
Hubert A. Allen, Jr., Hubert Allen and Associates

Real estate is a limited resource. Energy production invariably requires physical space. This analysis compares the energy production, maintenance and production costs, income and physical space required for the 185-megawatt Yankee Rowe Nuclear Power Station to the author’s own 1.72 kilowatt Rooftop PV system in Albuquerque, New Mexico. The two different cycles are presented. First, the lifetime stages of a complex nuclear power plant including construction, production years, plant decommissioning and fuel storage. Each step requires a dedicated area of real estate and this area is included in calculations of lifetime efficiency. Rooftop PV solar panels also require physical space but with a much simpler cycle. Results are presented showing the varying real estate requirements over the lifetime of the Yankee Rowe Nuclear Power Station and the Rooftop PV up to 2012. The lifetime unit of comparison proposed is the number of kilowatt hours produced per square foot, per year.
4. Adaptive Correlation Filter for Video Action Detection and Classification

Robert T. Arn, Colorado State University

Video action detection is one of the emerging challenges within computer vision. Here we will explore a Minimum Output Sum of Squared Error (MOSSE) filter for the purposes of action classification and detection. This adaptive correlation filter operates in real time and requires fewer training samples than other action detection approaches to obtain reliable results. This is joint work with David S. Bolme, Bruce A. Draper, and Michael Kirby.

5. Bayesian Nonparametric Methods for Solids Analysis of Long-wave Hyperspectral Image Data

Candace Berrett, Brigham Young University

With the continual advances in technology and data collection, the words “large”, “extended”, and “hyper” are being more frequently used to describe data. This is the case with remote sensing—now using tools such as hyperspectral cameras to collect data. Rather than a 2-D image recorded by a normal camera, a hyperspectral image is a large 3-D cube, containing observations over a wide range of spectral wavelengths. While the increase in the amount of data would hopefully provide an increase in information for identifying materials within the image, parsing through the large dataset can also make learning from the data more difficult. Ideally, the observed spectrum at each pixel of this image would exactly match the true spectral signal of the associated material; however, the observed spectrum is a mixture of many interacting signals and measurement error. The goal is to use this large, noisy and messy data to determine spectral and emissivity fingerprints for each material. One mechanism for doing this is to cluster the pixels by similar spectra, and determine the underlying signal within each cluster. The flexibility and feasibility of Bayesian nonparametric methods and the Dirichlet Process make them an ideal tool for doing this. Combining this stochastic process with the physical process of Plank’s Law, we use a large dataset containing 30 different solids, 258 wavelengths for each pixel, observed across a period of 3 months, to cluster and identify posterior distributions of spectra associated with each solid. This is joint work with William Baumann.

6. Scalable Bayesian Nonparametric Methods for Community Detection

Jonathan Berry, Sandia National Laboratories

Community detection algorithms are typically divisive or agglomerative, where the respective splits and/or merges are governed by some objective function such as modularity. At this conference, Dan Nordman will present Bayesian nonparametric alternative methods developed at Iowa State University. We extend the most basic of the Iowa State methods with two goals: (1) to model graphs generated by community detection benchmark generators such as Lancichinetti-Fortunato-Radicchi (LFR), and (2) to handle large instances in serial and in parallel. Our poster will be a progress report summarizing this ongoing work. This is joint work with Daniel Dunlavy and Cynthia Phillips.

7. Uncertainty Quantification for Carbon Capture Simulation

Sham Bhat, Los Alamos National Laboratory

The Carbon Capture Simulation Initiative (CCSI) is developing technology to accelerate the process of carbon capture from coal-fired power plants reliable and affordable by using simulation rather than physical testing for scaling up to commercial scale. Uncertainty quantification (UQ) capability is critical to simulation-based analysis of carbon capture systems due to the need to understand and manage complexity and economic impact of incorporation of carbon capture systems in current and future commercial operations. These uncertainty quantification tools include input sensitivity analysis, calibration of inputs, construction of surrogate models, and propagation of uncertainty. This talk will illustrate the use of a few UQ concepts in early stage development and evaluation of requirements for effective simulation of a solid sorbent process for carbon capture, using the NETL 32D model and thermogravimetric analysis (TGA) data from the National Energy Technology Laboratory (NETL). This is joint work with Joanne Wendelberger and Lisa Moore.
8. Practical Data Analysis in Two Renewable Energy Applications  
David Biagioni, University of Colorado

We present two practical applications of data analysis techniques applied to a) thin film photovoltaic (PV) reliability and b) biomass sugar release data. In the first application, a multiway regression analysis (N-way Partial Least Squares) is applied to capacitance-voltage (CV) data taken from several cadmium telluride solar cells subjected to stress over time. We identify a combination of factors that are highly correlated with degradation in efficiency of the cells. This study demonstrates the potential efficacy of multiway models in exploring CV data - to our knowledge a first in this field - and begins to address a central question in PV reliability: “can reliability of a module be predicted from pre- or early-stress measurements?”. In the second application, we use a combination of uncertainty quantification and sparse regression to identify compositional factors associated with biomass recalcitrance. We identify a handful of compositional signatures (specifically, mass-to-charge ratios in a mass spectrum) that correlate with total sugar release of treated samples in a switchgrass population. We also identify uncertainties in the experimental measurements that prohibit more complex models from being built in a meaningful way. This will allow us to specify “uncertainty bottlenecks” in the experimental pipeline that currently prevent a better quantitative understanding of the data.

9. Neutron Multiplicity Counting in Non Proliferation Applications  
Tom Burr, Los Alamos National Laboratory

In many nonproliferation applications it is important to use detected neutrons to assess whether an item is undergoing fission, which leads to neutron “chains.” A neutron chain is the starter neutron plus all the prompt (essentially instantaneous in our context) offspring neutrons arising from subsequent induced fission. The basis to distinguish other neutron sources from induced fission is that neutron chains should lead to strong temporal “bursts” of neutrons analogous to spikes of neurons in brain studies. However, most fielded neutron detectors detect slow neutrons while most fission-produced neutrons are born fast and must be slowed down prior to detection. The slowing down (“thermalization”) time for each neutron is random, well approximated by an exponential distribution, and complicates the main goal here of inferring the probability distribution function (pdf) characterizing the number of neutrons per chain.

This poster describes the context, the challenge of dealing with thermalization time, and provides a new approach to estimate the moments of the pdf characterizing the number of neutrons per chain. This is joint work with Nick Hengartner.

10. Fire Modeling of a Glovebox Fire  
Leigh J. Cash, Los Alamos National Laboratory

At the Los Alamos National Laboratory Plutonium Facility (TA 55) the spread of radiological contamination on surfaces, airborne contamination, and excursions of contaminants into the operator’s breathing zone are prevented through the use of a variety of gloveboxes (the glovebox, coupled with an adequate negative pressure gradient, provides primary confinement). A sustained fire in a glovebox can be large safety risk because of its potential to affect the glovebox’s containment integrity. Because experimental data does not currently exist on the response and survival time of gloveboxes (and glovebox components) used for plutonium process operations to conditions experienced in a facility fire, conservative assumptions are used to predict the consequences of a facility fire. These predicted consequences are in turn used as a basis for the design and selection of fire prevention and suppression features. Experimental data derived from exposure of an actual glovebox to a facility fire may support a reexamination of fire prevention and suppression requirements.

In collaboration with the University of Texas, computational fire modeling was used to accurately reproduce the experiments. In addition to typical glovebox fire loads and operating conditions, this problem requires computational models be used to predict system response to unknown atypical situations. To assist the modeling community, a general methodology for conducting simulations to unbounded high-dimensional problems was developed. One unique challenge associated with this problem is the degradation of materials within the system due to nuclear radiation absorption. A second challenge is developing model validation criteria from the experimental data so that the fire model can be used as a predictive tool for cases where a corresponding experimental data does not exist. In summary, a computational fire model has been developed to predict fire extinguisher activation time for a wide range of fire inputs and glovebox configurations. This project is important to the fire modeling community through the development of a validation suite of small compartment fire modeling in Fire Dynamics Simulator. A larger scientific impact would be that the methodology used for this problem could be applied to solve problems of similar structure in other fields. This is joint work with Michael E. Cournoyer and Ofodike Ezekoye.
11. Random Graphs with Latent Spatial Structure
Emily Casleton, Iowa State University

Many varieties of random graph generators and models have appeared in literature across a wide range of disciplines. A number of these generators utilize geography, or physical locations of nodes, to help determine the structure of a realized graph. These include the Watts-Strogatz, or small-world model, the Waxman model, the BRITE generator, and Yook’s method. These models incorporate geographic information that intuitively, but indirectly, affect the dependence structure of real world graphs (e.g. routers are easier to connect if they are closer to each other.) With respect to statistical modeling and analysis, Exponential Random Graph Models (ERGM) have been the most widely studied, particularly in the social sciences. The advantage these models provide is the ability to allow for complex inter-connections between edges. Through specification of a joint distribution in the form of a Gibbs distribution, a dependence structure is induced. The limitation of ERGMs is that, although they allow for dependence, they fail to model or explain it directly. The method proposed here combines the concepts of geography and an explicit specification of a dependence structure. Geographic information is incorporated through a latent spatial structure which determines Markov neighborhoods for edges that dictate conditional dependencies. A binary Markov random field (MRF) is then applied to the resulting configuration of edges. Under appropriate model restrictions, a joint distribution results from the conditional specification. This allows for explicit modeling and interpretation of the conditional dependence and independence between edges. Simulation of models from the proposed class exhibit features similar to those found in communication and social network data.

12. Applying Model-Based Clustering for Analysis of Community Atmospheric Model Output
Wei-Chen Chen, Oak Ridge National Laboratory

Modern Community atmosphere models (CAM5) run at high spatial and temporal resolution, effectively producing hundreds of terabytes of climate output. The size of these datasets impedes application of advanced statistical analysis such as model-based clustering, which can be used for extracting data features based on finite mixture models. We develop a parallel version of expectation and maximization (EM) algorithm to enable model-based clustering on such large datasets. We apply the technique to CAM5 output, isolate clusters corresponding to weather systems, and use a variety of visualization techniques to display the clusters and the correlations between selected climate variables. We perform a scalability analysis of the resulting code. The methods presented in this poster are broadly applicable for clustering This is joint work with George Ostrouchov, Dave Pugmire, Mr Prabhat, and Michael Wehner.

13. Algorithms and Applications of Sparse Support Vector Machines
Sofya Chepushtanova, Colorado State University

In this study we explore an L1-norm linear support vector machine (L1-norm SVM) for data classification and feature selection problems, and its applications to real-world data. The L1-norm SVM constructs an optimal separating hyperplane between two classes of data points with maximal margin measured by infinity-norm. The resulting optimization problem is in the form of a linear program which we solve using a primal dual interior point method. As the L1-norm suppresses lots of components of a weight vector w in the decision function \( f(x) = w'x+b \), the nonzero components indicate the features (spectral bands for hyperspectral data or biomarkers for medical data) that are effective at separating the data. We propose a statistical framework for the feature selection problem for multi-class data based on hypothesis testing. The behavior of the learning method is illustrated by modeling decision functions for classification of the Febry-Perot hyperspectral data set and sepsis medical data, with high accuracy rates achieved in both full and reduced feature spaces. This is joint work with Christopher Gittins and Michael Kirby.
14. An Information-Theoretic and Algebraic Approach to Financial Reconciliation

Peter Chew, Galisteo Consulting Group, Inc.

A basic financial audit function, reconciliation involves identifying correspondences between disparate datasets, and has applications in homeland security including the detection of fraud or illicit activity. Financial reconciliation is also an instance of a larger class of data-matching problems which the field of record linkage attempts to address. Reconciliation is a widespread challenge in the financial world, and for large datasets, is often approached using highly manual and time-consuming techniques. Though a seemingly run-of-the-mill problem, it in fact presents interesting challenges from a data analytics perspective. We present a highly general approach which we have developed (and implemented in production) which draws upon information theory and linear algebra, and which is inspired by techniques from natural language processing. When tested, a variant of the approach was found to have extremely high precision (agreeing with human judgment 99% of the time) and recall (reconciling over 75% of a sample dataset). Since the cost (in labor) of previously reconciling the same dataset was around half a million dollars – and the cost of developing our approach was just a fraction of that – clearly there are large rewards to be reaped from using this kind of approach in practice.

15. Multistate Stochastic Process Modeling of Nuclear Power Plant Reliability

David Collins, Los Alamos National Laboratory

Nuclear power plants (NPPs) play an important role in energy security and reduction of airborne pollutants. Given the slow rate of new plant construction, the U.S. Department of Energy’s reactor sustainability program is aimed at extending the life of existing NPPs beyond the currently licensed 60 years. This involves an extensive research agenda, investigating all aspects of NPP reliability and safety. We will present a general overview of NPP operation and issues related to collection and analysis of reliability data, along with our specific research.

We have investigated various types of multistate stochastic processes (MSPs) for predicting the reliability of NPP piping subsystems; this is an important area, since typical NPPs may have up to 40 miles of piping, much of it safety-critical (e.g., the main reactor coolant loop). We represent subsystems as statistical flowgraphs, with vertices representing states of partial or complete failure and edges representing probability distributions for transitions between states. Failure transitions are driven by processes based on material properties of the pipes, and the physical and chemical dynamics of the fluid being carried. Repair transitions are based on detection of leakage by visual inspection, or non-visible flaws by radiography or ultrasound. Given the complexity of the transition processes, we model the subsystems as semi-Markov processes, using the flowgraph framework to solve for quantities of interest such as the hazard rate for pipe rupture.

The poster will include examples of analysis at the level of individual state transitions as well as subsystems. We will also discuss a range of open issues, including incomplete data and the many uncertainties involved in system and subsystem-level models.

16. MapReduce-Enabled Model Reduction for Large Scale Simulation Data

Paul Constantine, Stanford University

Modern physics-based simulations produce a prodigious amount of data. Oftentimes, these simulations must be repeated at various input parameters to characterize model uncertainties or study sensitivities of model predictions. The resulting ensemble of results can require many terabytes of storage, rendering even simple queries and analyses challenging.

This poster will present our experiences using a Hadoop cluster to store the results of an ensemble of simulation results. We implement a thin singular value decomposition -- enabled by a communication-avoiding QR decomposition -- in the MapReduce framework for model reduction and interpolation. The MapReduce framework enables us to scale this method to datasets that do not fit into memory while providing automatic fault tolerance and parallelization. We apply this procedure to a terabyte-sized database of results from a large scale thermal simulation with varying material properties. We employ the resulting reduced order model to study the prediction of probability of loss of assured safety.
17. Are Megaquakes Clustered?
Eric G. Daub, Los Alamos National Laboratory

The last decade has seen an increase in the number of magnitude 8 and larger earthquakes worldwide, prompting speculation that large earthquakes cluster in time. We investigate the global earthquake catalog since 1900 to look for clustering in the largest earthquakes (magnitude M>7). We examine the catalog as a function of both minimum catalog magnitude M and time window T. Our results show that the earthquake record is consistent with a Poisson process, and hence are random in time. This result holds independent of whether aftershocks are included in the dataset except at minimum magnitudes below 7.3. Below M=7.3, the data is clustered when aftershocks are included in the catalog. We do find that even without aftershocks in the data, there is clustering for magnitude thresholds below 7.3 and time intervals of 2-5 years, though this clustering is due to a large number of events occurring in the early part of the century when magnitudes are less certain. Our results show that a transparent statistical approach can be used to investigate earthquake datasets to examine their behavior in time. This is joint work with E. Ben-Naim, R. A. Guyer, and P. A. Johnson.

18. A Predictive Model for Geographic Statistical Data
Jorge Diaz-Castro, University of Puerto Rico

Any planar map may be transformed into a graph. If we consider each country to be represented by a vertex (or node), if they are adjacent they will be joined by an edge. To consider how trends migrate across boundaries, we obtain relevant measures of the statistic we want to consider; namely, the index of prevalence, and the index of incidence. We define a cycle by a given unit of time, usually a year. We then propose various alternate equations whereby, by parametrizing various variables, such as population size, birth rate, death rate, and rate of immigration/emigration, we calculate a new index of prevalence/index of incidence, for the next cycle. For a given data set, each statistic we consider may propagate by a different equation, and/or a different set of parameters; this will be determined empirically. What we are proposing is, technically, to model how a discrete stochastic process propagates geographically, according to geographical proximity. Very often, statistics that depend on geographical proximity are tabulated by variables that are not; i.e., alphabetically. Such a predictive model would be relevant in areas such as public health; and/or crime mapping, for law enforcement purposes. We present an application using a GIS (geographic information system).

19. Image Registration for Change Detection Using Aerial and Satellite Imagery
Matt Fair, Lakshman Prasad, Los Alamos National Laboratory

The task of image registration is common first step in image analysis between more than one digital image. It is the process of bringing images into alignment between one another so that they can be compared to one another on a pixel-to-pixel basis. This poster walks people through the steps to attain a projective homography transformation, a common technique that produces a plane-to-plane mapping from one image to another. This commonly used technique works well for when the scene is planar, however in situations where the scene is not planar, misregistration for parts of the image where estimated plane does not match what is realistically in the scene, mainly caused by picture angle and terrain relief. In situations where the planar homography does not produce a satisfactory registration, a second stage registration is required. A discussion of several second stage registration methods that have been explored to achieve a better registration for change detection is done. Change detection is much more than a simple image subtraction technique between two registered image pairs. We explored change detection using the RADIUS algorithm, developed at Los Alamos National Laboratory, and registration effects on change detection.

20. Scalable Methods for Large-Scale Statistical Inverse Problems, with Applications to Subsurface Flow
Pearl Flath, The University of Texas at Austin

We address the challenge of large-scale nonlinear statistical inverse problems by developing an adaptive Hessian-based non-stationary Gaussian process response surface method to approximate the posterior pdf solution. We employ an adaptive sampling strategy for exploring the parameter space efficiently to find interpolation points and build a global analytical response surface far less expensive to evaluate than the original. The accuracy and efficiency of the response surface is demonstrated with examples, including a subsurface flow problem.
21. Functional Data Analysis in Cosmic Microwave Background Radiation Modeling  

Devin Francom, Brigham Young University

In this poster we present an analysis of functional outputs from computational models for cosmic microwave background radiation (CMBR) to assess cosmological evolution. These data are developed using computational models that require extensive tuning. We are interested in understanding what differences in CMBR are due to different types of tuning. Functional data analysis will allow us to consider continuous functions as we would usually consider disjoint data points. Thus, this type of analysis allows us to compare derivatives of these response functions. In order to create functions out of the data, we use cubic smoothing splines. When comparing functions, we are faced with variation in amplitude as well as altitude. We are interested in understanding both of these types of variation. In order to deal with amplitude variation, we warp the functions. We do this using landmarks, often local extrema, common to most of the functions. From the landmarks, we create warping functions that will warp CMBR so that landmarks occur at similar points on the time scale. We then identify clusters in the warped CMBR functions and the warping functions themselves to better understand CMBR over the early evolution of the universe.


Michael Fugate, Curtis Storlie, Los Alamos National Laboratory

At Los Alamos National Laboratory engineers conduct experiments to evaluate how well detonators and high explosives work. The experimental unit, often called an “onion skin”, is a solid hemisphere of high explosive materials with a detonator located directly under the pole at the equator of the hemisphere. When the detonator explodes a streak camera mounted above the pole records when the shock wave arrives at the surface. The output from the camera is a two-dimensional image that is transformed into curve that shows the arrival time as a function of polar angle.

The statistical challenge is to estimate the mean curve of a population of arrival time curves, and compare the mean of a baseline population to that resulting from a new population of onion skins. The engineering goal is to manufacture a new population of onion skins that generate arrival time curves with the same shape as the baseline. We present two statistical approaches which test for differences in mean curves, and provide simultaneous confidence bands for the difference: (i) a B-Spline basis approach, (ii) a Bayesian Gaussian process approach. In problems that involve complex modeling with modest sample sizes, it is important to apply multiple approaches with complementary strengths such as these to determine whether all approaches provide similar results. We compare the performances of the two approaches on several simulations that were constructed to mimic the actual analysis. Finally we present an analysis of historical onion skin data. This is joint work with Aparna Huzurbazar, Dave Higdon, Elizabeth Francois, and Douglas McHugh.


John Galbraith, Los Alamos National Laboratory

“Statistical electromagnetics” offers an experimental technique for studying electromagnetic coupling in complicated, real world environments. Maxwell’s equations are typically solved for (deterministic) electromagnetic quantities such as the electric field, but these computations require known “boundary conditions.” Many radio frequency (RF) remote sensing applications of interest to the DOE would operate in real, unconstrained environments, without any hope of modeling the boundary conditions accurately. Instead, we treat electromagnetic quantities as random values to mitigate our lack of detailed knowledge as we investigate difficult RF problems. Experimentally, we generate actual electromagnetic fields drawn from random distributions in a device called a “reverberation chamber”, and observe real circuit targets perturbed by the generated fields. Fields generated this way have rigorous statistical properties which we exploit. We have recently proposed a novel simulation approach to compliment the experiments, by sampling the boundary conditions using a Monte Carlo method. This is joint work with Max Light, Nick Hengartner, Dave Higdon, and Eric Raby.
24. Feature Extraction Metrics for Quantitative Comparison Between Simulations and Experimental Results
Mike Grosskopf, University of Michigan

In order to perform quantitative predictive analysis on simulations results of the radiative shock system, output metrics must be devised which can be meaningfully compared between the model and the experiment. The primary diagnostic for laboratory experiments on radiative shocks in xenon gas is x-ray radiography. For the radiographic data, we have explored output metrics related to the amount and distribution of dense xenon, including the shock location - as defined by a breakpoint in a piecewise constant fit to data integrated radially over a fixed window - and the total area and axial centroid of xenon which absorbs more than a set threshold of backlighter x-ray emission. Examples and results of the metrics are reported and discussed.

25. Monte Carlo Analysis for Repair/Restoration in Earthquake Scenarios
Christina Hamada, University of New Mexico Los Alamos/Los Alamos National Laboratory

Hazus is a software product developed by the Federal Emergency Management Administration (FEMA) for assessing the damage and consequences of floods, hurricanes and earthquakes. Hazus is frequently used to estimate damage to critical infrastructures in scenarios of interest. Hazus uses a collection of fragility curves to estimate the probability of different states of damage.

The probabilistic damage descriptions provided by Hazus for earthquake scenarios provide a challenge to interpret with respect to consequences. A structure, such as a roadway segment, is assigned a probability distribution over a finite set of damage states: None, Slight, Moderate, Extensive, or Complete. There are several methods to interpret the distribution: assign threshold levels, compute the “average” damage, or perform Monte Carlo analysis for an infrastructure system. We compare the average damage for roads to damage estimated from a Monte Carlo simulation. In addition, the Monte Carlo approach is explored for assessing the damage and subsequent repair and restoration times for highway transportation systems.

Michael Hamada, Los Alamos National Laboratory

We consider estimating the proportion of a growing population having a specified manufacturing anomaly from data obtained by repeatedly sampling the population as it grows. The observed proportion, while easily calculated, is biased. The Horvitz-Thompson (HT) estimator, commonly used in survey sampling, accounts for the unequal probabilities with which units are selected in this sampling design to produce an unbiased estimator of the population proportion having the anomaly.

We present the form of the HT estimator and an unbiased estimator of its variance that can be used to assess the uncertainty of an HT estimate. We consider an illustrative example to show the benefit using survey sampling theory and present how simulation might be used to estimate the inclusion probabilities. This is joint work with S.L. Lohr, C.A. Hamada, and T. Burr.

27. A Statistical Model for the Study of U-Nb Aging
Geralyn M. Hemphill, Los Alamos National Laboratory

In response to concerns that aging during long-term stockpile storage may change the microstructure and properties of U-6wt% Nb alloy components in ways that could adversely affect performance, a study was undertaken to model the aging response of two U-Nb alloys in order to quantify property and lifetime predictions and uncertainties. This presentation provides a brief background of the materials involved in the study, the mechanical property responses that were modeled, and information on the development of the semi-empirical model based on quasi-static, strain-rate tensile properties. An example of the resulting model and lifetime predictions for one of the properties is provided.
28. Statistical Comparison of Toolmarks in Forensics
Amy Hoeksema, Iowa State University

In forensics, fingerprints can be used to uniquely identify suspects in a crime. Similarly, a toolmark left at a crime scene can be used to identify the tool that was used. However, the current practice of identifying matching toolmarks involves visual inspection of marks which can be a very subjective process. As a result of this subjectivity, declared matches are often successfully challenged in court, so law enforcement agencies are particularly interested in encouraging research in more objective approaches. Our analysis is based on comparisons of profilometry data, essentially depth contours of a toolmark surface taken along a linear path. Through the use of a statistical algorithm that returns a measure of similarity for two specimens, we are working to develop criteria for a match. We are also enhancing the algorithm to model the angle at which a mark was made which will lead to angle determination for a toolmark left at a crime scene. With sufficient development, such methods may lead to more defensible forensic analyses.

29. Efficient Estimators for Sequential and Resolution-Limited Inverse Problems
Darren Homrighausen, Carnegie Mellon University

A common problem in the sciences is that a signal of interest is observed only indirectly, through smooth functionals of the signal whose values are then obscured by noise. In such inverse problems, the functionals dampen or entirely eliminate some of the signal’s interesting features. This makes it difficult or even impossible to fully reconstruct the signal, even without noise. In this paper, we develop methods for handling sequences of related inverse problems, with the problems varying either systematically or randomly over time. Such sequences often arise with automated data collection systems, like the data pipelines of large astronomical instruments such as the Large Synoptic Survey Telescope (LSST). The LSST will observe each patch of the sky many times over its lifetime under varying conditions. A possible additional complication in these problems is that the observational resolution is limited by the instrument, so that even with many repeated observations, only an approximation of the underlying signal can be reconstructed. We propose an efficient estimator for reconstructing a signal of interest given a sequence of related, resolution-limited inverse problems. We demonstrate our method’s effectiveness in some representative examples and provide theoretical support for its adoption.

30. Sensitivity of the Community Land Model (CLM4.0) to Key Modeling Parameters and Modeling of Key Physical Processes with Focus on the Arctic Environment
Elena Kalinina, Sandia National Laboratories

The purpose of this sensitivity study was to identify the major parameters and physical processes that have greatest impacts on the near the surface energy balance and consequently may drive climate changes in Arctic environment. We used Community Climate Model (CAM) to generate atmospheric data forcing files for historic calibration period and subsequent 100 years. These data provided an input into the Community Land Model (CLM 4.0). We selected a CLM point grid cell near Fairbanks, Alaska for this study. We considered a number of parameters in our sensitivity analysis, including hydrogeologic and thermal soil properties and vegetation characteristics. We modified a current approach used in CLM to simulate soil moisture to allow for more realistic water table representation and we investigated how this may affect the temporal distribution of soil moisture and ice. We classified the land model feedbacks to the climate model based on the correlation coefficients between the energy constituents and state variables. The work is underway to include in this analysis human impacts, such as water withdrawal and agricultural activities. This is joint work with William Peplinski, Vincent Tidwell, and David Hart.
31. Evaluating the Likelihood of a Fire-Following-An-Earthquake Scenario
Elizabeth J Kelly, Los Alamos National Laboratory

Department of Energy Standard DOE-STD-3009, *Preparation Guide for U.S. Department of Energy Nonreactor Nuclear Facility Documented Safety Analyses*, directs that earthquake induced fires be evaluated for non-reactor nuclear facilities. One challenge is to determine a conservative fire scenario, but not one that is so overly conservative that it is not useful. This talk describes a statistical approach to evaluate the likelihood of a fire-after-earthquake scenario. The approach is not meant to be exact, but to provide a context for assessing and comparing the conservatism of various fire scenarios. The first component of the approach is a statistical model to predict the number of ignitions following an earthquake event. This model is based on data for ignitions following earthquakes from 1906 to 1989 in Alaska and California. The data contain estimates for the mean peak ground acceleration (PGA) for each earthquake, an estimate of the built area affected in million square feet (MMSF) for each earthquake, and the number of ignitions within the estimated affected area (IGNS). The statistical model uses negative binomial regression to estimate the expected number of ignitions as a function of the explanatory variables, PGA and MMSF. To assure conservatism, the associated upper confidence and prediction limits are derived from the statistical model. An example of an application of this approach to a large single structure is presented as well as methods for adjusting results for possible under reporting of ignitions in the data. This is joint work with Raymond N Tell.

32. Uncertainty Quantification of 3D Geophysical Models
Carène Larmat, Monica Maceira, Los Alamos National Laboratory

For the last decade, several research institutions have been addressing the Earth’s 3D heterogeneities and complexities by improving tomographic methods. Utilizing dense array datasets, these efforts have led to 3D seismic models with the best resolution thus far, but little is done to provide any absolute assessment of the model uncertainty. The question of “How good is a model at representing the Earth’s true physics?“ remains largely not addressed in a time where 3D Earth models are used for national and security missions. We believe that the current computational power makes it now possible to tackle this important scientific question. Our proposal is based on the use of methods lately developed in statistical sciences (especially stochastic methods such as Monte-Carlo Markov Chain) combined with state-of-the-art methods developed in geosciences to accurately model the full waveform propagation in heterogeneous media. We will also revisit the problem of data gap and real resolution of 3D models with new numerical methods coming from the framework of adjoint problems.

We will show preliminary results from an on-going study on 3D geophysical models validation. We tested 3D tomographic models for the Western USA (DNA models) generated using both ray-theoretical and finite-frequency methods in order to get insight into the merit of these two imaging techniques. The model performances are evaluated by comparing synthetic seismograms and observed data. An accurate but computationally costly numerical method based on the Spectral Element Method (SEM) was used to generate the seismograms. Different measurement techniques for different parts of the signal were used. Preliminary results for DNA09 and four moderate-size events show a difference in performance between models obtained with the finite-frequency or ray-theory limited to smallest periods (<15s), with no perceptible difference at longer periods (50-200s). For the four events considered and for the period band between 50-200s, both finite-frequency and ray-theoretical DNA09 models predict the observations well. Data and synthetic seismograms fit the best for the horizontal components and smaller epicentral distances; the fit degrades as we go farther from the event. We are performing more research to interpret these results by looking for systematic path or radiation pattern effects. This is joint work with P. Loxley and Y. Kato.

33. Exploring Large Scale Scientific Data Using Information Theory
Teng-Yok Lee, The Ohio State University

Information theory was proposed by Claude E. Shannon in 1940s to quantitatively model the transmission of information through communication channel among data sources. Since then, its applications have extended from communication to different domains, including image processing, computer graphics, and visualization. This poster provides an overview of our ongoing project on integrating information theory with in situ analysis and visualization of large scale scientific data. We are in the process of developing a software library called Information Theory Library (ITL) which provides kernels for computing various fundamental information theory based metrics, such as Shannon’s entropy, conditional entropy and mutual information. This poster summarizes the software architecture of ITL for different types of scientific data, efficient parallelization of the computation of the metrics for large scale data, and the design of a general program interface for various types of in situ data analysis. The performance of ITL on supercomputers is also presented to demonstrate its capability of dealing with large data in distributed environment. In addition, this poster also presents our ongoing research direction toward entropy-driven data organization for large scale vector and scalar data. This is the joint work with Abon Chaudhuri, Han-Wei Shen, Tom Peterka, Cong Wang, Tiantian Xu, Bo Zhou, and Yi-Jen Chiang.
This research explores applications of wavelet analysis to solution monitoring (SM) for nuclear safeguards. SM data includes time series of levels and volumes from tanks at a nuclear facility. Large changes in these measurements are assumed to occur primarily during shipment or receipt events, which are monitored as one component of a system to detect nuclear material diversion. Three goals for SM evaluation systems are to smooth the measurement errors, detect change points that indicate start and stop times of events, and estimate signal changes during events.

Wavelets can be used to estimate a function by expanding it into a wavelet series. These wavelet transformations retain significant features of the function because wavelets are localized in frequency and time. Wavelet change detection uses wavelet coefficients to identify the start or stop time and quantity of an event. One challenge in using wavelets for change detection or data smoothing is the Gibbs phenomenon, which is the tendency of the wavelet-based reconstructed function to overshoot or undershoot the real value. This paper provides empirical evidence that the Haar wavelet basis function reduces the Gibbs effect seen in the reconstructed signal using other wavelets. An effective wavelet thresholding method for smoothing SM data is illustrated.

Real chaotic systems include the earths climate system, and the dynamics governing earthquakes. On initial examination, fluctuation dissipation like and embedology theorems from the theory of strange attractors suggest the usefulness of autoregressive models, but so far attempts to use such models have not met with much success with the harder problems such as prediction of regional precipitation.

In this paper we show that further examination of the relevant theory suggests some explanation why typically autoregressive models show at best intermittent prediction, and how to improve on that prediction. Examples from prediction of local precipitation and prediction of earthquake “climate” are given.

Techniques for automated feature extraction, including neuroscience-inspired machine vision, are of current interest in the areas of climate change impacts and Land Use/Land Cover classification using satellite image data. We describe an approach for automatic classification of land cover in multispectral satellite imagery of the Arctic using sparse representations over learned dictionaries. We demonstrate our method using DigitalGlobe Worldview-2 visible/near infrared high spatial resolution imagery. We use a Hebbian learning rule to build spectral-textural dictionaries that are adapted to the data. Sparse image representations of pixel patches over the learned dictionaries are used to perform unsupervised k-means clustering into land-cover categories. This approach combines spectral and spatial textural characteristics to detect geologic, vegetative, and hydrologic features. We compare our technique to standard remote sensing algorithms. Our results suggest that neuroscience-based models are a promising approach to practical pattern recognition problems in remote sensing, even for datasets using spectral bands not found in natural visual systems. This is joint work with Steven P. Brumby.

Consequences of an influenza outbreak and effects of factors determining mitigation strategies were studied by computer simulation using statistical design and analysis of experiments methods. Twelve input variables that specify differing degrees of disease mitigation were selected to be evaluated within the context of nine uncertain inputs that characterize influenza expression, e. g., reproductive number, case mortality rate, duration of infective period. Variables were characterized with probability distributions based on assessments of low, high, and typical values from expert opinion and literature review. An approach to experiment design called dispersion array based experiment was developed that incorporates orthogonal array based Latin hypercube sampling. The simulation experiment consisted of 128 combinations of mitigation factors each evaluated over 16 distinct disease characterizations and allowed a quick assessment of simulation results, particularly mitigation efficacy. Sensitivity evaluation of the mitigation factors showed effectiveness against the range of influenza expressions with little stress on the health care system. This is joint work with Dennis Powell, Jeanne Fair, Rene LeClaire, and Michael Wilson.
38. Scan Statistics for the Online Detection of Locally Anomalous Subgraphs
Josh Neil, Los Alamos National Laboratory

Identifying anomalies in computer networks is a challenging and complex problem. Often, anomalies occur in extremely local areas of the network. Locality is complex in this setting, since we have an underlying graph structure. To identify local anomalies, we introduce a scan statistic for data extracted from the edges of a graph over time. In the computer network setting, the data on these edges are multivariate measures of the communications between two distinct machines over time. We describe two shapes for capturing locality in the graph: the star and the k-path. Both of these shapes are motivated by hacker behaviors observed in real network attacks. To identify local anomalies, these shapes are enumerated over the entire graph, over a set of sliding time windows. Local statistics in each window are compared with their historic behavior to capture anomalies within the window. These local statistics are model-based, and we present two models motivated by network flow data to demonstrate the scanning procedure. Data speeds on larger networks require online detection to be nimble. We describe a full anomaly detection system that has been applied to a corporate-sized network, which achieves far better than real-time analysis speed. We present results on simulated data whose parameters were set to mimic data from Los Alamos National Laboratory’s (LANL’s) entire internal network (~ 20,000 hosts). In addition, we present a result from our analysis of an actual hacker attack on LANL’s internal network, which resulted in good quality detection of the malicious behavior.

39. Computational and Statistical Challenges in Cosmology
Adrian Pope, Argonne National Laboratory

The aim of cosmology is to infer what we can about the history and constituents of the universe from observations of the current state of objects in the night sky. The evolution of large-scale structures over cosmic time from tiny seed fluctuations is governed by relatively simple first-principles physics (e.g., gravity), but the non-linear nature of the eventual evolution necessitates a numerical approach to detailed predictions. The exponential growth of the number of pixels in the detectors in cameras on survey telescopes is driving a rapid increase in both the volume and quality of data from sky surveys. We can leverage a similar exponential growth in high performance computing power to provide theoretical predictions of equal quality. However, these improvements demand thoughtful choices in how we handle both observed and simulated datasets and in what statistical descriptions we use to compare them. We aim to learn more about the fundamental nature of dark matter and dark energy, the dominant components of the universe, via the solution of this cosmological inverse problem. We will present work related to several aspects of these challenges and attempt to prompt discussions about what statistical descriptions we might consider employing in the future.

40. Inverse Methods for Dimensionality Reduction
Juan Ramirez Jr., University of Colorado at Boulder

The analysis of structured high-dimensional data has been a research topic gaining interest in the engineering, computer science, statistics and mathematics communities. Methods have been developed to reduce the dimensionality of data in order to make computational operations feasible and extract the underlying structure in the data. These methods become relevant when attempting to processes large high-dimensional data sets for prediction, classification, or filtering. Processing high-dimensional data in the ambient space quickly becomes computationally expensive and often impossible. Until now inverse methods, outside of compressive sensing recovery, do not exist for linear and nonlinear dimensionality reduction techniques. In this work, we examine the problem of recovering a high-dimensional observation point from its low-dimensional embedding. In particular, we use the low-dimensional embedding to produce an out-of-sample interpolation point, and then recover its corresponding high-dimensional realization. We propose a least-square best-fit recovery method for the metric multidimensional scaling dimensionality reduction technique. We focus on synthetic data that exhibits a linear relationship between the high-dimensional observations and the low-dimensional latent variables. We present experiments on data drawn form statistical distributions and low-dimensional smooth manifolds. This is joint work with Francois G. Meyer.
41. Assessing NetworkPartitioning

David Robinson, Sandia National Laboratories

A number of recent research efforts have focused on identifying community structure in complex networks such as social and biological systems and the Worldwide web. For example, characterizing the community structure provides crucial insight into particular functional characteristics that may exist in metabolic networks. This paper outlines research related to the use of a hierarchical Bayesian approach for identifying the community structure and introduces the application of posterior predictive assessment methods to characterize the significance of the partitioning. The initial research goals are two-fold: confirm the underlying statistical modeling assumptions and assesses the statistical importance of the community structure predicted by the model. A long-term goal is to develop a metric that can be used to assess the performance of statistical network partitioning models. The current effort is focused on the construction of discrepancy functions based on mutual information between network edges and nodes. These discrepancy function is then used to identify nodes where the partitioning model has worked well and also those areas of the network where partitioning may be obscured by noise. They also provide metrics for the comparison of similarity between communities. The research is in the earliest stages and this paper provides an introduction to the current status.

42. Indago: A Novel Search and Analysis Tool for Electronically Stored or Transmitted Content

Jorge H Román, Los Alamos National Laboratory

As data generation and storage technologies have advanced, society itself has become increasingly reliant upon electronically generated and stored data. Digital content is proliferating faster than humans can consume it. Proliferation of electronic content requires automated tools to analyze large volumes of unstructured text. Network traffic as well as large corporate repositories need to be scanned for content of interest, either to stop the flow of unwanted information such as corporate secrets or to identify relevant documents for an area of interest. Network filters rely mostly on simple word list matches to identify “interesting” content, and searches rely on Boolean logic queries. Both approaches have their advantages and limitations. A keyword list is simple and can be implemented easily. Boolean logic with word proximity operators allows finer definition of target pattern of interest. However, both may retrieve too many false positives. A document may contain the right words, but not in the right context. For example, a web search for apple brings both references to the computer company and the fruit. We have developed an approach that searches for concepts-in-context with reduced number of false positives. Furthermore, by the use of commercial of the shelf hardware we have accelerated the process significantly so that the analysis can be done in near real time. Context-based search and analysis can greatly enhance the exploration of data and phenomena by essentially reducing the data deluge and increasing the efficiency and effectiveness of the human analyst and/or end-user to access and fully exploit the data-to-knowledge potential that is inherent but latent in nearly every collection. This is joint work with David DuBois, Shelly Spearing, Andrew DuBois, Mike Boorman, and Carolyn Connor.

43. Rhipe: R and Hadoop Integrated Programming Environment

Jeremiah Rounds, Purdue University

Use R to do intricate analysis of large data sets via Hadoop. Large complex data sets that can fill up several large hard drives (or more) are becoming commonplace, and many R using data analyst will have to confront that reality in the coming years. Data parallel distributed computing paradigms such as MapReduce have emerged as able tools to deal with large data, but until now very little has been done to put those paradigms into the hands of R users. Rhipe is a software package that allows the R user to create MapReduce jobs that work entirely within the R environment using R expressions. This integration with R is a trans-formative change to MapReduce; it allows an analyst to quickly specify Maps and Reduces using the full power, flexibility, and expressiveness of the R interpreted language. More information about Rhipe is available at www.rhipe.org.

In our use of Rhipe, we are developing the concept of Divide and Recombine (D&R) for the analysis of large complex data. In D&R data are divided into subsets in one or more ways, usually by using conditioning variables to partition the data into subsets. Numeric and visualization methods are applied to each of the subsets separately. Then the results of each method are recombined across subsets. By introducing and exploiting parallelization of data, D&R using Rhipe succeeds in making it possible to apply to large complex data almost any existing analysis method already available in R.

This poster outlines these ideas and presents current performance evaluations for D&R data analyses implemented on the Rhipe software system.
44. Analysis of Biothreat Detection Technologies
Ryan Roundy, Brigham Young University

Biothreat detector testing is primarily laboratory based with assessments of technology based on ability to detect agent concentrations with reasonable probability of detection and false alarm rates. Our approach provides innovative mathematical modeling in the form of flexible Gaussian process models for binary responses (as most testing conducted in laboratory, open air, and tunnel settings are binary responses) as the primary analysis tool. Furthermore, we propose a Bayesian hierarchical formulation of this model. The Gaussian process formulation based on probit link functions allows borrowing of strength across the various modalities of testing and across the statistical models.

45. Scalable Analysis of Network Measurements with Hadoop and Pig
Taghrid Samak, Lawrence Berkeley National Laboratory

The deployment of ubiquitous distributed monitoring infrastructure such as perfSONAR is greatly increasing the availability and quality of network performance data. Cross-cutting analyses are now possible that can detect anomalies and provide real-time automated alerts to network management services. However, scaling these analyses to the volumes of available data remains a difficult task. Although there significant research into offline analysis techniques, most of these approaches do not address the systems and scalability issues. This work presents an analysis framework incorporating industry best-practices and tools to perform large-scale analyses. Our framework integrates the expressiveness of Pig, the scalability of Hadoop, and the analysis and visualization capabilities of R to achieve a significant increase in both speed and power of analysis. Evaluation of our framework on a large dataset of real measurements from perfSONAR demonstrate a large speedup and novel statistical capabilities.

46. Topologies of the Conditional Ancestral Trees and Full Likelihood-based Inference in the General Coalescent Tree Framework
Ori Sargsyn, Los Alamos National Laboratory

The general coalescent tree framework represents a family of models for ancestries of random samples of DNA sequences at a non-recombining locus. Many ancestral models derived under various evolutionary scenarios fit into this framework. Using this framework and the infinite-sites model for mutations in DNA sequences, this paper presents a computationally tractable full likelihood-based inference method for neutral polymorphism data in a sample of DNA sequences. First, an exact sampling scheme is developed for the topologies of the conditional ancestral trees. Because this scheme has some computational limitations, a second scheme is developed based on an importance sampling. These sampling schemes are combined with Monte Carlo integrations for estimating the likelihood of the full polymorphism data, the ages of mutations in the sample, and the time of the most recent common ancestor of the sample. This paper shows, for example, how to estimate the likelihood of neutral polymorphism data in samples of DNA sequences that are completely linked to a mutant allele of interest, which could be neutral or under selection. The method is illustrated on the data in a sample of DNA sequences at the APOE gene locus.

47. Estimating the Central Direction of Random Rotations in SO(3)
Bryan Stanfill, Iowa State University

Random rotations in SO(3) (the space of 3x3 orthogonal matrices with determinant 1) are fundamental in representing the orientation and motion of three-dimensional objects in fields such as computer science, kinesiology or material science for example. Currently, various estimators of the central (mean) direction have been proposed in the literature but little attention has been paid to comparing these estimators with respect to desirable statistical properties such as bias and sampling variability. In this paper we compare four favorable estimators across three commonly used distributional models from the family of uniform-axis random spin distributions. Three of these estimators have previously been studied in the literature whereas we adapt Fishers’ median, originally proposed for spherical data, to rotation matrices for the first time. Through the means of an extensive simulation study we show that no estimator performs optimal across all distributions but rather that the performance of the estimators depends on underlying distribution characteristics.
Jon Stearley, Sandia National Laboratories

High performance computing (HPC) systems present unique problems for the detection and diagnostics of operational faults. Despite the presence of large system logs, the isolation to a particular root cause of job failure remains difficult. With rapid changes in job layout and communication paths, effective diagnostic algorithms are elusive. We present an alternative approach in which potential fault locations are identified automatically, within some credibility interval, and as more jobs are processed, the uncertainty bounds shrink to provide an increasing focus on the fault location. The information can then be used to schedule jobs so as to minimize the likelihood of failure and also schedule maintenance actions more accurately. A Bayesian statistical approach is proposed that accounts for the complex interaction of HPC jobs and hardware, via modeling the system as a graph of connected components. Our methods seek to integrate the distinct perspectives of users and operators, starting here with job pass/fail and system structure, and as part of our on-going research effort, adding data from physical sensors and unstructured text from system logs. The mathematical structure of the statistical algorithm is presented using data from a simulated HPC system. A comparison of the Bayesian approach and a traditional maximum likelihood is also discussed. This is joint work with David Robinson.

49. Scientist Attrition at LANL: A Look over Time, Using Survival Analysis Methods
Blair O. Stephenson, Los Alamos National Laboratory

Although not all attrition is necessarily adverse to an organization, the unexpected loss of employees is sometimes viewed as a negative to an organization (hiring costs, training, important skill sets). This exploratory study examines the attrition of scientists at LANL over nearly 3 decades. Specifically, survival analysis methods are utilized to model the voluntary (non-retirement) attrition of scientists. In the present context, the event of interest (voluntary non-retirement exit) may be viewed as competing with alternative modes of departure (e.g., exit due to retirement). Initially, we consider descriptive summary curves such as the survivor and hazard functions. To examine the influence of covariates, cause-specific Cox regression results are contrasted with a competing risks regression alternative. Also discussed are issues such as allowing for the inclusion of both invariant (e.g., gender) and time varying covariates (e.g., salary). Results suggest that a number of demographic factors are associated with a reduced risk of voluntary (non-retirement) departure from the organization. The influence of competing risks was found to be minimal.

50. Visualization at the Mesoscale
Richard Strelitz, Los Alamos National Laboratory

Decades of experience tell us that scientists see neither pixels nor images, but meaningful patterns derived from basic diagrams and fundamental solutions of underlying equations; the science then exists in the meso-scale. We propose to put visualization there too. The pattern paradigm we propose solves two problems of visualization: 1) the gulf between what a scientist sees and what they know, and 2) the increasing mismatch between tera-scale data size and megapixel screen size. Data sets may be large but are visibly full of dependency, in time, space and origin. We replace this redundancy with a rules, regression and residual. The patterns recognized by the scientist here become predictive formulations, to be applied to the data and tested for their explanatory power data (regression) and the remainder-the residual. Aggregating data into these parametric archetypes reduces the data pushed to the graphics card and makes the visualization scalable and resolution independent as encountered in AMR and multi-res grids. It changes analysis in visualization from a sole focus on values to one of scaling laws, spatial distribution and evolution, and leads to testable prediction, usable reductionism, and, we believe, a true chance for knowledge enhancing interactivity. Our goal is to help visualize science; by putting the science that ‘explains the data’ ahead of mere exposition of the data, we believe we are on the right track.

51. Spatio-Spectral Anomalous Change Detection in Hyperspectral Imagery
James Theiler, Los Alamos National Laboratory

Given two images of the same scene, taken at different times and under different conditions, the aim of anomalous change detection is to distinguish the pervasive differences (due for instance to illumination or calibration effects) from the actual changes that have occurred in the scene. This poster will present an approach for combining local spatial context with the spectral properties of a pixel for the purpose of identifying anomalous changes in pairs of images.
52. Uncertainty Analysis for GIS Estimates of CO2 Storage of the Oriskany Sandstone

Andrew C. Thomas, Carnegie Mellon University

The sequestration of carbon dioxide gas in deep saline-filled formations is potentially a major means for handling excess carbon dioxide production from fossil-fuel power plants. Whether this is a feasible long-term option depends largely on the ultimate storage capacity of geological formations, which we seek to estimate with appropriate considerations for uncertainty. Given data from a number of drill operators across Pennsylvania, we estimate the shape of the Oriskany sandstone layer, then calculate the storage capacity of the entire formation with respect to differing depth and thickness, as well as the accompanying changes in pressure and temperature. A Bayesian approach to thin-plate splines and geostatistical variation is used to provide an estimate of uncertainty on the total capacity. This is joint work with Bobak Karimi, William Harbert, Mitchell Small, Olga H. Popova, and Sean T. McCoy.

53. Analysis of Amplified Fragment Length Polymorphisms for Detection of Infectious Biological Agents

Lawrence Ticknor, Los Alamos National Laboratory

The detection of biological weapons or potentially infectious biological agents requires the identification of strains that have been identified previously and also strains that have not been seen before. Analysis techniques need to be able to alert decision makers to strains that are related to known agents. Amplified Fragment Length Polymorphisms (AFLP) provide genomic data about a strain's DNA and allow a reasoned analysis of the data to provide reliability for making decisions. We describe the development of an AFLP data analysis technique based on clustering that works with internal controls and replicates to show the usefulness of AFLP for determining related strains.

An AFLP dendrogram or ‘clustering tree’, which includes previously characterized pathogenic Bacillus anthracis, nonpathogenic Bacillus thuringiensis, and nonpathogenic Bacillus cereus was developed. We use this tree to illustrate three examples where AFLP has been used for understanding unknown Bacillus strains in comparison to pathogenic Bacillus anthracis. These examples include a Bacillus thuringiensis bio weapon strain from Iraq, an isolate from a Bosnian soldier infection that was thought similar to an organic pesticide, and Bacillus cereus strains that caused deaths in Texas welders. Trees for Clostridium, Yersinia, and Burkholderia have been developed but will not be presented. This is joint work with Karen Hill and Paul Jackson.

54. Automated Analysis of Large Numbers of Diffraction Datasets Using the Rietveld Method

Sven C. Vogel, Los Alamos National Laboratory

An increasing number of diffraction experiments are parametric or kinetics studies. Such experiments generate hundreds or even thousands of diffraction datasets, which pose substantial challenges to the data analysis. To address these challenges and other issues, the gssalanguage was developed at LANSCE. The software provides input to and processes output from the GSAS Rietveld package for full pattern diffraction data analysis. Gsas language allows the development of scripts for the automatic evaluation of large numbers of data sets and provides documentation of the refinement strategies employed, thus fostering the development of efficient refinement strategies as well as information exchange between experienced and novice users. Use of the bash shell and standard Unix text-processing tools, available natively on Linux and Mac OSX platforms and via the free cygwin software on Windows systems, make this software platform independent. Application examples are given and opportunities for future enhancements, e.g. by statistical analysis of the difference curve of a refinement for user guidance with respect to the refinement strategy or quality control, are identified.

55. Derivation of QMU-like formalisms from Maximum Entropy Arguments

Timothy Wallstrom, Los Alamos National Laboratory

I show that the method of Quantification of Margins and Uncertainties (QMU) can be derived from an appropriate Bayesian Network using maximum entropy. The crucial assumption is that one only possesses information about the bounds of the distribution. If, instead, one only possesses information about the variance, the same argument leads to a different formalism, which is essentially a first order reliability theory. I argue, furthermore, that the latter assumption may be more realistic for many applications.
Brian Weaver, Los Alamos National Laboratory

When measurement systems are assessed, a standard is used to determine a system's bias, its precision, or other quality characteristics. The true value of a standard is not known perfectly and an error is associated with it. With most measurement devices, this error can be ignored and the standard's value is treated as known. For highly precise measurement systems, these errors associated with the standard must be accounted for. In this paper we will assess calorimeters (a highly precise system) by estimating a predictive distribution of a future measured value by accounting for both the measurement error from the calorimeter and the error in the standard.

57. Neuromorphic Algorithms for Classifying and Locating Objects in a Fixed Camera Video Feed
Dylan Paiton, Los Alamos National Laboratory

The detection and localization of objects belonging to semantically distinct categories (i.e. Cyclists, Cars, Persons) remains a difficult problem for computer vision systems applied to natural video sequences. Traditionally, computer vision systems have made predictions using a single processing strategy, such as motion detection, template matching, or bag-of-feature-based representations followed by a classification stage. Here, we explored the improvements that result from combining multiple processing streams into a single integrated system. As in the visual cortex, we combined multiple processing streams representing detected motion, color/texture, and shape/form. We represented the classifications performed by each stream as weighted points on the image and used a classic clustering algorithm, DBSCAN, to combine them into a single detection. Confidences were then derived from the different streams into a single measure. The incentive for our efforts was the DARPA NeoVision2 challenge, which required us to localize and label multiple objects in a fixed high-definition camera at Stanford University. Our results were evaluated using scoring software and ground-truths provided by DARPA. This is joint work with SP Brumby, GT Kenyon, GJ Kunde, KD Peterson, MI Ham, PF Schultz, JM Galbraith, and JS George.

58. Sparse Principal Subspace Estimation
Vincent Vu, Carnegie Mellon University

Principal components analysis (PCA) is a widely used technique for dimensionality reduction and data visualization. The main idea is to capture the principal modes of variation of the data within a subspace spanned by the leading eigenvectors of a population covariance matrix. However, when applied to high-dimensional data, PCA can provide inconsistent estimates of the population eigenvectors. Our proposal, sparse principal subspace estimation, builds on the main idea of PCA by imposing the following sparsity constraint: the principal modes of variation of the data can be captured by a subspace spanned by a smaller number of the original variables. Sparsity constraints are especially reasonable when the signal of interest has sparse representation in some basis. Importantly, they can make estimation feasible and enhance interpretability in high-dimensions. We back these claims with: 1) theoretical results establishing rates of convergence and optimality, and 2) practical examples including visualization with a sparse principal components biplot.

59. In-situ Exploration and Sampling of Computer Simulation Models
Joanne Wendelberger, Los Alamos National Laboratory

Computer simulation models are used in a variety of disciplines to investigate and predict the behavior of systems with various inputs and outputs. With increasingly large and complex problems and access to advances in computing, traditional black box approaches to studying simulation models may become bottle-necked by the fact that the ability to calculate has outstripped the rate at which information can be transmitted. This calls for innovative approaches for incorporating analysis and data selection directly into the computational framework. A collaborative effort between statisticians and computer scientists is underway to integrate statistical sampling and analysis approaches with the computational and visualization strategies employed by computer scientists to efficiently extract information and accelerate scientific discovery. This is joint work with K. Myers, P. Fasel, J. Ahrens, J. Woodring, and J. Patchett.
60. Data Challenges for Scientific Data Analysis and Synthesis: An Example
Arthur Wiedmer, University of California Berkeley

Environmental monitoring data at DoE sites is usually archived in databases that were created to make data storage easy. However, building a system geared towards answering scientific inquiries usually requires changes in the data model, quality control, as well as new interfaces to access the data. This poster will show the process for the ingest of the Savannah River Site environmental data and an example of scientific question for the Savannah River Site using the ASCEM data management system.

61. Automatic Detection of PV System Configuration
Matthew Williams, Shawn Kerrigan, Locus Energy

Distributed solar energy is rapidly growing, and with that growth there is increasing need for software tools to validate data collected on PV system configuration. Clean configuration data is the foundation for most fleet performance modeling analytics, but large-scale deployments’ data on location, orientation, and tilt are generally subject to human error. In our paper we will cover a number of techniques to automatically validate location, orientation and tilt data.

Location detection: Combining three methodologies to determine location is often accurate within 10 miles

1. **Solar noon:** By statistically determining a system’s solar noon, the system’s longitude can be extracted from astronomical calculations
2. **Neighbor correlation:** System location can be triangulated by correlating the system’s production with production data from a network of solar projects with known locations
3. **Weather & irradiance model simulations:** By simulating production using historical weather data and satellite-based irradiance models, an installed system’s location can be determined by comparing simulated vs. actual production.

Orientation and tilt detection: Combining two methodologies to determine orientation and tilt provides solid data-quality guardrails.

1. **Daily start, peak, and end times:** System orientation is indicated by the time difference between start to peak, and peak to end, of a typical system production profile
2. **Weather & irradiance model simulations:** Orientation and tilt can be found by comparing actual vs. simulated output from a search through potential orientation and tilt angles

Our paper will illustrate the concepts with case examples and more information on the technique’s accuracy.

62. Using the Ensemble Kalman Filter for Computer Model Calibration
Matt Pratola, Los Alamos National Laboratory

The Ensemble Kalman Filter (EKF) is a popular approach for state and parameter estimation in the data assimilation literature. Yet while the problems in model calibration are similar to those in data assimilation, the EKF has not received much attention in model calibration problems. In this poster, we review the EKF and demonstrate when it may be useful for calibration problems. We also develop a resampling approach to refine the EKF estimates and outline an optimal design strategy when the ability to sample field observations is limited.
Interactive graphics progress us beyond the limitations of static statistical displays, particularly for exploring data, and analyzing models. A long missing feature in R graphics systems (either base or grid graphics) is support for interactivity. A number of standalone systems exist, GGobi, MANET and Mondrian, which support interactive displays of multivariate data, but they lack the extensibility, and tight integration with modeling that R furnishes. There have been several attempts to provide interactive graphics in R, from locate() (ages ago!) and getGraphicsEvent() in base R, to packages RGtk, RGtk2, rggobi, tcltk and iplots. Iplots, the most recent, provides extensive interactive graphics, using the Java backend. It is fast and includes most common types of plots; basic operations in interactive graphics such as selection and zooming are also supported. However, it lacks some features such as the tour or direct support for color palettes.

We have developed a new R package, cranvas, which is based on several other packages to make fast (Qt), flexible (R) and elegant interactive statistical graphics in R. The drawing is based on two packages qbase and qtpaint, which provide API's to the Qt libraries in R. The data structure is based on plumbr and objectSignals, which bring forward a new data structure called "mutable data". These packages set up signals and listeners on data (implemented purely with R), so that changes in data can trigger changes in plots. The mutable data is also the foundation of a "data pipeline" behind cranvas, where events of statistical analysis importance, like variable transformation or dimension reduction, can be propagated through to the displays. The cranvas package aims to borrow from the design of ggplot2, which is based on a grammar of graphics (Wilkinson, 2005). Currently this package includes common statistical graphics, histogram, scatterplot, bar plot, boxplot, parallel coordinate plot and map, and tours, and common interactions, brushing, identifying, deletion, zooming, panning and different types of linking. Color palettes are supported. This poster shows examples with the NRC rankings data, how the ideas relate to other interactive graphics work, what is under the hood, and what we are planning in the future.

There are increased concerns in health conditions of DOE security workers (guards and patrolmen) because they are potentially exposed to a broad range of occupational hazards such as chemical, radiation and physical exertion. Data from DOE Illness and Injury Surveillance Program (IISP) managed by Office of Health, Safety and Security is used to assess health conditions among the security workers and compare with those among other workers. IISP has been collecting data from the participating DOE sites since 1990 and the goal is to monitor morbidity in, and assess the overall health of, the DOE work force to identify groups that may be at increased risk for occupational illnesses and injuries. More than 20 years of IISP data allow trend analyses and comparisons within and between DOE sites. However, common limitations of health surveillance data exist. For example, data quality, data capture rate and data submitting requirements etc. change by time and vary by sites. The trend analyses and comparisons using conventional measures such as illness and injury rates may not be appropriate and may produce misleading results. Therefore, a rank approach is explored and a set of rank-based analytical methods are developed to assess health conditions among DOE security workers.

Modeling system reliability over time when binary data are collected both at the system and component level has been the subject of many papers. In a series system, it is often assumed that component reliability is linear in time through some link function. Often little or no prior information exists on the parameters of the linear regression, and in a Bayesian analysis they are modeled using very diffuse priors. This can have unintended consequences for the analysis, specifically in the prediction of system reliability. In this work, we consider negative log-gamma distributions as means of specifying prior information on reliability. We first show how our method can be implemented in modeling series system reliability over time and then introduce the idea of component selection as a mean of system size reduction in the modeling exercise.
<table>
<thead>
<tr>
<th>Name</th>
<th>Affiliation</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kayode S. Adekeye</td>
<td>Redeemer's University</td>
<td><a href="mailto:samadek_2017@yahoo.co.uk">samadek_2017@yahoo.co.uk</a></td>
</tr>
<tr>
<td>Deb Agarwal</td>
<td>Lawrence Berkeley National Laboratory</td>
<td><a href="mailto:daagarwal@lbl.gov">daagarwal@lbl.gov</a></td>
</tr>
<tr>
<td>James Albano</td>
<td>Rochester Institute of Technology</td>
<td><a href="mailto:jaa8397@rit.edu">jaa8397@rit.edu</a></td>
</tr>
<tr>
<td>Francis Alexander</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:fja@lanl.gov">fja@lanl.gov</a></td>
</tr>
<tr>
<td>Mohanad Alkhasawneh</td>
<td>Qatar University</td>
<td><a href="mailto:alkhasawneh@qu.edu.qa">alkhasawneh@qu.edu.qa</a></td>
</tr>
<tr>
<td>Hubert Allen</td>
<td>Hubert Allen and Associates</td>
<td><a href="mailto:hubertallen@comcast.net">hubertallen@comcast.net</a></td>
</tr>
<tr>
<td>Christine Anderson-Cook</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:candcook@lanl.gov">candcook@lanl.gov</a></td>
</tr>
<tr>
<td>Robert Arn</td>
<td>Colorado State University</td>
<td><a href="mailto:robert.arm@colorado.edu">robert.arm@colorado.edu</a></td>
</tr>
<tr>
<td>David Banks</td>
<td>Duke University</td>
<td><a href="mailto:banks@stat.duke.edu">banks@stat.duke.edu</a></td>
</tr>
<tr>
<td>Candace Berrett</td>
<td>Brigham Young University</td>
<td><a href="mailto:cberrett@stat.byu.edu">cberrett@stat.byu.edu</a></td>
</tr>
<tr>
<td>Jonathan Berry</td>
<td>Sandia National Laboratories</td>
<td><a href="mailto:jberry@sandia.gov">jberry@sandia.gov</a></td>
</tr>
<tr>
<td>K. Sham Bhat</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:bhat9999@lanl.gov">bhat9999@lanl.gov</a></td>
</tr>
<tr>
<td>David Biagioni</td>
<td>University of Colorado, Boulder / National Renewable Energy Laboratory</td>
<td><a href="mailto:biagioni@colorado.edu">biagioni@colorado.edu</a></td>
</tr>
<tr>
<td>Derek Bingham</td>
<td>Simon Fraser University</td>
<td><a href="mailto:dbingham@stat.sfu.ca">dbingham@stat.sfu.ca</a></td>
</tr>
<tr>
<td>Tan Bui</td>
<td>ICES, The University of Texas, Austin</td>
<td><a href="mailto:tanbui@ices.utexas.edu">tanbui@ices.utexas.edu</a></td>
</tr>
<tr>
<td>Vera Bulaevskaya</td>
<td>Lawrence Livermore National Laboratory</td>
<td><a href="mailto:vulaevs@yahoo.com">vulaevs@yahoo.com</a></td>
</tr>
<tr>
<td>Tom Burr</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:tburr@lanl.gov">tburr@lanl.gov</a></td>
</tr>
<tr>
<td>Leigh Cash</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:lcash@lanl.gov">lcash@lanl.gov</a></td>
</tr>
<tr>
<td>Emily Casleton</td>
<td>Iowa State University</td>
<td><a href="mailto:casleton@iastate.edu">casleton@iastate.edu</a></td>
</tr>
<tr>
<td>Charles Nakleh</td>
<td>Sandia National Laboratories</td>
<td><a href="mailto:cnakleh@sandia.gov">cnakleh@sandia.gov</a></td>
</tr>
<tr>
<td>Wei-Chen Chen</td>
<td>Oak Ridge National Laboratory</td>
<td><a href="mailto:wccsnow@gmail.com">wccsnow@gmail.com</a></td>
</tr>
<tr>
<td>Sofya Chepushtanova</td>
<td>Colorado State University</td>
<td><a href="mailto:svchepus@gmail.com">svchepus@gmail.com</a></td>
</tr>
<tr>
<td>Peter Chew</td>
<td>Galisteo Consulting Group, Inc.</td>
<td><a href="mailto:PeterAChew@aol.com">PeterAChew@aol.com</a></td>
</tr>
<tr>
<td>Hugh Chipman</td>
<td>Acadia University / Technometrics</td>
<td><a href="mailto:hugh.chipman@acadiau.ca">hugh.chipman@acadiau.ca</a></td>
</tr>
<tr>
<td>David Collins</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:dccollins@lanl.gov">dccollins@lanl.gov</a></td>
</tr>
<tr>
<td>Paul Constantine</td>
<td>Stanford University</td>
<td><a href="mailto:paul.constantine@stanford.edu">paul.constantine@stanford.edu</a></td>
</tr>
<tr>
<td>Eric Daub</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:edaub@lanl.gov">edaub@lanl.gov</a></td>
</tr>
<tr>
<td>Kim Denton-Hill</td>
<td>Sandia National Laboratories</td>
<td><a href="mailto:kmdenton@sandia.gov">kmdenton@sandia.gov</a></td>
</tr>
<tr>
<td>Jorge Diaz</td>
<td>University of Puerto Rico</td>
<td><a href="mailto:jdiaz_castro@hotmail.com">jdiaz_castro@hotmail.com</a></td>
</tr>
<tr>
<td>Thomas Donnelly</td>
<td>SAS Institute Inc.</td>
<td><a href="mailto:tom.donnelly@jmp.com">tom.donnelly@jmp.com</a></td>
</tr>
<tr>
<td>Ryan Elmore</td>
<td>National Renewable Energy Laboratory</td>
<td><a href="mailto:Ryan.Elmore@nrel.gov">Ryan.Elmore@nrel.gov</a></td>
</tr>
<tr>
<td>Matt Fair</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:mbfair@lanl.gov">mbfair@lanl.gov</a></td>
</tr>
<tr>
<td>H. Pearl Flath</td>
<td>The University of Texas at Austin</td>
<td><a href="mailto:pfflath@ices.utexas.edu">pfflath@ices.utexas.edu</a></td>
</tr>
<tr>
<td>Wilson Kwasi Fosu</td>
<td>Regent Univ. of Science and Tech., Ghana</td>
<td><a href="mailto:wfosu@opportunityghana.com">wfosu@opportunityghana.com</a></td>
</tr>
<tr>
<td>Devin Francom</td>
<td>Brigham Young University</td>
<td><a href="mailto:devinfrcsom@gmail.com">devinfrcsom@gmail.com</a></td>
</tr>
<tr>
<td>Mike Fugate</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:fugate@lanl.gov">fugate@lanl.gov</a></td>
</tr>
<tr>
<td>John Galbraith</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:jgalb@lanl.gov">jgalb@lanl.gov</a></td>
</tr>
<tr>
<td>Luke Gosink</td>
<td>Pacific Northwest National Laboratory</td>
<td><a href="mailto:luke.gosink@pnnl.gov">luke.gosink@pnnl.gov</a></td>
</tr>
<tr>
<td>Michael Grosskopf</td>
<td>University of Michigan</td>
<td><a href="mailto:mkgros@umich.edu">mkgros@umich.edu</a></td>
</tr>
<tr>
<td>Ryan Hafen</td>
<td>Pacific Northwest National Laboratory</td>
<td><a href="mailto:ryan.hafen@pnnl.gov">ryan.hafen@pnnl.gov</a></td>
</tr>
<tr>
<td>Aric Hagberg</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:hagberg@lanl.gov">hagberg@lanl.gov</a></td>
</tr>
<tr>
<td>Christina Hamada</td>
<td>University of New Mexico, Los Alamos / Los Alamos National Laboratory</td>
<td><a href="mailto:chamada@lanl.gov">chamada@lanl.gov</a></td>
</tr>
<tr>
<td>Michael Hamada</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:hamada@lanl.gov">hamada@lanl.gov</a></td>
</tr>
<tr>
<td>John Hathaway</td>
<td>Pacific Northwest National Laboratory</td>
<td><a href="mailto:john.hathaway@pnnl.gov">john.hathaway@pnnl.gov</a></td>
</tr>
<tr>
<td>Katrin Heitmann</td>
<td>Argonne National Laboratory</td>
<td><a href="mailto:heitmann@anl.gov">heitmann@anl.gov</a></td>
</tr>
<tr>
<td>Geraldyn Hemphill</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:gsh@lanl.gov">gsh@lanl.gov</a></td>
</tr>
<tr>
<td>Dave Higdon</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:dhigdon@lanl.gov">dhigdon@lanl.gov</a></td>
</tr>
<tr>
<td>Amy Hoeksema</td>
<td>Iowa State University</td>
<td><a href="mailto:ahoek01@iastate.edu">ahoek01@iastate.edu</a></td>
</tr>
<tr>
<td>James Holloway</td>
<td>University of Michigan</td>
<td><a href="mailto:hagar@umich.edu">hagar@umich.edu</a></td>
</tr>
<tr>
<td>Darren Homrighausen</td>
<td>Carnegie Mellon University</td>
<td><a href="mailto:dhomrigh@stat.cmu.edu">dhomrigh@stat.cmu.edu</a></td>
</tr>
<tr>
<td>Gardar Johannesson</td>
<td>Lawrence Livermore National Laboratory</td>
<td><a href="mailto:johannesson1@llnl.gov">johannesson1@llnl.gov</a></td>
</tr>
<tr>
<td>Elena Kalinina</td>
<td>Sandia National Laboratories</td>
<td><a href="mailto:eakalin@sandia.gov">eakalin@sandia.gov</a></td>
</tr>
<tr>
<td>Takeaki Kariya</td>
<td>Meiji University</td>
<td><a href="mailto:kariya@kisc.meiji.ac.jp">kariya@kisc.meiji.ac.jp</a></td>
</tr>
<tr>
<td>Elizabeth Kelly</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:ekelly@lanl.gov">ekelly@lanl.gov</a></td>
</tr>
<tr>
<td>Name</td>
<td>Affiliation</td>
<td>Email</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------------------</td>
<td>------------------------------</td>
</tr>
<tr>
<td>Paul Kidwell</td>
<td>Lawrence Livermore National Laboratory</td>
<td><a href="mailto:kidwellpaul@gmail.com">kidwellpaul@gmail.com</a></td>
</tr>
<tr>
<td>Wendy Kisch</td>
<td>Iowa State University</td>
<td><a href="mailto:wikisch@iastate.edu">wikisch@iastate.edu</a></td>
</tr>
<tr>
<td>Tammy Kolda</td>
<td>Sandia National Laboratories, California</td>
<td><a href="mailto:tkgolda@sandia.gov">tkgolda@sandia.gov</a></td>
</tr>
<tr>
<td>Carene Larmat</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:carene@lanl.gov">carene@lanl.gov</a></td>
</tr>
<tr>
<td>Earl Lawrence</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:earl@lanl.gov">earl@lanl.gov</a></td>
</tr>
<tr>
<td>Teng-Yok Lee</td>
<td>The Ohio State University</td>
<td><a href="mailto:leeten@cse.ohio-state.edu">leeten@cse.ohio-state.edu</a></td>
</tr>
<tr>
<td>Nathan Lemons</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:nlemons@lanl.gov">nlemons@lanl.gov</a></td>
</tr>
<tr>
<td>Kristin Lennox</td>
<td>Lawrence Livermore National Laboratory</td>
<td><a href="mailto:lennox3@llnl.gov">lennox3@llnl.gov</a></td>
</tr>
<tr>
<td>Rebecca Lilley</td>
<td>University of New Mexico</td>
<td><a href="mailto:becs921@gmail.com">becs921@gmail.com</a></td>
</tr>
<tr>
<td>Jason Loepky</td>
<td>University of British Columbia, Okanagan</td>
<td><a href="mailto:jason@stat.ubc.ca">jason@stat.ubc.ca</a></td>
</tr>
<tr>
<td>Claire Longo</td>
<td>University of New Mexico</td>
<td><a href="mailto:clongo01@unm.edu">clongo01@unm.edu</a></td>
</tr>
<tr>
<td>Michael LuValle</td>
<td>OFS Laboratories</td>
<td><a href="mailto:mj@mjsoptics.com">mj@mjsoptics.com</a></td>
</tr>
<tr>
<td>Monica Maceira</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:mmaceira@lanl.gov">mmaceira@lanl.gov</a></td>
</tr>
<tr>
<td>David Messinger</td>
<td>Rochester Institute of Technology</td>
<td><a href="mailto:messinger@cis.rit.edu">messinger@cis.rit.edu</a></td>
</tr>
<tr>
<td>Sarah Michalak</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:michalak@lanl.gov">michalak@lanl.gov</a></td>
</tr>
<tr>
<td>Angela Mielke</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:amielke@lanl.gov">amielke@lanl.gov</a></td>
</tr>
<tr>
<td>Nadine Miner</td>
<td>Sandia National Laboratories</td>
<td><a href="mailto:neminer@sandia.gov">neminer@sandia.gov</a></td>
</tr>
<tr>
<td>Daniela Moody</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:damoody@lanl.gov">damoody@lanl.gov</a></td>
</tr>
<tr>
<td>Lisa Moore</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:lmoore@lanl.gov">lmoore@lanl.gov</a></td>
</tr>
<tr>
<td>Max Morris</td>
<td>Iowa State University</td>
<td><a href="mailto:mmorris@iastate.edu">mmorris@iastate.edu</a></td>
</tr>
<tr>
<td>Kary Myers</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:karymyers@gmail.com">karymyers@gmail.com</a></td>
</tr>
<tr>
<td>Joshua Neil</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:jneil@lanl.gov">jneil@lanl.gov</a></td>
</tr>
<tr>
<td>Dan Nordman</td>
<td>Iowa State University</td>
<td><a href="mailto:dnordman@iastate.edu">dnordman@iastate.edu</a></td>
</tr>
<tr>
<td>Edward Ochoa</td>
<td>Sandia National Laboratories</td>
<td><a href="mailto:emochoa@sandia.gov">emochoa@sandia.gov</a></td>
</tr>
<tr>
<td>Chris Oehmen</td>
<td>Pacific Northwest National Laboratory</td>
<td><a href="mailto:Christopher.Oehmen@pnl.gov">Christopher.Oehmen@pnl.gov</a></td>
</tr>
<tr>
<td>George Ostrouchov</td>
<td>Oak Ridge National Laboratory</td>
<td><a href="mailto:ostrouchovg@ornl.gov">ostrouchovg@ornl.gov</a></td>
</tr>
<tr>
<td>Christopher Paciorek</td>
<td>University of California, Berkeley</td>
<td><a href="mailto:paciorek@stat.berkeley.edu">paciorek@stat.berkeley.edu</a></td>
</tr>
<tr>
<td>Dylan Paiton</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:dpaiton@gmail.com">dpaiton@gmail.com</a></td>
</tr>
<tr>
<td>Tom Peterka</td>
<td>Argonne National Laboratory</td>
<td><a href="mailto:tpeterka@mcs.anl.gov">tpeterka@mcs.anl.gov</a></td>
</tr>
<tr>
<td>Cindy Phillips</td>
<td>Sandia National Laboratories</td>
<td><a href="mailto:caphill@sandia.gov">caphill@sandia.gov</a></td>
</tr>
<tr>
<td>Rick Picard</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:picard@lanl.gov">picard@lanl.gov</a></td>
</tr>
<tr>
<td>Adrian Pope</td>
<td>Argonne National Laboratory</td>
<td><a href="mailto:apope@anl.gov">apope@anl.gov</a></td>
</tr>
<tr>
<td>Matthew Pratola</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:mpratola@lanl.gov">mpratola@lanl.gov</a></td>
</tr>
<tr>
<td>Juan Ramirez Jr.</td>
<td>University of Colorado, Boulder</td>
<td><a href="mailto:juan.ramirez@colorado.edu">juan.ramirez@colorado.edu</a></td>
</tr>
<tr>
<td>Jaideep Ray</td>
<td>Sandia National Laboratories, California</td>
<td><a href="mailto:jairay@sandia.gov">jairay@sandia.gov</a></td>
</tr>
<tr>
<td>Shane Reese</td>
<td>Brigham Young University</td>
<td><a href="mailto:reese@stat.byu.edu">reese@stat.byu.edu</a></td>
</tr>
<tr>
<td>Scott Robbins</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:srobbins@lanl.gov">srobbins@lanl.gov</a></td>
</tr>
<tr>
<td>David Robinson</td>
<td>Sandia National Laboratories</td>
<td><a href="mailto:drobin@sandia.gov">drobin@sandia.gov</a></td>
</tr>
<tr>
<td>Jorge Roman</td>
<td>Los Alamos National Laboratories</td>
<td><a href="mailto:jhr@lanl.gov">jhr@lanl.gov</a></td>
</tr>
<tr>
<td>John Roskosnensky</td>
<td>Sandia National Laboratories</td>
<td><a href="mailto:jkrosko@sandia.gov">jkrosko@sandia.gov</a></td>
</tr>
<tr>
<td>Jeremiah Rounds</td>
<td>Purdue University</td>
<td><a href="mailto:roundsjeremiah@gmail.com">roundsjeremiah@gmail.com</a></td>
</tr>
<tr>
<td>Ryan Roundy</td>
<td>Brigham Young University</td>
<td><a href="mailto:roundrys@gmail.com">roundrys@gmail.com</a></td>
</tr>
<tr>
<td>Taghrid Samak</td>
<td>Lawrence Berkeley National Laboratory</td>
<td><a href="mailto:tsamak@lbl.gov">tsamak@lbl.gov</a></td>
</tr>
<tr>
<td>Prabuddha Sanyal</td>
<td>Sandia National Laboratories</td>
<td><a href="mailto:psanyal@sandia.gov">psanyal@sandia.gov</a></td>
</tr>
<tr>
<td>Ori Sargsyan</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:sargsyan@lanl.gov">sargsyan@lanl.gov</a></td>
</tr>
<tr>
<td>Adam Shipman</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:ashipman@lanl.gov">ashipman@lanl.gov</a></td>
</tr>
<tr>
<td>Bryan Stanfill</td>
<td>Iowa State University</td>
<td><a href="mailto:stanfill@iastate.edu">stanfill@iastate.edu</a></td>
</tr>
<tr>
<td>Jon Stearley</td>
<td>Sandia National Laboratories</td>
<td><a href="mailto:jrstebar@sandia.gov">jrstebar@sandia.gov</a></td>
</tr>
<tr>
<td>Blair Stephenson</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:bblairs@lanl.gov">bblairs@lanl.gov</a></td>
</tr>
<tr>
<td>Curtis Storlie</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:storlie@lanl.gov">storlie@lanl.gov</a></td>
</tr>
<tr>
<td>Richard Strelitz</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:strelitz@lanl.gov">strelitz@lanl.gov</a></td>
</tr>
<tr>
<td>Matt Taddy</td>
<td>Chicago Booth</td>
<td><a href="mailto:taddy@chicagobooth.edu">taddy@chicagobooth.edu</a></td>
</tr>
<tr>
<td>Mark Tardiff</td>
<td>Pacific Northwest National Laboratory</td>
<td><a href="mailto:mark.tardiff@pnnl.gov">mark.tardiff@pnnl.gov</a></td>
</tr>
<tr>
<td>James Theler</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:jt@lanl.gov">jt@lanl.gov</a></td>
</tr>
<tr>
<td>Andrew C. Thomas</td>
<td>Carnegie Mellon University</td>
<td><a href="mailto:acthomas@stat.cmu.edu">acthomas@stat.cmu.edu</a></td>
</tr>
<tr>
<td>Sandy Thompson</td>
<td>DOE NNSA / Pacific Northwest National Laboratory</td>
<td><a href="mailto:Sandra.thompson@nnsa.doe.gov">Sandra.thompson@nnsa.doe.gov</a></td>
</tr>
<tr>
<td>Lawrence Ticknor</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:lot@lanl.gov">lot@lanl.gov</a></td>
</tr>
<tr>
<td>Mark Van Benthem</td>
<td>Sandia National Laboratories</td>
<td><a href="mailto:mhvanbe@sandia.gov">mhvanbe@sandia.gov</a></td>
</tr>
<tr>
<td>Scott Vander Wiel</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:scottv@lanl.gov">scottv@lanl.gov</a></td>
</tr>
<tr>
<td>Name</td>
<td>Affiliation</td>
<td>Email</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Vincent Vu</td>
<td>Carnegie Mellon University</td>
<td><a href="mailto:vqv@stat.cmu.edu">vqv@stat.cmu.edu</a></td>
</tr>
<tr>
<td>Timothy Wallstrom</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:tcw@lanl.gov">tcw@lanl.gov</a></td>
</tr>
<tr>
<td>Graham Walsh</td>
<td>New Mexico Tech</td>
<td><a href="mailto:gwalsh@emrtc.nmt.edu">gwalsh@emrtc.nmt.edu</a></td>
</tr>
<tr>
<td>Brian Weaver</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:theguz@lanl.gov">theguz@lanl.gov</a></td>
</tr>
<tr>
<td>Joanne Wendelberger</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:joanne@lanl.gov">joanne@lanl.gov</a></td>
</tr>
<tr>
<td>Arthur Wiedmer</td>
<td>University of California, Berkeley /</td>
<td><a href="mailto:arthur.wiedmer@berkeley.edu">arthur.wiedmer@berkeley.edu</a></td>
</tr>
<tr>
<td></td>
<td>Lawrence Berkeley National Laboratory</td>
<td></td>
</tr>
<tr>
<td>Matthew Williams</td>
<td>Locus Energy</td>
<td><a href="mailto:matthew.williams@locusenergy.com">matthew.williams@locusenergy.com</a></td>
</tr>
<tr>
<td>Alyson Wilson</td>
<td>IDA Science and Technology Policy Institute</td>
<td><a href="mailto:awilson@ida.org">awilson@ida.org</a></td>
</tr>
<tr>
<td>Jon Woodring</td>
<td>Los Alamos National Laboratory</td>
<td><a href="mailto:woodring@lanl.gov">woodring@lanl.gov</a></td>
</tr>
<tr>
<td>Joey Zhou</td>
<td>Department of Energy</td>
<td><a href="mailto:joey.zhou@hq.doe.gov">joey.zhou@hq.doe.gov</a></td>
</tr>
<tr>
<td>Roger Zoh</td>
<td>Iowa State University</td>
<td><a href="mailto:rszoh8@gmail.com">rszoh8@gmail.com</a></td>
</tr>
</tbody>
</table>